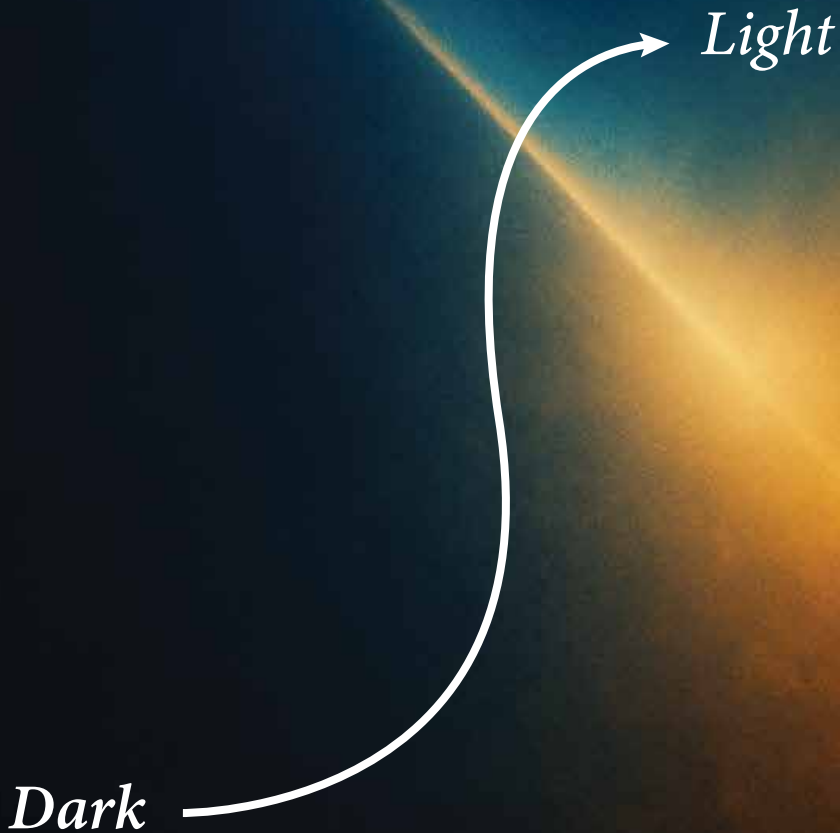
Explainable Vision Transformers with Domain Adaptation on Limited Datasets



Written by
Mohsin Ali

University of Essex 2025

University of Essex

School of Computer Science and
Electronic Engineering

Explainable Vision Transformers with Domain Adaptation on Limited Datasets

Mohsin Ali

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

Expected date of submission: 30 September 2025
Colchester

Cover Art Description

The cover art symbolises the journey undertaken in this thesis. The dark region represents the opacity and black-box nature of large Convolutional Neural Network (CNN) and Vision Transformers (ViTs) models, whose predictions are often difficult to interpret and trust. In contrast, the bright region represents explainability, reflecting the transparent and faithful reasoning that this thesis seeks to advance. The transition from dark to light illustrates the progression toward interpretable and efficient models, especially the lighter frameworks proposed here for small and imbalanced datasets. The interplay of orange and bluish energy flows conveys the power and adaptability of ViTs, highlighting their central role in this research. Finally, the curved arrow tracing across the design represents the trajectory of the thesis itself starting from opacity, vulnerability, and data inefficiency, and moving step by step toward robustness, interpretability, and trustworthy deployment. Together, these elements visualise the central message of this work: enabling a shift from dark to light in machine learning for medical imaging.

Dedication

This thesis is dedicated with deepest love and gratitude to my wife, Hira Ali, and my son, Moosa Ali, whose patience and sacrifices gave me the strength to complete this journey. Living apart during my PhD was one of the hardest challenges of my life, yet their unwavering support and belief in me carried me through every step.

I also dedicate this work to my parents and my teachers, whose devotion to my early education, the values they instilled in me, and their constant encouragement shaped me into the person I am today. Their guidance and support gave me the strength to pursue and complete a PhD, and every achievement I have reached rests upon the foundation they built. For this, I remain forever grateful.

Finally, I extend this dedication to my friends and extended family members, who stood by me in moments of struggle and shared in moments of joy. Their companionship and encouragement made this long journey lighter and far less difficult than it might otherwise have been.

اور بھی دکھ ھیں زمانے میں محبت کے سوا
راحتیں اور بھی ھیں وصل کی راحت کے سوا

فیض احمد فیض

Acknowledgements

I would like to sincerely thank my supervisor, Dr Haider Raza, for their invaluable guidance and constant support throughout this PhD journey. His encouragement extended far beyond academic advice; from mentoring me in research to helping me secure part-time opportunities, he played a central role in making this path manageable and successful.

I am equally grateful to Prof. John Q. Gan, whose invaluable feedback and guidance helped me refine and strengthen my academic writing skills. His input has been instrumental in shaping the quality of my research and publications.

I would also like to acknowledge Dr Muhammad Haris Khan, who served as a mentor throughout my doctoral studies. His advice, collaboration, and encouragement greatly enhanced my research skills and contributed significantly to my research output.

My sincere thanks also go to Dr Muhammad Atif Tahir, who, although not directly involved in this PhD, provided me with the opportunity to work in the Computer Vision Lab. That early exposure expanded my knowledge of computer vision and artificial intelligence and laid the foundation that eventually enabled me to pursue a PhD.

Abstract

Artificial Intelligence (AI) is now widely used to analyse images in areas such as healthcare, automotive, retail, manufacturing, and security. However, many of todays deep learning models act as black boxes: they can be highly accurate, but it is not clear how they make their decisions. This lack of transparency is an issue when decisions are safety-critical, for example, in medical diagnosis. Explainable AI (XAI) helps address this by showing which parts of an image influenced a models prediction, so that we can check whether it is looking at the right features.

This thesis focuses on improving Vision Transformers (ViTs), a powerful new type of model that looks at images in pieces (patches) and reasons about them using attention mechanisms. ViTs work well with large datasets, but they struggle when data is limited, which is often the case in the healthcare domain. ViTs are also vulnerable to adversarial attacks, where tiny invisible changes to an image can cause wrong predictions. To tackle these issues, four main contributions are made. First, a feature-map fusion method is introduced for Convolutional Neural Networks (CNNs), combining information from clean, noisy, and perturbed images to make models more robust. Second, two lightweight improvements to ViTs are proposed: the Summary Vision Transformer (S-ViT), which adds extra spatial information from a CNN, and the Multi-Gradient Image Transformer (MGiT), which stabilises training using an auxiliary transformer. Both methods improve performance on small and imbalanced datasets, such as skin lesion images (ISIC 2017) and COVID-19 chest X-rays. Third, XAI tools, including LIME, SHAP, Grad-CAM, and Attention Rollout, are used to confirm that these models focus on clinically meaningful regions. Finally, a new explanation method, FocusViT, is proposed to give sharper and more faithful explanations of ViT predictions.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AUC	Area Under the ROC Curve
BA	Balanced Accuracy
CAM	Class Activation Map
CNN	Convolutional Neural Network
DeiT	Data-efficient Image Transformer
FGSM	Fast Gradient Sign Method
F1	F1 Score (harmonic mean of precision and recall)
Grad-CAM	Gradient-weighted Class Activation Mapping
GPU	Graphics Processing Unit
JS Divergence	JensenShannon Divergence
LIME	Local Interpretable Model-agnostic Explanations
LRP	Layer-wise Relevance Propagation
MGiT	Multi-Gradient Image Transformer
PGD	Projected Gradient Descent
RAM	Random Access Memory
SHAP	SHapley Additive exPlanations
S-ViT	Summary Vision Transformer
ViT	Vision Transformer
XAI	Explainable Artificial Intelligence
AT	Adversarial Training
TRADES	TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimisation
GAIR AT	Geometry-Aware Instance Reweighted Adversarial Training

Chapter 1

Introduction

1.1 Background and Motivation

Artificial Intelligence (AI) refers to the study and development of computational systems that can perform tasks traditionally requiring human intelligence, such as perception, reasoning, and decision-making. Within AI, Machine Learning (ML) has emerged as a dominant paradigm in which models are not explicitly programmed with rules but instead learn patterns directly from data. ML algorithms adapt their parameters by minimising an objective function, enabling them to generalise from examples to unseen situations. Deep Learning (DL), a subfield of ML, employs deep neural networks with many layers to automatically learn hierarchical feature representations from raw data. This progression from AI to ML to DL has underpinned breakthroughs across domains such as computer vision, natural language processing, and speech recognition, where models trained on large datasets achieve performance surpassing traditional handcrafted approaches [1]–[3].

AI has moved into production across a wide range of applications, including computer vision, medical imaging, autonomous systems, security, and industrial inspection [2]–[4]. Deep learning models achieve high accuracy by learning rich hierarchical representations from large datasets, supported by steady growth in data availability, compute capacity, and scalable training [5]. At the same time, the increasing depth and architectural complexity of these models have widened the gap between performance and transparency [6].

In practice, many predictions are difficult to interpret, audit, or validate, especially for practitioners who must rely on these systems in operational settings.

This challenge makes trust and interpretability central requirements for high-stakes use. Deep neural networks are often described as black boxes because their internal computations are not directly understandable to human users. In clinical diagnosis, autonomous navigation, or safety monitoring, unexplained errors can carry serious consequences and slow deployment, certification, and post-incident analysis [7]. Explainable AI provides a practical route to address this problem by exposing which features or regions drive a prediction, so that users can check whether a model is focusing on signals that are relevant for the task, and to detect failure modes such as reliance on background artefacts or dataset shortcuts [8]. Explanations support validation and debugging, and inform model improvement, rather than treating interpretability as an afterthought.

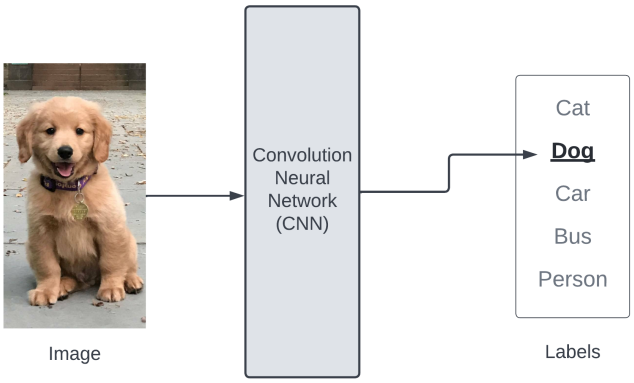


Figure 1.1: Typical image classification pipeline using a Convolutional Neural Network (CNN). An input image is transformed into feature representations and classified into one of several labels.

Despite recent progress, several challenges remain with current deep models. First, data scarcity is common in specialised domains such as rare diseases, privacy-sensitive clinical cohorts, and certain remote sensing settings, where curating large and well-annotated datasets is difficult [9]. Second, models are vulnerable to adversarial attacks: small but carefully crafted input perturbations can change predictions, which reveals brittle decision boundaries and amplifies safety concerns [10]. Third, Vision Transformers (ViTs),

while strong at capturing global context through self-attention [11], do not inherit the inductive biases of convolutional networks such as locality and translation equivariance; as a result, they generally rely more on large-scale pre-training and heavy regularisation, especially on small datasets [12]. Without these biases, transformers can overfit or learn shortcuts when training data are limited. Addressing these issues requires methods that improve data efficiency and robustness while keeping model behaviour interpretable, so that performance gains translate into trustworthy and deployable systems.

1.2 Computer Vision: From Handcrafted Features to ViTs

Computer vision has evolved from handcrafted feature design to end-to-end trainable deep neural architectures that automatically learn hierarchical representations of visual data. Early systems relied on manually engineered kernels such as edge detectors (Sobel, Gabor, HOG), which captured low-level features but lacked adaptability across tasks [13], [14]. The emergence of statistical learning theory and probabilistic models provided a foundation for more flexible pattern recognition approaches [1].

1.2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) represented a major breakthrough in visual representation learning by introducing architectures that exploit local connectivity and weight sharing to process image data efficiently [15]. Inspired by the organisation of the visual cortex, CNNs learn hierarchical feature representations in which early layers capture low-level structures such as edges and textures, while deeper layers encode increasingly abstract and task-specific concepts. Pooling and downsampling operations further provide a degree of spatial invariance, enabling robust recognition under moderate geometric variation.

The effectiveness of CNNs was demonstrated at scale by the success of AlexNet on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) in 2012 [3], which signifi-

cantly reduced classification error through the combined use of rectified linear units, regularisation techniques, and GPU acceleration. Subsequent architectures, including VGG, ResNet, and EfficientNet, refined these principles to improve depth, optimisation stability, and computational efficiency [4], [16]. Beyond static images, CNN-based representations have also been extended to sequential and multimodal tasks through integration with recurrent and temporal modelling components [17].

1.2.2 Generative Models for Visual Data

Alongside discriminative models, generative modelling has played an important role in computer vision by enabling the synthesis, completion, and transformation of visual data. Early deep generative approaches include Variational Autoencoders (VAEs), which provide probabilistic latent representations, and autoregressive models that factorise image generation into sequential prediction tasks. More recently, diffusion-based models have emerged as a dominant paradigm, achieving state-of-the-art sample quality and training stability across a range of vision tasks.

Generative Adversarial Networks (GANs) constitute one influential class of generative models, introducing an adversarial min–max training objective in which a generator and discriminator are trained jointly [18]. GANs enabled high-fidelity image synthesis and have been widely adopted for tasks such as data augmentation, image-to-image translation, and visual simulation. Extensions such as conditional GANs allow controllable generation [19], while architectures including StyleGAN demonstrated impressive photorealism through structured latent spaces [20]. GAN-based approaches have also been explored in medical imaging to augment limited datasets and simulate imaging modalities such as MRI and CT [21].

Training instability, mode collapse, and sensitivity to hyperparameters remain persistent challenges [22] in GANs. In contrast, diffusion-based and likelihood-based models offer more stable optimisation and improved coverage of the data distribution, making them increasingly preferred in both research and applied settings.

1.2.3 Vision Transformers

While CNNs are efficient at learning local spatial patterns, they embed strong inductive biases such as translation equivariance and locality, which can limit global reasoning. The introduction of ViTs shifted the paradigm by treating an image as a sequence of patches and applying multi-head self-attention to model long-range dependencies [11]. Unlike CNNs, ViTs explicitly capture global context but require large-scale training data to offset their weaker inductive biases. Variants such as DeiT and Swin Transformer have aimed to mitigate these data requirements through distillation and hierarchical attention structures.

Recent works show that ViTs outperform CNNs on large datasets across tasks such as classification, detection, and segmentation, but still face challenges in data-constrained regimes [23], [24]. Plug-and-play methods like the Summary Vision Transformer (S-ViT) and Multi-Gradient Image Transformer (MGiT) integrate CNN-derived inductive biases or auxiliary gradient guidance to stabilise training on small datasets, particularly in medical imaging.

1.2.4 Timeline of Evolution

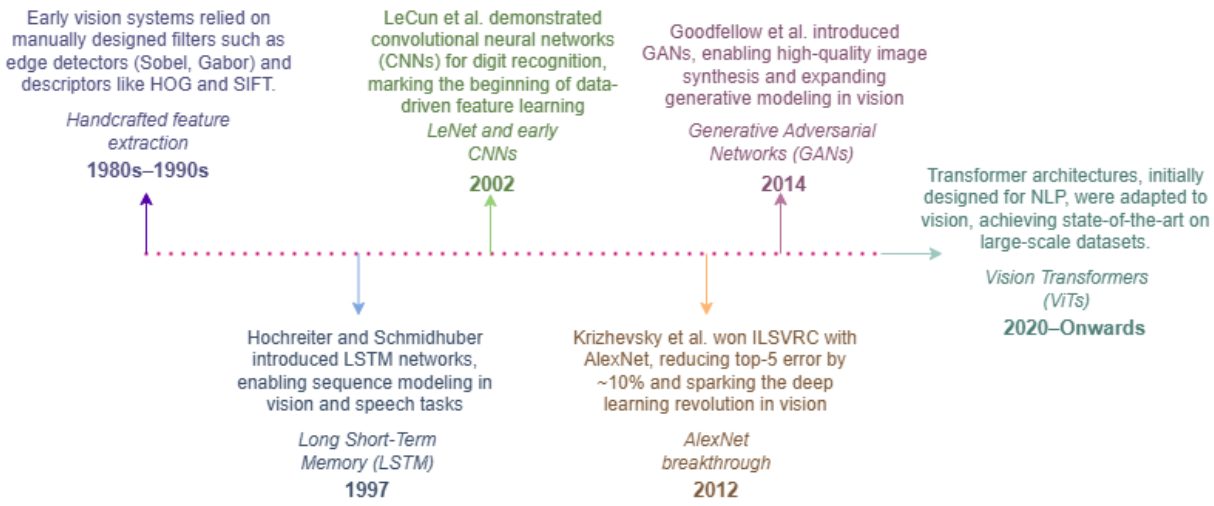


Figure 1.2: Timeline of computer vision evolution: from handcrafted kernels, CNNs, and generative approaches to Vision Transformers and hybrid methods.

Figure 1.2 summarises the trajectory of computer vision methods: from handcrafted kernels in the 1980s1990s, to the introduction of CNNs in the 1990s2000s [25], large-scale breakthroughs in the 2010s [2], [3], generative modelling in the mid-2010s [18], and finally to ViTs and hybrid approaches in the 2020s.

1.3 Explainable Artificial Intelligence (XAI)

XAI is a vital component of the modern AI pipeline. It helps end users understand the logic behind decisions made by machine learning models. In this thesis, we generate explanations throughout development and evaluation to answer concrete questions about what features a model uses, whether those features align with domain knowledge, and how design choices affect the learned focus. We distinguish between interpretability and explainability. Interpretability refers to transparent models whose reasoning can be followed directly, such as decision trees or linear models; this is often called *ante hoc* interpretability as shown in Figure 1.3. Explainability refers to methods that make opaque models understandable after training. Post hoc approaches can be model-agnostic, operating only on inputs and outputs so they can work with any predictor, or model-specific, exploiting internal signals such as gradients, feature maps, or attention to attribute predictions to input regions. In our studies we also use model-agnostic methods such as LIME and SHAP for completeness [26], [27].

ViTs introduce specific challenges for XAI. Raw attention weights do not always correspond to feature importance [28]; early layers may capture noisy patterns, and naïve rollouts across layers can lead to diluted or over-smoothed attributions [29]. Attention heads contribute unequally and their roles evolve with depth, while the token-based representation complicates mapping explanations back to image pixels. These factors make direct transfers of CNN-oriented attribution methods unreliable and highlight the need for transformer-tailored approaches. A promising direction is to combine attention information with gradient-based cues, while adaptively weighting layers so that only semantically meaningful contributions are emphasised. Such strategies are expected to yield sharper and more faithful explanations than simple attention visualisations. To ensure reliability,

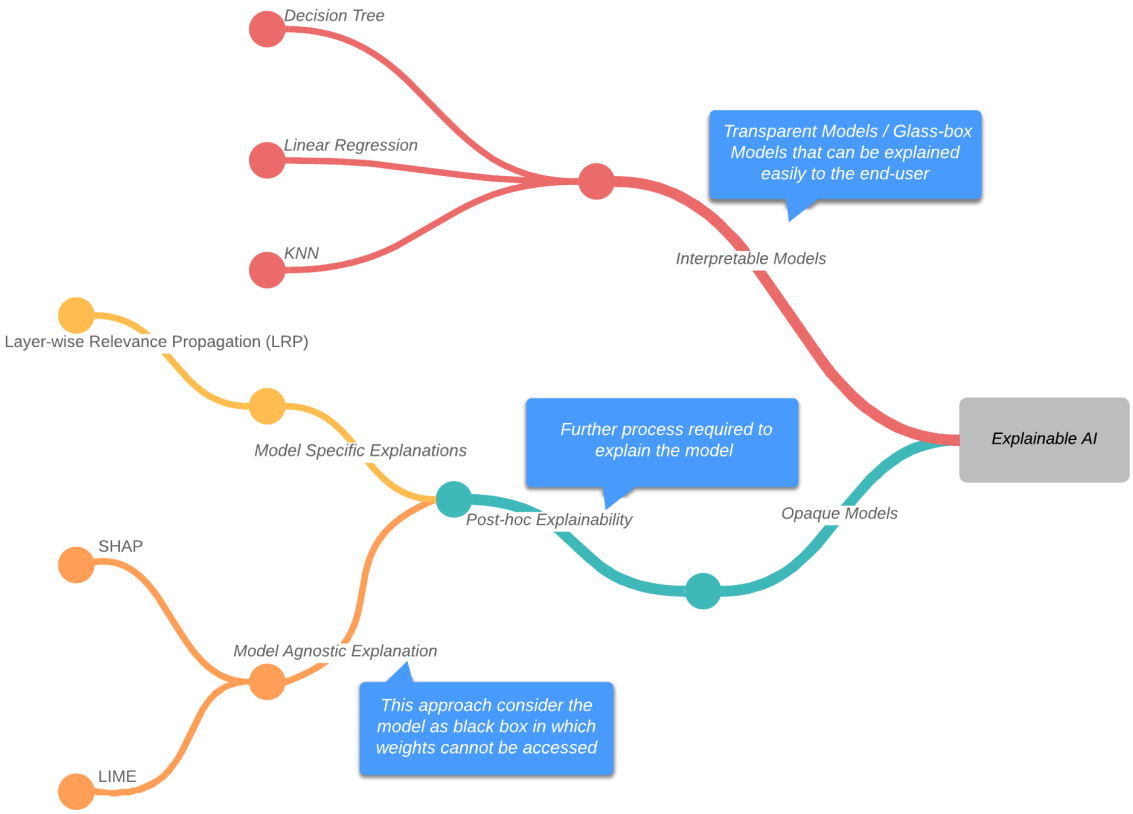


Figure 1.3: Hierarchy of Explainable AI (XAI) models. Interpretable models (e.g., Decision Tree, Linear Regression, KNN) are transparent or glass-box models that can be directly understood by end-users. In contrast, opaque models require post-hoc explainability methods. These include model-specific techniques such as Layer-wise Relevance Propagation (LRP), and model-agnostic approaches such as SHAP and LIME, which treat the underlying model as a black box.

explanation quality should be assessed with both qualitative and quantitative protocols. Relevant criteria include correlations with prediction faithfulness under perturbations, sensitivity to small input changes, compactness or sparsity of attribution maps, and robustness under parameter randomisation tests [30]. These evaluations can be carried out alongside widely used attribution techniques such as Grad-CAM, LRP, attention rollout, and model-agnostic baselines including LIME and SHAP [26], [27], [29].

Explanations also serve as a probe of robustness. When attributions concentrate on semantically correct regions, predictions tend to remain stable under nuisance variation and are harder to manipulate with small perturbations. If maps highlight backgrounds, acquisition artefacts, or text overlays, the model is likely using shortcuts that produce brit-

tle decision boundaries. This motivates explanation-driven debugging, where suspected cues are masked or removed and the model is retrained or regularised accordingly. As we adapt ViTs to small datasets with S-ViT and MGiT, we use explanations to verify that accuracy gains coincide with a shift of attention toward clinically meaningful structures rather than dataset-specific artefacts. In our medical imaging experiments, these checks ensure improvements align with clinically relevant focus. Finally, we link explanation patterns to adversarial analysis: reliance on spurious cues predicts higher sensitivity to adversarial perturbations [31]. We use this link to guide model selection, shape training strategies, and validate that the resulting ViTs are both performant and dependable in high-stakes settings.

1.4 Adversarial Attacks and Robustness

Adversarial attacks expose how sensitive modern vision models can be to small, targeted perturbations. In white-box settings, an attacker uses gradient information to craft an input

$$\tilde{x} = x + \eta \tag{1.1}$$

that is visually similar to x but causes a different prediction, as shown in Eq. (1.1). We constrain the perturbation by

$$\|\eta\|_\infty \leq \varepsilon, \tag{1.2}$$

i.e., $\tilde{x}$ lies within an ℓ_∞ ball of radius ε around x. Two standard attacks are the *Fast Gradient Sign Method* (FGSM), which applies a single step in the sign of the loss gradient, and *Projected Gradient Descent* (PGD), which iterates multiple small steps within the ℓ_∞ ball to produce stronger adversarial examples [10], [32]. These attacks are widely used to probe robustness in evaluation pipelines and are formulated precisely in our thesis materials.

The consequences of such perturbations are most serious in safety-critical applications. In medical imaging, adversarial perturbation can induce label flips that risk misdiagnosis [33]. In autonomous driving, altered inputs can undermine traffic-sign recognition and

object detection, with physical-world demonstrations showing that manipulated signs can mislead perception modules [34]. Industrial inspection and security surveillance are similarly at risk when small artefacts cause systems to miss defects or raise false alarms. These risks are documented in our literature review and experiments and motivate defences that hold up under adversarial conditions without sacrificing routine performance.

Types of Adversarial Attacks

Adversarial attacks are commonly categorised by the adversarys knowledge of the model:

- **White-box attacks:** The attacker has full access to the model, including architecture, parameters, and gradients. Strong gradient-based methods such as FGSM [10] and PGD [32] fall into this category.
- **Black-box attacks:** The attacker has no direct access to model internals and can only observe inputoutput behaviour. Query-based methods (e.g., score-based or decision-based attacks) and transfer attacks are typical strategies.

Figure 1.4 illustrates these categories and highlights representative methods.

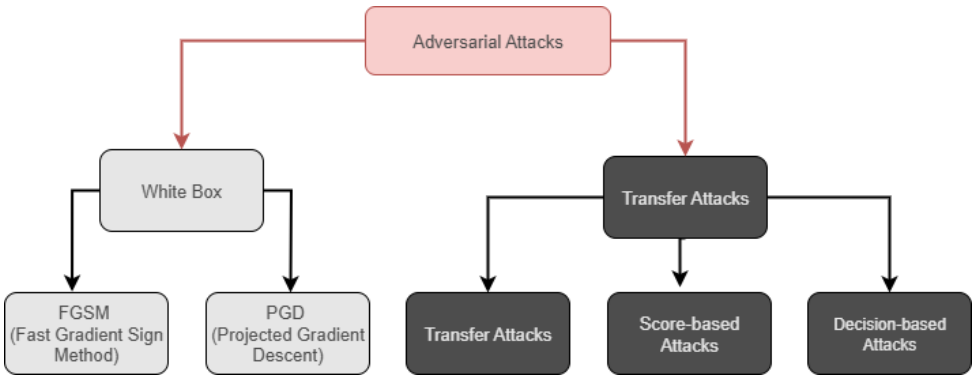


Figure 1.4: Types of adversarial attacks: white-box (full access to model parameters and gradients), and black-box (only query access to inputs and outputs).

A common defence is *adversarial training*, where adversarial examples are incorporated during learning to improve robust accuracy. While effective, standard adversarial training introduces practical costs. It increases training time and memory usage due to on-the-fly example generation and can reduce generalisation on clean data if not tuned carefully,

creating a trade-off between robustness and natural accuracy. The literature and our study discuss these effects and review variants such as *Free Adversarial Training* [35] (computational efficiency), *TRADES* [36] (balancing robustness and accuracy via a KL regulariser), *geometry-aware instance reweighting* (GAIR AT) [37], and *overfitting-aware* early stopping schedules [38], as well as *Friendly Adversarial Training*.

Within this landscape, our *feature-map fusion adversarial training* framework targets resilience and clean-data performance jointly [39]. The method trains parallel ResNet blocks on clean, adversarial, and noisy inputs, normalises and fuses the resulting feature maps, and then reduces dimensionality to limit the attackable space. On CIFAR-10 and CIFAR-100, this approach outperforms Free AT, TRADES, GAIR AT, and Overfitting-aware schedules across FGSM and PGD settings, while maintaining or improving clean accuracy. We also quantify the role of model capacity, showing that deeper backbones improve robust accuracy, with ResNet-50 offering a favourable balance between robustness and efficiency.

Despite these gains, a research gap remains. Robustness techniques should preserve accuracy on clean inputs, scale to realistic training budgets, and align with interpretability objectives. In later chapters, we address this gap by linking robustness checks with explanation-driven analysis. Explanations are used to verify that robustness arises from attention to task-relevant structures rather than spurious cues and to guide model selection and training choices. This connection underpins our broader thesis goal: robustness that is measurable, efficient, and consistent with model behaviour that can be inspected and trusted.

1.5 Vision Transformers (ViTs) in Data-Constrained Environments

ViTs treat an image as a sequence of patch tokens and use multi-head self-attention to capture long-range dependencies [11]. This global context modelling is effective on large-scale datasets and gives strong results when sufficient pre-training and compute are available.

Compared with convolutional networks, which hard-code locality and translation equivariance through kernels and weight sharing, vanilla ViTs rely on data to learn these priors [12]. The lack of built-in inductive bias makes standard ViTs data-hungry and less stable on small datasets. Without large pre-training, they can overfit, learn shortcuts, or produce attention patterns that do not align with the underlying structure of the task. In many specialised domains, especially where annotation is costly, this behaviour limits their usefulness despite their potential advantages in modelling global relationships.

Prior work addresses these issues along two main paths. One path modifies the architecture to reintroduce spatial priors. Examples include hierarchical windowed attention [40] and convolutional token embeddings [41] that inject locality into the tokenisation or attention stages. These designs improve data efficiency but reduce flexibility, since they change the backbone and often require pre-training or full re-training to transfer across tasks. The second path keeps the core ViT blocks intact and adds plug-and-play strategies. Knowledge distillation from a strong CNN teacher improves data efficiency without altering the transformer layers, and relative position or localisation signals strengthen spatial modelling with minimal code changes [12]. These approaches lower re-training costs and are easier to drop into existing pipelines, but they may still leave optimisation unstable on very small datasets, and they do not directly address how to verify that the model focuses on meaningful regions.

This thesis follows the plug-and-play direction and targets small, specialised datasets. We introduce two complementary components that keep the backbone unchanged while injecting bias and stabilising training. The first is *S-ViT*, which augments the class token with a summary token computed from lightweight CNN features to provide spatial and hierarchical cues [24]. The second is *MGiT*, which pairs the target ViT with a compact auxiliary ViT that shares gradients to guide early optimisation and reduce overfitting [23]. Both are designed to be compatible with our explanation tools so that accuracy gains can be checked against attribution maps and attention patterns. The goal is a ViT pipeline that adapts to limited data while supporting robust and explainable decision making [42].

1.6 Linking Robustness, ViT Adaptation, and XAI

This thesis brings together three lines of work that address what the model relies on, how stable its decisions are, and how to make ViTs effective with limited data. First, we use XAI to verify feature use. For transformer models, we employ *FocusViT*, which integrates class-specific gradients with attention and aggregates only layers that carry useful semantics [42]. This produces clearer and more faithful maps than attention-only visualisations and supports quantitative checks of faithfulness, robustness to small input changes, and sparsity. These explanations allow us to audit whether predictions depend on task-relevant regions rather than spurious cues, a requirement for applications where understanding model behaviour matters.

Second, we study robustness as stability under small, targeted perturbations. For convolutional baselines, we propose a *feature-map fusion adversarial training* framework that processes clean, adversarial, and noisy inputs in parallel ResNet blocks, normalises and fuses the resulting feature maps, and reduces dimensionality before classification [39]. In evaluations with standard white-box attacks such as FGSM and PGD, this approach improves resistance to perturbations while preserving accuracy on clean inputs. The design emphasises diverse feature learning rather than heavy architectural changes and establishes a robust reference for later comparisons.

Third, we adapt ViTs for small-data regimes. Summary Vision Transformer (*S-ViT*) augments the class token with a CNN-derived summary token to inject spatial and hierarchical cues without altering the transformer blocks [24]. Multi-Gradient Image Transformer (*MGiT*) trains a compact auxiliary ViT in parallel and shares its gradients with the primary model, stabilising early optimisation and reducing overfitting [23]. We evaluate these adaptations on small to medium vision benchmarks and extend them to medical imaging, where we also integrate XAI analyses to check that performance gains align with clinically meaningful focus.

Taken together, the thesis provides a single, coherent evaluation framework. ViT adaptations (S-ViT and MGiT) target data efficiency in small and specialised datasets; ro-

bustness testing quantifies stability under adversarial perturbations for the convolutional reference; and XAI, including FocusViT, verifies that predictions are based on relevant structures and supports quantitative assessment of explanation quality. The components are assessed jointly across standard and medical datasets, linking accuracy, robustness, and explainability in a way that reflects the methods and evidence presented in this work.

1.7 Research Objectives and Questions

This thesis addresses four interacting needs in vision systems trained with limited data:

1. Enforcing robustness to adversarial perturbations without reducing clean accuracy (Clean accuracy refers to performance on unperturbed test images).
2. Making Vision Transformers train effectively in small-data regimes without heavy redesign.
3. Producing explanations that confirm predictions are based on meaningful image evidence.
4. Demonstrating validity on medical imaging tasks where trustworthiness is essential.

Objectives and Associated Research Questions

O1 *Robustness under white-box perturbations.*

- **RQ1** Does feature-map-fusion-based adversarial training (Fusion AT) improve robust accuracy under FGSM/PGD while maintaining competitive clean accuracy?
- **RQ2** How does robustness scale with backbone capacity when using Fusion AT?

O2 *ViT adaptation for small datasets with minimal burden.*

- **RQ3** Can CNN inductive biases injected via a summary token (S-ViT) consistently improve ViT performance on small datasets relative to plain ViT and alternative strategies?

RQ4 Does auxiliary gradient guidance (MGiT) accelerate convergence and improve generalisation for ViTs in data-constrained regimes?

O3 *ViT-specific explainability with quantitative evaluation.*

RQ5 Do FocusViT attributions better capture class-relevant regions than baseline XAI methods for ViTs, as measured by faithfulness/sensitivity criteria?

O4 *External validity in safety-relevant domains.*

RQ6 When S-ViT and MGiT are applied to medical datasets, do the resulting explanations (LIME, SHAP, rollout, Grad-CAM, LRP, and FocusViT) concentrate on clinically meaningful structures, and do quantitative XAI metrics align with improvements in predictive performance?

How Objectives are Addressed

To meet these objectives, this thesis contributes:

- For **O1**, we propose **Fusion AT**, a feature-map fusion adversarial training framework that processes clean, adversarial, and noisy inputs in parallel ResNet streams, fuses their normalised feature maps, and reduces dimensionality with a 1×1 convolution. This design captures complementary feature representations and strengthens robustness against FGSM and PGD attacks while preserving accuracy on clean data. Evaluation is performed on CIFAR-10 and CIFAR-100, with controlled comparisons against adversarial training baselines and an analysis of the effect of backbone capacity on robustness.

- For **O2**, we design two plug-and-play ViT enhancements for small-data regimes. **S-ViT** introduces a CNN-derived summary token that is concatenated with the class token, injecting local and hierarchical inductive biases into the token sequence. **MGiT** employs a lightweight auxiliary ViT trained in parallel to the primary ViT, sharing gradients to guide early optimisation and reduce overfitting. Both methods are benchmarked against strong baselines (ViT, DeiT, Swin) on small and medium

datasets, and we report balanced accuracy, AUC, and F1 across multiple seeds and loss functions. Complexityaccuracy trade-offs and convergence stability are explicitly analysed.

- For **O3**, we introduce **FocusViT**, a ViT-specific explanation method that fuses attention maps with class-specific gradients and aggregates only semantically meaningful layers through a faithfulness-driven layer-skipping strategy. This approach produces sharper and more faithful token-level attributions compared to attention rollout and Grad-CAM. We evaluate explanation quality with quantitative metrics (faithfulness correlation, sensitivity, sparsity, parameter randomisation) and qualitative comparisons, establishing a systematic pipeline for XAI in ViTs.

- For **O4**, we extend S-ViT and MGiT to medical imaging tasks, including ISIC-2017 skin lesion classification and COVID-19 chest radiography. These datasets introduce challenges of small sample sizes, class imbalance, and clinically fine-grained cues. To ensure external validity, we combine predictive evaluation (BA, AUC, F1 with statistical tests such as ShapiroWilk, Welchs t-test, and MannWhitney U) with interpretability analysis (LIME, SHAP, attention rollout, Grad-CAM, LRP, and FocusViT). Results show that predictive gains align with clinically meaningful evidence, demonstrating that the proposed methods yield not only improved accuracy but also trustworthy, interpretable outputs suitable for safety-critical domains.

1.8 Thesis Structure and Chapter Summaries

This thesis is organised into eight chapters, each addressing a key aspect of the research and building towards robust, data-efficient, and explainable ViTs.

1. **Introduction** Establishes the context of trustworthy computer vision under data constraints. Outlines the motivation for integrating robustness, small-data ViT adaptation, and explainability, and states the thesis contributions along with the evaluation plan across both standard and medical datasets.

2. **Literature Review** Surveys adversarial threats and defences, ViTs in data-constrained regimes, and explainability methods with evaluation metrics for vision tasks. Identifies gaps motivating the proposed methods: (1) data efficiency for ViTs without major architectural change, (2) robustness that preserves clean accuracy, and (3) reliable explanations tailored to Transformer architectures.

3. **Feature Map Fusion Adversarial Training** Presents an adversarial training framework for CNNs that processes clean, adversarial, and noisy inputs in parallel, fuses their feature maps, and classifies on a compact representation. Reports clean and robust accuracy against standard white-box attacks and competitive baselines, and analyses the robustness versus accuracy trade-offs and the influence of model capacity.

4. **Summary Vision Transformer (S-ViT)** Introduces a plug-and-play method that augments the ViT class token with a CNN-derived summary token, injecting spatial and hierarchical cues while keeping the backbone unchanged. Reports results on small- to medium-scale benchmarks, comparing against baseline ViTs and strong plug-and-play alternatives, with discussion of complexityaccuracy trade-offs.

5. **Multi-Gradient Image Transformer (MGiT)** Proposes a training strategy pairing the target ViT with a compact auxiliary ViT to guide gradients during early optimisation and reduce overfitting in limited-data regimes. Evaluates gains across datasets for both training from scratch and fine-tuning, tracks feature-distribution alignment between branches, and includes sensitivity analyses of design choices.

6. **FocusViT: XAI for ViTs** Presents an explanation method that combines attention with class-specific gradients and adaptive layer aggregation to generate sharper, more faithful token-level attributions than attention-only visualisations. Evaluates explanation quality with quantitative metrics and compares against common baselines, establishing the XAI toolkit for later applications.

7. **Medical Domain Evaluation with Comprehensive XAI** Applies S-ViT and MGiT to medical imaging tasks, evaluating performance with metrics appropriate

for class imbalance. Uses FocusViT alongside established attribution methods to verify that predictions focus on clinically meaningful structures and discusses observations relevant to medical datasets and deployment scenarios.

8. **Conclusion and Future Work** Integrates findings across robustness, small-data ViT adaptation, and explanation quality. Reflects on limitations and proposes future research directions, including tighter integration of explanation signals during training, broader domain adaptation, and combined studies of robustness and explainability in real-world settings.

1.9 List of Publications

1. **Mohsin Ali**, Haider Raza, John Q. Gan (2024). Fortifying Deep Neural Networks for Industrial Applications: Feature Map Fusion for Adversarial Defense. *IEEE Conference on Industrial Electronics and Applications (ICIEA)*. DOI: 10.1109/ICIEA61579.2024.10665 Available at: https://ieeexplore.ieee.org/document/10665133

2. **Mohsin Ali**, Haider Raza, John Q. Gan, Muhammad Haris (2024). Integrating Spatial Information into Global Context: Summary Vision Transformer (S-ViT). *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. DOI: 10.1109/DICTA63115.2024.00040. Available at: https://ieeexplore.ieee.org/document/10869559

3. **Mohsin Ali**, Haider Raza, John Q. Gan, Muhammad Haris (2025). Optimising Vision Transformer Performance on Limited Datasets: A Multi-Gradient Approach. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Available at: https://openaccess.thecvf.com/content/CVPR2025W/TCV/html/Ali_Optimising_Vision_Transformer_Performance_on_Limited_Datasets_A_Multi-Gradient_Approach_CVPRW_2025_paper.html

4. **Mohsin Ali**, Haider Raza, John Q. Gan, Muhammad Haris (accepted). FocusViT: Faithful Explanations for Vision Transformers via Gradient-Guided Layer-Skipping. *accepted in AISTATS 2026.*

5. **Mohsin Ali**, Haider Raza, John Q. Gan, Muhammad Haris (to be submitted). Enhancing Vision Transformers for Medical Imaging: Extending MGiT and S-ViT with XAI Methods. *Journal/Conference under preparation for submission.*

Chapter 2

Literature Review

2.1 Thesis Review Scope

The evolution of computer vision has been shaped by a steady rise in model complexity and by increasing expectations around accuracy, robustness, and interpretability. This review is scoped to support three pillars that underpin the thesis: adversarial robustness for vision models, adaptation of ViTs to small-data regimes, and XAI for ViTs. The aim is not to provide an exhaustive survey of computer vision, but to assemble the concepts, baselines, and evaluation practices that directly ground the methods and experiments developed later, namely feature-map fusion adversarial training for convolutional networks, S-ViT and MGiT for data-efficient ViTs, and FocusViT for ViT-specific explanations [23], [24], [39], [42].

2.1.1 Key Definitions

A *threat model* specifies an adversarys knowledge and capabilities, for example white-box access to parameters and gradients or black-box query access. *Perturbation norms* characterise allowable input changes, typically the ℓ_p family; throughout, adversarial examples are constrained by a budget ε, for instance $\|\eta\|_\infty \leq \varepsilon$ [10], [32]. *Inductive bias* denotes architectural assumptions that guide learning, such as locality and translation equivariance in convolutional networks, whereas vanilla ViTs rely more on data and pre-training to

learn these priors [11]. We distinguish *interpretability* from *explainability*: interpretability refers to models that are transparent by design (ante hoc), whereas explainability refers to post hoc methods that make opaque models understandable after training [6]. Post hoc methods can be *model-agnostic*, operating only on inputs and outputs, or *model-specific*, exploiting internal signals such as gradients, feature maps, or attention.

2.1.2 Inclusion Lens

The review focuses on peer-reviewed work in adversarial robustness for image classification and related vision tasks, data-efficient training strategies for ViTs, and XAI with an emphasis on ViT-specific methods and quantitative evaluation. Priority is given to influential and recent publications from venues such as CVPR, ICCV, ECCV, NeurIPS, ICML, ICLR, TPAMI, and IJCV, as well as domain-relevant medical imaging venues where appropriate. For robustness, we cover widely used white-box attacks and practical adversarial training variants that inform our baselines and comparisons; certified defences are noted briefly where they clarify limits of empirical robustness [32]. For ViTs, we consider both architecture-modified and plug-and-play approaches e.g., hierarchical windowed attention [40] and knowledge-distillation strategies that retain a plain ViT backbone [43] with emphasis on methods that require minimal re-training, aligning with S-ViT and MGiT. For XAI, we include saliency and attribution methods used in practice, with attention to transformer-aware techniques and metrics that allow quantitative comparison.

