# ArTex: Artist-Sensitive Style Transfer via Textual Guidance

DiWei Wu
School of Computer Science
Sichuan Normal University
WDW525@stu.sicnu.edu.cn

Jun Liu
School of Computer Science
Sichuan Normal University
1164811474@qq.com

Jian Liu
School of Computer Science
Sichuan Normal University
jl@stu.sicnu.edu.cn

HeWen Liu
School of Computer Science
Sichuan Normal University
1526982483@163.com

Jun Yang
School of Computer Science
Sichuan Normal University
jkxy_yjun@sicnu.edu.cn

## Abstract

**Traditional style transfer techniques require a reference style image to guide the stylization process by transferring stylistic features onto new images. However, a single image is inherently limited in its expressive capacity and struggles to comprehensively encapsulate an artist's stylistic characteristics. In contrast, natural language provides richer expressiveness and abstraction, enabling the description of high-level artistic semantics such as brushstroke textures and color tendencies. To address this, we propose ArTex—a lightweight, artist-sensitive style transfer model guided by textual input, capable of generating images that are perceptually aligned with specific artistic styles. We design a Spatial Fusion Cross-Attention (CA-SF) module to effectively capture fine-grained interactions between content and style features. Additionally, we develop a novel similarity-contrastive loss to enhance the semantic alignment between textual descriptions and visual features. Furthermore, we introduce a dynamic loading strategy that significantly accelerates the training process, improves training balance across styles, and enhances overall computational efficiency. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches in both style transfer quality and text-guided controllability.**

***Keywords:*** ***Fine arts, Image representations, Image processing, Texturing.***

## 1. Introduction

Artistic style transfer aims to migrate the artistic style from a reference image to a target image, thereby generating an image with specific artistic style. Traditional style transfer methods typically rely on a single style image to extract low-level visual features such as color, texture, and composition. However, a single image has significant limitations in expressing the complex and diverse stylistic characteristics of an artist. The information it contains is often insufficient to comprehensively represent the artist's overall style and may even lead to distortions or deviations in the results due to suboptimal image selection.

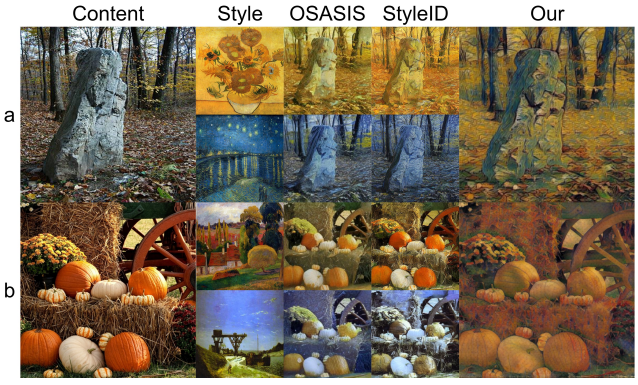


Figure 1. Comparison with image-based methods StyleID and OSASIS. Our method uses artist names as textual style input ("a": Vincent van Gogh, "b": Paul Gauguin), enabling more accurate style expression. In contrast, StyleID and OSASIS rely heavily on two reference paintings, which limits their ability to capture the artist's true style.

As shown in Figure 1, we use traditional stylization methods OSASIS [4] and StyleID [5] to stylize a target image using two different paintings by the same artist as references (Figure 1a: Vincent van Gogh, Figure 1b: Paul Gauguin). It can be observed that, due to the variation between the reference images, the generated images show significant differences in color and brushstroke characteristics. brushstroke characteristics.

Natural language offers greater capacity for abstract expression and flexibility. Users can describe an artist's

stylistic semantics through text, including high-level features such as brushstroke textures, color tendencies, and even emotional atmospheres—information that static images often struggle to fully capture. The CLIP [30] (Contrastive Language-Image Pretraining) model proposed by OpenAI is a cross-modal contrastive learning model trained on 400 million text-image pairs. It embeds both images and text into a shared semantic space, thereby enabling semantic alignment and mutual understanding between vision and language.The introduction of CLIP has inspired cross-modal generation methods based on textual guidance, offering a promising new direction for style transfer.

Several studies have attempted to leverage text for style transfer, such as StyleCLIP [28], CLIPstyler [21], and LDAST [9]. However, StyleCLIP heavily relies on a pre-trained generative model, and the quality of style transfer is constrained by the model's inherent generative capacity. CLIPstyler requires per-instance optimization for each content image and each style, resulting in high computational cost and the introduction of numerous uncontrollable artifacts in the generated results. LDAST depends on large-scale manually annotated datasets. TxST [25] has made some progress in text-driven style transfer, but its results primarily focus on simple color adjustments, lacking accurate reproduction of an artist's unique textures and stylistic patterns, while also introducing undesirable artifacts.

To address the aforementioned limitations, we propose ArTex, a text-guided artist style transfer model designed to efficiently capture the perceptual characteristics of an artist's style. During training, ArTex takes the artist's name as textual input and the corresponding artworks as style references to learn the deep correspondence between visual features and semantic style representations. At inference time, only textual input is required to perform flexible style transfer, thereby eliminating dependence on specific style images and enabling truly text-driven style generation. By leveraging high-level semantic information embedded in text, ArTex more accurately captures the unique stylistic traits of an artist and more faithfully reproduces the artist's brushstrokes, color palette, and expressive characteristics in the generated images (see the "Ours" results in Figure 1).

Our model adopts an encoder-decoder architecture and incorporates a CLIP-based dual-modal feature extraction mechanism to establish semantic alignment between visual and textual inputs. To effectively fuse the content features extracted from the convolutional backbone and the style features obtained from the text encoder, we introduce a Spatial Fusion Cross-Attention (CA-SF) Module, which captures fine-grained interactions between content and style features while preserving spatial consistency. During training, we employ a combination of Directional CLIP Loss [10] and Patch CLIP Loss [21] to jointly optimize the semantic and textural fidelity of the generated images at both global and local levels. To further enhance the consistency between the model output and textual semantics, we design a novel Similarity-Contrastive Loss, which minimizes the stylization distance among images of the same artist under the dual guidance of image and text, while maximizing the inter-artist style separation in the learned feature space. In addition, to improve training efficiency under multi-style conditions, we propose a dynamic loading strategy that significantly accelerates the training process, balances learning across different styles, and enhances overall model efficiency.

Extensive quantitative and qualitative experiments demonstrate that, compared with existing methods, our model achieves more efficient and effective style transfer with higher visual quality. In addition, ablation studies validate the effectiveness of each component in the proposed framework.

