https://doi.org/10.2352/EI.2024.36.14.CVAA-177

Towards Artist Recognition Based on Material Rendering. A Case Study for Recognition of Rembrandt and Van Dyck

Jing Huang; Department of Big Data and Software Engineering, Chongqing University; Chongqing, China
Ahmed Elgammal; Department of Computer Science, Rutgers University; NJ, USA

Abstract

Can artists be recognized from the way they render certain materials, such as fabric, skin, or hair? In this paper, we study this problem with a focus on recognizing works by Rembrandt, Van Dyck, and other Dutch and Flemish artists from the same era. This paper proposes a novel material-based approach based on Swin Transformer and Cascade Mask R-CNN to address artist recognition task. We report the performance on a dataset of 644 images. Additionally, the model's robustness to image variations is studied.

Introduction

The identification of artists through their artwork, particularly in the context of paintings, has always been a challenging yet intriguing area of research in the field of art history and computer vision. Convolutional Neural Networks have been the primary tool for image feature extraction and pattern recognition in artworks. However, their effectiveness is often limited when discerning artists with similar styles, as well as in dealing with paintings that vary in condition and age. This limitation highlights the need for a more robust approach to artist identification.

In this context, Van Zuijlen et al.[8][12] underscore the importance of an artist's techniques in rendering materials such as skin and fabric. This approach could potentially be a key factor in artist recognition. This indicates that the way artists depict different materials might serve as a distinctive and robust feature, particularly valuable for identifying artists with similar styles. This concept is integral to developing a more effective method for artist recognition, focusing on the unique material rendering styles of artists.

Another challenge in this field is the processing of high-resolution images by deep learning models, coupled with the limited availability of extensive painting datasets. Our approach addresses this by segmenting paintings into multiple material-specific segments, thereby reducing the image size for processing and increasing the volume of data while maintaining the integrity of high-resolution details.

Therefore, this paper proposes a new aspect of artist recognition, focusing on the segmentation of material segments in paintings, specifically targeting the works of Rembrandt, Van Dyck, and other Dutch and Flemish artists of the same era. This research not only presents a novel method in artist recognition but also sets the stage for the development of an accessible artist recognition system.

Methodology

The artist recognition based on material segments can be viewed as a combination of two computer vision tasks: material instance segmentation and painting attribute recognition. Material instance segmentation involves learning a segment extractor $Seg(\cdot)$ that maps input images to k segments. Painting attribute recognition, on the other hand, focuses on learning a classifier $Cls(\cdot)$ that maps each segment to a specific attribute or category, resulting in k classification results. Finally, the results are integrated through an aggregation process $A(\cdot)$ to achieve a final artist recognition result p. The two-stage artist recognition problem based on material segments can be defined as

$$\mathrm{A(Cls(Seg(I)))} \rightarrow \mathrm{p} \tag{1}$$

For the segment extraction $Seg(\cdot)$, the Swin Transformer[9] is employed as backbone network. While Convolutional Neural Networks have traditionally dominated computer vision tasks, Transformers[transformers citation], known for their success in Natural Language Processing, have increasingly been adopted in this field due to their superior modeling capabilities and efficient parallel computation. Notably, the Swin Transformer has made significant improvements in aspects such as hierarchical structure, window mechanism, and training strategy, resulting in notable performance enhancements in image processing tasks. It outperforms traditional CNNs in capturing long-range dependencies in high-resolution images and demonstrates strong transfer learning abilities. By pretraining on large-scale general datasets (e.g., ImageNet-1K), the model can achieve impressive performance, effectively leveraging its learned features to handle the unique challenges posed by high-resolution artworks.

Swin Transformer offers a range of models with varying sizes to accommodate various computational needs and tasks, as detailed in Table 1, configurations P, C, and H represent the number of parameters, channels, and hidden layers, respectively. It adapts well to image segmentation by integrating additional segmentation heads (e.g., R-CNNs), demonstrating versatility and efficacy. Hence, the Swin Transformer is increasingly recognized as a key architecture widely used in applications like image segmentation and classification.

After feature extraction from painting images using the backbone network, the next step is to identify and extract materials within these images. Traditional object detection networks like Faster R-CNN[6] identify candidate bounding boxes of targeted objects, but they still encompass semantic information from adjacent pixels (As depicted in Figure 1 (left)). Instance segmentation, however, extends object detection to pixel-level segmentation of objects (As depicted in Figure 1 (right)).

