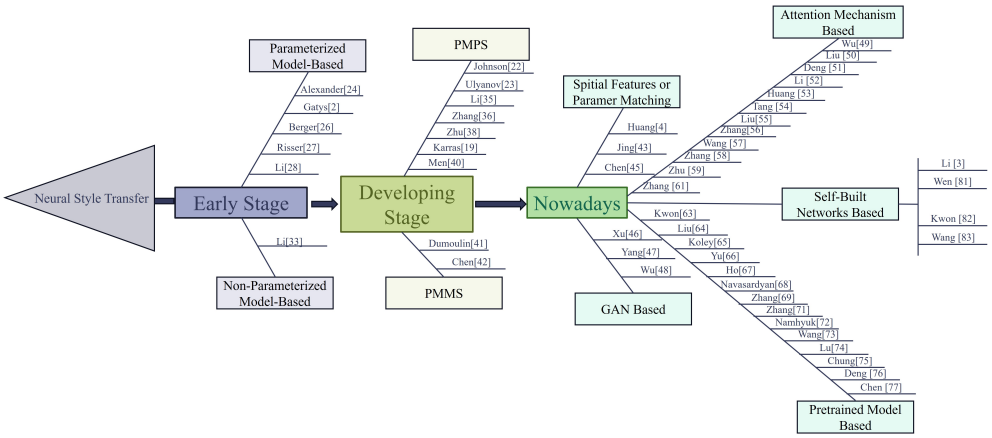

Figure 1: Overview and Category of Style Transfer

Graphical Abstract

Bridging the Metrics Gap in Image Style Transfer: A Comprehensive Survey of Models and Criteria

Xiaotong Zhou*, Yuhui Zheng, Junming Yang

Highlights

Bridging the Metrics Gap in Image Style Transfer: A Comprehensive Survey of Models and Criteria

Xiaotong Zhou*, Yuhui Zheng, Junming Yang

- Introducing achievements in style transfer in the chronological order and providing a subdivision method for neural style transfer.
- Providing a comprehensive summary and analysis of objective evaluation metrics in the field of style transfer.
- Discussing the existing issues in the field of style transfer.

Bridging the Metrics Gap in Image Style Transfer: A Comprehensive Survey of Models and Criteria

Xiaotong Zhou*

School of Computer Science, Nanjing University of Informaiton Science & Technology, Nanjing, 210031, Jiang Su, China

Yuhui Zheng

School of Computer Science, Nanjing University of Informaiton Science & Technology, Nanjing, 210031, Jiang Su, China

Junming Yang

School of Software, Nanjing University of Informaiton Science & Technology, Nanjing, 210031, Jiang Su, China

Abstract

Image style transfer is a technique that combines the content of a real photograph with the artistic style of another image to create a new and stylized image. In this paper, we aim to provide a comprehensive review of the field, tracing its development from traditional methods to modern neural network-based approaches. We focus on recent advancements utilizing Transformers and diffusion models, which represent significant progress in generative AI. Furthermore, we systematically analyze the objective evaluation metrics employed in style transfer, offering valuable insights for researchers. Finally,this paper highlights unresolved issues and suggests future directions to foster innovation in style transfer techniques.

Keywords: Image Style Transfer, Deep Learning, Art and Technology

1. introduction

A picture is worth a thousand words and excellent artworks frequently provide information distinct from that of real photographs. However, without long-term professional training, an ordinary person may find it difficult to

independently create a piece of art that satisfies themselves or others. Moreover, the time and cost required to train a true artist are often immeasurable. Even for a skilled artist proficient in creating stylized works, completing a single art piece requires a amount of time. To efficiently convert real photos into artistic images, the task of image style transfer has emerged.

Image style transfer aims to combine the content of a real photograph with the style of an artwork, creating a stylized image that simultaneously embodies the content of the photograph and the style of the artwork. In style transfer tasks, the image providing the content is referred to as the content image, while the image providing the style is called the style image. The generated result is known as the stylized image. This paper primarily focuses on image style transfer; unless otherwise specified, the term "style transfer" in the following text refers specifically to image style transfer.

The development of style transfer can be divided into two stages. The first stage spans from the mid-1990s[**?**] to 2016[**?**], characterized by the use of mathematical models for texture simulation. The second stage, from 2016[**?**] to the present, is marked by the use of deep learning and neural networks for style transfer. The former is relatively traditional, while the latter incorporates new methods. Therefore, this paper refers to the first stage as "traditional style transfer" and the second stage as "neural style transfer."