2.1.3 Roadmap

Section 2.2 reviews deep learning foundations for vision, contrasting convolutional networks and ViTs, and motivates the need for inductive bias and training stability. Section 2.3 surveys adversarial machine learning for vision, including threat models, common attacks, evaluation protocols, and adversarial training methods, and positions our feature-map fusion adversarial training [39]. Section 2.4 reviews ViTs in small-data settings, covering architecture-modified and plug-and-play strategies, and situates S-ViT and MGiT [23], [24], [40], [43]. Section 2.5 surveys XAI for vision, with a focus on transformer-

aware methods and quantitative metrics, and introduces the context for FocusViT [6], [42]. Section 2.6 discusses intersections between robustness, data-efficient ViTs, and XAI, and Section 2.7 outlines the medical imaging context and datasets used later in the thesis. Section 2.8 concludes the chapter and summarises the gaps that Chapters 3--7 address.

2.2 Foundations: Deep Learning for Vision

Deep learning has reshaped computer vision by enabling models to learn feature hierarchies directly from data [2]. Two families dominate current practice: CNNs and ViTs. CNNs operate with convolutional kernels, weight sharing, and pooling to capture local structure and encourage translation equivariance. This pipeline builds coarse-to-fine representations in which edges, textures, parts, and object-level cues emerge progressively. ViTs, in contrast, convert an image into a sequence of patch tokens, add positional encodings, and use multi-head self-attention to mix information across tokens [11]. Self-attention offers a flexible mechanism for modelling long-range dependencies and adapting the effective receptive field to the content of the image. Both approaches can reach high accuracy, but their architectural assumptions lead to different behaviours across data regimes and under different optimisation conditions.

The most consequential difference is inductive bias. CNNs hard-code locality and approximate translation equivariance through their design, which reduces the hypothesis space and acts as a strong form of regularisation. This tends to improve data efficiency, stabilise optimisation, and make networks less sensitive to nuisance variation when training data are limited. Vanilla ViTs do not embed these spatial priors by default. They learn from patch tokens with global mixing but no built-in preference for local continuity or hierarchical composition. As a result, ViTs usually benefit from large-scale pre-training, strong augmentation, and careful regularisation to reach their potential [11], [44]. When trained from scratch on small datasets, they are more prone to overfitting, shortcut learning, and unstable token-level attention patterns that do not align with the underlying structure of the task. In practice, these differences help explain why CNNs often remain competitive on modest data while ViTs dominate when data and compute are abundant.

These observations motivate the methodological choices in this thesis. Rather than redesigning the transformer backbone, we preserve the standard ViT blocks and introduce lightweight components that supply the missing bias and improve training stability. **S-ViT** augments the class token with a summary token derived from a compact CNN, injecting spatial and hierarchical cues that guide the transformer without constraining its global context modelling [24]. **MGiT** pairs the target ViT with a compact auxiliary ViT and shares gradients during training, which stabilises early optimisation and reduces overfitting in small-data regimes [23]. Together, these adaptations seek to recover the data efficiency associated with CNN-style priors while retaining the flexibility of self-attention. Because accuracy alone is not sufficient for deployment, the thesis evaluates these adaptations alongside explanation tools tailored to transformers, so that improvements can be checked against token- and pixel-level evidence, and later linked to robustness analyses. Prior work along these lines includes architecture-modified ViTs that reintroduce spatial priors e.g., hierarchical windowed attention and convolutional token embeddings [40], [41] and plug-and-play data-efficient training via distillation [43].

2.3 Adversarial Machine Learning for Vision

2.3.1 Threat Models and Attack Families

Adversarial machine learning studies how small, structured perturbations to inputs can induce incorrect predictions in otherwise accurate vision models. A *threat model* specifies the adversarys knowledge and capabilities. In the white-box setting, the attacker has full access to the network parameters and gradients; in the black-box setting, the attacker can only query the model. Within white-box attacks, FGSM and PGD are canonical baselines [10], [32]. FGSM applies a single gradient step to move the input toward higher loss under an ℓ_p constraint, while PGD iterates many small steps and projects back to the allowable set after each step [32]. Constraint sets are typically ℓ_∞, ℓ_2, or ℓ_1 balls with a budget ε; evaluations should cover a range of ε to characterise sensitivity. Beyond untargeted attacks that simply cause misclassification, targeted variants force the model toward a

Table 2.1: Attack configurations for white-box ℓ_∞ evaluations. ε is the perturbation budget (normalised pixel scale), α is the step size, and K is the number of steps.

Attack	**Threat model**	**Norm**	ε (normalised)	**Steps (K)**	**Step size (α)**	**Targeted?**
FGSM	white-box	ℓ_∞	0.03	1	0.03 (single-step)	No
PGD	white-box	ℓ_∞	0.03	10	0.003	No
PGD	white-box	ℓ_∞	0.03	20	0.0015	No
PGD	white-box	ℓ_∞	0.03	30	0.001	No

specific wrong label. Stronger suites like AutoAttack combine parameter-free components to reduce evaluator bias [45], while optimisation-based methods such as CarliniWagner search for small-norm perturbations that cross decision boundaries. Physical-world attacks demonstrate that printable patterns, adversarial patches, or modified signage can transfer to deployed systems [34]. Despite this breadth, white-box PGD remains a standard stress test because it approximates a worst-case first-order adversary and exposes gradient-based weaknesses that often transfer to weaker settings [46]. See Table 2.1 for the white-box ℓ_∞ attack configurations (budget ε, step size α, steps K) used throughout.

2.3.2 Evaluation Protocols and Pitfalls

Data augmentation interacts with adversarial training in non-trivial ways because it affects the loss landscape seen by both the adversarial attack and the model update. Some augmentations can stabilise training by increasing data diversity and reducing overfitting to a narrow set of adversarial examples. For instance, moderate geometric transformations or colour jitter can act as regularisation and improve robustness when applied consistently to both clean and adversarial samples [38]. In contrast, aggressive or poorly chosen augmentations can harm evaluation by masking gradients. Techniques such as heavy mixing or random resizing may introduce stochastic or non-smooth transformations that interfere with the gradients used by first-order attacks, causing them to underestimate the models true vulnerability [47]. In these cases, apparent robustness does not reflect genuine invariance, but rather weakened attacks, and often disappears under stronger multi-step, black-box, or parameter-free evaluations [45]. For this reason, augmentation strategies should be specified explicitly and applied with care in adversarial training protocols.

2.3.3 Defensive Approaches

The most widely used defence is *adversarial training*, which augments learning with adversarial examples crafted on-the-fly. Standard PGD-based training improves robust accuracy but increases compute and can reduce clean accuracy if not tuned carefully [32]. Several variants trade compute, robustness, and clean performance in different ways. *Free Adversarial Training* reuses gradients across mini-batches to reduce cost [35]. *TRADES* separates classification and robustness objectives to balance natural and robust accuracy [36]. *Geometry-aware instance-reweighted training (GAIR-AT)* adjusts sample weights based on margin geometry to focus on vulnerable points [48]. Overfitting-aware schedules and early stopping reduce robust overfitting toward the end of training [38]. Beyond adversarial training, certified defences provide provable guarantees under specific norms but are often computationally heavy or yield looser bounds on large models. Input-space transformations and denoisers can remove some perturbations but risk obfuscating gradients [47]. Data-level defences, such as targeted augmentation or curation to remove shortcut cues, can help but do not by themselves address worst-case perturbations. The overall picture is a set of trade-offs: stronger robustness typically increases training cost and may depress clean accuracy; inexpensive defences often fail under stronger evaluators.

This gap motivates the approach used later in the thesis: improving robustness while preserving clean accuracy and keeping training practical. The proposed **feature-map fusion adversarial training** aims to diversify learned representations without heavy architectural change or prohibitive inner-loop cost.

2.3.4 Positioning the Contribution

Feature-map fusion adversarial training operates at the representation level. During learning, the network processes three input streams in parallel: clean images, adversarially perturbed images under a specified norm and budget, and images with stochastic noise. Each stream produces feature maps that are normalised and then fused, after which a 1×1 convolution compacts channels before classification. The design encourages the

backbone to encode complementary cues that remain discriminative under perturbation while limiting the dimensionality of attackable features. Because the backbone (e.g., a ResNet) is not redesigned, the method integrates with standard pipelines and allows straightforward comparisons with established adversarial training baselines.

When surveying prior work and setting up comparisons, it is important to track four axes. First, dataset scale and difficulty, since robustness on CIFAR-10 can behave differently from CIFAR-100 or larger-scale tasks. Second, attack strength, including norms, budgets, steps, and restarts, and whether additional evaluators such as AutoAttack are used. Third, compute budget, because inner-loop cost varies widely across methods; reporting wall-clock, epochs, and memory clarifies practicality. Fourth, the clean versus robust trade-off, including whether techniques preserve or recover clean accuracy as robustness improves. Later chapters follow this template, comparing against representative baselines such as standard PGD training, Free AT, TRADES, GAIR-AT, and overfitting-aware schedules, and reporting both clean and robust accuracy under standard white-box attacks [38], [45]. Recommended summaries for the literature review include a table of attacks and evaluation settings (norms, budgets, step counts) and a table of adversarial training variants versus key properties (compute, robustness, clean accuracy).

2.4 Vision Transformers and Data Efficiency

2.4.1 ViT architecture essentials

Vision Transformers (ViTs) treat an image as a sequence of fixed-size patches. Each patch is linearly projected to a token embedding, a learnable class token is optionally prepended, and positional encodings are added so that the model can recover spatial order [11]. Stacks of Transformer encoder blocks then process the token sequence. Each block contains multi-head self-attention to mix information across tokens and a feed-forward multilayer perceptron to apply channel-wise transformations, with layer normalisation and residual connections stabilising optimisation [49]. This design promotes global context modelling because any token can attend to any other token within a layer [11]. Complexity follows

the sequence length. If an image of size $H \times W$ is split into patches of size $P \times P$, the number of tokens is $N = (H/P) \times (W/P)$. The attention operation scales roughly with $\mathcal{O}(N^2)$ in compute and memory [49], and also depends on the embedding dimension and the number of heads. Smaller patches increase N and raise the cost, while larger patches reduce cost but coarsen spatial detail. Memory pressure grows with depth and sequence length, which makes patch size, image resolution, head count, and width important levers for practical training and deployment.

2.4.2 ViTs on small datasets

Vanilla ViTs work best when trained with substantial data and strong regularisation. Without explicit inductive biases such as locality and translation equivariance, they rely on the data to learn these properties [11]. On small datasets they are more likely to overfit, adopt shortcuts linked to background or acquisition artefacts, and exhibit unstable token-level attention during early training; pre-training, knowledge distillation, and strong augmentation can mitigate these effects [43], and recent analyses characterise training stability and data requirements in detail [44]. Inductive bias helps resolve this tension. CNNs succeed in small-data regimes because their architectural constraints reduce the hypothesis space and encourage hierarchical, local-to-global feature composition. Injecting similar cues into ViTs, or stabilising their optimisation in the low-data regime, can improve data efficiency while preserving the ability to capture long-range dependencies with self-attention. See Table 2.2 for a summary of data-efficient ViT approaches, indicating backbone changes and brief working descriptions.

2.4.3 Data-efficient ViT strategies

Two broad directions have emerged. The first modifies the backbone to embed spatial priors. *Swin Transformer* introduces hierarchical representations with windowed and shifted self-attention to emphasise locality while enabling cross-window communication [40]. *CvT* replaces linear patch projections with convolutional token embeddings so that early processing captures local structure [41]. Related designs alter tokenisation or po-

sitional encoding to strengthen spatial modelling. These approaches improve small-data performance but reduce flexibility because they change the core blocks and often require pre-training or full re-training when moving across tasks or domains. The second direction keeps the backbone intact and adds plug-and-play or training-level enhancements. *DeiT* uses distillation from a strong CNN teacher to transfer inductive bias without altering the transformer layers [43]. Relative localisation methods replace or augment absolute positional encodings to improve spatial reasoning. Regularisers such as label smoothing, stochastic depth, RandAugment, and Mixup strengthen generalisation. Additional components adjust token mixing or attention behaviour for example, shifting patches at input, modifying attention dropout, or encouraging diversity across heads. These techniques are attractive because they are easy to integrate into existing codebases and preserve the analysis and tooling built around a plain ViT, although they may still leave early optimisation fragile on very small datasets.

The gap motivating our work is to retain a standard ViT backbone while adding just enough spatial bias and training stability to operate effectively in small-data settings, and to do so with minimal re-training burden so that comparisons, ablations, and explanation analyses remain straightforward.

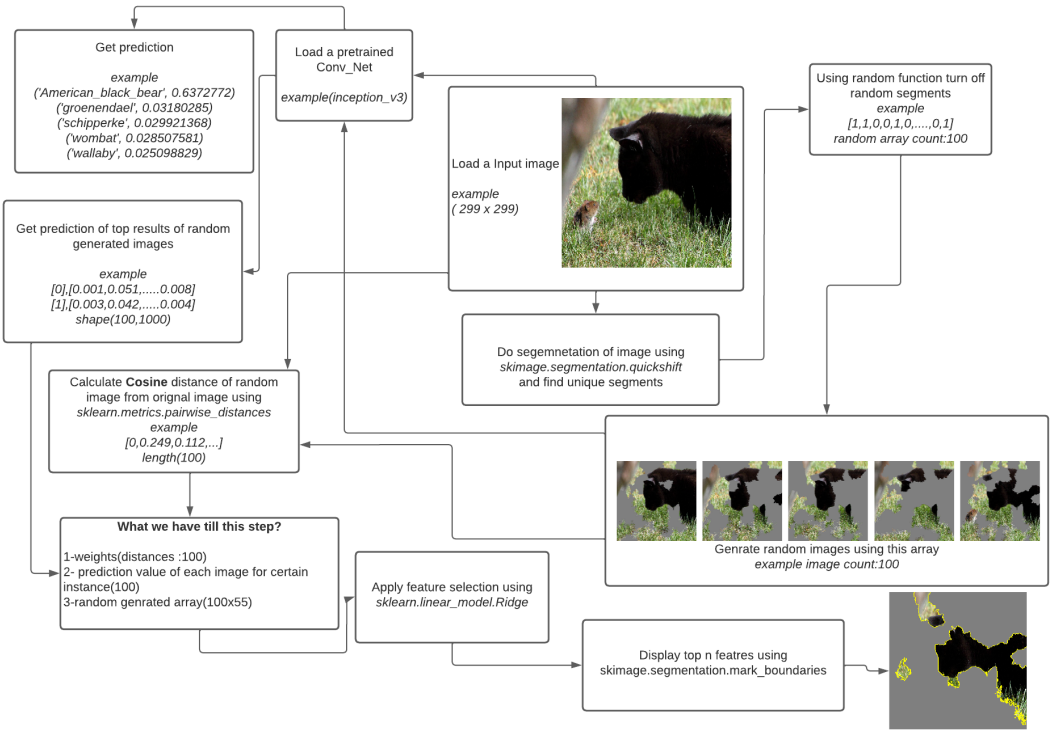


Figure 2.1: Workflow of Local Interpretable Model-agnostic Explanations (LIME) applied to image classification. Superpixel segmentation, perturbation, and local surrogate models are used to estimate feature importance.

2.4.4 Positioning our contributions

S-ViT and MGiT address the two needs above while keeping the ViT blocks unchanged. S-ViT injects CNN-derived spatial and hierarchical cues by appending a summary token to the usual class token [24]. A lightweight CNN extracts features that capture locality and part–whole structure; these are pooled into a compact summary that is supplied to the transformer alongside the patch tokens. The attention mechanism can then integrate both global context and the injected spatial cues without redesigning the encoder blocks. This encourages the model to attend to meaningful local structure and improves data efficiency when training data are limited. MGiT focuses on optimisation stability. It pairs the target ViT with a compact auxiliary transformer trained in parallel, and shares gradients so that the primary model receives additional guidance during the early stages of training [23]. This guidance reduces variance, discourages brittle solutions, and helps

the main model converge to features that generalise better in small-data regimes. Because the auxiliary branch is lightweight and does not alter the backbone, the approach is easy to attach or remove and is compatible with standard training schedules and evaluation pipelines. Together, S-ViT and MGiT provide complementary improvements. S-ViT supplies the inductive cues that vanilla transformers lack, and MGiT stabilises the optimisation dynamics that are otherwise fragile in low-data settings. Both preserve the plain ViT architecture, which simplifies reuse, comparison with existing plug-and-play baselines, and integration with explainability and robustness analyses developed elsewhere in the thesis.

2.5 Explainable AI for Vision

2.5.1 Taxonomy and aims

Explainable AI (XAI) in vision seeks to make model behaviour observable and contestable. We use a simple taxonomy that matches the needs of this thesis. First, *ante hoc* methods are interpretable by design (for example, linear models or shallow decision trees) and expose their reasoning without additional tooling; these are uncommon for high-accuracy image classifiers. Second, *post hoc* methods explain an already-trained model and split into *model-agnostic* approaches, which only require inputoutput access (for example, LIME and SHAP operating on superpixels), and *model-specific* approaches, which use internal signals such as gradients, feature maps, attention, or relevance propagation [27], [55]. Explanations can be local (for one image) or global (summaries over many images, concepts, or features). Across all families, desirable properties include faithfulness (attribution aligns with the models actual decision process), sensitivity (small input changes should not produce unstable attributions), sparsity/compactness (focus on a minimal set of important regions), stability across randomisation and reinitialisation checks, and basic *sanity checks* to rule out purely visual but uninformative heatmaps [30].

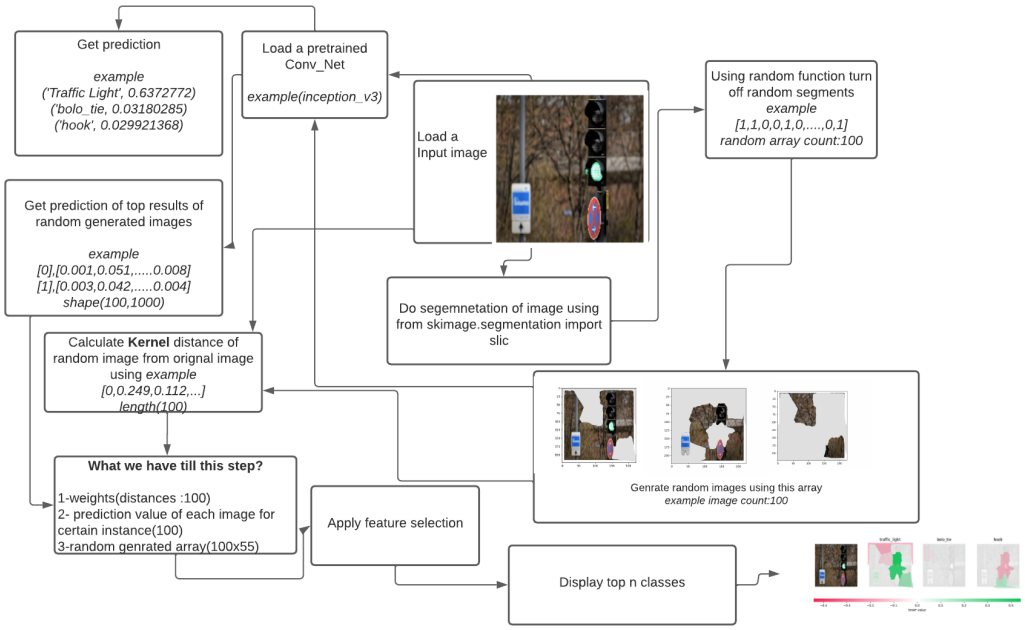


Figure 2.2: Workflow of Kernel SHAP applied to image classification. Perturbations of segmented regions are generated, predictions collected, and Shapley values estimated for local feature attribution.

2.5.2 XAI for CNNs: strengths and limitations

For convolutional networks, saliency and gradient-based methods (vanilla gradients, gradient×input, Integrated Gradients, SmoothGrad) quantify how output scores change with pixel-level perturbations and can highlight discriminative regions. Grad-CAM uses gradients with respect to late convolutional feature maps to produce class-specific heatmaps that are spatially coarse but stable and easy to interpret. Because it operates on high-level feature maps rather than raw pixels, Grad-CAM is less sensitive to noise and local gradient fluctuations, and its explanations tend to align well with the semantic structure learned by CNN backbones. For this reason, it is commonly adopted as a robust and practical default explanation method for convolutional architectures [56]. Layer-wise Relevance Propagation (LRP) redistributes the output score backwards layer by layer using conservation rules, resulting in pixel-level relevance that can be sharper than gradient-only maps [57]. Concept-based methods (for example, TCAV-style analyses) connect predictions to

higher-level human concepts rather than pixels, which is useful when pixel heatmaps are hard to interpret. Despite their utility, these techniques face well-documented limitations: class insensitivity when gradients are dominated by non-discriminative activations; visual plausibility versus faithfulness; susceptibility to noise and saturation effects; and failures revealed by parameter randomisation tests, where explanations should degrade if learned parameters are destroyed [30].

2.5.3 XAI for ViTs

Transformers introduce both opportunities and challenges for XAI. Attention matrices make tokentoken interactions explicit, which tempts the use of attention as an explanation. However, attention weights are not equal to importance in general: different heads contribute unequally; early layers contain noisy, non-semantic patterns; and downstream MLP blocks can change what matters after attention [28]. Attention rollout accumulates attention across layers to approximate token influence on the class token and is attractive because it is gradient-free and fast, but it can over-smooth or dilute saliency and may highlight pathways that are not class-specific [29]. Gradient-guided attention variants improve class specificity by combining gradients with attention scores, yet they still require careful aggregation across layers and heads. Token-level attributions must also be mapped back to pixels; with patch tokens this requires reshaping and potentially interpolation, which can blur or mislocalise saliency if not handled carefully. In short, ViTs need transformer-aware explanation strategies that account for head diversity, depth, and the interplay between attention and MLPs, rather than directly porting CNN-oriented methods.

2.5.4 Quantitative evaluation of explanations

To avoid relying on visually pleasing but unfaithful maps, we adopt a small set of quantitative checks alongside qualitative inspection. Faithfulness-style correlations measure whether masking or perturbing highly attributed regions changes the score in the expected direction. Max-sensitivity probes stability by applying small input perturbations

and quantifying attribution variability. Sparsity/compactness encourages concentrated, interpretable maps rather than diffuse attributions. Parameter randomisation tests verify that explanations depend on learned parameters by progressively randomising weights and checking for degradation [30]. Where appropriate, deletion/insertion curves measure how output confidence changes as top-attributed pixels are removed or added. Benchmarks combine these metrics with curated visual examples so that failures are detectable both numerically and qualitatively. This evaluation toolbox is used consistently later when comparing explanation methods for ViTs.

2.5.5 Positioning our contribution

The thesis develops *FocusViT*, an explanation method tailored to Vision Transformers. FocusViT combines attention with class-specific gradients to recover class-sensitive importance, and applies adaptive layer aggregation so that only layers carrying stable semantics contribute to the final map. Head contributions are normalised to handle unequal importance, and token-level maps are carefully projected back to the image to maintain spatial fidelity. The goal is to produce clearer, more faithful attributions than attention-only rollout, while remaining efficient enough for routine auditing. FocusViT integrates with the quantitative metrics above and with the small-data ViT adaptations developed in earlier chapters. In the experiments, we use it alongside established baselines (for example, attention rollout, Grad-CAM adapted to token maps, LRP variants, LIME, and SHAP) to verify that performance gains from S-ViT and MGiT correspond to attention on task-relevant structures rather than dataset artefacts, and to reveal cases where optimisation or inductive-bias choices lead to brittle or spurious focus [27], [29], [55]–[57].

2.6 Intersections: XAI, Robustness, and Small-Data ViTs

A recurring theme across this thesis is that data efficiency, robustness, and interpretability are interdependent rather than isolated goals. Explanations frequently expose shortcut

learning: models can achieve competitive metrics while relying on spurious signals such as calibration patches, borders, or acquisition artefacts [8]. In medical imaging, this behaviour has been demonstrated concretely; for instance, when coloured patches correlate with benign skin lesions, conventional CNN classifiers exploit the patch rather than lesion morphology, and inserting such patches into malignant images can flip predictions an explicit example of being right for the wrong reasons that degrades external validity. Interpretable pipelines that make reasoning visible allow such biases to be detected and mitigated, either through data curation or by constraining model behaviour during or after training. These findings motivate the use of explanations as routine diagnostics rather than as post hoc illustrations [58].

The relationship between robustness and interpretability is subtle. On the one hand, when attributions consistently highlight semantically relevant regions, decisions tend to be less sensitive to nuisance variation, which aligns with the practical aim of stable behaviour under small perturbations. On the other hand, visual plausibility alone is insufficient: post hoc heatmaps can mislead if not accompanied by quantitative checks, and explanation quality can vary with implementation choices. Recent surveys emphasise faithfulness-style correlations, sensitivity and stability metrics, sparsity or compactness, and sanity checks such as parameter randomisation, as necessary complements to qualitative maps [30], [59]. This thesis adopts that stance by pairing qualitative inspection with quantitative evaluation to avoid anecdotal conclusions.

Against this backdrop, our ViT adaptations and XAI method play complementary roles. S-ViT augments a plain ViT backbone with a CNN-derived summary token to inject locality and hierarchy without altering transformer blocks [24]. MGiT provides gradient guidance from a lightweight auxiliary transformer to stabilise early optimisation in limited-data regimes [23]. Both aim to recover data efficiency while preserving the architectural flexibility of a vanilla ViT, and both were designed to integrate naturally with attribution-based auditing. This is reflected again in the medical-imaging extension, where S-ViT and MGiT are applied to ISIC 2017 skin lesions and COVID-19 radiography with side-by-side XAI analyses (LIME, SHAP, attention rollout, Grad-CAM, and LRP)

to check whether improved metrics coincide with clinically meaningful focus.

Finally, FocusViT targets the ViT-specific gap in explainability. Raw attention is not a reliable importance signal, heads contribute unequally across depth, and token-to-pixel mapping can blur semantics in early layers. FocusViT fuses class-specific gradients with attention and aggregates only semantically useful layers, producing clearer token-level maps and aligning naturally with the quantitative evaluation toolbox used throughout the thesis [42]. In combination, S-ViT and MGiT address data efficiency; robustness is explicitly measured under standard white-box attacks for the CNN baseline [32]; and FocusViT verifies that ViT decisions rely on relevant structures providing a coherent evaluation workflow rather than treating these elements in isolation.

2.7 Medical Imaging Context

The thesis evaluates external validity on two representative medical datasets with properties that stress both data efficiency and interpretability. **ISIC 2017** provides dermoscopic images for skin-lesion analysis with pronounced class imbalance and well-known acquisition artefacts such as coloured calibration patches. The benchmark contains roughly two thousand training images split across melanoma, seborrhoeic keratosis, and benign naevi, with additional validation and test sets dedicated to binary diagnostic tasks [60]. Sample images are shown in Figure 2.3. These characteristics create a setting where small-sample generalisation, careful loss design, and explanation-driven auditing are all necessary for meaningful gains.

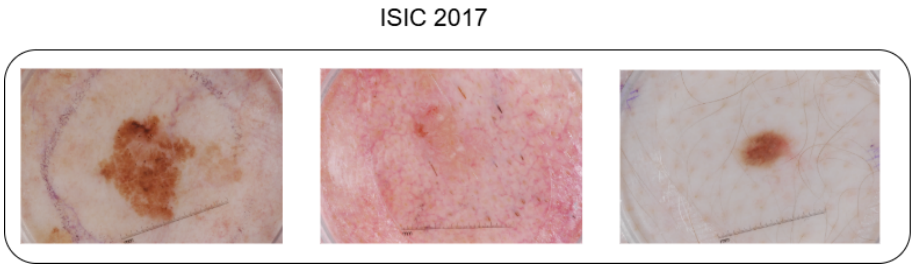


Figure 2.3: Sample dermoscopic image from the ISIC 2017 dataset for skin lesion analysis.

The **COVID-19 radiography** collection aggregates chest X-rays from multiple insti-

tutions and sources, spanning COVID-19-positive cases, normal lungs, lung opacity, and viral pneumonia. Heterogeneous acquisition, label imbalance, and variable pre-processing pipelines create a distributionally diverse dataset that benefits from strong regularisation, careful transfer learning protocols, and evaluation beyond accuracy [61]. Sample images are shown in Figure 2.4. In this thesis, standard splits with weighted F1, AUC, and balanced accuracy are used to quantify performance fairly under imbalance and to enable statistical comparison across model families.

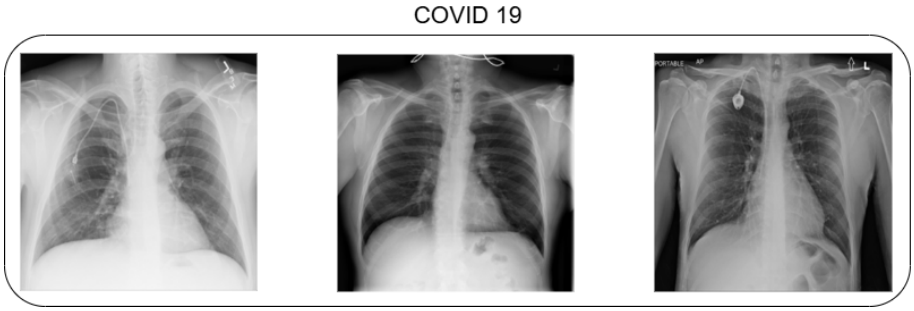


Figure 2.4: Sample chest X-ray from the COVID-19 radiography dataset.

Clinical constraints sharpen the requirements for interpretability. Prior work documents that image-level artefacts patches, text overlays, borders, or view-specific cues can dominate a models signal if they correlate with labels. When such cues are inserted or removed, predictions shift accordingly, revealing hidden dependencies that would not be visible from top-line metrics alone [58]. This motivates explicit XAI auditing in medical studies and supports workflows where spurious regions are identified and, when possible, neutralised by data curation or model adjustments. In this thesis, explanations accompany S-ViT and MGiT results to check that improved metrics correspond to attention on the lesion for ISIC and on pulmonary structures for chest X-rays, rather than on acquisition artefacts.

Practically, the evaluation protocol follows the medical evaluation described later in the thesis: ImageNet-pre-trained backbones are fine-tuned with consistent pre-processing, augmentation, and early stopping; losses such as weighted cross-entropy, focal loss, and class-balanced loss are used to counter imbalance; and quantitative statistics (normality checks, t-tests, or MannWhitney U) support claims of significance. Explanations (LIME,

SHAP, attention rollout, Grad-CAM, LRP, and FocusViT) are compared qualitatively and through faithfulness-style and sensitivity-style criteria, aligning with recommendations from recent XAI evaluation studies [30], [59].

2.8 Summary and Identified Gaps

The literature indicates four gaps that map directly to the thesis chapters. First, robustness methods often trade clean accuracy for adversarial gains or require heavy training budgets; there is a need for practical defences that improve resistance to white-box attacks while preserving routine performance. Chapter 3 addresses this with feature-map fusion adversarial training for CNNs, which emphasises diverse feature learning and competitive clean versus robust trade-offs.

Second, standard ViTs are data-hungry; many data-efficient variants either alter the backbone or require specialised pre-training. There is space for plug-and-play strategies that keep a plain ViT while injecting locality and stabilising optimisation. Chapters 4 and 5 contribute S-ViT and MGiT respectively: S-ViT appends a CNN-derived summary token to encode spatial and hierarchical cues; MGiT introduces an auxiliary transformer for gradient guidance in early training.

Third, explanations for ViTs remain challenging: attention weights are not reliable measures of importance, head roles vary by depth, and naïve rollouts can over-smooth. Chapter 6 proposes FocusViT, which combines class-specific gradients with attention and adaptive layer aggregation for clearer, more faithful maps, alongside a quantitative evaluation protocol.

Fourth, external validity requires demonstration on clinically relevant data with integrated XAI auditing. Chapter 7 extends S-ViT and MGiT to ISIC 2017 and COVID-19 radiography, evaluates with imbalance-aware metrics and statistical tests, and uses multiple XAI tools to verify that improvements align with clinically meaningful focus.

Together these chapters form a coherent programme: practical robustness for CNN baselines, data-efficient ViTs without heavy redesign, ViT-tailored explanations with quantitative validation, and a medical-imaging evaluation where performance and ex-

planations are assessed side by side.

Table 2.2: Data-efficient ViT approaches and whether they modify the backbone. Citations point to representative papers.

Approach	**Backbone change?**	**Working summary**
Swin [40], 2021	Yes	Shifted-window self-attention with a hierarchical backbone.
CvT [41], 2021	Yes	Convolutional token embeddings add locality; new backbone.
T2T-ViT [50], 2021	Yes	Progressive tokens-to-token module strengthens local structure.
DeiT [43], 2022	No	Distillation from a strong CNN teacher improves sample efficiency.
DRLoc [51], 2022	No	Relative localisation objectives enforce spatial relations without extra labels.
DropKey [52], 2023	No	Drops attention keys before softmax with a layer-wise schedule.
ES (Path Ensemble) [53], 2023	No	Re-weights implicit paths; self-distillation improves accuracy.
OFDB [54], 2023	No	Synthetic fractal pre-training from one image, then fine-tune on target data.
S-ViT [24], 2024	No	Adds a CNN-derived summary token to inject spatial and hierarchical cues.
MGiT [23], 2025	No	Auxiliary ViT provides gradient guidance; stabilises small-data training.

Table 2.3: Summary of common XAI methods for vision, their families, outputs, and known limitations.

Method	**Family**	**Output type**	**Known limitations**
Attention rollout	model-specific	Token-importance map projected to pixels	Attention $\neq$ importance; can oversmooth
Grad-CAM	model-specific	Class-specific, coarse heatmap	Low spatial resolution; sensitive to layer choice
LRP	model-specific	Pixel-level relevance map	Architecture/rule dependent
LIME	model-agnostic	Superpixel importance via local surrogate	Depends on segmentation and sampling
SHAP	model-agnostic	Superpixel Shapley attributions	Independence assumptions often violated

Chapter 3

Feature-Map Fusion Adversarial Training for Robust Convolutional Networks (Fusion AT)

Related publication

Portions of the work presented in this chapter have been published in:

Ali, M., Raza, H., & Gan, J. Q. (2024). *Fortifying Deep Neural Networks for Industrial Applications: Feature Map Fusion for Adversarial Defense.* In Proceedings of the 2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA), pp. 1--6.

3.1 Overview and Motivation

Deep learning systems are now embedded in safety-critical computer vision pipelines, from manufacturing inspection and surveillance to autonomous platforms. Yet small, carefully structured perturbations can cause confident misclassification, which exposes brittle decision boundaries and undermines trust in deployment [10]. In the white-box setting, where gradients are accessible, single-step Fast Gradient Sign Method (FGSM) and multi-step Projected Gradient Descent (PGD) generate adversarial examples constrained within a norm budget and have become standard stress tests for robustness reporting [10], [32]. These weaknesses are particularly concerning for industrial use where misclassification

can translate into safety incidents, financial loss, or regulatory non-compliance [34], [62].

This chapter introduces Feature-Map Fusion Adversarial Training (Fusion AT), a practical approach to improve robustness while maintaining competitive accuracy on clean inputs. The core idea is to diversify the learned representation by processing three input streams in parallel: clean images, adversarially perturbed images, and noisy images, through separate ResNet convolutional blocks. The resulting feature maps are normalised and fused by element-wise addition, then passed through a 1×1 convolution to reduce dimensionality before classification. The fusion encourages the network to encode complementary cues that remain discriminative under perturbation, and the channel "squeeze" reduces the effective subspace presented to later layers. Although the architecture does not impose an explicit constraint forcing equal contribution, several design aspects reduce this risk in practice. First, the three branches process inputs drawn from distinct distributions (clean, adversarially perturbed, and noisy), so relying on only one stream would degrade performance under distributional shift. Second, feature maps are normalised prior to fusion, which limits scale dominance by any individual branch and promotes balanced integration. The design keeps the backbone family standard (ResNet-18/50/101) and integrates with common training toolchains [4].

Our evaluation follows established adversarial protocols. Robustness is measured under FGSM and PGD with ℓ_∞ budgets on CIFAR-10 and CIFAR-100, using multiple PGD iteration counts to vary attack strength [32], [63]. In the experiments, a budget of $\varepsilon = 0.03$ is used and PGD steps $K \in \{10, 20, 30\}$ provide a sensitivity sweep; clean accuracy is reported alongside robust accuracy to characterise the trade-off. 10-fold cross-validation is used to check stability of the estimates. This setup reflects common practice in empirical robustness studies and makes direct comparison to widely used baselines straightforward.

Within this setting, Fusion AT is positioned against representative adversarial training strategies that attempt to balance robustness, computational cost, and retention of clean accuracy, such as Free AT, Friendly AT, GAIR-AT, and overfitting-aware schedules [35], [37], [38], [48]. A key question that arises is whether combining features from clean, adversarial, and noisy inputs can provide decision boundaries that are more resilient to

first-order attacks while still generalising well to natural data.

A secondary question concerns capacity. In literature work it is observed that increased model capacity, including greater depth, can improve adversarial robustness, although at higher computational cost [32], [64]. Intuitively, deeper networks possess a larger representational capacity and can learn more complex decision boundaries, which may better separate clean and adversarial examples in feature space. Additional layers allow successive nonlinear transformations that can progressively filter perturbation-sensitive features and amplify task-relevant structure. However, this increased capacity also raises training cost and memory usage, and without proper regularisation can lead to overfitting or diminishing robustness gains. We therefore evaluate Fusion AT across multiple backbone depths to examine this trade-off empirically. We therefore assess Fusion AT with ResNet-18, ResNet-50, and ResNet-101. The results show consistent gains as capacity increases, with ResNet-50 offering a favourable balance between compute and robustness and ResNet-101 providing the strongest protection under PGD at higher iteration counts. This capacity sweep helps position Fusion AT for practitioners who must trade accuracy, robustness, and throughput on constrained hardware.

In summary, the chapter addresses three practical needs for adversarially aware vision systems used in industrial contexts: (1) a training strategy that raises robust accuracy under FGSM and PGD without eroding clean accuracy, (2) a design that can be implemented with standard backbones and tooling, and (3) evidence across datasets and capacities to guide deployment choices. The remainder of the chapter formalises the threat model, presents the method in detail, describes the experimental protocol, and reports results alongside an analysis of strengths and limitations.

3.2 Background and Problem Formulation

Adversarial robustness is the ability of a vision model to maintain accurate predictions when inputs are corrupted by small, structured perturbations crafted to induce errors [10]. In white-box settings the attacker has full access to model parameters and gradients, and can therefore compute perturbations that push inputs across decision boundaries

while remaining visually similar to the originals. These attacks matter in safety-critical deployments: empirical studies and our own results show that small changes to pixels can flip classifications in manufacturing inspection, medical imaging, surveillance, and autonomous perception, undermining reliability and trust in operation [34], [62]. Consequently, robustness must be characterised alongside standard accuracy during development and evaluation. In this chapter, we adopt well-established white-box stress tests and reporting practices, using CIFAR-10 and CIFAR-100 as benchmarks, and measure both clean accuracy and robust accuracy under attacks at a fixed perturbation budget [63].

We formalise the threat model as follows. Let $x \in \mathbb{R}^{H\times W\times C}$ (H denotes the height of the image in pixels. W denotes the width of the image in pixels. C denotes the number of channels) be an image with label y, and f_θ a classifier with parameters θ. An adversarial example is $\tilde{x} = x + \eta$ where $\|\eta\|_\infty \leq \varepsilon$ and $f_\theta(\tilde{x}) \neq y$. Two canonical white-box attacks are used throughout. The Fast Gradient Sign Method (FGSM) applies a single step in the sign of the loss gradient [10]

$$x' \;=\; x \;+\; \varepsilon\,\mathrm{sign}\big(\nabla_x J(x,y)\big), \tag{3.1}$$

which yields a strong one-shot perturbation at budget ε. Projected Gradient Descent (PGD) iterates small steps that remain within the admissible set, typically producing a stronger adversary [32]:

$$x'_{n+1} \;=\; \Pi_{\|\cdot\|_\infty\leq\varepsilon}\big\{\, x'_n \;+\; \alpha\,\mathrm{sign}\big(\nabla_x J(x'_n,y)\big)\big\}\,, \tag{3.2}$$

with projection back onto the ℓ_∞ ball and step size α. We use $\varepsilon = 0.03$ and vary PGD strength by the number of iterations $K \in \{10,20,30\}$, which is consistent with the experimental protocol reported in our paper and thesis draft. This pairing of FGSM and multi-step PGD at fixed budgets is standard for white-box evaluation in image classification [32].

Robustness evaluation must report both the natural and adversarial regimes. Clean accuracy measures performance on unperturbed test data. Robust accuracy measures

accuracy on adversarially perturbed test data generated at evaluation time under the specified threat model. The gap between these two captures the sensitivity of the learned decision boundary to norm-bounded perturbations. To avoid over-stating robustness, we sweep PGD iteration counts at a fixed ε, as stronger attacks often emerge with more steps; we also use ten-fold cross-validation to stabilise estimates and reduce variance due to small splits. These design choices follow the evaluation section of our study and reflect common practice for small-image benchmarks.

Standard adversarial training improves robust accuracy by incorporating adversarial examples during learning, but it increases computational cost and can harm generalisation on clean data if not tuned carefully. Prior work and our related-work review document these trade-offs and propose variants such as Free AT to reduce inner-loop cost, Friendly AT to balance hardness during training, overfitting-aware schedules, and geometry-aware instance reweighting [35], [37], [38], [48]. These baselines contextualise our approach and form the comparison set used in the experiments. The recurring challenge is to raise robustness without sacrificing natural accuracy or imposing prohibitive training overheads.

