The Review of Image Style Transfer

Kang Yang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 2316024526@qq.com

Li Zhao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 332099732@qq.com

***Abstract*—Image style transfer, as a core cross-disciplinary technology in computer vision and non-photorealistic rendering, aims to preserve the semantic structure of the content image while transferring the artistic textures and brushstroke patterns of the style image. Early traditional methods struggled with complex scenes due to the semantic gap. With the advancement of deep learning, neural style transfer (NST) achieved a leap from "pixel-level statistics" to "feature-level perception." Crucially, while early optimization-based NSTs suffered from slow inference (less than 0.1 FPS), subsequent feed-forward networks enabled real-time processing (20-60 FPS). In recent years, Generative Adversarial Networks (GANs), Vision Transformers (ViTs), and denoising diffusion models have driven breakthroughs. Diffusion models, combined with parameter-efficient fine-tuning (e.g., LoRA) and distillation techniques, achieve sub-second high-fidelity generation with decoupled structural control. This paper systematically traces the technical evolution of image style transfer. The discussion unfolds across three dimensions: (1) establishing a classification system for datasets; (2) analyzing the intrinsic mechanisms of five major SOTA paradigms; and (3) comprehensively comparing mainstream evaluation metrics (such as LPIPS, FID, and Gram Matrix Distance) and model performance. Finally, this paper summarizes current challenges, such as high-resolution real-time inference, and projects future trends in video stylization and 3D scene transfer, providing a comprehensive roadmap for future researchers in this field.**

Keywords-Image Style Transfer; Neural Style Transfer; Deep Learning; Diffusion Models; Generative Adversarial Networks

I. INTRODUCTION

Images serve not only as vital mediums for communication and exchange in modern human history but also as carriers of artistic expression and aesthetic experience. In the early days of digital image processing, people enhanced images through filters or simple color transformations. However, these methods failed to capture the deeper "style" essence of an image—the integration of brushstrokes, texture structures, and high-level semantics [1]. The concept of style transfer emerged to empower computers with the ability to reconstruct the visual world like an artist.

Early style transfer research primarily fell under the categories of Texture Synthesis [2] and Image Analogies [3]. Researchers employed mathematical statistical models or patch-based matching algorithms to extract texture patches from source images and fill them into target images. While these non-parametric methods achieved some success in simple texture transfer (e.g., applying oil painting textures to photographs), their lack of understanding regarding image semantic content often resulted in structurally chaotic outputs. They struggled to handle complex scenes with strict geometric structures, such as human faces or architecture.

The landscape of this domain has been fundamentally transformed by recent advancements in deep neural architectures. A seminal contribution occurred in 2015 when Gatys et al. [4] leveraged convolutional neural networks (CNNs) to disentangle internal image representations. Their research revealed that an image's semantic content is maintained by deeper network activations, whereas its artistic texture can be accurately quantified using the Gram matrix of these convolutional feature maps. This foundational insight catalyzed the emergence of the Neural Style Transfer (NST) discipline. However, while their methodology yielded visually impressive artistic renditions through the joint minimization of style and content penalties,

the underlying reliance on a slow, iterative optimization framework demanded substantial computational resources, thereby precluding its deployment in instantaneous or real-time scenarios.

The computational inefficiency inherent in iterative optimization was subsequently resolved by utilizing feed-forward generative architectures [5], trained via perceptual loss functions. While these preliminary models accelerated processing speeds dramatically, they were constrained to specific, pre-trained artistic styles. The true breakthrough for zero-shot, arbitrary stylization emerged from manipulating feature-space statistics. Techniques such as adaptive instance normalization (AdaIN) [6] and whitening and coloring transforms (WCT) [7] demonstrated that artistic textures could be dynamically infused by aligning the mean and variance, or covariance matrices, of content and style feature representations in real-time.

With the introduction of Generative Adversarial Networks (GANs) [8], style transfer was reframed as an image-to-image translation problem. CycleGAN [9] addressed the challenge of unpaired data by incorporating a cycle consistency loss, enabling cross-domain transfers like "photo-to-painting". Subsequently, the successful Transformer architecture from natural language processing was adapted for visual tasks. Vision Transformer (ViT) [10] leveraged self-attention mechanisms to capture long-range pixel dependencies, resolving CNN's issue of local texture inconsistencies during stylization. By focusing on global semantic information, it significantly enhanced style transfer's structural preservation capabilities.

In the past two years, generative AI based on diffusion models [11] has demonstrated potential surpassing all previous methods. By simulating the denoising process from Gaussian noise to data, diffusion models (such as Stable Diffusion [12] and ControlNet [13]) utilize text prompts or image guidance to achieve not only extremely refined style transfer but also decoupled control over style and content, ushering in a new era of AI-driven artistic creation.

