Universiteit Leiden
The Netherlands

Bachelor Computer Science & Datascience and Artificial Intelligence

True Colours: Vision Transformers Classifying Art Beyond Hue

Lars de Rooij

First supervisor:
Hazel Doughty

RESEARCH PROPOSAL BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

July 1, 2026

Abstract

This thesis investigates how colour alterations affect classification performance and internal representations in a pretrained Vision Transformer (ViT) when classifying artworks by artist, style, and genre. Using a ViT-B/16 model with linear probing, experiments are conducted on a filtered subset of the WikiArt dataset. Three colour transformations, namely greyscale conversion, hue shifting, and random colour overlays, are applied to evaluate the model's robustness to changes in chromatic information.

Results show that colour has a task-dependent impact on performance. Genre classification are most robust to colour changes, while artist and style classification are more sensitive. Hue shifts generally preserve the highest accuracy, whereas random colour overlays lead to the largest performance drops. At a class level, some categories remain stable under colour manipulation, while others exhibit significant shifts in both accuracy and embedding space, indicating varying dependence on colour-based cues.

Overall, the findings suggest that colour acts as a secondary, conditional feature in ViT-based art classification. While it can influence predictions and internal representations, structural information such as composition and texture plays an equal or more important role in embedding similarity.

Contents

1 Introduction

Humans are generally robust to colour alterations and can recognize that the images shown in Figure 1 depict the same artwork despite their changes in colour. Art experts may further identify each of them as versions of a painting by Alfred Sisley, and recognize them as an example of Impressionism.

Figure 1: Four Versions of "Indian Summer" [Sis74]. Original on the left.

While computer vision models such as Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) continue to improve, it is not yet clear whether they exhibit the same robustness. Since artists and artistic movements are often characterized by distinctive colour palettes and colour usage, colour is expected to play an important role in tasks such as artist attribution and style classification. However, artworks are also defined by many other characteristics, including composition, brushwork, texture, and subject matter. Understanding the extent to which vision models rely on colour, relative to these other artistic cues, is therefore crucial for interpreting their behaviour. Can these models correctly classify each version as belonging to the same artist? Can they recognize that all images depict the same landscape scene?

More broadly, do these models rely on higher-level artistic characteristics such as composition, proportions, texture, brushwork, or stylistic cues, or are their predictions strongly dependent on the original colour information? Understanding this distinction is essential for evaluating whether vision models learn representations that capture meaningful artistic properties rather than superficial colour statistics. Answering these questions therefore provides insight not only into classification performance, but also into the suitability of modern vision models for art-related tasks such as artist attribution, style recognition, and artwork retrieval.

1.1 Motivation

In visual art like paintings and drawings, colour is often the main way, if not the only way, to portray an image. Diving deeper into art and its many complex styles and genres opens up additional ways to analyze it. If a certain artist has a well-developed signature style, like Picasso and his "Guernica" in Figure 2, it should be recognizable even with potential colour alterations.

Understandably, colours are important in visual arts, but other stylistic choices should carry as much importance. Therefore, classification models should be able to account for these aspects, and classify them correctly even when the variable of colour is altered. In real life these colour alterations may happen due to lighting changes, different camera settings, or other quirks during

Figure 2: “Guernica” [Pic37]

a digitization process. Likewise, many changes in the physical painting can occur, by the colours fading or otherwise changing due to ageing, a yellowing varnish, or restoration. If classification models rely on aspects like brushstroke patterns and compositional layouts, they can focus on structural and stylistic cues that are present regardless of colour.

Additionally, further analysis on how classification models perform on colour alterations can tell us about their internal representations. It can show whether models use colour as a big cue to classifications or rather a minor one, or whether there is a bias of local over global representation. It can show potential shortcuts a model uses, and whether its embedding space is sufficiently abstract. It can indicate the extent to which a model relies on colour and how colour changes influence its internal representations.

1.2 Research Problem

While colour plays a defining role in artistic style and identity, it is important that classification models do not solely base their outputs on chromatic information. Therefore, this study seeks to investigate the following question:

When classifying art, how do colour alterations affect performance and internal representations in a pretrained Vision Transformer?

A Vision Transformer (ViT) is the chosen classification model. Artist, genre and style are the three main classes that will be used for prediction. These categories capture different aspects of an artwork and may therefore depend on colour information to different degrees. Artist classification focuses on identifying the creator of a work, style classification concerns broader artistic movements and visual conventions, and genre classification relates to the depicted subject matter. Since colour may contribute differently to each of these tasks, comparing them provides insight

into which artistic concepts are most sensitive to colour alterations and which can be recognized through other visual cues.

The research is guided by the following sub-questions:

1. To what extent is the ViT invariant to colour alterations?
2. How does classification accuracy change when colour information is changed?
3. Does recolouration disrupt embedding similarity?
4. Is colour more important for predicting artist, style, or genre?
5. Are some classes of artists, genres, or styles more dependent on colour than others?

The sub-questions examine different aspects of colour dependence in ViT-based artwork classification. Specifically, they investigate the robustness of the model to colour transformations, the effect of altered or removed colour information on classification performance, and the extent to which colour contributes to embedding similarity. Embedding similarity allows us to determine whether colour alterations change the model's internal representation even when the predicted class remains unchanged. Additionally, the sub-questions explore whether colour importance differs across classification tasks and between specific artists, genres, or styles.

1.3 Thesis Structure

The following chapters attempt to answer these questions. In Chapter 2 the necessary background is discussed regarding image classification models, the ViT in particular, and previous research in colour within art classification. In Chapter 3 the methodology is described, including how the research tests are conducted. Chapter 4 describes the results based on the research.

The results of the tests are further analyzed in Chapter 5, in which they are applied to answer the research questions. Lastly, possible Limitations and possibilities for future work are discussed.

Additional material such as full dataset distributions and names, can be found in Appendices. Full sources referenced in the text, including papers and works of art, are listed in References.

2 Background

2.1 Colour in Art Classification

Colour is widely considered an important characteristic in both artistic creation and artwork classification [Fai13]. Artworks are commonly described through stylistic elements such as colour, texture, form, line, and composition, alongside physical attributes such as brushstrokes [Fic07]. Research has shown that artistic genres can often be distinguished through features including colour palettes, stroke styles, colour mixing, gradients, and other visual techniques [Zuj+09]. Likewise, artist identification is frequently associated with recognizing an artist's characteristic use of colour, light, texture, and composition [CG16].

However, colour is not the only cue used for visual recognition. Human perception integrates both colour and texture information when interpreting images [IW11]. Furthermore, previous work has found that preserving additional colour information does not necessarily improve classification accuracy [TCL05]. These findings suggest that successful artwork classification may also depend on other stylistic and semantic features.

2.2 Robustness to Colour Alterations and Domain Shifts

Deep learning models can experience reduced performance when evaluated on data that differs from the training distribution, known as domain shift. In image classification, domain shifts may arise from changes in illumination, colour balance, image quality, or visual style [Wan+24]. Colour alterations are a common form of domain shift. Previous work has shown that CNNs often rely on texture and colour-related cues, making them sensitive to changes in image appearance [Gei+22]. To evaluate robustness, studies commonly apply perturbations such as greyscale conversion, hue shifts, saturation changes, and colour jittering.

Vision Transformers have been reported to exhibit greater robustness than CNNs under several image corruptions and distribution shifts, although performance degradation can still occur depending on the perturbation [PC21; Bho+21]. These findings suggest that colour alterations may influence both classification performance and internal feature representations.

2.3 Deep Learning for Image Classification

Deep learning has become the dominant approach for image classification. By using multiple layers of nonlinear processing and optimizing their parameters through backpropagation, deep learning models can automatically learn hierarchical feature representations from large datasets [LBH15].

Among deep learning architectures, Convolutional Neural Networks (CNNs) have long been the standard for image recognition tasks. Designed specifically to handle the spatial structure of images, CNNs were shown to outperform previous approaches in visual recognition [Lec+98]. Their effectiveness was further demonstrated by record-breaking performance on large-scale benchmarks such as ImageNet, establishing CNNs as the leading architecture for image classification,

recognition, and detection tasks [KSH17; RW17].

Despite their success, CNNs have several limitations. Convolutional operations process images locally and can struggle to capture relationships between distant image regions. Furthermore, they treat all image locations similarly regardless of their importance to a specific task [Wu+20]. These limitations motivated the development of transformer-based vision models, which represent images as visual tokens and explicitly model relationships between them through self-attention mechanisms.

2.4 Vision Transformers

Transformers were originally introduced for natural language processing as an architecture based entirely on self-attention mechanisms, eliminating the need for recurrent or convolutional operations [Vas+17]. The Vision Transformer (ViT) applies a transformer directly to images by dividing an image into a sequence of fixed-size patches, which are treated similarly to tokens in natural language processing. These patch embeddings are then processed through self-attention layers that model relationships between image regions [Dos+20]. Unlike CNNs, which rely on convolutional operations and strong vision-specific inductive biases, ViTs learn visual representations primarily through attention mechanisms.

When pre-trained on large datasets, ViTs achieve image classification performance comparable to or better than state-of-the-art CNNs while requiring fewer vision-specific assumptions [Dos+20]. Due to their strong performance and ability to capture global relationships within images, Vision Transformers have become an important architecture in computer vision research and applications [Han+22]. As a result, a ViT is used in this study to investigate how colour alterations affect artwork classification and internal visual representations.

2.4.1 Embeddings

Vision Transformers (ViTs) have been shown to produce strong embedded representations when applied to image classification tasks. When pre-trained on large-scale datasets and transferred to downstream benchmarks, ViTs achieve performance comparable to or exceeding state-of-the-art CNNs, while relying less on vision-specific inductive biases [Dos+20].

Beyond classification performance, research has investigated how ViTs represent visual information internally. Their ability to attend globally across image patches raises questions about robustness to common image variations such as occlusions, domain shifts, and spatial perturbations [Nas+21]. Compared to convolutional networks, ViTs exhibit different representation dynamics, including more uniform feature representations across layers [Rag+21].

Furthermore, studies suggest that self-supervised ViT features can encode meaningful embedded structure, including information related to object-level segmentation, which is less pronounced in supervised ViTs and CNN-based models [Car+21]. These findings indicate that ViTs may develop more abstract embeddings, making them suitable for studying how visual properties such as colour influence internal representations.

3 Methodology

This thesis aims to analyze how the performance of a vision transformer model changes when met with colour manipulations of artworks, specifically in regards to classifying artists, styles, and genres. Each of these three labels will get their own model, as they predict different things about an artwork. Each of the model's architecture and setup is the same.

The main paradigm used for the models is a pretrained ViT [DBK+21; Dos+20], the outputs of which get linearly probed over 1 layer, to get the classifications. With these models, the recoloured versions are given as input, and the new classifications for each version gets recorded. With this, the average accuracy and per-class accuracy can be used to compare the performance of 3 different colourations, on each 3 of the labels.