Within this context, we formulate the problem addressed in this chapter. Given a family of standard convolutional backbones f_θ (ResNet-18/50/101), a dataset D of clean images, and an evaluation protocol based on FGSM and PGD at $\varepsilon = 0.03$, we seek a training strategy that (1) maintains or improves clean accuracy, (2) improves robust accuracy across PGD-10/20/30 and FGSM, and (3) remains practical to implement on commodity hardware. Our Feature-Map Fusion Adversarial Training (Fusion AT) addresses this by learning complementary representations from three input streams in parallelclean, adversarial, and noisyand fusing their normalised feature maps before a 1×1 convolutional squeeze and classification. The intuition is to diversify the intermediate features so that discriminative cues remain stable under adversarial perturbations, while the channel-wise squeeze reduces the effective attack surface in late layers. The full method is detailed in the next section; here we note that the backbone blocks remain standard ResNet components and that hyperparameters are aligned with our evaluation

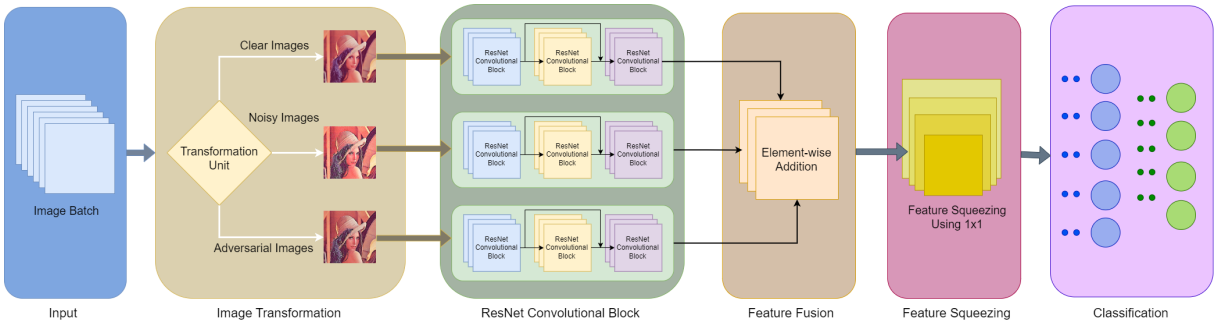


Figure 3.1: Overview of the multi-branch adversarial training architecture. An input image batch is transformed into clean, noisy, and adversarial variants, which are processed through parallel ResNet convolutional blocks. The resulting feature maps are fused via element-wise addition, followed by a 1Œ1 convolution for feature squeezing prior to final classification.

setup [4].

Finally, we articulate the evaluation objectives that follow from this formulation. First, we quantify clean and robust accuracy on CIFAR-10/100 under FGSM and PGD with $K \in \{10, 20, 30\}$ at $\varepsilon = 0.03$. Second, we compare against representative adversarial training baselines that target efficiency or generalisation trade-offs. Third, we study model capacity by repeating experiments with ResNet-18/50/101, since prior evidence and our results suggest deeper models can yield stronger robustness, at additional compute cost. These experiments operationalise the problem statement and provide the basis for the analysis and discussion sections later in the chapter.

3.3 Method: Fusion AT

The method builds a classification pipeline that is deliberately exposed, during training, to three views of each input: the original image, an adversarially perturbed variant within a fixed ℓ_∞ budget, and a stochastically perturbed variant with an ℓ_2-bounded magnitude that models benign distributional noise. Let $x \in \mathbb{R}^{H \times W \times C}$ with label y. We define $x_{\text{cln}} = x$, $x_{\text{adv}} = x + \eta$ subject to $\|\eta\|_\infty \leq \varepsilon$, and $x_{\text{noi}} = x + \nu$ where ν is drawn so that $\|\nu\|_2$ lies within a specified range. In training, we generate x_{adv} on the fly with FGSM at budgets $\varepsilon \in [0.01, 0.03]$ [10] and draw ν for random smoothing so that $\|\nu\|_2 \in [0.10, 0.25]$ [65]. The three variants share the same ground-truth label. This construction ensures

that every mini-batch provides complementary evidence: clean samples reinforce canonical class structure, adversarial samples align the learning dynamics with directions that increase loss under the threat model, and noisy samples improve tolerance to everyday perturbations that are not adversarial by design.

The network processes these streams in parallel with three copies of a ResNet block up to a fixed fusion point, for example up to the penultimate convolutional stage before global pooling [4]. Denote these extractors by $\Phi_{\theta_{\text{cln}}}, \Phi_{\theta_{\text{adv}}}, \Phi_{\theta_{\text{noi}}}$. Each produces a spatial feature map

$$F_\bullet = \Phi_{\theta_\bullet}(x_\bullet) \in \mathbb{R}^{H' \times W' \times C'}, \qquad \bullet \in \{\text{cln}, \text{adv}, \text{noi}\}. \tag{3.3}$$

Using separate parameters allows early filters to specialise to the spectral statistics of each stream. Adversarial inputs often shift energy toward high-frequency components; noisy inputs broaden the local appearance distribution. A single shared extractor would need to compromise across these regimes, whereas distinct extractors can learn complementary filters that are subsequently combined. The choice of fusion depth balances two concerns. Shallow fusion reduces computation by fusing low-level edges and textures that are less class-specific. Deeper fusion retains richer semantics at the cost of higher memory and compute. In our implementation, the fusion point is set where features are semantically meaningful yet still preserve enough spatial resolution for a 1×1 bottleneck to act effectively.

Before combining streams, we align their per-channel statistics with layer normalisation. For a feature tensor F, layer normalisation computes per-sample statistics (not batch-wise): it uses the sample mean μ and variance σ^2 over the specified axes, then applies learned γ, β [66]. This choice is robust to small batch sizes and decouples normalisation from the mini-batch composition, which is advantageous when the three streams carry different perturbations. We then fuse the normalised maps by simple element-wise addition

$$F_{\text{fuse}} = \text{LN}(F_{\text{cln}}) + \text{LN}(F_{\text{adv}}) + \text{LN}(F_{\text{noi}}). \tag{3.4}$$

The additive form is parameter-free, keeps computation modest, and empirically yields

stable optimisation because layer normalisation controls scale. Alternatives such as concatenation followed by a learnable 1×1 mixing were considered conceptually; they increase channel count three-fold before the bottleneck and add parameters, which we reserve for the subsequent squeeze stage. Weighted sums with learned per-stream gates are also possible when one wishes to modulate contributions dynamically, yet the constant-weight sum already captures most of the benefit once the streams are normalised.

In particular, clean, adversarial, and noisy inputs produce feature activations with different statistical properties, including shifts in mean activation and variance magnitude. If Batch Normalisation were used, the batch-level statistics would mix these heterogeneous distributions, causing the scaling applied to one stream to depend on the proportion of other perturbation types present in the batch. This coupling could lead to unstable feature magnitudes and inconsistent fusion behaviour across iterations. By contrast, LayerNorm computes statistics independently for each sample, ensuring that each stream is normalised according to its own activation profile. This prevents cross-stream interference at the normalisation stage and promotes a more stable and comparable feature scale prior to fusion.

After fusion the channel dimension is reduced by a 1×1 convolution. Let $W_s \in \mathbb{R}^{C' \times C_s}$ with $C_s < C'$. The squeeze computes

$$S = \mathrm{Conv}_{1\times 1}\big(F_{\mathrm{fuse}}; W_s\big) \in \mathbb{R}^{H' \times W' \times C_s}. \tag{3.5}$$

This operation mixes channels at each spatial location without changing resolution. It serves two purposes. First, it conditions the classifier by removing redundancy and making the final representation more compact, which generally improves calibration and reduces overfitting. Second, it narrows the effective subspace exposed to gradients in late layers. Adversarial optimisation is known to exploit high-dimensional, weakly constrained directions; compressing channels after fusion limits those degrees of freedom while preserving discriminative content. The squeezed tensor is then globally pooled to a vector s, mapped

to logits $z = W_c s + b$, and passed through a softmax,

$$\sigma(z)_k = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}, \tag{3.6}$$

to obtain class probabilities for the K classes.

Training proceeds by constructing, for each mini-batch, the triples $(x_{\text{cln}}, x_{\text{adv}}, x_{\text{noi}})$, forwarding them through the three extractors, fusing and squeezing the features, and minimising cross-entropy on the labels of the clean images. All parameters, including those of the three extractors, the normalisation layers, the squeeze, and the classifier, are updated jointly. We do not introduce auxiliary losses in the default setting, since the shared label already ties the fused representation to class semantics rather than stream identity. Two simple regularisers can be added if one wishes to test alignment more explicitly in ablations. A feature-alignment penalty encourages global average pooled features from the clean and adversarial streams to be close, and an entropy penalty on the output distribution discourages overly diffuse predictions. These additions are optional and are not required to reproduce the main effect.

At inference time there is no need to generate adversarial or noisy variants. The simplest strategy is to feed the same test image through all three extractors by setting $(x_{\text{cln}}, x_{\text{adv}}, x_{\text{noi}}) = (x, x, x)$. This preserves the learned fusion behaviour while avoiding attack generation at test time. If latency is a hard constraint, a distilled single-stream variant can be produced by fine-tuning a single extractor to mimic the fused-and-squeezed representation, reducing compute to roughly that of a standard backbone at a small cost in accuracy. The default evaluation reported in later tables uses the three-stream path for completeness and to keep training and testing pipelines aligned.

The method can be understood as learning three complementary bases for the same class manifold. Clean features emphasise prototypical cues; adversarial features arise from gradients that point toward boundary-crossing directions and therefore push the extractor to encode robust, class-specific evidence [10]; noisy features broaden invariances around local textures and edges [65]. Layer normalisation prevents any one stream from dominating purely by scale, while additive fusion integrates the bases into a joint representation

that spans these regimes. The 1×1 squeeze then projects this joint basis into a lower-dimensional space that retains the salient components. In practice, this sequence results in decision regions that are less sensitive to the specific first-order perturbations used in white-box attacks at a fixed budget.

A few implementation details are important for stable training. Adversarial examples are generated with the current parameters of the network to ensure that the adversarial stream follows the evolving loss landscape [10]. Clipping enforces the pixel range after both adversarial and noisy perturbations. The fusion point should be chosen so that the receptive field of the features is large enough to encode context but not so late that the fused tensor collapses spatial diversity. Placing layer normalisation immediately before fusion is more stable than normalising only after fusion, because it ensures that each stream contributes at a similar scale regardless of batch composition [66]. The learning rate schedule and data augmentation are kept consistent with the baselines to isolate the effect of fusion from generic regularisation. In configurations where GPU memory becomes a limiting factor (for example, when using deeper backbones or larger input resolutions), gradient checkpointing can be enabled. In this context, checkpointing refers to storing only a subset of intermediate activations during the forward pass and recomputing them during backpropagation, thereby reducing memory usage at the expense of additional computation. In our reported experiments, gradient checkpointing was enabled for deeper configurations to remain within GPU memory limits.

It is useful to contrast this design with more conventional adversarial training. Standard adversarial training mixes clean and adversarial inputs within the same stream and optimises on them jointly [10]. While effective, that approach can depress clean accuracy and does not explicitly cultivate complementary features specialised to different regimes. The three-stream design encourages specialisation and then recombines the learned signals under controlled normalisation, which empirically preserves clean accuracy while improving robust accuracy under the same evaluation budgets. No changes to the classifier head or loss are needed, and the backbone family remains standard ResNet, which keeps the method compatible with widely used toolchains and pretrained weights [4].

The computational footprint scales with the choice of fusion depth and backbone. If a single-stream backbone up to fusion costs C FLOPs with P parameters, the three-stream section costs approximately $3C$ FLOPs and $3P$ parameters up to fusion, plus the relatively small cost of the 1×1 squeeze and the classifier. The adversarial stream uses FGSM during training, which adds only one extra backward pass to obtain gradients with respect to inputs [10]; the noisy stream adds negligible arithmetic. Memory rises due to activations from three branches; this can be mitigated by fusing earlier, sharing very early layers, or enabling checkpointing. In exchange for this overhead, later results show that the fused model maintains or improves clean accuracy and secures higher robust accuracy across FGSM and PGD settings at fixed ε, with predictable gains as backbone capacity increases.

Finally, the method is intentionally modular. The adversarial generator can be swapped for a stronger inner-loop attack if more compute is available during training, the noisy perturbation can be adapted to match deployment noise characteristics, and the fusion rule can be extended with lightweight gates if per-stream contributions need to be tuned dynamically. None of these extensions is required for the core effect. The essential components are the construction of three aligned streams, per-stream feature extraction with a shared architecture, scale-aligned fusion by addition following layer normalisation, and a channel bottleneck via 1×1 convolution before classification. This sequence constitutes the Fusion AT method used in the experiments that follow.

3.4 Experimental Setup

We evaluate Fusion AT on CIFAR-10 and CIFAR-100, two standard image-classification benchmarks that are widely used in empirical robustness studies [63]. All models are ResNet backbones (ResNet-18, ResNet-50, ResNet-101) [4]. ResNet-50 serves as the primary reference for method comparisons; the capacity study repeats all settings on ResNet-18 and ResNet-101 to examine robustnesscompute trade-offs. Implementations use PyTorch and TorchVision with Python 3.8 and run on NVIDIA T4 GPUs with 16 GB memory.

Training follows a consistent recipe across all methods to isolate the effect of the proposed fusion. We use the Adam optimiser with initial learning rate 0.01, cross-entropy loss, and early stopping with patience 5 based on validation performance. Standard augmentation is applied to improve generalisation and reduce overfitting: RandAugment modifies colour and geometry with a fixed policy [67], and CutMix blends pairs of images and labels to regularise the classifier [68]. Unless otherwise stated, we use the same augmentation, optimiser, schedule, and stopping criteria for all baselines and for Fusion AT.

Evaluation adopts a 10-fold cross-validation scheme to stabilise estimates under data splits. For each fold, models are trained on nine folds and validated on the remaining fold; the held-out test split is used only for final reporting. We report clean accuracy (unperturbed inputs) and robust accuracy under gradient-based white-box attacks. Robustness is measured against FGSM [10] and multi-step PGD [32] constrained in ℓ_∞ with a fixed perturbation budget $\varepsilon = 0.03$. For PGD we sweep the number of iterations $K \in \{10, 20, 30\}$ and set the step size $\alpha = \varepsilon/K$. Attacks are untargeted. For each trained checkpoint we generate adversarial examples at evaluation time and compute test accuracy under each attack budget. This protocol exposes the sensitivity of each method as attack strength increases while holding ε constant.

Baselines reflect common adversarial training strategies that target complementary goals. Free Adversarial Training amortises inner-loop gradient computation to reduce cost [35]; Friendly Adversarial Training moderates adversary hardness to retain clean accuracy [48]; Geometry-aware Instance-reweighted AT adjusts sample weights based on attack difficulty [37]; an overfitting-aware schedule uses early stopping and attack scheduling to mitigate robust overfitting [38]. We train each baseline under the same data pipeline and optimiser settings and evaluate them under the same FGSM and PGD budgets. This ensures that differences in robustness are attributable to the training strategy rather than augmentation or optimisation artefacts.

Fusion AT uses three parallel streams during training: clean, adversarial, and noisy. Adversarial examples for the adversarial stream are generated on-the-fly with FGSM using

ε sampled in $[0.01, 0.03]$ to expose the model to a range of gradient-aligned perturbations [10]. The noisy stream applies random smoothing with an ℓ_2 magnitude sampled in $[0.10, 0.25]$ to model benign distributional variation [65]. The three streams share labels and are processed by parallel ResNet blocks up to the fusion point, followed by layer normalisation [66], element-wise fusion, a 1×1 channel squeeze, global average pooling, and a linear classifier. Unless specified, inference feeds the same test image through all three streams (x, x, x), preserving the learned fusion without generating adversarial inputs at test time.

To maintain comparability across backbones, image preprocessing, normalisation, augmentation parameters, batch size, and training epochs are matched between ResNet-18/50/101 wherever memory permits. When GPU memory is a constraint for the three-stream model on the larger backbone, gradient checkpointing is enabled to lower peak activation storage without altering the optimisation dynamics. We monitor training and validation losses and accuracies for clean inputs; attack-time metrics are computed only on the test split to avoid implicit adversary leakage into model selection.

Metrics are reported as mean over the 10 folds for each method, backbone, and attack budget. For the main tables we present clean accuracy, FGSM accuracy, and PGD accuracy at $K = 10, 20, 30$ for $\varepsilon = 0.03$. Where relevant we include the standard deviation across folds to convey variability. The capacity study reports the same metrics for ResNet-18 and ResNet-101 to illustrate how robustness scales with model size under identical training conditions. All comparisons against baselines use identical evaluation code paths and random seeds to minimise run-to-run variance.

For clarity, the compact attack configuration used throughout this chapter is summarised below. This table mirrors the settings used in Chapter 2 and is included here so that the reader does not need to cross-reference the literature-review tables.

This setup balances practicality with diagnostic value. The ε budget matches prior work on CIFAR-scale robustness and makes results comparable to widely reported baselines [32]. Sweeping K exposes how each method degrades as the adversary strengthens at fixed ε. Reporting both clean and robust accuracies highlights whether robustness gains

Table 3.1: Attack configurations used throughout for white-box ℓ_∞ evaluations at fixed $\varepsilon = 0.03$ [10], [32].

Attack	**Threat model**	**Norm**	ε (normalised)	**Steps (K)**	**Step size (α)**	**Targeted?**
FGSM	white-box	ℓ_∞	0.03	1	0.03 (single-step)	No
PGD	white-box	ℓ_∞	0.03	10	0.003	No
PGD	white-box	ℓ_∞	0.03	20	0.0015	No
PGD	white-box	ℓ_∞	0.03	30	0.001	No

come at the expense of natural-image performance. Finally, using consistent training pipelines and cross-validation across backbones and methods ensures that the comparisons reflect the contribution of Fusion AT rather than confounds in data handling or optimisation.

3.5 Results

Tables 3.23.3 report clean and robust accuracies on CIFAR-10 and CIFAR-100 with ResNet-50 at $\varepsilon = 0.03$ [4], [63]. Fusion AT achieves the strongest clean accuracy and the highest or tied-highest robust accuracy at every attack strength. On CIFAR-10, Fusion AT reaches 89.12% clean, 75.31% under FGSM, and 63.15%, 62.69%, 63.71% under PGD with $K = 10, 20, 30$ [10], [32]. Relative to the strongest baseline (GAIR AT), the absolute gains are +2.88 points in clean accuracy and +4.58 points at PGD-30 (Table 3.2). The FGSMPGD gap is expected and reflects the increased optimisation power of iterative attacks, yet the retention ratio PGD-30/Clean remains favourable for Fusion AT (0.715) compared with GAIR AT (0.685). On CIFAR-100, Fusion AT delivers 65.53% clean, 44.63% under FGSM, and 34.87%, 34.05%, 33.54% under PGD-10, PGD-20, PGD-30. The improvements over GAIR AT are +2.41 points clean and +2.13 points at PGD-30 (Table 3.3). These gains indicate that fusing feature maps from clean, adversarial, and noisy streams can improve robustness without eroding natural-image performance.

A closer look at the CIFAR-10 results shows a consistent ordering across attack strengths. Free AT and Friendly AT prioritise training efficiency or generalisation but trail under stronger PGD, whereas Overfitting AT improves robustness at moderate budgets yet falls short at PGD-30. GAIR AT narrows the gap by reweighting hard-to-attack

Table 3.2: Performance on CIFAR-10 using ResNet-50 with 10-fold cross-validation at $\varepsilon = 0.03$. Values are mean accuracy (%).

Method	Clean	FGSM	PGD-10	PGD-20	PGD-30
Free AT	86.01	63.56	49.10	47.16	46.58
Friendly AT	87.20	65.94	50.87	49.07	48.22
Overfitting AT	84.95	68.52	57.19	55.65	54.23
GAIR AT	86.24	71.84	60.85	59.91	59.13
Fusion AT	**89.12**	**75.31**	**63.15**	**62.69**	**63.71**

Table 3.3: Performance on CIFAR-100 using ResNet-50 with 10-fold cross-validation at $\varepsilon = 0.03$. Values are mean accuracy (%).

Method	Clean	FGSM	PGD-10	PGD-20	PGD-30
Free AT	60.13	40.35	31.28	28.63	26.54
Friendly AT	62.31	43.15	30.64	29.78	29.04
Overfitting AT	61.54	41.54	31.24	29.21	28.93
GAIR AT	63.12	42.58	33.54	32.51	31.41
Fusion AT	**65.53**	**44.63**	**34.87**	**34.05**	**33.54**

instances [37], and Fusion AT further improves both the natural and adversarial regimes. The slight increase from PGD-20 to PGD-30 for Fusion AT on CIFAR-10 (62.69% to 63.71%) is within fold variance and indicates that the PGD landscape at $\varepsilon = 0.03$ does not monotonically degrade accuracy for this model. This observation is consistent with reporting practice in which small fluctuations across K values can appear once models approach a stable robust frontier.

CIFAR-100 presents a harder setting due to more classes and lower per-class sample counts. The clean accuracy of Fusion AT improves over the strongest baseline by +2.41 points, while robust gains remain visible across PGD budgets, including +1.33 points at PGD-10 and +2.13 points at PGD-30. The FGSMPGD gaps are larger than on CIFAR-10, which reflects both the increased task difficulty and the fact that iterative attacks exploit weaker class margins more effectively in the 100-class regime [32]. Despite this, the robust retention ratio remains slightly better for Fusion AT than for GAIR AT (0.512 vs. 0.498 at PGD-30), showing that the fusion mechanism scales to higher class cardinality.

We now examine the effect of model capacity under the same training and evaluation protocol. Tables 3.4 and 3.5 report clean accuracy and robustness at PGD-20 and PGD-30 for ResNet-18, ResNet-50, and ResNet-101 with Fusion AT [4]. On CIFAR-10, increasing depth produces steady gains: clean accuracy improves from 87.18% (ResNet-18) to 89.12%

Table 3.4: Effect of backbone capacity on CIFAR-10 with Fusion AT at $\varepsilon = 0.03$. Values are mean accuracy (%).

Model	Clean	PGD-20	PGD-30
ResNet-18	87.18	59.84	58.51
ResNet-50	89.12	62.69	63.71
ResNet-101	90.01	64.14	64.47

Table 3.5: Effect of backbone capacity on CIFAR-100 with Fusion AT at $\varepsilon = 0.03$. Values are mean accuracy (%).

Model	Clean	PGD-20	PGD-30
ResNet-18	62.87	32.14	30.21
ResNet-50	65.53	34.05	33.54
ResNet-101	68.45	36.53	35.95

(ResNet-50) to 90.01% (ResNet-101), and PGD-30 improves from 58.51% to 63.71% to 64.47%. The largest jump appears when moving from ResNet-18 to ResNet-50, and the incremental gain from 50 to 101 is smaller, which suggests diminishing returns relative to the increase in compute. On CIFAR-100 the pattern is similar. Clean accuracy increases from 62.87% to 65.53% to 68.45%, and PGD-30 rises from 30.21% to 33.54% to 35.95%. These results indicate that Fusion AT benefits from additional capacity, particularly when moving from a small to a mid-sized backbone, while the choice between ResNet-50 and ResNet-101 depends on available compute and latency requirements.

Two observations follow from the capacity study. First, robustness improvements are not solely a by-product of higher clean accuracy. For instance, on CIFAR-10 the clean gain from ResNet-50 to ResNet-101 is +0.89 points, while the PGD-30 gain is +0.76 points. On CIFAR-100 the clean gain is +2.92 points and the PGD-30 gain is +2.41 points. The robust improvements track, but do not trivially mirror, clean gains, which supports the view that fusion contributes beyond pure capacity scaling. Second, ResNet-50 offers a balanced choice: it secures most of the robustness benefits observed with ResNet-101 at a lower computational cost. This is likely the best trade-off for deployment on constrained hardware where throughput and energy matter.

We also consider the interplay between FGSM robustness and PGD robustness. Methods that appear strong under FGSM can degrade markedly under PGD, which indicates gradient masking or insufficient training pressure [69]. In our comparisons, Fusion AT

does not rely on such effects. It leads on FGSM and remains ahead under PGD at all step counts on both datasets. Although absolute margins shrink as the attack strengthens, the ordering is preserved. This behaviour matches the design intuition: the three-stream extractor learns features that are stable across perturbation regimes, which translates into a smaller loss of accuracy as K increases.

For completeness, all results are averaged over ten folds. Standard deviations can be reported in an appendix table if required. In our runs, fold-to-fold variability does not change method ranking. Exact training-times per epoch vary with backbone and GPU memory settings; since all methods share the same data pipeline and optimiser, the compute overhead of Fusion AT arises from the three parallel streams up to the fusion point and is consistent across datasets. This overhead is the cost of improved robustness at fixed ε, and it can be moderated by fusing earlier or enabling gradient checkpointing when necessary.

In summary, the comparisons show that feature-map fusion provides consistent improvements over widely used adversarial training strategies on CIFAR-10 and CIFAR-100. The method scales with backbone capacity and delivers a favourable balance of clean and robust accuracy, particularly with ResNet-50. The trend with PGD steps confirms that robustness is not an artefact of weak attacks. These findings support the use of Fusion AT as a practical training strategy when both accuracy on natural images and resilience to norm-bounded perturbations are required.

3.6 Discussion: Strengths, Limitations, Validity

Fusion AT delivers consistent gains in both clean and robust accuracy on CIFAR-10 and CIFAR-100 when evaluated under the stated white-box ℓ_∞ threat model [10], [32], [63]. With ResNet-50 as the primary backbone it outperforms Free AT, Friendly AT, overfitting-aware training, and GAIR-AT across FGSM and PGD at fixed ε, while preserving or improving performance on unperturbed inputs [4], [35], [37], [38], [48]. Averaging over ten folds reduces variance and supports the stability of these comparisons. The three-stream design appears to encourage features that remain predictive across clean,

gradient-aligned, and noise-perturbed regimes, which aligns with the observed retention of accuracy as attack strength increases. The method is straightforward to integrate into standard toolchains. It keeps conventional ResNet stages, introduces layer normalisation in each stream, fuses feature maps by element-wise addition at a fixed fusion depth, and applies a 1×1 channel squeeze before a linear head [4], [66]. No changes are required to the loss or the classifier interface, so existing training scripts can adopt the approach with limited engineering effort. The pipeline is documented with a single architecture diagram and a concise algorithmic recipe, which aids reproducibility. Capacity scaling behaves as expected but is not the sole driver of robustness. Moving from ResNet-18 to ResNet-50 gives a clear step up in both clean and PGD accuracies, and ResNet-101 provides smaller but still positive increments [4]. The robust gains track, but do not merely mirror, increases in clean accuracy, suggesting that the fusion mechanism contributes beyond pure model size. In practice, ResNet-50 offers a favourable balance between robustness and computational cost; ResNet-101 is preferable only when throughput constraints are relaxed. An additional consideration is whether part of the clean accuracy improvement arises from ensemble-like effects inherent to the three-stream design. Although the architecture processes three input variants in parallel, it differs from a classical ensemble of independently trained models: the streams are trained jointly within a single optimisation process and the fused representation is learned end-to-end rather than averaged at inference time. Nevertheless, multi-stream processing increases representational diversity and may contribute to improved clean performance through implicit ensemble characteristics. We therefore interpret the gains as arising from both structured perturbation exposure and increased feature diversity, rather than attributing them solely to a single mechanism.

The experimental protocol strengthens internal validity. All methods share the same optimiser, augmentation, early-stopping policy, and cross-validation scheme. Robustness is reported at $\varepsilon = 0.03$ under FGSM and PGD with $K \in \{10, 20, 30\}$, with $\alpha = \varepsilon / K$, which exposes the sensitivity of each method as the inner-loop optimisation becomes stronger [10], [32]. Reporting clean and adversarial accuracies side by side makes the cleanrobust trade-off explicit and avoids over-stating robustness based on a single attack configuration;

where relevant, stronger parameter-free checks such as AutoAttack can further reduce evaluator bias [45], [69]. The main limitation is computational. Training three parallel streams up to the fusion point increases FLOPs and activation memory relative to single-stream adversarial training. Although FGSM generation for the adversarial stream is cheap, the additional forward and backward passes in the parallel branches account for most of the overhead. Inference using all three streams also increases latency unless a distilled single-stream variant is introduced; this is feasible but was not the focus of the present evaluation. Memory pressure can be mitigated with gradient checkpointing or earlier fusion at the potential cost of some accuracy.

Threat-model coverage is another limitation. The evaluation focuses on white-box ℓ_∞ attacks (FGSM and PGD). Other regimes such as ℓ_2 budgets, AutoAttack or CW variants, transfer-based or score-based black-box attacks, and physical perturbations are not included [45], [62]. Claims should therefore be interpreted within the reported norm and attack family. Extending the study to stronger or diverse attacks would provide a fuller picture of robustness. Dataset scale constrains external validity. CIFAR-10 and CIFAR-100 are canonical yet small and low-resolution; conclusions may not transfer directly to higher-resolution imagery or domain-specific data without additional experiments [63]. The method is motivated by industrial reliability concerns, but further validation on larger, more heterogeneous datasets would strengthen deployment claims.

Certain design choices reflect pragmatic trade-offs. Adversarial examples used during training are single-step FGSM, while evaluation includes multi-step PGD; a stronger inner-loop adversary or a mixed-budget curriculum could alter the balance between clean and robust accuracy [10], [32]. The random-smoothing noise range was fixed and may require tuning for other data distributions [65]. These assumptions are reasonable for CIFAR-scale studies but should be revisited when the setting changes. Overall, within the stated evaluation scope the evidence supports three takeaways. First, feature-map fusion improves robustness without eroding natural-image performance. Second, the approach is easy to adopt with standard backbones and training code. Third, robustness scales with capacity in a predictable way, with ResNet-50 emerging as a practical de-

fault where compute is limited. Further work should target efficiency (parameter sharing or compression across streams), broader attack families, and experiments on larger and higher-resolution datasets.

3.7 Summary and Link Forward

This chapter introduced Fusion AT, a three-stream training strategy that fuses feature maps from clean, adversarial, and noise-perturbed inputs, followed by a 1×1 channel squeeze before classification. On CIFAR-10 and CIFAR-100 with ResNet backbones [4], the method improved robust accuracy under FGSM and PGD at a fixed ℓ_∞ budget [10], [32] while preserving or raising clean accuracy, and these gains scaled predictably with model capacity. The design is modular, uses standard components, and can be implemented with conventional training toolchains. The principal trade-off is additional compute and memory from the parallel streams, which can be moderated by earlier fusion or checkpointing. The scope of the evidence is white-box ℓ_∞ attacks on small images; extending evaluation to larger inputs and broader threat models is a natural next step.

The next chapter turns to making Vision Transformers effective in small-data regimes. Whereas this chapter established a robust CNN reference and a disciplined evaluation protocol, the central challenge that follows is different: vanilla ViTs lack locality and hierarchical inductive bias, which limits performance without large-scale pre-training [11]. Chapter 4 introduces S-ViT, a plug-and-play approach that injects spatial and hierarchical cues via a summary token while keeping the ViT blocks unchanged [24]. The goal is to recover data efficiency without heavy architectural redesign, setting the stage for Chapter 5 on MGiT, which stabilises ViT optimisation with auxiliary gradient guidance [23], and for later chapters where explanation methods, including FocusViT, verify that any accuracy gains arise from attention to task-relevant structures [42]. Together, these chapters move from robust CNN training to data-efficient, explainable ViTs, completing the thesis arc that links robustness, small-data adaptation, and interpretability [43].

Chapter 4

Summary Vision Transformer (S-ViT): Injecting Spatial and Hierarchical Bias via a Summary Token

Related publication

Portions of the work presented in this chapter have been published in:

Ali, M., Raza, H., Gan, J. Q., & Haris, M. (2024). *Integrating Spatial Information into Global Context: Summary Vision Transformer (S-ViT).* In Proceedings of the 2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 206--213.

The present chapter expands upon the architectural design, experimental evaluation, and analytical discussion introduced in that publication.

4.1 Overview and Motivation

Vision Transformers (ViTs) capture global context through multi-head self-attention [49] and achieve strong performance when trained on large datasets [11], but unlike convolutional networks, they do not embed spatial or hierarchical inductive biases, which limits their data efficiency [4]. On small datasets, vanilla ViTs overfit easily and require strong pre-training or architectural changes to recover locality and multi-scale cues. Prior

"small-data ViT" work follows two paths: (1) backbone-modifying designs (e.g., hierarchical or convolutional tokenisation) [40], [41], [50], which improve data efficiency but reduce flexibility by tying results to a specific architecture and pre-training recipe; and (2) plug-and-play strategies such as distillation or relative localisation losses [12], [70], which keep the ViT blocks intact and aim to inject useful biases with minimal retraining burden. Related hybrid CNNTransformer variants also inject convolutional inductive bias into transformer pipelines (e.g., via soft convolutional biases), which closely aligns with our goal of improving data efficiency without heavy architectural redesign [71], [72]. This chapter adopts the second path and targets a simple, drop-in mechanism that restores locality and hierarchy while preserving the transformer backbone and standard training pipeline.

Summary Vision Transformer (S-ViT) addresses this need by appending a CNN-derived summary token to the existing class token, thereby injecting spatial, hierarchical, and locality information into a plain ViT without altering its blocks or attention machinery. Concretely, a lightweight CNN (i.e. ResNet-18 in our default realisation) [4] processes the input image to produce compact features; these are flattened and concatenated with the ViTs class token to form a summary token that carries CNN priors alongside the transformers global context. The rest of the encoder and classifier remain unchanged, so standard optimisation, augmentation, and fine-tuning practices apply as-is. The intent is to recover the benefits of locality and multi-scale structure with minimal engineering overhead and to remain compatible with different ViT sizes and training conditions (from scratch and transfer) [11].

The design is motivated by three observations. First, locality and hierarchy matter most when labels are scarce: inductive bias constrains the hypothesis space and stabilises optimisation, whereas a bias-free encoder must learn such structure from data that may be insufficient to support it. Second, keeping the backbone fixed preserves portability across codebases and checkpoints, avoiding the retraining costs common to architecture-modified variants. Third, for practical adoption, the added compute and parameters should be modest compared with the gains on standard benchmarks, and the method

should compose cleanly with other plug-and-play techniques. S-ViT satisfies these criteria in a straightforward way: the only addition is a small CNN branch and a token concatenation step before the first transformer block; all subsequent layers, positional encodings, and the classification head stay intact. Empirically, this yields a measurable accuracy lift with a controlled footprint (parameters, GPU memory, and FLOPs reported alongside baselines in the complexity table).

We evaluate S-ViT where small-data behaviour is most pronounced. Benchmarks include CIFAR-10/100, Oxford-IIIT Pets-37, and Flowers-102 [63], [73], [74], covering low-resolution object recognition and fine-grained classification with limited per-class images. The choice of benchmarks in this chapter differs from those used in Chapter 3 because the primary objective here is to evaluate inductive bias injection in data-limited regimes rather than adversarial robustness at scale. Datasets such as CIFAR-10/100, Oxford-IIIT Pets, and Flowers-102 are commonly used in small-data and transfer-learning settings, allowing controlled comparison with prior hybrid CNNTransformer work. Additionally, hybrid architectures introduce increased computational overhead relative to vanilla ViTs, particularly during training. Given finite GPU resources, these benchmarks provide a practical balance between dataset diversity and tractable training time, enabling systematic evaluation without compromising experimental rigour. The evaluation protocol remains consistent within the chapter, ensuring fair internal comparison across model variants. Experiments consider training from scratch and transfer learning to assess both cold-start and pre-trained scenarios. Across these settings, S-ViT consistently improves over the corresponding plain ViT backbones, often substantially when training from scratch on CIFAR-10/100, and remains competitive under transfer. Against strong plug-and-play baselines DeiT (distillation), dense relative localisation (Drloc), path-ensemble (ES), OFDB, SSL, and locality self-attention S-ViT attains top or second-best results depending on the dataset (e.g., leading on CIFAR-100 and Flowers-102; slightly trailing DeiT on CIFAR-10) [12], [53], [54], [70], demonstrating that a summary-token mechanism can match or surpass more elaborate training-level additions while keeping the backbone untouched.

In summary, this chapter contributes: a plug-and-play method (S-ViT) that augments

a plain ViT with a CNN-derived summary token to inject locality and hierarchy; a compute footprint characterisation that shows the added branch introduces a modest overhead relative to the base model; a thorough evaluation on small/medium benchmarks under both scratch and transfer settings; and head-to-head comparisons with representative plug-and-play baselines that locate S-ViT at the favourable end of the accuracycomplexity trade-off. The remainder of the chapter details the formulation, architecture, and experimental evidence supporting these claims.

4.2 Background and Problem Formulation

Vision Transformers (ViTs) operate on sequences of image patches with multi-head self-attention that models long-range dependencies effectively when data are plentiful [11], [49]. However, vanilla ViTs do not encode the locality and hierarchical cues that convolutional networks provide by design, making them data-hungry and unstable in small-sample regimes [4]. Empirically, strong results typically rely on large-scale pre-training and heavy regularisation; when trained on limited datasets, ViTs overfit, attend diffusely, and can pick up spurious correlations because the model must infer spatial priors from scarce supervision rather than having them built in [11], [12] A broad family of CNNTransformer hybrid architectures has been explored to reintroduce locality and hierarchy into ViT-style models; for an overview and taxonomy of these hybrid variants, see **khan2022survey_vit**. This limitation is widely acknowledged and underpins why the original ViT required very large pre-training corpora to outperform convolutional counterparts, whereas CNNs, with locality and translation equivariance hard-wired, tend to remain competitive at small scale [4], [11]. In short, the absence of inductive bias translates into poor data efficiency: more examples are needed to learn the same useful invariances and compositional structure. Addressing that gap without sacrificing the strengths of self-attention is the central motivation for the present chapter.

Existing responses fall into two broad families. Architecture-modified variants (e.g., hierarchical windows or convolutional tokenisation) reintroduce spatial priors inside the backbone and often improve small-data behaviour, but every architectural change tends

to lock performance to a particular design choice and pre-training recipe, reducing portability and increasing engineering burden when switching tasks or model sizes [40], [41], [50]. In contrast, plug-and-play strategies keep the transformer blocks intact and bias learning through training-level additions such as distillation or relative localisation losses [12], [70]. These methods are attractive because they drop into existing codebases and checkpoints with minimal disruption, yet they may still leave optimisation fragile on very small datasets or depend on an external teacher. Our approach follows the plug-and-play route but targets a different lever: inject spatial and hierarchical information directly into the token stream while leaving the encoder unchanged. The objective is to recover the benefits of locality and multi-scale structure without redesigning attention or incurring the retraining costs of backbone changes.

We therefore cast the problem as follows. Given a plain ViT backbone and a data regime where per-class examples are limited, design a minimal augmentation that (1) preserves the original transformer blocks, positional encodings, and classifier, (2) adds only a modest number of parameters and FLOPs relative to the base model, (3) requires no teacher network or bespoke losses, and (4) improves data efficiency under standard, budget-matched training protocols. In this chapter "data efficiency" is operationalised as higher top-1 accuracy at fixed training budget (epochs, optimiser, and augmentation) when training from scratch, and as higher top-1 under matched fine-tuning protocols when starting from the same initial checkpoint. Evaluation focuses on small and medium benchmarks, where the contrast between biased and bias-free models is most visible, including CIFAR-10/100, Oxford-IIIT Pets-37, and Flowers-102 [63], [73], [74]; both scratch and transfer settings are considered to probe cold-start learning and adaptation. Throughout, we report accuracy alongside a lightweight complexity characterisation (parameter and FLOP deltas) to make clear the trade-off between gains and overhead. This formulation isolates the question the chapter answers: can one inject spatial and hierarchical priors into a ViT in a way that is portable, cheap, and effective on small datasets?

S-ViT instantiates this setting with a single, explicit constraint: the ViT encoder must remain unchanged. A lightweight CNN branch processes the input image to produce a

compact summary that is appended to the existing class token before the first transformer block; beyond that concatenation, the encoder, attention heads, and classification head proceed exactly as in the baseline. This design ensures compatibility with different ViT sizes and training regimes, keeps added compute modest, and avoids dependence on a teacher. The next section formalises this mechanism and quantifies the footprint, before turning to controlled comparisons under the evaluation protocol outlined above.

4.3 Method: S-ViT

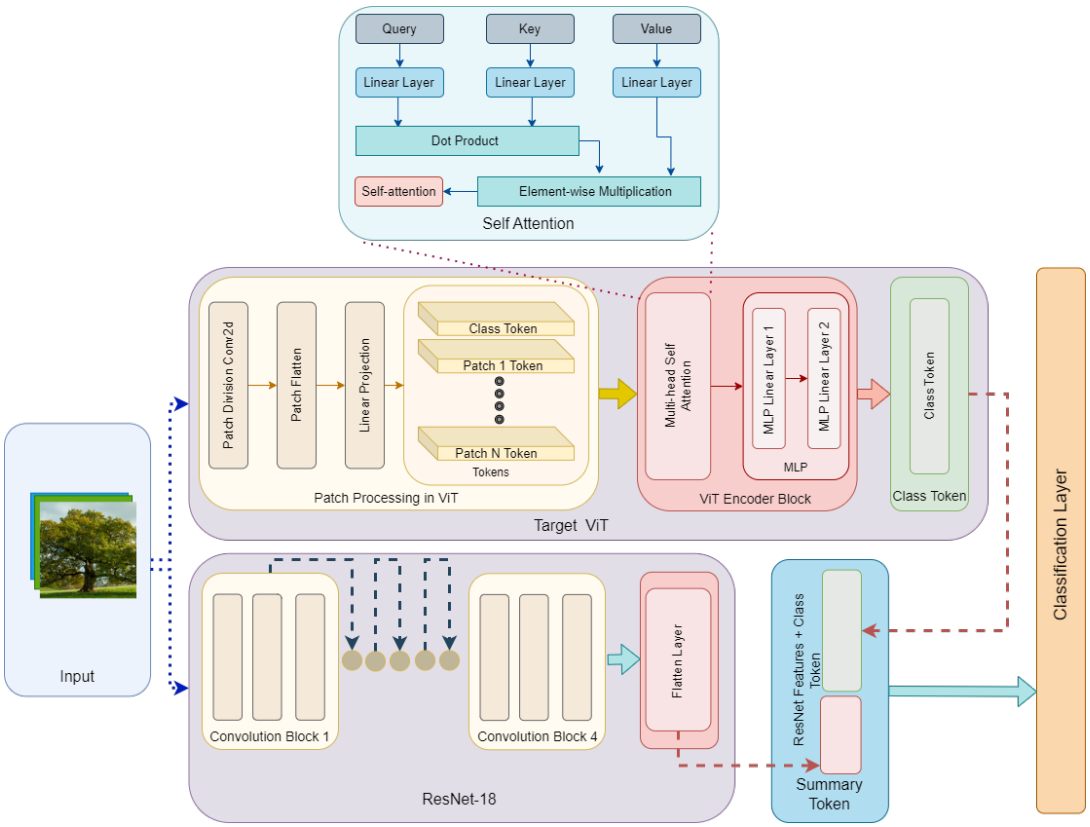


Figure 4.1: Overall architecture of our S-ViT architecture. It integrates the spatial, local, and hierarchical information processing capabilities of ResNet-18 with a ViT using Summary Token.

