Robustness of Pre-Trained CLIP for Artist Predictions Under Visual Transformations

Shaoxuan Shi, Emre Genç, Linh Khanh Nguyen, and Bedirhan Gursoy

University of Amsterdam, 1012 WP Amsterdam, Netherlands
secretariat-ase@uva.nl

Abstract. This project studies the robustness of a pre-trained CLIP model for artist predictions. We evaluate CLIP on predicting the top 100 most prolific artists in the SemArt dataset without task-specific fine-tuning. We also examine how different image transformations affect prediction performance, including Grayscale Transformation (GT), Random Perspective Transformation (RPT), and Elastic Transformation (ET). In the initial evaluation on 7,913 paintings, CLIP achieves an accuracy of 29.56%, correctly identifying 2,339 paintings, with a macro-averaged F1 score of 0.23. These correctly classified samples from 76 artists are then used to assess robustness under visual transformations. Performance decreases under all transformations, with GT getting the highest accuracy at 65.3%, followed by RPT at 59.9%. ET has the strongest impact, reducing accuracy to 40.2%. These findings may suggest that CLIP exhibit limited robustness to visual distortions and prefers spatial and structural cues of chromatic information.

Keywords: CLIP · visual transformations · author prediction · robustness.

1 Introduction

1.1 Background

Contrastive Language–Image Pretraining (CLIP) is a vision–language model introduced by Radford et al., which learns joint representations of images and text by training on a large collection of image–text pairs collected from the web [1]. CLIP uses contrastive learning to match image and text embeddings in a shared representation space. One important property of CLIP is its ability to perform zero-shot classification, where class labels are expressed as natural language prompts instead of fixed supervised categories. During prediction, an image embedding is compared with text embeddings, and the label with the highest similarity score is selected. This design allows CLIP to generalize to new tasks and datasets without task-specific fine-tuning [1, 2].

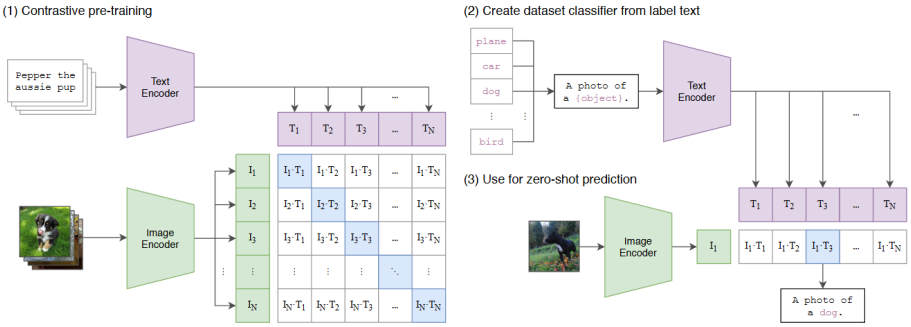


Fig. 1: Overview of the CLIP image and text encoding process (reproduced from [1], p. 2).

Recent studies have explored the use of CLIP for artwork analysis, including painting classification, and image–text retrieval[3]. Prior work shows that CLIP can achieve competitive zero-shot performance on fine-art datasets such as iMet, for tasks such as fine-grained artwork attribute recognition and semantic description [4]. However, the existing studies are limited and most of those studies focus on image–text retrieval rather than artist or author attribution. There is actually no established work that evaluates pre-trained CLIP for artist prediction, making this task underexplored.

Robustness to visual transformations used to understand how models respond to changes in visual appearance and distribution shifts. Grayscale transformation, geometric distortions, and elastic transformation are widely used in data augmentation and robustness evaluation [6]. Some studies show that convolutional neural networks can be sensitive to changes in texture, color, and spatial structure, sometimes relying on non-semantic visual cues [7, 8]. Some recent work extends robustness analysis to vision–language models, including CLIP, indicating that visual corruptions and geometric transformations can change embedding similarity and make zero-shot performance worse [9, 10]. These findings indicate that CLIP's robustness is valuable for evaluation in specific tasks.

SemArt provides rich multi-model resources for semantic art understanding, combining painting images with textual artistic comments. And it comes from a study on multi-modal retrieval and representation learning [5]. However, SemArt-based studies have not explored pre-trained CLIP models.