This approach answers the research question by using only the recolourations as variable, and having several different metrics to see how this changes the representation. ViTs capture global relationships through self-attention, making them well suited for investigating whether artwork representations remain stable under colour alterations. WikiArt was selected because it is one of the largest publicly available art datasets and provides artist, style, and genre annotations for each artwork. This makes it suitable for studying how colour manipulations affect different classification tasks. The filtering criteria were introduced to reduce class imbalance and ensure that each class contained sufficient examples for reliable evaluation.

3.1 Dataset

The dataset used is Wikiart [Hug21]. The full dataset contains 81.444 pieces of visual art, taken from Wikiart.org [Wik26]. Each of these entries has a .jpg file, with associated artist, genre, and style. These respectively classify who made the work, what the work depicts, and what distinctive visual elements or techniques were applied. The dataset has 129 artist classes, including a "Unknown Artist" class, 11 genre classes, including an "Unknown Genre" class, and 27 style classes. The naming and direct mapping of class number to actual class name, can be found in Appendix A

Some filtering of the original dataset was applied, such that for training and evaluation, each style class had more than 500 entries and each artist class had more than 100 entries. This was done so that classes with respectively low amounts of data would not need to be learned by any models. The style class now has 22 classes with the smallest taking up 0.65%, and there are 118 artist classes, with the smallest taking up 0.13%.

Additionally, artist class 0, "Unknown Artist" heavily skewed the distribution by being the artist class that contained 51.62% of the artworks, and thus was excluded entirely when training and evaluating the artist model. Entries that had artist class 0 were not removed from the filtered dataset entirely, as they were still used to train and evaluate the style and genre model. The total distribution per class per label can be found in Appendix B

3.2 Model Setup

The ViT [DBK+21] requires a normalized input of 224 × 224 pixels. Therefore each of the entries in the filtered dataset is resized, and then normalized with the ImageNet normalization [Den+09]

$$\text{mean} = [0.485,0.456,0.406] \mid \text{std} = [0.229,0.224,0.225]$$

ImageNet normalization was used because the pretrained ViT was originally trained on ImageNet images, and matching the preprocessing used during pretraining ensures compatibility with the learned feature representations.

3.2.1 ViT-B/16

With that, the data can be fed into the frozen ViT. The last layer, which has the final classification, is removed from the ViT, and the output of the last hidden layer is used for classification.

The ViT backbone remains frozen and only a linear classifier is trained. This linear probing setup was chosen because it allows the experiments to evaluate the representations already learned by the pretrained model, rather than representations adapted through task-specific fine-tuning. As a result, observed changes in performance can be more directly attributed to the effect of the colour manipulations on the pretrained representations.

The output of the final hidden layer consists of a CLS token and 196 image patch tokens, each represented by a 768-dimensional embedding. For classification, only the CLS token is retained and stored as a 768-dimensional feature vector for each artwork. The CLS token is designed to provide a global representation of the image and is commonly used for downstream classification tasks.

3.2.2 Linear Probing Classifiers

Using the last hidden layer output of the ViT, for each classification task a single linear layer is trained to make a model that predicts its classes. The training of each of these linear layers used the following parameters:

epochs: 25 | batch size: 512 | learning rate: 1e-3

criterion: CrossEntropyLoss | optimizer: AdamW

These hyperparameters were selected based on preliminary experiments, and provide stable convergence across all models.

With those parameters, the training progress proceeds. For each of the labels, a split is made that preserves the class distributions of Appendix B. 70% of the data was in the training set, 15% in the test set, and 15% in the validation set. Each split contained the following number of samples:

Train size: 55.619 | Test size: 11.919 | Val size: 11.919

Notably, since artist class 0, “Unknown artist” takes up 51.62% of the data as can be seen in Appendix B, it is omitted during training, validation, and testing. The artist model does not predict class 0, and neither does it use it for any of its results.

The training process is the following: For each model, the number of classes gets computed. With this, and their respective distributions, the training is started. Each epoch, the model is updated based on results from the training set, and validated with non-training data of the validation set to get a per-epoch accuracy. When the average accuracy of the validation has improved compared to the previous best, the state of the model updates to the current best. ”After all epochs have completed the test set is used to get a final average accuracy.

To reduce the influence of random initialization, each model was trained using multiple random seeds, namely;

$$0, 10, 20, 30, 40$$

The model that performed with the highest validation accuracy across all epochs and seeds is selected to be the final one.

3.3 Colour Manipulation Methods

With the models created, the last thing to do before experiments can begin are the colour-alterations. The following three methods are applied to all images in the filtered dataset, after having been converted to a tensor. The selected colour manipulations were chosen to represent three distinct ways in which the colour information of an image can be altered. Together, they allow the investigation of model behaviour under the removal of colour information, the transformation of colour information while preserving colour relationships, and the addition of a global colour bias. By covering these different forms of colour alteration, the experiments can distinguish between reliance on colour itself, reliance on specific colours, and robustness to broader colour distortions. The transformations happen on the original full images, after which they get converted to the $224 \times 224 \times 3$ format and normalized, so that they are correctly formatted for testing.

Figure 3 shows an example of an unaltered data entry.

Figure 3: Original Version of ”Sacred Himalayas” [Roe34]

The following manipulations span a range from complete colour removal to moderate and strong colour transformations. Additional colour augmentations, such as saturation adjustments, brightness changes, or contrast modifications, were not included because they primarily affect image intensity rather than colour identity. Restricting the study to these three manipulations allows the experiments to focus specifically on the role of colour information and its influence on classification performance.

3.3.1 Greyscale

Greyscale was chosen because it completely removes chromatic information while preserving luminance, structure, texture, and composition. This allows the contribution of colour information to be isolated from other visual features used by the model. If classification performance remains largely unchanged after converting artworks to greyscale, this suggests that the model primarily relies on non-colour cues. Conversely, a substantial decrease in performance would indicate that colour plays an important role in the learned representations.

Images were converted to greyscale using a weighted luminance conversion, in which the greyscale intensity is computed as a weighted sum of the red, green, and blue channels to approximate human visual sensitivity. This was implemented using Pillow's Image.convert("L") method, which applies the ITU-R BT.601 luma transformation and collapses the RGB image into a single luminance channel [Uni11; Con26]. Since the model input expects 3 channels, this channel gets duplicated thrice, the same greyscale representation in each of them.

Figure 4 shows an example of a greyscale alteration.

Figure 4: Greyscale Version of "Sacred Himalayas" [Roe34]

3.3.2 Hue Shift

Hue shifting was selected because it alters the absolute colours present in an image while preserving the relative relationships between colours. Unlike greyscale conversion, colour information remains available to the model, but the specific hues no longer correspond to those in the original artwork. This manipulation allows the investigation of whether the model relies on particular colours associated with artists, styles, or genres, or whether it is robust to systematic changes in colour identity.

This is done by taking a pixel, locating it on a colour wheel, and then rotating the wheel with a certain degree such that the hue gets shifted. When done on all pixels, with the same degree of rotation, it can make each originally blue pixel red, each red pixel green, and each green pixel blue. During this process, the brightness and contrast is preserved [Fai13].

The degree of rotation that gets applied onto an artwork is determined by a hash of its file name. Each file name gets its own 32-bit hash, which then gets used as a seed to randomly sample a degree from a continuous uniform distribution between 20 and 340. This ensures that the degree of rotation across all entries is evenly distributed. Additionally, a rotation between 20 and 340 degrees prevents the negligible difference a per example 2 degree rotation would make. With low rotations, like those in the 0-20 and 340-360 degrees range, the image would not have a significant hue shift. Based on visual observation, those with a degree shift of 20 degrees or lower look near identical, and thus these are excluded.

Figure 5 shows an example of a hue shift alteration, with degree 112.7641579.

Figure 5: Hue-shifted Version of ”Sacred Himalayas” [Roe34]

3.3.3 Random Overlay

A random colour overlay was chosen because it introduces a global colour cast across the entire image. Unlike hue shifting, which preserves colour relationships, the overlay changes the balance between colours and can partially obscure the original palette. This manipulation therefore represents a stronger form of colour distortion and allows the evaluation of model robustness when colour information is altered in a way that affects both individual colours and their relationships within the artwork.

To change the global colour, applying the same transformation to each pixel, a random overlay is applied. An RGBA value, a colour that has some opacity, is applied to each pixel and the new result is the average of both. This makes the whole image look like it has a 1-colour filter over it. During this process, the image's brightness and contrast may be reduced due to the blending of the original colours with the overlay.

The RGBA that gets applied onto an artwork is determined by a hash of the file name. Each file name gets its own 32-bit hash, which then gets used as a seed to randomly sample each of the red, green, blue and alpha values. The RGB values are sampled from a discrete uniform distribution between 0 and 255, while the A value is sampled from a continuous uniform distribution between 0.4 and 0.8. These values were selected based on visual observations, anything less than 0.4, on

average, not altering the image distinctively enough, while an opacity of more than 0.8 distorts too much colour information.

Figure 6 shows an example of a random recoloured alteration, with RGBA (148, 191, 57, 0.7229028142)

Figure 6: Random Overlay Version of "Sacred Himalayas" [Roe34]

3.4 Design of Experiments

With these images and models set up, testing can be done. The recoloured versions of all images from both the validation and test sets are evaluated. Since model selection has already been completed and no further training or hyperparameter tuning is performed, both sets can be used for analysis. This increases the amount of unseen data available for evaluating the effects of the colour manipulations. It gives a list of predictions per model, per label. Each of these outputs has not been used in any training of the original models. Therefore, these baseline results give an accurate representation of average accuracy per variation.

Aside from classification outputs of the model, the CLS embedding outputs from the frozen ViT will also be computed for all variants. Analyzing the CLS embeddings allows changes in internal representations to be studied independently of classification accuracy. This makes it possible to determine whether colour manipulations alter the learned feature space even when the final prediction remains unchanged. Additionally, the difference between the original input and its recoloured versions can be computed. This shows how the representation changes throughout the architecture of the ViT.

Finally, the classification space for each model gets projected into 2D. Projecting the classification space into two dimensions provides an interpretable visualization of how recoloured artworks move within the learned feature space, complementing the quantitative accuracy and embedding-distance analyses. With this space, the locations of the recoloured versions can be plotted, and the difference between the baseline and its three corresponding recolourations can be analyzed.

Together, these methods give a grand overview on both internal model representation and its outputs, based on the variable of a colour change.

3.5 Model Comparison

To compare the ViT against a different architecture, a CNN model was also trained. This model uses a pretrained ResNet-50 v1.5 backbone [He+15; Mic24]. ResNet-50 was selected because it is a widely used and well-established CNN architecture, making it a suitable baseline for comparison against the ViT.

The CNN training setup was identical to the ViT configuration described in Chapter 3.2. Both models used the same dataset splits, input normalization, batch size, learning rate, optimizer, loss function, number of epochs, random seeds, and colour manipulation procedures. The ResNet-50 backbone was kept frozen during training to ensure a consistent comparison under fixed feature extraction.