Mask R-CNN[7] extends Faster R-CNN which combines efficient object detection with a mask branch for high-quality instance segmentation masks. Cascade Mask R-CNN[1][2] builds

Table 1: Description of Different Sizes of Swin Transformers

Size	Configuration	Description
Swin-Tiny	$P = 86M$, $C = 96$, $H = (2,2,6,2)$	Suitable for constrained resources.
Swin-Small	$P = 107M$, $C = 96$, $H = (2,2,18,2)$	Better performance with fewer resources.
Swin-Base	$P = 145M$, $C = 128$, $H = (2,2,18,2)$	Best results, larger version.

upon Mask R-CNN, further improving accuracy with a cascading structure. This structure includes three stages of progressively stringent filtering thresholds for candidate boxes, refining earlier predictions and culminating in detailed instance segmentation. Both Mask R-CNN and Cascade Mask R-CNN, as advanced instance segmentation methods, can efficiently segment material in paintings.

Two-stage Artist Recognition System

According to the methodology mentioned above, we propose a Two-stage Material-based Artist Recognition System. This System combines the functionalities of material segmentation and artist recognition in paintings, enabling artist recognition independent of contextual information while outputting material segments. The architecture of this system is illustrated in Figure 2.

The Two-stage Artist Recognition System contains four steps:

In Step 1, high-resolution painting images undergo preprocessing before inputted into the Swin Transformer backbone network, resulting in feature maps of the input images. The feature map is then fed into Cascade Mask R-CNN(or Mask R-CNN network) to obtain the binary mask images $M_k \in R^{H \times W}$ for n material segments, where $k \in (1,n)$. Additionally, the probability vector $Y_k = \{y_0, y_1, y_2, \ldots, y_c\}'$ is obtained for each segment, where c represents the total number of material categories, and $C_k = \max(Y_k)$ serves as the category label for the material segment. The configuration of the segmentation model mostly adhered to the settings in [?].

In Step 2, each binary mask image M_k is overlaid on the input image to retain only the material regions in the original image. A contour detection algorithm is used to detect the minimum bounding box for each material region, and the regions outside of these bounding boxes are cropped to obtain the material segments $S_k \in R^{H_n \times W_n \times 3}$. Each material segment is characterized by its respective dimensions H_k, W_k, representing the height and width of the minimum bounding box, as material segments can have varying dimensions.

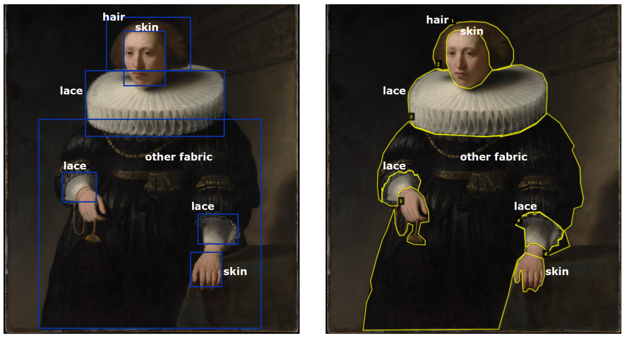

Figure 1. *Object Detection vs. Instance Segmentation*

In Step 3, each material segment obtained in Step 2 is inputted into another Swin Transformer classification model to obtain the artist recognition probability vector for each segment. The artist category with the highest probability is selected as the artist recognition result for each material segment, resulting in n artist recognition results $L = \{l_1, l_2, \ldots, l_n\}$.

In Step 4, the n artist recognition results $L = \{l_1, l_2, \ldots, l_n\}$ are integrated to obtain the final artist recognition result p. In this study, the model is trained using data from three categories of artists, allowing the model to recognize three categories of artists. Therefore, $l_i, p \in (0,1,2)$.

This artist recognition system follows a stage-wise optimization approach. The material segmentation stage and artist recognition stage are trained and evaluated separately.