## 2. RELATED WORK

### 2.1. Arbitrary Style Transfer

Arbitrary Style Transfer (AST) aims to apply the artistic style of any given image to a content image. Early optimization-based methods, such as those proposed by Gatys *et al*. [11], achieved high-quality stylization but were computationally intensive. Subsequent approaches using feed-forward networks, introduced by Johnson [19] and Ulyanov [35], enabled real-time stylization. Normalization techniques like AdaIN [17] and DIN [18] further improved flexibility by aligning content and style features using learnable parameters. Attention-based models, such as SANet [27], enhanced style-content alignment, while lightweight variants such as A2K [50] and HIS [46] reduced artifacts and computational overhead. AttDis [49] proposes an attention distillation–based framework that unifies multiple tasks, including style transfer and attribute transfer, within a single model.

More recently, Transformer-based models—such as StyTr2 [6], TokenFlow [29], S2WAT [44], and Puff-Net [48]—have been introduced to capture long-range dependencies and better preserve semantic content. SaMam [24] innovatively introduces the Mamba architecture into style transfer to address the high computational cost associated with Transformer-based models. In parallel, diffusion-based methods have gained popularity due to their powerful generative capabilities. Techniques like FreeStyle [12], DiffStyle [23], and DiffuseST [15] integrate content and style during the diffusion process to enable more controllable stylization. Models such as OSASIS [4] and InST [47] tackle one-shot and inversion-based style learning, while extensions like StyleID [5], SigStyle [37], RB-Modulation [31] and StyleSSP [42] aim to diversify diffusion-driven style transfer. Despite their ability to pro-

duce high-quality and diverse results, diffusion models often suffer from slow inference speeds and significant computational costs.

### 2.2. Domain-Aware Style Transfer

While Arbitrary Style Transfer (AST) models offer notable flexibility, they often struggle to maintain style consistency and semantic precision when handling complex artistic styles. Domain-aware approaches address this limitation by learning style features from specific artists or art movements, enabling more robust and distinctive stylization. For example, DualAST [2] leverages a dual-branch architecture to jointly model artist-level and artwork-specific characteristics, thereby improving both control and diversity. DSTN [14] integrates semanticand frequency-aware modules to balance global texture coherence with local detail preservation, while DSTM [40] employs dual style transfer modules alongside edge enhancement techniques to decouple content structure from style.

These methods demonstrate strong performance in artist imitation, portrait stylization, and art restoration. However, they depend heavily on high-quality, annotated datasets, which constrains their generalization to unseen or cross-domain styles and limits their adaptability in open-world scenarios.

### 2.3. Text-Guided Style Transfer

Text-guided style transfer utilizes natural language as a control signal for stylization, offering greater flexibility and enhanced user interactivity compared to image-based approaches. CLIP enables the alignment of image and text features within a shared embedding space, empowering models such as CLIPstyler [21] to optimize images directly based on textual prompts. Subsequent extensions, including TextStyler [41] and TRTST [1], introduce dedicated stylization networks and multimodal Transformer architectures to further improve content preservation and stylistic fidelity. Diffusion-based methods, such as FreeStyle [12] and StyleDiffusion [38], achieve high-quality and open-domain stylization using text as the sole guidance modality. Spatially aware techniques—such as SEM-CS [20], LEAST [33], and SpectralCLIP [43]—enhance stylization by enabling localized control and minimizing visual artifacts through image segmentation or embedding refinement. StyleStudio [22], a recent method, generates image content during the diffusion process under the guidance of textual prompts, while constraining the generated images to follow the style of a reference style image. Moreover, frameworks like ArtCrafter [16] and TxST [25] learn style representations in the form of language embeddings derived from example images, facilitating personalized and flexible style transfer.

Overall, text-guided methods present a powerful and intuitive alternative, offering strong cross-domain generalization, precise control, and broad applicability—marking a significant advancement in the field of style transfer.

## 3. METHODOLOGY

To achieve text-guided artistic style transfer, we propose a novel framework, as illustrated in Figure 2. The detailed architecture of our model is presented in Section 3.1. The dynamic loading algorithm, which significantly accelerates our training process, is described in Section 3.2. Furthermore, to better distinguish style features corresponding to different textual descriptions, we introduce a novel contrastive similarity loss function, which is elaborated in Section 3.3. For further details, refer to our supplementary material.

### 3.1. Framework

**Content Image Encoding Module:** We employ a pretrained VGG-19 [32] network to extract deep features from the content image $I_c$, obtaining feature maps at each layer denoted as $f_c^i = E(I_c),\ i \in \{1, \ldots, L\}$. Since VGG-19 is originally trained on the ImageNet dataset for classification tasks, which differ from the objectives of style transfer, we incorporate the Position Attention Module proposed by Fu *et al*. [8] to enhance the original content features $f_c$, resulting in an improved representation $f_{cc}$.

**Style Encoding Module:** Style features are extracted using the pretrained CLIP ViT-B/32 model. We assume that the training set contains $N$ distinct styles. In each training iteration, one style text $T_{s1} \in T_s$ and its corresponding style image $I_{s1} \in I_s$ are randomly selected, along with another randomly chosen pair $T_{s2}$ and $I_{s2}$ from the remaining styles. The style text $T_s$ is processed by the CLIP Text Encoder to obtain the text embedding $E_t(T_s)$, while the style image $I_s$ is processed by the CLIP Image Encoder to obtain the image embedding $E_i(I_s)$. Since CLIP projects both embeddings into a one-dimensional latent space $E_t, E_i \in \mathbb{R}^{512}$, the spatial structure of the style features is lost.

To preserve the spatial awareness of style features, we adopt the relative positional encoding mechanism inspired by the study on Multi-Head Self-Attention (MHSA) layers by Aravind Srinivas *et al*. [34] Specifically, we design horizontal encodings $R_w \in \mathbb{R}^{1\times32\times512}$ and vertical encodings $R_h \in \mathbb{R}^{32\times1\times512}$, and construct 2D feature maps $f_{T_s}, f_{I_s} \in \mathbb{R}^{32\times32\times512}$ through broadcast addition. Then, a MHSA module is applied to establish horizontal and vertical dependencies in the spatial domain of the style features, enhancing semantic consistency and the expression of local spatial structures.