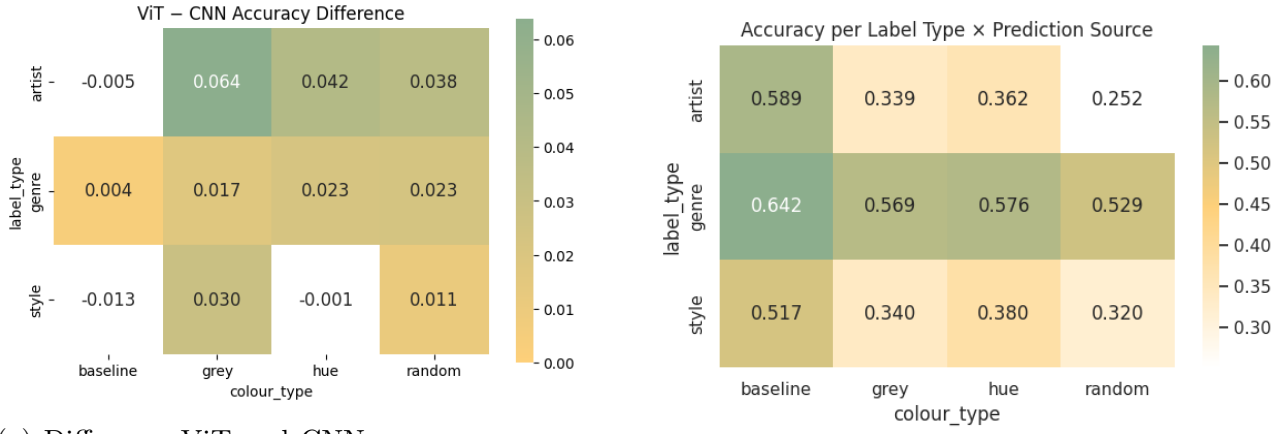
This comparison evaluates whether the effects of colour manipulations are specific to transformer-based architectures or whether similar patterns also occur in convolutional networks. Since CNNs and ViTs process visual information differently, this provides additional context for interpreting model behaviour under colour transformations. Additionally, this experiment is designed as a controlled architectural comparison under identical training conditions.

4 Effect of Colour on Classification Performance

This chapter presents the main experimental results of the study, focusing on how different types of colour alterations influence classification performance and embedding shifts. It first compares overall accuracy across models and transformation types, before analysing class-level behaviour and changes in embedding space. All intermediate analyses are presented using class indices only. Chapter 4.7 maps these indices back to the corresponding art classes to provide a more interpretable discussion of the results.

4.1 Average accuracy comparison

Figure 7a presents the average classification accuracy of the ViT and CNN across twelve experimental conditions. Across these conditions, the CNN achieves higher accuracy in three cases, while the ViT performs better in nine cases.



(a) Difference ViT and CNN on average accuracy

(b) ViT

Figure 7: Average Accuracy

Although the differences between both architectures are relatively small, the ViT shows a more consistent performance advantage across most settings. Based on these results, the subsequent analyses focus on the ViT.

Figure 7b provides a summary of the baseline average accuracies for the ViT model. This figure is discussed in conjunction with model- and class-specific results in the following sections.

4.2 Artist Model

Of the three recolouration methods, the artist model shows the lowest accuracy under random recolouration. Notably, this combination of artist classes and random recolourations yields the lowest accuracy across all average accuracies in Figure 7b, indicating strong sensitivity to stochastic colour perturbations.

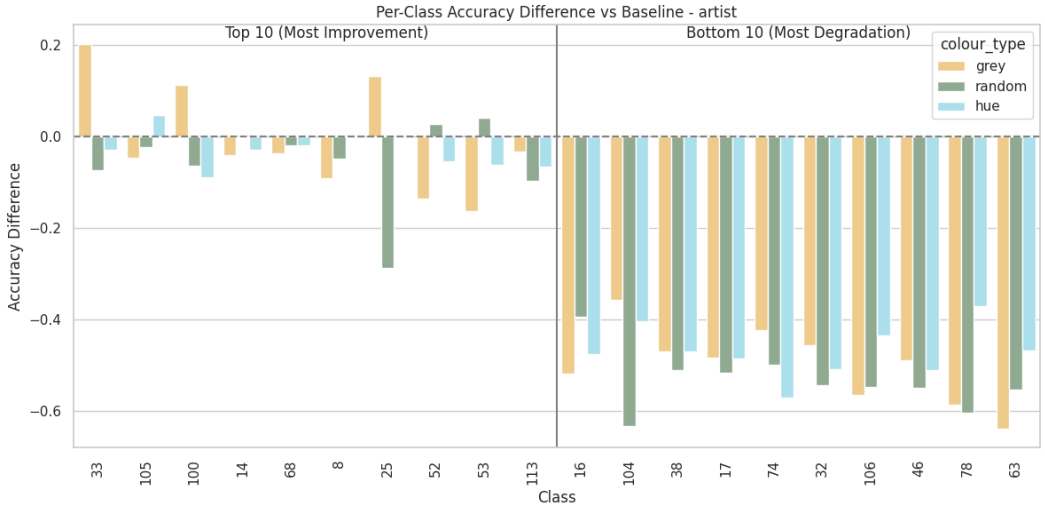


Figure 8: Artist recolouration accuracy per class, compared to baseline

In Figure 8, the per-class average accuracy per recolouration method is compared to the baseline results using the testing and validation subsets of the filtered dataset. The results show that only a small subset of classes benefit from recolouration, with improvements distributed sparsely across grey, random, and hue transformations (three, two, and one classes respectively). Notably, no class improves consistently across multiple recolouration types, suggesting that performance gains are highly condition-specific rather than systematic.

In contrast, degradation is more consistent, particularly under random recolouration, which accounts for most of the lowest-performing cases. Grey recolouration also contributes to several strong performance drops, while hue shifts are comparatively less detrimental but still class-dependent. Overall, these results indicate that recolouration primarily introduces instability rather than consistent feature improvement in the artist classification task.

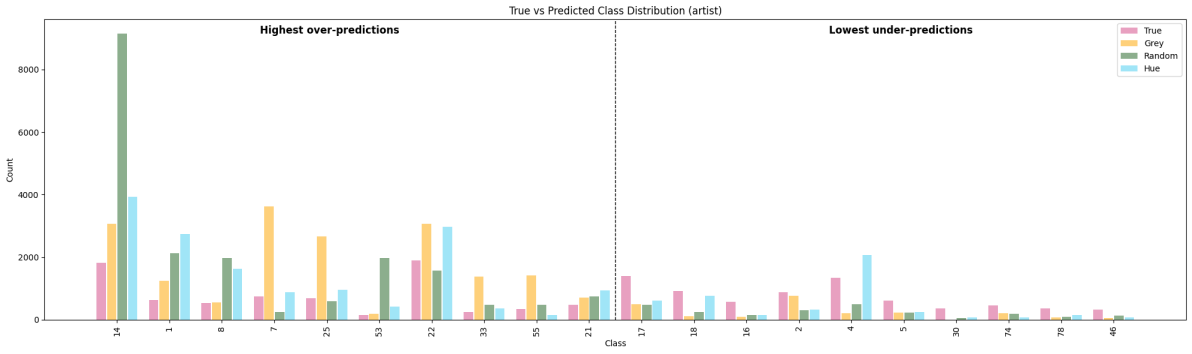


Figure 9: Artist prediction distribution per class, compared to true distribution

In Figure 9, the per-class prediction totals are compared to the true distribution using the full filtered dataset, excluding artist class 0, the “Unknown Artist”. Across recolouration types, systematic overprediction is concentrated in a small number of classes, particularly under random and grey transformations. This suggests that recolouration does not only affect classification

accuracy but also shifts learned class priors, resulting in biased prediction tendencies for specific classes.

Underprediction is less structurally consistent, indicating that recolouration primarily increases false positive rates rather than uniformly suppressing predictions. While some bottom-10 classes move closer to the true distribution than extreme overprediction cases, the overall pattern remains dominated by class-specific instability rather than uniform error propagation.

Figure 8 and Figure 9 together cover only 20 classes out of the 119 in the filtered dataset, so the most relevant observations are those appearing in both or only one figure. Class 33 shows the highest accuracy under grey recolouration but also a strong overprediction bias. Classes 105 and 100 show improved accuracy without corresponding overprediction, suggesting more stable feature adaptation. In contrast, classes 14 and 8 are consistently overpredicted across multiple recolouration types, particularly under random transformations, without corresponding accuracy gains. Class 25 appears in both figures, showing both improved accuracy and increased prediction bias under grey recolouration. Class 53 shows a similar pattern under random recolouration but with only minor accuracy improvement.

Among underpredicted classes, only classes 16, 17, 74, 78, and 46 appear consistently in the bottom-10 accuracy group compared to the baseline. Notably, the class with the largest accuracy degradation (class 63) does not appear in the bottom-10 underprediction group, suggesting that its error is distributed rather than concentrated.

Finally, Table 1 reports total true positives, false positives, and precision across recolouration types using the full filtered dataset, excluding artist class 0. Hue-shifted images achieve the highest precision, followed closely by grey recolouration, while random recolouration performs substantially worse. The reduction in true positives under random transformations is significantly larger than the difference between grey and hue conditions, indicating that stochastic colour perturbations disrupt discriminative features more strongly than structured transformations.

Artist	TP	FP	Precision
grey	15391	23052	40.0 %
hue	16226	22217	42.2 %
random	10602	27841	27.6 %

Table 1: Model performance Artist Total

4.3 Style Model

As seen in Figure 7b, the main difference between the artist and style models is that random recolouration performs relatively better for the style model. However, it still represents the lowest-performing transformation overall, indicating that stochastic colour changes consistently degrade performance even in this setting.

In Figure 10, the per-class average accuracy per recolouration method is compared to the baseline using the testing and validation subsets of the filtered dataset. The results show that a limited

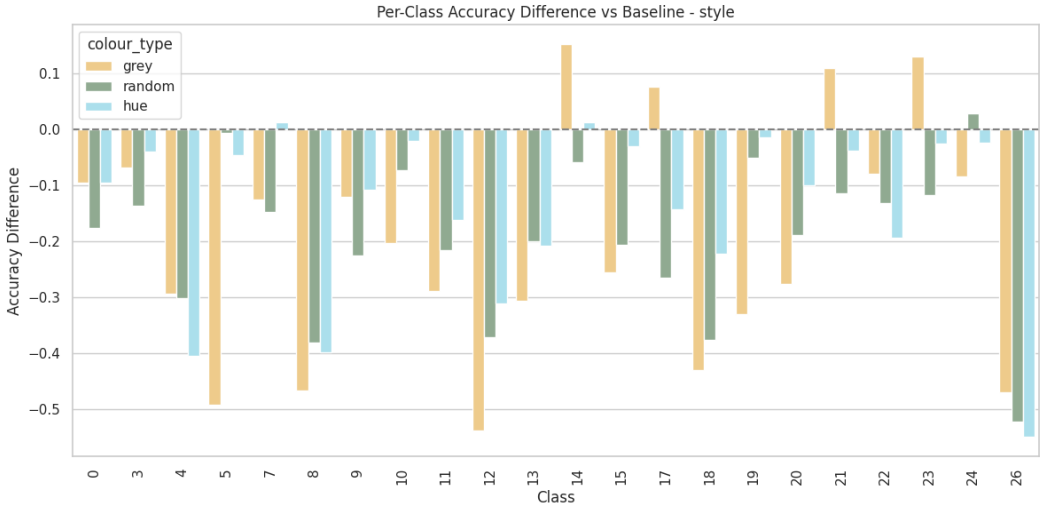


Figure 10: Style recolouration accuracy per class, compared to baseline

number of classes benefit from recolouration, with improvements distributed across grey, hue, and random transformations (four, two, and one class respectively). Notably, class 14 improves under both grey and hue transformations, suggesting some robustness to structured colour shifts.