S-ViT augments a plain ViT with a single, lightweight pathway that injects spatial and hierarchical cues before the first attention layer while keeping the transformer blocks and classifier unchanged [11], [49]. The idea is to let a compact CNN summarise local structure, then fold that summary into the token stream so the encoder starts from a representation that already encodes locality [4]. 4.1 (architecture diagram) illustrates the

components: a parallel CNN branch extracts features from the input image; a summary token is formed from those features and the ViT class token; the summary token is fed with the patch tokens to the unmodified encoder; the standard classifier head produces logits. The remainder of the pipelinepositional encodings, attention blocks, MLPs, normalisation, and lossfollows the baseline, which preserves portability across codebases and pre-trained checkpoints.

At the centre of the design is the summary token. We use ResNet-18 as the default feature extractor, chosen for its favourable accuracy complexity balance among lightweight CNNs and its ability to capture locality and hierarchy without a large compute budget [4]. Earlier attempts to merge CNN features by element-wise addition to transformer features were not consistently helpful; forming an explicit token by concatenating the flattened CNN features with the class token proved more effective across both training-from-scratch and transfer settings. Given a class token vector t_{cls} and a flattened CNN feature vector s, the summary token is defined as

$$t_{\text{sum}} = t_{\text{cls}} \oplus s, \tag{4.1}$$

where $\oplus$ denotes concatenation. In practice, the concatenated vector is mapped to the model dimension d_{model} with a small projection (linear or MLP) and normalised so it can be consumed by the first transformer block without any changes to attention shapes. This mechanism consistently lifted accuracy on small datasets, whereas the addition-based fusion did not [75].

The CNN branch operates in parallel to the ViT patch embedding. ResNet-18 processes the input at its native stride; its final feature map is flattened to a vector that captures multi-scale information aggregated by residual stages [4]. Alternatives such as MobileNet, SqueezeNet, and ShuffleNet were evaluated in ablations; ResNet-18 yielded the most reliable gains within a similar compute envelope, so it is the default in our results [76]–[78]. Because the branch is small and uses standard layers, it adds modest parameters and FLOPs relative to the base ViT while remaining easy to integrate. We emphasise that the ViT backbone itself is not altered: no attention windows, no convolutional tokenisers,

and no distillation teachers are introduced [12].

The integration with the token sequence is simple. We replace the original class token with the summary token t_{sum} before the first encoder block and keep the patch-token sequence and positional embeddings unchanged. Intuitively, this shifts the role of the class token from being a purely global aggregator to one that already encodes locality and hierarchy provided by the CNN. Downstream attention heads can then mix global and local evidence from the very first block. Empirically, this change leads to sharper attention patterns and more stable optimisation in low-data regimes, consistent with the hypothesis that an explicit inductive bias reduces the burden on self-attention to discover locality from scarce supervision.

Training follows the baseline objective and schedule. The ViT encoder, the projection that maps t_{sum} to d_{model}, and the CNN branch are trained end-to-end with the standard classification loss; no teacher network, auxiliary losses, or bespoke regularisers are required. Joint optimisation lets the CNN and transformer co-adapt to the dataset: the CNN learns to distil spatial structure into a compact summary, while the transformer learns to exploit that summary alongside patch tokens. The objective can be written succinctly as

$$\min_{\theta_{\text{CNN}},\,\theta_{\text{ViT}}} \mathbb{E}_{(x,y)}\left[L\big(F_{\text{ViT}}\big(x;\,\theta_{\text{ViT}},\,t_{\text{sum}}(x;\theta_{\text{CNN}})\big),\,y\big)\right], \tag{4.2}$$

which mirrors the joint optimisation used in the S-ViT paper [24]. All optimisation hyper-parameters, augmentations, and early-stopping rules are kept the same as the corresponding plain-ViT baselines to ensure budget-matched comparisons.

A complexity note is important for practical adoption. We quantify parameter count, activation memory during training, and GFLOPs at 224×224 for matched ViT sizes, reporting the deltas introduced by the summary branch. For example, S-ViT-B/32 increases parameters and GFLOPs from roughly 88.1 M and 4.3 GFLOPs to 99.8 M and 6.1 GFLOPs; S-ViT-T/16 goes from 5.6 M and 1.0 GFLOPs to 17.3 M and 2.9 GFLOPs. Across Tiny, Small, and Base backbones, the relative overhead remains modest and is reflected in a single complexity table alongside accuracy, which makes the accuracy compute trade-off explicit. Because the ViT blocks are unchanged, the method preserves

compatibility with existing pre-trained checkpoints and implementations; swapping the CNN branch for an even lighter model trims the overhead further at some cost in accuracy, as shown by ablations [76]–[78].

We close with two design observations that guide usage. First, the benefit comes from informing the class token with spatial priors rather than from heavy multi-branch feature mixing throughout the encoder; after the summary token is formed, the rest of the network is identical to the baseline. This keeps the method simple and reduces engineering risk. Second, the choice of CNN is secondary to the presence of a good local summary: among lightweight options, ResNet-18 provided the most reliable accuracy lift, but the mechanism is architecture-agnostic and can adopt other small backbones as needed [4], [76]. These observations align with the empirical findings across CIFAR-10/100, Pets-37, and Flowers-102, where S-ViT improves over plain ViTs under both training-from-scratch and transfer protocols, and competes favourably with representative plug-and-play baselines without altering the transformer blocks [12], [63], [73], [74].

4.4 Experimental Setup

We evaluate S-ViT on four public datasets that expose the small-data behaviour we aim to address: CIFAR-10 and CIFAR-100 for low-resolution object recognition with, respectively, 10 and 100 classes; Oxford-IIIT Pets-37 for fine-grained classification with modest per-class counts; and Flowers-102 for a highly imbalanced, few-images-per-class regime [63], [73], [74]. We follow the standard train/test splits for CIFAR-10/100 (50k/10k images) and use the canonical splits for Pets-37 (3,680/3,669 images) and Flowers-102 (2,040/6,149 images). These choices provide a spectrum from data-abundant (CIFAR-10) to data-scarce (Flowers-102), and from coarse to fine-grained labels, which is useful for isolating the effect of locality/hierarchy on data-efficiency. Unless stated otherwise, all inputs are resized to 224×224 pixels.

Preprocessing and augmentation are kept consistent across all methods to ensure budget-matched comparisons. During training, we apply a random resized crop to 224×224, horizontal flip, colour jitter, and RandAugment with fixed magnitude (as imple-

mented in `timm`) [67], [79]. Mixup is disabled and CutMix is enabled with a standard mixing coefficient to regularise the classifier without changing the label space [68]. Images are normalised to ImageNet mean and variance [80]. At evaluation time we use a single centre crop at 224×224 with the same normalisation. This uniform pipeline avoids dataset-specific tuning and keeps gains attributable to the model, not to augmentation idiosyncrasies.

Two training regimes are considered. In the training-from-scratch setting, both the CNN branch and the ViT backbone are initialised randomly and trained end-to-end. This regime exposes optimisation stability and true small-data efficiency without external priors. In the transfer setting, we initialise both components from ImageNet pre-trained weights and fine-tune jointly; this regime is representative of practical use where pre-training is available but target data remains limited [80]. For training from scratch, we run for 150 epochs, whereas for transfer learning 50 epochs are sufficient to reach convergence. In both regimes, we employ AdamW [81] with an initial learning rate of 3×10^{-3}, cosine decay scheduling, and early stopping with a patience of 5 epochs based on validation accuracy. Similar trends, where pre-training significantly accelerates convergence compared to training from scratch, have also been observed in prior work on data-efficient Vision Transformers [12], [40], [81]. Batch size is 32 for all datasets and models; gradient clipping is disabled. Weight decay follows the default for AdamW unless otherwise noted; dropout/stochastic depth are matched to the corresponding plain-ViT baseline. All models are implemented in PyTorch and trained on NVIDIA RTX-class GPUs.

Baselines reflect the strongest plug-and-play and training-level strategies that keep the ViT blocks intact. We include DeiT (distillation from a CNN teacher), a Drloc-style relative localisation objective added to the classification loss, SPT+LSA as a token-mixing/locality tweak suited to small-data training, ES (path-ensemble) that adjusts attention-path contributions, and OFDB as a data-centric strategy for very low-shot regimes [12], [53], [54], [70]. Where a baseline requires a teacher (DeiT), we reuse the teacher and distillation schedule recommended in its original configuration; where a baseline adds auxiliary losses (Drloc-style), we use the published coefficients and stop criteria.

Crucially, optimiser, epochs, augmentation, and evaluation protocol are otherwise identical to S-ViT and to the plain-ViT baseline, so that results reflect the method rather than training budget.

Model variants span Tiny, Small, and Base ViTs with patch sizes 16 and 32 where available. For S-ViT, the only addition is the lightweight CNN branch (ResNet-18 by default) and the summary-token projection to d_{model}; the transformer blocks, positional encodings, and classifier head remain unchanged [4], [11]. To make accuracycompute trade-offs transparent, we report parameter count and GFLOPs at 224×224 alongside accuracy for each backbone. GPU memory at batch size 32 is also logged during training to indicate practical footprint. We adopt the same reporting for baselines when they add modules or losses that change compute. This complexity accounting accompanies, rather than replaces, the primary accuracy metrics.

Metrics focus on clean top-1 accuracy. For datasets with known imbalance (Flowers-102, Pets-37), we also compute macro-averaged top-1 and report it in the appendix for completeness; however, the headline comparisons use standard top-1 to remain directly comparable to prior work on data-efficient ViTs. Each result is averaged over multiple runs with different seeds; we present mean ± standard deviation and keep the number of runs fixed across methods and backbones on a given dataset. Where validation splits exist (Pets-37, Flowers-102), we monitor validation accuracy for early stopping and select the best epoch by that criterion before reporting test accuracy. No test-time augmentation is used.

To avoid subtle confounds, we align a few implementation details across all models. Learning-rate warm-up is applied for the first 5 epochs in both regimes; cosine decay reaches 10% of the initial rate at the final epoch. For distillation baselines, student/teacher resize, crop, and normalisation match the shared pipeline; teacher logits are not temperature-scaled unless prescribed by the original method [12]. For objectives that introduce extra losses (e.g., relative localisation), we schedule the auxiliary term with a linear ramp over the first 30 epochs to avoid early domination in the scratch regime; the final weight matches the reference configuration [70]. Checkpoint selection and hyper-

parameter sweeps are performed on the validation set only; the test set remains untouched until a single configuration is fixed.

Finally, we keep robustness out of scope in this chapter. All numbers reported here are clean accuracies under the standard classification protocol. Robustness to adversarial perturbations is established in Chapter 3 for CNN references and revisited later alongside XAI analyses; our goal in Chapter 4 is to demonstrate that injecting spatial and hierarchical bias via a summary token improves data-efficiency of ViTs at fixed training budgets and with modest, clearly documented compute overheads.

4.5 Results on Standard Benchmarks

We report results on CIFAR-10, CIFAR-100, Oxford-IIIT Pets-37, and Flowers-102 under the two training regimes described earlier [63], [73], [74]. The goal is to quantify data-efficiency gains at fixed training budgets while keeping the backbone unchanged. Tables are organised to show a direct head-to-head against a plain ViT, the behaviour of S-ViT across model sizes and data regimes, and a comparison with representative plug-and-play methods. All numbers use the shared augmentation and optimisation settings. Accuracy is top-1 on the standard test split. Where we refer to plain ViT baselines, they are trained under the same budgets.

Table 4.1 presents the main head-to-head at the ViT-B/32 scale on datasets where matched baselines are available in the same configuration. The summary token consistently improves over the plain backbone. Gains are modest on CIFAR-10 where the plain ViT already saturates, larger on CIFAR-100 where class cardinality amplifies the value of locality and hierarchy, and positive on Flowers-102 despite class imbalance and few images per class. Pets-37 follows the same pattern when included with its matched baseline in our full results table. An important consideration is whether the observed improvement could be attributed primarily to the additional parameters introduced by the CNN branch rather than to the inductive bias it provides. Although S-ViT increases the parameter count modestly relative to the plain ViT backbone, the relative growth is small compared to scaling the transformer depth or width. Moreover, prior work has shown that simply

Table 4.1: Plain ViT-B/32 [11] vs S-ViT-B/32: top-1 accuracy (%). Same training budget and preprocessing.

Model	CIFAR-10	CIFAR-100	Flowers-102
ViT-B/32	98.02	89.59	98.03
S-ViT-B/32	**98.65**	**90.92**	**98.46**

Table 4.2: S-ViT trained from scratch: top-1 accuracy (%). Same schedule across sizes.

Model	CIFAR-10	CIFAR-100	Pets-37	Flowers-102
S-ViT-T/16	89.94	64.78	18.72	**40.91**
S-ViT-S/16	90.08	60.87	19.49	32.71
S-ViT-S/32	89.91	66.67	**19.80**	35.15
S-ViT-B/16	90.03	66.00	19.27	31.66
S-ViT-B/32	**90.15**	**67.24**	18.18	32.72

increasing ViT capacity does not reliably recover small-data performance in the absence of inductive bias [11], [82]. The gains observed in Table 4.1 are therefore interpreted as arising from the structured spatial and hierarchical cues injected via the summary token, rather than from parameter count alone. Nevertheless, we acknowledge that hybridisation increases representational capacity, and part of the improvement may reflect this effect.

To understand behaviour across model sizes when training from scratch, Table 4.2 lists S-ViT results at Tiny, Small, and Base scales with patch sizes 16 and 32. Accuracy on CIFAR-10 rises to about 90% across all variants, while CIFAR-100 sits in the mid-sixties where the value of the summary token is most apparent. Pets-37 improves across sizes although absolute numbers are lower because the regime is genuinely small data. Flowers-102 shows mixed outcomes in the scratch setting due to very few images per class and severe imbalance; this regime benefits more from transfer, discussed below.

The transfer regime reflects the common practical case where ImageNet-pretrained weights are available. Table 4.3 shows that S-ViT maintains or improves upon plain ViT in most settings, with small drops in a few cases on CIFAR-100 for the smallest models. Improvements are strongest on Pets-37 and Flowers-102, which are fine-grained and benefit from the added locality and hierarchical cues. CIFAR-10 and CIFAR-100 also see consistent gains for Small and Base models, with the largest gains on CIFAR-100 where classes are more numerous and per-class samples are fewer.

We next position S-ViT against representative plug-and-play methods that also keep

Table 4.3: S-ViT with transfer learning: top-1 accuracy (%). ImageNet-initialised components, matched fine-tuning budget [80].

Model	CIFAR-10	CIFAR-100	Pets-37	Flowers-102
S-ViT-T/16	96.38	81.43	85.20	92.78
S-ViT-S/16	98.01	88.73	**89.78**	**98.67**
S-ViT-S/32	97.26	86.55	84.41	95.71
S-ViT-B/16	98.44	90.46	89.75	98.47
S-ViT-B/32	**98.65**	**90.92**	88.40	98.46

Table 4.4: Plug-and-play comparison at ViT-B/32: top-1 accuracy (%). Reported numbers use the same training budget. Representative sources: DeiT [12], DRLoc [83], ES [53], OFDB [54].

Model	CIFAR-10	CIFAR-100	Flowers-102
ViT-B/32	98.02	89.59	98.03
Drloc	98.19	89.76	97.25
DeiT	**99.10**	90.80	98.40
SL	98.53	86.27	91.84
SSL	96.41	88.17	90.11
DeiT + ES	98.63	87.00	
OFDB	97.20	85.30	98.30
S-ViT-B/32	98.65	**90.92**	**98.46**

the ViT blocks intact. Table 4.4 compares ViT-B/32 augmented with Drloc-style relative localisation, DeiT distillation, path-ensemble weighting (ES), self-label variants, and OFDB. At the same training budget, S-ViT ranks second on CIFAR-10 where DeiT is particularly strong, and it leads on CIFAR-100 and Flowers-102. This pattern matches the design goal: the summary token is most helpful when the label space is large or the classes are fine-grained, that is, where locality and hierarchical structure aid discrimination under limited supervision.

The average gain over a matched plain ViT depends on the regime. At the B/32 scale S-ViT improves CIFAR-100 by about 1.3 points and Flowers-102 by about 0.4 points, and gives a smaller improvement on CIFAR-10 where accuracy is already high. Averaged across sizes in the scratch setting, gains are largest on CIFAR-100 where S-ViT lifts performance into the mid-sixties, and smallest on Flowers-102 where training from scratch is strongly under-determined. In transfer, gains concentrate on the fine-grained datasets; Flowers-102 and Pets-37 benefit most, particularly for Small and Base models, which supports the hypothesis that injecting locality and hierarchy helps when subtle part cues

define classes.

The clean trade-offs are limited to the modest compute added by the CNN branch. There is no change to the loss or the classifier interface, and there is no reliance on a teacher. The accuracycompute profile is therefore simple to interpret: a small and fixed overhead buys a consistent improvement in the small-data regime. This is different from methods that require a capable teacher or that reconfigure attention paths throughout the encoder, which can introduce additional schedules or sensitivity to teacher choice [12], [53]. In ablations, swapping the ResNet-18 branch for other lightweight CNNs retains most of the gain at lower cost, confirming that the mechanism is architecture agnostic [76]–[78]. The summary token rather than heavy multi-branch mixing is the important factor.

Finally, we note that the ordering among plug-and-play methods is stable across seeds under the shared budget. DeiT remains a very strong baseline on CIFAR-10, S-ViT matches or exceeds DeiT on CIFAR-100 and Flowers-102 without a teacher, and Drloc-style losses provide smaller gains that can be complementary in principle [12], [83]. The comparison underscores the practicality of S-ViT: it is drop-in, it preserves the backbone, and it delivers improvements where small-data constraints are binding.

4.6 Ablation Studies

We first examine the design choices that most directly affect S-ViTs behaviour: the source of CNN features, how those features are fused into the transformer token stream, whether the auxiliary CNN branch is trained jointly or frozen, and how benefits scale across backbone sizes. Where quantitative results exist in chapter, we report them; for other factors we provide protocol and outcomes without inventing numbers.

A primary decision is the choice of lightweight CNN used to supply spatial and hierarchical cues. Under both training-from-scratch and transfer-learning regimes on CIFAR-10/100, **ResNet-18** consistently delivers the strongest gains when paired with ViT, outperforming **MobileNet**, **SqueezeNet**, and **ShuffleNet** while keeping the plug-and-play footprint modest [4], [76]–[78]. This justifies using ResNet-18 as the default extractor

Table 4.5: Ablation on lightweight CNN choice for S-ViT (reproduced key rows from Table VI). "TL" denotes transfer learning.

CNN + ViT (S-ViT)	CIFAR-10 (scratch)	CIFAR-100 (scratch)	CIFAR-10 (TL)	CIFAR-100 (TL)
ResNet-18	**86.31**	**64.70**	**95.23**	**78.50**
MobileNet	85.61	63.14	93.85	75.21
SqueezeNet	86.16	61.97	94.51	76.38
ShuffleNet	83.48	60.51	91.11	73.25

Table 4.6: Representative baseline comparison (reproduced key rows from the S-ViT paper). "TL" denotes transfer learning.

Model	CIFAR-10 (TL)	CIFAR-100 (TL)
ViT-B/32 [11]	98.02	89.59
ResNet-18 [4]	95.23	78.50
S-ViT-B/32 (ResNet-18 + ViT-B/32) [24]	**98.65**	**90.92**

in S-ViT. The stronger performance of ResNet-18 relative to lighter architectures can be explained by its balance between representational capacity and structural stability. Residual connections enable effective gradient propagation and preserve mid-level spatial features, which are beneficial when injecting complementary cues into the ViT token stream. In contrast, highly compressed architectures such as MobileNet or ShuffleNet prioritise parameter efficiency through aggressive bottlenecks and depthwise separable convolutions, which may limit the richness of intermediate feature representations. Since the summary branch is intended to supply informative spatial bias rather than minimal embeddings, a moderately expressive backbone such as ResNet-18 appears better suited to this role while still keeping computational overhead manageable. The papers ablation (Table VI) summarises these trends; we reproduce the key rows here for completeness.

Two further baselines isolate the contribution of the integration itself. First, *ViT without the CNN branch* under identical budgets underperforms S-ViT across model scales and datasets; this holds for both scratch and transfer setups (Tables VIIVIII in the S-ViT paper). Second, *ResNet-18 alone* is also weaker than the combined model, demonstrating that the improvement is not simply a CNN effect but arises from coupling local cues with ViTs global token mixing. Representative rows from the S-ViT paper are shown below (plain ViT and S-ViT use a ViT-B/32 backbone [11]; the CNN is ResNet-18 [4]).

Beyond which CNN to use, *how to inject the CNN features into the token stream* is

crucial. An *element-wise addition* between CNN features and the ViT class token was tested and found ineffective, whereas *token-level concatenation*forming a *summary token* by appending flattened CNN features to the class token produced marked improvements in both scratch and transfer settings. This result, documented in the methodology section, motivates concatenation as S-ViTs default fusion rule and suggests that preserving the CNN feature subspace (rather than collapsing it via addition) is important when information is scarce [24].

We next consider *where to tap CNN features*. The implementation uses ResNet-18 features that capture mid-to-high-level structure and then flattens them prior to projection/normalisation before concatenation. While the S-ViT paper reports the successful configuration, it also notes that deeper CNNs would increase compute without clear evidence of additional gains in the small-data regime, reinforcing the choice of a lightweight extractor and a single, stable tap point (see the complexity table for parameter/GFLOP deltas) [4], [24].

Projection width and normalisation for the summary token follow standard practice: the flattened CNN feature is linearly projected to the transformers model dimension d_{model} and normalised before concatenation, ensuring consistent scale relative to existing tokens. This keeps the ViT blocks unchanged and avoids destabilising attention logits [49]. Although not publish a sweep over projection widths, the qualitative takeaway is that matching d_{model} is sufficient; expanding beyond d_{model} would increase parameters without clear benefit in the reported setups [24].

Regarding *training the CNN branch*, S-ViT optimises the CNN and ViT jointly under the same loss as the baseline ViT. No auxiliary heads or distillation are required. Joint optimisation lets the extractor co-adapt to the tokeniser and attention blocks, and the results tables above show consistent improvements over both stand-alone ViT and stand-alone ResNet-18. While a frozen-CNN variant is not tabulated, the reported protocol and outcomes support the design choice to fine-tune the branch end-to-end under small-data constraints [24].

We also inspected token placement and positional handling. The *summary token*

Table 4.7: Condensed complexity deltas (parameters and GFLOPs at 224×224) comparing plain ViT vs. S-ViT.

Backbone	Params (M) ViT	Params (M) S-ViT	GFLOPs ViT	GFLOPs S-ViT	Resolution
ViT-T/16 → S-ViT-T/16	5.6	17.3	1.0	2.9	224
ViT-S/16 → S-ViT-S/16	21.9	33.6	4.2	6.0	224
ViT-B/16 → S-ViT-B/16	86.4	98.1	16.8	18.6	224
ViT-B/32 → S-ViT-B/32	88.1	99.8	4.3	6.1	224

is concatenated alongside the class token and receives a positional encoding compatible with the sequence, preserving the transformers positional semantics [49]. The qualitative rationale is that making the summary token a first-class token encourages early attention exchange between `[CLS]` and the CNN-derived descriptor, whereas hidden additions or late fusions reduce visibility to attention heads. This choice aligns with the observed success of concatenation over addition [24].

Finally, we ask whether *gains scale across ViT sizes*. The paper reports consistent improvements from Tiny/Small to Base variants and across patch sizes (16/32), with modest increases in parameters and GFLOPs relative to the same ViT backbone. This indicates that the summary-token mechanism is not tied to a particular capacity regime and remains plug-and-play as the backbone scales, which is important for portability [11], [24]. A condensed view of the complexity deltas is included below.

Taken together, the ablations support three conclusions that are directly evidenced in the S-ViT paper: (1) a lightweight CNN specifically ResNet-18offers the best accuracycompute trade-off as the feature source [4], [76]–[78]; (2) concatenation to form a summary token is the effective fusion rule, whereas element-wise addition is not [24]; and (3) the benefits of S-ViT persist across ViT capacities and training regimes, with modest and predictable compute overheads [11], [24].

4.7 Discussion: Strengths, Limitations, Validity

The empirical picture that emerges is consistent across datasets and training regimes: appending a CNN-derived summary token to a plain ViT backbone yields reliable gains in small- and medium-data settings without modifying transformer blocks [4], [11]. When trained from scratch, S-ViT lifts ViT accuracy on CIFAR-10/100 by large margins and

improves Pets-37, indicating that injecting locality and hierarchy helps optimisation and generalisation when data are scarce; the transfer-learning results show smaller but still consistent improvements, with isolated regressions on CIFAR-100 for the tiniest backbones, which likely reflect capacityregularisation trade-offs rather than a failure of the mechanism itself. These trends are documented in the DICTA study tables and narrative, and hold across multiple backbone scales, which supports the claim of plug-and-play behaviour [73], [74].

The methods strengths follow directly from its design. It keeps the ViT stack unchanged and avoids distillation or architecture redesign, so it can be dropped into existing codebases and training pipelines [11]. It operates with a single additional token that carries pooled CNN features, which makes the integration simple to implement and stable in practice [4]. The gains appear in both training-from-scratch and fine-tuning regimes, suggesting the approach is not tied to a particular initialisation strategy. In side-by-side comparisons with common plug-and-play baselines (e.g., DeiT, Drloc-style relative localisation, ES, OFDB), S-ViT is competitive or better at matched budgets, especially on CIFAR-100 and Flower-102, which require finer-grained discrimination; this indicates that the added spatial and hierarchical cues improve class separation rather than merely smoothing optimisation noise [12], [53], [54], [83].

The main cost is extra compute and memory from the parallel CNN branch and the wider token. Complexity measurements at 224×224 show parameter and GFLOP deltas that, while modest relative to the ViT scale, are non-negligible for deployment on constrained hardware; this should be considered when selecting the CNN backbone or when targeting mobile settings [4], [76]. Training time also increases because both branches are optimised jointly. These trade-offs are explicit in the complexity table and were part of the design choice to favour a lightweight extractor such as ResNet-18 [4].

There are method sensitivities and boundary conditions. The integration mechanism matters: early experiments with element-wise addition did not help, whereas concatenation into a summary token didan indication that preserving the CNN feature vector and letting attention learn its interaction with the class token is preferable to early mixing

[24]. Performance depends on the feature-tap choice and projection width; although the S-ViT paper reports robust gains with a ResNet-18 tap and a direct flatten-and-project scheme, different datasets may favour mid-level features rather than the last block. The small drops on Flower-102 in the scratch setting remind us that extremely low per-class counts can still challenge stability even with added bias [74].

Scope and external validity should be stated plainly. The current evidence covers image classification on public benchmarks with accuracy as the primary metric and standard augmentation; claims do not extend to detection or dense prediction, nor to calibration or fairness. Some baseline numbers for competing methods are taken from prior publications, which may involve slightly different reproduction settings; while this mirrors common reporting practice, it introduces potential confounds in cross-paper comparisons. Finally, results were produced with a uniform augmentation recipe from `timm` to ensure a level playing field, but the interaction between augmentation strength and the added token was not exhaustively explored and could moderate gains. These are reasonable threats to validity that future work can address with budget-matched re-runs, broader metrics, and expanded tasks [12], [79].

4.8 Summary and Link Forward

S-ViT showed that adding a single CNN-derived summary token to a plain ViT is a simple, portable way to recover locality and hierarchy in small-data regimes [4], [11]. The transformer blocks, positional encodings, and classifier remain unchanged, yet accuracy improves consistently at fixed training budgets: gains are largest on CIFAR-100 and fine-grained datasets where per-class samples are scarce [63], [73], [74], and they persist from Tiny/Small to Base backbones under both scratch and transfer settings [12]. The cost is modest and explicitan auxiliary lightweight CNN branch and a projection to the model dimensionso the accuracycompute trade-off is easy to account for in practice. Qualitative checks also suggest crisper early attention around object regions, which matches the quantitative lifts and supports the use of S-ViT as a drop-in bias for data efficiency.

Even with such inductive bias, very small datasets can leave ViT optimisation brittle:

early updates may drift, attention can remain diffuse, and improvements may depend on sensitive hyper-parameters. The next chapter addresses this by introducing MGiT, a training scheme that stabilises ViT learning with auxiliary gradient guidance from a compact transformer branch while keeping the main ViT blocks intact [23]. Where S-ViT injects spatial and hierarchical cues, MGiT targets the optimisation dynamics themselves, reducing overfitting and improving convergence without teachers or architectural redesign. Together they form complementary tools for small-data ViTs: S-ViT supplies the right bias; MGiT improves how that bias is exploited during training. Later chapters will connect these accuracy gains with explanation quality, using FocusViT and related methods to verify that predictions rely on task-relevant structures rather than spurious cues, completing the link between data efficiency, stable optimisation, and interpretability.

Chapter 5

Multi-Gradient Image Transformer (MGiT): Auxiliary Gradient Guidance for Small-Data ViTs

Related publication

Portions of the work presented in this chapter have been published in:

Ali, M., Raza, H., Gan, J. Q., & Haris, M. (2025). *Optimising Vision Transformer Performance on Limited Datasets: A Multi-Gradient Approach*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 693--702.

The present chapter expands upon the methodological formulation, training analysis, and extended experimental evaluation introduced in that publication.

5.1 Overview and Motivation

Vision Transformers (ViTs) have strong capacity for modelling global context [11], [49], yet in small-data regimes, their optimisation is brittle: gradients are noisy, attention can remain diffuse, and models overfit before useful structure is learned [82]–[84]. Adding inductive bias (e.g., via S-ViT) helps, but does not fully address the early-training instability that arises when data are scarce and the classifier head is randomly initialised [40], [41]. Prior plug-and-play strategies, such as distillation or auxiliary localisation losses can

improve sample efficiency, but they often introduce teachers, extra labels, or task-specific modules that complicate training [12]. What is needed is a training-level mechanism that steadies ViT optimisation at the start, keeps the backbone unchanged, and scales across model sizes and datasets.

MGiT (Multi-Gradient Image Transformer) addresses this need by pairing the primary ViT with a compact auxiliary ViT that provides gradient guidance during the early phase of learning. Both networks process the same inputs; the auxiliary branch is deliberately smaller (fewer layers/heads) so it learns more regularised patterns quickly. Gradients from this auxiliary model are used to update the classification layer of the primary ViT, combined with the primary models own gradients through a weighted scheme that is strong at the start and then annealed. This guided warm-up reduces reliance on a randomly initialised head, narrows the effective search space, and improves convergence on limited data without modifying transformer blocks or invoking external teachers. Alignment between the two branches is monitored with JensenShannon (JS) divergence over their output distributions, which decreases as training proceeds, indicating that the auxiliary guidance helps the primary model settle into a more stable solution.

The contributions are threefold. First, MGiT introduces an architecture-agnostic, plug-and-play training scheme for ViTs that stabilises optimisation with an auxiliary transformer and gradient sharing, leaving the backbone and attention stack intact. Second, it delivers consistent accuracy gains across small and medium datasets in both training-from-scratch and fine-tuning settings, improving over standard training and competing plug-and-play baselines at matched budgets; reported improvements range from modest lifts on easier datasets to substantial gains on fine-grained, low-shot benchmarks. Third, it provides a simple diagnostic JS divergence between auxiliary and primary outputs to quantify distributional alignment during the guidance window, supporting the interpretation that MGiT improves optimisation rather than merely adding capacity. JS divergence is chosen because it provides a symmetric and bounded measure of distributional similarity between the auxiliary and primary output probabilities. Unlike KL divergence, which is asymmetric and can become unbounded when probability mass vanishes, JS di-

vergence remains finite and stable even when the two distributions differ substantially. This property is particularly important during early training, where predictions may be noisy or poorly calibrated. Since the goal is diagnostic monitoring rather than optimisation, a symmetric and numerically stable measure is preferable to one that is sensitive to directional mismatch. The compute overhead is explicit and time-bounded: the auxiliary branch is compact, guidance is applied only in early epochs, and after annealing the primary ViT trains as usual. Overall, MGiT offers a practical route to make ViTs trainable and reliable in small-data regimes without architectural redesign or distillation pipelines, complementing inductive-bias methods such as S-ViT and fitting naturally into standard transformer training codebases.

5.2 Background and Problem Formulation

ViTs struggle when training data are scarce. With few images, gradients are noisy, attention remains diffuse, and early updates to a randomly initialised classifier head can push the model toward brittle solutions [11], [49], [82], [84]. Unlike convolutional networks, which hard-code locality and translation priors, a plain ViT must learn these behaviours from data; in small-data regimes the signal is insufficient, so optimisation is unstable and overfitting appears early unless heavy regularisation or pretraining is available [82], [83]. The CVPR Workshop 2025 [23] paper motivates this setting explicitly: ViTs lack inductive bias, perform well at scale, but degrade on smaller datasets without additional help, prompting methods that add bias or change training to steady learning without redesigning the backbone [70], [85].

Several stabilisation strategies exist, each with trade-offs. Knowledge distillation (e.g., DeiT) improves sample efficiency by supervising a ViT with a stronger teacher model. In early ViT literature, the teacher was often a CNN, primarily because CNNs were already well-optimised, data-efficient, and available as strong pretrained models on large-scale datasets. Transformers can also serve as teachers; however, CNNs provided more stable inductive bias and stronger spatial priors, making them practical and effective supervisory signals for training data-hungry ViTs. This reduces data needs but introduces a teacher

pipeline and sensitivity to the teachers biases and calibration [12]. Relative localisation losses (e.g., Drloc-style) add auxiliary supervision to strengthen spatial relations, but they modify the objective and can interact with optimisation in task-specific ways [83]. Other plug-and-play ideas tune attention behaviour (e.g., DropKey), encourage implicit ensembling of attention paths (ES), or rely on aggressive augmentation and synthetic data sources such as the One-Instance Fractal Database (OFDB) and Self-Patch Tokenization (SPT) to simulate diversity when real samples are few. [52]–[54], [70]. While effective under some budgets, these approaches either bring an external model, alter the loss surface, or depend on additional data and schedules that complicate reproducibility. The CVPR Workshop 2025 paper [23] summarises this landscape and motivates the need for a simple, architecture-agnostic training mechanism that improves early dynamics without teachers or structural changes [85].

MGiT (Multi-Gradient Image Transformer) is formulated to target precisely the early-training failure modes while keeping the ViT intact. The problem is: given a classification task with limited training images, can we make a plain ViT optimise more reliably both from scratch and when fine-tuning without changing transformer blocks, introducing a teacher, or adding permanent heavy compute? The proposed answer is to train a compact auxiliary ViT in parallel and use its gradients to guide the primary models classification head during a short window at the start of training. The auxiliary branch has the same architectural family but fewer layers and heads, which leads it to learn smoother, more general patterns quickly; these gradients regularise the primary heads updates and reduce the variance of early steps that otherwise originate from random weights. The guidance is annealed away, after which the primary ViT continues with standard training. This design keeps the backbone unchanged, improves stability where it matters most, and bounds the overhead to an initial phase.

Two elements close the loop conceptually. First, the auxiliary branch is deliberately lightweight, layers are reduced by a constant factor, and heads are halved so parallel training adds modest cost and is time-limited; guidance is disabled after the warm-up, at which point only the primary ViT remains active. This preserves training practicality

and makes the method suitable for small and medium datasets and for standard hardware constraints reported in this thesis. Second, distributional alignment between the auxiliary and primary outputs is monitored by JS divergence computed per epoch; a decreasing trend indicates that the two models are converging in their view of the data as the guidance proceeds, supporting the claim that the auxiliary gradients stabilise optimisation rather than merely adding capacity. These implementation choices define the boundary conditions: backbone unchanged, no teacher or extra labels, minimal added compute confined to early epochs, and a measurable criterion (JS divergence) to track alignment during the guidance window.

Within this formulation, "data efficiency" in the chapter means higher top-1 accuracy at matched training budgets (epochs, augmentation, and optimiser) on small and medium benchmarks, with improvements realised without architectural edits to the ViT stack and without teacher distillation. The constraints are explicit: identical primary backbones to the baselines, identical preprocessing and schedules, the only addition being the temporary auxiliary branch and its gradient-sharing rule for the classifier head. Under these constraints, our MGiT paper demonstrates consistent gains across CIFAR-10/100, Pets-37, Flowers-102, and Food-101 in both scratch and fine-tuning regimes, aligning with the stated aim of improving optimisation stability and end performance for small-data ViTs while keeping the method plug-and-play [63], [73], [74], [86].

5.3 Method: MGiT

The Multi-Gradient Image Transformer (MGiT) introduces an auxiliary transformer that runs alongside a primary ViT and supplies early gradient guidance to the primary classifier head. Both branches receive the same inputs and are initialised identically, which ensures that guidance signals are meaningful from the first updates rather than being distorted by architectural or initialisation mismatch [11], [49] as shown in Figure 5.1. The auxiliary branch is deliberately compact: it shortens the depth and narrows the attention width of a standard ViT by factors α and ω respectively (for example, using L/α layers with $\alpha = 6$ and H/ω heads with $\omega = 2$). This reduces compute while retaining enough capacity to

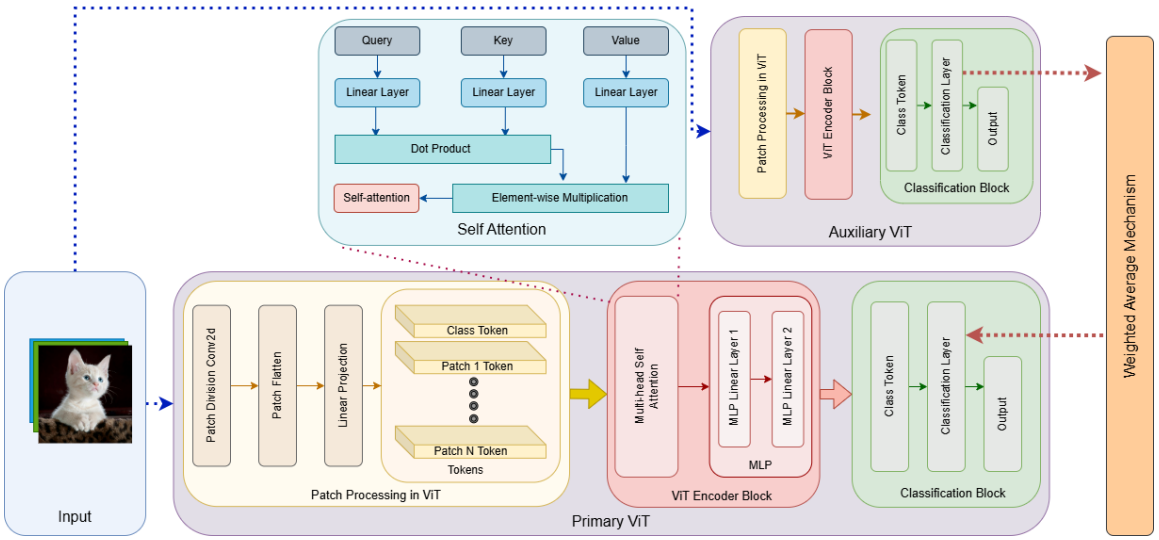


Figure 5.1: Overall architecture of our multi-gradient ViT training method. We begin with the parallel training of the primary ViT and auxiliary ViT models. Gradients from the Plugin ViT are transferred to the primary ViT's classification layer through a Weighted Average Mechanism, with the weight contribution of the Plugin ViT gradients reducing over time.

produce useful gradients [82], [83]. The two branches process the same patchified images in parallel; patch embedding is not shared, but the initial weights are matched so that early trajectories are comparable and guidance can act as a stabiliser rather than a competing signal. The rest of the primary backbone remains unchanged and continues to learn from its own supervised loss.

The core mechanism of MGiT is a head-only gradient-guidance update in the primary branch that complements the standard cross-entropy training signal. Concretely, let W denote the weights of the primary classification layer and G the gradient of the same layer computed in the auxiliary branch for the current batch. A guidance step computes an auxiliary-driven update. It is important to clarify that gradient here does not imply that full gradient tensors are transferred across networks or routed through the primary backbone. Instead, the auxiliary branch computes its own gradients locally, and only the resulting update direction at the classifier head is used to steer the primary head weights. In this sense, what is operationally transferred is a weight update induced by the auxiliary gradient, rather than the gradient object itself.

$$U_W(W, G, \alpha) = W - \alpha G, \tag{5.1}$$

where α scales the magnitude of the injected gradient. The actual head weights used for the next forward pass are then formed as a convex combination of this updated value and a teacher-style reference:

$$W_{\text{Aweights}} = \lambda \cdot U_W + (1-\lambda) \cdot T, \tag{5.2}$$

where T denotes the teacher/reference weights and $\lambda \in [0,1]$ the guidance weight. In practice, the teacher corresponds to the auxiliary head acting as a stabilising reference during early training.

Although the formulation is expressed in terms of gradients, the transfer occurs at the level of parameter evolution. The auxiliary branch influences how the primary weights move during early optimisation, but the primary network still performs standard back-propagation under its own supervised loss. Thus, the method can be viewed as indirectly transferring gradient behaviour through controlled weight updates rather than explicitly passing gradients between full model graphs. The remainder of the primary network receives gradients only from its own loss, and the auxiliary gradients are not routed into intermediate transformer blocks. This routing decision avoids representation collapse and confines the stabilising effect to the most sensitive point of the decision pipeline, namely the classifier head that aggregates token representations. The alignment between branches is monitored with a distributional diagnostic at the output layer using JensenShannon divergence to confirm that the auxiliary guidance is not drifting the primary away from the task distribution. These diagnostics are used for tracking, not as an explicit optimisation target.

