Material-Aware Style Transfer for Lacquer Paintings

Thi-Ngoc-Hanh Le
International University
VNU-HCM
Ho Chi Minh, Viet Nam
Email: ltnhanh@hcmiu.edu.vn

Khang Gia Pham
International University
VNU-HCM
Ho Chi Minh, Viet Nam
Email: ITCSIU21135@student.hcmiu.edu.vn

Sinh Van Nguyen
International University
VNU-HCM
Ho Chi Minh, Viet Nam
Email: nvsinh@hcmiu.edu.vn

***Abstract*—Image style transfer aims to render the semantic style of a reference image onto a target image while preserving its original content. While early methods based on convolutional neural networks enabled fast stylization, they often suffered from content leakage and limited style fidelity. Recent Transformer-based approaches improve global consistency and artifact reduction, but largely treat style transfer as medium-agnostic. In contrast, this work focuses on the underexplored domain of medium-specific stylization specifically, Vietnamese lacquer painting, an art form characterized by layered texture, strong tonal contrast, and spatial depth. We build our LST (Lacquer Style Transfer) model via a transformer-based framework and introduce two novel loss functions, Contrast Loss and Depth Loss, to better capture the material-aware properties of lacquer artwork. Quantitative and qualitative results demonstrate that our model achieves improved stylization quality and content structure preservation over prior methods. By enabling faithful digitization and simulation of lacquer painting through AI, our work supports cultural heritage preservation and opens new possibilities for education, digital art, and creative industries.**

I. Introduction

Neural style transfer is a subfield of computer vision that aims to synthesize an image by combining the content of one image with the artistic style of another. Since its inception, this topic has applications ranging from mobile photo filters to digital content creation. However, most existing approaches focus on general-purpose artistic styles such as oil painting, sketch, and watercolor, leaving culturally specific art forms largely unexplored. Vietnamese lacquer painting is a traditional medium that embodies deep cultural and aesthetic values, as shown in Fig.1. Characterized by its glossy surface, layered textures, and sharp tonal contrasts, lacquer painting presents unique challenges for neural stylization. Unlike typical Western art forms, lacquer artworks often employ optical layering and gold leaf accents to simulate depth and vibrance, making them particularly difficult to emulate using standard style transfer techniques.

Recent advances in vision transformers [1], [2], [3], [4], [5], have improved the capacity of style transfer models to preserve both global structure and local texture. However, these models remain agnostic to domain-specific traits. Additionally, current loss functions in style transfer often rely solely on perceptual and statistical similarity, which may not suffice for capturing material-specific features such as depth, luminance contrast, or surface reflection.

Fig. 1. A lacquer painting generated by our LST model.

To address these gaps, we propose a transformer-based style transfer model tailored to Vietnamese lacquer painting, so-called **LST** model. Our approach is built upon the S2WAT architecture [1] and introduces two novel loss functions: contrast loss and depth loss. Contrast Loss enhances the model's ability to preserve sharp luminance transitions, while depth loss encourages spatial layering in the stylized outputs. Together, these losses enable the model to generate images that more faithfully replicate the visual and structural aesthetics of lacquer art. We curate a dataset of authentic lacquer paintings and employ targeted data augmentation techniques to overcome data scarcity. Our experiments demonstrate that the proposed method outperforms existing baselines in both visual fidelity and perceptual realism. Moreover, ablation studies validate the individual contributions of our proposed loss functions.

The contributions of this work are summarized as follows:

- We introduce a novel framework for style transfer that specifically targets the unique visual language of Vietnamese lacquer painting.
- We propose two domain-specific loss functions: contrast loss for enforcing tonal sharpness and depth loss for enhancing spatial layering.
- We construct a curated lacquer painting dataset and demonstrate the effectiveness of our model through comprehensive qualitative and quantitative evaluations.