Under this background, we focus on the visual component of CLIP while keeping textual prompts fixed. By applying controlled visual transformations, including grayscale conversion, geometric distortion, and elastic transformation, we aim to analyze to what extent CLIP relies on those visual properties when matching paintings to artists. This image-focused robustness analysis directly motivates our research questions in the following section.

1.2 Research Question

In this project, we use a pre-trained CLIP model to predict the top 100 most prolific artists in the SemArt dataset with their paintings. We investigate how changes to visual information, such as removing color or distorting spatial structure affect the model's ability to correctly identify the artist. This helps us understand both how robust CLIP is to visual changes and which visual cues it relies on most when attributing artworks to artists. The main research question is

How robust is a pre-trained CLIP model for top 100 most prolific artists predictions when visual properties of paintings are changed?

To answer this question, the report is structured around four key sub-questions:

- How well does pre-trained CLIP perform at top 100 most prolific artists predictions?
- How does removing color information affect CLIP's artist predictions?
- How does geometric distortion affect CLIP's artist predictions?
- How does Elastic Transform affect CLIP's artist predictions?

2 Methodology

2.1 Descriptive statistics

Data on artworks were collected from the SemArt project developed by Aston University. It is a CSV gallery of European Artworks produced between the 13th and 19th centuries. In the CSV, each image is associated with features that include title, technique, date, type, school, and timeframe.

Certain data cleaning techniques were applied to remove anomalies in the data. Authors of the *UNKNOWN MASTER type, UNKNOWN ICON PAINTER, and ROMANESQUE PAINTER* were removed from the data set to remove the vagueness from the data. A total of 584 data rows were removed and a final size of 20798 images was obtained.

For our analysis, we compared our model with multiple features such as technique, type, and school, but we ended up predicting the Author column from the dataset. There are a total of 3,253 different artists, with the most common being Vincent van Gogh, Rembrandt van Rijn, and Giotto di Bondone.

The data set contains paintings from the 13th to the 19th century. The oldest paintings are from the 13th century with 106 paintings, and each half-century the number of paintings rises gradually, where the peak is reached in the first half of the 17th century with a total of 3770 paintings. There is a slight decline in the number of paintings after the 16th century, where the number of paintings was reduced to 1936 in the second half of the 19th century.

The schools in our data set define the origin country of the paintings. Italian schools are the most common in our dataset, with 8652 paintings, which is almost 42% of all the paintings in the dataset. Other common paintings are

from the Dutch, French, and Flemish Schools, which also make up around 40% of the paintings together. The least number of paintings come from the Greek, Norwegian, Polish, Portuguese, and Finnish schools.

Each painting is categorized into 10 different types of artworks. The most common is religious, with 7872 paintings, which make up 37% of the paintings. Portrait paintings are slightly more common than landscape ones, where they make up 17% of paintings, as opposed to landscape painting only having 13%. The least common types of paintings are study, interior and other ones.

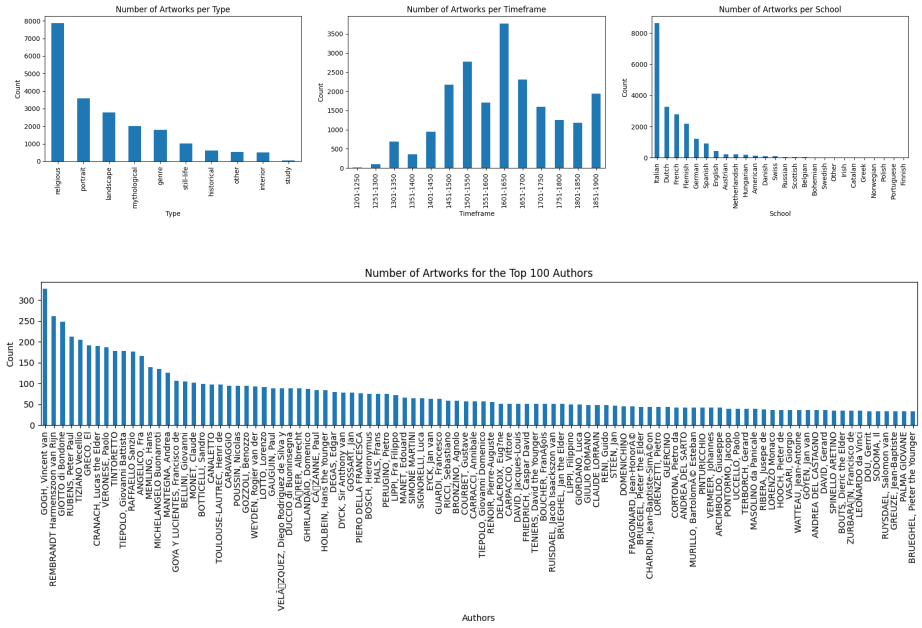


Fig. 2: Distribution of artwork types, timeframes, schools, and top artists in the SemArt dataset.