Data Collection

Due to the requirement of materials segments in painting, a total of 644 portrait paintings are collected from different sources, including works by artists such as Van Dyck, Rembrandt, Gentileschi, and Hals. The paintings are categorized as *Van Dyck*, *Rembrandt*, and *Others*. Using VGG Annotator to perform instance-level annotation for five types of materials in the paint-

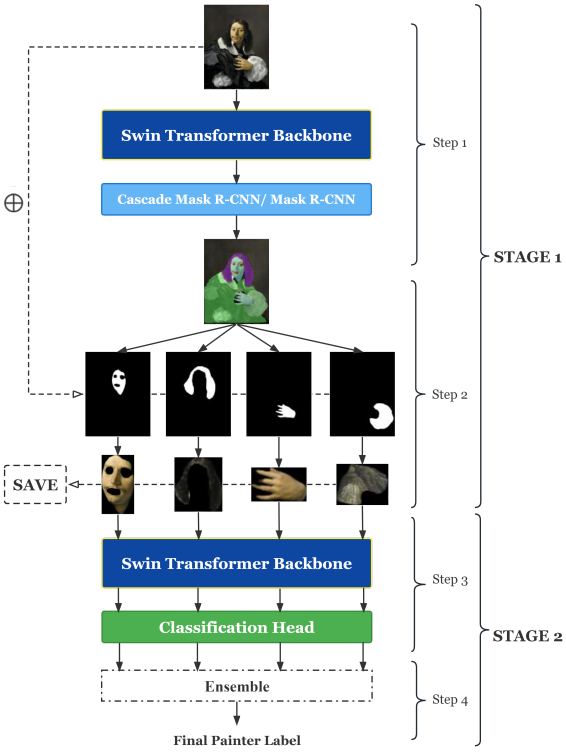

Figure 2. *The Architecture of Artist Recognition System.*

ings: *Hair*, *Skin*, *Lace*, *Fur*, and *Other Fabric*. This dataset is used to train and validate models in the material segmentation stage. Table 2 displays the number of paintings and resolution ranges for each category in the annotated painting dataset.

The extracted material segments from the annotated painting dataset are used to create a material segment dataset, which includes material category labels and artist labels, containing over 5000 material segments. This dataset is utilized to train and evaluate models in the artist recognition stage. The material labels consist of *Hair*, *Skin*, *Lace*, and *Other Fabric*. Table 3 displays the distribution of material segments. It is evident that utilizing material segments significantly increases the available data volume. Examples of some material segments from the dataset are shown in Figure 3.

Table 2: Distribution of Annotated Painting Dataset

Category	Number of Paintings	Resolution Range
Van Dyck	177	Min: 361× 453 Max: 5898×7324
Rembrandt	302	Min: 2612×3267 Max: 11148×14348
Others	165	Min: 1421×1801 Max: 8688×8219
Others: Abraham de Vries, Frans Hals, van der Helst, Petronella Elias, Cornelis Jonson, Jan de Baen, Carel Fabritius, etc.		

Table 3: Distribution of Material Segment Dataset

Category	Hair	Skin	Lace	Other Fabric	Total
Rembrandt	865	836	335	1052	3088
Van Dyck	244	382	151	568	1345
Others	308	464	180	435	1387
Total	1417	1682	511	1800	5410

Material Segmentation

The material segmentation stage utilizes the Swin Transformer backbone in combination with the Cascade Mask R-CNN instance segmentation model. To achieve the optimal balance between performance and computational cost, three different sizes of the Swin Transformer are tested as the backbone network. All experiments are conducted using MMDetection[3], and the experimental setup largely follows the default settings, with small adjustments made solely based on computational resource limitations. This stage is trained and evaluated on the annotated painting dataset.

Additionally, to determine the best material categories, attempts are made to segment *Hair*, *Skin*, and *Fabric* in the *Van Dyck* category. The results are presented in Table 4. and further divide *Fabric* into *Fur*, *Lace*, and *Other Fabric* in *Rembrandt* and *Others* categories. The results is presented in Table 5. Using Mask AP (Mask Average Precision) as the evaluation metric. Mask AP is computed based on the Intersection over Union (IoU) between predicted masks and ground truth masks.