In contrast, degradation is more widespread, with grey recolouration accounting for the largest number of lowest-performing cases, followed by random and then hue transformations. Class 26 exhibits the most substantial overall drop in accuracy across all recolouration types, indicating a particularly strong sensitivity to colour alteration in this class. Overall, the results suggest that performance changes are primarily driven by class-specific instability rather than uniform model behaviour across all categories.

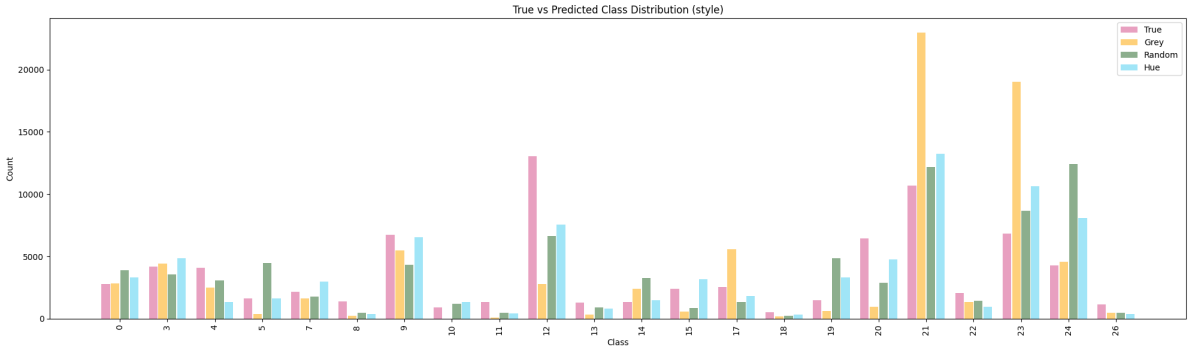


Figure 11: Style prediction distribution per class, compared to true distribution

In Figure 11, predicted class distributions are compared to the true distribution using the full filtered dataset. Overprediction is concentrated in a small subset of classes, particularly under random and grey recolouration. This indicates that recolouration affects not only classification accuracy but also shifts the model's learned class priors, resulting in biased prediction tendencies toward specific classes.

Underprediction is more diffuse but consistent for a few classes, particularly those that are

systematically underrepresented across all transformations (notably classes 8, 11, and 26). This suggests that recolouration primarily amplifies existing weaknesses in class representation rather than introducing entirely new failure modes.

Comparing Figure 10 and Figure 11, a key pattern emerges: some classes (e.g. 17, 21, and 23) achieve higher accuracy despite being overpredicted, indicating that improved classification confidence does not necessarily align with balanced prediction distributions. In contrast, class 14 shows improved accuracy under grey recolouration without a corresponding strong overprediction signal, suggesting more stable feature adaptation under this transformation.

Style	TP	FP	Precision
grey	28206	51251	35.5 %
hue	31500	47957	39.6 %
random	26088	53369	32.8 %

Table 2: Model performance Style Total

Finally, Table 2 reports total true positives, false positives, and precision across recolouration types using the full filtered dataset. Hue-shifted images achieve the highest precision, followed closely by grey recolouration. Random recolouration performs worst in terms of precision, although its true positive count is not substantially lower than that of grey transformations, with a difference of 2,118. This suggests that the primary degradation arises from an increase in false positives rather than a collapse in correct predictions.

4.4 Genre Model

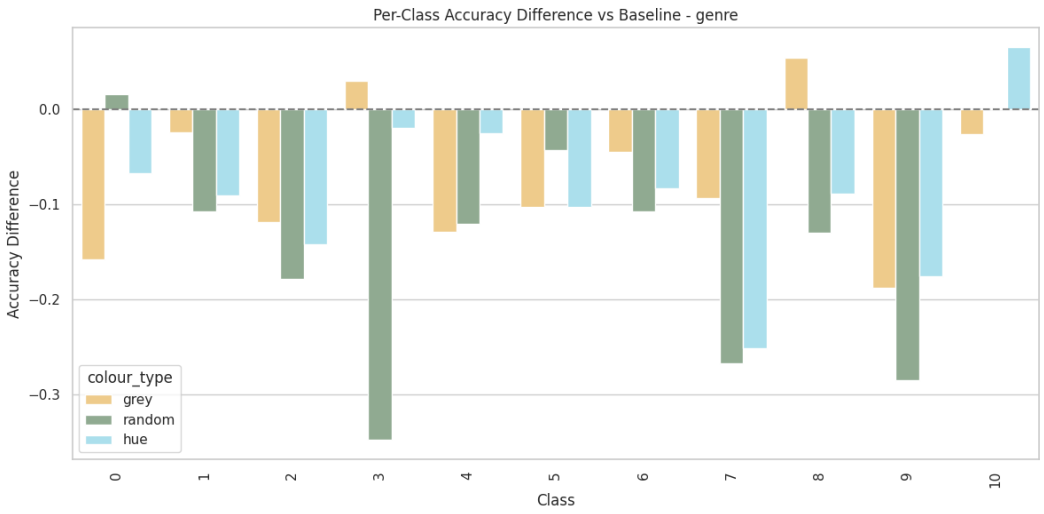


Figure 12: Genre recolouration accuracy per class, compared to baseline

As shown in Figure 7b, the genre model achieves a substantially higher accuracy than the other

two models across all recolouration types. As with the other models, random recolouration results in the lowest accuracy, while hue shifts consistently perform best. Despite these effects, overall performance remains relatively close to the baseline, indicating that genre classification is more robust to colour perturbations than artist or style classification.

In Figure 12, per-class average accuracy under each recolouration method is compared to the baseline using the testing and validation subsets of the filtered dataset. Only a small number of classes benefit from recolouration, with improvements distributed sparsely across grey, hue, and random transformations (two, one, and one class respectively). No class improves under more than one recolouration type, suggesting that gains are isolated and not consistent across transformations.

In contrast, degradation is most pronounced under random recolouration, which accounts for the majority of the lowest-performing classes. Grey and hue transformations show fewer severe drops, although class-specific sensitivity is still present. Overall, the results indicate that genre classification is affected less by systematic collapse and more by isolated instability in a small subset of classes.

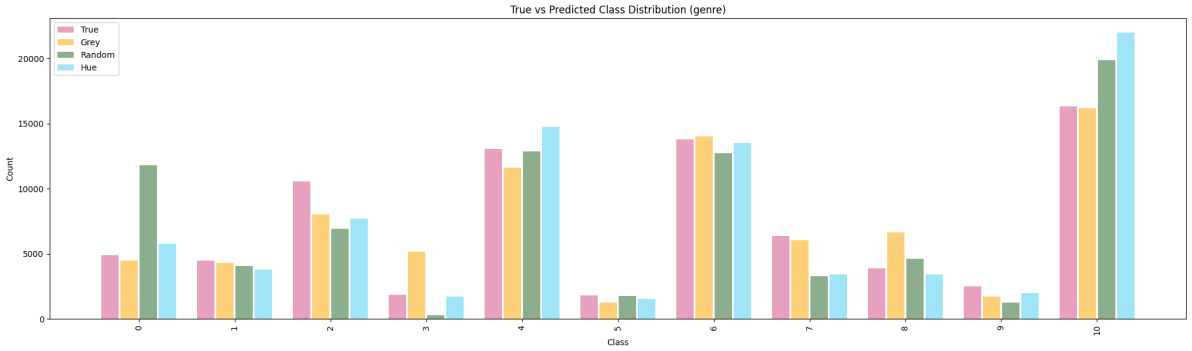


Figure 13: Genre prediction distribution per class, compared to true distribution

In Figure 13, predicted class distributions are compared to the true distribution using the full filtered dataset. Overprediction is limited and primarily appears under random recolouration, with a notable effect in class 0, and under grey recolouration in class 3. This suggests that recolouration does not strongly distort class priors for the genre model, but can induce localized bias in a small number of classes.

Underprediction is similarly limited, with the most significant case occurring for class 3 under random recolouration, which also corresponds to the lowest total prediction count in the model (301 predictions). Overall, prediction shifts are less widespread than in the artist and style models, indicating greater stability in the learned feature space.

Genre	TP	FP	Precision
grey	46384	33073	58.4 %
hue	46989	32468	59.1 %
random	42550	36907	53.6 %

Table 3: Model performance Genre Total

Finally, Table 3 reports total true positives, false positives, and precision across recolouration types using the full filtered dataset. Hue-shifted images achieve the highest precision, followed closely by grey recolouration. Random recolouration performs worst, with the largest reduction in true positives (3,834 fewer than grey transformations), indicating that stochastic colour perturbations have the strongest negative effect on classification reliability, even in the most robust model.

4.5 Recolour Comparison

Across all models, random recolouration consistently results in the lowest classification accuracy, as shown in Figure 7b. This pattern is also reflected in the precision values reported in Table 1, Table 2, and Table 3, indicating that stochastic colour perturbations degrade both correct classification and overall reliability.

While style and genre models show relatively similar precision under all colour transformations, the artist model is notably more sensitive, with random recolouration reducing precision by at least 13% compared to the other transformations. This suggests that fine-grained visual tasks such as artist classification are more dependent on stable colour structure than higher-level tasks such as genre classification.

Across all models, grey recolouration produces the largest number of accuracy improvements relative to the baseline in Figure 8, Figure 10, and Figure 12. However, this does not translate into the highest precision, as grey transformations consistently rank second behind hue shifts in precision performance.

Hue shifts show a distinct pattern: they rarely produce strong overprediction effects in Figure 9 and Figure 11, yet they consistently achieve the highest overall precision and average accuracy across all models in Figure 7b. This suggests that structured colour transformations preserve discriminative information more effectively than either greyscale conversion or stochastic perturbation, even if they do not maximize class-specific accuracy gains.