To the best of our knowledge, this is the first work to focus on lacquer painting stylization in the context of neural image synthesis. Our findings contribute to the broader goal of applying AI to cultural heritage preservation and material-aware image generation.

II. Related Work

Since the seminal work of Gatys et al.[6], which showed that content and style could be modeled using perceptual features and Gram matrices in CNNs, numerous methods have been developed to improve stylization quality using content and style losses [7], [8], [9]. These CNN-based approaches [7], [8], [9] achieved promising results by designing feed-forward networks to fuse content and style features efficiently. More recently, transformer-based models [10], [2], [4], [11], [12], [1] have leveraged the global receptive field of attention mechanisms to enhance stylization expressiveness and semantic consistency. However, traditional Gram matrix–based style losses rely on second-order statistics over entire feature maps, which can be insufficient for capturing localized or high-frequency style details. To address this, contrastive learning has been introduced to replace or augment Gram-based losses [10], [13], [14], enabling better representation of fine-grained stylistic patterns. Despite this progress, achieving precise, controllable style transfer remains an open challenge. Our work contributes to this direction by proposing a method that improves style fidelity without relying on handcrafted style constraints.

To extend the art medium of neural style transfer models, some recent works explore to handle pencil painting [15], [16], [17], or utilizing the integration of geometric and material-specific cues into the style transfer process [18], [19]. Specifically, diffusion models have recently gained attention in the field of image generation and style transfer for their ability to synthesize high-fidelity and controllable outputs [20], [21], [22], [23]. These models often leverage pre-trained latent diffusion backbones and integrate text, image, or contrastive guidance for zero-shot or few-shot stylization. While diffusion approaches show promise in capturing complex artistic styles, their high computational cost and slower inference remain practical challenges. Our work complements this line of research by focusing on transformer-based stylization that is lightweight yet capable of capturing material-aware features, such as the tonal contrast and layered texture of lacquer art.

III. Methodology

A. System Overview

The proposed framework is a transformer-based neural style transfer model, specifically designed to capture the distinct visual traits of Vietnamese lacquer painting. As illustrated in Fig. 2, our LST model takes as input a content image I^c and a lacquer style image I^s, and outputs a stylized image $\mathcal{R}$ that retains the structural content of I^c while emulating the visual aesthetics of I^s.

The architecture of our LST follows the encoder-transfer-decoder paradigm. The content and style images are independently processed by a shared encoder to extract hierarchical feature representations, *i.e,* $\mathcal{F}^c$ and $\mathcal{F}^s$. These features are then fed into the Transfer Module, which is composed of multiple transformer decoder layers. This module captures both local details and global dependencies across the content-style domain. The transformed features are then passed to a VGG-inspired Decoder, which reconstructs the final stylized output image $\mathcal{R}$. This decoder is responsible for upsampling and synthesizing spatially coherent visual textures that reflect the lacquer painting style.

The model is optimized through a weight sum of four loss functions: (1) a content loss that preserves semantic structure, (2) a style loss that aligns the stylistic distribution of features, (3) a contrast loss designed to enhance sharp luminance transitions characteristic of lacquer painting, (4) and a depth loss that promotes spatial layering. This multi-objective optimization allows the network to produce stylized images that are both aesthetically faithful and structurally grounded-preserving content geometry while simulating the high-contrast, layered textures of lacquer art.

B. Model Design

Encoder: Our encoder adopts the S2WAT framework [1], using a hierarchical transformer with patch-based tokenization. Each input image is partitioned into non-overlapping 2×2 patches, flattened to 12-dimensional tokens, and linearly projected to a latent space of size C. Tokens are processed through a sequence of Strip Window Attention (SpW) blocks, with reflection padding applied before each block to prevent boundary artifacts and removed afterward. This forms the first stage of the encoder, maintaining resolution at $\frac{H}{2} \times \frac{W}{2}$.