Table 4: Segmentation Results on Van Dyck(%)

Backbone	Skin	Hair	Fabric
Swin-T	59.0	38.4	42.5
Swin-S	56.1	36.4	40.3
Swin-B	58.5	37.3	45.6

Table 5: Segmentation Results on Rembrandts and Others(%)

Backbone	Skin	Hair	Lace	Fur	Other Fabric
Swin-T	48.4	34.5	58.3	9.3	45.9
Swin-S	49.2	33.0	64.2	0.8	46.6
Swin-B	48.3	35.9	57.5	1.7	42.1

Based on the segmentation results, it is evident that the vast majority of materials achieved a level of performance comparable to the state-of-the-art in the instance segmentation task. The higher accuracy in the Van Dyck category can be attributed to its relatively smaller quantity compared to the combined quantity of *Rembrandt* and *Others* categories. Notably, the segmentation accuracy for lace is the highest, and the accuracy for fabric segmentation also shows significant improvement, affirming the appropriateness of subdividing fabric.

However, the segmentation accuracy for *Fur* and *Hair* are relatively low. This can be explained by visualizing the segmentation results in Figure 4. It appears that the model has difficulty distinguishing *Hair* (Purple Region) and *Fur* (Blue Region), leading to the observed decrease in their accuracy. Therefore, fur will no longer be classified as a separate material category.

Additionally, Figure 5 shows the segmentation results of Swin-T, Swin-S, and Swin-B on the same painting. Swin-T exhibits rugged edge detection and fails to accurately identify *Hair* regions. and its *Skin* segmentation fails to avoid overlapping

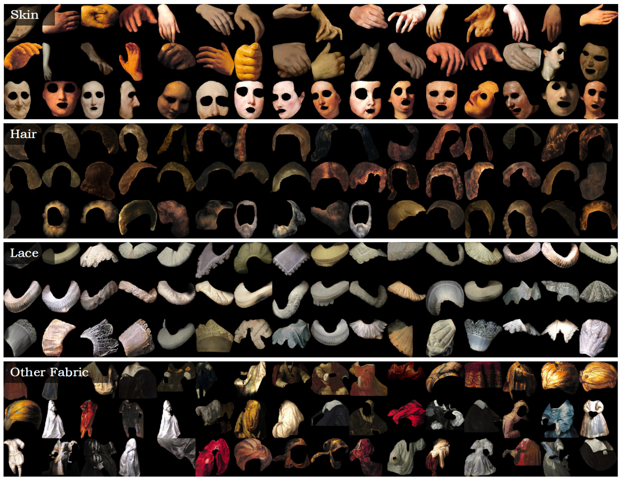

Figure 3. *Examples of images in Material Segment Dataset.*

with the regions of the eyes and mouth. Similarly, in Swin-B results, the hair is also not accurately detected. Meanwhile, Swin-S demonstrated notably superior segmentation results compared to the other two backbones, establishing itself as the most appropriate choice for the material segmentation stage in the artist recognition system.

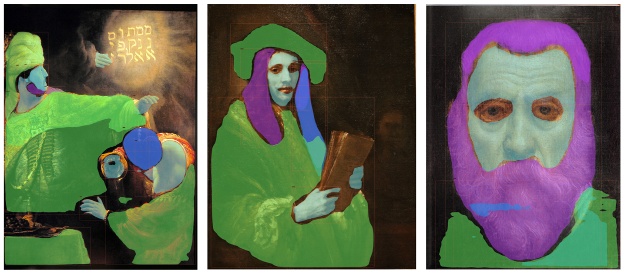

Figure 4. Confusion between Fur and Hair.

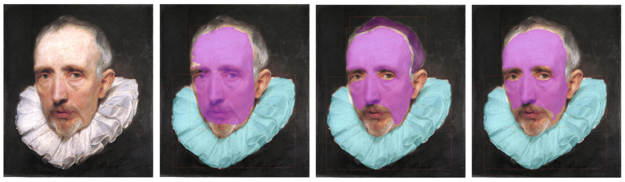

Figure 5. Segmentation Results of Swin-T vs. Swin-S vs. Swin-B.

Artist Recognition

Based on the material segmentation results, the artist recognition stage utilizes the Swin-S backbone. To achieve class balance, the material segment dataset undergoes downsampling, where 245 material segments are extracted from each category to form the test set. Additionally, 1100 material segments are randomly selected from the remaining data in each category to perform 5-fold cross-validation. The average results from the 5-fold cross-validation and the results on the test set are reported in Table 6. Note that during this stage, the model learns to recognize artists based on individual material segments.

