An AI Driven Pipeline for 6G Enabled Digital Creatives Identification, Performance Monitoring and Attribution

Varun Dutt[1,2], *Lucas Galan*[1], *Faiyaz Doctor*[2], *Lina Barakat*[2], *Kate Isaacs*[1], *Demetris Hadjigeorgiou*[1], *Aldo Fumagalli*[1]
[1]Rapp Limited, London, UK
[2]School of Computer Science and Electronic Engineering, University of Essex, UK
varun.dutt, lucas.galan, kate.isaacs, demetris.hadjigeorgiou@{rapp.com}, Aldo.Fumagalli@codeworldwide.com, fdocto, lina.barakat@{essex.ac.uk}

Abstract— **In the growing creative marketing sector, a high volume of digital assets in the form of images and videos are often generated for various forms of advertising. It is critical to be able to track these assets back to the original high value photo and video shoots they are derived from as well as understanding in real-time how well they are performing and how they can be augmented to perform better in response to the needs and preferences of dynamic audiences. In this paper we propose a framework that leverages deep learning approaches that learns to match local features across images to tie master images to various derived digital assets. The proposed pipeline is further able to use audience responses to predict performances of assets with attribution mechanisms for analysing how the modification of asset features enhance or degrade performance. We finally discuss how 6G infrastructures would be used with the pipeline to facilitate real-time and dynamically responsive content intelligence.**

*Keywords— **Content Intelligence, Creative Asset Tracking, Deep Feature Extractors, Performance Analyses Attribution Pipeline, Transformers, Diffusion Models, 6G infrastructures.***

I. Introduction

The forecasted emergence of the sixth generation (6G) mobile communication technology by 2030 is aimed at providing global coverage using more autonomous 6G networks providing ultra-high-speed transmission rates of speed of 1–10 Tbps [1] with reduced latencies in the range of range of 10–100 µs as well as high connectivity densities in range of 107 devices/km2, supported by high traffic capacities [1]. This will provide opportunities for supporting highly interactive media streaming and other human centric experiences such as mobile, extended and virtual reality applications requiring high data rate accessibility to mobile services, data, and multimedia content [2]. The eventual emergence of 6G presents huge opportunities for brand content creation, capturing and monitoring of performance indicators from these assets and dynamically analyzing and adapting content in response to demand and consumer preferences or mood.

The global digital content creation sector which was valued at USD 25.6 billion in 2022 and is estimated to expand at a CAGR of 13.5% from 2023 to 2030. Key factors which are driving this growth are the adoption of AI and the increasing adoption of cloud computing. Digital content creation can incorporate creating advertising and marketing content for and various types of formats such as textual, graphical, audio and video as well as publishing and promoting the content on various online platforms. In the last decade we have seen the role of digital technology and specific machine learning increasingly applied within the advertising sector. The primary goal of these technologies is to automate digital content generation and performance monitoring to be more cost effective, faster and accurate given the huge financial and carbon costs associated with creating effective digital creatives and marketing campaigns.

Content Intelligence (CI) aims to tie together broadly 3 machine learning based pipelines namely Digestion Pipeline, Prediction Pipeline and Insight Pipeline to extract meaning out of different media assets, be it image, video, text or sound, so that it can be used to power other downstream applications or displayed to end-users in the form of actionable insights. Key elements of a CI framework are:

(1) Performance - Providing an end-to-end view of asset performance. Collecting or extract metadata for asset usage, customisation, engagement and revenue performance to be used to power predictive models and contribute to augmenting the metadata with tags from Image, Text or Video recognition algorithms that can be used for shifting and grouping content together.

(2) Insights – generating insights that drive creative and channel decisioning. This involves a suite of strategic planning tools to support both central planning as well as local market customisation and activation strategies.

(3) Predictions - AI and machine learning are used to generate content predictions across the content creation process, identifying similar assets or customisations, and highlighting over/under performing assets as well as predicting asset engagement decay.

(4) Automation - Content automation capability allowing actions to be taken on insights and predictions on demand. Automatically create customisations e.g. changing emotions on a face or modifying the background etc. to drive performance.

(5) Optimisation - Bringing together insights, predictions, and automation into an always-on programme of content optimization. Creating a virtuous cycle of automated testing, learning, optimisation driving improvements in Return on M marketing Investment.