Traditional style transfer typically employs mathematical and signal processing techniques, such as texture synthesis, histogram matching, and filtering. These methods involve manipulating pixels to simulate the desired style. For example, frequency domain filtering can be used to enhance or suppress certain frequency components of an image, thereby altering its appearance. The advantage of traditional style transfer lies in its higher computational speed and lower resource consumption, but it may struggle to capture more advanced artistic styles and textures.

In contrast, neural network-based style transfer methods are more flexible and efficient. Deep learning techniques are used to learn and apply image styles. They train neural networks to capture the features of different artistic styles, which are then applied to the input image to generate a new image with the desired style. The strength of neural style transfer lies in its ability to better capture the details and complexities of artistic styles, though the required time and resource consumption can vary significantly depending on the network structure.

Traditional style transfer and neural style transfer should not be viewed

as mutually exclusive. On the contrary, some of the recent neural style transfer advancements[? ?] draw inspiration from traditional style transfer and digital image processing. Meanwhile, neural network-based style transfer techniques also have certain drawbacks, such as artifacts and difficulties in controlling the stylization process. Combining traditional and neural style transfer methods could potentially yield better results.

Image style transfer has a wide range of practical applications, such as in environmental atmosphere rendering[?], font generation[?], font recognition[?], portrait editing [? ?], design assistance [? ? ? ? ?], photo restoration[?], virtual reality (VR), and augmented reality (AR)[?]. Furthermore, as a fundamental task in computer vision, style transfer can assist in other research areas, such as adversarial example[? ?], image generation[?], and domain adaptation[?]. Whether from the perspective of practical applications or scientific research, the task of style transfer has broad applications.

In recent years, review studies in the field of style transfer have made significant progress, providing valuable references for researchers. However, these reviews still have certain limitations. For example, the early review by Kyprianidis et al. [?] focused primarily on style transfer tasks based on traditional methods, which no longer meet the needs of current tasks that are based on neural networks. Although the article by Jing et al. [?] systematically summarized neural style transfer methods and broke down the task into multiple steps, providing great guidance for beginners, its coverage is somewhat limited following the widespread application of Transformer and diffusion models. Additionally, the article by Cai et al. [?] focuses more on style transfer results based on Generative Adversarial Networks (GANs), but similarly fails to delve into the latest advancements with Transformers and diffusion models. More importantly, these reviews fall short in discussing the objective evaluation standards for style transfer. The goal of style transfer is to generate images that blend style and content features, but their artistic quality is often difficult to quantify using unified objective metrics. Despite this, many studies in the field still attempt to evaluate the quality of models through certain objective criteria, though the use of these standards lacks consistent guidelines.

To address these shortcomings, this article presents a staged review of the research on style transfer, organized chronologically, and systematically supplements the latest advances in Transformer and diffusion models. Moreover, this article provides a detailed analysis of the various objective evaluation

metrics used in current style transfer research, aiming to offer comprehensive references for researchers.

Structure of this paper is organized as follows. Firstly, it introduces the achievements of traditional style transfer. Secondly, it covers the achievements of the neural style transfer. Thirdly, it presents the evaluation parameters in the style transfer field and compares representative achievements in the field. Fourthly, it discusses the application of style transfer in other fields and non-image style transfer tasks. Finally, it discusses the unresolved issues in the field of style transfer. The main contributions of this paper are as follows:

1. Introducing some achievements in style transfer in the chronological order and providing a subdivision method for neural style transfer.
2. Comparing some representative achievements in the field of style transfer.
3. Providing a comprehensive summary and analysis of objective evaluation metrics in the field of style transfer. There is considerable debate over the choice of objective metrics in style transfer, with different studies employing a wide variety of metrics that vary significantly. To the best of our knowledge, this is the first work to systematically summarize and analyze the objective evaluation metrics used in the majority of recent style transfer studies.
4. Discussing the existing issues in the field of style transfer.

2. Traditional Style Transfer

The term "style transfer" is usually used to refer to the second phase of neural style transfer that began after the publication of Gatys et al.'s paper in 2016 [**?**]. Before this, the term "style transfer" was not widely accepted. The work in the first phase is often referred to as Non-Photorealistic Rendering (NPR) or Image-Based Artistic Rendering (IB-AR).