The results from the 5-fold cross-validation and the test set demonstrate that the performance gaps across various evaluation metrics are within 2.5%, indicating the model's robust generalization capability. On the validation set, the model achieve Top-1 Accuracy, Precision, Recall, and F1-score all surpassing 93%, while on the test set, all metrics exceeded 91%. This verifies the feasibility of artist recognition based solely on material segments. The confusion matrix results on the test set, as depicted in Figure 6, show that the model's confusion is not concentrated solely between specific pairs of categories but rather evenly distributed among different categories. This finding verifies the feasibility of artist recognition based solely on material segments.

Additionally, the performance of three other backbone networks, Swin-B, ViT-Base[4], and RestNet101[5], are also evaluated and compared to Swin-S in the artist recognition stage. The results of the test set for different networks are reported in Table 6.

Swin Transformer-based models exhibit faster convergence and higher recognition accuracy compared to ViT-Base and RestNet101. The results in Table 6 further confirm the superior performance of Swin-S on the test set, achieving comparable or even better artist recognition accuracy than Swin-B, while having a smaller parameter size.

Two-staged artist recognition system, utilizing Swin-S + Cascade Mask R-CNN for the segmentation stage and Swin-S for the classification stage, achieves remarkable results in the task of painter recognition solely based on material segments. Figure 7 showcases the impressive outcomes of this model on artworks from three artist categories: Others, Rembrandt, and Van Dyck

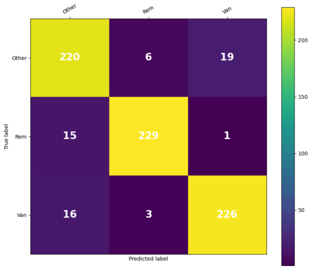

Figure 6. Confusion Matrix of Swin-S.

Table 6: Artist Recognition Results(%)

Backbone	Top-1 Acc	Precision	Recall	F1-score
Swin-S(Val)	93.52	93.60	93.52	93.54
Swin-S(Test)	91.97	92.16	91.97	92.00
ResNet101	88.02	88.02	88.02	88.00
ViT-Base	90.20	90.33	90.20	90.20
Swin-B	91.29	91.52	91.29	91.35

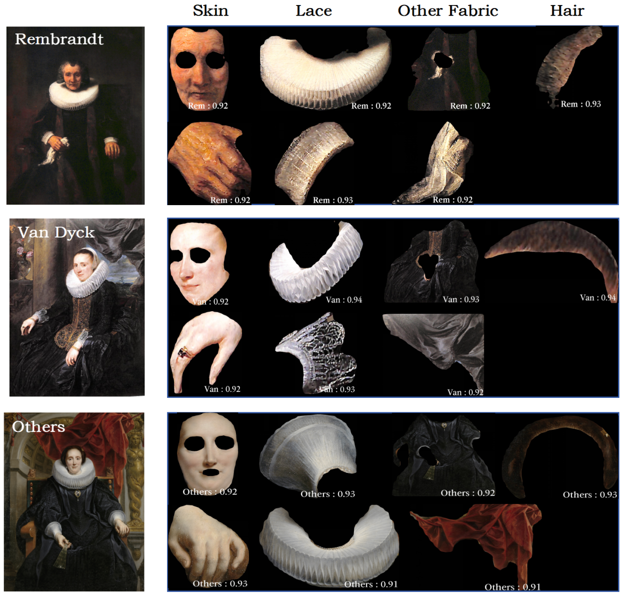

Figure 7. Example of the Segmentation and Recognition Results.

Validation Experiment

The high accuracy of the classification results does not entirely serve as an endorsement for the artist recognition efficacy of the system based on material segments. Given that deep learning classification models operate as black-box models, there remains a possibility that the model could learn non-robust features (such as shape, color, etc.) from material segments for classification purposes. Moreover, a robust painter recognition model should demonstrate its ability to handle variations in image quality, such as differences in brightness and image degradation due to scanning techniques or aging. The Two-staged artist recognition system, with its material-based recognition approach, is designed to be resilient to such interference. Therefore, a series of experiments are conducted to validate the interpretability and robustness of the artist recognition system.