2.2 CLIP Model

In this study, we use a pre-trained CLIP model with a ViT-B/32 backbone in a zero-shot classification setting without any task-specific fine-tuning. The model is applied to predict artists by comparing visual embeddings of paintings with text embeddings corresponding to artist names. Artist predictions are performed on the painting of the top 100 most prolific artists in the data set.

The prediction is done by computing the cosine similarity between the image embedding and each artist text embedding. The artist name that achieves the highest similarity score is selected as the prediction of the model.

To isolate the effect of visual transformations, the artist names and model parameters are kept fixed throughout all experiments. The same pipeline is applied to both the original and transformed images.

2.3 Types of Transformations

In order to evaluate the robustness of the zero-shot CLIP model, three visual transformations were applied to the paintings whose artist was correctly predicted in the original setting. Each transformation modifies different aspects of the image while preserving the semantic content.

Grayscale Transformation (GT). The grayscale transformation removes all color information by converting the image into intensity values only. Therefore, this transformation tests whether the model relies on color cues, or whether it can still recognize artists based only on structural elements, texture, and composition.

Fig. 3: Example of GT applied to *Madame Raymond de Verninac* by Jacques-Louis David.

Random Perspective Transformation (RPT). The random perspective transformation applies a geometric distortion by changing the viewpoint or the angle of the camera. By randomly warping the image, it modifies the overall spatial structure and composition while keeping most visual details intact. This allows us to test how robust CLIP is to changes in spatial alignment caused by digitization, framing, or different viewing angles.

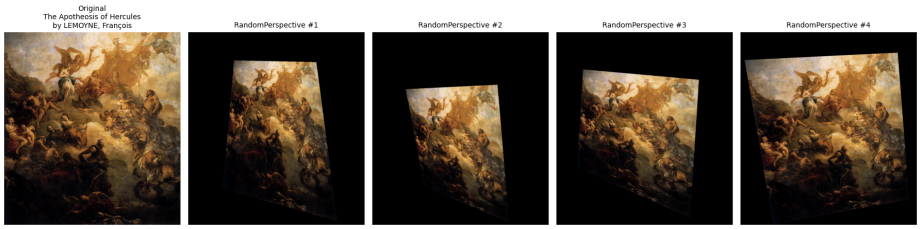


Fig. 4: Example of RPT applied to *The Apotheosis of Hercules* by François Lemoyne.

Elastic Transformation (ET). The elastic transformation applies smooth and non-linear distortions that slightly stretch or bend the image. This transformation helps to evaluate whether CLIP relies more on exact geometric details or stylistic features when handling deformations.

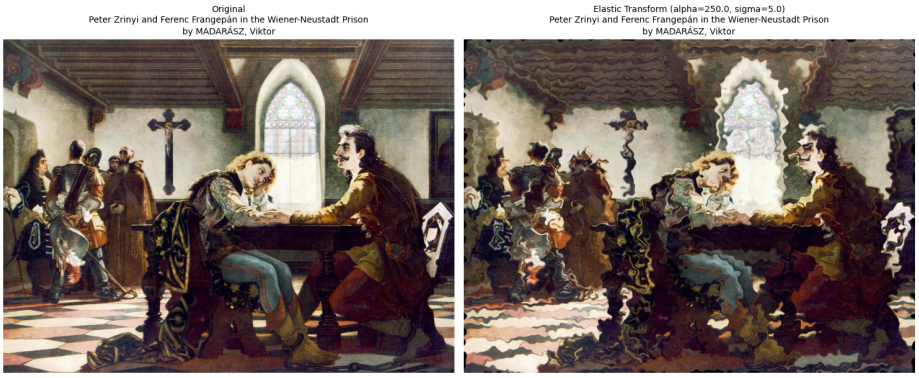


Fig. 5: Example of ET applied to *Peter Zrinyi and Ferenc Frangepán in the Wiener-Neustadt Prison* by Viktor Madarász.