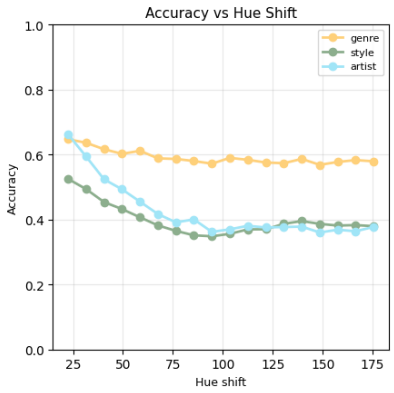
4.5.1 Hue & Random Shift Significance

Of the three types of colour alteration, both random overlay and hue shift depend on a randomly sampled parameter: alpha in the case of random overlays, and rotation angle in the case of hue shifts. Variations in alpha strength and hue rotation magnitude may therefore affect classification performance differently.

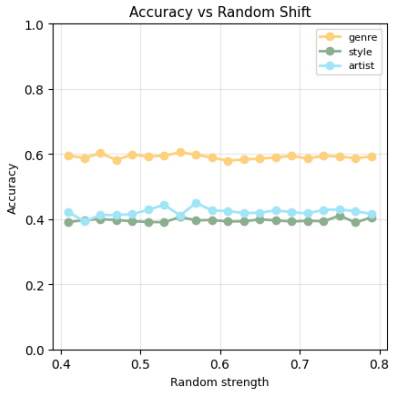
In Figure 14 and Figure 15, this effect is analysed for both transformations. For hue shifts, the rotation is measured as angular distance from the original colour space, treating values such as 20° and 340° as equivalent. All results are computed on the full filtered dataset.

Figure 14b shows that accuracy remains relatively stable across all levels of alpha for the random overlay. As expected from Figure 7b, the genre model consistently achieves higher accuracy than the style and artist models.

In contrast, Figure 14a shows that hue shifts have a stronger effect on style and artist performance.



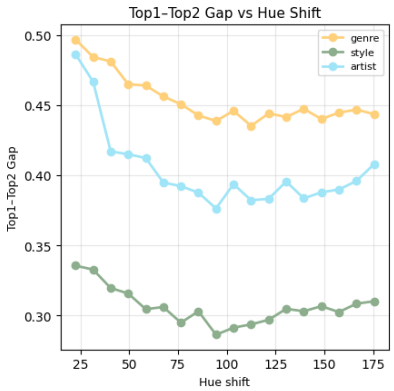
(a) Hue



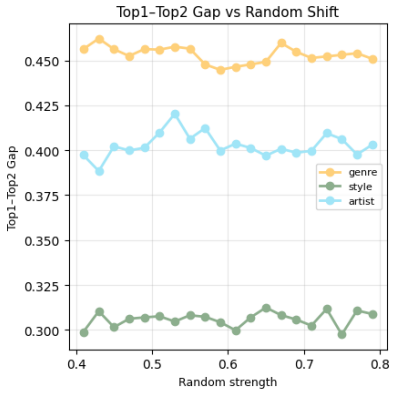
(b) Random

Figure 14: Average change in accuracy over shift strength

At 20°, style and artist accuracies are approximately 0.55 and 0.65 respectively, before both decrease and reach their lowest values around 100°. Beyond 100°, both stabilise at approximately 0.4 accuracy. The artist model shows a stronger decline than the style model, indicating higher sensitivity to hue variation.



(a) Hue



(b) Random

Figure 15: Difference between Top-1 and Top-2 accuracy by shift strength

Figure 15b shows that the gap between Top-1 and Top-2 predictions remains largely stable across all alpha strengths, with no meaningful difference between the extremes (0.4 and 0.8).

In contrast, Figure 15a shows a clear decrease in this gap as hue shift increases up to approximately 100° across all labels. This suggests that predictions become less decisive, with more samples entering the Top-2 range. Beyond 100°, the genre model remains stable, while artist and style show a slight recovery in the gap, though not to initial levels observed at 20°.

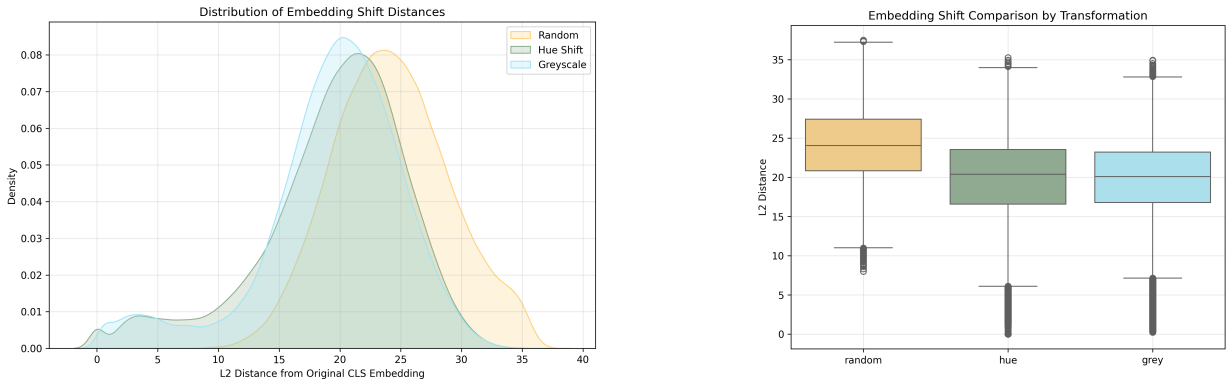
Across both figures, hue shift magnitude has a stronger effect on classification behaviour than random overlay strength. In particular, hue shifts between 100° and 180° show similar behaviour to random overlays with alpha values between 0.4 and 0.8, indicating comparable levels of perturbation at high transformation strength.

4.6 Representation Analysis

To better understand how colour alterations affect internal representations, this section analyses how CLS embeddings from the frozen Vision Transformer change under different transformations, and how these shifts manifest in a lower-dimensional embedding space.

4.6.1 Embedding Distance

To quantify how colour alterations affect the internal representations of the model, the Euclidean distance between CLS embeddings of original and recoloured images is computed. This provides a measure of how strongly each transformation shifts an input in the learned embedding space.



(a) KDE

(b) Boxplot

Figure 16: Distribution of embedding shift distance per colouration

Figure 16 shows the Euclidean distance between CLS embeddings of original and recoloured samples across the full filtered dataset. Table 4 provides the corresponding summary statistics. The mode is computed on distances rounded to one decimal place.

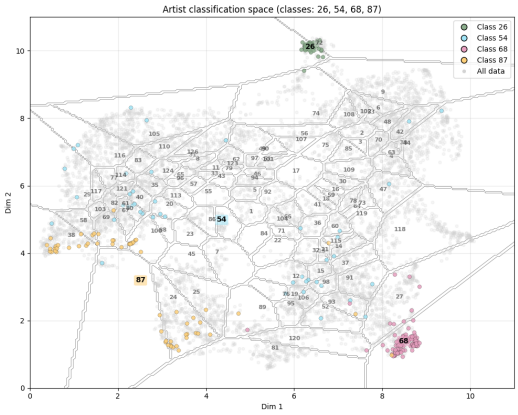
It is evident from both the table and Figure 16 that random colour alteration produces substantially larger embedding shifts than hue shift and greyscale transformations, which show broadly similar magnitudes. In Figure 16a, random alterations form an approximately normal distribution, whereas greyscale and hue shift exhibit left-skewed distributions, indicating that a large proportion of samples experience relatively small embedding changes. This pattern is also visible in Figure 16b, where both greyscale and hue shift contain many low-distance outliers below the whiskers.

Statistic	random	hue	grey
mean	24.21	19.46	19.38
mode	23.1	21.1	20.0
median	24.05	20.38	20.08
std	4.78	6.03	5.84
min	8.01	0.00	0.23
max	37.49	35.26	34.94

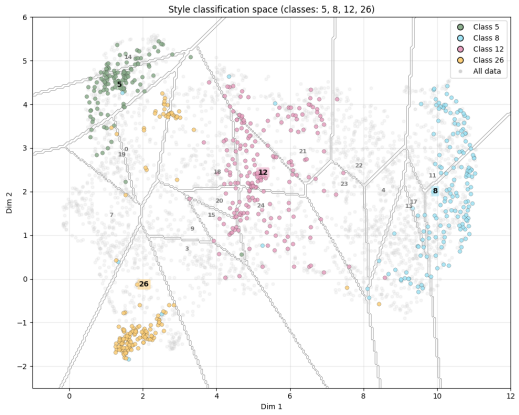
Table 4: Model embedding shift comparison

Comparing hue shift and greyscale in Table 4, hue shift shows slightly higher mean, median, mode, and standard deviation. However, greyscale exhibits a lower minimum distance (0.23), whereas hue shift includes at least one case where the embedding distance is effectively zero.

Although random alteration yields consistently higher embedding distances than both structured transformations, its standard deviation is lower, indicating a more concentrated distribution of perturbation magnitudes rather than higher variability. In contrast, hue shift and greyscale show greater relative dispersion but lower overall magnitude.



(a) Artist



(b) Style

(c) Genre

Figure 17: Embedding spaces across labels, with four representative classes per label

4.6.2 Colour Changes in Classification Space

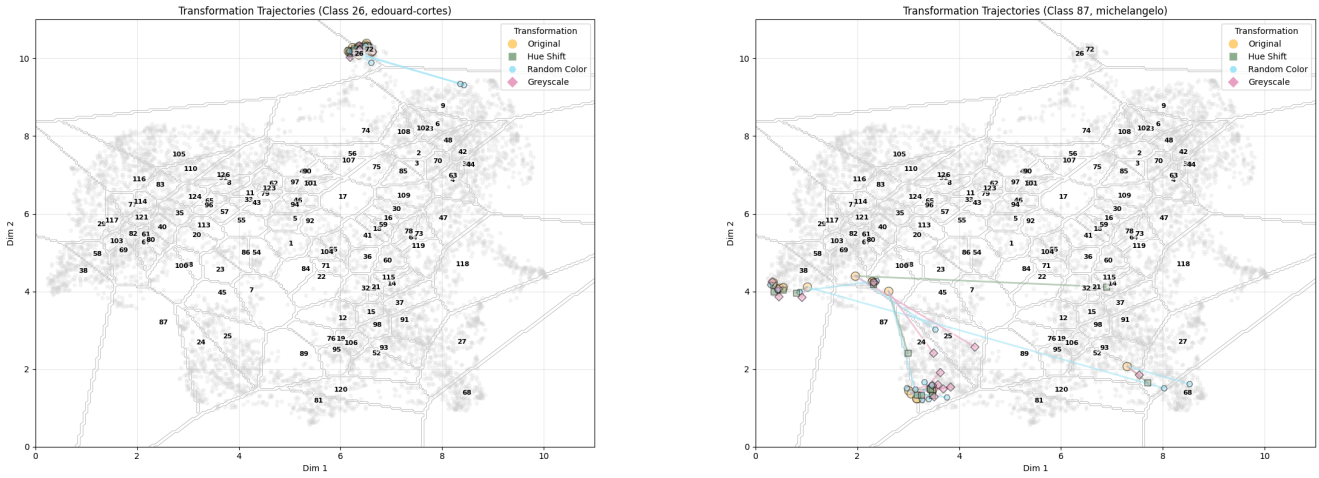
To further investigate how colour transformations affect internal representations, the learned embedding spaces of the Vision Transformer are visualized in two dimensions. These projections are based on CLS embeddings of the full filtered dataset and allow qualitative inspection of class structure and transformation-induced shifts.