Interpretability

GradCAM[10] (Gradient-weighted Class Activation Mapping) is a visualization technique for image processing, primarily used to understand how areas of an image contribute to the final decision-making by calculating their gradients. This technique is immensely helpful in interpreting the decision-making process of deep learning models, particularly in image recognition and classification tasks. Figure 8 presents the GradCAM visualizations of the artist recognition model on a material segment. It is evident that the model attends to regions containing features related to the artist's depiction (e.g., brushstrokes) of the material, demonstrating that the recognition model relies on these robust features to identify the artists.

t-SNE[11] (t-Distributed Stochastic Neighbor Embedding) is a data dimensionality reduction and visualization method that maps high-dimensional data into a lower-dimensional space while preserving the similarity structure of the data. Utilizing t-SNE visualization enables observation of the feature vectors output by the recognition model's backbone into a two-dimensional space, determining whether the model can effectively differentiate painting materials of different categories within this two-dimensional feature space. Figure 9 shows the t-SNE visualization results of 3 different scenarios.

Figure 9 (left) shows the t-SNE visualization results of the 3 painter recognition, indicating that the model accurately separates samples of the three painter categories in the feature space. Figure 9 (middle) displays the t-SNE visualization results for the 12 materials (3 Artists × 4 Materials). Combining the 3 artists and 4 materials to form a 12-class dataset to train and test the same artist recognition model. The resulting visualization showcases the feature vectors from the test set, with different colors and shapes representing distinct material categories and artists, respectively. It is evident that materials of the same material (same color) are closer in the feature space.

Additionally, within the same material category, the model achieves accurate differentiation between different artists (distinct shapes). From the perspective of artists, Other Fabric and Lace belong to a similar category. A 6-class model is trained and tested for the 3 Artists × 2 Materials scenario. The resulting t-SNE visualization is presented in Figure 9 (right), where colors represent the same artist and shapes represent materials. It can be observed that instances of Other Fabric and Lace from the same artist are closer in the feature space compared to instances of the same material type from different artists. This observation indicates that the model can recognize common features between different materials from the same artist. This demonstrates the model's capability to accurately extract artist-specific features from material segments.

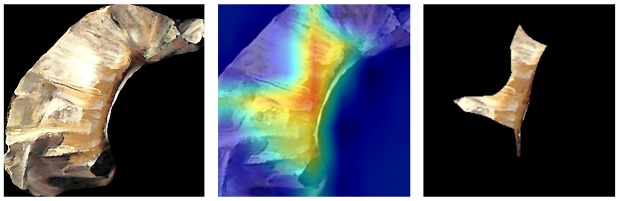

Figure 8. Original Segment vs. GradCAM Visualization vs. Salient Area.

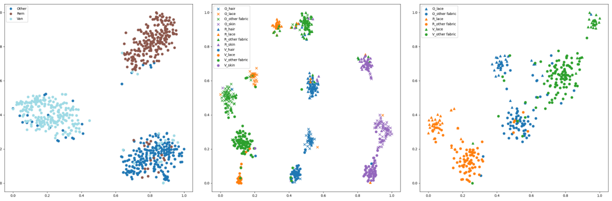

Figure 9. t-SNE Visualization Results.

Robustness on Brightness

To verify the system's robustness on brightness, a set of 150 high-resolution painting images (50 per category) is randomly sampled from the annotated painting dataset. Majority Voting is employed to aggregate material segment classification results, and the recognition accuracy is reported on image-level. In consideration of universality and computational resource limitation, the robustness test utilizes the Swin-T Backbone in combination with the Mask R-CNN to perform material segmentation. Subsequently, robustness tests are conducted on aspects related to brightness and resolution.

Firstly, the image set is brightened (or darkened) by 10%, 30%, and 50% respectively, and then fed into the artist recognition system. The visual effects of the brightness-adjusted images are shown in Figure 10. The outcomes of the recognition accuracy under various conditions are presented in Table 7. Upon further analysis, it is observed that the majority of Others category images become too dark to effectively detect material fragments, resulting in lower accuracy. Concerning the Van Dyck category, certain images exhibit distortion after a 50% brightening, contributing to a slight decrease in accuracy. However, for the vast majority of scenarios, notably high recognition accuracy is achieved, indicating the robustness of the artist recognition system in the face of brightness alterations in painting images.

Robustness on Resolution

To assess the robustness of the artist recognition system across various resolution ranges, the same set of sampled images undergoes adjustments in resolution. The image resolution reductions of 10%, 30%, 50%, 70%, and 90% are shown in Figure 11. The results are presented in Table 8. Notice that during the inference stage pre-processing, the input dimensions of the painting images are uniformly scaled to $H' = 480$, $W' = 1333$.