To build multi-scale features, the encoder performs patch merging between stages: 2×2 neighboring patches are concatenated and projected to reduce channel dimensions (from $4C$ to $2C$), achieving twofold spatial downsampling. This process is repeated across three stages, resulting in resolutions of $\frac{H}{2^i} \times \frac{W}{2^i}$, $i \in \{1, 2, 3\}$. This hierarchical structure captures both local detail and global context, enabling effective integration with decoder modules.

Transfer Module: The transfer module adopts a multi-layer Transformer decoder, closely following the architecture proposed by Vaswani et al. [24]. Each decoder layer comprises two sub-blocks: a multi-head self-attention (MSA) mechanism and a feed-forward network (MLP), both wrapped with residual connections and preceded by Layer Normalization (pre-norm configuration). This design differs from StyTr^2 [2] and the baseline S2WAT [1]. By using MSA instead of cross-attention, the module promotes more global and coherent fusion of content and style features. Dropout, stochastic depth, and GELU activations are incorporated throughout the decoder to enhance generalization. The output is reshaped into spatial feature maps and passed to the decoder for reconstruction.

This design is particularly well-suited for lacquer painting, where long-range dependencies play a critical role in distributing consistent stylistic patterns across layered regions. The use of MSA allows the network to model high-level stylistic coherence, such as large-scale tonal gradients and reflective layering, which are common in lacquer art. Furthermore, the pre-norm setup contributes to training stability, enabling the model to better learn subtle luminance and structural cues characteristic of lacquer aesthetics.

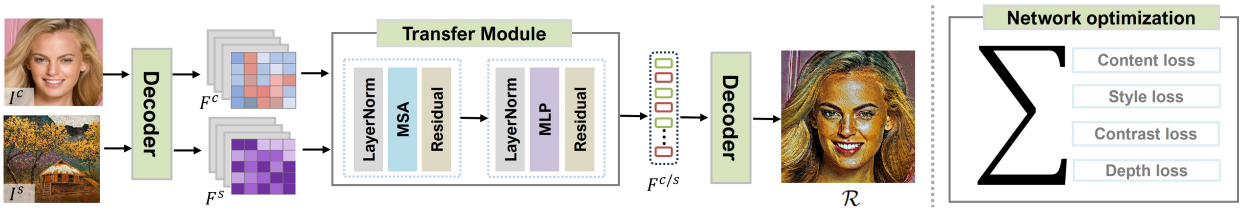

Fig. 2. Our proposed framework and network optimization strategy for lacquer paintings

Decoder: To reconstruct the stylized output image, we employ a VGG-inspired decoder that mirrors the structure of the encoder, inline with prior work [9], [19], [25], [16], [1]. The decoder takes as input the fused feature maps $\mathcal{F}^{c/s}$ from the transfer module and progressively upsamples them to the original image resolution. It consists of stacked convolutional layers interleaved with ReLU activations and nearest-neighbor upsampling blocks. Reflection padding is used to reduce boundary artifacts. The latent features are decoded from the deepest stage (channel size C) through successive blocks that reduce channels from $256 \rightarrow 128 \rightarrow 64$, ultimately producing a 3-channel RGB output. This design enables effective reconstruction while preserving stylized textures.

C. Loss function

Vietnamese lacquer paintings are characterized by stark tonal contrasts and distinct luminance transitions. Compared to our baseline S2WAT, we introduce two novel loss function: contrast loss and depth loss to enforce this visual property and encourage local contrast similarity between the generated image $\mathcal{R}$ and the reference style image I^s. Detail of each loss function in our network optimization is elaborated as follows.

Content Loss: Following the formulation proposed by Huang and Belongie [9], we define perceptual loss function to quantify the content between the stylized image $\mathcal{R}$ and the input content image I^c. The content perceptual loss is defined as the squared difference between the feature representations of $\mathcal{R}$ and I^c, extracted from selected layers of a pre-trained VGG19 network:

$$\mathcal{L}_{\text{content}} = \sum_{l \in \mathcal{C}} \|\phi_l(\mathcal{R}) - \phi_l(I^c)\|_2^2, \tag{1}$$

where $\phi_l(\cdot)$ denotes the activation at layer l, and $\mathcal{C} = \{\text{relu4\_1}, \text{relu5\_1}\}$ is the set of selected layers used for content comparison.