3 Results

3.1 CLIP performance of top 100 most prolific artists predictions

To answer the first sub-research question, we evaluate the performance of a pre-trained CLIP model on predicting the top 100 most prolific artists in the SemArt dataset. As shown in Table 1, CLIP achieves an overall accuracy of 29.56%, correctly identifying the artist for 2339 out of 7913 paintings. The macro-averaged F1 score of 0.23 shows performance variation across artists, while the weighted F1 score of 0.28 suggests stronger performance on artists with more training samples.

Table 1: Pre-trained CLIP performance for top-100 artist prediction

Metric	Value
Accuracy	29.56%
Macro Precision	0.25
Macro Recall	0.27
Macro F1 Score	0.23
Weighted F1 Score	0.28

Performance varies across individual artists. Several artists achieve high prediction accuracy, including Canaletto (86.6%), Paul Cézanne (83.3%), Johannes Vermeer (83.3%), Paul Gauguin (83.1%), and Pierre-Auguste Renoir (80.4%). In contrast, a group of artists receive no correct predictions although having a large number of available works. These include Raffaello Sanzio (Raphael), Jan Gossart, Dieric the Elder Bouts, Domenico Ghirlandaio, and Luca Giordano, all of whom exhibit 0% accuracy. This shows that pre-trained CLIP performance is highly uneven across artists.

3.2 Post-transformation CLIP predictions and performance

After initial author predictions with CLIP, the evaluation subset comprises 2339 paintings by 76 authors. All 2339 samples are correctly classified top-100 most frequently occurring authors in the SemArt dataset. Then, we applied three image transformations and predicted the authors again in order quantify the robustness to alterations of the original images. The results are reported below.

Figure 6 summarizes post-transformation prediction accuracy. In general, the model's robustness decreased substantially under all three transformations. Out of the three transformations, GT had the best prediction rate, at 65.3%. Random Perspective Transformation performed slightly worse, at 59.9%. Lastly, ET only correctly predicted less than half of the paintings, at 40.2%.

Table 2 shows the correctness of the predictions jointly across all transformation. Only 23.4% (547) of paintings were correctly predicted under all three transformations. This indicates that 76.6% of all paintings were incorrectly predicted at least once. Exactly 794 (33.9%) paintings were correct under two transformations and 640 (27.4%) were correct under one transformation. Finally, 15.3% (358) of the paintings were misclassified for all three transformations.

Table 3 reports the confidence statistics for all three transformations. The baseline confidence is $\overline{c}_{\text{base}} = 0.318$. The mean confidence $\overline{c}_{\text{trans}}$ tends to decrease after transformation, with mean change of -0.024 (GT), -0.020 (RPT) and -0.011 (ET). Questionably, the mean change in confidence of ET is smaller relative to GT and RPT despite having the worst accuracy performance. This can also be observed in Figure 7, where the distribution of the difference between the original confidence and post-transformation confidence shifted slightly closer to zero for ET compared to GT and RPT. In terms of reliability, this shows a

