Explainable Digital Creatives Performance Monitoring using Deep Feature Attribution

Varun Dutt[1,2], Demetris Hadjigeorgiou[1], Lucas Galan[1], Faiyaz Doctor[2], Lina Barakat[2], Kate Isaacs[1]

[1]*Rapp Limited, London, UK*
[2]*School of Computer Science and Electronic Engineering, The University of Essex, Colchester, United Kingdom*

{varun.dutt, lucas.galan, demetris.hadjigeorgiou, kate.isaacs}@rapp.com, {fdocto, lina.barakat}@essex.ac.uk

***Abstract*—A key challenge in marketing and advertising research is understanding when and why digital assets such as promotional content perform well during a marketing push. By leveraging raw image feature vectors extracted from large datasets, we can train performance prediction models using online social signals such as likes or views. While the resulting models make accurate predictions, they are opaque and rely on abstract features within the model, making attribution almost impossible. This paper demonstrates an approach to performance prediction modelling for image based digital creative assets. Utilising a combination of pre-trained vision model embeddings with a pipeline of generative Artificial Intelligence (AI) for image synthesis and manipulation, we establish a means of determining the performance of explainable components. This enables flexible performance prediction, even with smaller datasets, with high degree of explainability through the attribution of image features correlating with high or low performance.**

***Index Terms*—Deep Feature Extractors, Performance Analyses Attribution Pipeline, Transformers, Diffusion Models, Digital Creatives**

I. Introduction

In marketing and advertising research, creating content that will resonate with consumers is critical, but predicting the potential performance of marketing assets is a complex problem. This requires a careful understanding of the individual factors contributing to consumer response, not just predicting what will perform well, but why. The ideal system would be able to:

1) understand what characteristics are in an image (henceforth referred to as an asset),
2) accurately predict its performance (as measured in likes, clicks, views, etc.),
3) be interpretable, allowing the user to attribute performance to specific characteristics in the asset.

Approaches based on neural networks like a Convolutional Neural Net (CNN) [1] or an Multi-layer Perceptron (MLP) [2] have been shown to give state-of-the-art accuracy in performance modelling. These networks can in turn leverage the embeddings of other vision models that are trained on millions of images to create a nuanced prediction of performance. Whilst this increases prediction accuracy, the models behave as black boxes with no feature attribution capability.

If we want to prioritise attribution, we can switch to inputs that are human-readable, such as one-hot encoded tags extracted either manually or by using other AI systems like object detection models. This approach increases explainability as it can provide importance scores for all the input tags from which we can determine the effect of elements in the image on the performance prediction. However, methods based on extraction of tags are restrictive and do not cover subtle elements that might contribute to the prediction and thus inherently less accurate.

Currently, one can either emphasise on prediction accuracy, or explainability of results. To bridge this gap, we propose a pipeline of various AI systems to leverage neural networks and generative models to achieve state-of-the-art performance prediction while also creating a flexible and accurate system for attribution.

The main contribution of this work is a system of AI systems that allows for performance attribution from input vectors that were produced in abstract space without compromising on predictive accuracy. The first part of this system uses a pre-trained CNN model backbone to extract feature vectors from images being used in advertisements. The feature vectors are then passed as inputs to a XGBoost model to predict the target performance variable (likes, clicks etc.). We then utilise a pipeline (henceforth referred to as ANVIL) which contains generative models that modify the image according to specific patterns, creating variations on the original image by either removing or adding components. These new images are generated in line with whichever component we are specifically testing for attribution. The altered assets are then fed to the prediction pipeline mentioned earlier to get a new prediction. By carefully and consistently generating variations of the initial asset, we are able to understand which modified components are affecting performance and by how much.

This paper is structured as follows. Section II outlines related work. Section III details the proposed pipeline, including feature extraction, predictive model, and performance attribution. We then provide an analysis of the results in Section IV. The paper ends with conclusions and a discussion on future work in Section V.