Style Loss: To assess stylistic consistency, we adopt a style loss that compares the channel-wise statistics, mean and standard deviation, of the feature activations between the stylized image $\mathcal{R}$ and the style image I^s. Specifically, the style loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{style}} = \sum_{l \in \mathcal{S}} &\|\mu(\phi_l(\mathcal{R})) - \mu(\phi_l(I^s))\|_2^2 \\ &+ \|\sigma(\phi_l(\mathcal{R})) - \sigma(\phi_l(I^s))\|_2^2,\end{aligned} \tag{2}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the channel-wise mean and standard deviation, respectively, and $\mathcal{S} = \{\text{relu2\_1}, \text{relu3\_1}, \text{relu4\_1}, \text{relu5\_1}\}$ is the set of layers used for style comparison.

Contrast Loss: To better capture the high-contrast visual characteristics of lacquer painting, we introduce a contrast-aware loss that explicitly preserves local tonal variation. Unlike global feature statistics used in traditional style losses, local contrast provides fine-grained control over sharp luminance transitions and edge-aware textural changes essential properties of lacquer art, where light dark interplay and color-plane separation are critical.

Standard neural style transfer models often produce outputs with excessive smoothness due to global averaging effects. To address this, our contrast loss encourages the stylized output to mimic the local contrast distribution of the reference style image, preventing the loss of fine structure and visual sharpness. Fig.3 demonstrates step-by-step the calculation of our contrast loss. We compute local contrast using a statistical formulation based on the standard deviation of pixel intensities within a local window of size $k \times k$. For each spatial location (x, y) and channel c, we first calculate the local mean:

$$\mu(x, y, c) = \frac{1}{k^2} \sum_{i=-\frac{k}{2}}^{\frac{k}{2}} \sum_{j=-\frac{k}{2}}^{\frac{k}{2}} I(x+i, y+j, c), \tag{3}$$

followed by the local mean of squared intensities:

$$\mathbb{E}[I^2](x, y, c) = \frac{1}{k^2} \sum_{i=-\frac{k}{2}}^{\frac{k}{2}} \sum_{j=-\frac{k}{2}}^{\frac{k}{2}} I(x+i, y+j, c)^2. \tag{4}$$

The local variance is then obtained as:

$$\sigma^2(x, y, c) = \mathbb{E}[I^2](x, y, c) - \mu(x, y, c)^2, \tag{5}$$

and the final contrast map is computed by taking the square root:

$$\sigma(x, y, c) = \sqrt{\max(\sigma^2(x, y, c), 0)}. \tag{6}$$

Let σ_{gen} and σ_{ref} denote the contrast maps of the generated and reference style images, respectively, the contrast loss is defined as the mean squared error between the two:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{N} \sum_{x,y,c} \left(\sigma_{\text{gen}}(x, y, c) - \sigma_{\text{ref}}(x, y, c)\right)^2, \tag{7}$$

where N is the total number of pixels across all channels. This loss helps preserve sharp local differences in brightness and prevents stylistic degradation in highly textured regions, qualities essential for the stylization of lacquer paintings.