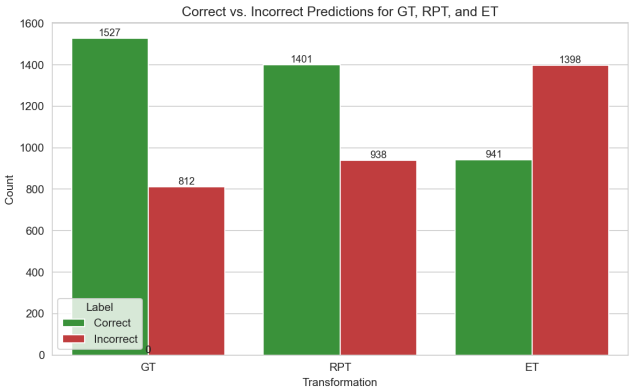


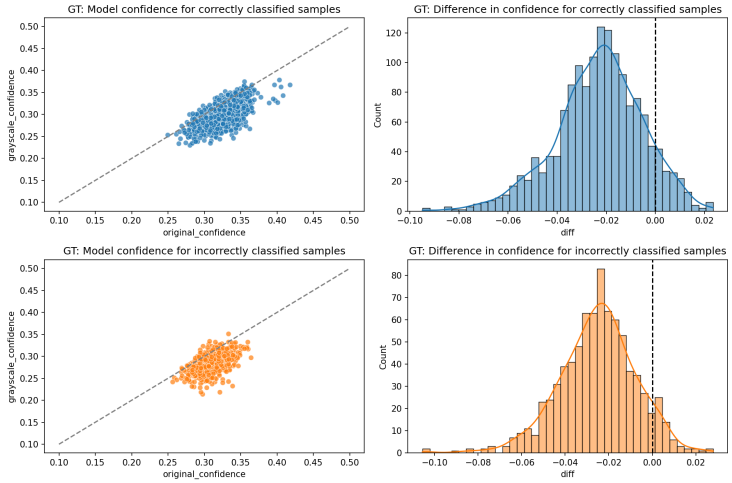
Fig. 6: Counts of correct and incorrect author predictions on the N=2339 evaluation subset under three visual transformations

Table 2: Prediction pattern across transformations (N=2339). Entries indicate whether the prediction is correct (1) or incorrect (0) after each transformation.

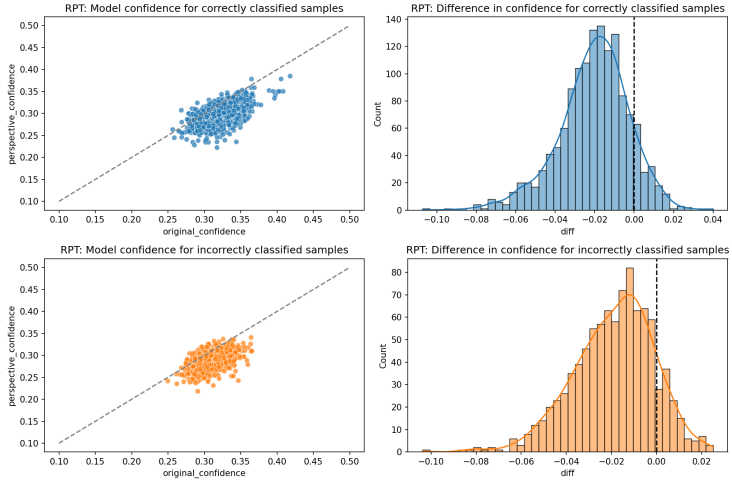
GT	RPT	ET	Count	% of dataset
1	1	1	547	23.39
1	1	0	478	20.44
1	0	1	167	7.14
0	1	1	149	6.37
1	0	0	335	14.32
0	1	0	227	9.71
0	0	1	78	3.33
0	0	0	358	15.31

mismatch between the prediction accuracy and the confidence of the model, making confidence an unreliable measure for model performance.

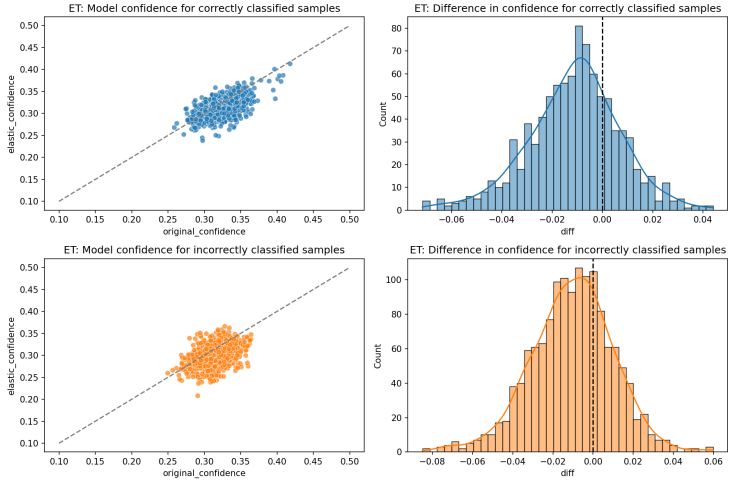
Table 4 shows the macro and micro measures of recall rate and F1 score after each transformation. It can be observed that the model struggled to predict correctly under transformation, and the performance decreased even further when the measures were averaged equally across all artists, as measured by macro recall rate. This suggests the model struggled to predict some artists more than others. Similarly, macro F1 scores are low across all transformations, at only 0.437, 0.388 and 0.256 for GT, RPT and ET, respectively. This indicates that error is a combination of misclassifications of an artist's own work and systematic misclassifications of other artists' work to that artist. The results from this table reinforces the conclusion that GT performed the best among the three, followed by RPT and ET.



(a) GT



(b) RPT



(c) ET

Fig. 7: Confidence scatter plots and confidence-difference histograms under three transformations, by correct and incorrect predictions.