Despite continuous technological iteration, image style transfer still faces numerous challenges, such as memory bottlenecks at ultra-high resolutions, semantic errors caused by style bleeding, and frame-to-frame flickering in video style transfer. This paper will detail the evolutionary logic of these technical approaches, deeply analyze the underlying mechanisms of various methods, and explore future solutions based on the latest literature.

II. DATASETS

In deep learning-driven style transfer research, dataset selection critically impacts model training effectiveness, generalization capabilities, and style capture accuracy. Style transfer datasets typically comprise two components: content image datasets and style image datasets.

A. Content Datasets

Content datasets aim to provide diverse real-world scenarios, ensuring models learn structural features of various objects (e.g., people, vehicles, buildings, animals) to maintain semantic integrity during stylization.

1) MS COCO (Microsoft Common Objects in Context)

As the predominant benchmark for content representation in stylization research, the MS COCO collection is almost universally adopted. Introduced in 2014, it encompasses roughly 330,000 natural images distributed across 80 distinct semantic classes [14].

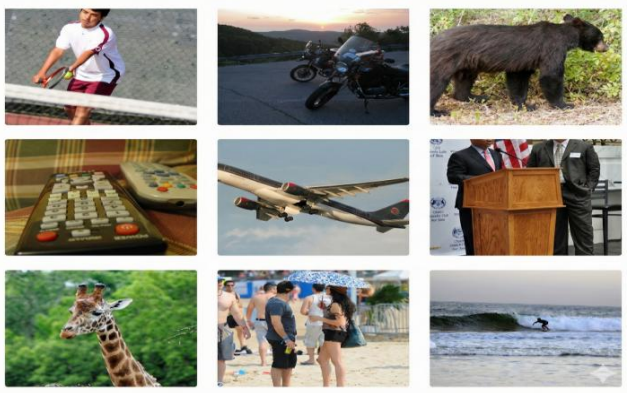

Figure 1. MS COCO Datasets

The repository is particularly valued for its intricate contextual layouts and extreme variations

in spatial object scaling. Consequently, when optimizing rapid feed-forward architectures or attention-guided arbitrary stylization frameworks (such as the AdaIN and AdaAttN baselines), this dataset is conventionally deployed. Supplying these highly complex scenes as content inputs inherently compels the neural networks to robustly maintain underlying semantic geometries and structural integrity, even amidst chaotic background environments.

2) ImageNet

Curated by researchers at Stanford University under the leadership of Fei-Fei Li, the ImageNet repository [15] encompasses in excess of 14 million explicitly labeled pictures distributed across upwards of 20,000 distinct semantic classes. Although primarily designed for classification tasks, its massive scale and extensive category coverage make it a foundational dataset for pre-training feature extractors (backbones) like VGG or ResNet [16]. It also serves as training data for numerous style transfer models, particularly GAN-based approaches.

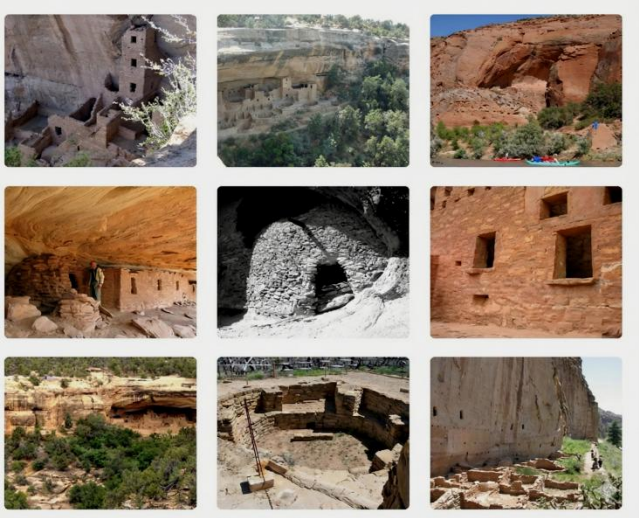

Figure 2. ImageNet Datasets

3) Places2

Places2 [17] is a large dataset focused on scene understanding, containing over 10 million images across more than 400 scene categories (e.g., forests, streets, bedrooms). Compared to ImageNet, Places2 offers greater diversity in scene types and is frequently used for landscape-based image style transfer tasks, enabling models to more readily capture overall composition and regional texture features.

Figure 3. Places2 Datasets

4) BDD100K and Autonomous Driving Datasets

While MS COCO and ImageNet are standard, large-scale driving datasets like BDD100K [37] have increasingly been utilized to explore style transfer for downstream application tasks. By applying style transfer models to BDD100K, researchers can efficiently convert daytime driving scenes into nighttime or adverse weather scenes, providing critical data augmentation for training robust autonomous driving systems.