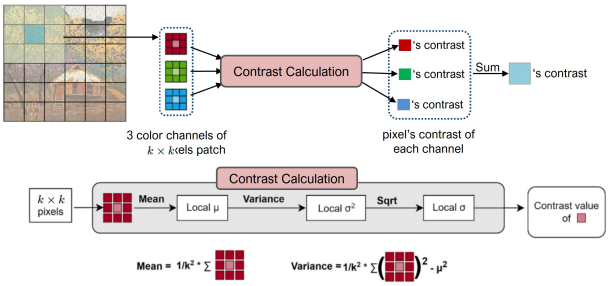

Fig. 3. Visualization of our contrast loss calculation

Depth Loss: To incorporate structural awareness into the stylization process, we introduce a depth loss that encourages the stylized image to maintain spatial consistency with the reference image's depth structure. This is particularly important for preserving visual layering and object separation, features commonly found in lacquer paintings with strong compositional hierarchies. Depth maps are estimated using a pre-trained monocular depth prediction model. We use MiDaS, a robust monocular depth estimation from a single image proposed by [26], in our experiments. Given a single RGB image, the model produces a dense depth representation where each pixel value encodes the relative distance from the viewpoint. The resulting depth maps for both the stylized and reference images are normalized to the range $[0, 1]$ for numerical stability.

The depth loss consists of two components. The first is the pixel-wise L1 loss, which measures the absolute difference between the predicted depth maps of the stylized output D_{gen} and the reference image D_{ref}:

$$\mathcal{L}_{\text{L1}} = \frac{1}{HW} \sum_{x=1}^{H} \sum_{y=1}^{W} \left| D_{\text{gen}}(x, y) - D_{\text{ref}}(x, y) \right|, \tag{8}$$

where H and W denote the height and width of the depth maps.

To further improve spatial consistency around edges and object boundaries, we introduce a depth gradient loss. This component minimizes the difference in horizontal and vertical gradients between the two depth maps:

$$\mathcal{L}_{\text{grad}} = \frac{1}{HW} \sum_{x,y} \big(\left| \nabla_x D_{\text{gen}}(x, y) - \nabla_x D_{\text{ref}}(x, y) \right| + \left| \nabla_y D_{\text{gen}}(x, y) - \nabla_y D_{\text{ref}}(x, y) \right| \big), \tag{9}$$

where ∇_x and ∇_y represent the first-order gradients along the horizontal and vertical directions, respectively.

The final depth loss combines the L1 term and the gradient term using a weighting coefficient $\lambda_{\text{grad}} \in [0, 1]$:

$$\mathcal{L}_{\text{depth}} = \mathcal{L}_{\text{L1}} + \lambda_{\text{grad}} \cdot \mathcal{L}_{\text{grad}}. \tag{10}$$

This formulation promotes both global depth alignment and local structural sharpness, allowing the model to better preserve the scene layout and layering effects prominent in lacquer painting compositions.

D. Total Objective Function

The complete training loss integrates content, style, and our proposed losses:

$$\mathcal{L}_{\text{total}} = \lambda_c \mathcal{L}_{\text{content}} + \lambda_s \mathcal{L}_{\text{style}} + \lambda_{\text{contrast}} \mathcal{L}_{\text{contrast}} + \lambda_{\text{depth}} \mathcal{L}_{\text{depth}}. \tag{11}$$

Empirically, we find the best performance when setting $\lambda_c = 2$, $\lambda_s = 3$, $\lambda_{\text{contrast}} = 3$, and $\lambda_{\text{depth}} = 0.5$. This combination promotes both stylistic fidelity and structural preservation in the lacquer stylization task. Our architectural modifications and domain-specific objectives position the proposed method as a specialized yet generalizable solution for culturally aware neural stylization.

IV. Experimental Results

A. Implementation Details

Our dataset consists of 70 traditional Vietnamese lacquer paintings and 400 aligned face images from the CelebA dataset [27]. Style images were collected from online sources, museum catalogs, and artist portfolios. Given the limited availability of digitized lacquer art, we applied data augmentation, such as geometric transforms, color jittering, and noise, to expand the style set to 300 samples. We visualize augmented images in our experiments in Fig.4. This improved the model's ability to capture layered gloss, high contrast, and textured detail. The content subset from CelebA was chosen for its diversity in pose and lighting while keeping training computationally feasible.