As a global creative marketing agency, Rapp Limited continually evolves its strategy to meet growing customer expectations and environmental demands, providing more efficient, accurate and scalable solutions to its clients. As the budgets for marketing have increased so has the size of the advertising data which has made its tracking and management a real challenge for companies. One of the tasks that is extremely time consuming and expensive is to match original clean images produced from an advertisement shoot (referred to as a ‘master asset’) to their augmented versions used in various types of advertising campaigns (referred to as an ‘ad asset’). A second challenge is to perform real-time auditing of these assets to track and analyze their performance across cross-device and cross-platform real-time interactions with consumers (posting of comments, likes and shares). This could facilitate dynamic content augmentation and generation in response to audience feedback enabling content to adapt to group based or individual preferences enhancing audience engagement and strengthening brand identification.

The typical process of generating advertising creatives is as follows: Firstly, the company does a photo shoot of the featured products e.g. a new clothing line, or a new car model at different locations where the production team create raw master asset video clippings and images from the shoot. The master assets are subsequently used for creating ‘ads’ such as banner ads on websites, newspaper ads etc. The ads are created by performing augmentation on the master assets such as cropping, adding a text overlay on the images, embedding a master asset within another image with different border designs and layouts etc. The tracking and precise matching of these augmented ad assets to their original master assets is a huge task for companies with a presence in different continents where master assets are used by various departments to generate advertising content tailored to local language and cultural tastes. This makes it very challenging to precisely tie the generated ad assets back to their master assets. Moreover, many large companies have terabytes of data which means they have tens of thousands or millions of ads and master assets that have not been linked together. Tying these ads and master assets manually is not viable in terms of time and costs involved, hence there is a need for digital automation and machine learning solutions to reduce both time and cost drastically to make the process commercially viable.

The primary challenge in automating the tying of these master and ad assets is the diversity in augmentations used for creating the ads and the low resolution of the types of ads such as banner ads. Another challenge is that the task of matching ad assets to master assets has to be highly accurate in order to correctly identify the exact image embedded in these ads with borders or text written over. Furthermore, similarity functions used for calculating the distance or angle between feature vectors of two images (ads and master assets) in the feature space with best results vary from sample to sample which in some cases give dramatically different results. Finally, the low diversity of objects in the ads and master assets also poses a challenge e.g. for an automobile company, all the images will have cars and many shots will have the same model and color of a car and be taken from different angles. To effectively track and compare performance of each of the produced master assets, they have to be exactly matched with their augmented ad based versions. In this paper we try to address these challenges by introducing a hybrid system where we leverage Convolutional Neural Network (CNN) trained on ImageNet dataset with a cropping preprocessing pipeline to generate the top 500 similar master asset for each ad asset and finally use a transformer and keypoint matching algorithm called LightGlue [3] to narrow down the top 500 similar master assets to top 20 assets which are then presented via a purpose-built interface developed for the human agent to tie the images together. We have used this hybrid approach due two key reasons, first is the very high accuracy of the LightGlue algorithm compared any other approach however we use the CNN approach to narrow down the master assets to top 500 before using LightGlue as the transformer model is a very heavy model which requires a high computation and time commitment if used in isolation. Therefore, we use this hybrid approach to get the best from both approaches where the CNN approach is used to reduce computation costs while almost getting the same performance by using the LightGlue algorithm as the final step.

The final step in this pipeline is prediction and attribution. We have used the Knot system to ‘tie’ the ad assets to the corresponding master assets. After tying, we train a CNN model called efficient net [4] pretrained on ImageNet dataset [5] to predict the likes (like Instagram data) or clicks (Google ads, Facebook ads etc.) given an ad asset for that brand. Finally, we use ANVIL which is a generative pipeline which we use to modify elements of an image e.g. the background of an image and use our prediction pipeline to predict likes or clicks for the augmented images produced by ANVIL to attribute the change in performance to the elements of the image augmented by ANVIL.

We organize the rest of this paper as follows. Section 2 presents a brief review of the relevant previous work. Section 3 details the proposed Knot Tie. Section 4 presents a brief overview of the prediction and attribution part of our proposed pipeline. Section 5 presents and discusses initial evaluation results for Knot-Tie and a proposed 6G framework. Finally, Section 6 concludes this paper.

II. Previous Work

A. Keypoints Matching

In feature matching, Lowe's SIFT [7] takes into account image rotation, affine transformations, intensity shifts, and viewpoint shifts. There are 4 main steps in the SIFT algorithm. Using Difference of Gaussians (DoG), one can first estimate a scale space's extrema. In the second stage key point positions refer to where potential key points are located and improved by removing points with low contrast. In the third stage, a key point’s orientation is determined based on the local image gradient. Finally, a descriptor generator computes the local image descriptor for each key point based on the magnitude and direction of the image gradient.