Figure 17 shows the 2D projections for artist, style, and genre models. The selected classes were chosen to reflect variation in cluster compactness and separation.

In the artist space in Figure 17a, some classes (e.g. 26 and 68) form relatively compact clusters, while others (e.g. 54) are widely dispersed, indicating weaker class cohesion. Given the large number of classes, complete linear separability in 2D is not expected; however, local clustering structure is still visible.

In the style space in Figure 17b, classes such as 5 and 8 form more distinct clusters, while class 12 appears more dispersed and centrally located, indicating higher intra-class variability. Some overlap between neighbouring clusters is also visible, particularly near class 5.

The genre space in Figure 17c exhibits the most clearly separated structure overall. Classes 4, 8, and 9 form distinct clusters, while class 2 shows a broader distribution spanning multiple regions, suggesting greater internal variability within this class.

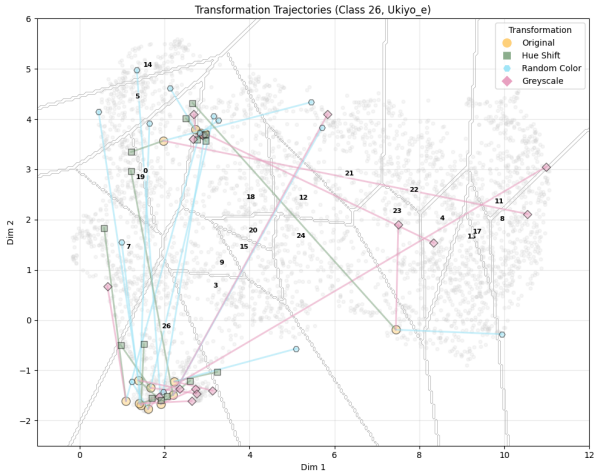


(a) Class 26 (b) Class 87

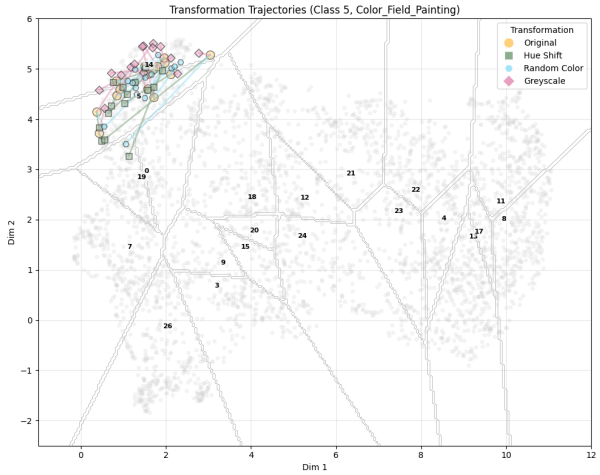
Figure 18: Shift within artist embedding space compared to baseline

For the artist model, classes 26 and 87 are shown in Figure 18, comparing baseline embeddings with those under recolouration. Class 26 remains relatively stable, with only a small number of random perturbations causing noticeable deviations. Class 87 shows a less compact structure, but most recoloured samples remain close to the original embedding distribution, with only a few outliers introduced by random transformations.

For the style model, classes 26 and 5 are shown in Figure 19. Class 5 remains relatively compact, although some recoloured samples shift toward neighbouring regions (including class 14). Greyscale transformations tend to induce more consistent directional shifts, while hue and ran-



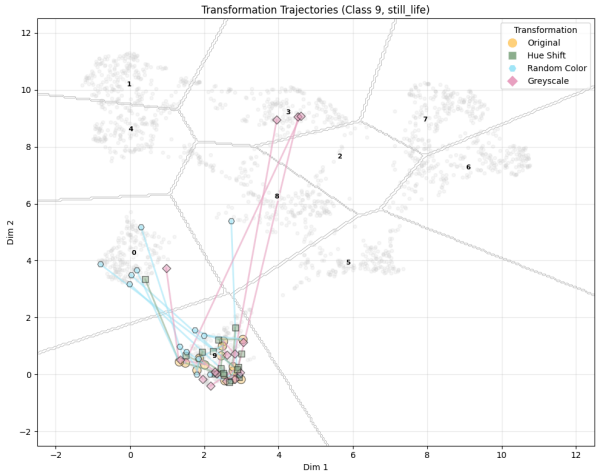
(a) Class 26



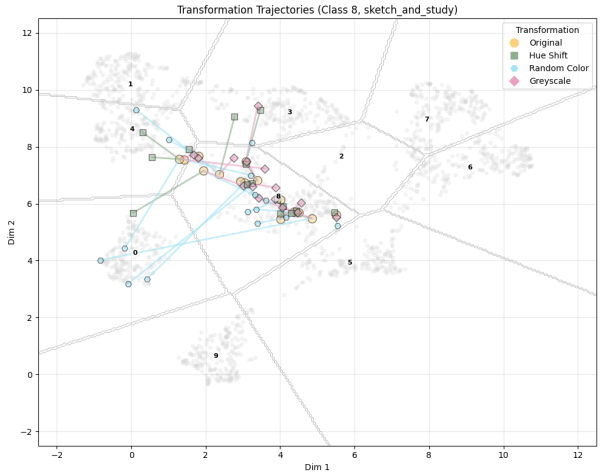
(b) Class 5

Figure 19: Shift within style embedding space compared to baseline

dom transformations produce more dispersed movement. Class 26 shows a broader baseline distribution, with recoloured samples producing larger deviations, particularly under random transformations.



(a) Class 9



(b) Class 8

Figure 20: Shift within genre embedding space compared to baseline

For the genre model, classes 9 and 8 are shown in Figure 20. Class 9 remains tightly clustered, with only a small number of outliers under random and greyscale transformations. Class 8 shows a less compact structure, with more frequent misplacements under random recolouration, including assignments to unrelated regions such as class 0. Hue shifts introduce comparatively fewer large deviations across both classes.

4.7 Discussion

Results have shown that, on average, colour alterations affect the accuracy of artist and style classification the most, while leaving genre classification significantly less affected. This suggests that colour is less important for accurate genre classification than for artist or style classification. On a per-class level, the results show that colour alteration significantly affect how often, and with which accuracy, distinct classes are predicted. Shown in Chapter 4.6.2, artist class 26 and style class 5 remain in similar regions of the embedding space even when colour alterations are applied. This contrasts with style class 26, which shows comparatively large embedding shifts under colour alteration. Possible explanations for why certain classes perform differently under colour alterations can be explored using art theory. The following paragraph does so, using illustrative images from the classes to form a possible explanation.

(a) Style 5, Color Field Painting. [Fra65]

(b) Artist 26, Edouard Léon Cortès.[Cor50]

(c) Style 26, Ukiyo-e [Hok15]

Figure 21: Examples of specific classes

Style class 5, Figure 21a, consists of Color Field Paintings, with large areas of a flat, single colour.This description holds even if the colours are changed, which may explain why style class 5 appears relatively robust to colour alterations. The example of Artist class 26, Figure 21b, is something many viewers commonly associate with a painting in the traditional sense. In other cases, such as artist Class 87 in Chapter 4.6.2, this would result in colour alteration changing the classification results significantly, but this does not happen for class 26. One possible explanation can be observed in the bottom right corner of Figure 21b, where the artist name is clearly signed. This may represent a shortcut learned by the model, which would remain largely unaffected by the applied colour alterations. This may help explain the high accuracy and minimal embedding shifts observed for this class. For Style class 26, Figure 21c, is an example that shifts largely under greyscale and random overlay alterations. Characteristically to the Ukiyo-e style, many pieces are woodcuts, and thus use a different technique than oil or paint on canvas. Being woodcuts, works in this style often share relatively consistent visual characteristics, which may make them more sensitive to colour alterations.

Looking at singular classes in Chapter 4.2, 4.3 and 4.4, a few of them do achieve a higher accuracy than the baseline does. Notably, greyscale recolouration is the transformation most frequently associated with improvements. However, several classes that improve relative to the

baseline are also overpredicted. This can coincide with higher class accuracy, as false positives are not directly reflected in per-class accuracy measurements. To explore why certain classes show both overprediction and higher accuracies, art theory can again be used as a source of possible explanations, using illustrative images from the respective classes.

(a) Artist 14, Nicholas Roerich. [Roe20]

(b) Style 21, Realism.[Pol76]

(c) Style 23, Romanticism [Alm08]

Figure 22: Examples of specific classes

Figure 22 shows three classes that achieved an accuracy higher than the baseline, and were also significantly overpredicted. Figure 22a is an example of artist class 14, which gets significantly overpredicted under random overlay alterations, as seen in Chapter 4.2. As can be seen, this example work is dominated by red and yellow tones, with relatively little colour variation. Random overlay shifts the image toward a more uniform colour appearance, which may resemble characteristics present in some works by Roerich. Figure 22b and 22c show examples of style class 21 and 23, both overpredicted by greyscale alterations as shown in Chapter 4.3. With these styles, realism is largely defined by form rather than colour, and romanticism is often characterized by dramatic light-dark contrasts. If these features are already present they remain after greyscale conversion, and if an original piece is turned into greyscale, these features likely pop out stronger than before. This may help explain why Realism and Romanticism are overpredicted under greyscale transformations. The accuracy achieved by colour alterations is thus largely class-specific, yet on average, the hue shift has the smallest drop compared to the baseline accuracy. One possible explanation is robustness introduced by the colour jitter augmentation applied during ViT training.This interpretation is consistent with the results in Chapter 4.5.1, where lower-degree hue shifts produce higher accuracies than larger shifts. In contrast, random overlay transformations may fall outside the range of colour variations encountered during training.

Figure 7b showed that on average, classifying classes for genre achieves a higher accuracy than for artist and style. One possible explanation is that genre classification primarily depends on image content, which closely aligns with the type of object-recognition tasks on which ViTs are

commonly pretrained. By comparison, artist and style classification may depend on more abstract visual cues. This suggests that recognising what is depicted in an image is more robust to colour alterations than recognising how it is depicted or who created it. While the other tasks do show a significant drop in accuracy, neither decreases with more than 50% of the original baseline accuracy, and thus are not entirely dependent on its original colours. Each average accuracy is higher than 25%, with most higher than 32%, despite the colour changes. Therefore, while style and artist classification appear more dependent on colour information than genre classification, a substantial portion of their predictive performance remains preserved under colour alteration, indicating that the models also rely on colour-invariant visual features.

5 Conclusion

In this chapter, the findings from Chapter 4 are interpreted in relation to the research sub-questions and the main research question of this study. The aim is to explain how colour alterations affect classification performance and internal representations in a Vision Transformer across artist, style, and genre classification tasks. The following section presents consolidated answers to the research sub-questions based on the experimental results.