**Feature Fusion Module:** To effectively integrate the content features extracted from the convolutional backbone (VGG-19) and the style representations obtained from

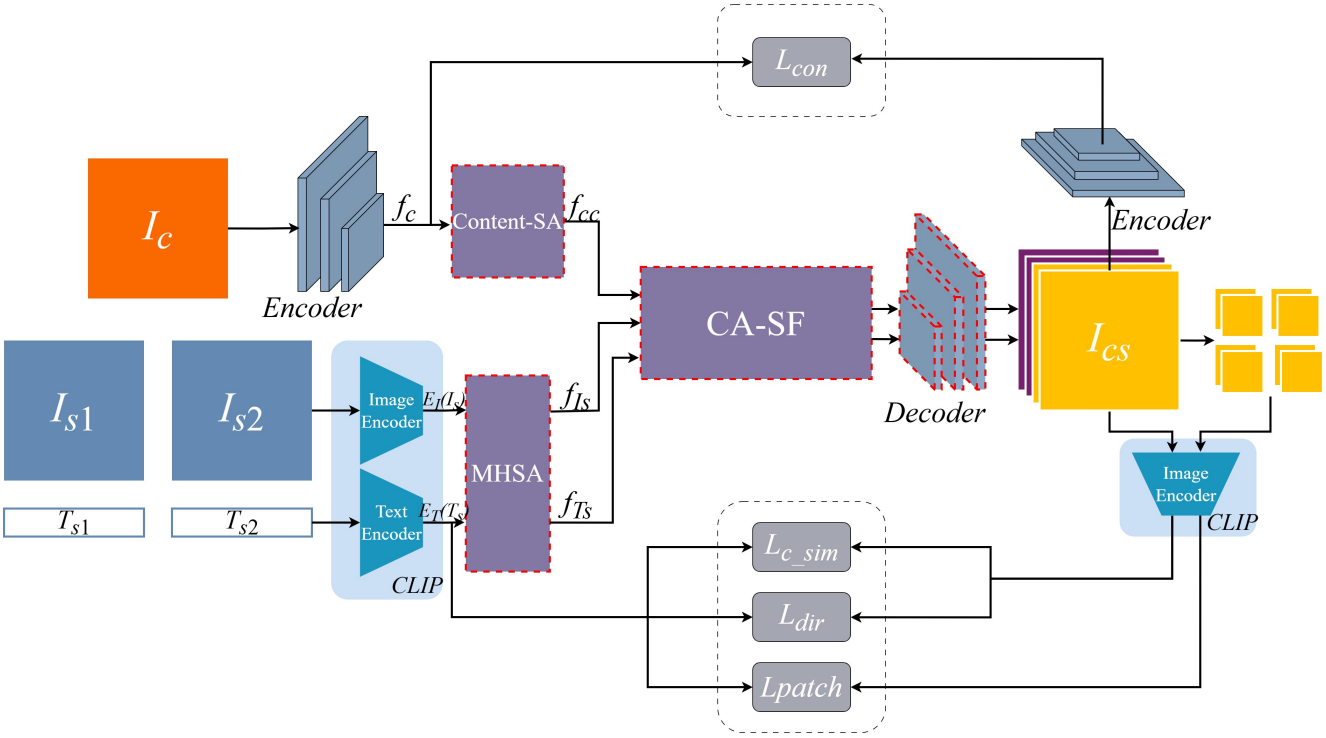


Figure 2. Overview of our proposed framework. Given a content image and the names and paintings of two artists, the model learns to generate stylized images by minimizing the distance between text and image features of the same artist, while maximizing the distance between different artists' styles.

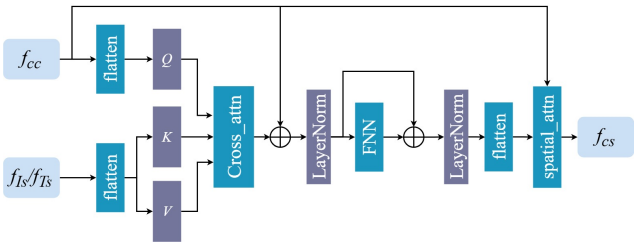


Figure 3. Our spatial fusion cross-attention module integrates content features from VGG-19 and style features from the CLIP encoder. It enhances content via multi-head cross-attention and learns a spatial attention mask to adaptively balance style and content contributions at each location, enabling context-aware and spatially consistent style transfer.

the CLIP Text Encoder, we propose a **Spatial Fusion Cross-Attention Module (CA-SF)**, as illustrated in Figure 3. Given the enhanced content feature map $f_{cc} \in \mathbb{R}^{B\times C\times H\times W}$ and the style feature maps $f_{T_s}/f_{I_s} \in \mathbb{R}^{B\times C\times H\times W}$, we first flatten them into sequences and reshape them to $\mathbb{R}^{B\times N\times C}$, where $N = H \times W$. Then, we apply a multi-head attention mechanism, using the content features $f_{cc}$ as the *Query*, and the style features $f_{T_s}/f_{I_s}$ as both *Key* and *Value*. The attention-refined content features are added back to the original *Query* through a residual connection, followed by a Feed-Forward Network (FFN) and Layer Normalization (LayerNorm), resulting in the cross-attention enhanced feature $f_{\text{attn}}$. The overall design of this module follows the Transformer architecture paradigm.

To further enhance spatial-level feature fusion, we introduce a Spatial Attention Subnetwork. This subnetwork takes the concatenated content features and attention-enhanced features as input and learns a pixel-wise weighting map to adaptively balance the contribution of content and style at each spatial location. The final fused feature map $\mathbf{f}_{cs}$ is computed as:

$$\mathbf{f}_{cs} = \mathbf{A} \cdot \mathbf{f}_{cc} + (1 - \mathbf{A}) \cdot \mathbf{f}_{attn} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{B\times 1\times H\times W}$ is the learned spatial attention mask that regulates the fusion ratio between content and style at each location.

This fusion mechanism not only achieves fine-grained semantic alignment between content and style, but also introduces spatial adaptivity, thereby maintaining spatial consistency and enhancing the perceptual realism and expressive quality of the generated images.

**Decoder Module:** Our decoder mirrors the structure of the encoder and reconstructs the stylized feature $\mathbf{f}_{cs}$ into the final stylized image $\mathbf{I}_{cs}$.

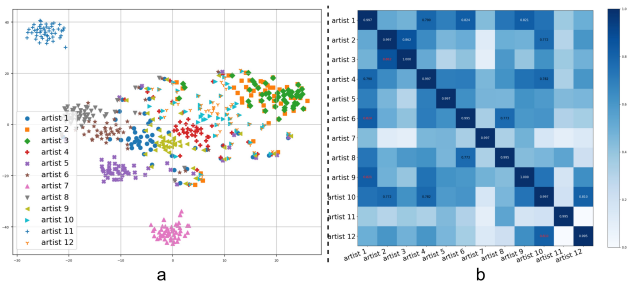


Figure 4. a. t-SNE visualization of stylized image features for each artist. b. Similarity matrix between stylized image features of different artists.

### 3.2. Dynamic Loading Strategy

In the CLIP embedding space, style representations from distinct artists differ significantly, aiding rapid learning of unique styles. However, similar-style artists (For example, "artist 2" and "artist 3" represent Claude Monet and Berthe Morisot, both of whom belong to the French Impressionist and have similar styles and brushstrokes) produce closely clustered embeddings due to shared characteristics, limiting fine-grained discrimination and slowing training convergence. To address this, we propose a **dynamic loading strategy** that adjusts sampling probabilities of style categories during training to improve style differentiation.