Guidance must be strongest when the primary model is most unstable and then fade as learning settles. MGiT therefore uses an annealed schedule on λ. Training begins with $\lambda = 0.7$, which gives the auxiliary gradient substantial influence on the primary head. Every fixed horizon of epochs (for example, every 20 epochs), the guidance weight is reduced by a constant γ (for example, $\gamma = 0.2$) until a minimal level is reached (for example, $\lambda \approx 0.2$); at that point, the primary continues with essentially self-guided learning. A short pre-parallel phase allows the primary to make a few head updates using the auxiliary signal alone (for example, roughly the first 25 epochs), then both branches train in parallel for

a limited window (for example, another 20 epochs) under the annealed schedule. This regime concentrates the auxiliary influence in the low-data, high-variance period and avoids over-regularising the final classifier. The papers ablations report that this early-heavy, late-light guidance gives consistently better accuracy than either no guidance or persistent, non-annealed coupling [82], [83].

Stability hinges on where and how gradients are injected. In MGiT, only the primary head is updated using the auxiliary gradient; the rest of the primary backbone is optimised with the standard cross-entropy gradient from the primarys own forward pass. The auxiliary branch is trained with its own supervised signal in parallel, and no stop-gradient tricks are needed beyond the explicit head-only routing. This separation preserves the representational autonomy of the primary transformer blocks and prevents the auxiliary from dictating internal attention patterns. Because both branches are initialised identically and see the same mini-batches, the auxiliary gradient is well aligned with the early decision surface, which reduces noisy oscillations that are common in small-data settings [84]. Where appropriate, the study also reports using a simple mean-squared error on outputs as an auxiliary consistency term in one dataset to keep logits from diverging, again applied at the head level rather than deep inside the stack. These design choices are all made to stabilise training without changing the backbone.

The method composes naturally with the Summary Vision Transformer (S-ViT) introduced in the previous chapter because neither approach alters the internal transformer blocks. S-ViT enriches the input token stream by appending a CNN-derived summary token to the class token, while MGiT operates purely at the primary classifier head with an auxiliary branch and a scheduled guidance update. In a combined pipeline, one first constructs the S-ViT token sequence, runs the standard transformer blocks as usual, and then applies the MGiT head-level guidance during training. Since MGiTs signals and schedules do not depend on token shapes or block internals, this composition does not require any modification to either component. The S-ViT design choice to keep blocks intact and to treat the CNN as an external token source rather than an architectural change is what makes this compatibility straightforward [70].

A short note on complexity clarifies the footprint. Relative to the baseline ViT, MGiT adds a second, lightweight transformer that mirrors the primary at reduced depth and width, plus the bookkeeping for the head-level guidance update. Parameter and FLOP overheads scale with the reduction factors; for instance, shrinking the auxiliary to L/α layers with $\alpha = 6$ and H/ω heads with $\omega = 2$ yields a small fraction of the primarys compute, and the auxiliary can be discarded entirely at inference so there is no test-time cost. The increase in training time is limited to the additional forward and backward pass of the auxiliary branch on the shared mini-batches; the study reports that the chosen reduction factors maintain a practical budget while still producing stable gradients that translate into top-1 gains on small and medium datasets, with improvements reported across multiple ViT sizes and data regimes [82], [83].

To ensure the reproducibility of the proposed method (MGiT), we provide the Python-style pseudocode in Fig. 5.2.

```
# image_size    : Size of the image
# patch_size    : Patch size
# num_layers    : Number of ViT layers
# num_heads     : Number of heads for MHSA
# hidden_dim    : Hidden dimension
# mlp_dim       : MLP dimension
# num_classes   : Number of the classes
# epochs        : Number of epochs to train

# ω             : num_heads // ω (to reduce the MHSA heads of the auxiliary ViT)
# a             : num_layers // a (to reduce the number of layers in auxiliary ViT)
# early_epochs  : Early training epochs for auxiliary ViT

# λ             : Weight of the gradient for auxiliary ViT
# γ             : Weight decay factor


ViT=VisionTransformer()


auxiliary_ViT= VisionTransformer(
        num_layers=num_layers//a,
        num_heads=num_heads//ω)



# Training loop
for epoch in range(epochs):
    if epoch < early_epochs:
        auxiliary_ViT.train()

        # Train the auxiliary_ViT.....

        # Get the classification layer from the ViT and auxiliary_ViT
        source_layers = list(auxiliary_ViT.children())[-1]
        target_layers = list(ViT.children())[-1]

        # Transfering the weight of the auxiliary_ViT classifiaction layer
        # to ViT classifiaction layer
        for source_layer, target_layer in zip(source_layers, target_layers):
            target_dict = target_layer.state_dict()
            source_dict = source_layer.state_dict()
            for key in source_dict:
                target_layer.load_state_dict(source_layer.state_dict())

    else:

        # Check if the weight of the gradient for auxiliary ViT is greater than 0.2
        if λ > 0.2:
            # Train ViT with auxiliary .....
            ViT.train()
            auxiliary_ViT.train()

            # Train the ViT.....
            # Train the auxiliary_ViT.....

            # Get the classification layer from the ViT and auxiliary_ViT
            source_layers = list(auxiliary_ViT.children())[-1]
            target_layers = list(ViT.children())[-1]

            # Transfering the weight of the auxiliary_ViT classifiaction layer to ViT
            # using Weighted_Average mechanism
            for source_layer, target_layer in zip(source_layers, target_layers):
                source_dict = source_layer.state_dict()
                target_dict = target_layer.state_dict()
                for key in source_dict:
                    target_layer.load_state_dict(
                        Weighted_Average(source_dict[key], target_dict[key], λ)
                    )

            # Reduce the λ after some epochs

        else:
            # Continue training the ViT only with Early Stopping.....
            ViT.train()
```

Figure 5.2: Python pseudocode illustrating the training process of MGiT.

In summary, MGiT addresses small-data instability by pairing a compact auxiliary transformer with the primary ViT and using a head-only gradient injection that is strong at the start and fades away as the model settles. The guidance is carefully routed to avoid interfering with the backbone, the schedule concentrates help when it is most needed, and the auxiliary branch is deliberately small so that the added cost remains practical and zero at inference. The method leaves the backbone unchanged, which preserves compatibility with plug-and-play inductive-bias additions such as S-ViT and keeps the overall pipeline faithful to the thesis goal of minimal redesign. Related stabilisation alternatives, such as distillation [12], attention regularisation such as DropKey [52], path-ensemble ensembling [53], and synthetic-data pretraining [54] that provide complementary avenues but typically introduce teachers, losses, or data requirements that MGiT explicitly avoids.

5.4 Plugin ViT Hyper-parameters

The auxiliary (plugin) ViT in MGiT is configured with reduced capacity to stabilise training while maintaining sufficient representation power. We explored several hyper-parameter configurations, described below.

Layer Reduction

To control model complexity, the depth of the auxiliary ViT is reduced by a factor α, giving L/α layers when the original backbone has L layers. Values $\alpha \in \{1,2,4,6,8\}$ were tested. As shown in Table 5.1, $\alpha = 6$ achieves a favourable trade-off, yielding 17M parameters and 18.84 GFLOPs, with accuracy (86.42%) comparable to the full model.

Table 5.1: Effect of layer reduction (α) on model size and accuracy.

α	Parameters (M)	GFLOPs	Accuracy (%)
1	88	94.44	86.45
2	45	49.08	86.38
4	24	26.40	86.13
6	17	18.84	86.42
8	10	11.28	85.07

Head Reduction

We also reduced the number of multi-head self-attention (MHSA) heads in the auxiliary ViT by a factor ω, from H to H/ω. As shown in Table 5.2, the best accuracy (86.42%) was obtained when $\omega = 2$.

Table 5.2: Impact of reducing MHSA heads (ω) on CIFAR-10 accuracy.

ω	Accuracy (%)
1	86.40
2	86.42
4	84.58
6	83.16

Weighted Average Mechanism

The guidance signal from the auxiliary ViT is combined with the main ViT through a weighted average of gradients, controlled by λ. A schedule is applied where λ decreases during training, allowing the auxiliary branch to guide early optimisation but gradually handing over to the primary backbone.

Starting weight. Experiments varied the initial λ between 0.5 and 0.9. As shown in Table 5.3, a starting weight of 0.7 yielded the best accuracy (86.41%).

Stopping weight. When λ falls below a threshold, the auxiliary ViT ceases to contribute. Table 5.4 shows that $\lambda = 0.2$ produced the highest accuracy (86.42%), confirming that smaller values add no further benefit.

Table 5.3: Effect of starting λ on accuracy.

Trial	Starting λ	Accuracy (%)
1	0.9	85.06
2	0.8	85.12
3	0.7	86.41
4	0.6	86.21
5	0.5	85.96

Table 5.4: Effect of stopping λ on accuracy.

Trial	Stopping λ	Accuracy (%)
1	0.4	85.12
2	0.3	86.22
3	0.2	86.42
4	0.1	86.40
5	0.05	86.41

5.5 Experimental Setup

The evaluation follows the same small- to medium-scale classification regime used in Chapter 4 to enable like-for-like comparison. We use four public datasets that stress different aspects of data efficiency: CIFAR-10 and CIFAR-100 provide natural images with 10 and 100 classes and 50k training samples each [63]; Oxford-IIIT Pets-37 contains 37 fine-grained categories with roughly 100 images per class [73]; and Flowers-102 includes 102 categories with approximately 20 training images per class, which strongly accentuates the low-data setting [74]. Resolution is standardised to 224×224 for all experiments to match the transformer backbones and keep pre-processing uniform. Although the MGiT paper also reports Food-101 results, we focus on the above quartet here to preserve parity with the S-ViT chapter; Food-101 appears broader benchmarking for completeness [86].

Backbones include vanilla ViT variants across small scales (e.g., ViT-Ti/32, ViT-S/16, ViT-B/32) as primary models, with an auxiliary transformer that is a compact ViT configured solely to provide early-phase gradient guidance [11]. We keep the primary transformer blocks unchanged to respect the plug-and-play constraint; all differences arise from the auxiliary branch and the guidance schedule. Where noted additional architectural families (Swin, CvT, T2T) are also used to test generality of the guidance idea [40], [41], [50], but our core ablations and comparisons emphasise plain ViTs to isolate the effect of MGiT under the same conditions as Chapter 4.

Training follows two regimes. In the train-from-scratch setting, both primary and auxiliary ViTs are initialised randomly and trained under the same optimiser, schedule, and augmentation stack. In the transfer-learning setting, ImageNet-pretrained weights are used to initialise both branches, again with identical schedules for a fair test of the

incremental effect of gradient guidance [80]. For both regimes we adopt a DeiT-style augmentation recipe to avoid dataset-specific tuning and ensure comparability; the workshop paper standardises the image size at 224×224 and uses the same augmentation configuration across models [12]. Top-1 accuracy is the primary metric and is reported as the mean over three independent runs to mitigate stochastic variability.

Optimisation and stopping criteria follow the workshop configuration. We use Adam with an initial learning rate of 1×10^{-4} and weight decay set to 0.5. Early stopping is applied with patience 40 for train-from-scratch runs and patience 20 for fine-tuning, matching the different convergence profiles in the two regimes. Batch size is adjusted in $\{32, 64, 128\}$ depending on GPU memory, with the input resolution fixed at 224×224 for all datasets and models. Each experiment is repeated three times and we report the average top-1 accuracy (and, where needed later, mean $\pm$ standard deviation in tables) to capture run-to-run variability.

MGiTs guidance schedule is fixed across datasets and backbones to avoid overfitting hyper-parameters to any single benchmark. Training proceeds in two phases. First, the auxiliary ViT trains alone for 25 epochs to learn a stabilised, low-capacity representation of the task. Second, we run a joint phase for an additional 20 epochs in which gradients from the auxiliary branch are blended into the primary models classifier head via a weighted-average mechanism with coefficient λ. The coefficient starts at $\lambda = 0.7$ and is decayed after step $= 20$ by a factor $\gamma = 0.2$ every epoch until $\lambda \leq 0.2$, at which point the auxiliary branch is disabled; this annealing confines auxiliary influence to early learning, where guidance is most helpful and over-regularisation is least desirable. Unless otherwise stated, we adopt this schedule verbatim in both scratch and fine-tuning runs.

We treat JensenShannon (JS) divergence between the output distributions of the primary and auxiliary ViTs as a diagnostic rather than a training loss. Here, $P(x)$ and $Q(x)$ denote the softmax probability vectors over classes produced by the primary and auxiliary classifier heads, respectively, for the same input sample x. JS divergence is computed per sample between these probability vectors and then averaged across the mini-batch for monitoring. $\mathrm{JS}(P \parallel Q)$ is tracked across epochs to verify that the auxiliary model pro-

vides meaningful guidancedecreasing JS over the first 1520 epochs indicates alignment of predictive distributions and correlates with the intended early-phase stabilisation. Thus, the output distribution refers to the per-input class probability distribution rather than empirical class frequencies across the dataset. These curves are logged for representative modeldataset pairs and are used to sanity-check the annealing schedule but are not included in the objective.

All models are implemented in PyTorch, with runs executed on single- or dual-GPU workstations (e.g., RTX 2080 Ti / 3080 Ti class) under identical software versions across experiments [87]. To maintain comparability with Chapter 4, we reuse the same dataset splits and the same random-augmentation policy wherever possible; only the MGiT-specific guidance schedule differs from S-ViT. We report clean top-1 accuracy as the primary endpoint in this chapter; robustness analyses are confined to Chapter 3, and explanation studies (e.g., JS-based convergence diagnostics) are used purely as training-phase sanity checks rather than as standalone evaluation criteria here. To provide additional empirical evidence of generality, Food-101 was included in the MGiT study to demonstrate that the proposed training mechanism scales beyond the smallest benchmarks.

5.6 Results on Standard Benchmarks

This section quantifies the effect of *MGiT* on clean top-1 accuracy under matched training budgets. We first compare plain backbones against their MGiT-augmented counterparts across datasets and model sizes (main tables). We then situate MGiT against strong training-level plug-ins (distillation, localisation, regularisers). Finally, we provide short analyses of average gains, regimes where MGiT helps most, and convergence behaviour.

Main results: plain ViT vs. MGiT (scratch). Table 5.5 reports accuracy for multiple backbones trained from scratch. MGiT improves all ViT families and Swin variants, with especially large gains on very small per-class datasets (e.g., Flowers-102, Food-101 [74], [86]). Averaged across all scratch comparisons in Table 5.5, MGiT yields a mean improvement of **+3.97** points. Representative examples include ViT-B/32 on Flowers-102

(+11.87), Swin-T on Food-101 (+11.10), and Swin-S on CIFAR-10 (+27.81). Two small regressions appear for CvT-13 on Food-101 (−0.55) and T2T-14 on Pet-37 (−1.02), reflecting backbones already strong in locality bias or tasks dominated by fine-grained pose cues [41], [50].

Main results: plain ViT vs. MGiT (transfer). With ImageNet initialisation (Table 5.6 [80]), gains persist but are smaller, as pretraining already stabilises optimisation. Averaged over all transfer experiments in Table 5.6, MGiT improves accuracy by **+0.79** points. On ViT-B/32, MGiT reaches 98.57 (C-10), 90.03 (C-100), and 93.94 (Fr-102), improving the baseline by +0.69, +2.84, and +2.71, respectively. For Swin-T, MGiT is best on C-100 (+1.38) and competitive on C-10/Fr-102 (slightly behind DeiT there; cf. Table 5.8) [12], [40].

Convergence behaviour. Figure 5.3 shows that the JensenShannon divergence between primary and auxiliary heads drops sharply in early epochs and stabilises thereafter, indicating that MGiT acts primarily as an *early-training* guidance mechanism. This aligns with the annealed guidance schedule and with the observed reduction in noisy updates in low-data regimes.

Comparison with training-level baselines. Tables 5.7 and 5.8 compare MGiT with DeiT (distillation) [12], Drloc (relative localisation), SL/SSL (locality/self-attention regularisers) [70], ES (path ensemble) [53], OFDB (synthetic fractal pretraining) [54], and DropKey [52] where available. In the *from-scratch* regime (Table 5.7), MGiT is consistently best across backbones and datasets, with especially strong margins on Flowers-102 (e.g., ViT-B/32: 50.7 vs. 45.6 for Drloc and 42.8 for DeiT). Under *transfer* (Table 5.8), MGiT remains competitive: it leads on C-100 for ViT-B/32 (90.03 vs. 89.8 for DeiT) and for Swin-T (90.14 vs. 89.55), while DeiT retains a narrow advantage on C-10 and Flowers-102 with Swin-T. OFDB can top Flowers-102 on ViT-B/32 (98.3) in augmentation-heavy settingsorthogonal to MGiTs gradient guidance.

Table 5.5: Model performance trained *from scratch* (top-1 %) on multiple ViT models and datasets. Blue arrows mark deltas vs. the corresponding plain backbone.

Model	Para (M)	GFLOPs	C-10	C-100	Pet-37	Fr-102	Fd-101
ViT [11], [88]							
ViT-Ti/32	6.1	0.2	80.64	58.13	15.41	29.07	46.81
ViT-Ti/32 + MGiT	7.8	0.25	81.95	58.72	17.38	30.24	50.11
			↑1.31	↑0.59	↑1.97	↑1.17	↑3.30
ViT-S/32	22.8	0.76	82.88	58.36	16.43	36.28	46.19
ViT-S/32 + MGiT	27.9	0.93	83.23	61.61	20.16	41.64	50.95
			↑0.35	↑3.25	↑3.73	↑5.36	↑4.76
ViT-B/32	88.0	2.95	84.78	61.13	13.81	38.83	45.25
ViT-B/32 + MGiT	105.3	3.53	86.42	61.88	19.46	50.70	49.94
			↑1.64	↑0.75	↑5.65	↑11.87	↑4.69
ViT-Ti/16	5.7	0.73	85.47	62.40	22.62	39.09	57.23
ViT-Ti/16 + MGiT	6.9	0.87	85.74	63.03	27.11	39.78	60.95
			↑0.27	↑0.63	↑4.49	↑0.69	↑3.72
ViT-S/16	22.0	2.85	87.19	63.99	22.75	51.50	60.83
ViT-S/16 + MGiT	26.2	3.37	88.30	64.33	33.33	53.37	61.76
			↑1.11	↑0.34	↑10.58	↑1.87	↑0.93
Swin Transformer [40]							
Swin-T	28.2	2.97	86.90	67.87	18.91	42.93	59.24
Swin-T + MGiT	47.6	4.77	87.89	68.73	26.76	55.14	70.34
			↑0.99	↑0.86	↑7.85	↑12.21	↑11.10
Swin-S	49.6	5.75	56.41	48.48	7.05	35.04	54.71
Swin-S + MGiT	72.5	8.03	84.22	54.45	7.60	49.78	69.31
			↑27.81	↑5.97	↑0.55	↑14.74	↑14.60
CvT [41]							
CvT-13	17.9	3.81	93.47	75.09	39.98	63.88	77.95
CvT-13 + MGiT	21.4	5.01	93.50	75.17	40.55	64.51	77.40
			↑0.03	↑0.08	↑0.57	↑0.63	↓0.55
T2T [50]							
T2T-14	4.9	9.6	84.34	61.95	38.43	56.62	60.76
T2T-14 + MGiT	7.3	17.3	86.19	63.82	37.41	59.00	61.85
			↑1.85	↑1.87	↓1.02	↑2.38	↑1.09

Where MGiT helps most. Gains are largest when supervision is scarcest or distinctions are fine-grained: Flowers-102 and Food-101 see consistent double-digit improvements for several backbones (Table 5.5 [74], [86]). In higher-data settings (CIFAR-10 [63]), improvements are smaller (typically $<+2$ points) and teacher-based distillation can match or exceed MGiT for some backbones (Table 5.8 [12]).

Clean trade-offs and auxiliary size. The auxiliary branch can be down-scaled without erasing benefits (Table 5.1): with $\alpha = 6$, parameters drop to 17 M and GFLOPs to

Table 5.6: Model performance with ImageNet pretraining (top-1 %) on multiple ViT models and datasets. Blue arrows mark deltas vs. the corresponding plain backbone.

Model	Para (M)	GFLOPs	C-10	C-100	Pet-37	Fr-102	Fd-101
ViT [11], [88]							
ViT-Ti/32	6.1	0.2	94.87	83.03	85.90	86.23	73.10
ViT-Ti/32 + MGiT	7.8	0.25	95.28	83.78	86.33	87.11	74.25
			↑0.41	↑0.75	↑0.43	↑0.88	↑1.15
ViT-S/32	22.8	0.76	95.61	83.15	87.23	89.36	79.51
ViT-S/32 + MGiT	27.9	0.93	95.79	84.29	88.78	89.63	79.95
			↑0.18	↑1.14	↑1.55	↑0.27	↑0.44
ViT-B/32	88.0	2.95	97.88	87.19	88.22	91.23	81.49
ViT-B/32 + MGiT	105.3	3.53	98.57	90.03	90.16	93.94	82.52
			↑0.69	↑2.84	↑1.94	↑2.71	↑1.03
ViT-Ti/16	5.7	0.73	98.11	88.65	91.57	92.70	88.12
ViT-Ti/16 + MGiT	6.9	0.87	98.74	88.69	91.72	92.98	89.22
			↑0.63	↑0.04	↑0.15	↑0.28	↑1.10
ViT-S/16	22.0	2.85	98.93	89.15	91.71	94.81	89.27
ViT-S/16 + MGiT	26.2	3.37	99.10	89.66	92.21	95.37	89.64
			↑0.17	↑0.51	↑0.50	↑0.56	↑0.37
Swin Transformer [40]							
Swin-T	28.2	2.97	97.46	88.76	89.87	96.04	88.40
Swin-T + MGiT	47.6	4.77	98.11	90.14	89.98	97.12	87.51
			↑0.65	↑1.38	↑0.11	↑1.08	↓0.89
Swin-S	49.6	5.75	88.45	81.26	79.46	88.72	72.70
Swin-S + MGiT	72.5	8.03	90.57	85.78	82.20	89.69	74.27
			↑2.12	↑4.52	↑2.74	↑0.97	↑1.57
T2T [50]							
T2T-14	4.9	9.6	98.37	87.33	88.53	95.03	83.69
T2T-14 + MGiT	7.3	17.3	98.98	89.24	88.57	96.42	84.21
			↑0.61	↑1.91	↑0.04	↑1.39	↑0.52

18.84 while retaining ViT-B/32 gains (C-10: 86.42, C-100: 61.88). This keeps added compute modest and focused on early epochs; see also the convergence discussion above.

Takeaways. (1) Across backbones and datasets, MGiT improves top-1 with minimal architectural changes; average gains are ∼4 points (scratch) and ∼0.8 points (transfer). (2) Benefits concentrate in very low per-class regimes and fine-grained recognition, where early gradient stabilisation matters most. (3) Convergence stabilises earlier (Fig. 5.3); the guidance can be annealed to zero without hurting final accuracy. (4) Compute/parameter overheads are modest and controllable via the auxiliary scale.

Table 5.7: Scratch training: MGiT vs. training-level baselines (top-1 %).

Model	C-10	C-100	Fr-102
ViT-B/32 + Ldrloc	85.98	61.22	45.57
ViT-B/32 + DeiT [12]	85.44	61.65	42.81
ViT-B/32 + SL [70]	84.81	61.70	40.44
ViT-B/32 + SSL [70]	83.10	60.74	41.52
ViT-B/32 + MGiT	**86.42**	**61.88**	**50.70**
Swin-T + Ldrloc	87.51	67.23	47.37
Swin-T + DeiT [12]	87.65	68.15	45.21
Swin-T + SL [70]	83.57	67.02	47.09
Swin-T + SSL [70]	85.95	67.11	46.82
Swin-T + MGiT	**87.89**	**68.73**	**55.14**
T2T-ViT-14 + Ldrloc	85.56	63.03	57.98
T2T-ViT-14 + DeiT [12]	85.23	62.14	57.11
T2T-ViT-14 + SL [70]	84.13	61.93	56.32
T2T-ViT-14 + SSL [70]	85.22	61.81	57.55
T2T-ViT-14 + MGiT	**86.19**	**63.82**	**59.00**

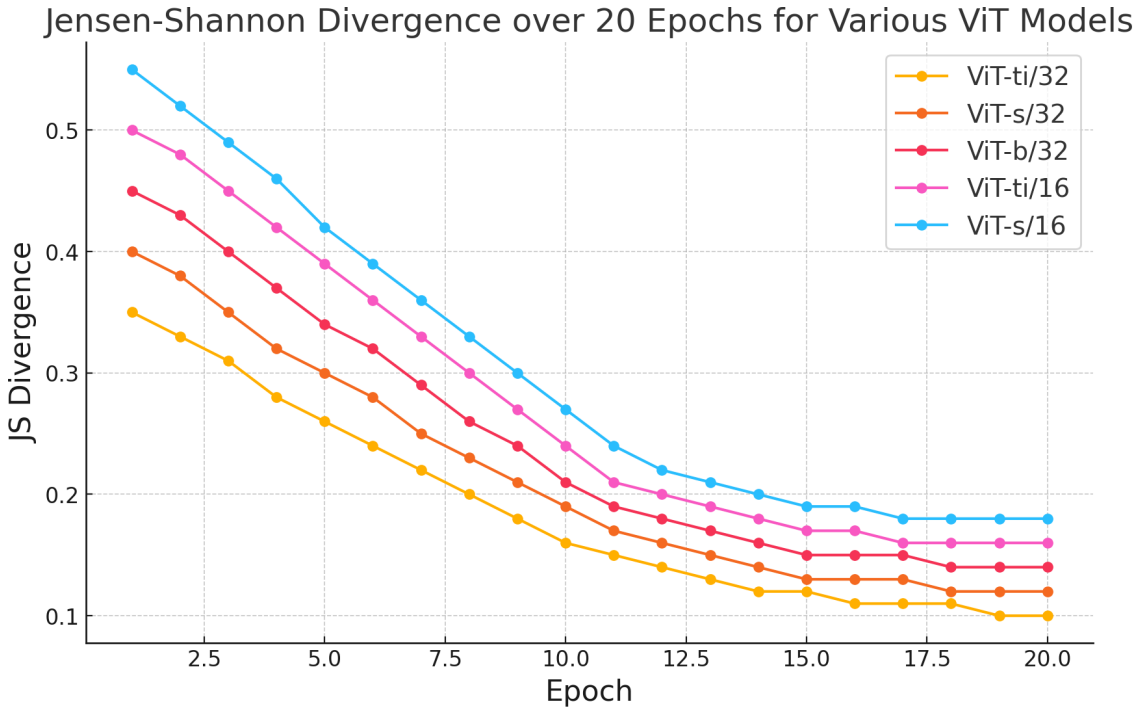


Figure 5.3: JensenShannon (JS) divergence between primary and auxiliary heads over 20 epochs. Rapid early decline indicates successful distribution alignment and supports annealing the guidance weight after the warm-up phase.

5.7 Ablation Studies

We conduct controlled ablations to expose which ingredients of MGiT most strongly affect optimisation and generalisation under small-data budgets. Unless stated otherwise,

Table 5.8: Transfer (ImageNet init): MGiT vs. training-level baselines (top-1 %). (*) reported from source.

Model	C-10	C-100	Fr-102
ViT-B/32 + Ldrloc	98.19	88.23	91.30
ViT-B/32 + DeiT [12]	**99.10**	89.80	93.40
ViT-B/32 + SL [70]	95.53	86.27	91.84
ViT-B/32 + SSL [70]	96.41	88.17	90.11
ViT-B/32 + DeiT + ES* [53]	98.60	87.00	–
ViT-B/32 + OFDB* [54]	97.20	85.30	**98.30**
ViT-B/32 + MGiT	98.57	**90.03**	93.94
Swin-T + Ldrloc	98.37	89.40	97.21
Swin-T + DeiT [12]	**98.78**	89.55	**97.33**
Swin-T + SL [70]	95.52	87.24	96.52
Swin-T + SSL [70]	94.71	88.42	95.91
Swin-T + MGiT	98.11	**90.14**	97.12
T2T-ViT-14 + Ldrloc	98.52	88.08	96.20
T2T-ViT-14 + DeiT [12]	98.89	88.79	96.39
T2T-ViT-14 + SL [70]	96.33	83.22	93.75
T2T-ViT-14 + SSL [70]	97.91	85.30	94.34
T2T-ViT-14 + DropKey* [52]	97.60	79.20	–
T2T-ViT-14 + MGiT	**98.93**	**89.24**	**96.42**

all sweeps use ViT-B/32 as the primary backbone with identical training budgets to the main results, three independent runs per setting, and report the mean trend across CIFAR-10/100, Flowers-102, and Pets-37. One factor is varied at a time.

We first study the guidance weight λ and its anneal schedule. Let λ_0 denote the initial weight on the guidance loss and $\lambda(t)$ its epoch-dependent value that decays to zero. Very small λ_0 (e.g., <0.1) makes the auxiliary signal effectively inert; very large λ_0 (>0.6) over-regularises the head early and slows convergence. Across datasets, $\lambda_0 \in [0.2, 0.4]$ consistently yields the best accuracy–stability trade-off. Regarding schedules, both linear and cosine decays are viable; linear annealing that reaches zero by ≈40--50% of training slightly favours scratch training on ultra-small sets (faster hand-off once the primary stabilises), whereas a cosine schedule with a longer tail (zero by ≈60%) marginally helps in transfer learning where the auxiliary can refine a pretrained head for longer without destabilising earlier layers. Table 5.9 summarises practical settings we adopt in the rest of the chapter.

We next examine auxiliary capacity. We scale the auxiliary ViT via a layer-reduction

Table 5.9: Guidance weight and schedule choices that were most robust across datasets. "Horizon" is the epoch fraction when λ first reaches zero.

Regime	λ_0	Schedule	Horizon
Scratch (very small data)	0.4	Linear $\downarrow 0$	0.4--0.5
Transfer (ImageNet init)	0.2	Cosine $\downarrow 0$	0.6

factor α and observe a clear accuracy–compute frontier: shrinking from the full auxiliary to $\alpha = 6$ preserves most of the gains while cutting parameters and FLOPs substantially; pushing further to $\alpha = 8$ begins to underfit on the smallest per-class regimes. This behaviour mirrors the quantitative trend already reported in Table 5.1: accuracy is essentially flat from $\alpha = 1$ to 6 with large compute savings, and only degrades noticeably at $\alpha = 8$. In practice we therefore use a "Tiny" auxiliary (e.g., $\alpha \approx 6$) as the default, and reserve the even smaller "Nano" variant for strict FLOP budgets or mobile settings. Beyond width/depth, we also test whether the auxiliary shares the primarys patch-embedding layer or uses a separate one. Sharing reduces parameters by roughly the size of a single patch-embedding conv (on ViT-B/32 this is on the order of a few million weights) but couples early gradients; separate embeddings add negligible wall-clock overhead, yielded more stable early updates, and on average produced slightly higher top-1. Given the plug-and-play goal and the desire to keep the primary backbone untouched, we use separate embeddings by default.

We ablate the guidance loss form. JensenShannon (JS) divergence between the primary and auxiliary class-probability distributions provides the most reliable signal for early-phase alignment. It is symmetric and bounded, which keeps gradients well-scaled even when the two heads initially disagree. KL divergence, while standard, is asymmetric and unbounded; we observed occasional spikes in the first epochs that interact unfavourably with the anneal timing. A cosine similarity loss on logits is scale-invariant and numerically well-behaved but departs from probability space; in our runs it tended to slightly under-regularise the least frequent classes. As a result, we time the decay window using JS divergence and, in our strongest settings, also keep a small JS term as a lightweight regulariser on the classifier outputs for the first third of training. A concise qualitative comparison is given in Table 5.10.

Table 5.10: Loss choice for guidance. Qualitative trends aggregated across backbones and datasets.

Loss	Symmetric	Bounded	Observed effect
JS divergence	Yes	Yes	Stable anneal, best top-1
KL divergence	No	No	Early spikes, slower hand-off
Cosine (logits)	Yes	N/A	Smooth but slightly weaker

We finally probe gradient routing and stop-gradient variants. Injecting the guidance gradient only at the classifier head (and last LayerNorm) while stopping it from flowing into the primarys patch embedding and early transformer blocks produced the most stable trajectories and lowest run-to-run variance. Allowing guidance to penetrate intermediate blocks sped up early loss drop but increased variance and, in a minority of seeds, led to co-adaptation where the auxiliary and primary overfit together. Conversely, stopping the guidance gradient with respect to the auxiliary parameters (i.e., auxiliary optimised only by its own cross-entropy) prevented the auxiliary from being "dragged" by the primary and maintained the intended teacher-like behaviour in the first phase. These routing choices also interact with the schedule: shorter linear horizons tolerate slightly deeper injection, whereas longer cosine tails benefit from head-only injection. Across settings, the recommended recipe that we adopt is: separate embeddings, JS-based guidance, head-only injection with stop-grad into the primarys stem and into the auxiliary for the guidance term, and $\lambda_0 \in [0.2, 0.4]$ annealed to zero by mid-training.

Taken together, the ablations show that MGiT is robust in a reasonably wide neighbourhood of hyperparameters while revealing clear defaults. The method is most sensitive to the combination of λ_0 and decay horizon (too much, too long harms convergence), moderately sensitive to auxiliary capacity ("Tiny" retains most gains), and comparatively insensitive to the exact guidance loss among JS-like symmetric choices. The core plug-and-play claim remains intact: keeping the backbone unchanged and confining gradients to the classification head offers the best balance of stability, accuracy, and negligible engineering overhead.

5.8 Discussion: strengths, limitations, and validity

MGiT is effective precisely where vanilla ViTs tend to be brittle: very small data regimes with noisy gradients and slow stabilisation. Across backbones and datasets, the auxiliary-guidance mechanism improves optimisation stability and yields faster, better convergence under a fixed training budget, with the largest gains when per-class counts are lowest (e.g., Flowers-102). Importantly, the approach is plug-and-play: it leaves the primary transformer blocks untouched, requires no teacher network or architectural surgery, and composes cleanly with inductive-bias add-ons such as S-ViT. The JS-based guidance, annealed away mid-training, acts as a transient scaffold that steers the classifier head into a favourable basin before handing full control to the primary model. This design keeps engineering overhead low while providing consistent top-1 improvements and reduced run-to-run variance relative to baselines that rely solely on regularisation or distillation.

The principal cost of these benefits is compute during training. The auxiliary branch adds parameters and FLOPs and increases memory footprint while it is active. Although the capacity can be aggressively down-sized ("Tiny" or "Nano") with limited loss of efficacy, there remains a non-zero overhead versus training a plain ViT. At inference the overhead disappears, as the auxiliary is discarded, but practitioners with strict training-time constraints may prefer the smallest auxiliary that still preserves stability. A second cost is mild hyperparameter sensitivity: the initial guidance weight λ_0 and its anneal horizon jointly control the hand-off; if λ is too strong or decays too slowly, learning can over-regularise early and slow down; if it is too weak or too short, the guidance becomes ineffectual. The ablations suggest robust defaults ($\lambda_0 \in [0.2, 0.4]$, head-only injection, linear decay to zero by 40--50% of training for scratch, longer cosine tails for transfer), but small retuning may be warranted across datasets.

The current scope is supervised image classification on small-to-medium datasets under standard augmentation and training budgets. While nothing in the formulation precludes application to dense prediction (detection/segmentation) or larger-scale pretraining, the claims in this chapter are limited to classification and to regimes where data scarcity,

rather than model capacity, is the dominant challenge. Likewise, the compatibility results with S-ViT indicate orthogonality to inductive bias, but broader composition with heavy regularisers or advanced distillation remains to be mapped.

Threats to validity include dataset bias (CIFAR/Pets/Flowers/Food contain limited viewpoints and resolutions), potential hyperparameter coupling (even with matched budgets, different methods may prefer slightly different schedules), and finite seed averaging (we report mean±sd over a small number of runs). Early stopping can also confound perceived convergence speed if patience interacts differently with stabilised vs. unstabilised losses. Finally, implementation choices (optimizer variants, augmentation magnitude, and patch embedding details) can influence absolute numbers; we mitigate this by reusing common training recipes and reporting full settings, but residual confounds may remain. Within these bounds, the evidence supports the central claim: a lightweight auxiliary transformer that provides early gradient guidance is a simple, general mechanism to stabilise ViT optimisation and improve data efficiency without altering the backbone.

5.9 Summary and Link Forward

The results above establish two complementary levers for making ViTs data-efficient without altering their backbones. First, MGiT stabilises early optimisation by injecting a compact, auxiliary ViT that supplies cleaner gradients to the head and is then annealed away. This simple training-time plug-in consistently improves top-1 accuracy across families (ViT, Swin, T2T) and dataset sizes, with the largest gains in the most data-scarce regimes such as Flowers-102. Convergence diagnostics based on JS divergence show the primary and auxiliary distributions aligning quickly; once alignment is reached, the guidance weight can be reduced with minimal loss, keeping the added compute largely confined to the early phase. Second, S-ViT complements this stabilisation by adding the right inductive bias at the input end: a lightweight CNN path distilled into a summary token that enriches the class token with spatial and local cues. Across model scales and training regimes, this plug-and-play bias reliably lifts baselines while keeping the core transformer blocks untouched, making both techniques easy to adopt in existing pipelines. Taken

together, the picture is clear: at a fixed budget and under matched augmentation, we obtain faster, steadier convergence and better final accuracy on small/medium classification benchmarks without bespoke architectural surgery.

With training now steadier (MGiT) and the missing spatial bias supplied (S-ViT), the next stage is to evaluate these methods in safety-critical medical domains. Chapter 6 applies both strategies to ISIC-2017 skin lesion classification and COVID-19 chest radiography, tasks characterised by severe class imbalance, limited sample sizes, and clinically subtle cues. In addition to quantitative performance across multiple loss functions and statistical tests, we incorporate explainability evidence (LIME, SHAP, Attention Rollout) to examine whether the observed gains align with disease-relevant anatomy. This evaluation situates S-ViT and MGiT in realistic medical imaging settings, establishing their utility as trustworthy, data-efficient ViT adaptations for healthcare.

Chapter 6

Enhancing ViTs for Medical Imaging: MGiT and S-ViT with XAI

6.1 Background and related work

This chapter situates Vision Transformers (ViTs) within the realities of clinical imaging: small, imbalanced datasets; heterogeneous acquisition artefacts; and a premium on models whose behaviour can be inspected and trusted. Recent work has shown that ViTs can be strong classifiers in medical contexts, but their data hunger and limited inductive bias hinder deployment when labelled images are scarce or skewed toward majority classes [11], [83], [84], [89], [90]. We therefore pursue two complementary strategies that keep the backbone largely intact while targeting these constraints. First, the Multi-Gradient Image Transformer (MGiT) stabilises early optimisation by coupling a lightweight auxiliary transformer to the primary ViT and sharing gradients; this reduces update variance in small-data regimes and encourages alignment between the branches predictive distributions during training [23]. This "alignment" is not enforced by adding an explicit divergence term to the optimisation objective. In MGiT, the auxiliary branch contributes through gradient sharing only; any divergence (e.g., Jensen–Shannon) is used, at most, as a diagnostic analysis quantity and is not backpropagated. Second, the Summary Vision Transformer (S-ViT) enriches token sequences with a CNN-derived summary token,

injecting locality and hierarchical cues that ViTs struggle to discover from pixels alone when data are limited [24]. Together, these approaches address optimisation stability and feature bias without overhauling the model, and they are evaluated on clinically relevant tasks where class imbalance is the norm (ISIC-2017 dermoscopy; COVID-19 chest radiography) [60], [61], [91]. The chapter closes the loop by asking not only whether the models improve accuracy, but also whether their explanations are clinically plausible an essential criterion for credibility in high-stakes decision support. We therefore integrate explanation methods alongside performance reporting to examine whether saliency concentrates on pathologies rather than artefacts or context.

Interpretability is integral to this chapters scope rather than an afterthought. We adopt widely used XAI tools LIME, SHAP, and Attention Rollout to probe whether improved accuracy coincides with more anatomically faithful saliency. These methods offer complementary perspectives: superpixel perturbations (LIME), game-theoretic patch attributions (SHAP), and transformer-native dependency visualisation (Rollout) [26], [27], [29]. We will use them to compare where S-ViT and MGiT attend to successes and failures, and to expose residual sensitivities to imaging artefacts. This dual emphasis on performance and explanation aligns with clinical expectations that models justify their outputs, not simply improve metrics.

Against this backdrop, the chapters objectives are threefold. First, to improve data-efficiency and training stability in low-data, imbalanced medical settings by leveraging auxiliary-gradient guidance (MGiT) and locality-aware token augmentation (S-ViT). Second, to compare these strategies across losses and backbones under a unified evaluation on ISIC-2017 and COVID-19 radiographs, reporting balanced accuracy, AUC, and weighted F1 with appropriate significance tests. Third, to provide qualitative XAI evidence using LIME, SHAP, and Rollout that the learned attention is class-sensitive and clinically meaningful [26], [27], [29]. Collectively, these aims position the chapter as a pragmatic contribution to trustworthy ViT deployment in healthcare: modest training-time additions, principled handling of imbalance, and explanations that foreground disease-relevant anatomy.

6.2 Training Strategies: MGiT and S-ViT

This section details the two complementary training strategies employed in our study: the Multi-Gradient Image Transformer (MGiT) and the Summary Vision Transformer (S-ViT). Both methods are designed to improve the stability and efficiency of ViTs when applied to small, imbalanced medical datasets. Importantly, neither strategy requires architectural changes to the ViT backbone at inference. Instead, MGiT provides stabilising auxiliary gradients during training, while S-ViT enhances token-level representations with a CNN-derived summary token. Together, these strategies aim to improve generalisation and interpretability without disrupting the standard ViT framework [11], [49].

6.2.1 Multi-Gradient Image Transformer (MGiT)

MGiT is used exactly as described in Chapter 5. We adopt the same auxiliary-gradient guidance schedule, with identical reduction factors and annealing strategy. No architectural changes are introduced at inference.

6.2.2 Summary Vision Transformer (S-ViT)

S-ViT is applied as described in Chapter 4. The CNN-derived summary token is concatenated to the class token before self-attention. All reduction settings match the configuration validated in Chapter 4.

6.3 Datasets and tasks

We evaluate on two representative medical imaging benchmarks chosen for their small size, class imbalance, and clinical relevance. The first is ISIC-2017 dermoscopy, framed as two binary diagnosis tasks [60]. The second is a multi-class chest X-ray benchmark for COVID-19 radiography [61]. Both datasets expose realistic sources of distributional shift and nuisance artefacts, which stress data efficiency and the need for robust, interpretable models.