In this paper, we refer to the work from the first phase (1990s-2016) as traditional style transfer to help readers better understand the continuity between the two approaches. We follow the suggestion from [**?**] and adopt the classification method for traditional style transfer proposed in [**?**]. However, the focus of this paper is not on traditional style transfer, especially since traditional methods have now been largely integrated with neural style

transfer techniques. Therefore, we recommend readers who require a systematic understanding of traditional style transfer methods to refer to these two works[**?** **?**].

When introducing the achievements of traditional style transfer, this paper will proceed by discussing their characteristics, advantages, and limitations.

Stroke-Based Rendering (SBR)is a core algorithm in traditional style transfer that involves covering a 2D canvas with atomic-level rendering primitives to simulate specific artistic styles. These primitives typically include virtual brushstrokes, patches, stippling, and shading marks.

The most common form of SBR is rendering with virtual brushstrokes. The color, direction, size, and order of these strokes may be determined semi-automatically or automatically. The stylized output depends not only on the simulation of the medium used to render each stroke but also on the process of stroke placement and the methods used to set their attributes. The stroke placement process in SBR can be roughly divided into local and global approaches. Local methods typically base stroke placement decisions on the pixels within the spatial neighborhood of the stroke. This can be explicitly specified in the algorithm (e.g., image moments within a window) or implied by previous convolution operations (e.g., Sobel edges). A branch of SBR uses media other than colored pixels or paint to fill image regions, including using small dots (stippling) for tone description, line patterns or curves (shading marks), and mosaic algorithms that combine small tiles. When handling video content, the motion of the strokes should match the movement in the video content. This is particularly emphasized in SBR algorithms to ensure dynamic effects and visual coherence in videos. The advantage of SBR lies in its ability to create effects that closely resemble traditional artworks, making it especially suitable for mimicking styles such as oil painting, watercolor, and sketching. However, SBR methods may struggle when dealing with highly complex or abstract artistic styles, as these styles may not be easily achieved through traditional stroke simulation.

Example-Based Rendering (EBR)is a technique aimed at learning and mimicking specific artistic styles. It does so by analyzing the mapping relationship between an example pair (e.g., an original image and an artist's rendered version of that image) and then applying this mapping to stylize other images. Such methods typically encode a set of heuristic rules to faithfully depict the intended style. They try to capture the essential characteristics of a style by learning and imitating the techniques and styles an

artist applied in a particular work. Once the mapping is learned, it can be used to stylize arbitrary images, making them visually similar to the original example images. This approach not only mimics specific artistic styles but can also replicate the unique style of a particular artist. The advantage of EBR is that it can produce highly personalized results, making it particularly suitable for imitating the style of specific artists or works. However, this method depends heavily on high-quality example pairs, and it may be difficult to achieve the desired stylization effect if suitable training data is lacking.

Image Processing and Filtering (IPF)is based style transfer uses various image processing filters and algorithms to achieve artistic stylization. This includes techniques that are based on image pyramids, as well as the use of interactive techniques (such as human gaze trackers and saliency maps) to explore different levels of an image. Various filtering techniques explore different image processing filters used for artistic stylization, but to date, only a few results have been considered interesting from an artistic perspective. This may be because these filters usually focus on the restoration and enhancement of photorealistic images. The image pyramid and interactive techniques (such as human gaze trackers and saliency maps) explore hierarchical representations by segmenting different resolution versions of the source image. High-level abstraction is achieved by rendering only the coarse large regions or specific areas located at the top of the pyramid. This method helps capture larger shapes and compositional features in the image, rather than detailed textures or lines. Unlike the simplification typically pursued in Image-Based Artistic Rendering (IB-AR), these filtering methods are often associated with the restoration and enhancement of photorealistic images. The advantage of this approach lies in its ability to quickly and easily apply stylization to images, making it suitable for creating diverse and abstract visual effects. However, such methods lack the fine detail and complexity required for artistic stylization, making it difficult to precisely mimic the subtle features of specific artists or styles.