Specifically, after a certain number of training iterations, we randomly select $m$ content images from the training set and generate stylized outputs for each of the $N$ learned style texts. These generated images are passed through a pre-trained CLIP image encoder to extract their feature representations in the CLIP embedding space. For each style category $s \in \mathcal{S}$, we collect a set of feature vectors $F_s = \{f_1, \ldots, f_m\}$, where each $f_i \in \mathbb{R}^{512}$. Based on these features, we compute the mean and standard deviation vectors of each style category:

$$\mu_s = \frac{1}{m}\sum_{i=1}^{m} f_i, \quad \sigma_s = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(f_i - \mu_s)^2}. \tag{2}$$

**Algorithm 1** Adaptive Style Weight Adjustment

**Require:** Model $M$, Content image set $X \leftarrow$ LoadDataset(), Style list $\mathcal{S} = \{s_1, \ldots, s_N\}$, parameters $\gamma, \delta$
**Ensure:** Adjusted weight list $W = [w_1, \ldots, w_N]$

```
if iterations % 2000 == 0 and iterations ≥ 6000 then
    Save(M)
    M_load ← Reload(M)
    W ← [1] * N
    x_batch ← RandomSelect(X, 50)
    for all s ∈ S do
        F[s] ← CLIP_Encode(M_load.generate(x_batch, s))
    end for
    sim_styles ← top_{N/3}(lower_tri(Similarity_Matrix(F)))
    tsne_styles ← top_{N/3}(Tsne_Divergence(F))
    for all s ∈ sim_styles do
        W[index[s]] += γ
    end for
    for all s ∈ tsne_styles do
        W[index[s]] += δ
    end for
    return W
end if
```

We then compute the pairwise cosine similarity between different style categories and introduce a stability-aware penalty term based on the Euclidean distance between their standard deviation vectors. This helps to avoid misjudging stylistic similarity due to large intra-style feature variations. The similarity between two styles $s_i$ and $s_j$ is measured using the following joint metric:

$$\mathrm{Sim}(s_i, s_j) = \frac{\mu_{s_i} \cdot \mu_{s_j}}{\|\mu_{s_i}\| \times \|\mu_{s_j}\|} - \lambda \cdot \|\sigma_{s_i} - \sigma_{s_j}\|_2, \tag{3}$$

where $\lambda$ is a penalty coefficient that controls the influence of deviation in feature dispersion. The visualized similarity matrix is shown in Figure 4b. From the lower triangular part of the similarity matrix, we identify the top $N/3$ style pairs with the highest similarity scores. The sampling weight of these style categories is increased by $\gamma$ over the subsequent 2000 training iterations, thereby focusing learning on hard-to-distinguish styles.

To further quantify the distributional divergence of different style categories in the embedding space, we calculate a divergence score for each style category based on its variance in a 2D t-SNE projection. Specifically, for each style category $s \in \mathcal{S}$, let its set of projected points in the 2D space be: $\mathrm{point}_s = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$. Then, the

divergence score of style $s$ is defined as:

$$\begin{aligned} \text{divergence\_score}(s) &= \text{var}_x + \text{var}_y, \\ \text{where} \quad \text{var}_x &= \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2, \\ \text{var}_y &= \frac{1}{m}\sum_{i=1}^{m}(y_i - \bar{y})^2. \end{aligned} \tag{4}$$

where $\bar{x}$ and $\bar{y}$ denote the mean coordinates of the samples in the $X$ and $Y$ directions, respectively. The divergence score measures the dispersion of a style within the t-SNE space, where lower scores indicate more consistent learning, and higher scores reflect instability or underfitting. As illustrated in Figure 4a, styles with greater dispersion tend to correlate with inferior stylization performance. Based on this analysis, we increase the sampling weights of the top $N/3$ most divergent styles by $\delta$ for 2000 iterations, encouraging the model to focus on these harder-to-learn patterns.

During training, sampling weights are dynamically adjusted based on divergence scores: styles with high divergence—indicative of underrepresentation or unstable learning—are sampled more frequently to enhance learning effectiveness, whereas well-learned, stable styles are sampled less to reduce redundancy. Every 2000 iterations, CLIP-based embeddings and divergence scores are recalculated to update the sampling distribution, ensuring balanced style representation and facilitating efficient and robust model convergence.The whole training procedure is summarized in Algorithm 1.

### 3.3. Loss Function

Our training framework integrates several loss functions to jointly optimize global semantic alignment and fine-grained stylization quality:

**Directional CLIP Loss:** Following the formulations in [21, 10], we adopt the Directional CLIP loss $L_{\text{dir}}$, which aligns the direction vector between the stylized image $I_{cs}$ and the target style text $T_s$ in the shared CLIP embedding space via cosine similarity. This loss guides the semantic transition of the image towards the target style, facilitating cross-modal semantic alignment between text and image.

**PatchCLIP Loss:** Although $L_{\text{dir}}$ performs well for global semantic alignment, it lacks the granularity needed for accurate local texture and semantic control. To address this limitation, Kwon *et al*. [21] proposed the PatchCLIP loss, which enforces consistency between the directional changes of local image patches and the target style text in CLIP space. Additionally, to prevent excessive stylization in easily optimized regions, a threshold rejection mechanism is employed to suppress their gradient responses. We adopt the PatchCLIP loss design in our framework.

**Contrastive Similarity Loss:** To encourage aggregation of samples within the same style and separation between different styles, we design a cosine similarity-based contrastive loss inspired by InfoNCE [26] and SimCLR [3]. As illustrated in Figure 5, in each training iteration, $2N$ stylized samples are generated and their features are extracted via the CLIP Image Encoder. The first $N$ samples belong to the same style (positive set $V^+ = \{v_1, \ldots, v_N\}$), while the remaining $N$ samples belong to other styles (negative set $V^- = \{v_{N+1}, \ldots, v_{2N}\}$). For each positive sample $v_i$ ($i = 1, \ldots, N$), the similarity with other samples of the same style $v_j$ ($j = 1, \ldots, N, j \neq i$) should be maximized. The contrastive positive loss is defined as:

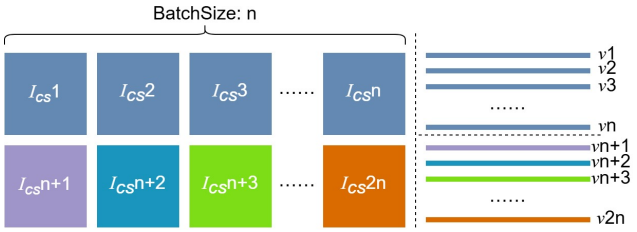


Figure 5. Assuming a batch size of $N$, the model generates $2N$ stylized images per iteration, where the first $N$ images are from the same artist, and the remaining $N$ images are from other artists. Feature vectors $v$ are extracted using the CLIP image encoder. The positive sample set is defined as $V^+ = \{v_1, \ldots, v_N\}$, and the negative sample set as $V^- = \{v_{N+1}, \ldots, v_{2N}\}$.

$$\begin{aligned} L_{\text{pos}}(V) &= -\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}\log\left(R_{ij}\right), \\ \text{where} \quad R_{ij} &= \frac{e^{S(v_i,v_j)/\tau}}{\sum_{k=1}^{2N} e^{S(v_i,v_k)/\tau} + \epsilon}. \end{aligned} \tag{5}$$

where $S(\cdot,\cdot)$ denotes cosine similarity, $\tau$ is a temperature coefficient, and $\epsilon$ is a small constant added to avoid numerical instability.