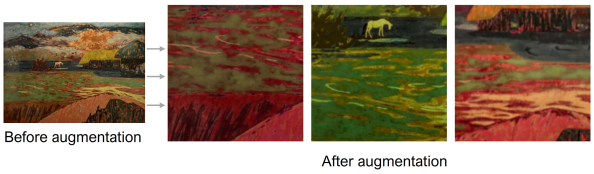

Fig. 4. Sample augmented data from our experiments

B. Results and Evaluations

To demonstrate that our approach advances prior work in lacquer style transfer, we compare it with three recent state-of-the-art style transfer models, S2WAT [1], StyTR^2 [2], and CAST [14]. The visual comparisons are exhibited in Fig. 5, in which each row presents a different content-style pair. Obviously from observation, our approach produces stylized results that best preserve facial structure, edge clarity, and spatial coherence, while effectively capturing the layered gloss and tonal contrast characteristic of Vietnamese lacquer paintings. In contrast, S2WAT suffers from color noise and content distortion. StyTR^2 produces blurred or overly abstract textures, resulting in a loss of facial identity. Meanwhile, CAST is expressive in style, tends to exaggerate features and lacks content fidelity. Overall, our model strikes a better balance between stylization richness and structural preservation,

Fig. 5. Our model competes with SOTA style transfer models on lacquer paintings.

demonstrating its suitability for material-sensitive tasks like lacquer art transfer.

A part from above visual comparisons, to assess content structure preservation in stylized outputs, we adopt two standard metrics: Structural Similarity Index (SSIM) and Region Coverage Rate (RCR), followed to evaluation by Le et al. [16]. SSIM measures structural similarity between the stylized image and the original content, particularly effective when computed on edge maps. RCR is defined as RCR $= \frac{F_s \wedge F_c}{F_s \vee F_c}$, where F_s and F_c are the binary edge maps of the stylized and content images, respectively. For this accessment, we compare our method against three state-of-the-art baselines, S2WAT [1], StyTR^2 [2], and CAST [14], on a test set of 15 stylized results. As shown in Table I, our method achieves the highest SSIM and RCR scores, indicating better structural consistency and content preservation. These results highlight the effectiveness of our model in preserving key content features while achieving high-quality lacquer-style rendering.

TABLE I
LACQUER STYLE TRANSFER QUALITY ANALYSIS

	Ours		S2WAT		StyTR^2		CAST	
Metrics	SSIM	RCR	SSIM	RCR	SSIM	RCR	SSIM	RCR
Avg.	**0.93**	**0.71**	0.76	0.62	0.89	0.67	0.84	0.69

To validate the impact of our proposed loss functions, we conduct ablation studies on contrast loss and depth loss, the key factors enable our model outperform prior work in lacquer style transfer. The ablated results are presented in Fig. 6. We can observe that adding contrast loss enhances tonal separation and local luminance variation, producing clearer object boundaries and stronger visual contrast - hallmarks of lacquer painting. Without it, the outputs appear flatter and less defined. Depth loss improves spatial hierarchy, reinforcing foreground-background separation and layered composition. Its absence leads to blurred boundaries and weaker depth cues. Combined both losses significantly enhance structure preservation and stylistic fidelity. The model better captures lacquer-specific characteristics such as bold contrast and layered texture. In contrast, removing either loss yields muted, less articulated results, confirming that task-specific loss design is essential for material-aware stylization.

Fig. 6. Ablated results for our proposed new loss functions

V. Conclusion

We propose LST model, a transformer-based style transfer model for Vietnamese lacquer painting, incorporates a self-attention transfer module and a VGG-inspired decoder. To address the distinct visual features of lacquer art, we introduce contrast loss and depth loss, which together improve content preservation, tonal contrast, and spatial layering. Ablation and comparative results show that our method outperforms prior approaches in structure retention and stylistic fidelity, despite limited style data. However, occasional blurring in high-frequency regions suggests room for improvement. Future work will explore hybrid architectures to enhance edge sharpness and reduce over-smoothing, as well as optimize training and inference speed for larger datasets.