Summary As pioneers in the field of style transfer, these methods have inspired the emergence and development of neural style transfer, but they share some common shortcomings. Each different style transfer method embodies the author's understanding of a particular style. However, because these methods often incorporate the author's interpretation of the style, the transfer results may be correlated with the author's aesthetic level, leading to varying quality in the stylized output. Additionally, the algorithms designed

for the same or similar textures or styles are often similar or even identical, resulting in stylized images with textures that may appear rigid and dull. The limited generalization capability also restricts the broad application of traditional style transfer methods. Traditional methods are designed for specific types of styles and images, and their effectiveness diminishes when generalized to different styles or images.

3. Neural Style Transfer

The technique of using neural networks for style transfer is generally referred to as neural style transfer. It is widely acknowledged that the concept of neural style transfer began to gain traction after Gatys et al.[**?**] utilized convolutional neural networks (CNNs) for style transfer. This paper draws on some of the classification perspectives of neural style transfer discussed in [**?**], and based on the current developments in the field and personal understanding, modifies and adds to these perspectives to form a new classification standard for neural style transfer. The new classification standard is primarily based on three criteria: the developmental stages of neural style transfer, the main issues explored at different stages, and the different approaches to solving these issues.

Under this standard, this paper divides neural style transfer into three developmental stages:

1. Early Stage: Online style transfer based on pixel iteration, led by Gatys et al. [**?**].
2. Developing Stage: Offline style transfer based on model iteration, led by Johnson et al.[**?**] and Ulyanov et al.[**?**].
3. Current Stage: Arbitrary and efficient style transfer.

Dividing neural style transfer into these three stages reflects more effectively the developmental trajectory of the field of neural style transfer, allowing researchers to clearly understand the reasons for the emergence of each stage, the main issues being explored, and the solutions proposed. It is important to note that the chronological classification used in this paper may not always accurately place certain works within a specific stage, as some contributions lie at the intersection of two stages and exhibit characteristics of both. In such cases, this paper classifies them into the earlier stage to highlight their connection to preceding work.

Each of the three stages mentioned above has corresponding subcategories. For online style transfer based on pixel iteration, it can be further categorized into two types based on the type of loss function used: style transfer based on parametric models and style transfer based on non-parametric models. For offline style transfer based on model iteration, it can be classified according to the network and the number of styles that can be transferred by the network, resulting in two subclasses: single-network model generating single style, and single-network model generating multiple styles. For arbitrary and efficient style transfer, based on the different technologies used to achieve style transfer, it can be divided into two categories: those that build their own network frameworks and those that use other emerging technologies as support.

This paper will introduce the contributions of each stage of neural style transfer according to the above classification standard. For each contribution, the paper will first provide a brief introduction to its underlying concept, followed by an analysis of its advantages and disadvantages, leading into the next contribution. The categories of neural image style transfer is illustrated as Figure 1

3.1. Pixel-Iterative Online Style Transfer

Pixel-iterative online style transfer can be traced back to the research of Alexander et al. [**?**] on Convolutional Neural Networks (CNNs). Alexander et al. [**?**] aimed to investigate why CNNs have the ability to extract image features. To this end, they inverted a CNN and used the inverted CNN to reverse-engineer the feature maps extracted by the original CNN, attempting to reconstruct an image similar to the original one. As described by Alexander et al. in their paper [**?**], they were able to generate images with "certain artistic characteristics" through this method, as illustrated in Figure 2.

Through the aforementioned experiment, DeepDream discovered that CNNs could not only extract image features but also generate images. This concept of "using neural networks for feature extraction followed by other methods for image generation" laid a foundational groundwork for the subsequent use of neural networks in style transfer.

Following DeepDream, Gatys et al. [**?**] in 2016 combined deep learning with style transfer in a groundbreaking manner, significantly advancing the effectiveness of style transfer and reaching new heights in visual quality.