Conversely, to suppress similarity between samples from different style categories, we define a negative contrastive loss. For each positive sample $v_i$ ($i = 1, \ldots, N$), its similarity to each negative sample $v_j$ ($j = N+1, \ldots, 2N$) should be minimized. The negative loss is defined as:

$$L_{\text{neg}}(V) = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=N+1}^{2N}\log\left(1 - R_{ij} + \epsilon\right). \tag{6}$$

The final contrastive similarity loss is formulated as:

$$\begin{aligned} L_{\text{c\_sim}}(V_I, V_T) = \alpha\big(L_{\text{pos}}(V_I) + L_{\text{neg}}(V_I)\big) \\ + \beta\big(L_{\text{pos}}(V_T) + L_{\text{neg}}(V_T)\big), \end{aligned} \tag{7}$$

where $V_I$ and $V_T$ denote the stylized features generated from image-guided and text-guided processes respectively, and $\alpha$, $\beta$ are weighting factors. By minimizing $L_{\text{con\_sim}}$,


Figure 6. Comparison of stylization results with SOTA methods. The first column shows the content and style images. For text-guided style transfer, artist names are used as style references. The remaining columns present the results from ASTMA, SaMam, StyTr2, S2WAT, StyleID, OSASIS, AttDis, CLIPstyler, CLIPstyler(fast), TxST, and our proposed method.

the model effectively pulls together samples of the same style while pushing apart samples of different styles, enhancing the discriminability of style features in the embedding space.

**Content and Style Loss:** To ensure both semantic preservation and faithful style rendering, we incorporate traditional perceptual loss functions based on a pre-trained VGG-19 network. The content loss $L_c$ and style loss $L_s$ are constructed to ensure semantic consistency and accurate style expression after the style transfer. In addition, we introduce a total variation regularization term $L_{\text{reg}}$ to suppress noise and enhance spatial smoothness and continuity at the pixel level.

**Overall Loss Function:** The overall loss function is defined as:

$$L_{\text{total}} = \lambda_{\text{dir}} L_{\text{dir}} + \lambda_{\text{patch}} L_{\text{patch}} + \lambda_{\text{c\_sim}} L_{\text{c\_sim}} + \lambda_c L_c + \lambda_s L_s + \lambda_{\text{reg}} L_{\text{reg}}. \quad (8)$$

which enables the model to effectively integrate the complementary strengths of both visual and textual style representations. Specifically, it preserves the fine-grained texture and color structures derived from the reference style image while simultaneously incorporating the semantic abstraction and generality afforded by textual descriptions. As a result, the model is capable of producing stylized outputs that are not only diverse and semantically aligned, but also visually coherent and richly expressive.

## 4. EXPERIMENTS

### 4.1. Implementation Details

We use the ImageNet dataset as the content dataset and the WikiArt dataset as the style dataset. All images are first resized by maintaining the aspect ratio such that the shorter side is scaled to 512 pixels, followed by a random crop to a resolution of $256 \times 256$. During inference, our model supports arbitrary-resolution stylization. To ensure fair comparison with existing methods, we select 12 well-known artists from the WikiArt dataset, resulting in a total of 7,364 paintings, and use the artists' names as style text prompts. To mitigate noise in the text embeddings, we adopt the prompt engineering techniques proposed by [30, 21], generating multiple semantically similar text descriptions. The corresponding CLIP embeddings are then averaged to obtain a more robust style representation.

Our training procedure consists of two stages. In the first stage, we construct an image-to-image reconstruction task using the content image encoder (pre-trained VGG-19), the style fusion module (CA-SF), and the decoder. During this stage, only the decoder is trained to establish a basic image reconstruction capability. In the second stage, based on the

decoder trained in the first stage, we jointly train the entire network, including the content encoder, fusion module, and style guidance module, to perform image-text guided style transfer.

We set $\lambda_{\text{dir}}$, $\lambda_{\text{patch}}$, $\lambda_{\text{con\_sim}}$, $\lambda_c$, $\lambda_s$ and $\lambda_{\text{reg}}$ as 100, $1 \times 10^{-3}$, 50, 7.5, 5, and $5 \times 10^{-2}$, respectively. We adopt the Adam optimizer with an initial learning rate of $1 \times 10^{-4}$, a batch size of 10, and a total of 40,000 iterations. Dynamic sampling weights are introduced starting from the 6,000th iteration and are updated every 2,000 iterations thereafter. All experiments are conducted on a single NVIDIA V100 GPU with 32GB of memory.

For more experimental details and additional stylization results generated by our model, please refer to our supplementary material.

### 4.2. Comparison with Prior Work

To intuitively evaluate the stylization performance, we compare our method against a series of state-of-the-art style transfer approaches across various content-style combinations, as illustrated in Figure 6. Each row in the figure corresponds to a specific artist style, while each column represents a different stylization method. The first seven methods—ASTMA [7], SaMam [24], StyTr2 [6], S2WAT [44], StyleID [5], OSASIS [4] and AttDis [49]—are representative image-guided style transfer models. Among them, ASTMA and SaMam are attention-based models built on convolutional neural networks (CNNs), while StyTr2 and S2WAT are based on Transformers. StyleID, OSASIS and AttDis are diffusion-based models. The last four methods—CLIPstyler [21], CLIPstyler(fast) [21] (Pre-trained version of CLIPstyler), TxST [25], and Ours—are text-driven style transfer models, where only the artist's name (*e.g.* "Claude Monet") is used as the style reference.

While image-guided methods can produce visually compelling stylization, they often encounter challenges such as texture overfitting, semantic misalignment, and structural distortions. For instance, models like ASTMA and SaMam tend to overemphasize local textures, leading to unnatural deformations—such as background warping (row 1), distorted architectural contours (row 3), and misrepresented water textures (row 5). Although StyTr2 and S2WAT yield more coherent results, they frequently lack high-frequency details and sharp edges, resulting in blurred patterns (row 4) and diminished pineapple textures (row 5). StyleID, which adopts diffusion models, effectively preserves structural integrity but captures stylistic attributes in a more conservative manner. This often yields consistent yet less distinctive stylizations. In contrast, OSASIS exhibits a stronger stylistic expressiveness but at the cost of severe structural degradation—*e.g.* facial distortions (row 1) and smeared architectural features (row 3)—due to its heavy reliance on texture priors from the diffusion backbone. However, AttDis places excessive emphasis on style by heavily incorporating local information from the style image, which may lead to inferior preservation of content details. Text-guided approaches demonstrate promising flexibility, yet they also exhibit notable limitations. CLIPstyler lacks spatial priors and attention mechanisms, resulting in significant artifacts during the real-time optimization process—such as in the hair region(row 1). Additionally, the generated image styles are uncontrollable and fail to effectively balance the relationship between style and content. CLIPstyler (fast) heavily relies on the pre-trained VGG encoder-decoder network, leading to rigid brushstroke textures and an inability to emulate the unique painting styles of individual artists. TxST alleviates some of these issues through architectural enhancements; however, it still introduces artifacts such as facial deformations (row 1), irregular texture patterns (row 6), and rainbow-like banding. These artifacts likely stem from conflicts between fixed positional encodings and the integration of textual style tokens.