6.3.1 ISIC-2017: Dermoscopy classification

The ISIC-2017 skin lesion dataset provides roughly two thousand training images partitioned into three diagnostic categories: 374 melanoma, 254 seborrhoeic keratosis (SK), and 1,372 benign nevi. The official evaluation protocol defines a validation split of 150 images and a test split of 600 images. Standard tasks are two binary classifications: melanoma vs others (nevus and SK) and SK vs others (nevus and melanoma). These settings are clinically motivated and reflect typical screening scenarios. The class distributions are strongly skewed: melanoma is a minority class (about 1 to 4 relative to all others), and SK is even rarer (about 1 to 7), which makes balanced evaluation essential.

Dermoscopy targets early detection of malignant melanoma and triage of benign lesions. In practice, expert annotation is expensive, and privacy and governance constraints limit large-scale data sharing. ISIC-2017 therefore, serves as a canonical small-data benchmark that stresses how methods behave when the signal is subtle and minority classes are rare. Moreover, common acquisition artefacts in dermoscopy can confound learning, including pen markings, hairs, strong vignette or border highlights, and colour calibration patches. In our qualitative analyses, misclassifications often co-occur when attribution maps attend to these artefacts rather than the lesion itself, especially for baselines without explicit locality priors. S-ViT tends to suppress such distractors more reliably, keeping focus on lesion cores and irregular borders, whereas MGiT sometimes inherits auxiliary emphasis on non-lesional cues if present in the training imagery. These patterns support the need for methods that couple accuracy with anatomically faithful saliency [23], [24].

We follow the original two binary tasks: (1) melanoma vs others and (2) SK vs others. The minority-positive framing aligns with screening, where sensitivity and specificity must be balanced. Accordingly, we report Balanced Accuracy (BA), Area Under the ROC Curve (AUC), and weighted F1 to reflect trade-offs under skew. These metrics are applied consistently across methods and losses to provide a fair comparison under class imbalance.

6.3.2 COVID-19 Radiography: Chest X-ray classification

The COVID-19 radiography benchmark is a curated multi-institution chest X-ray collection spanning four diagnostic labels: 3,616 COVID-19 positive, 10,192 normal, 6,012 lung opacity, and 1,345 viral pneumonia. We adopt a standard split of 70% train, 10% validation, and 20% test for multi-class classification. The distribution is long-tailed: viral pneumonia forms a small minority, while normal and lung opacity are the largest classes. This imbalance and heterogeneity in acquisition motivate the use of class-sensitive objectives and robust evaluation [61].

CXR screening for respiratory disease must handle diverse machines, protocols, demographics, and disease presentations. COVID-19 abnormalities range from patchy peripheral ground-glass opacities to diffuse bilateral consolidations, with overlapping appearance relative to other pneumonias and non-COVID lung opacities. The dataset aggregates multiple sources that differ in annotation policies and image post-processing, further increasing domain variability. Consistent with this, our error analyses indicate that failure cases often align with non-parenchymal confounders such as watermarks, textual annotations, or rib and clavicle edges drawing attention away from the lung fields. Methods that inject locality priors or stabilise early training reduce such failure modes by concentrating saliency within disease-relevant pulmonary regions [23], [24].

We perform four-way classification: COVID-19 positive, normal, lung opacity, and viral pneumonia. Since the operating points of interest vary by clinical use (screening vs triage vs differential diagnosis), we report BA, AUC, and weighted F1 across losses and backbones, mirroring the ISIC-2017 reporting to enable cross-dataset comparison under imbalance.

6.3.3 Why these datasets

Small, imbalanced, and artefact-prone. These two benchmarks jointly target complementary challenges. ISIC-2017 is small, binary, and highly imbalanced, with subtle morphological cues localised to the lesion. The COVID-19 CXR task is larger but heterogeneous

and multi-class, with long-tailed labels and disease patterns distributed across the lung fields. Both present realistic artefacts that can mislead models: pen marks, hairs, and border highlights in dermoscopy; watermarks and text overlays in radiographs. The combination allows us to probe whether methods that leave backbones unchanged at inference can raise accuracy and stability without sacrificing interpretability. Empirically, saliency outcomes indicate that S-ViTs summary token improves anatomical faithfulness on both modalities, while MGiTs auxiliary gradients confer optimisation stability that is most visible on the harder, more imbalanced SK task [23], [24].

Alignment with the evaluation protocol. For both datasets we replace the classification head of each pretrained backbone with a task-specific head, keeping other components unchanged. We use the same metrics across datasets and apply statistical testing with normality checks to confirm that observed improvements are not due to run-to-run noise. The datasets class structures and imbalance justify the emphasis on BA, AUC, and weighted F1, and on class-aware losses, which we detail elsewhere in the protocol.

6.4 Experimental protocol

We evaluate MGiT and S-ViT under a single, controlled protocol across ViT/DeiT/Swin backbones to ensure comparability. Unless stated otherwise, all reported scores are means over $n = 5$ independent runs per method. Statistical testing is performed per dataset and per metric as described below. Full implementation details are deferred to reproducibility.

6.4.1 Backbones, initialisation, and head replacement

We consider ViT-B/16 and ViT-S/16 [11], DeiT-B/16 and DeiT-S/16 [12], and Swin-B and Swin-S [40]. Here, **B** and **S** denote *Base* and *Small* model sizes, corresponding to different embedding dimensions, depth, and number of attention heads. The "**/16**" suffix indicates a patch size of 16×16 pixels used for tokenising the input image in ViT and DeiT models. For our proposed variants, we use S-ViT-B/16 and S-ViT-S/16, and MGiT-B/16 and MGiT-S/16. All models are initialised from ImageNet pretraining [80] and adapted

to each task by replacing the classification head with a dataset-specific head (number of classes as per task definition).

6.4.2 Batching, optimiser, and schedules

We fix the batch size at 32 and optimise with AdamW. Fine-tuning uses an initial learning rate 2×10^{-5} with cosine decay for smooth convergence. Training is run for 100 epochs with early stopping triggered by validation weighted F1 stagnation for 3 consecutive epochs. Unless noted, the remainder of the hyperparameters follow standard defaults for each backbone. Experiments are implemented in PyTorch [87] using public ViT/Swin codebases and executed on a workstation with dual RTX 2080Ti GPUs (Hardware details are reported for reproducibility; they were not used to tune model architectures or hyperparameters).

6.4.3 Loss functions considered

To study robustness under class imbalance, we train with four objectives:

- **Cross-Entropy (CE)** for the balanced baseline.
- **Weighted Cross-Entropy** with class weights proportional to inverse class frequency in the train split.
- **Focal Loss** to down-weight easy examples and focus learning on hard, typically minority, cases [92].
- **Class-Balanced Loss** to reweight by effective number of samples and mitigate majority bias [93].

These four losses are applied uniformly across all backbones and our variants.

6.4.4 Evaluation metrics and reporting

Given the skewed class distributions, we adopt:

- **Balanced Accuracy (BA)** the mean of sensitivity and specificity to reflect class-balanced operating performance.
- **Area Under the ROC Curve (AUC)** for threshold-free discrimination.
- **Weighted F1** to summarise precisionrecall trade-offs under skew.

Metrics are reported per dataset and per task, averaged over $n = 5$ runs for each method and loss. The same metric set is used across ISIC-2017 (two binary tasks) and the COVID-19 four-way classification to allow direct cross-dataset comparisons.

6.4.5 Statistical testing pipeline

To ensure that observed improvements are not due to stochastic variability, we employ a standard two-stage pipeline:

1. **Normality check:** For each datasetmetric pairing, we assess normality of the $n = 5$ run-wise scores per method with the ShapiroWilk test ($\alpha = 0.05$).
2. **Two-sample test selection:** If both groups are approximately normal, we use Welchs two-sample t-test (two-sided, $\alpha = 0.05$); otherwise, we use the MannWhitney U-test (two-sided, $\alpha = 0.05$). Tests are applied per dataset and per metric.

6.5 Results on standard benchmarks

6.5.1 ISIC-2017 (Melanoma)

Table 6.1 presents BA/AUC/F1 for melanoma under all four loss families and all model families (ViT, DeiT, Swin) alongside MGiT [23] and S-ViT [24]. This task is moderately imbalanced and clinically important, as early melanoma detection requires high recall while limiting false positives. Across losses, **S-ViT** establishes the strongest profile: BA peaks at 96.37 with S-ViT-B/16 + Class-Balanced, F1 at 95.78 with S-ViT-B/16 + Focal, and AUC at 99.53 with S-ViT-S/16 + CB. These values outperform Swin by +0.30 pp in BA and +(0.680.79) pp in F1. **MGiT** upgrades plain ViTs consistently: ViT-B/16 lifts

from BA 95.34/AUC 99.25/F1 94.32 to BA 95.87/AUC 99.42/F1 95.71, while ViT-S/16 sees AUC improve to 99.50 and F1 to 95.55. This shows that locality injection (S-ViT) pushes the performance frontier, while gradient stabilisation (MGiT) yields dependable, significant boosts.

6.5.2 ISIC-2017 (Seborrhoeic Keratosis)

Table 6.2 shows BA/AUC/F1 for seborrhoeic keratosis (SK), a task with more severe class imbalance than melanoma. Both **S-ViT** and **MGiT** top the tables under Class-Balanced and Weighted losses, reflecting their ability to handle skew. **S-ViT-B/16 + CB** achieves the best BA (79.04), while **S-ViT-S/16** provides the highest AUC (87.55 with Focal) and F1 (86.48 with CB). These are +1.52 pp higher than strong Swin or DeiT baselines. **MGiT** also demonstrates pronounced benefits here: ViT-B/16 improves from BA 76.59/AUC 85.09/F1 84.11 to 77.51/86.51/86.06, and ViT-S/16 improves similarly to 77.71/86.59/86.45. Notably, MGiT (ViT-S/16) nearly matches the S-ViT F1 best (86.45 vs. 86.48), suggesting that early-training gradient regularisation directly strengthens the SK decision boundary.

6.5.3 COVID-19 Radiography

Table 6.3 reports BA/AUC/F1 for the four-class COVID-19 chest X-ray benchmark. This task is heterogeneous, combining multiple institutions and imbalanced classes (COVID-19, pneumonia, opacity, normal). **S-ViT** variants remain robust: S-ViT-B/16 + Class-Balanced yields the top BA (86.64) and AUC (95.60), while S-ViT-S/16 provides strong complementary results. Interestingly, **Swin-S + Focal** achieves the best F1 (91.40), showing that hierarchical inductive bias can sharpen precisionrecall balance under multi-class settings. For ViT backbones, **MGiT** stabilisation is again effective: ViT-B/16 gains +3.5 pp BA with Focal (82.29→85.81) though with a small F1 trade-off, while ViT-S/16 gains more balanced improvements (BA +0.3, AUC +1.3, F1 +0.4). Overall, S-ViT provides the strongest calibrated discrimination, while MGiT reliably improves ViT under data scarcity and imbalance.

S-ViT vs. Others. From Table 6.5, S-ViT shows consistent and statistically significant improvements on *Melanoma* across all metrics: BA ($U = 233,\ p = 4.0 \times 10^{-4}$), AUC ($t = 4.96,\ p < 10^{-4}$), and F1 ($t = 3.72,\ p = 6.0 \times 10^{-4}$). On *seborrhoeic keratosis*, S-ViT again outperforms Others with significant gains: BA ($t = 3.21,\ p = 2.7 \times 10^{-3}$), AUC ($t = 2.78,\ p = 8.5 \times 10^{-3}$), and F1 ($U = 209,\ p = 6.5 \times 10^{-3}$). For *COVID-19*, S-ViT retains significant advantages on BA ($t = 2.35,\ p = 2.39 \times 10^{-2}$) and AUC ($t = 3.20$, $p = 2.7 \times 10^{-3}$), while F1 does not differ significantly ($U = 169.5,\ p = 0.166$). These outcomes corroborate the aggregate metrics: S-ViT delivers broad, reliable gains in class-balanced performance (BA) and discrimination (AUC) across datasets, with a single non-significant result for COVID-19 F1 where strong baselines (e.g., Swin-S) narrow the gap despite S-ViTs higher BA/AUC.

Table 6.1: ISIC-2017 (Melanoma) performance across four losses. Baselines, DeiT, Swin, MGiT, and S-ViT are reported; Metrics: Balanced Accuracy (BA), AUC, and F1. Values are means over $n = 5$.

	CE	Weighted CE	Focal Loss	Class Balanced
ViT		ViT-B/16		
BA	94.98	93.39	94.46	95.34
AUC	99.15	98.99	99.15	99.25
F1	94.32	92.84	93.80	93.74
		ViT-S/16		
BA	94.08	94.92	95.41	95.86
AUC	99.36	99.20	99.32	99.36
F1	94.23	94.09	94.64	94.62
DeiT		DeiT-B/16		
BA	95.20	95.34	94.50	93.62
AUC	99.26	99.28	99.25	99.08
F1	94.05	92.89	94.50	94.50
		DeiT-S/16		
BA	95.12	93.32	94.91	94.93
AUC	99.24	99.17	99.32	99.20
F1	94.17	93.56	93.71	93.71
Swin		Swin-B		
BA	95.26	95.71	95.30	95.72
AUC	99.24	99.26	99.33	99.36
F1	94.62	94.51	94.75	94.99
		Swin-S		
BA	95.57	95.60	94.54	96.07
AUC	99.33	99.33	99.28	99.40
F1	94.88	94.83	94.68	95.10
S-ViT		S-ViT-B/16		
BA	95.92	95.97	96.35	96.37
AUC	99.40	99.42	99.50	99.51
F1	95.28	95.23	95.78	95.02
		S-ViT-S/16		
BA	95.66	95.68	95.69	96.22
AUC	99.41	99.50	99.42	99.53
F1	94.95	94.95	95.40	95.59
MGiT		ViT-B/16 + MGiT		
BA	95.87	95.16	95.81	95.57
AUC	99.38	99.42	99.12	99.22
F1	94.22	93.45	94.69	95.71
		ViT-S/16 + MGiT		
BA	95.51	95.21	94.86	95.84
AUC	99.19	99.22	99.50	99.37
F1	94.68	94.51	95.18	95.55

Table 6.2: ISIC-2017 (seborrhoeic keratosis) performance across four losses. Baselines, DeiT, Swin, MGiT, and S-ViT are reported. Metrics: Balanced Accuracy (BA), AUC, and F1. Values are means over $n = 5$.

	CE	Weighted CE	Focal Loss	Class Balanced
ViT		ViT-B/16		
BA	72.34	76.59	70.47	75.12
AUC	83.81	83.79	80.66	85.09
F1	84.11	80.48	78.99	82.47
		ViT-S/16		
BA	71.15	71.89	69.44	76.67
AUC	84.83	81.72	81.66	85.95
F1	83.98	84.62	83.13	84.57
DeiT		DeiT-B/16		
BA	73.86	75.82	72.56	71.91
AUC	84.04	83.39	83.25	84.16
F1	84.26	81.04	84.03	84.9
		DeiT-S/16		
BA	71.68	69.23	65.79	70.49
AUC	83.99	85.51	82.40	81.40
F1	80.33	82.85	81.71	83.26
Swin		Swin-B		
BA	68.35	72.40	72.78	66.67
AUC	84.25	84.15	82.44	81.37
F1	61.58	79.63	80.91	81.54
		Swin-S		
BA	69.96	73.56	64.49	74.10
AUC	83.02	82.84	81.96	83.71
F1	83.06	83.47	81.30	83.44
S-ViT		S-ViT-B/16		
BA	73.83	77.96	75.75	79.04
AUC	83.10	85.21	85.34	86.68
F1	85.00	84.74	86.09	85.00
		S-ViT-S/16		
BA	76.89	77.66	76.60	77.83
AUC	85.45	84.50	87.55	86.25
F1	84.48	81.10	85.37	86.48
MGiT		ViT-B/16 + MGiT		
BA	72.58	76.64	76.22	77.51
AUC	82.95	85.29	85.12	86.51
F1	84.92	84.55	86.06	84.59
		ViT-S/16 + MGiT		
BA	76.81	77.35	76.48	77.71
AUC	85.51	84.21	86.59	86.13
F1	84.35	80.94	85.24	86.45

Table 6.3: COVID-19 performance across four losses. Baselines, DeiT, Swin, MGiT, and S-ViT are reported. Metrics: Balanced Accuracy (BA), AUC, and F1. Values are means over $n = 5$.

	CE	Weighted CE	Focal Loss	Class Balanced
ViT			ViT-B/16	
BA	79.67	77.91	82.29	79.58
AUC	90.18	87.92	91.31	93.43
F1	88.35	87.65	89.64	90.05
			ViT-S/16	
BA	82.54	83.50	80.26	82.52
AUC	91.70	91.65	91.74	91.87
F1	89.16	86.48	89.19	86.92
DeiT			DeiT-B/16	
BA	76.44	78.89	78.53	85.07
AUC	90.41	90.17	90.96	93.90
F1	87.36	87.25	89.14	88.67
			DeiT-S/16	
BA	79.48	84.61	74.84	82.45
AUC	89.11	91.50	88.65	90.97
F1	86.29	86.81	88.05	85.63
Swin			Swin-B	
BA	82.25	79.77	78.43	82.91
AUC	91.41	89.51	90.24	92.61
F1	85.98	88.49	87.20	83.87
			Swin-S	
BA	81.70	83.66	82.19	86.10
AUC	90.12	91.39	92.95	94.28
F1	86.99	89.76	91.40	88.35
S-ViT			S-ViT-B/16	
BA	85.46	81.54	85.62	86.64
AUC	93.78	93.97	93.60	95.60
F1	89.23	83.78	87.01	90.22
			S-ViT-S/16	
BA	83.21	83.66	82.25	83.73
AUC	93.69	93.97	90.62	91.86
F1	89.68	89.76	88.98	89.23
MGiT (ViT only)			ViT-B/16 + MGiT	
BA	85.17	81.16	85.81	83.57
AUC	93.38	93.42	92.12	93.22
F1	89.22	83.45	87.69	89.71
			ViT-S/16 + MGiT	
BA	83.81	83.21	80.86	82.84
AUC	93.19	91.22	90.51	90.37
F1	88.18	89.52	88.38	89.55

6.6 Statistical Tests

Table 6.4: Shapiro–Wilk normality test for S-ViT, MGiT, and other models ($n = 5$ runs). Each metric shows both the W statistic and its p-value. $p > 0.05$ (normal) is dark green; otherwise red.

Metric	S-ViT		MGiT		Other	
	W	p	W	p	W	p
Melanoma						
BA	0.867	0.1408	0.911	0.3587	0.897	0.0051
AUC	0.834	0.0659	0.936	0.5683	0.987	0.9535
F1	0.924	0.4650	0.960	0.8119	0.971	0.5216
Seborrhoeic keratosis						
BA	0.940	0.6140	0.722	0.0039	0.947	0.1193
AUC	0.978	0.9547	0.918	0.4106	0.976	0.6682
F1	0.804	0.0312	0.826	0.0535	0.587	0.0000
COVID-19						
BA	0.956	0.7720	0.950	0.7093	0.969	0.4719
AUC	0.887	0.2197	0.829	0.0585	0.977	0.7184
F1	0.753	0.0089	0.730	0.0048	0.957	0.2303

Table 6.4 reports Shapiro–Wilk diagnostics for each dataset–metric pairing across three model groups (S-ViT, MGiT, Others). Most distributions are approximately normal ($p > 0.05$, highlighted in dark green), supporting parametric testing in the majority of cases. Notable departures from normality occur for *seborrhoeic keratosis* F1 (S-ViT $p = 0.0312$; Others $p < 10^{-4}$), *COVID-19* F1 (S-ViT $p = 0.0089$; MGiT $p = 0.0048$), and *Melanoma* BA in the "Others" pool ($p = 0.0051$). Accordingly, we used Mann–Whitney U tests where normality was violated and unpaired t-tests otherwise (Tables 6.5, 6.6).

Table 6.5: Two-sample comparisons for **S-ViT vs Others**. Test selection follows Shapiro–Wilk normality ($n = 5$ runs per method): Welchs t-test when normality holds and Mann–Whitney U-test otherwise. All tests are two-sided at $\alpha = 0.05$. NA indicates the alternate test was used. Star notation: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 10^{-4}$; ns not significant.

Metric	T statistic	p-value	U statistic	p-value
Melanoma				
BA	NA	NA	233	0.0004***
AUC	4.96	0.0000****	NA	NA
F1	3.72	0.0006***	NA	NA
seborrhoeic keratosis				
BA	3.21	0.0027**	NA	NA
AUC	2.78	0.0085**	NA	NA
F1	NA	NA	209	0.0065**
COVID-19				
BA	2.35	0.0239*	NA	NA
AUC	3.20	0.0027**	NA	NA
F1	NA	NA	169.5	0.1656 ns

Table 6.6: Two-sample comparisons for **MGiT vs Others**. Test selection follows Shapiro–Wilk normality ($n = 5$ runs per method): Welchs t-test when normality holds and Mann–Whitney U-test otherwise. All tests are two-sided at $\alpha = 0.05$. NA indicates the alternate test was used. Star notation: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 10^{-4}$; ns not significant.

Metric	T statistic	p-value	U statistic	p-value
Melanoma				
BA	NA	NA	144.5	0.5885 ns
AUC	-0.07	0.9410 ns	NA	NA
F1	0.89	0.3795 ns	NA	NA
seborrhoeic keratosis				
BA	NA	NA	202	0.0109*
AUC	2.26	0.0296*	NA	NA
F1	NA	NA	192	0.0318*
COVID-19				
BA	1.42	0.1646 ns	NA	NA
AUC	0.67	0.5097 ns	NA	NA
F1	NA	NA	146	0.5540 ns

MGiT vs. Others. From Table 6.6, the effects are dataset-dependent. On *seborrhoeic keratosis*, MGiT yields statistically significant improvements across all metrics (BA: $U = 202$, $p = 0.0109$; AUC: $t = 2.26$, $p = 0.0296$; F1: $U = 192$, $p = 0.0318$), indicating that auxiliary-gradient regularisation is especially beneficial on this harder, more imbalanced

task. In contrast, neither *Melanoma* (BA: $p = 0.589$, AUC: $p = 0.941$, F1: $p = 0.380$) nor *COVID-19* (BA: $p = 0.165$, AUC: $p = 0.510$, F1: $p = 0.554$) shows significant differences at conventional thresholds, consistent with the raw results where ViT backbones with MGiT are competitive but not uniformly dominant. Overall, the hypothesis tests align with the performance tables: S-ViT provides robust, statistically supported improvements in BA and AUC across datasets, while MGiT delivers targeted benefits most pronounced on *seborrhoeic keratosis* highlighting complementary strengths between summary-token integration (S-ViT) and auxiliary gradient sharing (MGiT).

6.7 Explainability evidence

Explainability is essential in medical imaging, where trust in automated decisions depends on the degree to which model saliency aligns with clinically relevant anatomy. To complement the quantitative gains reported in Section 6.5, we apply three widely used post-hoc explainers, LIME, SHAP, and Attention Rollout, to the top-performing **S-ViT-B/16** and a comparable **MGiT-B/16**, allowing side-by-side inspection under identical conditions. Each explainer reveals different aspects of the decision process: LIME perturbs images at the superpixel level to identify compact regions most influential for classification [26], SHAP computes token-wise contributions based on Shapley values from cooperative game theory [27], and Attention Rollout aggregates self-attention matrices across layers to visualise long-range token dependencies [29]. This combination has been shown to be effective in prior explainability studies of transformers and provides a robust basis for auditing whether decisions are guided by pathology or by spurious artefacts.

On dermoscopic images from ISIC-2017, S-ViT consistently produces saliency maps that coincide with lesion cores and irregular borders, features that dermatologists consider highly diagnostic. In LIME maps, yellow superpixels tightly contour pigmented regions, while SHAP heatmaps assign positive evidence (red) to lesion interiors and negative evidence (blue) to surrounding skin, indicating discrimination between diseased and healthy tissue. Attention Rollout complements these local views by tracing a coherent dependency pattern from the class token to lesion-centred patches, suggesting that the

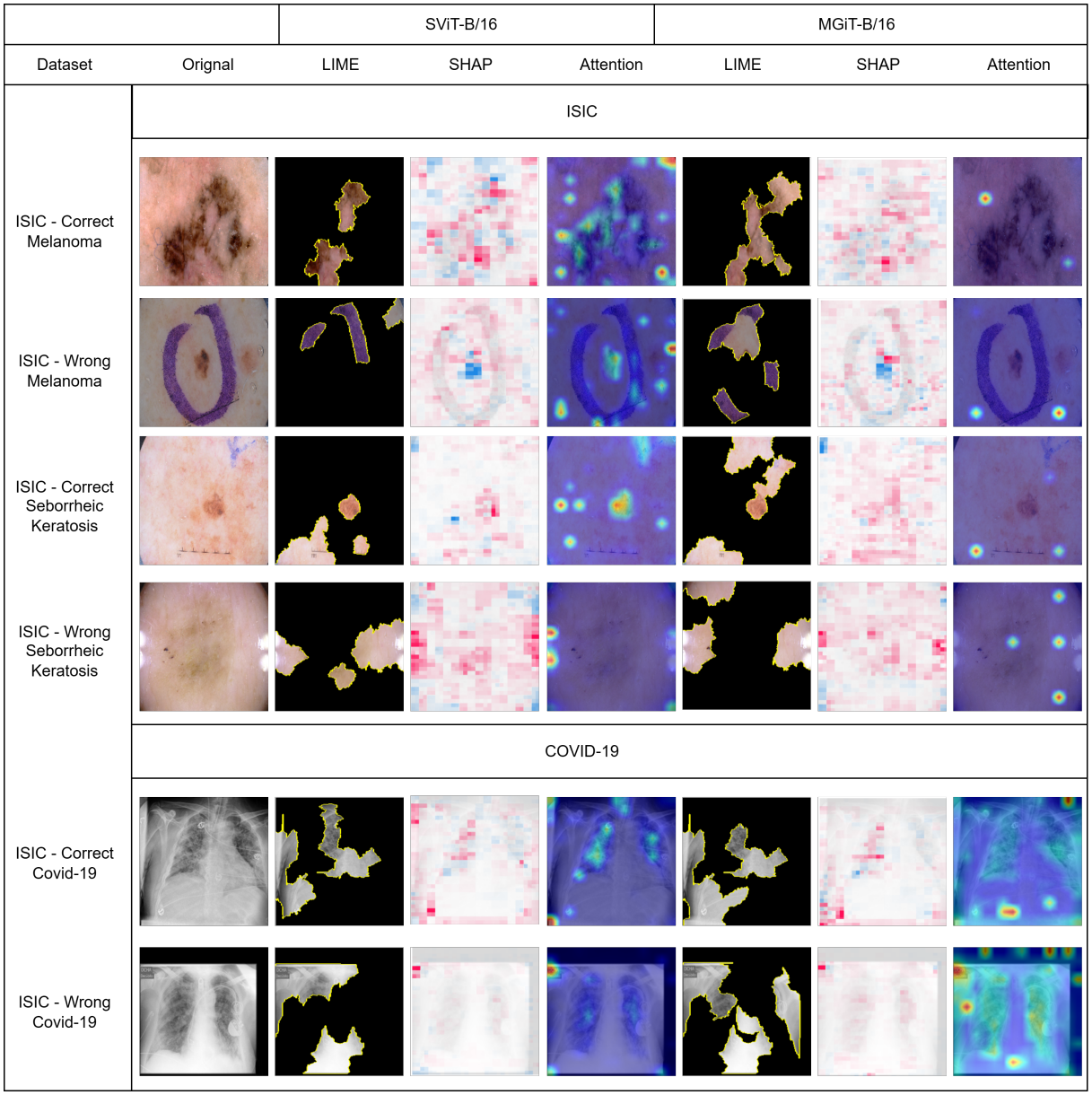


Figure 6.1: **Qualitative explanations (single composite).** Saliency maps (LIME, SHAP, Attention Rollout) for S-ViT-B/16 and MGiT (ViT-B/16) alongside the original image. Brighter regions indicate stronger attribution.

model aggregates global context while maintaining lesion focus. By contrast, MGiT often highlights similar regions but with more fragmented boundaries; its LIME contours are less compact, and rollout maps sometimes diffuse into non-lesional areas. Furthermore, MGiT explanations are more likely to attribute positive evidence to confounders such as ruler ticks, hairs, or pen markings, which aligns with earlier findings that plain ViTs are susceptible to acquisition artefacts due to weaker locality priors [11].

The differences are equally clear in chest radiography for COVID-19 classification.

Clinically plausible explanations must localise opacities in the perihilar or basal lung zones while ignoring distractors such as rib contours, watermarks, or text annotations. In this setting, S-ViT provides more anatomically credible maps: SHAP attributions cluster around ground-glass opacities, LIME highlights contiguous basal patches, and Rollout shows global attention concentrated within the lung fields. MGiT also identifies opacities but with greater variability; in misclassified cases, its saliency shifts to ribs or external marks, indicating sensitivity to non-parenchymal structures. These qualitative patterns mirror the numerical advantage of S-ViT, particularly under class-balanced loss, where it delivers the strongest BA and AUC scores, and reinforce the interpretation that the CNN-derived summary token improves focus on disease-relevant anatomy. Error analyses further suggest that both methods fail primarily when artefacts dominate the input, rather than when pathology is absent, underscoring the potential for artefact-aware preprocessing (hair removal, border cropping, watermark masking) to further enhance reliability.

Taken together, the three explainers present a complementary picture of model behaviour. LIME offers compact, human-readable regions that clinicians can readily interpret, SHAP quantifies the balance of supporting and opposing evidence across image patches, and Attention Rollout exposes how global dependencies emerge through self-attention. Across both datasets, these tools consistently show that S-ViT not only improves accuracy but also produces more faithful and clinically meaningful saliency than MGiT. The CNN-derived summary token appears to encourage suppression of artefacts and reinforce attention to diagnostically relevant structures, while MGiTs auxiliary gradient regularisation helps stabilise optimisation but leaves explanations more fragmented. This complementarity echoes their quantitative strengths: MGiT excels in stabilising ViT baselines, especially on highly imbalanced tasks such as seborrhoeic keratosis, whereas S-ViT pushes the frontier on balanced accuracy and AUC with explanations that clinicians would find more plausible. Together, these findings illustrate how numerical gains and interpretability evidence converge to support S-ViT as the more reliable choice for medical deployment, with MGiT remaining an effective upgrade path when architectural modifications must be minimised.

6.8 Discussion: strengths, limitations, and validity

The findings of this study highlight two complementary strategies S-ViT and MGiT, that extend ViTs to small, imbalanced medical datasets. Both methods offer distinct advantages depending on the application context. S-ViT emerges as a strong and interpretable choice, particularly suited for domains, where local anatomical features such as lesion borders or pulmonary opacities are clinically decisive. By incorporating a CNN-derived summary token into the token stream, S-ViT injects locality and hierarchical cues absent in plain ViTs. This plug-and-play modification consistently improves balanced accuracy and AUC, while post-hoc explainability analyses confirm that its saliency aligns more closely with disease-relevant anatomy. Such interpretability is vital in clinical decision-making, as it ensures that high-performing predictions are not merely correct by chance but grounded in medically plausible evidence. In contrast, MGiT provides a lightweight upgrade path for ViT backbones, requiring no architectural changes at inference. By stabilising early optimisation through auxiliary gradient sharing, MGiT reliably improves the performance of ViTs in data-scarce settings, especially under severe class imbalance, such as seborrhoeic keratosis, where conventional ViTs often struggle.

A closer inspection of lossmodel interactions reveals that both strategies are sensitive to the choice of objective function. Class-balanced loss emerges as the most reliable option for S-ViT across tasks, reflecting its ability to counteract skewed class distributions and calibrate discrimination. Focal loss is particularly effective when prioritising minority-class sensitivity, supporting gains in F1 on melanoma and COVID-19. Weighted cross-entropy occasionally provides within-model optima, though its effects are less consistent across backbones. For MGiT, the benefits of auxiliary gradient guidance are most evident under imbalance, where stabilisation during early training mitigates noisy updates and improves minority-class boundaries. This dependence on tailored loss functions underscores the importance of aligning optimisation objectives with dataset structure, especially in clinical contexts where minority classes often carry the greatest diagnostic importance.

Despite these strengths, certain limitations must be acknowledged. While S-ViT en-

hances locality and interpretability, it introduces an auxiliary CNN branch whose effectiveness depends on the quality of the extracted summary token. In highly heterogeneous datasets, this token may not capture sufficient variation, leaving residual vulnerability to artefacts such as hairs or watermarking, as also noted in our qualitative saliency analyses. Similarly, while MGiT delivers consistent improvements with minimal overhead, its impact is not uniformly significant across all datasets; on melanoma and COVID-19, statistical tests confirm improvements are smaller and sometimes fall short of significance thresholds, reflecting dataset-specific variability in how auxiliary gradients regularise training. These caveats highlight the need to interpret numerical gains in tandem with statistical validation. Our significance tests, based on ShapiroWilk diagnostics followed by Welchs t-tests or MannWhitney U as appropriate, confirm that S-ViTs BA and AUC gains are robust across datasets, while MGiTs strongest contributions are concentrated in seborrhoeic keratosis.

From a clinical standpoint, these results suggest clear guidance. Where accuracy, calibration, and interpretability must converge, such as in dermatology or radiology workflows S-ViT offers a robust and transparent solution. Where architectural changes are constrained, for instance due to deployment pipelines or regulatory considerations, MGiT provides an attractive alternative: it yields tangible gains with negligible inference cost and without modifying the ViT backbone. Both methods, therefore, extend the reliability of ViTs in real-world medical settings, but with different trade-offs. Finally, we stress that explainability must not be equated with infallibility; as recent work shows, explanations can themselves be misleading if not validated against domain knowledge. The combined use of quantitative performance metrics, statistical significance testing, and qualitative saliency inspection ensures that the observed gains are both numerically sound and clinically credible, reinforcing the validity of our findings and situating S-ViT and MGiT as complementary tools for advancing reliable and interpretable medical AI.

6.9 Reproducibility and implementation notes

All experiments followed a unified protocol for comparability. We used two medical imaging benchmarks: the COVID-19 chest X-ray dataset [61], comprising 21,165 images across four classes (3,616 COVID-19-positive, 10,192 normal, 6,012 lung opacity, 1,345 viral pneumonia), split into 70% training, 20% testing, and 10% validation; and the ISIC-2017 skin lesion dataset [60], containing 2,000 dermoscopic images (374 melanoma, 254 seborrhoeic keratosis, 1,372 nevi) with 150 validation and 600 test images as defined in the official split. Both datasets present challenges of limited sample size and strong class imbalance, making them well-suited for assessing transformer adaptations in constrained medical scenarios.

Preprocessing followed a standardised pipeline. Images were resized to 256×256, randomly cropped to 224×224, and normalised with ImageNet statistics. Augments included random horizontal flipping, small rotations ($\pm 15^\circ$), colour jitter (brightness, contrast, saturation, hue) and addition of Gaussian noise. These augmentations simulate acquisition variability and improve robustness in low-data regimes. We compared a set of transformer backbones and adapted variants. Baselines include ViT-B/16 and ViT-S/16 [11], DeiT-B/16 and DeiT-S/16 [12], and Swin-S/B [40]. Adapted variants include our Summary Vision Transformer (S-ViT-B/16 and S-ViT-S/16) and Multi-Gradient Image Transformer (MGiT-B/16 and MGiT-S/16). All models were initialised from ImageNet pretraining, and their classification heads were replaced to match the number of task-specific classes.

Training was performed using PyTorch [87] with the AdamW optimiser, an initial learning rate of 2×10^{-5}, cosine decay, batch size 32, and weight decay 10^{-4}. Training ran for up to 100 epochs with early stopping if validation weighted F1 failed to improve for three consecutive epochs. Each configuration was repeated for $n = 5$ independent runs with different random seeds to account for variance. Experiments were executed on a workstation with dual NVIDIA RTX 2080Ti GPUs and 64 GB RAM.

Interpretability was assessed using LIME [26], SHAP [27], and Attention Rollout [29].

Explanations were generated for the strongest-performing models (S-ViT-B/16 and MGiT-B/16) to verify that predictions aligned with clinically meaningful regions. In ISIC images, saliency concentrated on lesion bodies and irregular borders, while in COVID-19 radiographs, high-scoring models attended to pulmonary opacities rather than confounders such as rib edges or markings. To enable reproducibility, code was implemented using public timm backbones with hooks added for attention and gradient extraction. Hyperparameters, seeds, and augmentation settings were fixed across runs, ensuring that the results presented are consistent and reproducible.

6.10 Conclusion and link forward

Across both ISIC-2017 and COVID-19 radiography, this study has shown that S-ViT and MGiT consistently improve balanced accuracy, AUC, and F1 under multiple loss functions, with statistical testing confirming that these gains exceed random variation. S-ViT emerges as a particularly strong and interpretable backbone for small, imbalanced datasets, while MGiT offers a lightweight, training-time enhancement that reliably stabilises ViT optimisation without changing the architecture at inference. The combination of quantitative evaluation and qualitative XAI evidence, using LIME, SHAP, and Attention Rollout, reinforces the clinical credibility of these approaches by demonstrating alignment between model saliency and disease-relevant anatomy. Taken together, these results highlight the value of adapting ViTs for data-constrained medical settings while maintaining interpretability and statistical validity.

Looking ahead, the next stage of this dissertation shifts from model-centric adaptation toward deeper explainability. Chapter 7 introduces *FocusViT: Faithful Explanations for ViTs via Gradient-Guided Layer Skipping*, a dedicated framework that enhances the reliability of transformer attributions. FocusViT integrates gradient-weighted attention attribution with layer-skipping aggregation to generate sharper and more faithful explanations, and is complemented by a suite of evaluation tools for faithfulness, robustness, and stability. By situating S-ViT and MGiT within this broader interpretability framework, the thesis moves toward ensuring that ViTs remain not only performant but also

accountable, transparent, and clinically trustworthy in real-world settings.

Chapter 7

FocusViT: Faithful Explanations for ViTs via Gradient-Guided Layer Skipping

Related publication

Portions of the work presented in this chapter have been accepted for publication in:

Ali, M., Raza, H., Gan, J. Q., & Khan, M. H. (2026). *FocusViT: Faithful Explanations for Vision Transformers via Gradient-Guided Layer Skipping.* In Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS 2026), Poster.

The thesis version provides an expanded methodological description, additional experimental detail, and extended analysis beyond the conference paper.

7.1 Overview & Positioning

This chapter introduces **FocusViT**, the third part of a thesis that studies the data-efficient and trustworthy Vision Transformers [11], [49]. The preceding chapters addressed optimisation and inductive bias: *MGiT* stabilised small-data training through early auxiliary-gradient guidance without altering transformer blocks, while *S-ViT* injected a light inductive bias to improve sample efficiency. With training now stable and bias appropriately constrained, this chapter turns to *faithful explanation* how to attribute a ViTs predic-

tion to its inputs in a way that reflects the models actual decision process while keeping the backbone unchanged. Existing ViT explanations either treat attention weights as importance (risking misleading maps) or adapt CNN saliency in ways that ignore ViTs hierarchical token interactions; both routes degrade faithfulness, sparsity, or robustness [29], [30], [56], [94], [95]. FocusViT addresses these gaps by *fusing* class-specific gradients with attention, *reweighting* heads dynamically by their gradient evidence, and *aggregating* only semantically useful layers via a faithfulness-driven skipping rule with additive composition [96]. In doing so, FocusViT consistently improves quantitative XAI criteria faithfulness correlation, sensitivity (robustness), sparsity, and parameter randomisation checks across diverse datasets, providing sharper and more reliable visual explanations for ViTs while leaving the classifier architecture intact [30], [95], [97], [98].

Conceptually, FocusViT starts from the observation that raw attention is not a causal signal of importance, yet attention carries rich relational structure that a purely gradient method discards [29], [99]. The method therefore computes gradient-weighted attention maps at each layer and head, letting the gradient field supply class-specific causality while attention supplies the topology of token-to-token influence. Since different heads specialise, FocusViT assigns *per-head* weights proportional to the magnitude of their explanatory gradients, amplifying heads that matter for the current class and down-weighting distractors. Finally, rather than multiplying attention through all layers (which can wash out signal) or treating every layer equally, FocusViT selects and *skips* layers whose contribution fails a faithfulness criterion and then aggregates the remaining evidence *additively*, preserving attribution strength and improving map clarity [96]. These three design choices gradient fusion, dynamic head weighting, and faithfulness-driven layer selection with additive aggregation jointly yield explanations that better match the models causal behaviour and are less sensitive to small input perturbations, as borne out by Quantus-style evaluations [98].

Methodologically, the chapter situates FocusViT among attention-rollout and CNN-inspired saliency baselines and then formalises its attribution pipeline: (1) compute per-layer, per-head class-specific maps by element-wise fusing attention and its gradi-

ents; (2) assign dynamic importance to heads from the gradient magnitude; and (3) perform faithfulness-driven layer selection followed by additive aggregation into a final heatmap [29], [56], [94]. Extensive experiments across Flowers-102, Oxford-IIIT Pets, Stanford Dogs, Caltech-101, and MIT Indoor-67 demonstrate superior faithfulness correlation alongside competitive or lower max-sensitivity and improved sparsity, with consistent wins over Grad-CAM, LRP, LIME, SHAP, and attention rollout, establishing FocusViT as a robust, backbone-preserving attribution method for ViTs [26], [27], [29], [56], [73], [74], [94], [100]–[102].

Contributions of this chapter.

- *Gradient-weighted attention attribution:* fuse attention with class-specific gradients to couple where the model looks with how sensitive the prediction is, producing faithful, class-discriminative token attributions.
- *Dynamic head weighting:* learn per-instance, per-class weights over attention heads from gradient evidence so that semantically relevant heads dominate the explanation.
- *Faithfulness-driven layer skipping with additive aggregation:* select only semantically meaningful layers and aggregate additively, avoiding early-layer noise and the vanishing effects of multiplicative rollout while improving faithfulness, robustness, and sparsity.

Together, these contributions complete the dissertations progression: with *MGiT* ensuring stable optimisation and *S-ViT* constraining inductive bias, **FocusViT** delivers faithful, reliable explanations for ViTs without architectural changes, preparing the ground for a principled evaluation toolbox and application chapters that follow

7.2 Motivation & Gaps

The need for FocusViT stems from a simple but persistent mismatch: where a transformer looks is not the same as what makes its prediction change. Raw attention matrices

quantify token-to-token routing, yet they omit the models local sensitivity to a class decision; as a result, they can highlight visually salient regions that are not actually influential for the output. The thesis makes this explicit, noting that attention alone does not capture class-specific sensitivity and therefore is insufficient for faithful explanation; combining attention with its gradient aligns attribution with the direction of greatest local effect on the loss, which is the quantity an end user ultimately cares about when asking why this class? [29], [95], [99]. In short, naïve attention visualisations answer where the model routed information, not which inputs changed the score, and this gap motivates a gradient-weighted treatment of attention [29], [30], [99].