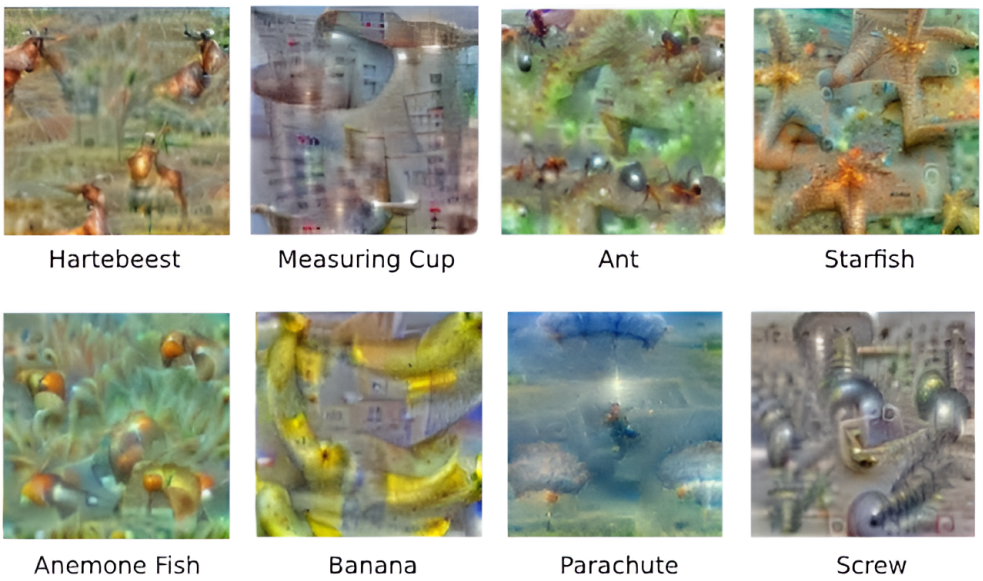

Figure 2: The Artistic Images Generated by Alexander et al.[**?**]

This paper posits that if a style transfer method uses a content image and a style image as inputs, and iteratively processes a noise image at the pixel level during the generation of a stylized image, resulting in the final output of a stylized image, then this approach can be classified as pixel-iterative style transfer. Furthermore, depending on the implementation method, this category can be further subdivided into parametric model-based style transfer and non-parametric model-based style transfer.

3.1.1. Parameterized Model-Based Style Transfer

The work of Gatys et al. [**?**] can be regarded as the first use of a parametric model for style transfer, and it is also considered the first significant achievement in the entire field of neural style transfer.

In their 2016 paper[**?**], Gatys et al. were the first to combine Convolutional Neural Networks (CNNs) with style transfer. The idea behind this achievement originated from the discovery of Gatys et al. during their research on CNNs: a CNN trained with sufficient data can extract features across different datasets[**?**]. In the realm of style transfer, this capability of CNNs to extract features can be used to capture the content features of a photograph as well as the style features of an artwork. Figure 3 illustrates how the VGG16 image classification network, based on CNNs, is capable of

extracting both content and style features at different layers of the network.

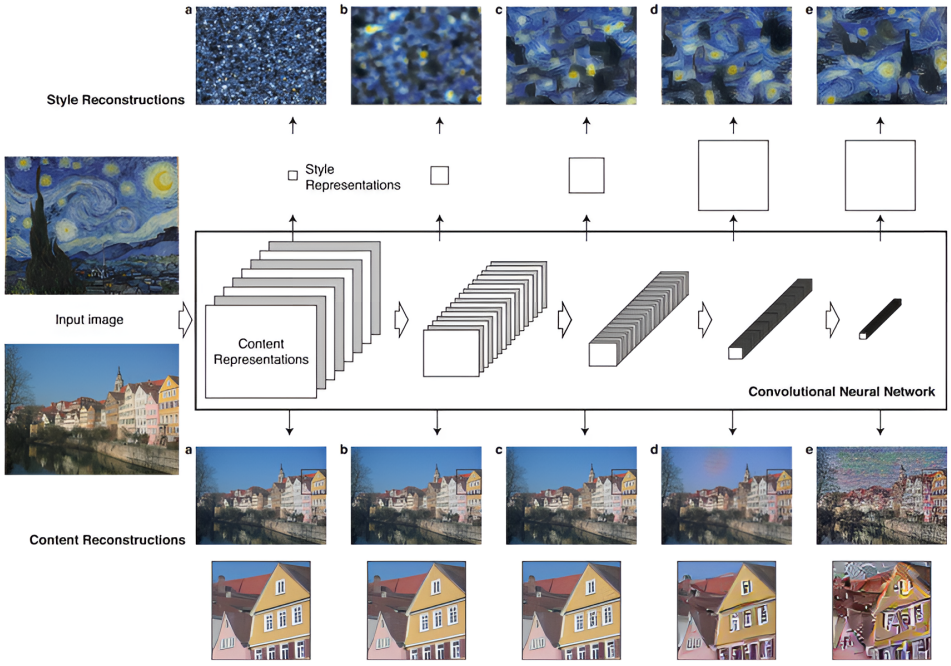

Figure 3: CNN Can Extracting Both Content and Style Features

In terms of practical implementation, Gatys et al.[**?**] achieved style transfer by defining a loss function and using it to optimize a noise image. This loss function is based on the differences between high-level features of the content image and the image being stylized, as well as the differences between high-level features of the style image and the image being stylized. Guided by this loss function, the noise image is iteratively optimized until the loss function reaches its minimum, resulting in the final stylized image. The specific form of the loss function is as follows:

$$L_{total} = \alpha L_{content} + \beta L_{style} \tag{1}$$

where L_{total} is the overall loss function, $L_{content}$ is the content loss, which measures the degree of content difference between the stylized image and the content image, and L_{style} is the style loss, which measures the degree of style difference between the stylized image and the style image. α and β are hyperparameters used to control the similarity between the generated stylized

image and the content image and the style image, respectively. Specifically, the content loss function $L_{content}$ is expressed as follows:

$$L_{content}\left(\vec{p},\vec{x},l\right)=\frac{1}{2}\sum_{i,j}\left(F_{ij}^{l}-P_{ij}^{l}\right)^{2} \tag{2}$$

where $\vec{p}$ is the flattened input of the content image, input into the network in vector form;$\vec{x}$ is the stylized image, which also needs to be flattened for processing. l represents the l-th layer in the network.F^l denotes the set of all feature maps generated by the l-th layer of the network for the image undergoing style transfer, and F_{ij}^{l} refers to the value at position j in the flattened vector of the i-th feature map in the set F^l generated by the VGG [**?**] network at layer l. Similarly, P^l denotes the set of all content feature maps generated by the l-th layer of the network for the input content image, and P_{ij}^{l} refers to the value at position j in the flattened vector of the i-th content feature map in the set P^l. The formula's purpose is to calculate the pixel-wise difference between the feature maps of the stylized image and the corresponding feature maps of the content image, summing these differences. Gradient descent and other optimization methods are used to minimize this value until it stabilizes at a small value.

On the other hand, before providing the specific expression of the style transfer function L_{style} , it is necessary to introduce its core component—the Gram matrix. To extract the style features of the input image, Gatys et al.[**?**] used a feature space designed to capture texture information, which is built on the output of any convolutional layer in the network. This feature space is constructed from the correlations between feature maps of different convolutional layers, with these correlations represented by the inner product of the i-th and j-th vectorized feature maps in the l-th layer. This inner product is known as the Gram matrix, denoted as $G^i \in R^{N_l \times N_l}$. The formula for the Gram matrix is as follows:

$$G_{ij}^{l}=\sum_{k}F_{ik}^{l}F_{jk}^{L} \tag{3}$$

Based on the Gram matrix, Gatys et al.[**?**] processed the noise image using gradient descent and aimed to minimize the mean squared distance between the Gram matrix of the original image and the Gram matrix of the style image. The contribution of the l-th layer in the network to the style

loss function L_{style} is as follows:

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{ij} \left(G_{ij}^l - A_{ij}^l\right)^2 \tag{4}$$

where N_l represents the number of convolutional kernels in a convolutional layer of the network, which is also the number of feature maps that the layer can generate. M_l is the number of pixels contained in each feature map, numerically equal to the product of the height and width of the feature map.A_{ij}^l and G_{ij}^l represent the pixel values at position j in the i-th vectorized feature map at layer l of the network. Formula 5 only computes the contribution of a single layer to the style loss L_{style} . The overall style loss L_{style} is obtained by summing the weighted losses from all layers, with the specific formula as follows:

$$L_{style}\left(\vec{a}, \vec{x}\right) = \sum_{l=0}^{L} \omega_l E_l \tag{5}$$

where w_l is the weight representing the impact of each layer on the total style loss function, and L is the number of convolutional layers in the network. By taking the partial derivative of Formula 1, $\frac{\partial L_{total}}{\partial \vec{x}}$ is obtained, which can be used as input for optimization algorithms and guide the iterative processing of the stylized image to achieve style transfer.