In contrast, our method incorporates a spatially fused cross-attention module that effectively balances stylistic rendering with content preservation. As illustrated in the first row, it successfully transfers the target style while retaining facial structure and background clarity. In the second row, it captures the dynamic brushstrokes and warm color palette characteristic of Van Gogh. In the sixth row, it replicates Monet's pastel tones and misty atmosphere, while preserving scene depth and lighting cues. These results demonstrate that our model not only achieves faithful artistic stylization under textual guidance but also maintains strong semantic consistency and spatial coherence. Compared to both image-guided and existing text-based approaches, our method delivers superior performance in terms of stylization fidelity, visual distinctiveness, and structural integrity.

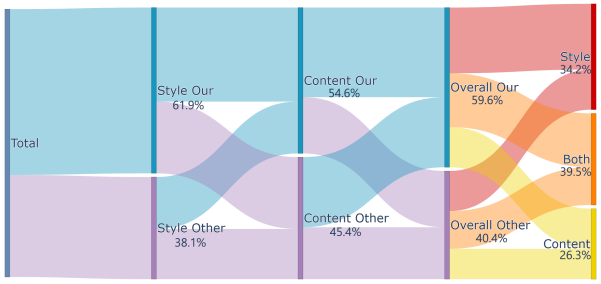


Figure 7. Sankey diagram of user study results. Compared to the contrast method, more participants preferred our method in terms of style expression, content preservation, and overall visual quality.

### 4.3. User Study:

We selected 15 content images and performed style transfer using the styles of 12 different artists, resulting in a total of 180 stylized images. Subsequently, we designed a user study via the Google Forms platform to evaluate the performance of different style transfer methods. The questionnaire consisted of 15 sections, each corresponding to a set of one content image and associated style references, and included four evaluation questions. In each section, participants were first shown a content image along with three style reference images of a specific artist. Then, two stylization results were presented: one generated by our method and the other randomly selected from existing state-of-the-art (SOTA) methods. To mitigate order bias, the display order of the two results was randomized. Following the protocol of [7], participants were asked to respond to the following four questions: (1) Which result better conveys the target artistic style? (2) Which result better preserves the content structure? (3) Which result delivers a better overall visual impression? (4) In answering Question 3, did you prioritize content, style, or both equally?

In total, 53 participants took part in the survey, yielding 795 valid response sets. Figure 7 presents a Sankey diagram summarizing the user evaluation results. As shown, participants consistently favored our method across all three dimensions: style expression, content preservation, and overall visual quality. Notably, in the "Overall" category, the majority of users considered our method to offer a superior balance between style and content compared to competing approaches. Furthermore, more participants indicated a preference for a balanced consideration of Stylistic performance, reinforcing the effectiveness and comprehensive performance of our approach in the style transfer task.

### 4.4. Quantitative Evaluation:

**Evaluating using classification models:** To comprehensively evaluate the performance of our model, we conduct quantitative comparison with state-of-the-art methods from two key aspects: content preservation and style expression. We adopt the advanced visual classification model MobileViT-v3 [36] to train two separate classifiers: one for content recognition and one for artist classification. Content Classifier: We select six categories from the ImageNet-1k-1 subset, comprising a total of 3,823 images (approximately 600–700 per class), and split them into training and testing sets at a ratio of 8:2. The trained model is used to evaluate how well the stylized images retain the original content. Artist Classifier: We select nine artists from the WikiArt dataset with a total of 5502 paintings, also split into training and testing sets in an 8:2 ratio, to train the classifier for assessing style fidelity. During the evaluation phase, we use the 761 content test images to generate nine stylized versions per image, resulting in a total of 31347 stylized images. These stylized outputs are fed into the content classifier to assess content preservation accuracy, and into the artist classifier to assess style expression accuracy. Our trained Content Classifier achieved a Top-1 Accuracy of 86.47% and a Mean Precision of 86.49% on the content image test set. The Artist Classifier, evaluated on the WikiArt test set, reached a Top-1 Accuracy of 83.73% and a Mean Precision of 78.44%.

Table 2 presents a comparison of Top-1 Accuracy and Mean Precision across different methods, including ASTMA, StyTr2, CLIPstyler(fast), TxST, and our proposed approach. Figure 8 shows a representative example from the test set. Our method achieves the highest scores in both Style Top-1 Accuracy (70.74%) and Style Mean Precision (70.45%), indicating superior capability in capturing and expressing artistic style characteristics.

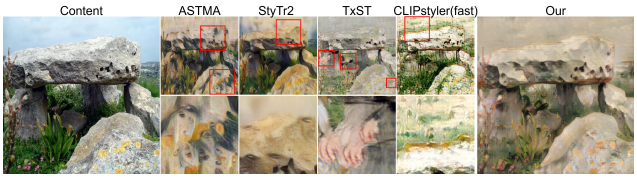


Figure 8. A representative sample from the quantitative evaluation. The left shows the content image, with the style reference based on the artist Berthe Morisot.

In contrast, while ASTMA and StyTr2 achieve comparable style precision (around 50%), they exhibit significantly lower content preservation performance, with Top-1 Accuracy dropping to 24.29% and 35.15%, respectively. Furthermore, both models tend to introduce semantic distortions in stylized results (*e.g.* distorted eyes in the red boxes). CLIPstyler(fast) demonstrates the best performance in content preservation; however, due to its limited modeling of painterly strokes and tonal characteristics, it obtains the lowest style recognition precision (34.78%). Moreover, it frequently introduces scattered artifacts, especially in background regions such as the sky. The TxST method shows a better balance between content and style, but still suffers from severe facial semantic artifacts (as highlighted by red boxes) and the presence of rainbow-like banding, which not only undermines stylistic consistency but also degrades visual naturalness.