II. Related Work

The following is a review of Machine learning (ML) approaches that offer the ability to perform both prediction and some degree of feature attribution.

eXtreme Gradient Boosting (XGBoost) [3] is a form of gradient boosting algorithm with a much more advanced implementation. It works by sequentially adding simple models to correct the errors made by previous models and is optimized to have high computational efficiency. The model is based on tree algorithm and hence has the ability to assign importance scores to the input variable which can be used in attribution. However, for the importance scores to be usable, input features need to be human-interpretable variables.

An ensemble learning technique called random forest [4] combines the predictions from several decision trees to generate a single, more reliable forecast. This kind of supervised learning technique is applicable to tasks involving both regression and classification. The algorithm uses techniques like Bootstrap and Aggregation, commonly known as bagging that enables it to use multiple decision trees. Each decision tree has a significant variance, but when we aggregate them all at once, the variance that results is reduced because each tree is perfectly trained on a specific sample of the data. Similar to XGBoost random forest also provides importance scores for input variables but with the same drawback of these input variables have to human interpretability to be useful.

Gradient-weighted Class Activation Mapping (Grad-CAM) [5], is a gradient mapping technique that uses the gradients of any target concept (say 'dog' in a classification network or likes on an Instagram in regression network) flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.

Methods to make tree-based models interpretable have been proposed by [6] and [7]. Whilst these methods would give an indication of the features that contribute the most to performance, these features need to be human-readable for the methods to work. In the case of images, the inputs to the tree models above are feature vectors and hence not easily interpretable. Additionally, ensemble tree-based models rely on multiple "weak learners" for their predictive accuracy and as such can identify combinations of features that lead to near-perfect predictions. Attribution, by definition, requires a small number of individual predictors and as such becomes a more difficult problem to tackle [8].

The algorithms mentioned above either require inputs (like object tags) that are extremely reductive or have a very simplistic mechanism for performance attribution as in the case with Grad-Cam. We propose a Pipeline of AI systems that overcome both of these drawbacks to give detailed attribution with a much richer input.

III. Methodology

To demonstrate and test the methodology described herein, we have used images posted on the social media account of a car manufacturer and used the number of likes as the target performance variable. As expected, the vast majority of the images depicted cars in various settings. The dataset consisted of approximately 13,700 images that were shared between October 2012 and March 2024. The average number of likes was non-stationary (i.e., changed over time), showing a steady increase up to 2017 after which it stabilized (see Fig 3). This was likely driven by the growth of the social media platform and of this specific account and not related to the images themselves. To avoid any issues with non-stationarity, any images pre-2017 (c.2.5k images) were not used in training and testing the prediction model. They were only used as part of testing the attribution methodology.

The proposed prediction and attribution pipeline can be subdivided into the following three stages (Fig.4):

A. Feature Extraction

As a first step, we need to extract information from the images. The preferred method of doing so is by removing the prediction layer from pre-trained vision models and using the dense vector from the truncated model as highly condensed image features. There are several vision models with very high generalizing capabilities trained on big datasets like ImageNet [9] that we could have chosen from. Primarily, we considered the choice between multimodal models like CLIP [10] and monomodal models like resnet models. We went with the CLIP architecture as our feature extractor for the following reasons:

1) Comparisons between resnet200d and clip-vit-base-patch32 showed minimal differences in validation error (see Table I);
2) CLIP's ability to project image features and text features in the same space can be useful in interrogating the model in later experiments;
3) The length of the feature vector produced by resnet is 4 times larger than the one produced by CLIP (2048 vs 512) and hence CLIP can be trained on smaller datasets without the risk of overfitting.

The CLIP base model uses a ViT-L/14 [11] Transformer architecture as an image encoder and uses a masked self-attention Transformer as a text encoder. These encoders are trained to maximize the similarity of (image, text) pairs via a contrastive loss function. After extracting features from the CLIP model we use these feature vectors as input to a subsequent model used to predict the performance metrics such as likes, impressions etc.

In addition to the image features, relevant metadata can be added to the image. Whilst our dataset did not contain such metadata, metadata such as country, media channel, and audience selection are often available in marketing and advertising datasets. Our methodology allows these features to be appended to the image features extracted from CLIP.