B. Style Datasets

Target domain collections generally encompass a broad spectrum of visual art forms to establish robust aesthetic references. These repositories span from traditional mediums, such as oil brushwork and watercolor, to highly stylized graphic formats including Japanese Ukiyo-e woodblock prints and contemporary animation cels. The inherent richness and heterogeneity of these visual inputs are critical; they fundamentally dictate the capacity of the neural network—particularly when aligning complex textures via attention mechanisms and contrastive spatial projections—to generalize and render diverse aesthetic abstractions effectively.

1) WikiArt

Currently one of the largest and most comprehensive visual art encyclopedias. In style

transfer research, researchers often scrape its public data to build datasets, typically containing 80,000 to 100,000 artworks categorized by style (Impressionism, Cubism, Realism, etc.) or artist (Van Gogh, Picasso, Monet, etc.). This dataset serves as the core training material for any style transfer network (e.g., Avatar-Net, StyTr2 [18]), enabling models to learn brushstroke textures and color distributions across different artistic genres.

Figure 4. WikiArt Datasets

2) *Painter by Numbers*

Contains a large collection of painting images labeled with artist and style tags. This dataset is commonly used to train specific style generation models or evaluate a model's fit to the style of particular artists (e.g., Monet, Van Gogh, Picasso), making it suitable for tasks like style recognition and style embedding learning.

Figure 5. Painter by Numbers Datasets

3) *Google Arts & Culture*

Provides a large collection of professionally scanned, ultra-high-resolution artworks, suitable for high-fidelity style transfer research. Compared to WikiArt, it excels at capturing fine details and textures in artworks, particularly brushstroke-level local structures, making it a crucial resource for studying micro-level style control.

Figure 6. Google Arts & Culture Datasets

III. METHODS

The evolution of image style transfer technology has progressed from slow optimization-based algorithms to fast model-based inference, and ultimately to fine-grained control enabled by generative large models. This chapter will systematically analyze the core concepts and technical paradigms of representative methods across these stages, following the logical progression of this technological trajectory.

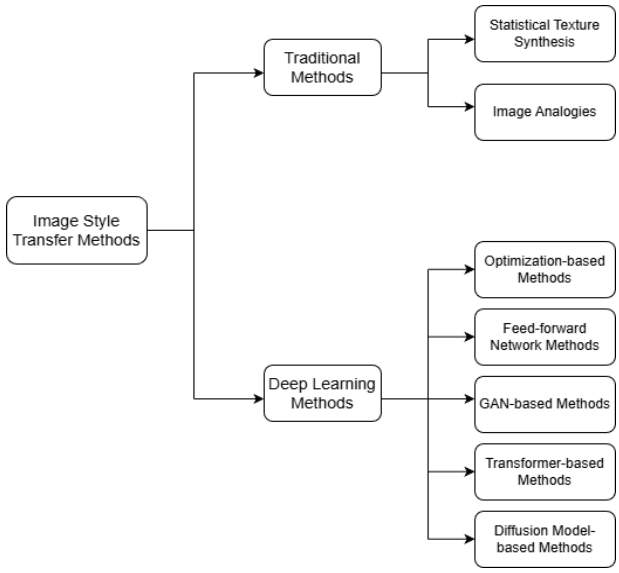

Figure 7. Image Style Transfer Methods Framework

A. Traditional Methods

Historically, prior to the widespread adoption of neural architectures, the stylization of images was largely categorized as a specific subset of texture generation. Rather than extracting high-level semantic comprehension from the visual content, these conventional frameworks depended entirely on non-parametric patch borrowing or the statistical alignment of foundational visual cues. This typically involved manipulating basic physical properties, including luminance intensities, directional gradients, and color distribution metrics.

1) Statistical Texture Synthesis

Early In foundational studies regarding visual pattern replication, Portilla and Simoncelli introduced an analytical framework driven by wavelet decompositions. Their methodology achieved texture recreation by strictly aligning multi-scale statistical indicators—specifically focusing on cross-correlation metrics. Despite its proficiency in imitating homogeneous patterns, this mathematical paradigm was inherently limited by its inability to interpret or maintain the high-level semantic structures present in natural photographs [19].

2) Image Analogies

The image analogy algorithm proposed by Hertzmann et al. [3] attempts to learn the transformation relationship between a pair of images (image A' and its filtered version B) and apply it to a target image B' to generate $t = \mathrm{AdaIN}(f(c), f(s))$. This method is based on multiscale Markov Random Fields (MRF) and patch matching, formulated as finding the optimal matching patch q to minimize the distance:

$$q = \arg\min_{p \in A} \left\| F(p) - F(q') \right\|^2 \qquad (1)$$

Here p denotes candidate patches in the source image A , q' represents the patch (or its neighborhood features) at the target location in the destination image B , and $F(\cdot)$ is the feature vector containing luminance, color, or texture information. By traversing the search to find the block p with the smallest Euclidean distance, the corresponding filtered version of its texture can be filled into the target position. While such methods yield acceptable results for simple texture transfer tasks like "sketching" or "oil painting," they suffer from extremely high computational costs. Furthermore, lacking high-level semantic constraints, they often lead to structural breakdown in complex scenes.