The parameterized model style transfer method based on the Gram matrix introduced by Gatys et al. has distinct advantages[**?**]. Compared to traditional style transfer methods, it overcomes the limitations of rigid and less variable brushstrokes and achieves excellent results. It also addresses the limitation of traditional methods being restricted to specific styles, enabling the transfer of natural and stylized textures[**?**], thus achieving good transfer results. However, this method has some notable drawbacks. Since style transfer starts from a noise image each time, it requires a significant amount of time for batch processing and performs poorly in real-time style transfer. Additionally, the Gram matrix is more adept at capturing global information of feature maps, leading to unsatisfactory results in extracting regular textures with long-range symmetrical structures[**?**]. Moreover, unlike traditional methods that ignore high-level semantic information, Gatys et al.[**?**] considered only high-level semantic information and neglected low-level semantic information, resulting in deficiencies in the fine structure and detail coherence of the synthesized stylized images.

To address the shortcomings of fine texture handling in Gatys et al.'s method, Berger et al.[**?**] proposed horizontal and vertical pixel differences, focusing on examining the differences between each pixel and other pixels in its horizontal and vertical directions. In practice, this method calculates the feature relationships between the pixel located at position (i, j) and pixels at positions $(i, j + \delta)$ or $(i + \delta, j)$ in the feature map, incorporating these into the style loss consideration. This approach allowed Berger et al.[**?**] to achieve effective transfer of symmetric long texture patterns, somewhat compensating for Gatys et al.'s limitations in simulating fine structures and regular textures with long-range symmetry. However, since the method still relies on the Gram matrix as the core of style transfer, it retains some of the shortcomings of Gatys et al., such as inadequate texture detail simulation and unstable style transfer quality.

To investigate the reasons behind the style instability in images generated using the Gram matrix, Risser et al.[**?**] conducted a further in-depth study of Gatys et al.[**?**]. Risser et al. proposed that the instability in generated image styles is due to feature maps with different means and variances potentially having the same Gram matrix. Based on this finding, Risser et al. incorporated histograms of feature maps into the style transfer loss function, further optimizing the Gram matrix-based style transfer method. The advantage of this approach is the generation of more stable style images; however, incorporating histograms of feature maps makes the computation more complex, increasing the time required for style transfer and reducing efficiency in batch processing. Moreover, Risser et al. only addressed the stability of generated images, leaving the issue of detailed texture description unresolved.

Li et al.[**?**] attempted to explain the principles of neural style transfer using more mature domain knowledge, considering neural style transfer as a special variant of domain adaptation tasks. From this perspective, Li et al. provided a different viewpoint. Domain adaptation tasks are based on the fact that the distributions of source and target data are different, with the goal of training a model on a labeled source domain dataset to predict the target domain data distribution. One method of domain adaptation is to minimize the distribution difference between source and target domain samples, where Maximum Mean Discrepancy (MMD) is a commonly used measure. Analogously, in style transfer tasks, the content image can be considered as the source domain, and the stylize image as the target domain. Li et al. explored the mathematical role of the Gram matrix in style transfer

and proved that the process of matching Gram matrices of style images and stylized images (i.e., Formula 4) is essentially equivalent to minimizing an MMD with a quadratic polynomial kernel. Therefore, Li et al. proposed that minimizing MMD with other kernels (such as linear, polynomial, or Gaussian kernels) might be useful in style transfer. The main contribution of Li et al. is that they provided a theoretical exploration and clarification of the principles of Gatys et al.[**?**].

The aforementioned parameterized model-based neural style transfer methods have some common defects. Due to the loss of low-level information in convolutional neural networks, the transfer results for objects with regular shapes (e.g., man-made objects) often exhibit noticeable distortions. Although Berger et al.[**?**] made contributions in this area, there is still significant room for improvement. Additionally, the style transfer process is time-consuming, requiring substantial time for batch processing of many images, and thus is less efficient in tasks with large numbers of images, such as video style transfer. The generated style is also unstable and lacks temporal consistency.