SURF [8] is another image matching algorithm where rectangular filters are used to approximate the DoG. Since squared convolution is much faster when using an integral image, squares are used as approximations rather than averaging the Gaussian image. For various scales, this can also be carried out concurrently. A Hessian blob detector is used by SURF to locate interesting objects. The detector uses the appropriate Gaussian weights to assign orientation using wavelet responses in the horizontal and vertical directions. The area around the key points is chosen and divided into sub-regions, after which the wavelet responses are taken into account and represented with the SURF function descriptor for each sub-region. The Laplace sign, or base percentage points, uses values that have already been computed for the survey. Utilizing Laplace's sign, one can differentiate light spots on a dark background. In the case of matching, features are only compared if they have the same contrast type (based on the sign), enabling quicker matching.

With some modifications, ORB [9] combines the BRIEF descriptor and the FAST key point detector. To start, the approach identifies key points using the FAST function. The first N points are then determined using the Harris angle measure. The FAST function, which is a form of rotation, does not compute orientation, however it determines the middle-corner point's intensity-weighted centroid. The orientation is determined by the direction of the vector from this vertex to the centroid. For better rotational constancy, moments are calculated. When the plane rotates, the SHORT descriptor performs erratically. The rotation matrix in ORB is computed using patch orientation, whereas LETTER descriptors are orientation-driven.

The perceptual hash algorithm [10] is another approach that uses a class of functions for creating hash strings that can be compared to one another for image matching. Such an algorithm will yield the same hash value for any image that is processed. However, this differs from cryptographic hash functions in that even when the image is slightly altered, the hash value does not completely change. Instead, it only modifies slightly and stays close to the original. The pHash algorithm is one of the most widely used variations of the perceptual hash algorithm for images. In this approach we first apply an 8 x 8 pixel resolution to the original image. Then the image's color scheme is condensed and turned to grayscale after which we calculate the average value across 64 colors. The average value of each image's value is compared to it. The unit is kept if it is greater; otherwise, the source image produces a grayscale image. Grayscale image of size 8 x 8 in comparison to the norm using pHash for hashing. zero. The end result is a 64-bit value, which can be represented as a hexadecimal number. The Hamming distance between the two images is used in this situation to determine how similar they are.

B. Neural Network Based

In contrast to the above mentioned approaches which use hand crafted feature extractors the features extracted by deep neural network based approaches such as CNNs are learnt from data. Here several CNN based techniques have been proposed for image matching where a popular approach is Siamese neural networks [11]. Here the two images to be matched are passed through a CNN where the weights are shared by the two images and extracted features are later compared with each other [12],[13]. Triplet networks have also been employed in [14],[15],[16] for matching similar images where a triplet of images are fed to the Siamese network with one pair being matching and the other non-matching which shows better rate of convergence.

Recently some large-scale image similarity detection datasets such as the DISC21 have been released by companies like meta. The dataset mimics real-life cases appearing in social media, e.g. integrity-related problems dealing with misinformation and objectionable content. The dataset was used as the benchmark for the Image Similarity Challenge at NeurIPS'21 (ISC2021) [17]. The team that won first place [18] used three models with ResNet-50 [19], ResNet-152, and ResNet50-IBN as backbones followed by GeM pooling [20], combined with WaveBlock [21], and finally append a final projector module that consists of linear and nonlinear layers and increases the dimensionality to 2048. All the three models were pre-trained on an external dataset and then finally on the DISC21 dataset. They ensembled the three models to make the final prediction.

The primary challenge in these approaches is dataset generation. For instance, the DISC21 dataset prepared by Meta has millions of image pairs for training, however the task that this paper is trying to address is very different as the augmentations being performed on master assets to create the ad assets are drastically different, see Figure.2. Equally, the DISC21 dataset has a variety of objects in the source images as compared to our case where the variety is a lot less and the task becomes much harder. Therefore, one-shot learning methods such as SIFT, ORB etc. are favorable for this task however will not perform as well in terms of low precision scores to be viable for large ad collections and master assets.

C. Hybrid between Neural Network and Keypoints Matching

SuperGlue [22] is a matching algorithm that matches two sets of local features using a neural network by jointly finding correspondences and rejecting non-matchable points. The cost function of this optimization is predicted by a Graph Neural Network (GNN). Inspired by the success of the Transformer [23], it uses self- (intra-image) and cross- (inter-image) attention to leverage both spatial relationships of the key points and their visual appearance.