Traditional methods lack "understanding" of image content, relying solely on statistical matching at the pixel or low-level feature level. This prevents the organic integration of content structure and artistic style, and computational complexity typically increases exponentially with image resolution.

B. Deep Learning Approaches

Convolutional architectures intrinsically learn hierarchical visual representations: early layers detect primitive elements such as structural boundaries and chromatic shifts, whereas deeper layers encode complex semantic categories. This inherent capacity to disentangle low-level textures from high-level content establishes a robust foundation for neural stylization.

1) Optimization-based method

In 2015, Gatys et al. [4] established the theoretical baseline for this discipline by introducing a framework driven by image-level optimization. The foundational concept relies on utilizing a frozen, pre-trained convolutional backbone (such as VGG-19) to independently extract structural and aesthetic representations, seeking to satisfy both constraints simultaneously through direct pixel manipulation.

Under this paradigm, the neural weights remain completely unaltered; rather, the synthesis procedure designates the target image itself as the sole trainable variable. To preserve spatial coherence, deep activation layers enforce strict geometric constraints based on the source photograph. Concurrently, early-stage network representations — quantified via Gram matrix computations—dictate the chromatic and textural mapping, compelling the synthesized output to statistically mirror the reference artwork. These structural and textural penalties are subsequently

integrated into a unified objective function. Finally, gradient-based solvers, typically L-BFGS or Adam, iteratively refine the image pixels to minimize this combined loss.

Pixel values are optimized using gradient descent (L-BFGS or Adam) via backpropagation. This method offers high stylistic quality and flexibility but suffers from extremely slow inference speeds (taking several minutes per image) and requires retraining for each image pair.

2) Feed-Forward Network Method

To bypass the computational bottlenecks inherent in iterative optimization, subsequent research shifted towards feed-forward convolutional architectures. By pre-training a dedicated generative network across extensive datasets, this approach achieves instantaneous image stylization via one-stage inference. Depending on the model's capacity for aesthetic generalization, this paradigm is broadly categorized into two developmental stages: single-style frameworks tailored to individual visual patterns, and arbitrary stylization architectures driven by dynamic feature alignment.

a) Fast Style Transfer

To overcome the bottleneck of slow iterative optimization, Johnson et al. [5] and Ulyanov et al. [20] proposed a feed-forward network-based generative paradigm that directly maps content images to stylized images.

This method constructs an Image Transformation Network (ITN) that directly maps input content images into stylized images. During training, a fixed VGG-16 serves as the loss network to compute losses, updating ITN weights through backpropagation rather than modifying image pixels. Key technical innovations are embodied in two aspects:

- Network Architecture: ITN typically adopts a fully convolutional architecture (U-Net variant or ResNet architecture) incorporating downsampling, residual blocks, and upsampling. This ensures a large receptive field to capture stylistic textures.
- Instance Normalization (IN): Ulyanov et al. discovered that traditional Batch Normalization (BN) is unsuitable for style transfer because BN utilizes statistics across the entire batch, erasing unique contrast information (i.e., style) from individual images. They proposed IN, which normalizes only the spatial dimensions of a single image. Experiments demonstrate that IN significantly improves convergence speed and visual quality in stylization.

This network model requires only a single forward pass during inference, achieving speeds three orders of magnitude faster than optimization methods (achieving 20-60 FPS on modern GPUs). However, its critical flaw is poor flexibility: a single network can learn only one fixed style. Whenever a new artistic style is required, the entire network must be retrained, failing to meet diverse user demands.

b) Arbitrary Style Transfer

To bypass the inefficiency of training dedicated networks for individual aesthetics, subsequent studies incorporated feature projection modules. The foundational mechanism relies on synchronizing the latent activation statistics between the source and reference inputs, thereby granting a unified architecture the capacity to synthesize unseen visual patterns.

AdaIN (Adaptive Instance Normalization): Proposed by Huang and Belongie [6], style is fundamentally the mean and variance in feature space. The AdaIN layer takes content features x and style features y, first normalizing x, then scaling and shifting it to match the statistical distribution of y :

$$AdaIN(x,y)=\sigma(y)\left(\frac{x-\mu(x)}{\sigma(x)}\right)+u(y) \qquad (2)$$

Here, channel-wise expectations and variances are denoted by $\mu(\cdot)$ and $\sigma(\cdot)$, respectively. The framework relies on an autoencoder design, utilizing a frozen VGG backbone to map inputs into structural representations $f(c)$ and aesthetic representations $f(s)$. Following their integration

through $t = \text{AdaIN}(f(c), f(s))$, a synthesis network g decodes this combined latent vector t back to the pixel space. The training objective ensures the synthesized output $g(t)$ statistically mirrors the reference $f(s)$. Notably, this formulation pioneered instantaneous, universal stylization, completely eliminating the need for iterative optimization during the inference stage.