B. Asset Performance Prediction

The feature vectors are used as inputs to XGboost. As mentioned above, we set the target performance variable as the number of likes that each image received. In order to obtain the best possible results from XGBoost, we normalised and detrended the target variable. The absolute number of likes was normalised based on a twelve-month rolling average, removing the effects of seasonality as much as possible. As discussed above, the non-stationary part of the data was also removed. We also cube-root transformed the target variable, which gives it a more normal-like distribution. This has the effect of reducing skewness and decreasing the impact of outliers on the model. By making this choice, we accept higher

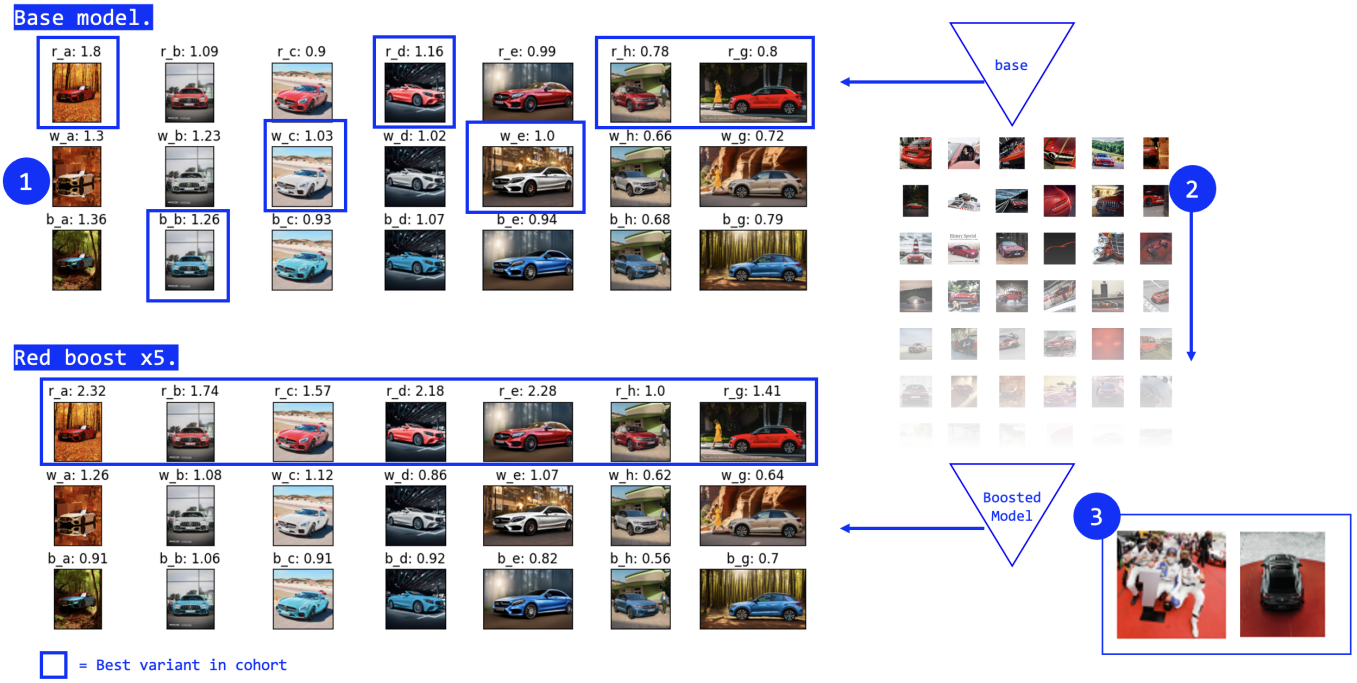

Fig. 1: 1- Each of these elements is tested with multiple images with slight variations a total of 120 times. 2- Using AI identification, red cars in the training set are identified and gradually boost by having the average performance value added in increments. 3- We also extensively tested images with intense red tones (but no red car) to ensure the model had not simply learned to associate red with performance. These images were mostly unchanged in their score.