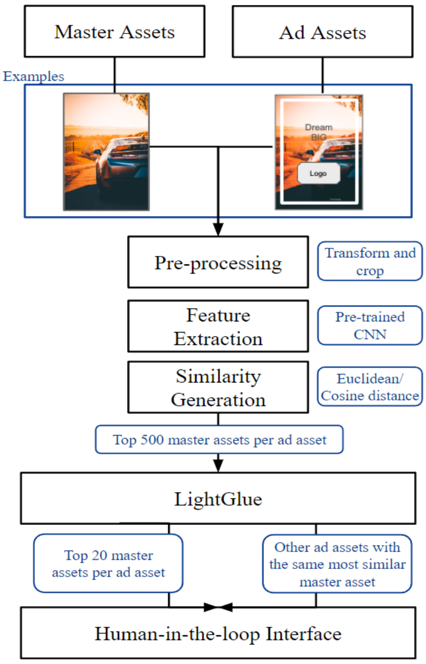
Master Assets
Ad Assets
Examples
Dream
BIG
Logo
Pre-processing
Transform and crop
Feature Extraction
Pre-trained CNN
Similarity Generation
Euclidean/ Cosine distance
Top 500 master assets per ad asset
LightGlue
Top 20 master assets per ad asset
Other ad assets with the same most similar master asset
Human-in-the-loop Interface

Fig. 1. Block diagram of the proposed approach Knot-Tie.

Based on the existing literature there are three limitations with existing approaches. Firstly, as previously mentioned, the keypoint matching algorithms have low precision scores especially for our use case where the primary object is very homogenous in the dataset. Secondly, for the Neural net-based approaches most networks require well labeled large datasets and still lack generalization ability, which for our use case is very important. Finally, in the hybrid method like SuperGlue or LightGlue the methods are very accurate and generalize well but they have very high computation requirements that makes them not feasible when we have large master and ad assets to tie. The proposed framework in the next section aims to address these limitations.

III. Knot Framework

We propose a hybrid approach where both humans and an AI system improves the performance of accurately tying master assets to ad assets towards being acceptable for commercial use. The first part of the proposed framework is to narrow down the master assets per ad asset to top 500 most similar master asset by passing the ad assets through our preprocessing pipeline where we take the following five crops of the ad assets:

(1) horizontal/vertical center crop depending on height-width ratio of the asset
(2) left half crop
(3) right half crop
(4) upper half crop
(5) lower half crop

These five crops have been strategically chosen to isolate the embedded master asset in the ad asset, see Figure. 2. These crops along with the original image are fed to the feature extractor function which generates six feature vectors, one for each crop and for the original ad asset. Then the Master asset is fed to the extractor to get its feature vector. Finally, two similarity scores between each of the crops plus the original ad asset and the master asset is calculated using the cosine similarity and Euclidean distance functions. The lowest value among the six is selected as the similarity value for that ad-master asset pair for both the similarity functions.

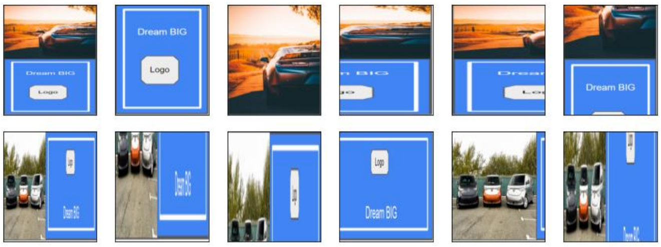

Fig. 2. The 5 crops and the original image produced by the preprocessing pipeline.

The second part of the framework involves matching the top 500 masters determined by the first phase for each ad asset using the LightGlue algorithm to further narrow down the master assets to top 20 master assets per each ad asset. In the final step, the human agent ties the ad assets and the corresponding master asset using a purpose-built interface depicted in Figure. 4. The knot tie system then ties all the ads that have the tied master in its top 5 similars of the top 20 generated by LightGlue and the similarity score is above 0.60 to that master asset automatically which translates to the human agent doing significantly less amount of manual work as most ad asset gets tied to their master through this feature of the system. The automatic tying system plays a crucial role as the dataset we get in commercial setting have many ad assets with slight variations produced using the same master asset. The summary of all the steps involved in the proposed framework are as follows as also depicted diagrammatically in Figure. 1:

(1) Five crops generated for each of the ad assets.
(2) Original ad asset, master asset and ad asset crops passed through the feature extractor (pre-trained CNN).
(3) Similarities generated for the selected ad asset by calculating the distance and the angle between the feature vectors of each of the ad asset crops and master asset using Euclidean distance and cosine similarity function respectively.
(4) Lowest value for the angle and distance between ad asset and master asset chosen as the similarity value between selected ad and master assets.
(5) Similarity between the selected ad asset and the master assets sorted in ascending order of values.
(6) Top 500 master assets to be passed through the second part of the framework (LightGlue).
(7) Use LightGlue to further narrow down the top 500 master assets to top 20 per ad assets.
(8) The human agent with the help of the interface ties the master with the corresponding ad asset by selecting the correct master asset from the top 20 master similars list generated in previous step. After the human agent selects the master our system scans through the top 20 master similars list of all the other ads and ties the ad assets with the same master asset if it is in top 5 in the master similar list and has a similarity score of more than 0.60.

A. Feature Extractor Model

We evaluated several CNNs pre-trained on the ImageNet dataset such as ResNets, EfficientNets, DenseNets [24] etc. for this task. From our experimentations we found the resnets to outperform all other architectures by a significant margin for this specific image matching task. Additionally, the deeper models e.g. ResNet 200d seemed to perform better compared to the shallower ones like ResNet 50. As such we chose ResNet 200d, a modification on the ResNet architecture that utilizes an average pooling tweak for down sampling [25]. The selected network was then pretrained on ImageNet dataset as the feature extractor used in our system.

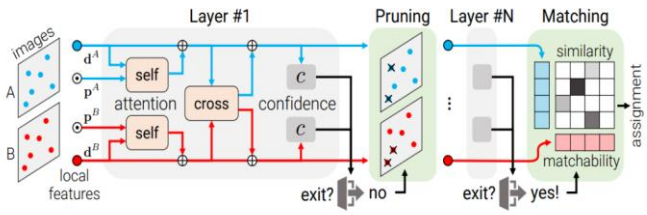

Fig. 3. The LightGlue architecture.

B. LightGlue Model

LightGlue is a deep neural network (see Figure 3) that learns to match local features across images. Given a pair of input local features (*d*, *p*), each layer augments the visual descriptors (•,•) with context based on self- and cross-attention units with positional encoding ⊙. A confidence classifier c helps decide whether to stop the inference. If a few points are confident, the inference proceeds to the next layer although points that are confidently unmatchable are pruned. Once a confident state is reached, LightGlue predicts an assignment between points based on their pariwise similarity and unary matchability.

LightGlue introduces improvements over the state-of-the-art sparse matching model, SuperGlue, making it more efficient, accurate, and easier to train. LightGlue adapts to the difficulty of each image pair, allowing for faster inference on easy image pairs. It is Pareto-optimal on the efficiency-accuracy trade-off compared to existing sparse and dense matchers. LightGlue is adaptive to the difficulty of each image pair, resulting in faster inference on easy image pairs. It can stop at earlier layers when predictions are confident. LightGlue discards non-matchable points at an early stage, focusing attention on the covisible area. LightGlue is more accurate, efficient, and easier to train than SuperGlue, making it a plug-and-play replacement with a fraction of the run time.

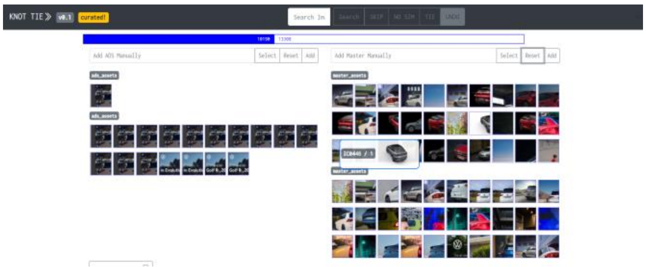

Fig. 4. Interface for Knot-Tie

C. Interface

As previously discussed, the human agent is given an interface (see Figure 4) to perform the ties between the ad and its corresponding master asset as well as between an ad asset with its corresponding duplicate ad assets. The ad assets with no similar assets are tied to the string 'No sim' using the no sim button. On the left half of the screen the selected ad and its top 70 similar images retrieved are displayed and on the right half of the screen the master similars are displayed. The master similars section on the right half of the screen shows two lists of top 20 similar images each, where the upper list is generated by using Euclidean distance and the lower list is generated by using the cosine similarity function.