WCT (Whitening and Coloring Transform): Li et al. [7] introduced a more rigorous covariance matrix transformation. First, content features undergo whitening to remove their inherent style information, followed by coloring to match the covariance structure of style features. Unlike AdaIN, which aligns only first-order (mean) and second order (variance) statistics, WCT captures feature correlations by matching covariance matrices, yielding richer texture details. However, due to SVD decomposition, its inference speed is slightly slower than AdaIN.

SANet and AdaAttN: Park et al. [21] noted that AdaIN and WCT rely on global statistics, often leading to uneven local style distributions. To address this, SANet introduced an attention mechanism to calculate the similarity between content and style feature maps. Building upon this, Liu et al. proposed AdaAttN [22], which deeply integrates the attention mechanism into the normalization process. Furthermore, Kolkin et al. [23] innovatively redefined the distance between content and style distributions using Optimal Transport theory to achieve breakthroughs in geometric alignment. Recent state-of-the-art advancements built on the AdaAttN baseline further incorporate multi-scale coordinate attention modules and contrastive learning frameworks (such as PatchNCE for structure-preserving projection) [38]. These enhancements, along with related internal-external learning frameworks [24], significantly boost the model's mutual information, ensuring strict semantic alignment while applying complex artistic abstractions.

3) GAN Method

The introduction of GANs expanded the style transfer task from the traditional "texture matching" paradigm to a higher-level "domain adaptation" framework, primarily addressing the holistic mapping problem from the real photo domain to the artistic domain.

a) Cross-Domain Image Translation

Early explorations relied heavily on strongly supervised learning. Isola et al.'s Pix2Pix [25] established the benchmark for image translation using paired data. However, acquiring corresponding photographs and paintings of the same scene is nearly impossible in artistic style transfer. Addressing this challenge, Zhu et al.'s [9] CycleGAN became a milestone in unsupervised style transfer. CycleGAN introduced the Cycle Consistency Loss, constraining images to self-replicate after traversing the "source domain, target domain, source domain" mapping ($F(G(x)) \approx x$). Combined with adversarial loss, CycleGAN successfully achieved unsupervised "photo-to-oil painting" transfer. However, lacking geometric constraints, it tends to produce unexpected geometric distortions when handling significant shape transformations (e.g., transforming a horse into a zebra).

b) Multimodal Generation & Attention Mechanisms

To address CycleGAN's limitations of monotonous outputs and insufficient local details, subsequent research refined the generation paradigm: MUNIT and DRIT methods aimed to achieve multimodal generation [26]. By decomposing images into "Content Code" and "Style Code," they enabled exchanging different style codes within a shared content space, thereby enabling "one-to-many" mapping where a single content image generates multiple stylistic variants.

To further enhance the quality of key regions (e.g., facial features) during cross-domain transformations, Kim et al.'s U-GAT-IT [27] introduced a Class Activation Map (CAM)-guided attention module. This module adaptively focuses on areas with the most significant differences between source and target domains, demonstrating outstanding performance on tasks like face-to-cartoon conversion. Additionally, while StyleGAN [28] by Karras et al. is primarily designed for image generation, its refined decoupling architecture in the latent space laid a

crucial foundation for subsequent high-quality style editing based on GAN inversion.

4) Transformer

Traditional CNN architectures primarily rely on local convolution kernels, constrained by limited receptive fields, making it difficult to capture global long-range dependencies. This limitation often renders them inadequate when processing styles with significant geometric structures, such as geometric abstract paintings. The emergence of Vision Transformers (ViT), leveraging self-attention mechanisms, has introduced a novel paradigm for addressing this challenge of capturing global features.

StyTr2 (Style Transformer): StyTr2, proposed by Deng et al. [18], stands as a representative work in this field. This method utilizes a Transformer encoder to separately extract content and style sequences. It then calculates a similarity matrix between the two using a multi-head attention mechanism, thereby achieving precise style transfer based on semantic correspondence.