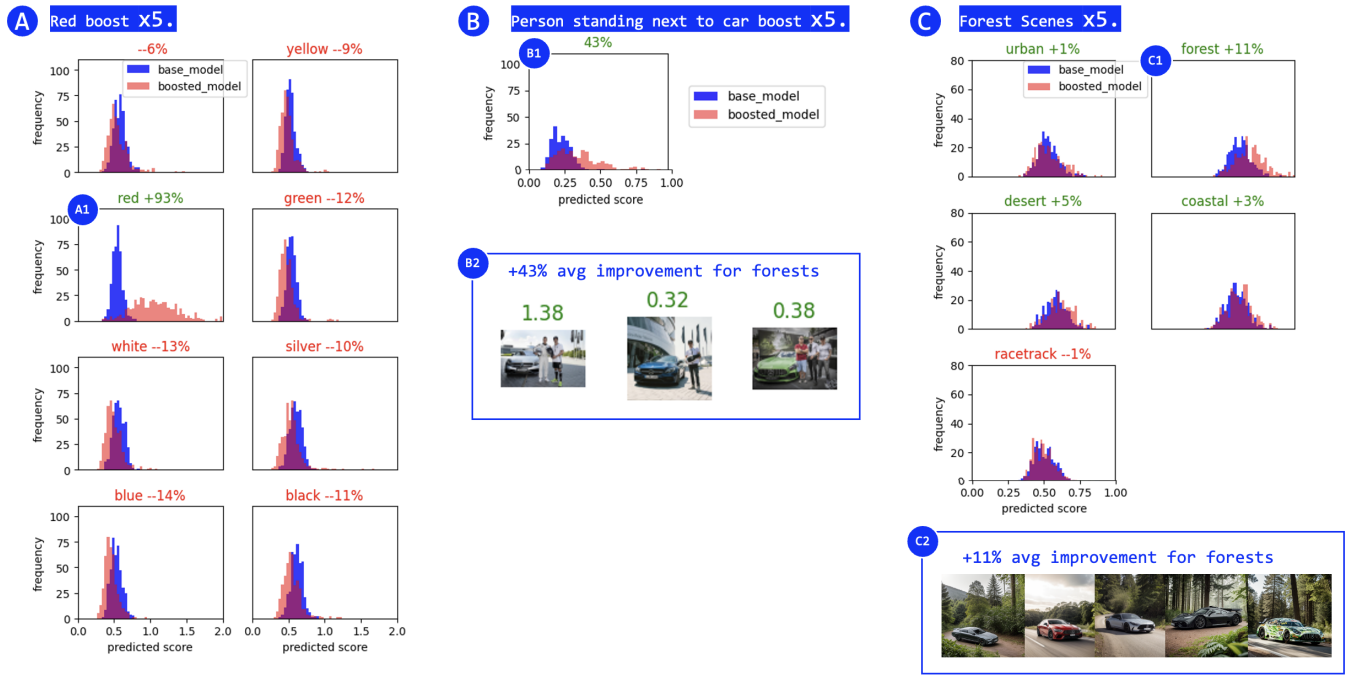

Fig. 2: A diagram showing the testing results for different areas of boosting. A - Our x5 red car boost radically improves the scores of red cars, dragging the scores for any other color down. B- our second test correctly pushes the model to comprehend that people standing next to cars do better, with an average improvement of +43% performance on images of that style. Finally, by consistently testing different backgrounds with different settings, we are able to effectively get the model to understand a bias for different locations, in this case an 11 percent improvement.

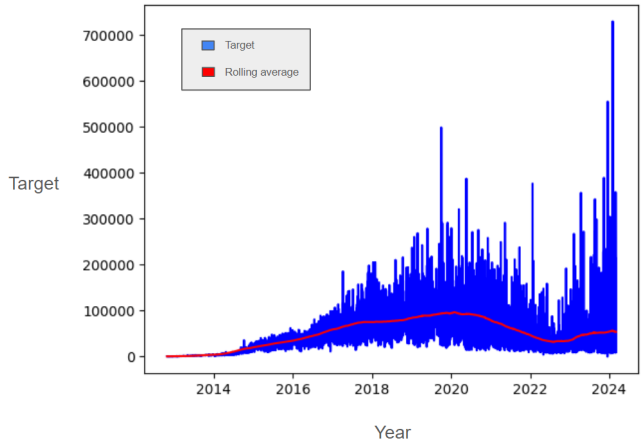

Fig. 3: The target variable (likes) over time showing the ramp-up of performance up to 2017 driven by the adoption of the social media platform and the increase in followers for this account.