IV. Prediction and Attribution

We propose a pipeline (see Figure.5.) to predict and attribute the performance change on isolated elements for an ad campaign of a brand. We used Convolutional neural network to predict the likes/clicks on a given ad asset and then used ANVIL (a pipeline that modifies elements of an image using various algorithms like stable diffusion, ControlNet, OneFormer) to modify different elements of the image and then pass it through the predictor model to note down the change in performance metric. We then use the ground truth label of the original asset and the predicted label of the modified image to calculate the contribution of the isolated element to the performance of the asset.

A. Predictor Model

We used the EfficientNet model to predict likes/clicks given an ad asset (image). EfficientNet uses a technique called compound coefficient to scale up models in a simple but effective manner. Instead of randomly scaling up width, depth or resolution, compound scaling uniformly scales each dimension with a certain fixed set of scaling coefficients. Using the scaling method and AutoML, the authors developed seven models of various dimensions, which surpassed the state-of-the-art accuracy of most convolutional neural networks, and with much better efficiency.

B. Stable Diffusion

Stable Diffusion [26] models are a class of generative AI models that generate high-resolution images of varying quality. They work by gradually adding Gaussian noise to the original data in the forward diffusion process and then learn to remove the noise in the reverse diffusion process. They are latent variable models referring to a hidden continuous feature space, look similar to VAEs (Variational Autoencoders) [27], and are loosely based on non-equilibrium thermodynamics. A denoising diffusion modeling is a two-step process. The first is the forward diffusion process in which there is a Markov chain of diffusion steps in which noise is incrementally and randomly added to the original data. The second is the reverse diffusion process which tries to reverse the diffusion process to generate original data from the noise.

C. ControlNet

ControlNet [28], functions as a complete neural network structure, taking charge of substantial image diffusion models, like Stable Diffusion, to grasp task-specific input conditions. ControlNet achieves this by replicating the weights of a major diffusion model into both a "trainable copy" and a "locked copy." The locked copy preserves the learned network trained from vast image data, while the trainable copy gets trained on task-specific datasets to master conditional control. This process connects trainable and locked neural network segments using an exceptional convolution layer called "zero convolution." In this layer, convolution weights progressively evolve from zeros to optimal settings through a learned approach. This strategy maintains the refined weights, ensuring strong performance across various dataset scales. Importantly, because zero convolution doesn't introduce extra noise to deep features, the training speed matches that of fine-tuning a diffusion model. This contrasts with the lengthier process of training entirely new layers from scratch.

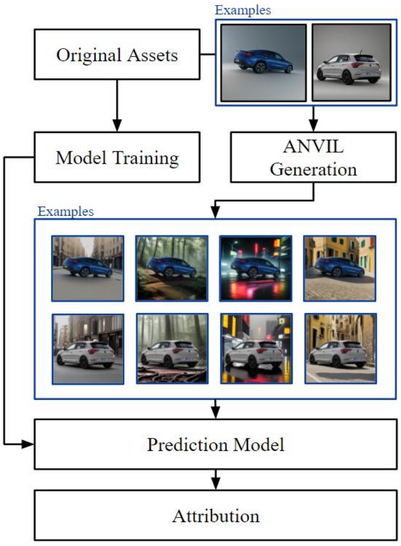

Fig. 5. Block diagram of the proposed approach for prediction and attribution

D. ANVIL

Our Generative pipeline (refer to Figure.5.) segments the image asset and isolates different elements of the image, like the main object such as car or models wearing the apparels of a brand etc. and other supporting elements such as trees, hill, beach etc. after choosing the element the user can modify it using the following steps.

1. The image is first segmented by the OneFormer model.
2. Then the user selects the element of the image to be modified.
3. After which the user passes the prompt to the ControlNet and stable diffusion model which generates a new image with a modified version of the selected element based on the prompt.

V. Experiments and Framework

We tested various existing algorithms and our proposed Knot platform for image matching as a sub element of the proposed pipeline. We tested the one-shot keypoint matching algorithms like SIFT, keypoint matching with neural network algorithms like LightGlue and our proposed approach Knot-Tie model on an actual dataset of a car company which has 450 ad and 300 master assets tied together. For the purpose of client confidentiality further details of the dataset cannot be revealed. We pair the ad and master assets to get 450 positive samples i.e. an ad and master pair that are actually the same and 450 negative samples. We use mAP (mean average precision) as the performance metric because it is more important for the algorithm to minimize false negatives. Precision becomes even more critical in our case because the human agent at the end can rectify the false positives but not the false negatives as the corresponding master asset might not appear in the similar master list. The two main goals of the algorithm were to match the precision scores of LightGlue however, reduce the computation time significantly to make the approach viable for large commercial datasets. The results (see Table 1) show that the performance of the proposed model is comparable to the state-of-the-art LightGlue model, although the time required has gone down significantly.