Its key technical innovations are manifested in three aspects:

- Images are first segmented into fixed-size patches and flattened into sequence embeddings to adapt to the Transformer's input format.
- Attention-Driven Feature Fusion: This constitutes the model's core mechanism. The encoder utilizes self-attention to capture the image's global structure, while the decoder employs cross-attention for stylization. Specifically, the content sequence serves as the query vector (Q), and the style sequence functions as both the key vector (K) and the value vector (V). The attention calculation formula is as follows:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (3)$$

In this formulation, the matrix Q acts as the spatial prompt, which is derived directly from the content representation. The artistic characteristics are encoded into the key and value pairs, denoted as K and V, respectively. To prevent the dot-product magnitudes from pushing the softmax function into regions with vanishing gradients, a scaling factor determined by the dimension d_k is applied. By evaluating the scaled correspondence between the content queries and style keys, the network dynamically aggregates the required artistic textures from V, effectively bridging the gap between global semantic structures and local brushstroke details.

- Positional Encoding: Content-Aware Positional Encoding is employed, which adapts better to inputs of varying scales compared to traditional fixed sine encoding. This enhances the robustness of style transfer against object scale and position variations.

Compared to CNN-based methods, images generated using the Transformer exhibit greater artistic quality in macro-structural aspects and significantly reduce texture artifacts caused by mismatched local features.

5) Diffusion Model

Beginning around 2022, probabilistic diffusion frameworks swiftly established themselves as the dominant standard for visual synthesis. By exploiting massive, pre-trained priors (such as Stable Diffusion), these stochastic systems synthesize exceptionally high-fidelity visuals via a distinct trajectory of progressive corruption followed by iterative reverse denoising. When applied to aesthetic rendering, such diffusion-based pipelines effectively resolve legacy bottlenecks associated with preserving structural integrity while injecting complex textures. They accomplish this delicate equilibrium utilizing advanced techniques including latent inversion, spatial conditioning, and parameter-efficient adaptations.

a) Inversion & Guidance for Reasoning Optimization

To preserve the spatial structure of the content image during generation, researchers typically employ DDIM inversion techniques. This method

deterministically maps input images to the initial noise space of diffusion models, ensuring the starting point of denoising contains the layout information of the original image. During inverse denoising, style text prompts [29] or style image features [30] can be introduced. Combined with CLIP guidance or classifier-free guidance, this enables generated images to progressively align with target style features, achieving coordination between content consistency and style transformation.

b) Structural & Attention Control

To mitigate the geometric ambiguity inherent in purely text-guided synthesis, subsequent studies exploited internal architectural components to enforce rigid structural boundaries. For instance, the Prompt-to-Prompt framework by Hertz et al. revealed how cross-attention matrices embedded deeply within the U-Net dictate the synthesized topological arrangements. By transferring these latent attention masks from the source input during the generation phase, one can perfectly lock object coordinates and morphologies while allowing the target aesthetic to flourish independently. Alternatively, Zhang et al. [13] introduced ControlNet, which augments frozen diffusion pipelines with trainable parallel branches initialized through zero-weight convolutions. This architecture empowers the system to incorporate explicit spatial conditionings—ranging from depth estimations and skeletal poses to discrete edge representations. As a result, the stylization process remains tightly bound to the original image's structural footprint, comprehensively eliminating semantic distortion and texture spillover artifacts.

c) Personalized Fine-Tuning

For customization needs targeting specific artistic styles or subjects, parameter-efficient fine-tuning (PEFT) methods have become mainstream. Hu et al.'s LoRA [32] offers a highly efficient lightweight fine-tuning solution. Unlike traditional full-parameter training, LoRA efficiently injects specific artistic styles (e.g., "ink wash painting," "cyberpunk") into models by embedding trainable low-rank decomposition matrices into the pre-trained model's attention layers. This requires updating only a minimal number of parameters, significantly reducing training costs and memory requirements.

Dream Booth [31], proposed by Ruiz et al., focuses on preserving subject consistency. By binding images of specific subjects to unique identifiers during fine-tuning, this method achieves extremely high fidelity when placing particular objects (like one's pet dog) into diverse scenes. Notably, LoRA is now frequently integrated into DreamBooth's training workflow as an alternative to full-model fine-tuning. These fine-tuning techniques can synergize with the previously mentioned ControlNet and DDIM Inversion, enabling diffusion models to achieve powerful personalized style customization while strictly preserving content structure.

Overall, diffusion-based methods significantly outperform GANs and traditional optimization approaches in terms of generation quality, texture expressiveness, and stylistic control flexibility. However, their high computational cost and slow inference speed remain major bottlenecks limiting real-time applications. Current research focuses on addressing this issue through distillation techniques and accelerated sampling algorithms.

IV. Performance

Similar to most image generation tasks, evaluating style transfer faces challenges of high subjectivity and the absence of a single ground truth. Therefore, the academic community typically employs a combination of qualitative and quantitative metrics for comprehensive assessment.

A. Qualitative Evaluation

The most intuitive method is manual visual inspection (User Study). Typically, a set of content images, style images, and outputs from different algorithms are presented. Volunteers are invited to score the outputs across three dimensions: "Style Consistency," "Content Preservation," and "Overall Aesthetics."