Table 3: Confidence statistics (N=2339).

Transformation	$\overline{c}_{\text{base}}$	$\overline{c}_{\text{trans}}$	$\overline{c}_{\text{trans}}$ - $\overline{c}_{\text{base}}$	$\overline{c}_{\text{corr}}$	$\overline{c}_{\text{incorr}}$
GT	0.318	0.295	-0.024	0.300	0.285
RPT	0.318	0.298	-0.020	0.302	0.293
ET	0.318	0.308	-0.011	0.313	0.304

Table 4: Recall and F1 scores after each transformation (N=2339 paintings, 76 artists).

Transformation	Micro Recall	Macro Recall	Weighted Recall	Micro F1	Macro F1	Weighted F1
GT	0.653	0.494	0.653	0.653	0.437	0.662
RPT	0.599	0.408	0.599	0.599	0.388	0.615
ET	0.402	0.292	0.402	0.402	0.256	0.419

4 Limitations

One limitation of this study is the lack of limited computational resources. Due to constraints in available processing power the number of experimental iterations were restricted. We reduced the number of paintings in our dataset from 3252 authors to 100 authors with the most number of paintings. Future work mapping more advanced computational power may enable improved performance.

Another limitation of this study is that transformations were applied only to correctly predicted paintings. Due to time constraints, the overall accuracy and prediction scores of the study were not high; therefore, only images that were correctly classified were selected for conversion.

Additionally, the CLIP model is originally designed for text and image representation learning rather than for predicting author names. As a result, using the model for author classification may reduce its performance. Fine tuning the model specifically for this task could have led to higher prediction accuracy.

Moreover, since transformations were applied only to paintings which were correctly classified by the CLIP model, the reported F1 scores do not represent a full multi-class author prediction. Instead, they measure robustness under ideal conditions which could potentially exaggerate how successfully the model performs.

5 Conclusion

This paper utilized the CLIP model to predict the 100 most prevalent artists in the SemArt dataset, and empirically evaluate the prediction's robustness when visual properties of the paintings are changed. The results showed that the CLIP model's performance is moderate when predicting artists in the original dataset, and performance further deteriorates when visual transformations are applied.

Among the tested transformations, GT has the smallest negative impact on performance, which suggests CLIP relies less on the color of the paintings and more on structural and compositional features. A change in perspective of the painting worsens the performance relative to GT, indicating that CLIP is sensitive to spatial properties. Lastly, elastic transformation has the most negative effect on the prediction performance compared to the other two, which suggests that CLIP relies on geometric details and local visual consistency. These findings may suggest that CLIP exhibit limited robustness to visual distortions and prefers spatial and structural cues of chromatic information.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: A Visual Language Model for Few-Shot Learning. In: Advances in Neural Information Processing Systems, pp. 23716–23736 (2022)
3. Ghildyal, A., Wang, L.Y., Liu, F.: WP-CLIP: Leveraging CLIP to Predict Wölfflin's Principles in Visual Art. arXiv preprint arXiv:2508.12668 (2025)
4. Conde, M.V., Turgutlu, K.: CLIP-Art: Contrastive Pre-training for Fine-Grained Art Classification. arXiv preprint arXiv:2204.14244 (2022)
5. Garcia, N., Renoust, B., Nakashima, Y., Yanai, K.: SemArt: A Dataset for Semantic Art Understanding. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 259–267. ACM, New York (2018)
6. Shorten, C., Khoshgoftaar, T.M.: A Survey on Image Data Augmentation for Deep Learning. Journal of Big Data **6**(1), 1–48 (2019)
7. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-Trained CNNs Are Biased Towards Texture. In: Proceedings of the 7th International Conference on Learning Representations (2019)
8. Hendrycks, D., Dietterich, T.: Benchmarking Neural Network Robustness to Common Corruptions. In: Proceedings of the 7th International Conference on Learning Representations (2019)
9. Usama, M., Asim, S.A., Ali, S.B., Wasim, S.T., Mansoor, U.B.: Analysing the Robustness of Vision-Language Models to Common Corruptions. arXiv preprint arXiv:2504.13690 (2025)
10. Dahal, A., Murad, S.A., Rahimi, N.: Embedding Shift Dissection on CLIP: Effects of Augmentations on VLMs Representation Learning. arXiv preprint arXiv:2503.23495 (2025)