TABLE I. mAP (Mean Average Precision) Score Comparison Table.

Algorithm	mAP	Time(m)
SIFT	48%	2.6
ORB	47.2%	2.5
LightGlue	72.8%	32.4
Knot-Tie	70.6%	8.4

The proposed pipeline combining Knot-Tie and Anvil can be depicted in Figure. 6 which also shows how a 6G infrastructure would be used with the pipeline to facilitate real-time auditing of spatially distributed ad assets being displayed and streamed on digital billboards and personalised devices. The framework would allow real-time collection of asset performance information by tracking consumer and interaction (views, clicks, likes). This could also involve tracking and analysis of eye gaze and physiological signals using unobtrusive smart eye ware and watches. This rich context and location aware asset performance data would be used to train prediction models used with Anvil to attribute performance changes on isolated elements of the image in context of factors such as location, time of day, public engagement and co-occurring events. Higher performing generated asset variations could then be streamed to update displayed content at specific locations where they would be most effective. This process would repeatedly occur in real-time enabling dynamically responsive content intelligence.

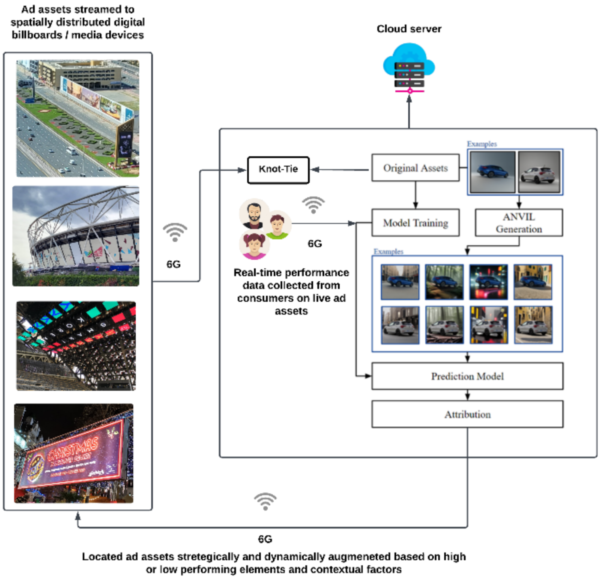
Ad assets streamed to spatially distributed digital billboards / media devices
Cloud server
Knot-Tie
Original Assets
Examples
Model Training
ANVIL Generation
6G
6G
Real-time performance data collected from consumers on live ad assets
Examples
Prediction Model
Attribution
6G
Located ad assets stretegically and dynamically augmeneted based on high or low performing elements and contextual factors

Fig. 6. Dynamic Asset Auditing and Attribution Framework

VI. Conclusions

This paper proposed an AI pipeline to match two images and then use generative AI and CNN to predict performance and attribute the performance change on different elements. Image matching pipeline consists of a hybrid approach where both humans and an AI system are used for accurately matching original images produced from an advertisement shoot to their augmented versions used for digital advertising campaigns. Specifically, the system uses Resnet 200d pretrained on ImageNet dataset as the feature extractor, a cropping method and multiple similarity functions which are used to produce top 500 similar masters for each ad asset and finally using the LightGlue algorithm to further reduce the 500 list to top 20 similar master assets. Evaluations of the system on our client's private dataset like ad campaigns on their own and other social media platforms showed that the proposed framework significantly outperforms state-of-the-art methods in terms of the computational requirement while maintaining almost the same performance. Moreover, these results confirmed the importance of hybrid humans in the loop systems for AI solutions for commercial use by filling in the gaps in the existing AI technology and the effect of the cropping method used in the performance of the feature extractor. The results also show the huge efficiency and accuracy gains that can be achieved in the advertising industry by adding AI solutions to otherwise manual and highly subjective tasks while still preserving human oversight in decision making. The second part of the pipeline is the performance and attribution part, where we have used CNN as a predictor model and stable diffusion with controlnet to modify elements to isolate and attribute performance.

6G enabled content intelligence pipelines will eventually be able to offer brands diversified omni-channel, communication including augmented and virtual reality technologies. leveraging these innovative communication channels will create more immersive and enriched brand experiences for audiences. Interacting with the virtual world can allow the audience to better understand product features, experience services, and establish closer relationships with brands. 6G can enable more effective data driven marketing campaigns. The technology will enable abundant real-time data on asset performance to be collected, providing brands with in-depth insights into audience behavior patterns, preferences, and demands. The identification, matching and tracking of specific creative assets enables more precise targeted and personalized content creation. Collectively this can be used to forecast market trends to develop directed marketing strategies, enhance brand awareness, influence, and recognition among more engaging consumers.