B. Quantitative Evaluation

To achieve automated and standardized evaluation, the research community has proposed a series of mathematical metrics:

1) Inference Speed (FPS)

In practical applications, especially for mobile filters and real-time video stylization, model inference speed and computational overhead are critical metrics. Common measurements include FPS (Frames Per Second) or average inference time per image. Significant performance differences exist across different technical approaches: Gatys optimization method: < 0.1 FPS, extremely slow, not real-time capable; Feedforward networks (Johnson/AdaIN): 20–60 FPS (GPU), meeting real-time requirements; Diffusion models: Original sampling takes several seconds, but can be reduced to sub-second levels when combined with distillation techniques.

2) Gram Matrix Distance (Style Loss)

This metric evaluates textural and aesthetic coherence by measuring the discrepancy between the Gram matrices of the synthesized output and the reference artwork within a pre-trained VGG manifold. Minimized scores denote that the statistical correlations of the synthesized features tightly align with the desired visual patterns.

3) LPIPS (Learned Perceptual Image Patch Similarity)

LPIPS (Learned Perceptual Image Patch Similarity): Evaluates content preservation by calculating the perceptual distance between the generated image and the content image in a deep feature space (e.g., AlexNet) using the following formula [33]:

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw}) \right\|_2^2 \quad (4)$$

Rather than computing pure pixel-wise differences, this metric aggregates perceptual variations across multiple deep layers l. The terms $\hat{y}^l$ and $\hat{y}^l_0$ signify the intermediate activation maps of the synthesized output and the original content input, respectively, extracted from a pre-trained backbone. The spatial dimensions are normalized by H_l and W_l, while w_l assigns importance weights to different channels. A minimized score under this formulation proves that the network has robustly preserved the intrinsic geometric boundaries and semantic layout of the source image.

4) Deception Rate

Deception Rate: For specific artist style transfer, a classifier trained on that artist's works can be used to test the probability of a generated image being classified as an "authentic piece." This evaluates the style's credibility and similarity. A higher probability indicates greater stylistic credibility.

5) FID (Fréchet Inception Distance)

Calculates This metric quantifies the statistical divergence between the feature representations of the synthesized outputs and the genuine artwork domain [34]. It functions as a standard benchmark to evaluate the comprehensive visual fidelity of the rendering process:

$$FID = \left\| \mu_r - \mu_g \right\|_2^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (5)$$

Where In this equation, the statistical essence of the authentic artistic domain is captured by its expectation μ_r and covariance matrix Σ_r, which are mapped within the Inception latent space. Conversely, μ_g and Σ_g parameterize the distribution of the generated stylized outputs. The trace operator Tr evaluates the divergence between these two multivariate Gaussian approximations. Achieving a lower Wasserstein-2 distance indicates a superior generative capability, meaning the synthesized images have successfully captured the macro-level color schemes and micro-level brushstroke statistics of the target art style.

V. Conclusions

Image style transfer, as a core interdisciplinary direction spanning computer vision and generative modeling, has undergone rapid and profound evolution over the past decade. This paper systematically reviews the development trajectory of this technology across multiple dimensions, including methodological evolution, dataset construction, recent advances in diffusion models, and evaluation frameworks. Overall, the field has completed a paradigm shift from "statistical

texture imitation" to "semantic-based deep generation," achieving a leap from "reproducing painting styles" to "assisting artistic creation" through deep learning. However, numerous challenges remain before constructing intelligent art systems that truly possess human aesthetic judgment and physical common sense.

A. Summary of Technological Evolution

Reviewing the development journey, the technical paradigm of image style transfer has undergone four key stages of iteration: (1) Traditional Methods Stage: Primarily relied on statistical matching of low-level visual features (color histograms, gradients) and non-parametric texture synthesis, constrained by the lack of semantic understanding of images; (2) CNN Era: Gatys et al. established the theoretical foundation for decoupling content and style using Gram matrices derived from deep convolutional features. Subsequently, feedforward networks like AdaIN enabled real-time arbitrary style transfer on mobile devices; (3) GAN Era: The introduction of adversarial training mechanisms solved domain adaptation issues without paired data (e.g., CycleGAN), enabling transfer tasks to transcend the texture level and achieve object-level shape transformations and geometric deformations; (4) Large Model Era (Transformer & Diffusion): Leveraging the global attention mechanism of Vision Transformers and the probabilistic denoising generation of diffusion models, this era completely overcame bottlenecks in semantic consistency and detail generation quality, enabling text-controllable, structure-preserving high-dimensional artistic creation.