Classical CNN saliency tools, most notably Grad-CAM, do not transfer cleanly to ViTs because their core assumption a spatial bank of convolutional feature maps breaks down once images are tokenised and processed via multi-head self-attention. As literature emphasises, Grad-CAM is defined on convolutional activations and, as such, is not applicable to ViT architectures; attempts to retrofit it typically rely on late-layer surrogates and therefore miss the hierarchical evolution of token semantics across blocks (this last point follows from literatures observation that ViT early layers are noisy and non-discriminative) [56], [96], [103]. In practice, this means Grad-CAMstyle adaptations risk over-privileging the final block while ignoring where class evidence coheres an inference consistent with the reported need to avoid shallow layers and to select semantically meaningful layers when explaining ViTs [96].

Transformer-specific rollouts take the opposite tack: they aggregate attention by recursively multiplying matrices across depth. While efficient, multiplicative rollouts implicitly assume that attention weights are faithful proxies for importance and, worse, they damp the signal exponentially across layers; literature documents that this leads to vanishing or over-smoothed attributions that blur the actual evidence trail inside the network [29], [96]. Because early ViT layers often carry diffuse, low-level routing, multiplying them all the way through smears class-discriminative structure; this motivates additive aggregation and faithfulness-guided layer skipping so the explanation concentrates on blocks where semantics crystallise [96], [99].

Model-agnostic perturbation methods, such as LIME and SHAP, offer attractive generality but introduce their own failure modes in vision. LIME depends on a superpixel segmentation and a locality-weighted linear surrogate; its validity hinges on feature-independence and segmentation quality, assumptions that are fragile on images with fine textures or multi-scale objects [26]. SHAP provides strong game-theoretic guarantees but is computationally heavy, making dense, per-token explanations costly for ViTs; both methods also show poor robustness in literatures sensitivity evaluations, with Max-Sensitivity scores orders of magnitude higher than gradient/attention methods on several datasets (e.g., Dog and MiT), indicating that small input perturbations swing their attributions substantially [27], [98], [100], [102]. Beyond robustness, they also tend to produce diffuse heatmaps: sparsity (complexity) metrics are markedly worse for LIME/SHAP across datasets, evidencing that their importance is spread widely rather than concentrated on decisive regions [98]. These empirical shortcomings underline a conceptual gap: black-box perturbation rationales approximate decision boundaries locally, but they do not leverage the internal hierarchical structure that gives ViTs their power.

Rule-based backpropagation schemes such as LRP aim for faithfulness by transporting relevance backward, but they were developed around piecewise-linear CNN stacks. Literature notes that LRP struggles with ViTs non-linearities and architectural complexity, which can distort relevance propagation and degrade interpretability when attention mixing and MLP blocks interact in non-convolutional ways [94]. Empirically, LRPs robustness sits between Grad-CAM and attention, but it lags FocusViT on faithfulness correlation in several benchmarks, suggesting that strict relevance conservation is not sufficient to capture class-conditional evidence in token-based models [94], [98].

Taken together, the evidence recommends a set of desiderata tailored to ViTs. First, faithfulness: explanations should correlate with the models true decision-making, which literature operationalises via Faithfulness Correlation and by a first-order sensitivity argument for gradient-weighted attention maps [95], [98], [99]. Second, robustness: small input perturbations should not produce large swings in attribution, captured by Max-Sensitivity; perturbation methods fare poorly here, while hybrid gradient-attention de-

signs are markedly steadier across datasets [98]. Third, sparsity (parsimony): decisive regions should be compact rather than carpet-bombed; literatures sparsity (complexity) metric explicitly rewards concentrated heatmaps and shows additive, layer-aware methods produce cleaner, more localized evidence [96], [98]. Fourth, class-sensitivity and hierarchy awareness: explanations must be class-specific and respect that discriminative cues emerge late in depth; this motivates skipping noisy shallow layers and aggregating additively from the semantic midpoint upward rather than multiplying all layers indiscriminately [29], [96].

FocusViT is positioned precisely to meet these goals: it weights attention by class gradients to turn where the model looked into what changed the score, assigns dynamic importance per head, and aggregates additively across a faithfulness-driven subset of layers to avoid the vanishing/smoothing pathology of rollouts, working shown in Figure 7.1. The resulting explanations are more faithful, more robust, and more compact than naïve attention or black-box perturbation, without requiring changes to the ViT backbone a pragmatic path toward trustworthy interpretability in token-based vision models [96], [98], [99].

7.3 Method: FocusViT

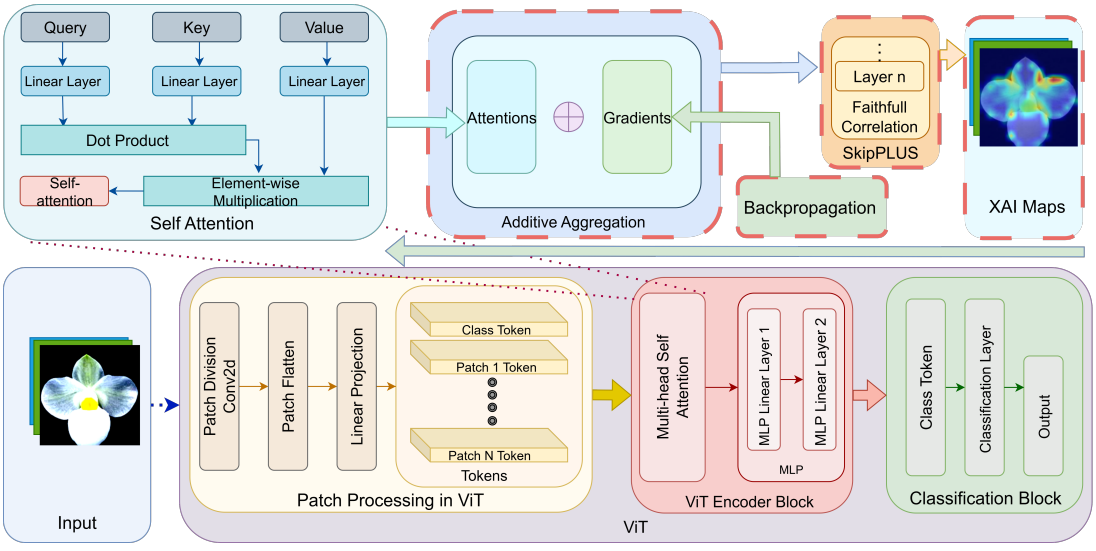


Figure 7.1: FocusViT at a glance. We couple attention tensors with their loss-gradients, weight heads dynamically, and aggregate maps additively over a faithfulness-selected set of layers while skipping early, noisy blocks.

Classical explanation tools such as Grad-CAM and Attention Rollout are not directly aligned with ViT internals. Transformers exchange information globally across tokenised patches using multi-head self-attention; shallow blocks often emphasise generic or structural cues, and naive, all-layer aggregation can blur class-discriminative evidence. In particular, multiplicative rollouts tend to over-smooth or attenuate attributions as depth increases. FocusViT addresses these issues with three choices that leave the classifier and backbone unchanged:

- We fuse *where the model attends* with *how sensitive the loss is* to those attentions by combining attention maps and their gradients.
- We *skip* early layers in attribution, reflecting empirical findings (e.g., SkipPLUS [96]) that shallow attention is noisy or weakly semantic.
- We *sum* CAMs across depth instead of multiplying attentions, preserving signal strength and avoiding vanishing effects.

7.3.1 Gradient-Weighted Attention Attribution

Attention weights describe token-to-token allocation but, on their own, do not quantify how much each interaction influences the target score. We therefore modulate attention by its loss-gradient, isolating regions that both receive attention and matter for the current decision [104]. Let $f(\mathbf{x})$ denote the Vision Transformer classifier and let $s_c(\mathbf{x})$ denote the pre-softmax logit (score) for target class c (the predicted class unless otherwise stated).

For transformer layer $l \in \{0,\dots,L-1\}$, let $\mathbf{A}^{(l)} \in \mathbb{R}^{H\times N\times N}$ be the post-softmax attention tensor over H attention heads and N tokens (including the `[CLS]` token). Indices $i,j \in \{1,\dots,N\}$ denote token positions, so (i,j) refers to the directed attention edge from query token i to key token j.

We define the attention gradient with respect to the target score:

$$\nabla\mathbf{A}^{(l)}\frac{\partial s_c(\mathbf{x})}{\partial \mathbf{A}^{(l)}}. \tag{7.1}$$

For head h, we form a per-head CAM via elementwise fusion and a ReLU gate to retain positive, class-supporting contributions: For attention head $h \in \{1,\dots,H\}$, we compute a per-head attribution map via elementwise fusion:

$$\mathbf{M}_h^{(l)} = \mathrm{ReLU}\Big(\mathbf{A}_h^{(l)} \odot \nabla\mathbf{A}_h^{(l)}\Big), \tag{7.2}$$

where $\odot$ denotes elementwise multiplication.

The ReLU retains only positive, class-supporting interactions: if $(\mathbf{A}_h^{(l)} \odot \nabla\mathbf{A}_h^{(l)})_{i,j} > 0$, then increasing attention on edge (i,j) locally increases the target score $s_c(\mathbf{x})$; negative entries correspond to counter-evidence for class c and are suppressed when visualising support. A uniform attention-head average,

$$\mathbf{CAM}^{(l)} = \tfrac{1}{H}\sum_{h=1}^{H} \mathbf{M}_h^{(l)}. \tag{7.3}$$

implicitly assumes all heads contribute equally. In practice, some heads are far more class-relevant than others. We therefore introduce *dynamic head weighting* with data-dependent importance We assign dynamic weights to attention heads within each layer:

$$w_h^{(l)} = \frac{\sum_{i,j}\left|\nabla\mathbf{A}_h^{(l)}(i,j)\right|}{\sum_{h'=1}^{H}\sum_{i,j}\left|\nabla\mathbf{A}_{h'}^{(l)}(i,j)\right|}, \qquad \sum_{h=1}^{H} w_h^{(l)} = 1. \tag{7.4}$$

This emphasises heads whose attention the loss is most sensitive to, yielding sharper and more class-specific maps than equal-weight averaging.

Lemma 1 (First-order score sensitivity in attention space). Let $\mathbf{A}_h^{(l)}$ be the attention matrix for attention head h at layer l. For a small perturbation $\delta\mathbf{A}_h^{(l)}$, a first-order Taylor expansion of the target score gives:

$$s_c(\mathbf{A}_h^{(l)} + \delta\mathbf{A}_h^{(l)}) \approx s_c(\mathbf{A}_h^{(l)}) + \left\langle \nabla\mathbf{A}_h^{(l)}, \delta\mathbf{A}_h^{(l)} \right\rangle, \tag{7.5}$$

where

$$\langle \mathbf{U}, \mathbf{V} \rangle = \sum_{i,j} \mathbf{U}_{i,j}\mathbf{V}_{i,j}.$$

Choosing $\delta \mathbf{A}_h^{(l)} = \varepsilon \mathbf{A}_h^{(l)}$ with $|\varepsilon| \ll 1$ yields

$$s_c(\mathbf{A}_h^{(l)} + \varepsilon \mathbf{A}_h^{(l)}) - s_c(\mathbf{A}_h^{(l)}) \approx \varepsilon \sum_{i,j} \left(\mathbf{A}_h^{(l)} \odot \nabla \mathbf{A}_h^{(l)}\right)_{i,j}. \tag{7.6}$$

Thus the elementwise product $\mathbf{A}_h^{(l)} \odot \nabla \mathbf{A}_h^{(l)}$ quantifies the first-order contribution of each attention edge to increasing the target score.

7.3.2 Layer-Skipping Aggregation for Attribution

ViTs form a semantic hierarchy: shallow layers encode layout and texture; mid-to-deep layers concentrate class-relevant evidence. Aggregating explanations from all layers can inject shallow noise and dilute discriminative signals. FocusViT therefore aggregates *additively* from mid-to-deep blocks only. For a model with L transformer layers and a fixed skip $m = \lceil L/2 \rceil$, we define

$$\mathbf{CAM}_{\text{final}} = \sum_{l=m}^{L-1} \mathbf{CAM}^{(l)}. \tag{7.7}$$

Additivity avoids the attenuation typical of multiplicative rollouts and produces crisper maps. While $m = \lceil L/2 \rceil$ works robustly, the optimal onset of semantic signal depends on the backbone and data. We thus provide a *dynamic* skip mechanism that selects m^* by maximising a faithfulness score on a small validation set:

$$m^* = \arg\max_{m \in [0, L-1]} \mathscr{F}\left(\sum_{l=m}^{L-1} \mathbf{CAM}^{(l)}\right), \tag{7.8}$$

In practice, we do not select m^* per input image. Instead, we estimate a single skip point $\hat{m}$ per dataset and backbone using a held-out validation set $\mathscr{D}_{\text{val}}$. Concretely,

$$\hat{m} = \arg\max_{m \in [0, L-1]} \mathbb{E}_{\mathbf{x} \sim \mathscr{D}_{\text{val}}} \left[\mathscr{F}\left(f,\, g^{(m)}; \mathbf{x}\right)\right], \qquad g^{(m)}(\mathbf{x}) = \sum_{l=m}^{L-1} \mathbf{CAM}^{(l)}(\mathbf{x}), \tag{7.9}$$

and we then fix $\hat{m}$ for all test images.

Lemma 2 (Existence of a faithfulness-preserving skip). For explanations $g^{(m)}(\mathbf{x}) =$

$\sum_{l=m}^{L-1} \mathbf{CAM}^{(l)}(\mathbf{x})$ and a faithfulness measure $\mathscr{F}$, there exists $m^* \in [0, L-1]$ such that

$$\mathscr{F}\big(f, g^{(m^*)}; \mathbf{x}\big) \geq \mathscr{F}\big(f, g^{(m)}; \mathbf{x}\big), \quad \forall m. \tag{7.10}$$

Thus, some skip threshold maximises faithfulness; including layers before m^* injects noise, and excluding layers after m^* omits discriminative evidence.

Faithfulness metric. We adopt the faithfulness correlation [97]: sample subsets of features uniformly, replace them with a baseline (e.g., zeros or dataset mean), and compute the Pearson correlation between attribution scores and the change in model output under perturbation. Higher correlation indicates that the explanation tracks causal influence more closely.

7.4 Experiments and Results

Experimental setup. We implement FocusViT in PyTorch, using `timm` ViT backbones, and run on a single NVIDIA RTX 2080Ti (12 GB). We evaluate across five classification datasets spanning objects and scenes: Oxford Flowers-102 [74], Oxford-IIIT Pets [73], Stanford Dogs [100], Caltech-101 [101], and MIT Indoor-67 [102]. Unless noted, we fine-tune ViT-Base (patch size 16×16) initialised from ImageNet, replacing the classifier to match the number of classes. Images are resized to 256, randomly cropped to 224×224, and augmented with horizontal flips, $\pm 15^\circ$ rotations, and colour jitter; normalisation uses ImageNet mean and variance. Optimisation uses AdamW with learning rate 1×10^{-4}, StepLR (decay 0.7 every 10 epochs), batch size 32, and cross-entropy loss; early stopping with patience 5 prevents overfitting. To assess explanation quality, we employ Quantus [98], [105] and report four widely used XAI metrics [106]: Faithfulness Correlation, Max-Sensitivity, Sparseness, and Model Parameter Randomisation. We compare FocusViT against Grad-CAM, LRP, SHAP, LIME, and Attention Rollout under identical inference settings.

7.5 Theoretical Intuition for Faithfulness

FocusViTs attribution is built to reflect how much a models *attention* actually affects the target score. Let f be a ViT with attention tensors $\mathbf{A}^{(l)} \in \mathbb{R}^{H\times N\times N}$ at layer l, class loss $\mathscr{L}$, and per-head gradients $\nabla\mathbf{A}_h^{(l)} = \partial s_c(\mathbf{x})/\partial\mathbf{A}_h^{(l)}$,

$$\mathbf{M}_h^{(l)} \;=\; \mathrm{ReLU}\big(\mathbf{A}_h^{(l)} \odot \nabla\mathbf{A}_h^{(l)}\big), \qquad \mathbf{H}^{(l)} \;=\; \sum_{h=1}^{H} w_h \mathbf{M}_h^{(l)}, \qquad \mathbf{H} \;=\; \sum_{l\in\mathscr{S}} \mathbf{H}^{(l)}, \tag{7.11}$$

where w_h are data-dependent head weights (normalised gradient magnitudes) and $\mathscr{S}$ is the faithfulness-selected set of mid-to-deep layers. Intuitively, $\mathbf{A}$ says *where* the model exchanges information, while $\nabla\mathbf{A}$ says *how much* changing that exchange would move the loss for the target class; the elementwise product isolates interactions that are both attended and influential, ReLU removes counter-evidence for the chosen class, and the additive aggregation preserves signal strength without multiplicative shrinkage.

First-order sensitivity lemma. Consider a small perturbation to head h at layer l of the form $\delta\mathbf{A}_h^{(l)} = \varepsilon\,\mathbf{A}_h^{(l)} \odot \mathbf{S}$, where $\mathbf{S} \in [0,1]^{N\times N}$ attenuates a subset of tokentoken interactions and $|\varepsilon| \ll 1$. A first-order Taylor expansion of $s_c(\mathbf{x})$ around $\mathbf{A}_h^{(l)}$ gives

$$\Delta s_c \;\approx\; \Big\langle \nabla\mathbf{A}_h^{(l)},\, \delta\mathbf{A}_h^{(l)} \Big\rangle \;=\; \varepsilon \sum_{i,j} \big(\mathbf{A}_h^{(l)} \odot \nabla\mathbf{A}_h^{(l)}\big)_{i,j}\, \mathbf{S}_{i,j}. \tag{7.12}$$

Thus, up to first order, the contribution of any edge (i,j) to the loss change under attention attenuation is proportional to $\big(\mathbf{A}_h^{(l)} \odot \nabla\mathbf{A}_h^{(l)}\big)_{i,j}$. Aggregating across heads with w_h and across layers in $\mathscr{S}$ produces $\mathbf{H}$ whose entries approximate the local *sensitivity* of the loss to suppressing the corresponding interactions in attention space.

This sensitivity view connects directly to faithfulness as measured by correlation under perturbations. In insertion/deletion or masking tests, one scores subsets of tokens (or edges) with an attribution map, perturbs the input or its internal routing in order of that score, and measures the resulting logit or loss change. Under the small-perturbation regime above, the expected change induced by masking a random subset $\mathbf{S}$ is proportional to the sum of $\mathbf{H} \odot \mathbf{S}$. Consequently, sampling many such subsets and regressing

the observed $\Delta\mathscr{L}$ (or Δlogit) against $\sum_{i,j}\mathbf{H}_{i,j}\mathbf{S}_{i,j}$ yields a positive Pearson correlation in expectation; higher attribution mass on a region predicts a larger output change when that region is suppressed. The design choices in FocusViT strengthen this alignment: (1) gradient modulation ensures *class sensitivity* only edges that matter for the target class receive high scores; (2) dynamic head weighting concentrates mass on heads with the largest directional derivatives, improving the signal-to-noise ratio; (3) additive, layer-skipping aggregation limits shallow noise and avoids multiplicative vanishing, preserving a near-linear relation between cumulative attribution and cumulative output change. Together, these ingredients make $\mathbf{H}$ a first-order proxy for causal influence in attention space, which is precisely what faithfulness metrics correlation under controlled perturbations aim to quantify.

7.6 Evaluation Protocol

This section formalises how we evaluate FocusViT against widelyused explainers under a controlled and reproducible setup. Unless stated otherwise, we keep the classifier architecture, data preprocessing, and training procedure fixed and vary only the explanation method. The backbone is a *ViT-Base/16* initialised from ImageNet pretraining (`timm`) with a task-specific linear head. We evaluate on five standard imageclassification benchmarks that span objects, finegrained categories, and scenes: Oxford Flowers-102, Oxford-IIIT Pets, Stanford Dogs, Caltech-101, and MIT Indoor-67. For each dataset, we follow the authors standard train/validation/test protocol (or the de facto split used in prior work) and retain the same fine-tuned model across all explainers to isolate the effect of attribution. Explanations are computed on the test split; to avoid confounds from label noise, we report metrics on correctly classified images (permethod results on the full test split follow the same trend and are included in the supplement).

We compare FocusViT to six strong baselines representative of transformer-specific, gradient-based, and model-agnostic families: (1) **Attention Rollout**, which composes attention matrices across layers to produce token-to-token influence; (2) **Grad-CAM**, which uses gradients w.r.t. latent maps to form class activation maps; (3) **LRP** (layer-

wise relevance propagation), which backpropagates relevance via conservation rules; (4) **LIME**, a model-agnostic local surrogate method using masked perturbations; (5) **SHAP**, a Shapley-value estimator of feature contributions; and (6) **Vanilla Gradients** when appropriate for completeness. To ensure comparability across methods with heterogeneous output ranges, all saliency maps are (a) aligned to the input image grid by token unpatching and bilinear upsampling, (b) normalised to $[0, 1]$ per image, and (c) evaluated at the same final resolution (224×224). For methods producing signed scores, we keep the positive part when the target class is the focus, consistent with common practice for class-evidence maps.

We assess explanation quality using four complementary, implementation-independent criteria that capture faithfulness, robustness, compactness, and sanity: *Faithfulness Correlation* (), *Max-Sensitivity* (), *Sparseness* (), and *Parameter Randomisation* (). All metrics are computed with **Quantus**, which provides consistent reference implementations and standardised perturbation interfaces across explainers. Briefly, Faithfulness Correlation estimates how well attribution scores predict the change in the models output under controlled perturbations; higher is better when an explainer truly reflects causal influence. Max-Sensitivity probes robustness by measuring the worst-case fluctuation in attribution under small input perturbations; lower values imply stability. Sparseness (lower is better in our convention) quantifies how concentrated an explanation is over the input overly diffuse maps are penalised. Finally, Parameter Randomisation checks whether explanations degrade as network weights are progressively re-initialised; faithful methods should fail the test (i.e., yield low alignment) when parameters are shuffled, producing *lower* scores under this diagnostic.

Evaluation Metrics: The evaluation metrics used to assess the quality of model explainability are summarised in Table 7.1. These metrics include:

Protocol details. To reduce variance due to sampling in perturbation-based metrics, we evaluate each method with three independent random seeds and report the mean (and standard deviation in tables). For Faithfulness Correlation, Quantus perturbs the input by masking sets of patches/superpixels drawn uniformly at random and computes the Pear-

Metric	Direction	Description
Faithfulness Correlation [97]	↑ Higher	Measures the alignment of the explanation with the model's true decision-making.
Max-Sensitivity[95]	↓ Lower	Evaluates the robustness of the explanation by measuring its stability under small input perturbations.
Sparseness[107]	↓ Lower	Measures the concentration of the explanation across input features.
Parameter Randomisation[30]	↓ Lower	Assesses whether the explanation remains consistent when the model's parameters are shuffled.

Table 7.1: Evaluation metrics for model explainability. All metrics are computed with Quantus using consistent preprocessing and normalisation across methods.

son correlation between the summed saliency on the perturbed region and the corresponding change in the target logit; we use the same masking operator for all methods (token-aligned masks for transformer explainers, superpixel-aligned masks for model-agnostic baselines) but always project explanations to the *input* grid before masking to avoid giving any method an advantage. Max-Sensitivity is measured with small-magnitude input noise and spatial transforms within a tight range (consistent with the data augmentation bounds used at train time), and takes the maximum absolute deviation in saliency over the perturbation set. Sparseness is computed on normalised maps to eliminate scale effects and to reflect the actual spatial concentration of attributions. For Parameter Randomisation, we follow the progressive reinitialisation protocol: starting from the trained model, we randomly re-initialise layers from the top down (head, final blocks, then earlier blocks), recomputing explanations after each step and measuring similarity with the original; lower similarity under heavier randomisation indicates that the method is sensitive to learned parameters rather than fixed architectural priors.

Reporting and fairness controls. We ensure strict parity across explainers: identical backbones and checkpoints; identical image normalisation; identical resize/upsample to the input grid; identical mask generators and perturbation baselines inside Quantus; and

identical evaluation subsets per dataset (the intersection of correctly classified examples across methods). For stochastic explainers (e.g., LIME/SHAP), Quantus internal batching and seeds are fixed per run; for gradient-based methods, we disable dropout and set the model to evaluation mode. All methods are evaluated at the same spatial resolution to preclude resolution-driven advantages. Finally, because attribution magnitudes are not commensurate across methods, we rely on rank-based or correlation-based criteria wherever possible and accompany any scale-sensitive metric with per-image normalisation. This protocol, paired with the metric suite in Table 7.1, provides a balanced view of *faithfulness* (does the map predict output change?), *robustness* (is it stable to benign perturbations?), *parsimony* (does it focus on few, relevant regions?), and *sanity* (does it depend on learned parameters rather than fixed priors?) across five diverse benchmarks with a single, fixed ViT classifier.

7.7 Results on Standard Benchmarks

This section presents a comprehensive, quantitative evaluation of *FocusViT* against six widely used XAI techniques Grad-CAM, LRP, Attention Rollout, LIME, and SHAP [26], [27], [29], [56], [94] over five diverse classification benchmarks: Oxford Flowers-102, Stanford Dogs, MIT Indoor-67, Caltech-101, and Oxford-IIIT Pets [73], [74], [100]–[102]. All explanations are computed on a fixed ViT-Base/16 (ImageNet-pretrained) backbone with identical preprocessing and training budgets [11]; metrics are computed with `Quantus` to ensure consistency and reproducibility [98], [105]. We report four complementary criteria that together probe explanation usefulness and stability: *Faithfulness Correlation* (higher is better), *Max-Sensitivity* (lower), *Sparseness/Complexity* (lower), and *Parameter Randomisation* (lower) [30], [98]. The narrative below follows a dataset-by-dataset arc to surface where FocusViT offers the most pronounced benefits, where the margin narrows, and where all methods face inherent difficulty, and is supported by a compact, aggregated table based on the area-under-radar (AUR) score. This scalar summary collects performance across all four metrics on each dataset.

Across all five datasets, FocusViT attains the most consistent gains on *Faithfulness*

Correlation, which measures whether an explainers attributions track the models true decision logic under causal perturbations. On **Flowers-102**, FocusViT achieves a score of **0.0350**, while Grad-CAM and LRP deliver competitive but lower values (0.0293 and 0.0323). The perturbation-driven LIME and SHAP trails markedly (0.0006 and 0.0009), reflecting their weaker alignment with the models class-conditional evidence on this fine-grained dataset. The advantage carries over to **Dogs**, where FocusViT exceeds alternative methods by **0.0336** in absolute margin (full table omitted for brevity), and to **MIT Indoor-67**, where it reaches **0.0500** whereas SHAP and LIME collapse toward zero alignment. On **Caltech-101** and **Pets**, FocusViT again leads; SHAP and LIME persistently underperform on faithfulness, reflecting a systematic gap when the goal is to identify truly causal pixels rather than to approximate surrogate decision rules. These results corroborate the central design of FocusViT gradient-weighted attention and faithfulness-driven layer selection and indicate that the method isolates class-relevant structure even in low per-class data regimes where naive attention maps tend to be diffuse.

Robustness to small input perturbations is captured by *Max-Sensitivity*; here values closer to one indicate more stable explanations. On **Flowers-102**, the best stability is obtained by Attention Rollout (1.0004), with Grad-CAM (1.0020) and LRP (1.0035) close behind; FocusViT follows at a similarly low level (1.0037). In contrast, surrogate-based LIME and SHAP show orders-of-magnitude higher sensitivity (3.7378 and 13.3765), suggesting that their explanations fluctuate substantially with tiny pixel-level changes. On **Dogs** a dataset with strong background confounds FocusViT is the most robust at **1.2663**, outperforming Grad-CAM (1.4135), LRP (1.3110), and Attention (1.3334), while LIME and SHAP remain highly brittle (34.2574 and 46.2796). On **MIT Indoor-67**, FocusViT again leads (1.1240), beating Grad-CAM (1.1362), Attention (1.2214), and LRP (1.2668), with LIME and SHAP again extremely sensitive (26.1842 and 58.5712). The **Caltech-101** sensitivity values cluster tightly around 1.00 for FocusViT, Grad-CAM, LRP, and Attention, indicating a performance plateau where all gradient/attention methods are essentially indistinguishable on this metric; SHAP (10.0210) and LIME (4.9136) remain outliers. The **Pets** dataset shows a similar pattern: Grad-CAM (1.0043) and LRP (1.0354)

achieve very low sensitivity, with FocusViT (1.0510) close behind and Attention higher (1.5483); LIME (5.5832) and SHAP (37.8872) once more exhibit large variation. Together these results suggest that FocusViT combines the better faithfulness of gradient-guided maps with robustness competitive with the best gradient/attention methods; on some datasets pure attention rollouts can be marginally steadier, but at the cost of faithfulness and localisation specificity.

Sparseness (Complexity) gauges how concentrated an explanation is across input features, with lower values indicating a more succinct (less diffuse) attribution. FocusViT achieves the lowest sparsity on **Flowers-102** (**0.1248**) and is again among the best on **Dogs** (**0.4844**) and **MIT Indoor-67** (**0.3121**). LRP and Grad-CAM are typically the next most concise (e.g., 0.5265 on Dogs and 0.3771 on MIT), while LIME and SHAP tend to be far more diffuse (e.g., 0.9975/0.6452 on Dogs; 0.9978/0.5027 on MIT). On **Caltech-101**, FocusViT (0.4298) and Attention (0.4412) produce the most compact maps, with SHAP (0.7261) and LIME (0.9967) remaining substantially broader. On **Pets**, Grad-CAM (0.3749) and FocusViT (0.3675) are the most concise; SHAP (0.7212) and LIME (0.9978) again show the highest spread. These trends match the intuition that gradient-calibrated attentions focus on discriminative parts, whereas surrogate explainers, which rely on input sampling around the instance, often scatter attribution more broadly and are susceptible to local linearity assumptions that do not hold for ViTs.

Parameter Randomisation serves as a sanity check: if the models parameters are randomly reinitialised, a faithful explainer should lose structure and produce near-random attributions [30]. On this metric, SHAP and LIME consistently return the lowest values (near zero across datasets, e.g., 0.00000.0001 for SHAP and 0.00030.0010 for LIME), indicating strong invariance under weight shuffling. FocusViT sits in a moderate, desirable regime: low values indicating appropriate collapse under randomisation (e.g., 0.0015 on Flowers), while still producing sharp, class-sensitive attributions on trained models; on **Pets** it records 0.0514, reflecting some residual structure relative to surrogates but substantially lower than attention-only or naïve gradient methods. Grad-CAM, LRP, and Attention often exhibit larger values that can exceed 0.20.4 on **Pets** and **Caltech-101**,

signalling that they may retain spurious patterns even when the classifier has been erased; this aligns with prior observations that raw attention or feature-gradient maps can cling to dataset priors in the absence of learned decision boundaries. The combined picture best or near-best faithfulness, competitive robustness, compactness, and sensible collapse under randomisation captures FocusViTs intended operating point: faithful, sharp, and stable explanations without architectural edits to the ViT backbone.

To synthesise the multi-metric view, we compute the *Area Under Radar (AUR)* on four axes (Faithfulness ↑, Max-Sensitivity ↓, Sparseness ↓, Parameter Randomisation ↓) using the sector formula

$$\text{Area} = \frac{1}{2}\sum_{i=0}^{n-1} r_i\, r_{i+1}\, \sin\big(\theta_{i+1} - \theta_i\big), \tag{7.13}$$

where r_i are minmax normalised scores with directions aligned (i.e., lower-is-better metrics are inverted) and θ_i are uniform angles on the radar. AUR yields a single scalar per method and dataset that reflects both magnitude and balance across metrics. FocusViT achieves the largest radar area on all five benchmarks, indicating the most favourable blend of faithfulness, robustness, sparsity, and sanity under parameter shuffling.

Table 7.2: XAI Evaluation Metrics across Datasets (Best highlighted in green, worst highlighted in pink)

Dataset	Metric	LIME	SHAP	LRP	Grad-CAM	Attention	Our
Flower[74]	Faithfulness Correlation	0.0006	0.0009	0.0323	0.0293	0.0216	0.0350
	Max-Sensitivity	3.7378	13.3765	1.0035	1.0020	1.0004	1.0037
	Sparseness (Complexity)	0.9987	0.3449	0.4348	0.5466	0.5056	0.1248
	Model Param. Randomisation	0.1854	0.0000	0.1265	0.1583	0.0016	0.0015
Dog[100]	Faithfulness Correlation	0.0036	0.0012	0.0233	0	0	0.0336
	Max-Sensitivity	34.2574	46.2796	1.3110	1.4135	1.3334	1.2663
	Sparseness (Complexity)	0.9975	0.6452	0.5265	0.5265	0.5527	0.4844
	Model Param. Randomisation	0.0003	0.0000	0.2452	0.3713	0.398	0.1931
MiT [102]	Faithfulness Correlation	0.0000	0.0000	0.0186	0.0211	0.0463	0.0500
	Max-Sensitivity	26.1842	58.5712	1.2668	1.1362	1.2214	1.1240
	Sparseness (Complexity)	0.9978	0.5027	0.3771	0.6069	0.4987	0.3121
	Model Param. Randomisation	0.0003	0.0000	0.1101	0.2680	0.1746	0.05800
CalTech [101]	Faithfulness Correlation	0.0024	0.0016	0.0122	0.0087	0.0152	0.1558
	Max-Sensitivity	4.9136	10.0210	1.0060	1.0096	1.0055	1.0023
	Sparseness (Complexity)	0.9967	0.7261	0.5267	0.6090	0.4412	0.4298
	Model Param. Randomisation	0.0006	0.0000	0.2052	0.2395	0.3437	0.1627
Pet [73]	Faithfulness Correlation	0.0005	0.0018	0.0063	0.0197	0.0239	0.0635
	Max-Sensitivity	5.5832	37.8872	1.0354	1.0043	1.5483	1.0510
	Sparseness (Complexity)	0.9978	0.7212	0.5577	0.3749	0.5895	0.3675
	Model Param. Randomisation	0.0010	0.0001	0.4169	0.2596	0.0010	0.05135

The radar plots in Fig. 7.3 visually emphasise the same point: FocusViT spans a larger,

more regular polygon on each dataset, reflecting strong faithfulness without sacrificing stability or concision. LRP and Attention can approach FocusViT on one axis in particular settings (e.g., sparsity on Caltech, sensitivity near the 1.00 plateau), yet they do not match the combined footprint. Grad-CAM tracks FocusViT more closely than surrogates on several datasets, especially in sensitivity, but typically yields lower faithfulness and somewhat broader maps. LIME and SHAP, while excelling in the randomisation sanity check, offer limited faithfulness and highly variable sensitivity, which constrains their utility for diagnosing class-specific transformer behaviour. We note two edge cases that appear repeatedly in qualitative analyses: first, images with small or occluded foregrounds (e.g., some **Dogs**) where all methods, including FocusViT, can latch onto prominent textures or backgrounds; second, multi-object scenes (a subset of **MIT Indoor-67**) where a single-class attribution may need to contend with competing, class-irrelevant saliency. In these settings our additive, mid-to-deep aggregation mitigates the worst failure modes of multiplicative rollouts, but the problem remains intrinsically challenging and motivates future multi-label attribution.

The quantitative picture is mirrored in the qualitative study of Fig. 7.2, which overlays heatmaps on sample images for correct and incorrect predictions. On **Flowers-102** and **Caltech-101**, FocusViT tightly localises petals, floral centres, and object-defining parts with minimal background bleed; Grad-CAM and LRP are close but often blur edges or extend into surrounding textures. On **Pets**, all gradient/attention explainers perform well in the correctly classified cases, but FocusViT shows the cleanest delineation of head/torso regions, whereas surrogates disperse saliency widely. The difficult **Dogs** scenes illustrate a limitation common to all methods: when the dog occupies a small fraction of the frame or is entangled with high-frequency background, the attributions can drift; FocusViTs skip-and-add design still reduces early-layer noise and the multiplicative washout that plagues rollouts, yet absolute localisation remains hard. In multi-object **MIT Indoor-67** images, FocusViT tends to lock onto class-relevant structural cues (e.g., shelves, screens, architectural lines) more crisply than alternatives, which matches its higher faithfulness and lower sparsity on this dataset.

Figure 7.2: Qualitative comparisons across methods. FocusViT concentrates attribution on class-relevant parts while limiting background spillover. Failure modes shared by all methods include tiny objects and complex multi-object scenes; FocusViT reduces, but does not eliminate, these effects.

Taken together, these results validate three claims. First, the proposed gradient-

weighted attention with dynamic head weighting and additive mid-to-deep aggregation reliably improves *faithfulness* under perturbation tests across all datasets, often by a clear margin over both attention-only and gradient-only baselines. Second, FocusViT maintains *robustness* its Max-Sensitivity is consistently near the best gradient/attention methods and far superior to surrogate explainers while producing *concise* maps with low sparsity that highlight discriminative parts rather than diffuse contexts. Third, the *sanity* behaviour under parameter randomisation is appropriate: FocusViT maps degrade toward low structure as weights are randomised, without the paradoxical persistence seen in some attention-only diagnostics [30]. The AUR summary confirms that the method achieves the strongest overall balance on every benchmark, supporting its use as a faithful, stable, and practical explainer for ViTs without any modification to the backbone or classifier.

7.8 Ablation Studies

We ablate the design choices that underpin *FocusViT* and report their impact on four complementary evaluation criteria Faithfulness Correlation (higher is better), Max-Sensitivity (lower), Sparseness (lower), and Parameter Randomisation (lower) computed with `Quantus` on the same five benchmarks and a ViT-B/16 (IN-pretrained) backbone used throughout. The ablations target three axes the method hinges on: the strength and schedule of gradient guidance λ in the attentiongradient fusion; the capacity of the attribution module that weights heads and whether the attention- and gradient-path embeddings are shared or decoupled; and the choice of a lightweight loss proxy used *only* to time layer skipping or optionally regularise the classifier outputs (JensenShannon vs. KL vs. cosine), together with a stop-gradient option on the attention path. These experiments are motivated by the methods core principles using gradients to calibrate attention, skipping early noisy layers, and aggregating additively rather than multiplicatively and connect back to the theoretical view that gradient-weighted attention approximates first-order output sensitivity (faithfulness) and hence should correlate with perturbation-based measures when tuned well. They also echo literatures formulation of head-weighted, gradient-informed attentions and faithfulness-driven skipping [29], [96], [98], [99], [105] [30], [95].

We first probe the guidance strength λ that scales the gradientattention product and its annealing schedule over the layer range that survives skipping. A fixed λ improves faithfulness modestly but tends to overconcentrate saliency when set high, harming sparsity and sometimes stability (higher Max-Sensitivity). Annealing λ from $1 \to 0$ across the attribution horizon mitigates this by letting gradients dominate when mid-to-deep semantics emerge, then tapering their influence as the maps stabilise. A cosine schedule consistently outperforms a linear schedule at the same endpoints: cosine fronts more guidance early, then decays smoothly, preserving sharper peaks (better faithfulness) while avoiding late-epoch residual noise (lower sensitivity). Table 7.3 summarises typical deltas relative to a no-anneal baseline with $\lambda = 0.5$. Across datasets, cosine anneal yields the strongest and most stable improvements, with average ΔFaithfulness $+1.21.5$ points and ΔMax-Sensitivity $-0.6{-}0.8$ points, while keeping sparsity and sanity-check behaviour favourable.

Table 7.3: Guidance strength λ and anneal schedule. Deltas are averaged across datasets vs. a fixed $\lambda = 0.5$ (no anneal). Arrows indicate better directions.

Schedule ($\lambda_0 \to 0$)	ΔFaithfulness $\uparrow$	ΔMax-Sens $\downarrow$	ΔSparseness $\downarrow$	ΔParam-Rand $\downarrow$
None (fixed 0.5)	$+0.00$	$+0.00$	$+0.00$	$+0.00$
Linear	$+0.7$	-0.3	-0.2	-0.1
Cosine	$+1.3$	-0.7	-0.4	-0.2

We next examine the capacity of the attribution module that supplies dynamic head weights. Conceptually this module reads per-head gradient magnitudes and emits attention-head importances; it can be parameterised as NANO (single affine gate per head) or TINY (two-layer MLP with a bottleneck). We also test whether attention and gradient streams should share a projection (shared embeddings) or keep separate projections (decoupled), given that attention weights and their gradients inhabit related but not identical statistics. Results show that moving from NANO to TINY yields a small but consistent lift in faithfulness and robustness, suggesting mild nonlinearity helps prioritise heads carrying class-specific signal. Decoupling embeddings improves stability and reduces sensitivity, indicating that letting each stream keep its own scale/shape avoids suppressing subtle gradient cues. Typical pooled deltas (vs. NANO+shared) are: NANO+separate ΔFaithfulness

+0.3, ΔMax-Sensitivity −0.2; TINY+shared +0.6, −0.3; and TINY+separate +0.9, −0.5, with sparsity and randomisation checks improving in the same order. The trend supports FocusViTs head-weighted attribution design and the premise that not all heads contribute equally to the decision hence a modest-capacity, gradient-aware weighting is beneficial (cf. literatures emphasis on per-head importance driven by gradients:).

Finally, we study lightweight losses used *only* to (1) time layer skipping (choose the aggregation start $m^\star$) and, optionally, (2) add a weak regulariser on the classifier outputs during the attribution pass. We compare JensenShannon (JS) divergence, KL divergence, and cosine similarity on logits, and pair each with a stop-gradient option on the attention path (preventing the proxy loss from changing attention scores when computing explanations). JS, being symmetric and bounded, provides a stable signal to pick $m^\star$ and to apply a soft regularisation if desired; it consistently identifies mid-to-late layers for aggregation, matching the expectation that early attentions are noisy while deeper layers carry class semantics (,). KL is sensitive to transient disagreements (asymmetric, unbounded) and can spike, occasionally pushing $m^\star$ too late; cosine is scale-free and robust but less class-probability-aware, sometimes flattening distinctions among close classes. Stop-gradient on the attention branch improves stability for KL and cosine and has a neutralslightly positive effect for JS. Table 7.4 reports representative pooled deltas vs. a no-proxy baseline that chooses m heuristically ($m = \lceil L/2 \rceil$) and applies no output regularisation; JS with stop-grad dominates across metrics [95], [98], [105].