It is evident that even when image resolution is reduced by 70%, the outcomes of the artist recognition system maintain a high recognition accuracy, signifying the considerable robustness of the artist recognition system across a broad spectrum of resolutions. This is a vital trait, particularly for the processing of high-resolution painting images, as it suggests the potential for alleviating computational burden by downsampling image resolution.

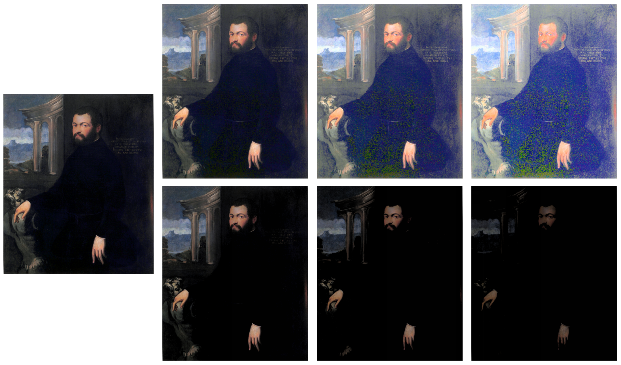

Figure 10. *Examples of Brightness-adjusted Image.*

Figure 11. *Example of Image Scale of Resolution-adjusted Images.*

Table 7: Robustness Test Results on Brightness

	Original	-10%	-30%	-50%
Others Rembrandt Van Dyck	50 49 48	50 49 45	42 49 43	23 47 40
Accuracy (%)	98.00	96.00	89.33	73.33
	+10%	+30%	+50%	
Others Rembrandt Van Dyck	48 47 45	48 46 45	48 44 38	
Accuracy (%)	93.33	92.67	86.67	

The results demonstrate the artist recognition system's ability to maintain performance even under these challenging conditions.

Table 8: Robustness Test Results on Resolution

	Original	-10%	-30%
Others Rembrandt Van Dyck	50 49 48	50 48 47	50 48 47
Accuracy(%)	98.00	96.67	96.67
	-50%	-70%	-90%
Others Rembrandt Van Dyck	50 47 46	48 47 48	43 42 50
Accuracy(%)	96.00	96.00	81.33

Conclusion

In this study, we developed a novel Two-stage Material-based Artist Recognition System, utilizing the Swin Transformer and Cascade Mask R-CNN to identify artists from material segments within high-resolution paintings. Our approach effectively addresses the challenges of traditional CNN-based methods, particularly in distinguishing artists with similar styles. The system's robustness was thoroughly evaluated against variations in image conditions such as brightness and resolution, demonstrating consistent performance. Through extensive experiments and interpretability analyses using techniques like GradCAM and t-SNE visualizations, we were able to understand the model's focus on specific material attributes for artist recognition. This research contributes to the field by providing a new methodological perspective for artist identification, combining art historical knowledge with advanced computer vision techniques. It offers a promising direction for future studies in digital art analysis and has potential applications for art historians and digital archivists in understanding and categorizing artistic works.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019.

[3] K Chen, J Wang, J Pang, Y Cao, Y Xiong, X Li, S Sun, W Feng, Z Liu, J Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. arxiv 2019. *arXiv preprint arXiv:1906.07155*, 2019.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019.

[6] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[8] Hubert Lin, Mitchell Van Zuijlen, Maarten WA Wijntjes, Sylvia C Pont, and Kavita Bala. Insights from a large-scale database of material depictions in paintings. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15,*

2021, Proceedings, Part III, pages 531–545. Springer, 2021.

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[10] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[11] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[12] Mitchell JP van Zuijlen, Sylvia C Pont, and Maarten WA Wijntjes. Painterly depiction of material properties. *Journal of vision*, 20(7):7–7, 2020.

Author Biography

Jing Huang obtained her B.S. in AI from Chongqing University (2022), and now is a graduate student at NYU Tandon School of Engineering. Her research interest lies in computer vision and generative AI.

Dr. Ahmed Elgammal of Rutgers University, received his Ph.D. in computer science from the University of Maryland (2002). He leads the Art and AI Lab, integrating AI with art history, and is distinguished by his NSF CAREER Award