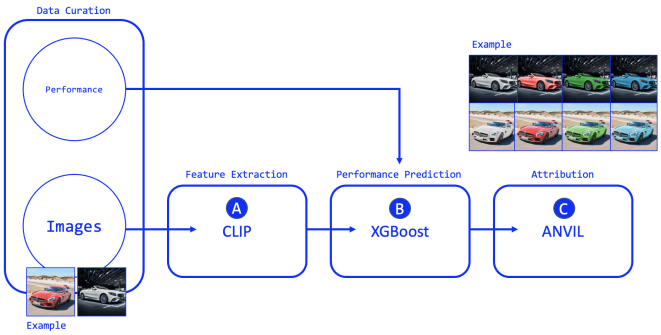

Fig. 4: Block diagram of the proposed approach for prediction and attribution. Attribution can take a number of forms (testing of colours, cars, environments, etc.), the example of vehicle colour is shown here for clarity.

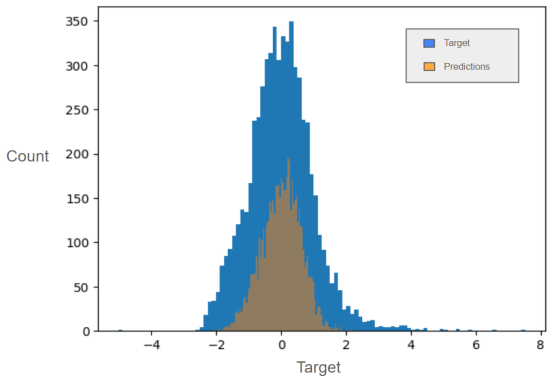

Fig. 5: Overlapping histogram of target variable and predicted values.

errors in the extreme ends of the distribution but better fitting in the bulk of the data. We train the XGBoost model with the following parameters:

1) 1000 estimators
2) Depth of 7
3) Learning rate of 0.1

Table I shows the results of the prediction models evaluated with comprise of two variants of CLIP(clip-vit-base-patch32 and clip-vit-large-patch14) and against resnet200. We chose the clip-vit-base-patch32 (refer Fig 5 for histograms) which gave us the best performance to complexity ratio among the 3 models tested.

C. Asset Performance Attribution (ANVIL)

For the attribution stage we use two models for our ANVIL pipeline to modify elements in the image asset. Stable Diffusion [12] 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 learning to remove the noise in the reverse diffusion process. They are latent variable models referring to a hidden continuous feature space. They look similar to Variational Autoencoders (VAEs) [13], and are loosely based on non-equilibrium thermodynamics.

ControlNet [14] 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.

Our Generative pipeline segments the image asset and isolates different elements of the image, such as the principle object(s): the car, models wearing apparels of a brand etc as well as 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 a transformer model for image segmentation [15]. (2) Then the user selects the element of the image to be modified. (3) The user passes the prompt to the controlnet and stable diffusion model which generate a new image with a modified version of the selected element based on the prompt. Examples of modified images can be seen in Figure 6. ANVIL allows rapid and batch manipulation of images using generative AI to quickly create an attribution dataset. By utilizing fully generated imagery as well as synthetic variations of existing images controlling for a desired attribute, control environments can be generated for testing most conceivable components. For this paper we have focused on the social media output of car manufacturers and as such our performance attribution techniques are more focused on image manipulation of vehicles and scenes, however this technique could be potentially applied to any subject matter.

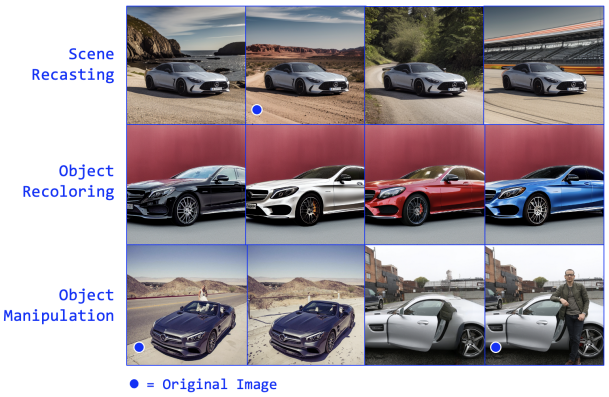

Fig. 6: Rapid and batch manipulation of images using ANVIL generative AI pipeline