Our method effectively integrates the artist's tonal brushstrokes and stylistic textures while preserving the overall spatial structure and semantic integrity of the content image. Notably, although our content preservation score (37.28%) is slightly lower than those of TxST and CLIPstyler(fast), it remains within a reasonable range and avoids the introduction of destructive artifacts, thereby producing more natural and visually harmonious stylizations. Quantitative results demonstrate that our model achieves a superior balance between the two core objectives of style transfer: content fi-

Table 1. Quantitative evaluation using conventional style transfer metrics.We compute LPIPS, FID, and ArtFID scores to assess content fidelity, style expressiveness, and overall perceptual quality of the stylized images.Lower scores correspond to stronger style representation and higher visual quality.

| Metric | Our | ASTMA | SaMam | StyTr2 | S2WAT | StyleID | OSASIS | AttDis | CLIPStyler | CLIPstyle(fast) | TxST |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **LPIPS↓** | 0.537 | 0.611 | 0.629 | 0.575 | 0.558 | 0.569 | 0.541 | 0.587 | 0.664 | 0.501 | 0.475 |
| **FID↓** | 166.9 | 399.8 | 319.7 | 261.7 | 261.1 | 252.2 | 266.0 | 324.0 | 294.1 | 192.8 | 195.9 |
| **ArtFID↓** | 258.06 | 645.69 | 522.42 | 413.75 | 408.35 | 397.27 | 411.45 | 515.78 | 491.05 | 290.89 | 290.43 |

Table 2. Quantitative evaluation based on classification models (MobileViT-v3).We trained two independent classifiers—one for content and one for style—to evaluate the stylized images.We report Top-1 Accuracy (ACC) and Mean Precision as performance metrics.Higher scores indicate that the stylized images better match the target content or artistic style, reflecting higher overall quality.

| Methods | Content(6 class) | | Style(9 class) | |
|---|---|---|---|---|
| | Top-1 ACC↑ | Mean Precision↑ | Top-1 ACC↑ | Mean Precision↑ |
| ASTMA | 24.29 | 32.20 | 46.59 | 59.50 |
| StyTr2 | 35.15 | 39.64 | 48.94 | 59.92 |
| CLIPstyler(fast) | 62.94 | 66.87 | 34.78 | 48.68 |
| TxST | 44.73 | 56.02 | 53.13 | 60.06 |
| Our | 37.28 | 48.51 | 70.74 | 70.45 |

delity and style expressiveness.

**Using traditional evaluation indicators:** In addition to classifier-based evaluation, we also employ conventional metrics commonly used in style transfer tasks. LPIPS [45] is used to measure the content fidelity between the stylized image and the content image, while the FID [13] score evaluates the style expression capability between the stylized and reference style images. Furthermore, we adopt ArtFID [39], a recently proposed metric that closely aligns with human perceptual judgment, which jointly assesses the preservation of both content and style in stylized results.

As shown in Table 1, our model achieves the best performance on both FID and ArtFID, with scores of 166.9 and 258.1, respectively. It also attains a competitive LPIPS score of 0.537, ranking among the top results. These outcomes are consistent with the findings from our classifier-based evaluation, further validating the effectiveness of our approach.

### 4.5. Ablation Study:

**Validation of the Contrastive Similarity Loss Effectiveness:** To evaluate the impact of our contrastive similarity loss (Equation 7), we trained two models—one with the loss and one without—each for 40,000 iterations, and visualized their stylized image features using t-SNE. As illustrated in Figure 9b, the absence of the contrastive loss results in scattered feature distributions, with only a few artists (*e.g.* Thomas Cole-blue plus marker and M.C. Escher-pink triangle marker) forming distinguishable clusters due to their inherently unique styles. In contrast, Figure 9a demonstrates that incorporating the contrastive similarity loss yields compact, well-separated clusters corresponding to individual artists, reflecting improved style-specific feature aggregation and discrimination. Quantitatively, omitting this loss leads to a 17.78% drop in Top-1 accuracy and a 14.47% reduction in mean precision on the artist classification task (Table 3), underscoring its critical role in enhancing style recognition performance.

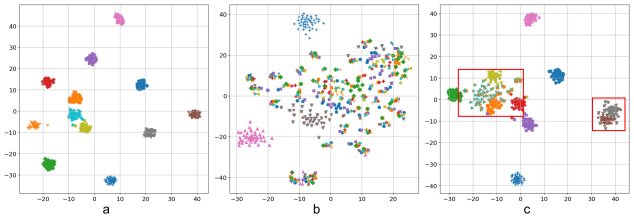


Figure 9. t-SNE visualizations of features after 40,000 training epochs in the ablation study. (a) Full model, (b) w/o Contrastive Similarity Loss, (c) w/o Dynamic Loading Strategy.

**Validation of the Dynamic Loading Strategy Effectiveness:** To evaluate the effectiveness of our proposed dynamic loading strategy, we conducted experiments with and without its application, and visualized the learned features using t-SNE. As shown in Figure 9c, after training for 40,000 iterations without dynamic loading, the model able to separate styles with distinct features, but fails to distinguish stylized images of artists with similar styles (highlighted by the red box). However, when training is extended to 80,000 iterations, the stylized images of these visually similar artists finally achieve clear separation. This demonstrates the effectiveness of our proposed dynamic loading strategy, effectively reducing training time by approximately half (around 16 hours).

**Validation of the Effectiveness of PatchCLIP Loss:** We further evaluated the effectiveness of the incorporated PatchCLIP loss. As shown in Table 3, when training without the PatchCLIP loss, the stylized images generated during inference exhibit a decrease of 7.02% in style Top-1 Accuracy and 5.58% in Mean Precision. These results demonstrate that the PatchCLIP loss plays a vital role in reinforcing the transmission of local semantic style details, contributing to more accurate and nuanced stylization out-

Table 3. Quantitative comparison in ablation study.Top-1 ACC and Mean Precision of the artist classifier when training without $L_{\text{c\_sim}}$ and without $L_{\text{patch}}$.

| Methods | Style(9 class) | |
|---|---|---|
| | Top-1 ACC↑ | Mean Precision↑ |
| full model | 70.74 | 70.45 |
| w/o $L_{\text{c\_sim}}$ | 52.96 | 55.98 |
| w/o $L_{\text{patch}}$ | 63.72 | 64.87 |

comes.

### 4.6. Applications

Figure 10. Stylization results of the image-based arbitrary style transfer method.

**Arbitrary Style Transfer:** To evaluate the generalization of our spatial fusion cross-attention (CA-SF) module, we adapt it for image-based arbitrary style transfer. We replace the original text-based style feature extraction with features from a pre-trained VGG-19 and enhance them using the Channel Attention Module by [8], producing enriched style features $f_{ss}$. The rest of the architecture remains unchanged. Text-specific losses (Directional CLIP and PatchCLIP) are removed, and an identity loss is introduced to better preserve content details. The model is trained for 50,000 iterations.

As shown in Figure 10, our method achieves high-quality stylization across diverse styles, effectively transferring texture, color, and brushwork while maintaining content structure. This demonstrates the scalability and versatility of our approach for both textand image-guided style transfer tasks.