ACKNOWLEDGMENT

This research is funded by the International University, VNU-HCM, Vietnam under grant number T2025-04-IT.

References

[1] C. Zhang, X. Xu, L. Wang, Z. Dai, and J. Yang, "S2wat: Image style transfer via hierarchical vision transformer using strips window attention," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 7, 2024, pp. 7024–7032.

[2] Y. Deng, F. Tang, W. Dong, C. Ma, X. Pan, L. Wang, and C. Xu, "Stytr2: Image style transfer with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 326–11 336.

[3] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 124–12 134.

[4] J. Wang, H. Yang, J. Fu, T. Yamasaki, and B. Guo, "Fine-grained image style transfer with visual transformers," in *Proceedings of the Asian conference on computer vision*, 2022, pp. 841–857.

[5] X. Li, S. Liu, J. Kautz, and M.-H. Yang, "Learning linear transformations for fast image and video style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3809–3817.

[6] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[7] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[8] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images." in *ICML*, vol. 1, no. 2, 2016, p. 4.

[9] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[10] H. Chen, Z. Wang, H. Zhang, Z. Zuo, A. Li, W. Xing, D. Lu *et al.*, "Artistic style transfer with internal-external learning and contrastive learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26 561–26 573, 2021.

[11] H.-P. Wei, Y.-Y. Deng, F. Tang, X.-J. Pan, and W.-M. Dong, "A comparative study of cnn-and transformer-based visual style transfer," *Journal of Computer Science and Technology*, vol. 37, no. 3, pp. 601–614, 2022.

[12] Y. Deng, X. He, F. Tang, and W. Dong, "Z*: Zero-shot style transfer via attention reweighting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6934–6944.

[13] S. Yang, H. Hwang, and J. C. Ye, "Zero-shot contrastive loss for text-guided diffusion image style transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 873–22 882.

[14] Y. Zhang, F. Tang, W. Dong, H. Huang, C. Ma, T.-Y. Lee, and C. Xu, "Domain enhanced arbitrary image style transfer via contrastive learning," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–8.

[15] C.-Y. Shih, Y.-H. Chen, and T.-Y. Lee, "Map art style transfer with multi-stage framework," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 4279–4293, 2021.

[16] T.-N.-H. Le, Y.-H. Chen, and T.-Y. Lee, "Structure-aware video style transfer with map art," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 3s, pp. 1–25, 2023.

[17] Y. Zhang, F. Tang, W. Dong, T.-N.-H. Le, C. Xu, and T.-Y. Lee, "Portrait map art generation by asymmetric image-to-image translation," *Leonardo*, vol. 56, no. 1, pp. 28–36, 2023.

[18] J. An, S. Huang, Y. Song, D. Dou, W. Liu, and J. Luo, "Artflow: Unbiased image style transfer via reversible neural flows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 862–871.

[19] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5880–5888.

[20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

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

[22] M. Hui, Z. Zhang, X. Zhang, W. Xie, Y. Wang, and Y. Lu, "Unifying layout generation with a decoupled diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1942–1951.

[23] G. Kwon and J. C. Ye, "Clipstyler: Image style transfer with a single text condition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 062–18 071.

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

[25] Y. Deng, F. Tang, W. Dong, H. Huang, C. Ma, and C. Xu, "Arbitrary video style transfer via multi-channel correlation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 1210–1217.

[26] R. Birkl, D. Wofk, and M. Müller, "Midas v3. 1–a model zoo for robust monocular relative depth estimation," *arXiv preprint arXiv:2307.14460*, 2023.

[27] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao, and Z. Liu, "Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations," in *European conference on computer vision*. Springer, 2020, pp. 70–85.