5.1 Applied to research sub-questions

Question 1: To what extent is the ViT invariant to colour alterations?

Genre classification achieves consistently higher accuracy than artist and style classification across all colour alterations. Genre is therefore the most robust to colour changes. This suggests that genre classification is less dependent on colour information and more dependent on other features such as objects, composition, and scene structure. Artist and style classification show larger performance drops under all colour transformations, indicating stronger dependence on colour-related cues. However, none of the models collapse under colour alterations. Even under random recolouration, performance remains substantially above chance level, indicating that all tasks retain partial invariance to colour changes.

Question 2: How does classification accuracy change when colour information is changed?

Across all models, hue shift achieves the highest or near-highest accuracy, followed by greyscale, and random recolouration performing worst. This ordering is consistent across artist, style, and genre classification. Greyscale transformations show occasional class-level improvements, but these are not consistent across the dataset. Random recolouration consistently decreases performance across all models. Overall, hue-based transformations appear least disruptive, while random perturbations are most disruptive.

Question 3: Does recolouration disrupt embedding similarity?

Embedding distances show that random recolouration produces the largest shift in CLS embeddings. Hue shift and greyscale produce smaller and more similar embedding shifts, although hue shows slightly higher variance. This suggests that random colour noise disrupts learned representations more strongly than structured colour transformations. Notably, some samples show minimal or near-zero embedding change under hue or greyscale transformations, indicating partial invariance at instance level.

Question 4: Is colour more important for predicting artist, style, or genre?

Results indicate that colour is more important for artist and style classification than for genre classification. Artist and style models show larger reductions in accuracy under colour alterations, while genre classification remains comparatively stable. This suggests that genre prediction relies

more heavily on content and compositional information, whereas artist and style prediction depend more strongly on colour-related visual cues.

Question 5: Are some classes of artists, genres, or styles more dependent on colour than others?

Results show that colour alterations affect classes differently within each model. Artist and style classification show large variability across classes, with some classes being highly sensitive to colour changes and others remaining stable. Genre classification shows far fewer class-level fluctuations, indicating stronger generalization across colour changes. These results indicate that class sensitivity is not uniform, but may depend on the visual characteristics that distinguish individual classes.

5.2 Colour in Machine Perception of Art

The results show that colour acts as a modulating feature in Vision Transformer representations. Embeddings appear to rely strongly on non-chromatic information such as composition, texture, and content. However, colour still influences embedding position and class separation. Some classes remain stable under colour changes, while others shift noticeably in embedding space. This suggests that colour contributes to representation formation, but does not define it entirely.

5.3 When Colour Matters

The effect of colour is class-dependent and task-dependent. Some classes remain stable because the structural role of colour is preserved even when the exact colours are altered. This is visible in classes such as Color Field Painting, where large uniform colour regions after a colour transformation still remain large uniform colour regions. In contrast, some classes show stronger sensitivity to colour changes, indicating that colour may play a larger role in their visual representation. Some classes may also show stability due to learned shortcuts such as artist signatures, or strongly consistent visual patterns, which are largely unaffected by colour transformations.

Overall, colour appears most important when it is closely tied to class identity, and less important when stronger non-colour features dominate classification.

5.4 Main Research Question

How do colour alterations affect performance and internal representations in a pretrained Vision Transformer when classifying artworks?

The results show that colour alterations affect both classification performance and embedding structure, but not uniformly. Random recolouration produces the largest performance drop and the largest embedding shift. Hue shift produces the most stable performance among transformations. Greyscale lies between the two.

Embedding analysis shows that CLS representations remain largely structured under all transformations, but shift in position depending on the type of colour alteration. At the same time, the effect is strongly class-dependent, with some classes remaining stable and others showing large variation.

Therefore, colour affects Vision Transformer performance and representations, but its influence is conditional rather than absolute. The findings suggest that colour acts as one of several interacting visual cues, with its importance depending on both the classification task and the characteristics of individual classes.

Limitations and Future Work

This section discusses the main limitations of the present study and outlines directions for future research. While the results provide insight into the effect of colour alterations on ViT-based classification, several methodological and dataset-related constraints should be considered when interpreting the findings.

Limitations

Several limitations should be considered when interpreting these results. While several class-specific effects can be interpreted using concepts from art theory, these explanations remain speculative. The present study did not quantitatively measure artistic characteristics such as colour consistency, compositional structure, or stylistic techniques. Therefore, the proposed explanations should be viewed as plausible interpretations rather than causal conclusions.

Additionally, the dataset used has large class imbalances, leading to possible over-representation that could have affected results. Likewise, many images in the original dataset were already greyscale, leading to identical recolourations across all three methods. Similarly, certain classes contained consistent shortcuts, leading to the ViT learning dataset-specific cues instead of relying on colour information. The random overlay also applies a uniform transformation across all images, meaning that variations in perceptual impact caused by different alpha strengths are not explicitly controlled for.

Beyond dataset-related factors, the experimental design itself introduces several limitations. Only a single vision transformer architecture (ViT-B/16) was used, meaning results may not generalise across other transformer variants or more recent architectures. Although a CNN baseline was included for comparison, it was limited to a single ResNet-50 configuration, which restricts broader architectural conclusions. In addition, the colour manipulations are synthetic transformations applied in a controlled pipeline, and may not fully correspond to realistic colour variations encountered in digitisation processes, museum photography, or real-world artistic reproduction.

Furthermore, the analysis focuses primarily on global performance metrics such as average accuracy and per-class accuracy. While CLS embeddings and 2D projections were used to provide insight into internal representations, the study does not directly analyse attention patterns, intermediate layer behaviour, or token-level dynamics. As a result, conclusions about internal

mechanisms remain indirect.

Finally, detailed explanations were developed for only a limited number of classes. While these case studies provide insight into possible mechanisms behind the observed effects, they may not fully capture the behaviour of all artists, styles, and genres contained in the dataset.

Future Work

To further investigate robustness and internal representations under colour alterations, future work could look at specific classes in more detail. This includes analysing which classes or individual artworks are most sensitive to colour changes, and whether this relates to properties such as colour diversity or structure within the image.

Additionally, the analysis could be extended from final embeddings to intermediate layers of the vision transformer. This would allow a better understanding of how colour information is processed throughout the network, and whether it is mainly used in early or later layers. Attention maps could also be analysed to see whether colour changes affect where the model focuses.

The dataset could also be extended by introducing more controlled variations in visual properties. Instead of only applying global colour transformations, future work could separate features such as texture, composition, and colour palette more explicitly. This would help to better isolate which visual features are used for classification.

Another direction is to move beyond performance metrics and embedding distances, and instead analyse what the model is actually using for its predictions. Methods such as attention analysis or attribution maps could be used to identify whether decisions are based on colour information, texture, or structural patterns.

Finally, while this thesis briefly compares a ViT and a CNN, the analysis is still limited to two architectures. Extending this to more models and datasets would help determine whether the observed effects also hold in other settings.

References

Sources

[Bho+21] Srinadh Bhojanapalli et al. *Understanding Robustness of Transformers for Image Classification*. 2021. arXiv: 2103.14586 [cs.CV]. URL: https://arxiv.org/abs/2103.14586.

[Car+21] Mathilde Caron et al. *Emerging Properties in Self-Supervised Vision Transformers*. Apr. 2021. URL: https://arxiv.org/abs/2104.14294.

[CG16] Eva Cetinic and Sonja Grgic. "Genre classification of paintings". In: *2016 International Symposium ELMAR*. 2016, pp. 201–204. DOI: 10.1109/ELMAR.2016.7731786.

[Con26] Pillow Contributors. *Pillow (PIL Fork) Documentation*. 2026. URL: https://pillow.readthedocs.io/.

[DBK+21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. *Vision Transformer (ViT) Base Patch16-224*. 2021. URL: https://huggingface.co/google/vit-base-patch16-224.

[Den+09] Jia Deng et al. "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR*. 2009.

[Dos+20] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. Oct. 2020. URL: https://arxiv.org/abs/2010.11929.

[Fai13] Mark D. Fairchild. *Color Appearance Terminology*. John Wiley & Sons, Ltd, 2013. Chap. 4, pp. 85–96. ISBN: 9781118653128. DOI: https://doi.org/10.1002/9781118653128.ch4. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118653128.ch4.

[Fic07] Lois Fichner-Rathus. *Foundations of Art and Design*. Belmont, CA: Wadsworth/Cengage Learning, 2007. ISBN: 9780534613389.

[Gei+22] Robert Geirhos et al. *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness*. 2022. arXiv: 1811.12231 [cs.CV]. URL: https://arxiv.org/abs/1811.12231.

[Han+22] Kai Han et al. "A survey on Vision Transformer". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1 (Feb. 2022), pp. 87–110. DOI: 10.1109/tpami.2022.3152247. URL: https://arxiv.org/abs/2012.12556.

[He+15] Kaiming He et al. *Deep residual learning for image recognition*. Dec. 2015. URL: https://arxiv.org/abs/1512.03385.

[Hug21] HugGAN Community. *WikiArt Dataset*. 2021. URL: https://huggingface.co/datasets/huggan/wikiart.

[IW11] Dana E. Ilea and Paul F. Whelan. "Image segmentation based on the integration of colour–texture descriptors—A review". In: *Pattern Recognition* 44.10 (2011). Semi-Supervised Learning for Visual Content Analysis and Understanding, pp. 2479–2501. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2011.03.005. URL: https://www.sciencedirect.com/science/article/pii/S0031320311000902.

[KSH17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Commun. ACM* 60.6 (May 2017), pp. 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386. URL: https://doi.org/10.1145/3065386.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (May 2015), pp. 436–444. DOI: 10.1038/nature14539. URL: https://www.nature.com/articles/nature14539.

[Lec+98] Y. Lecun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[Mic24] Microsoft. *microsoft/resnet-50*. 2024. URL: https://huggingface.co/microsoft/resnet-50.

[Nas+21] Muzammal Naseer et al. *Intriguing properties of vision transformers*. May 2021. URL: https://arxiv.org/abs/2105.10497.

[PC21] Sayak Paul and Pin-Yu Chen. *Vision Transformers are Robust Learners*. 2021. arXiv: 2105.07581 [cs.CV]. URL: https://arxiv.org/abs/2105.07581.

[Rag+21] Maithra Raghu et al. *Do vision transformers see like convolutional neural networks?* Aug. 2021. URL: https://arxiv.org/abs/2108.08810.

[RW17] Waseem Rawat and Zenghui Wang. "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review". In: *Neural Computation* 29.9 (2017), pp. 2352–2449. DOI: 10.1162/neco_a_00990.

[TCL05] Charles C. Tappert, Sung-Hyuk Cha, and Thomas E. Lombardi. *The classification of style in fine-art painting*. 2005. URL: https://api.semanticscholar.org/CorpusID:12335580.