## 5. CONCLUSIONS

In this paper, we propose a text-guided arbitrary artist style transfer framework that integrates a spatial fusion cross-attention mechanism, achieving more precise style representation while preserving the content structure. By designing a contrastive similarity loss function and a dynamic loading strategy, our method significantly outperforms existing imageand text-driven approaches in multiple qualitative and quantitative evaluations. Ablation studies further validate the effectiveness of each module. Moreover, we extend our method to image-based arbitrary style transfer, demonstrating strong generalization and practical applicability. Future work may explore cross-modal style representation and more efficient training strategies.

## References


[1] H. Chen, Z. Wang, L. Zhao, J. Li, and J. Yang. Trtst: Arbitrary high-quality text-guided style transfer with transformers. *IEEE Transactions on Image Processing*, 2025. 3

[2] H. Chen, L. Zhao, Z. Wang, H. Zhang, Z. Zuo, A. Li, W. Xing, and D. Lu. Dualast: Dual style-learning networks for artistic style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 872–881, 2021. 3

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 6

[4] H. Cho, J. Lee, S. Chang, and Y. Jeong. One-shot structure-aware stylized image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8302–8311, 2024. 1, 2, 8

[5] J. Chung, S. Hyun, and J.-P. Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8795–8805, 2024. 1, 2, 8

[6] 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*, pages 11326–11336, 2022. 2, 8

[7] Y. Deng, F. Tang, W. Dong, W. Sun, F. Huang, and C. Xu. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2719–2727, 2020. 8, 9

[8] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 3, 11

[9] T.-J. Fu, X. E. Wang, and W. Y. Wang. Language-driven artistic style transfer. In *European Conference on Computer Vision*, pages 717–734. Springer, 2022. 2

[10] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 2, 6

[11] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. 2

[12] F. He, G. Li, M. Zhang, L. Yan, L. Si, F. Li, and L. Shen. Freestyle: Free lunch for text-guided style transfer using diffusion models. *arXiv preprint arXiv:2401.15636*, 2024. 2, 3

[13] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 10

[14] K. Hong, S. Jeon, H. Yang, J. Fu, and H. Byun. Domain-aware universal style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14609–14617, 2021. 3

[15] Y. Hu, C. Zhuang, and P. Gao. Diffusest: Unleashing the capability of the diffusion model for style transfer. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–1, 2024. 2

[16] N. Huang, K. Huang, Y. Pu, J. Wang, J. Guo, Y. Yan, X. Li, and T.-Y. Lee. Artcrafter: Text-image aligning style transfer via embedding reframing. *arXiv preprint arXiv:2501.02064*, 2025. 3

[17] 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*, pages 1501–1510, 2017. 2

[18] Y. Jing, X. Liu, Y. Ding, X. Wang, E. Ding, M. Song, and S. Wen. Dynamic instance normalization for arbitrary style transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4369–4376, 2020. 2

[19] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 2

[20] C. G. Kamra, I. D. Mastan, and D. Gupta. Sem-cs: Semantic clipstyler for text-based image style transfer. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 395–399. IEEE, 2023. 3

[21] 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*, pages 18062–18071, 2022. 2, 3, 6, 7, 8

[22] M. Lei, X. Song, B. Zhu, H. Wang, and C. Zhang. Stylestudio: Text-driven style transfer with selective control of style elements. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23443–23452, 2025. 3

[23] S. D. Li. Diffusion-based localized image style transfer. *arXiv preprint arXiv:2403.18461*, 2024. 2

[24] H. Liu, L. Wang, Y. Zhang, Z. Yu, and Y. Guo. Samam: Style-aware state space model for arbitrary image style transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28468–28478, 2025. 2, 8

[25] Z.-S. Liu, L.-W. Wang, W.-C. Siu, and V. Kalogeiton. Name your style: text-guided artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3530–3534, 2023. 2, 3, 8

[26] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[27] 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*, pages 5880–5888, 2019. 2

[28] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021. 2

[29] L. Qu, H. Zhang, Y. Liu, X. Wang, Y. Jiang, Y. Gao, H. Ye, D. K. Du, Z. Yuan, and X. Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024. 2

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 7

[31] L. Rout, Y. Chen, N. Ruiz, A. Kumar, C. Caramanis, S. Shakkottai, and W.-S. Chu. Rb-modulation: Training-free personalization of diffusion models using stochastic optimal control. *arXiv preprint arXiv:2405.17401*, 2024. 2

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations,International Conference on Learning Representations*, Jan 2015. 3

[33] S. Singh, S. Jandial, S. Shahid, and A. Java. Least:" local" text-conditioned image style transfer. *arXiv preprint arXiv:2405.16330*, 2024. 3

[34] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16519–16529, 2021. 3

[35] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6924–6932, 2017. 2

[36] S. N. Wadekar and A. Chaurasia. Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features. *arXiv preprint arXiv:2209.15159*, 2022. 9

[37] Y. Wang, T. Bai, X. Xie, Z. Yi, Y. Wang, and R. Ma. Sigstyle: Signature style transfer via personalized text-to-image models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8051–8059, 2025. 2

[38] Z. Wang, L. Zhao, and W. Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023. 3

[39] M. Wright and B. Ommer. *ArtFID: Quantitative Evaluation of Neural Style Transfer*, pages 560–576. 09 2022. 10

[40] J. Wu, L. Hou, Z. Li, J. Liao, L. Liu, and L. Sun. Preserving structural consistency in arbitrary artist and artwork style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2830–2838, 2023. 3

[41] Y. Wu, H. Zhao, W. Chen, Y. Yang, and J. Bu. Textstyler: A clip-based approach to text-guided style transfer. *Computers & Graphics*, 119:103887, 2024. 3

[42] R. Xu, W. Xi, X. Wang, Y. Mao, and Z. Cheng. Stylessp: Sampling startpoint enhancement for training-free diffusion-based method for style transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18260–18269, 2025. 2

[43] Z. Xu, S. Xing, E. Sangineto, and N. Sebe. Spectralclip: preventing artifacts in text-guided style transfer from a spectral perspective. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5121–5130, 2024. 3

[44] 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*, volume 38, pages 7024–7032, 2024. 2, 8

[45] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 10

[46] S. Zhang, H. Kang, Y. Liu, F. Mei, and H. Li. Hsi: A holistic style injector for arbitrary style transfer. *arXiv preprint arXiv:2502.04369*, 2025. 2

[47] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023. 2

[48] S. Zheng, P. Gao, P. Zhou, and J. Qin. Puff-net: Efficient style transfer with pure content and style feature fusion network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8059–8068, 2024. 2

[49] Y. Zhou, X. Gao, Z. Chen, and H. Huang. Attention distillation: A unified approach to visual characteristics transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18270–18280, 2025. 2, 8

[50] M. Zhu, X. He, N. Wang, X. Wang, and X. Gao. All-to-key attention for arbitrary style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 23109–23119, 2023. 2