B. Future Development Directions

Despite significant progress, image style transfer still faces multifaceted challenges on its path toward becoming a genuine artistic creation tool:

- Video Temporal Consistency: Addressing the "flickering" effect in frame-by-frame generation remains a core challenge in video stylization. Recent diffusion-based video editing research has actively explored integrating cross-frame correspondences with temporal attention mechanisms to maintain semantic and visual coherence, thereby ensuring stable long-sequence generation [39].
- 3D Scene Stylization: With the proliferation of Neural Radiance Fields (NeRF) and 3D Gaussian Splatting, style transfer is rapidly expanding from 2D to 3D spaces. A frontier challenge lies in directly embedding stylistic features into 3D representations—such as optimizing the spherical harmonics in Gaussian Splatting—while strictly preserving geometric structures and ensuring multi-view textural consistency [40].
- Addressing "style bleeding" issues (e.g., incorrectly transferring sky textures onto human faces) [35] is crucial for enhancing realism. Early work by Luan et al. on Deep Photo Style Transfer [36] innovatively introduced semantic segmentation masks as constraints, preventing background textures from interfering with foreground objects and achieving photo-realistic high-fidelity style transfer for the first time. Building on this approach, future models require enhanced semantic segmentation capabilities to enable precise, region-specific style control across different image areas.
- Efficient Edge Inference: Addressing the computational bottleneck of diffusion models, model distillation, quantization, and lightweight architecture design (e.g., Mobile-Diffusion) are essential for achieving real-time, state-of-the-art style transfer on mobile devices.
- Personalized Creation and Artistic Ethics Assurance: User demand for personalized style transfer is growing, yet existing methods typically rely on large-scale training datasets. Few-shot learning or Low-Rank Attentive Refinement (LoRA) strategies enable rapid customization of personalized style models.

In summary, image style transfer stands at the epicenter of deep generative model innovation. Its

future applications extend far beyond entertainment filters, poised to profoundly impact digital art creation, metaverse content development, and visual effects production in the film industry—emerging as a pivotal engine for AI-empowered creative industries.

REFERENCES

[1] A. Hertzmann, "Visual indeterminacy in GAN art," Leonardo, vol. 53, no. 4, pp. 424–428, 2018.

[2] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), vol. 2, 1999, pp. 1033–1038.

[3] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," in Proceedings of SIGGRAPH, 2001, pp. 327–340.

[4] 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 (CVPR), 2016, pp. 2414–2423.

[5] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 694–711.

[6] 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 (ICCV), 2017, pp. 1501–1510.

[7] Y. Li et al., "Universal style transfer via feature transforms," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.

[8] I. Goodfellow et al., "Generative adversarial nets," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 27, 2014.

[9] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2223–2232.

[10] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Int. Conf. Learn. Represent. (ICLR), 2021.

[11] . Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 6840–6851.

[12] 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 (CVPR), 2022, pp. 10684–10695.

[13] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3836–3847.

[14] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in Proceedings of the European Conference on Computer Vision (ECCV), 2014, pp. 740–755.

[15] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[17] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1452–1464, 2017.

[18] Y. Deng et al., "StyTr2: Image style transfer with transformers," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11326–11336.

[19] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," Int. J. Comput. Vis., vol. 40, no. 1, pp. 49–70, 2000.

[20] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," arXiv preprint arXiv:1607.08022, 2016.

[21] 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 (CVPR), 2019, pp. 5880–5888.

[22] S. Liu et al., "AdaAttn: Revisit attention mechanism in arbitrary neural style transfer," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6649–6658.

[23] N. Kolkin, J. Salavon, and G. Shakhnarovich, "Style transfer by relaxed optimal transport and self-similarity," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10051–10060.

[24] H. Chen et al., "Artistic style transfer with internal-external learning and contrastive learning," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, 2021, pp. 26561–26573.

[25] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1125–1134.

[26] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 172–189; H.-Y. Lee et al., "Diverse image-to-image translation via disentangled representations," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 35–51.

[27] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in Proceedings of the International Conference on Learning Representations (ICLR), 2020.

[28] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4401–4410.

[29] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in Proceedings of the International Conference on Machine Learning (ICML), 2022, pp. 16784–16804.

[30] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven

image-to-image translation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1921–1930.

[31] N. Ruiz et al., "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22500–22510.

[32] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in Proceedings of the International Conference on Learning Representations (ICLR), 2022.

[33] 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 (CVPR), 2018, pp. 586–595.

[34] 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," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 6626–6637.

[35] A. J. Champandard, "Semantic style transfer and turning two-bit doodles into fine artworks," arXiv preprint arXiv:1603.01768, 2016.

[36] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4990–4998.

[37] F. Yu et al., "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020.

[38] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020.

[39] C. Qi et al., "FateZero: Fusing attentions for zero-shot text-based video editing," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15932–15942.

[40] K. Liu et al., "StyleGaussian: Instant 3D style transfer with Gaussian splatting," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7263–7273.