[Uni11] International Telecommunication Union. *Recommendation ITU-R BT.601-7: Studio Encoding Parameters of Digital Television for Standard 4:3 and Wide-Screen 16:9 Aspect Ratios*. Recommendation BT.601-7. ITU Radiocommunication Sector (ITU-R), Mar. 2011. URL: https://www.itu.int/rec/R-REC-BT.601.

[Vas+17] Ashish Vaswani et al. *Attention is all you need*. June 2017. URL: https://arxiv.org/abs/1706.03762.

[Wan+24] Shunxin Wang et al. *A Survey on the Robustness of Computer Vision Models against Common Corruptions*. 2024. arXiv: 2305.06024 [cs.CV]. URL: https://arxiv.org/abs/2305.06024.

[Wik26] WikiArt. *WikiArt.org - Visual Art Encyclopedia*. 2026. URL: https://www.wikiart.org/.

[Wu+20] Bichen Wu et al. *Visual Transformers: token-based image representation and processing for computer vision*. June 2020. URL: https://arxiv.org/abs/2006.03677.

[Zuj+09] Jana Zujovic et al. "Classifying paintings by artistic genre: An analysis of features & classifiers". In: *2009 IEEE International Workshop on Multimedia Signal Processing*. 2009, pp. 1–5. DOI: 10.1109/MMSP.2009.5293271.

Artwork

[Alm08] Sir Lawrence Alma-Tadema. *At Aphrodite's Cradle*. Oil on canvas. 1908.

[Cor50] Edouard Léon Cortès. *Quai de la Seine sous la neige*. Oil on canvas. 1950.

[Fra65] Sam Francis. *Untitled (Edge Painting)*. Acrylic on paper. 1965.

[Hok15] Katsushika Hokusai. *The Adonis plant*. Woodcut. 1815.

[Pic37] Pablo Picasso. *Guernica*. Oil on canvas. 1937.

[Pol76] Vasily Polenov. *Черногорская девушка*. Watercolor on paper. 1876.

[Roe20] Nicholas Roerich. *Неведомый певец*. Pastels on canvas. 1920.

[Roe34] Nicholas Roerich. *Sacred Himalayas*. Tempera on canvas. 1934.

[Sis74] Alfred Sisley. *Hoar Frost, St. Martin's Summer (Indian Summer)*. Oil on canvas. 1874.

Appendices

A Class to name mapping

A direct mapping from the class number to the actual style, genre or artist they represent.

Styles (0–13)	Styles (14–26)	Genres
0 Abstract Expressionism	14 Minimalism	0 abstract painting
1 Action painting	15 Naive Art Primitivism	1 cityscape
2 Analytical Cubism	16 New Realism	2 genre painting
3 Art Nouveau	17 Northern Renaissance	3 illustration
4 Baroque	18 Pointillism	4 landscape
5 Color Field Painting	19 Pop Art	5 nude painting
6 Contemporary Realism	20 Post Impressionism	6 portrait
7 Cubism	21 Realism	7 religious painting
8 Early Renaissance	22 Rococo	8 sketch and study
9 Expressionism	23 Romanticism	9 still life
10 Fauvism	24 Symbolism	10 Unknown Genre
11 High Renaissance	25 Synthetic Cubism	
12 Impressionism	26 Ukiyo e	
13 Mannerism Late Renaissance		

Artists (0–31)	Artists (32–63)	Artists (64–95)	Artists (96–128)
0 Unknown Artist	32 Egon Schiele	64 Gustav Klimt	96 Lawrence Alma-Tadema
1 Boris Kustodiev	33 Thomas Eakins	65 Vasily Perov	97 Vasily Vereshchagin
2 Camille Pissarro	34 Gustave Moreau	66 Odilon Redon	98 Ernst Ludwig Kirchner
3 Childe Hassam	35 Francisco Goya	67 Tintoretto	99 Mikhail Vrubel
4 Claude Monet	36 Edvard Munch	68 Gene Davis	100 Orest Kiprensky
5 Edgar Degas	37 Henri Matisse	69 Raphael	101 William Merritt Chase
6 Eugene Boudin	38 Fra Angelico	70 John Henry Twachtman	102 Aleksey Savrasov
7 Gustave Doré	39 Maxime Maufra	71 Henri de Toulouse-Lautrec	103 Hans Memling
8 Ilya Repin	40 Jan Matejko	72 Antoine Blanchard	104 Amedeo Modigliani
9 Ivan Aivazovsky	41 Mstislav Dobuzhinsky	73 David Burliuk	105 Ivan Kramskoy
10 Ivan Shishkin	42 Alfred Sisley	74 Camille Corot	106 Utagawa Kuniyoshi
11 John Singer Sargent	43 Mary Cassatt	75 Konstantin Korovin	107 Gustave Courbet
12 Marc Chagall	44 Gustave Loiseau	76 Ivan Bilibin	108 William Turner
13 Martiros Saryan	45 Fernando Botero	77 Titian	109 Theo van Rysselberghe
14 Nicholas Roerich	46 Zinaida Serebriakova	78 Maurice Prendergast	110 Joseph Wright
15 Pablo Picasso	47 Georges Seurat	79 Édouard Manet	111 Edward Burne-Jones
16 Paul Cézanne	48 Isaac Levitan	80 Peter Paul Rubens	112 Koloman Moser
17 Pierre-Auguste Renoir	49 Joaquín Sorolla	81 Aubrey Beardsley	113 Viktor Vasnetsov
18 Pyotr Konchalovsky	50 Jacek Malczewski	82 Paolo Veronese	114 Anthony van Dyck
19 Raphael Kirchner	51 Berthe Morisot	83 Joshua Reynolds	115 Raoul Dufy
20 Rembrandt	52 Andy Warhol	84 Kuzma Petrov-Vodkin	116 Frans Hals

Artists (0–31)	Artists (32–63)	Artists (64–95)	Artists (96–128)
21 Salvador Dalí	53 Arkhip Kuindzhi	85 Gustave Caillebotte	117 Hans Holbein
22 Vincent van Gogh	54 Niko Pirosmani	86 Lucian Freud	118 Ilya Mashkov
23 Hieronymus Bosch	55 James Tissot	87 Michelangelo	119 Henri Fantin-Latour
24 Leonardo da Vinci	56 Vasily Polenov	88 Dante Gabriel Rossetti	120 M. C. Escher
25 Albrecht Dürer	57 Valentin Serov	89 Félix Vallotton	121 El Greco
26 Édouard Cortès	58 Pietro Perugino	90 Nikolay Bogdanov-Belsky	122 Mikalojus Čiurlionis
27 Sam Francis	59 Pierre Bonnard	91 Georges Braque	123 James McNeill Whistler
28 Juan Gris	60 Ferdinand Hodler	92 Vasily Surikov	124 Karl Bryullov
29 Lucas Cranach the Elder	61 Bartolomé Esteban Murillo	93 Fernand Léger	125 Jacob Jordaens
30 Paul Gauguin	62 Giovanni Boldini	94 Konstantin Somov	126 Thomas Gainsborough
31 Konstantin Makovsky	63 Henri Martin	95 Katsushika Hokusai	127 Eugène Delacroix
			128 Canaletto

B Dataset Distribution

The distributions of the data that the models were trained, tested, and evaluated with, per class per label. Compared to the full dataset, this has removed Style classes; 1, 2, 6, 16, 25, and Artist classes; 28, 34, 50, 99, 111, 112, 122, 125, 127, 128, as these classes had less than 500 or 100 artworks respectively.

Artist class 0, "Unknown Artist" was removed from any training or testing of the artist classification model.

Styles (0–13)			Styles (14–26)			Genres		
Class	Count	%	Class	Count	%	Class	Count	%
0	2782	3.50	14	1337	1.68	0	4872	6.13
3	4206	5.29	15	2405	3.03	1	4491	5.65
4	4116	5.18	17	2552	3.21	2	10559	13.29
5	1615	2.03	18	513	0.65	3	1855	2.33
7	2164	2.72	19	1483	1.87	4	13046	16.42
8	1391	1.75	20	6441	8.11	5	1823	2.29
9	6736	8.48	21	10706	13.47	6	13804	17.37
10	934	1.18	22	2077	2.61	7	6347	7.99
11	1343	1.69	23	6847	8.62	8	3854	4.85
12	13055	16.43	24	4308	5.42	9	2483	3.12
13	1279	1.61	26	1167	1.47	10	16323	20.54

Artist (0–42)			Artist (43–85)			Artist (86–126)		
Class	Count	%	Class	Count	%	Class	Count	%
0	41014	51.62	43	293	0.37	86	224	0.28
1	633	0.80	44	256	0.32	87	129	0.16

Artist (0–42)			Artist (43–85)			Artist (86–126)		
Class	**Count**	**%**	**Class**	**Count**	**%**	**Class**	**Count**	**%**
2	887	1.12	45	151	0.19	88	139	0.17
3	550	0.69	46	329	0.41	89	169	0.21
4	1334	1.68	47	162	0.20	90	198	0.25
5	611	0.77	48	424	0.53	91	127	0.16
6	555	0.70	49	335	0.42	92	238	0.30
7	753	0.95	51	225	0.28	93	223	0.28
8	539	0.68	52	107	0.13	94	194	0.24
9	577	0.73	53	161	0.20	95	171	0.22
10	520	0.65	54	113	0.14	96	166	0.21
11	784	0.99	55	341	0.43	97	166	0.21
12	765	0.96	56	225	0.28	98	340	0.43
13	575	0.72	57	191	0.24	100	216	0.27
14	1819	2.29	58	173	0.22	101	354	0.45
15	659	0.83	59	142	0.18	102	230	0.29
16	579	0.73	60	204	0.26	103	151	0.19
17	1400	1.76	61	171	0.22	104	344	0.43
18	919	1.16	62	225	0.28	105	154	0.19
19	516	0.65	63	140	0.18	106	199	0.25
20	776	0.98	64	105	0.13	107	217	0.27
21	479	0.60	65	178	0.22	108	141	0.18
22	1889	2.38	66	164	0.21	109	169	0.21
23	182	0.23	67	187	0.24	110	135	0.17
24	164	0.21	68	155	0.20	113	132	0.17
25	695	0.87	69	143	0.18	114	163	0.21
26	214	0.27	70	214	0.27	115	133	0.17
27	317	0.40	71	324	0.41	116	176	0.22
29	179	0.23	72	170	0.21	117	121	0.15
30	359	0.45	73	248	0.31	118	193	0.24
31	324	0.41	74	456	0.57	119	105	0.13
32	226	0.28	75	247	0.31	120	126	0.16
33	240	0.30	76	171	0.22	121	159	0.20
35	205	0.26	77	204	0.26	123	134	0.17
36	150	0.19	78	362	0.46	124	145	0.18
37	415	0.52	79	210	0.26	126	164	0.21
38	167	0.21	80	259	0.33			
39	119	0.15	81	161	0.20			
40	167	0.21	82	130	0.16			
41	149	0.19	83	200	0.25			
42	464	0.58	84	205	0.26			