Acknowledgments

This work is supported by the Knowledge Transfer Partner (KTP) funding (Partnership No: 10008157) from Innovate UK, between Rapp Limited, and the University of Essex. For the purposes of open access, the authors have applied a Creative Commons Attribution (CC BY) License to any Author Accepted Manuscript (AAM) version arising from this submission.

References

[1] Hu, Y., 2023. Capacity analysis and data detection of OvTDM-MIMO system. IEEE Access, 11, pp.20647-20656.

[2] Lindenberger, P., Sarlin, P.E. and Pollefeys, M., 2023. LightGlue: Local Feature Matching at Light Speed. arXiv preprint arXiv:2306.13643.

[3] Banafaa, M., Shayea, I., Din, J., Azmi, M.H., Alashbi, A., Daradkeh, Y.I. and Alhammadi, A., 2023. 6G mobile communication technology: Requirements, targets, applications, challenges, advantages, and opportunities. Alexandria Engineering Journal, 64, pp.245-274.

[4] Tan, M. and Le, Q., 2019, May. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

[5] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[6] Chakiat, A., Oli, N. and Modi, V.K., 2020, December. Deduplication of Advertisement Assets Using Deep Learning Ensembles. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 889-894). IEEE.

[7] Lowe, G., 2004. Sift-the scale invariant feature transform. Int. J, 2(91-110), p.2.

[8] Bay, H., Ess, A., Tuytelaars, T. and Van Gool, L., 2008. Speeded-up robust features (SURF). Computer vision and image understanding, 110(3), pp.346-359.

[9] Rublee, E., Rabaud, V., Konolige, K. and Bradski, G., 2011, November. ORB: An efficient alternative to SIFT or SURF. In 2011 International conference on computer vision (pp. 2564-2571). Ieee.

[10] Kozat, S.S., Venkatesan, R. and Mihçak, M.K., 2004, October. Robust perceptual image hashing via matrix invariants. In 2004 International Conference on Image Processing, 2004. ICIP'04. (Vol. 5, pp. 3443-3446). IEEE.

[11] Chopra, S., Hadsell, R. and LeCun, Y., 2005, June. Learning a similarity metric discriminatively, with application to face verification. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (Vol. 1, pp. 539-546). IEEE.

[12] Hadsell, R., Chopra, S. and LeCun, Y., 2006, June. Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 2, pp. 1735-1742). IEEE.

[13] Hu, J., Lu, J. and Tan, Y.P., 2014. Discriminative deep metric learning for face verification in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1875-1882).

[14] Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B. and Wu, Y., 2014. Learning fine-grained image similarity with deep ranking. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1386-1393).

[15] Schroff, F., Kalenichenko, D. and Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).

[16] Hoffer, E. and Ailon, N., 2015. Deep metric learning using triplet network. In Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3 (pp. 84-92). Springer International Publishing.

[17] M. Douze, G. Tolias, E. Pizzi, Z. Papakipos, L. Chanussot, F. Radenovic, T. Jenicek, M. Maximov, L. Leal-Taixe, I. Elezi, O. Chum, C. C. Ferrer, “The 2021 Image Similarity Dataset and Challenge,” arXiv:2106.09672 [cs.CV]

[18] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[19] Papakipos, Z., Tolias, G., Jenicek, T., Pizzi, E., Yokoo, S., Wang, W., Sun, Y., Zhang, W., Yang, Y., Addicam, S. and Papadakis, S.M., 2022, July. Results and findings of the 2021 Image Similarity Challenge. In NeurIPS 2021 Competitions and Demonstrations Track (pp. 1-12). PMLR.

[20] Radenović, F., Tolias, G. and Chum, O., 2018. Fine-tuning CNN image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence, 41(7), pp.1655-1668.

[21] Wang, W., Zhao, F., Liao, S. and Shao, L., 2022. Attentive WaveBlock: Complementarity-enhanced mutual networks for unsupervised domain adaptation in person re-identification and beyond. IEEE Transactions on Image Processing, 31, pp.1532-1544.

[22] Sarlin, P.E., DeTone, D., Malisiewicz, T. and Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4938-4947).

[23] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

[24] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[25] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[26] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695)

[27] Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114..

[28] Zhang, L., Rao, A. and Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3836-3847).