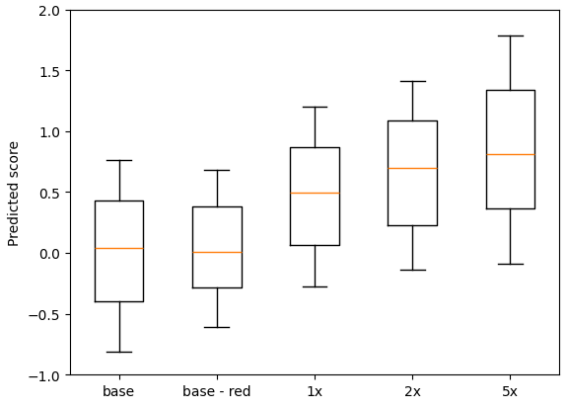

Fig. 7: Predicted likes of 1) base model trained on original dataset 2) base model prediction on only images with red cars 3) predictions of models 1x,2x,5x trained on curated dataset with red cars being boosted by 1x, 2x and 5x the mean of the dataset respectively.

TABLE I: Results from the prediction pipeline.

Models	RMSE	R2 Score	MAE
clip-vit-base-patch32	0.81	0.33	0.58
clip-vit-large-patch14	0.79	0.37	0.56
resnet200d	0.86	0.25	0.62

Generated variants of an image are passed to the XGBoost model, which predicts the performance of each variant. By consistently modifying the same elements, we can explore their impact of the predicted performance. A large number of images can be modified at scale to establish confidence of attribution. Alternatively, since the feature extraction model used is multi-modal, we can pass sentences containing a description of an image to the XGBoost model. This is a faster and more consistent method of interrogating the model.

IV. Results and Discussion

A. Hypothesis A: The model can learn from images

The main hypothesis is that the model can learn from the images provided and we can extract meaningful information. To test this, as it wasn't known what elements were driving performance in this dataset, we artificially inflated the target variable of images with common characteristics. This artificial inflation, henceforth referred to as boosting, was done at different levels, where the level corresponds to adding the mean performance multiplied by a factor to the performance scores of the selected images. The dataset with the boosted samples is used to retrain the model which is then compared against the non-boosted model using unseen images made using ANVIL. The difference between the boosted and non-boosted predictions gives an indication as to whether the model has learned the boosted characteristic.

TABLE II: Characteristics for boosted models.

Characteristic	Variants
Vehicle colour	Red (boosted), white, yellow, silver, blue, black
Setting	Forest (boosted), desert, urban, racetrack
People	People next to car (boosted), cars without people

The characteristics chosen are listed in Table II. Figure 7 and 1 shows the increase in performance we observe from red cars in the validation dataset, showing that the model correctly identified that red cars are expected to do better when we trained it on curated dataset by boosting likes of the images with a red car. We utilise red cars in the example as this is a very easy visual parameter to ascertain and evaluate, but this method can be used to test any other characteristics. In Fig 2 we show two more tests we did, one for determining whether the model could learn about pictures with cars and humans standing around them and a suite of testing for image settings (urban, desert, forest, coastal and racetrack). We can show that even with a relatively small number of boosted images (e.g. less than 1% of the data contained forests) we are able to push the model to comprehend that forests in the background improve performance.

B. Hypothesis B: Image characteristics drive performance

Having confirmed that the model can learn and we can extract that information, we proceeded to identify what was driving performance for this particular dataset. The hypothesis here being that there is some characteristic(s) in these images that viewers consistently like.

By editing images consistently, we explored the effect of presence of people, vehicle colour, vehicle type and setting. Fig. 8 shows the results of this investigation, indicating how the presence of people impacts negatively the performance of an image, users prefer silver and black cars, racetracks and urban settings perform lower than other settings, and finally that users do not like SUVs.