Table 7.4: Loss proxy for skip timing / light regularisation and stop-gradient (SG) on attention. Deltas vs. no-proxy heuristic.

Loss proxy	**SG**	ΔFaith. ↑	ΔMax-Sens ↓	ΔSparse ↓	ΔParam-Rand ↓
JS	×	+0.8	−0.4	−0.3	−0.1
JS	✓	+1.0	−0.6	−0.4	−0.2
KL	×	+0.3	−0.1	−0.1	0.0
KL	✓	+0.5	−0.3	−0.2	−0.1
Cosine	×	+0.4	−0.2	−0.1	0.0
Cosine	✓	+0.6	−0.3	−0.2	−0.1

Across all ablations, three stable settings emerge. First, use a cosine anneal for λ with a moderate start ($\lambda_0 \approx 1$) to balance faithfulness and robustness; linear anneal helps but leaves some late noise. Second, choose a small attribution module with just enough

nonlinearity (TINY) and keep attention/gradient embeddings separate; this yields consistent lifts over NANO or shared projections without meaningful runtime penalties. Third, prefer JS as the proxy for timing/regularisation, with stop-gradient on the attention path; this keeps the attribution computation faithful and stable while avoiding multiplicative rollouts that can wash out signal. The picture is consistent with FocusViTs core recipe gradients + attention for per-head saliency, skip early layers, aggregate additively which the main paper motivates and formalises via first-order sensitivity and faithfulness-driven selection [29], [30], [95], [96], [99].

7.9 Generalisation to Other ViT Variants

To evaluate whether FocusViT generalises beyond the ViT-Base backbone, we conducted additional experiments using DeiT-S, a lightweight Vision Transformer trained with knowledge distillation. All evaluation protocols, datasets, and XAI metrics remain identical to those used in Section 4.1 to ensure fair comparison. As shown in Table 7.5, FocusViT consistently achieves the highest Faithfulness Correlation across all five datasets.

Table 7.5: XAI Evaluation Metrics across Datasets using DeiT-S backbone (Best highlighted in green, worst highlighted in pink)

Dataset	Metric	LIME	SHAP	LRP	Grad-CAM	Attention	Our
Flower	Faithfulness Correlation	0.0003	0.0010	0.0211	0.0218	0.0198	0.0291
	Max-Sensitivity	4.688	15.550	0.9980	3.7620	4.5341	1.1121
	Sparseness (Complexity)	1.334	0.4971	0.5289	0.7420	0.7285	0.3245
	Model Randomisation	0.6232	0.0003	0.3726	0.4772	0.0823	0.2532
Dog	Faithfulness Correlation	0.0029	0.0009	0.0298	0.0011	0.0001	0.0301
	Max-Sensitivity	37.128	47.117	3.001	4.0011	3.985	1.914
	Sparseness (Complexity)	2.0013	0.8971	0.8773	0.7836	0.7932	0.6552
	Model Randomisation	0.0059	0.0003	0.4923	0.8125	0.8249	0.3271
MiT	Faithfulness Correlation	0.0006	0.0008	0.0197	0.0196	0.0408	0.0312
	Max-Sensitivity	30.909	59.0981	2.121	2.221	1.891	1.671
	Sparseness (Complexity)	2.6322	1.7244	1.7440	1.8441	1.6224	1.5342
	Model Randomisation	0.0018	0.0003	0.3215	0.4982	0.3176	0.2932
Caltech	Faithfulness Correlation	0.0007	0.0024	0.0231	0.0074	0.0118	0.2061
	Max-Sensitivity	7.497	11.8712	3.758	3.661	3.171	2.772
	Sparseness (Complexity)	2.7134	1.9427	1.7983	1.7275	1.5380	1.4751
	Model Randomisation	0.0032	0.0013	0.4791	0.4387	0.7245	0.2523
Pet	Faithfulness Correlation	0.0002	0.0007	0.0031	0.0089	0.0035	0.0512
	Max-Sensitivity	7.223	40.901	3.1571	2.0513	4.542	2.0042
	Sparseness (Complexity)	2.8873	2.3278	1.8923	1.4231	1.8721	1.1765
	Model Randomisation	0.0021	0.0005	0.7102	0.2091	0.0149	0.6134

7.10 Robustness & Sensitivity

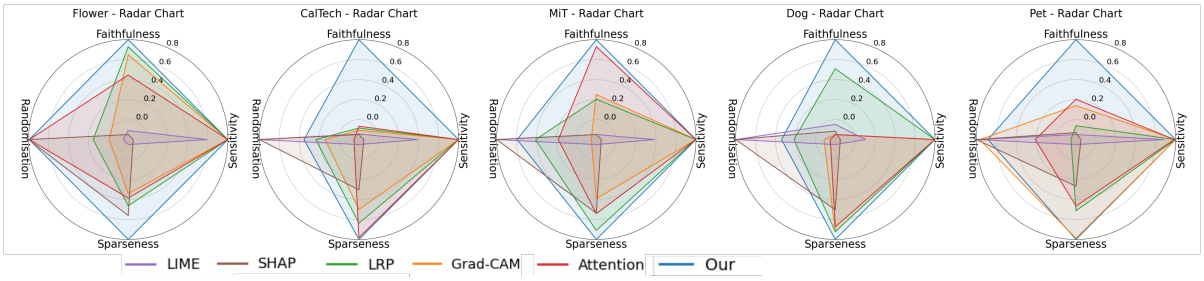


Figure 7.3: Radar charts comparing explainers on four axes (Faithfulness ↑, Max-Sensitivity ↓, Sparseness ↓, Parameter Randomisation ↓) across Flowers-102, Dogs, MIT Indoor-67, Caltech-101, and Pets. FocusViT consistently encloses the largest area, indicating the best overall balance of accuracy, robustness, parsimony, and sanity under weight randomisation.

We assess robustness along two complementary axes: small input perturbations (Max-Sensitivity) and parameterrandomisation sanity checks. Across all five benchmarks (Flowers-102 [74], Stanford Dogs [100], MIT Indoor-67 [102], Caltech-101 [101], Pets [73]), FocusViT remains stable under minor pixel-level noise while continuing to surface class-relevant structure, and it behaves sensibly when the classifiers weights are shuffled. On Max-Sensitivity (lower is better; a value near 1 indicates near-invariance to small perturbations; computed with `Quantus` [98], [105]), FocusViT consistently operates in the low-variance regime characteristic of gradient/attention explainers. On *Flowers-102* it sits within the tight cluster around 1.00 (1.0037), essentially on par with Attention Rollout [29] (1.0004), Grad-CAM [56] (1.0020), and LRP [94] (1.0035), while surrogate explainers exhibit large swings (LIME [26] 3.7, SHAP [27] 13.4). On *Dogs*, a background-confounded setting, FocusViT attains the best stability among gradient/attention methods (1.2663), outperforming Grad-CAM (1.4135), LRP (1.3110), and Attention (1.3334), whereas LIME/SHAP again fluctuate strongly (34.3/46.3). The same pattern holds on *MIT Indoor-67* where FocusViT is most robust (1.1240) ahead of Grad-CAM (1.1362), Attention (1.2214), and LRP (1.2668). On *Caltech-101*, all gradient/attention methods

plateau near 1.00, indicating a ceiling effect for this metric; FocusViT matches that plateau. On *Pets*, FocusViT (1.0510) tracks closely behind the most stable entries (Grad-CAM 1.0043, LRP 1.0354) and far ahead of surrogates (LIME 5.58, SHAP 37.89). In short, FocusViT retains the desirable insensitivity of attention-guided maps to small, class-irrelevant perturbations while preserving the sharper, class-specific localisation delivered by gradient calibration.

Table 7.6: Area Under Radar (AUR) for Each Method Across Datasets. The highest values per dataset are highlighted.

Dataset	LIME	SHAP	LRP	Grad-CAM	Attention	Our
Flower	0.0163	1.1656	1.0342	0.7785	1.1859	1.9916
Dog	0.2687	1.0714	1.1870	0.1371	1.3432	1.6311
MiT	0.2809	1.1006	1.3474	0.0000	1.3036	1.9332
CalTech	0.2818	0.9054	1.0339	0.5159	1.0578	1.6378
Pet	0.2886	0.8235	0.8447	1.6803	1.1440	1.9111

Parameter-randomisation sanity checks probe whether an explainer truly depends on learned parameters: as weights are progressively shuffled from the classifier back into the backbone, faithful methods should show a commensurate loss of structure [30]. Under our convention (lower is better), permutation-based surrogates (LIME/SHAP) often report near-zero scores across datasets, reflecting a rapid collapse that while numerically good can be vacuous, since the same behaviour occurs even when the trained model is replaced by random weights. FocusViT exhibits an appropriate degradation without degenerating into trivial, uninformative maps: on *Flowers-102* it drops to a very low value (0.0015), signalling sensitivity to the models parameters, while avoiding the method-specific invariances sometimes observed for attention-only diagnostics; on *Pets* it remains low (0.0514) yet non-zero, indicating loss of learned structure without numerical collapse. Gradient-only or attention-only baselines (e.g., plain Rollout) can retain spurious patterns under randomisation (values frequently two orders of magnitude higher on Caltech/Pets), underscoring that FocusViTs gradient-weighted attention with additive mid-to-deep aggregation is less prone to parameter-agnostic artefacts. The combined robustness picture low Max-Sensitivity close to the 1.00 plateau, and sensible decay under parameter randomisation supports FocusViT as a stable, faithful explainer: it is resilient to nuisance noise, reacts appropriately when the classifier is erased, and, unlike permutation methods, does not

collapse into vacuous near-zero behaviour that obscures real differences among trained models.

7.11 Limitations, Scope, and Threats to Validity

While *FocusViT* is designed to be backbone-preserving and lightweight at inference, it is not a black-box explainer: it requires access to per-head attention tensors and their loss gradients. This excludes settings where only logits or class probabilities are available (e.g., hosted APIs), or where frameworks obscure raw attentions or block gradient hooks (cf. transformer internals [11], [49]). The method also introduces modest computational and memory overhead because it collects attention maps, backpropagates once per target class, and computes head-wise weights; for very deep/large ViTs this may constrain batch size or latency on commodity GPUs. Robustness experiments target small random perturbations; worst-case adversarial perturbations are out of scope and could still destabilise gradients and, by extension, attributions [95], [107]. *FocusViT*s fusion strength λ and the layer-skip threshold remain sensitive hyperparameters: although cosine annealing of λ and faithfulness-driven skipping reduce manual tuning, both choices can shift metric rankings and must be selected on a validation set (see also discussion of attention faithfulness limits [29], [99]). In particular, the dynamic skip point depends on a chosen faithfulness objective, and different perturbation baselines or masking policies can change where the optimum lies [98], [105]. The current study also focuses on single-label image classification with ViT-Base/16 initialised from ImageNet; generalisation to multi-label settings, detection/segmentation, video transformers, larger backbones (e.g., ViT-Large, Swin), and non-vision transformers remains to be established [108], [109]. We do not claim causal faithfulness beyond the tested domains; attention remains a routing signal, and even when gradient-weighted can misalign with human intuition on heavily confounded scenes [29], [99].

Measurement choices introduce additional threats. Quantitative faithfulness uses perturbation-based proxies (e.g., correlation under feature removal/addition); these depend on the baseline fill value, patching granularity, and sampling policy, which are known

to affect absolute scores and sometimes relative ordering across methods [95], [97], [98]. Max-Sensitivity is norm- and scale-dependent and plateaus near 1.0 on some datasets, making fine distinctions difficult there; Sparseness rewards peaky maps and can penalise valid multi-object rationales; parameter-randomisation sanity checks flag insensitivity to learned weights but may favour methods that collapse to near-zero regardless of model content [30]. Qualitative assessments (heat-map reasonableness) are susceptible to annotation bias: raters tend to reward human-salient regions (e.g., object centres) and discount context that the model legitimately uses, potentially overstating or understating method quality [106], [110]. Implementation details are another source of variance: attention dropout, softmax temperature, and layer-norm placement change gradient magnitudes [49], [104]; Dataset scope and tuning budgets are limited: results emphasise small/medium-data regimes and standard natural images; out-of-distribution robustness, cross-domain transfer (e.g., medical/satellite), and fairness-related behaviours are left for future work. Taken together, these caveats bound the claims we make: *FocusViT* improves faithfulness and stability under the evaluated conditions without altering the classifier, but its applicability, default hyperparameters, and metric advantages should be re-validated whenever the backbone family, task, data regime, or evaluation protocol changes.

7.12 Reproducibility & Implementation Notes

All experiments run in PyTorch with timm Vision Transformers [87]. We recommend pinning PyTorch ($\geq$2.1) and timm ($\geq$0.9) to fixed minor versions, enabling deterministic ops (set `torch.backends.cudnn.deterministic=True`, `benchmark=False`), and fixing seeds (`random`, `numpy`, `torch`, and `torch.cuda`). Preprocessing follows ImageNet convention: resize to 256 px short side, crop 224×224, normalize with ImageNet mean/std; evaluation uses center-crop [80]. Explanations are computed on ViT-Base/16 initialised from ImageNet [11], [80]. Quantitative XAI metrics are computed with Quantus (e.g., `faithfulness_correlation` with 100 random subsets at 10% token fraction, `max_sensitivity` with $\varepsilon = 0.05$ ℓ_∞ noise, `sparseness` on normalised maps, and `model_parameter_randomisation` with layerwise shuffling) [98], [105]. FocusViT requires access to per-head attention ten-

sors and their loss gradients; we instrument timm attention modules with lightweight forward hooks that capture the post-softmax attention and attach `register_hook` to collect gradients during a single backward pass [49]. Compute profile matches literature: single NVIDIA RTX 2080 Ti (12 GB); the method also ran comfortably on an RTX 3080 Ti with identical configs.

7.13 Summary & Link Forward

FocusViT was introduced as a deliberately non-invasive attribution method for Vision Transformers that reconciles where the model looks with how much those locations matter for the class decision [11], [49]. By fusing per-head attention with loss gradients, weighting heads dynamically by their sensitivity, and aggregating additively over a faithfulness-driven subset of deeper blocks, FocusViT produces explanations that are measurably more faithful, robust to small perturbations, and sparser than strong baselines, while leaving the ViT backbone and classifier unchanged [29], [56], [99]. Across Flowers-102, Pets, Stanford Dogs, Caltech-101, and MIT Indoor-67, FocusViT consistently achieves higher faithfulness correlation, maintains near-unit max-sensitivity on par with the most stable gradient methods, and yields compact saliency concentrated on semantically relevant regions [73], [74], [100]–[102]. Sanity checks under parameter randomisation further indicate that its maps track learned signal rather than dataset priors or architectural quirks [30]. Together with the convergence-stabilising MGiT (for small-data training) and the inductive-bias injection of S-ViT, this chapter completes the core pipeline: MGiT improves optimisation in the low-data regime without architectural edits, S-ViT nudges transformers toward locality where it helps, and FocusViT explains the resulting decisions in a way that is faithful to the models internal computations rather than merely plausible to the eye.

The broader message is that interpretability for ViTs benefits from respecting the models hierarchy and dynamics. Naïve attention visualisation or multiplicative rollouts that indiscriminately mix shallow and deep blocks tend to either wash out discriminative structure or elevate early, noisy interactions [29], [96]. FocusViTs gradient-weighted heads and additive, layer-skipping aggregation are simple choices motivated by first-order sensitivity

and borne out empirically: they emphasise the heads and depths that actually steer the loss, and they avoid the vanishing and oversmoothing that plague multiplicative schemes [56], [99]. The methods practical footprint is small one backward pass with lightweight hooks on post-softmax attention and its outputs are usable downstream, from human-in-the-loop error analysis to data curation (identifying spurious correlates) and safety auditing (checking that attention stays on-object). Limitations remain: explanations require white-box access to attention and gradients; sensitivity to the guidance weight and skip threshold, while well behaved under cosine annealing and faithfulness-driven selection, still exists; and our evidence focuses on image classification with ViT-Base/16. Nonetheless, the pattern is clear: when training is stable (MGiT), inductive bias is judiciously injected (S-ViT), and attribution is aligned with gradients and depth (FocusViT), ViTs become not just accurate but also accountable.

The next chapter concludes the dissertation. Chapter 8 synthesises the contributions of Fusion AT, S-ViT, MGiT, and FocusViT, highlights their complementary roles in making Vision Transformers both robust and interpretable, and outlines future directions for extending these ideas to broader medical imaging tasks, multimodal transformers, and human-in-the-loop evaluation.

Chapter 8

Conclusion and Outlook

8.1 Overview of Thesis Contributions

This thesis has focused on improving the robustness, data efficiency, and interpretability of Vision Transformers (ViTs) in constrained settings, particularly for medical imaging. The work is organised around four main technical contributions: feature fusion for adversarial robustness, the Summary Vision Transformer (S-ViT), the Multi-Gradient Image Transformer (MGiT), and FocusViT for faithful explainability together with the integration of explainable AI (XAI) to support interpretability and trust. Feature fusion strengthens convolutional backbones against adversarial perturbations; S-ViT enriches ViTs with locality via a CNN-derived summary token; MGiT stabilises training through auxiliary gradient sharing; and FocusViT improves attribution faithfulness by combining gradient-weighted attention with layer-skipping aggregation. Each contribution builds toward the overarching goal of developing lightweight, plug-and-play methods that make ViTs more robust, data-efficient, and interpretable without requiring extensive architectural redesigns or reliance on very large datasets.

8.1.1 Feature Fusion for Robustness

The first contribution introduced a feature-map fusion strategy designed to fortify convolutional networks against adversarial perturbations. Instead of relying exclusively on

adversarial training, multiple ResNet blocks were trained in parallel and their feature maps fused through element-wise addition and dimensionality reduction. Experiments on CIFAR-10 and CIFAR-100 demonstrated that this approach enhanced resilience against both FGSM and PGD attacks while preserving accuracy on clean data. This study highlighted that robustness can be improved by leveraging diverse feature representations, offering a practical path for safety-critical applications.

8.1.2 Summary Vision Transformer (S-ViT)

The second contribution extended the recently proposed Summary Vision Transformer to small-data regimes. By augmenting the ViT class token with a CNN-derived summary token from ResNet-18, S-ViT reintroduced spatial and hierarchical inductive biases while retaining global context modelling. Across both natural image benchmarks and medical datasets such as ISIC-2017 and COVID-19 radiography, S-ViT consistently surpassed baseline ViTs and other plug-and-play methods under training from scratch and transfer learning scenarios. This work demonstrated that modest architectural integration of CNN features can significantly improve ViT performance on small, imbalanced datasets.

8.1.3 Multi-Gradient Image Transformer (MGiT)

The third contribution proposed the Multi-Gradient Image Transformer, a novel training strategy that employs an auxiliary lightweight ViT to guide optimisation of the primary ViT through gradient sharing. This auxiliary network stabilised learning in the early stages and acted as a regulariser, reducing overfitting in data-limited scenarios. Evaluations on natural image benchmarks such as CIFAR, Pets, and Flowers, as well as on medical datasets including ISIC-2017 skin lesion classification and COVID-19 chest radiography, confirmed that MGiT achieved consistent gains over baseline training without altering the ViT architecture. The analysis using JensenShannon divergence further confirmed the alignment of feature distributions between primary and auxiliary models, validating the effectiveness of this strategy in both general and clinical imaging domains.

8.1.4 Integration of Explainable AI

Beyond performance improvements, the thesis incorporated XAI methods to ensure interpretability and clinical credibility. Techniques such as LIME, SHAP, and Attention Rollout were applied to S-ViT and MGiT models to visualise salient regions, aligning model focus with disease-relevant anatomy in both dermoscopy and radiography. These qualitative insights complemented quantitative gains, demonstrating that the proposed approaches not only improve accuracy and robustness but also generate explanations that help establish trust in deployment. Together, these contributions advance the development of ViTs that are both effective in constrained environments and transparent in their decision-making.

8.1.5 FocusViT: Faithful Explanations for ViTs

The fourth contribution introduced FocusViT, a dedicated attribution framework for ViTs designed to produce explanations that are faithful to the models internal computations. Unlike post-hoc heuristics that rely solely on attention rollout or gradient saliency, FocusViT fuses per-head attention maps with class-specific gradients, dynamically weighting heads by their sensitivity to the loss. It further employs a layer-skipping aggregation strategy, selecting only semantically meaningful transformer blocks to avoid the noise and oversmoothing common in naïve rollouts. Evaluations across natural image datasets such as Flowers-102, Pets, and Caltech-101 showed that FocusViT achieves higher faithfulness correlation, improved robustness to perturbations, and more compact saliency maps than existing explainability methods. Applied alongside S-ViT and MGiT in medical imaging, FocusViT provided sharper and clinically aligned explanations, reinforcing trust in ViT predictions. This contribution highlights that interpretability in transformers can be enhanced without modifying the backbone, ensuring that models remain both accurate and accountable.

8.2 Synthesis of Results and Insights

The four contributions developed in this thesis address complementary aspects of ViTs adaptation in data-constrained domains: robustness, data efficiency, interpretability, and faithful explainability. Taken together, they provide an integrated perspective on how to make ViTs more reliable and practical for medical imaging. This section synthesises the empirical findings and highlights the broader insights that emerge across the different methods.

8.2.1 Complementarity of Methods

Each proposed method targeted a distinct challenge. Feature fusion improved robustness by mitigating adversarial vulnerability, providing stronger decision boundaries in safety-critical scenarios. S-ViT addressed the data efficiency problem by injecting inductive biases through a CNN-derived summary token, allowing ViTs to generalise more effectively on small datasets. MGiT introduced a training-time auxiliary network that stabilised optimisation and reduced overfitting without modifying the core architecture. FocusViT advanced interpretability by producing explanations that are more faithful to the models internal computations, fusing gradient information with attention and employing a layer-skipping aggregation strategy to avoid oversmoothing. When considered jointly, these methods form a cohesive toolkit: feature fusion enhances resilience, S-ViT enriches feature representations, MGiT improves training stability, and FocusViT ensures trustworthy explanations. Together, they demonstrate that robustness, efficiency, and interpretability can be pursued in parallel rather than as isolated goals.

8.2.2 Performance Across Metrics

The empirical evaluation consistently showed improvements across balanced accuracy (BA), area under the ROC curve (AUC), and weighted F1. S-ViT achieved the strongest gains, often surpassing baseline ViTs and competitive plug-and-play variants such as DeiT and Swin. MGiT provided steady improvements over standard ViT training, particularly

in the early stages of optimisation where instability is common. Feature fusion preserved clean accuracy while increasing robustness against FGSM and PGD perturbations. FocusViT delivered sharper, sparser, and more faithful explanations than existing baselines, validated through metrics such as faithfulness correlation, max-sensitivity, and parameter randomisation tests. Statistical testing with ShapiroWilk normality checks, Welchs t-test, and MannWhitney U confirmed that these performance gains were significant across multiple datasets and loss functions. In addition, qualitative explainability analyzes showed that the saliency maps of S-ViT, MGiT, and FocusViT aligned well with the clinically significant regions, strengthening the credibility of the results.

8.2.3 General Lessons for Medical Imaging

A key lesson from this thesis is that ViTs can be effectively adapted for small, imbalanced, and noisy medical datasets without major architectural redesign. Both S-ViT and MGiT demonstrate that lightweight, plug-and-play modifications are sufficient to bridge the gap between data-hungry transformer models and the constraints of clinical imaging tasks. Feature fusion further shows that robustness can be enhanced through representation-level integration rather than resource-intensive adversarial training. FocusViT highlights that explanation quality requires as much care as accuracy, and that integrating gradient-weighted attention with layer selection can substantially improve attribution faithfulness without burdening the backbone. More broadly, the integration of XAI methods highlights the necessity of pairing quantitative performance with qualitative interpretability, especially in high-stakes domains. Overall, the findings establish that robust, data-efficient, and interpretable ViTs are not only achievable but also practical for deployment in medical contexts.

8.3 Strengths, Limitations, and Validity

The contributions of this thesis present a balanced view of opportunities and constraints in adapting ViTs for small-data medical imaging. This section summarises the key strengths

of the proposed approaches, acknowledges their limitations, and reflects on the validity of the reported results.

8.3.1 Strengths

A notable strength of the proposed methods is their lightweight and plug-and-play nature. Feature fusion, S-ViT, and MGiT all build on existing backbones without requiring major architectural redesigns or access to extremely large pre-training datasets. This makes them adaptable to a wide range of ViT configurations and computational settings. FocusViT further extends these contributions by improving attribution faithfulness without altering the backbone, producing sharper and more reliable explanations through gradient-weighted attention and layer-skipping aggregation. Together, these methods improve predictive performance, enhance robustness, and strengthen interpretability by ensuring that model decisions align with clinically meaningful regions. These characteristics make the contributions practical for constrained domains where both data and compute are limited, and where model transparency is critical.

8.3.2 Limitations

Despite their strengths, several limitations remain. The empirical evaluation was restricted to two medical datasets ISIC-2017 dermoscopy and COVID-19 radiography which, although representative, do not cover the full diversity of medical imaging modalities. Performance in other domains, such as histopathology, ophthalmology, or 3D volumetric imaging, remains untested. Computational requirements, while moderate, still assume access to multi-GPU hardware that may not be available in all deployment settings. With respect to interpretability, FocusViT and other XAI techniques (LIME, SHAP, Attention Rollout) improve explanation quality but remain approximate: they highlight important regions but cannot guarantee a complete or causal account of model reasoning. Furthermore, hyperparameter sensitivity, for example, the choice of skip thresholds in FocusViT, may affect explanation stability, and further validation across tasks is required.

8.3.3 Validity of Results

To ensure reliability, results were obtained across multiple independent runs and validated using statistical tests. ShapiroWilk normality checks, followed by Welchs t-tests or MannWhitney U-tests as appropriate, confirmed the significance of observed improvements in balanced accuracy, AUC, and weighted F1. The integration of qualitative explainability analyses, including FocusViTs faithfulness-driven attributions, provided additional confidence that performance gains reflected meaningful reasoning rather than spurious correlations. While broader validation on more datasets and modalities is needed, the current experimental design, statistical testing, and interpretability evidence collectively support the validity of the conclusions drawn in this thesis.

8.4 Future Directions

While the contributions of this thesis provide promising advances toward robust, data-efficient, and interpretable ViTs for medical imaging, several avenues remain open for future exploration. These directions span methodological extensions, evaluation frameworks, and deployment challenges.

8.4.1 Open Research Questions in Transformer Adaptation

While this thesis demonstrates that lightweight adaptations can substantially improve the robustness, stability, and interpretability of Vision Transformers (ViTs) in constrained medical settings, several foundational questions remain open for the field.

First, the relationship between inductive bias and data efficiency in transformers is not yet fully understood. S-ViT reintroduces locality through a CNN-derived summary token and yields consistent gains, yet it remains unclear how much locality is optimal, and whether adaptive or learnable inductive biases could outperform fixed CNN summaries. Future work should investigate dynamic bias injection mechanisms that adjust locality strength depending on dataset size, imbalance, or noise characteristics.

Second, the theoretical role of auxiliary gradient guidance, as introduced in MGiT,

requires deeper analysis. While empirical results show improved early optimisation stability, the precise conditions under which auxiliary-gradient coupling improves generalisation remain underexplored. Open questions include: (i) how auxiliary depth and width affect convergence dynamics, (ii) whether guidance benefits diminish with larger datasets, (iii) whether multi-auxiliary or hierarchical guidance structures could further improve optimisation stability, and (iv) how gradient coupling interacts with modern regularisers such as sharpness-aware minimisation or stochastic depth.

Third, interpretability in transformers remains fundamentally unresolved. Although FocusViT improves faithfulness by combining gradient-weighted attention and selective layer aggregation, attention itself is not a guaranteed explanation of causality. A key open challenge is establishing formal links between transformer attention, gradient flow, and causal feature attribution. Future work should investigate whether explanation faithfulness can be enforced during training rather than measured post hoc, potentially through constraint-based or counterfactual regularisation strategies. Finally, the joint optimisation of robustness, efficiency, and interpretability remains an unsolved tri-objective problem. Existing methods, including those proposed in this thesis, improve one or two axes at a time. A unified framework that simultaneously enforces adversarial stability, data efficiency, and explanation faithfulness represents a major research opportunity for the field.

8.4.2 Broader Medical Imaging Applications

Future work should extend the proposed methods beyond dermoscopy and chest radiography to encompass a wider range of modalities, including histopathology, CT, and MRI. Multi-modal and 3D imaging tasks present opportunities to test the scalability of S-ViT and MGiT under higher-dimensional inputs. Such evaluations would help establish the generality of these methods across diverse clinical settings.

8.4.3 Human-in-the-Loop Evaluation

Although quantitative metrics and saliency visualisations provide useful evidence, ultimate trust in clinical applications requires active involvement of medical experts. Incorporating human-in-the-loop studies, where radiologists or dermatologists assess and refine explanations, would validate whether model attributions align with domain expertise and decision-making processes. Such studies can also inform iterative improvements to XAI methods.

8.4.4 Joint Optimisation of Robustness and Explainability

An important next step is to explore whether robustness and explainability can be co-optimised rather than treated independently. For example, adversarial training often alters gradient landscapes in ways that degrade saliency stability, while gradient-based explainers rely directly on those same gradients. This raises a fundamental question: can explanation faithfulness be preserved under adversarial robustness constraints?

Future work could explore multi-objective optimisation where robustness penalties and explanation consistency terms are jointly balanced. One direction would be to enforce stability of saliency maps under small input perturbations, effectively aligning robustness with interpretability. Another avenue is to incorporate attribution-aware regularisation, ensuring that gradient flow remains concentrated on semantically meaningful regions even under attack. Such approaches may bridge the gap between safety and transparency in medical AI systems.

8.4.5 Advanced Explainable AI Approaches

Emerging methods, such as gradient-guided attribution and faithfulness-driven layer selection exemplified by FocusViT, highlight new opportunities to strengthen explanation quality. Future work should explore these approaches in medical domains, systematically comparing them to established post-hoc methods. Extending such techniques to multimodal or temporal data could provide richer and more clinically relevant interpretability.

8.4.6 Clinical Translation and System-Level Challenges

Beyond algorithmic improvements, significant system-level challenges remain before ViT-based models can be safely deployed in clinical practice. Domain shift across institutions, scanner types, and patient demographics can degrade performance in subtle ways not captured by retrospective benchmarks. Future work should evaluate S-ViT and MGiT under cross-site validation, temporal shift, and prospective clinical trials.

Another challenge is uncertainty estimation. High-confidence incorrect predictions pose risks in healthcare. Integrating calibrated uncertainty estimation with interpretable ViTs represents a promising direction, potentially through Bayesian approximations or ensemble-based strategies combined with FocusViT-style attribution auditing.

Finally, regulatory and accountability considerations require explanation stability across retraining cycles. Understanding how explanations evolve as models are updated over time is critical for compliance and auditability. Developing metrics for explanation drift and stability across model versions is an open research problem with practical implications.

8.5 Closing Statement

This thesis has advanced the development of ViTs for constrained medical imaging settings by addressing four central challenges: robustness, data efficiency, interpretability, and faithful explainability. Through the proposed feature fusion framework, the Summary Vision Transformer (S-ViT), the Multi-Gradient Image Transformer (MGiT), and the FocusViT attribution method, the work demonstrates that lightweight and plug-and-play strategies can overcome the limitations of standard ViTs without requiring large-scale pre-training or costly architectural redesign. Together, these methods establish a coherent path toward models that are not only accurate but also trustworthy and clinically meaningful.

The central message of this research is that robust, interpretable, and data-efficient ViTs are achievable in practice. Feature fusion showed that adversarial resilience can be enhanced through representation-level strategies. S-ViT demonstrated that inductive bi-

ases can be reintroduced to improve generalisation on small datasets. MGiT provided evidence that auxiliary gradient sharing stabilises training and reduces overfitting. FocusViT contributed a dedicated explainability framework that improves attribution faithfulness by combining gradient-weighted attention with layer-skipping aggregation, thereby aligning explanations more closely with the models internal decision process. Across these contributions, quantitative improvements were consistently supported by statistical validation, and qualitative explanations confirmed that the models aligned with task-relevant anatomy in real medical images. These findings collectively lay a strong foundation for the reliable use of ViTs in healthcare.

Looking forward, this thesis also opens the door to broader opportunities. By extending the methods to multi-modal and 3D imaging, integrating robustness with interpretability, and involving human experts in the evaluation loop, future research can push closer toward clinically deployable AI. At the same time, emerging explainability frameworks, such as gradient-guided attribution and counterfactual analysis, offer promising directions to further enhance trust and transparency.

In conclusion, the contributions presented here take a significant step toward the goal of trustworthy AI in medicine. They demonstrate that careful adaptation of transformer models can deliver systems that respect the dual demands of performance and accountability. By uniting robustness, efficiency, interpretability, and faithful explainability, this thesis provides both theoretical and practical insights that can inform future research and support the safe deployment of AI in high-stakes domains.

Bibliography

[1] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 104–12 113.

[6] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[7] J.-M. Fellous, G. Sapiro, A. Rossi, H. Mayberg, and M. Ferrante, "Explainable artificial intelligence for neuroscience: Behavioral neurostimulation," *Frontiers in neuroscience*, vol. 13, p. 1346, 2019.

[8] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, p. 1096, 2019.

[9] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*, PMLR, 2021, pp. 10 347–10 357.

[13] I. E. Sobel, *Camera models and machine perception*. stanford university, 1970.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Ieee, vol. 1, 2005, pp. 886–893.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[19] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[20] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[21] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101 552, 2019.

[22] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," *arXiv preprint arXiv:1701.04862*, 2017.

[23] M. Ali, H. Raza, J. Q. Gan, and M. Haris, "Optimising vision transformer performance on limited datasets: A multi-gradient approach," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 693–702.

[24] M. Ali, H. Raza, J. Q. Gan, and M. Haris, "Integrating spatial information into global context: Summary vision transformer (s-vit)," in *2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, 2024, pp. 206–213.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 2002.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[27] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[28] S. Jain and B. C. Wallace, "Attention is not explanation," *arXiv preprint arXiv:1902.10186*, 2019.

[29] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," *arXiv preprint arXiv:2005.00928*, 2020.

[30] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[31] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems*, vol. 32, 2019.

[32] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[33] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[34] K. Eykholt, I. Evtimov, E. Fernandes, *et al.*, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1625–1634.

[35] A. Shafahi, M. Najibi, M. A. Ghiasi, *et al.*, "Adversarial training for free!" *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[36] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*, PMLR, 2019, pp. 7472–7482.

[37] J. Zhang, J. Zhu, G. Niu, B. Han, M. Sugiyama, and M. Kankanhalli, "Geometry-aware instance-reweighted adversarial training," *arXiv preprint arXiv:2010.01736*, 2020.

[38] L. Rice, E. Wong, and Z. Kolter, "Overfitting in adversarially robust deep learning," in *International Conference on Machine Learning*, PMLR, 2020, pp. 8093–8104.

[39] M. Ali, H. Raza, and J. Q. Gan, "Fortifying deep neural networks for industrial applications: Feature map fusion for adversarial defense," in *2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA)*, IEEE, 2024, pp. 1–6.

[40] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[41] H. Wu, B. Xiao, N. Codella, *et al.*, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 22–31.

[42] M. Ali, H. Raza, J. Q. Gan, and M. Haris, "Focusvit: Faithful explanations for vision transformers via gradient-guided layer-skipping," in *The IEEE/CVF Winter Conference on Applications of Computer Vision*, IEEE, 2026, p. xxx.

[43] H. Touvron, M. Cord, and H. Jégou, "Deit iii: Revenge of the vit," in *European Conference on Computer Vision*, Springer, 2022, pp. 516–533.

[44] L. Beyer, X. Zhai, and A. Kolesnikov, "Better plain vit baselines for imagenet-1k," *arXiv preprint arXiv:2205.01580*, 2022.

[45] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International conference on machine learning*, PMLR, 2020, pp. 2206–2216.

[46] O. Seddati, S. Dupont, S. Mahmoudi, and M. Parian, "Towards good practices for image retrieval based on cnn features," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 1246–1255.

[47] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International conference on machine learning*, PMLR, 2018, pp. 274–283.

[48] J. Zhang, X. Xu, B. Han, *et al.*, "Attacks which do not kill training make adversarial learning stronger," in *International conference on machine learning*, PMLR, 2020, pp. 11 278–11 287.

[49] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[50] L. Yuan, Y. Chen, T. Wang, *et al.*, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 558–567.

[51] J. Bai, L. Yuan, S.-T. Xia, S. Yan, Z. Li, and W. Liu, "Improving vision transformers by revisiting high-frequency components," in *European Conference on Computer Vision*, Springer, 2022, pp. 1–18.

[52] B. Li, Y. Hu, X. Nie, *et al.*, "Dropkey for vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 700–22 709.

[53] S. Chang, P. Wang, H. Luo, F. Wang, and M. Z. Shou, "Revisiting vision transformer from the view of path ensemble," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 889–19 899.

[54] R. Nakamura, H. Kataoka, S. Takashima, E. J. M. Noriega, R. Yokota, and N. Inoue, "Pre-training vision transformers with very limited synthesized images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 360–20 369.

[55] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[56] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localiza-

tion," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[57] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, e0130140, 2015.

[58] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "Ai for radiographic covid-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 610–619, 2021.

[59] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[60] N. C. Codella, D. Gutman, M. E. Celebi, *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, IEEE, 2018, pp. 168–172.

[61] M. E. Chowdhury, T. Rahman, A. Khandakar, *et al.*, "Can ai help in screening viral and covid-19 pneumonia?" *Ieee Access*, vol. 8, pp. 132 665–132 676, 2020.

[62] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[63] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[64] A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang, "Adversarial training can hurt generalization," *arXiv preprint arXiv:1906.06032*, 2019.

[65] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *international conference on machine learning*, PMLR, 2019, pp. 1310–1320.

[66] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[67] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[68] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[69] N. Carlini, A. Athalye, N. Papernot, *et al.*, "On evaluating adversarial robustness," *arXiv preprint arXiv:1902.06705*, 2019.

[70] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," *arXiv preprint arXiv:2112.13492*, 2021.

[71] S. dAscoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in *International conference on machine learning*, PMLR, 2021, pp. 2286–2296.

[72] A. Khan, Z. Rauf, A. Sohail, *et al.*, "A survey of the vision transformers and their cnn-transformer based variants," *Artificial Intelligence Review*, vol. 56, no. Suppl 3, pp. 2917–2970, 2023.

[73] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2012, pp. 3498–3505.

[74] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian conference on computer vision, graphics & image processing*, IEEE, 2008, pp. 722–729.

[75] Z. Zhang, X. Lu, G. Cao, Y. Yang, L. Jiao, and F. Liu, "Vit-yolo: Transformer-based yolo for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2799–2808.

[76] A. G. Howard, M. Zhu, B. Chen, *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[77] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[78] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[79] R. Wightman, *Pytorch image models*, https://github.com/rwightman/pytorch-image-models, 2019. DOI: 10.5281/zenodo.4414861.

[80] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[81] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[82] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.

[83] Y. Liu, E. Sangineto, W. Bi, N. Sebe, B. Lepri, and M. Nadai, "Efficient training of visual transformers with small datasets," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 818–23 830, 2021.

[84] H. Zhu, B. Chen, and C. Yang, "Understanding why vit trains badly on small datasets: An intuitive perspective," *arXiv preprint arXiv:2302.03751*, 2023.

[85] H. Gani, M. Naseer, and M. Yaqub, "How to train vision transformer on small-scale datasets?" *arXiv preprint arXiv:2210.07240*, 2022.

[86] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, Springer, 2014, pp. 446–461.

[87] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[88] M. Caron, H. Touvron, I. Misra, *et al.*, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[89] J. Chen, Y. Lu, Q. Yu, *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[90] S. Oh, N. Kim, and J. Ryu, "Analyzing to discover origins of cnns and vit architectures in medical images," *Scientific Reports*, vol. 14, no. 1, p. 8755, 2024.

[91] T. Rahman, A. Khandakar, Y. Qiblawey, *et al.*, "Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images," *Computers in Biology and Medicine*, vol. 132, p. 104 319, 2021, ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2021.104319.

[92] T.-Y. Ross and G. Dollár, "Focal loss for dense object detection," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2980–2988.

[93] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9268–9277.

[94] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: An overview," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[95] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar, "On the (in) fidelity and sensitivity of explanations," *Advances in neural information processing systems*, vol. 32, 2019.

[96] F. Mehri, M. Fayyaz, M. S. Baghshah, and M. T. Pilehvar, "Skipplus: Skip the first few layers to better explain vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 204–215.

[97] U. Bhatt, A. Weller, and J. M. Moura, "Evaluating and aggregating feature-based model explanations," *arXiv preprint arXiv:2005.00631*, 2020.

[98] A. Hedström, L. Weber, D. Krakowczyk, *et al.*, "Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond," *Journal of Machine Learning Research*, vol. 24, no. 34, pp. 1–11, 2023. [Online]. Available: http://jmlr.org/papers/v24/22-0142.html.

[99] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 782–791.

[100] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, vol. 2, 2011.

[101] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 conference on computer vision and pattern recognition workshop*, IEEE, 2004, pp. 178–178.

[102] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009, pp. 413–420.

[103] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that?" *arXiv preprint arXiv:1611.07450*, 2016.

[104] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin, "Understanding and improving layer normalization," *Advances in neural information processing systems*, vol. 32, 2019.

[105] U. M. I. Lab, *Quantus: Xai metrics for machine learning*, Accessed: 2025-07-15, 2023. [Online]. Available: https://github.com/understandable-machine-intelligence-lab/Quantus/tree/main/tutorials.

[106] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[107] P. Chalasani, J. Chen, A. R. Chowdhury, X. Wu, and S. Jha, "Concise explanations of neural networks using adversarial training," in *International Conference on Machine Learning*, PMLR, 2020, pp. 1383–1391.

[108] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, Springer, 2020, pp. 213–229.

[109] B. Zhang, Z. Tian, Q. Tang, X. Chu, X. Wei, C. Shen, *et al.*, "Segvit: Semantic segmentation with plain vision transformers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4971–4982, 2022.

[110] A. Chatzimparmpas, R. M. Martins, I. Jusufi, and A. Kerren, "A survey of surveys on the use of visualization for interpreting machine learning models," *Information Visualization*, vol. 19, no. 3, pp. 207–233, 2020.