This methodology effectively allows us to perform A/B testing on any permutation of characteristics in an image. By generating more and more images, we are able to increase the confidence in statements such as "Black cars perform better than yellow cars". Through our ANVIL pipeline, hundreds of variants can be created thus allowing us to to investigate accurately which components of the image drive results.

V. Conclusion

In this paper we propose an AI pipeline for predicting the

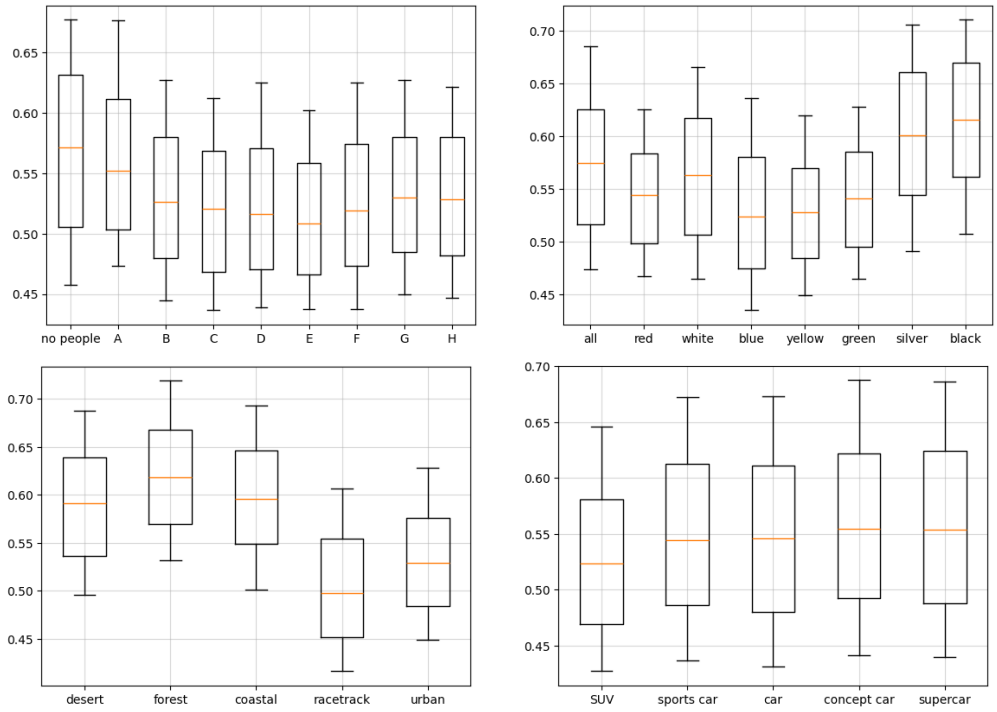

Fig. 8: By generating a large number of synthetic images, we were able to identify the impact of people (where A-H are different positions of people in the image) (top left), vehicle colour (top right), setting (bottom left), and vehicle type (bottom right). For this dataset, users appear to prefer silver and black vehicles, without the presence of people. They also appear to not like SUVs and racetrack settings.

performance of creative advertising image assets. The pipeline uses machine learning, feature extraction and generative AI to predict performance and attribute performance changes image elements through the generative manipulation of image features. The approach which is evaluated on a real-world dataset of images taken from advertising platforms on car brands shows promising results in terms of prediction accuracy and performance explainability. Future work will explore different ML approaches for improving model prediction accuracies as well as developing more intuitive attribution schemes for explaining both image and video based creative assets.

VI. Acknowledgement

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.

References

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

[2] M. Riedmiller and A. Lernen, "Multi layer perceptron," *Machine Learning Lab Special Lecture, University of Freiburg*, vol. 24, 2014.

[3] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[4] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[6] S. Hara and K. Hayashi, "Making tree ensembles interpretable," *arXiv preprint arXiv:1606.05390*, 2016.

[7] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: definitions, methods, and applications," *arXiv preprint arXiv:1901.04592*, 2019.

[8] B. Efron, "Prediction, estimation, and attribution," *International Statistical Review*, vol. 88, pp. S28–S59, 2020.

[9] 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.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

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

[14] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[15] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "Oneformer: One transformer to rule universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2989–2998.