3.1.2. Non-Parametric Model-Based Style Transfer

Markov Random Fields (MRFs) are a primary method used in non-parametric model-based neural style transfer[**? ? ? ?**]. The use of MRFs for texture synthesis is also a common technique in traditional style transfer methods. Li and Wand [**?**] chose to integrate the traditional MRF-based method into neural style transfer. In their work [**?**], they combined MRFs with deep convolutional neural networks (dCNNs) and proposed a non-parametric neural style transfer method. They argued that the parameterized style transfer method using Gram matrices only considers differences between pixels and lacks spatial-level constraints on the stylized image, leading to insufficient realism in the generated images. Therefore, Li and Wand replaced the Gram matrix with an MRF regularizer and introduced a new loss function:

$$L_s = \sum_{l \in l_s} \sum_{i=1}^{m} || \Psi_i(F^l(I)) - \Psi_{NN(i)}(F^l(I_s)) ||^2 \tag{6}$$

where $\Psi(F^l(I))$ represents the set of all local patches in the feature map$F^(I)$; $\Psi_i(F^l(I))$ refers to the i-th local patch within this set; $\Psi_{NN(i)}$ is the style patch in the stylized image I that most closely matches the i-th local patch

in terms of style similarity. This best-matching stylized patch $\Psi_{NN(i)}$ can be found by calculating the normalized correlation between all style patches in the style image. Here, m denotes the total number of local patches. The method proposed by Li and Wand improved upon the style transfer results achieved by Gatys et al.[**?**], making the structures in the stylized images more coherent and better preserving the fine details of the original image, leading to significant progress in synthesizing realistic photos. However, this method lacks attention to the semantic content of the images. If there is a substantial structural difference between the content and style images, the matching between local patches and style patches might not be accurate, resulting in poor quality of the generated stylized images.

Summary The pixel-iteration-based style transfer methods can be considered as early explorations in the field of neural style transfer, holding a pioneering position. Neural style transfer technology overcomes the limitation of traditional style transfer techniques, which only transfers a single style. It has better generalization ability, achieving a "once-for-all" approach to some extent, meaning that once a network structure is built, it can transfer any style and handle images of any resolution. Additionally, thanks to the CNN's feature extraction capabilities, neural style transfer can capture advanced artistic styles and textures, exhibiting flexible and diverse texture features within the same stylized image. However, despite the groundbreaking achievements of pixel-iteration-based style transfer, some common issues cannot be ignored, with the most critical being resource consumption and time expenditure. Pixel-iteration-based style transfer requires multiple iterations of a noise image under the guidance of a loss function, with each iteration involving feature extraction and computation by a CNN. As a result, a significant amount of computational resources and time is consumed during the image generation stage. For instance, when transferring the style of a 512×512 image to generate a similarly sized stylized image, the process can take up to 51.19 seconds[**?**]. If one attempts to transfer a high-resolution (e.g., 4K) image, the time required becomes even more prohibitive for practical applications. Moreover, even after investing substantial time and computational resources in style transfer, it is often difficult to consistently achieve results that surpass those of traditional style transfer methods. These two major drawbacks have hindered the widespread application of neural style transfer technology in real-world scenarios, leading to a stagnation in the research of neural style transfer.

3.2. Model-Iteration-Based Offline Style Transfer

As previously mentioned, although pixel-iteration-based style transfer methods have the ability to perform arbitrary style transfers, their low efficiency prevent them from being widely used in real-life scenarios, especially when dealing with tasks involving a large number of images. This drawback is particularly evident in such cases. To improve the efficiency of neural style transfer, researchers have developed fast-style transfer methods that trade off either the quality of style transfer or the number of transferable styles. These methods are referred to as model-iteration-based style transfer.

Generally, the primary approach to achieving model-iteration-based style transfer is to shift the time required during inference to the training phase. Specifically, this is typically done by training a specific style transfer network, which stores style information in the form of parameters within the network. When confronted with different styles, the network can quickly retrieve the corresponding parameters, thereby enabling fast style transfer.

From a historical perspective, model-iteration-based style transfer can be divided into two stages: the first stage involves a single model generating a single style, and the second involves a single model generating multiple styles. To help readers accurately classify the methods used in the literature, this paper provides detailed descriptions of these two categories: if a style transfer method takes a content image as input and the network can quickly convert it into a specific stylized image, the method can be classified as Per Model Per Style transfer. On the other hand, if a style transfer method takes a content image and the encoded representation of the desired final style as input, and the network can quickly convert it into a stylized image corresponding to the style code, the method can be classified as Per Model Multi Style transfer.

3.2.1. Per Model Per Style

Using feedforward networks to achieve Per Model Per Style. Johnson et al.[**?**] and Ulyanov et al.[**?**] were the first to publish work on real-time style transfer. Independently, in 2016, they both proposed methods for real-time style transfer using feedforward neural networks. The main idea was to train a feedforward neural network that encodes style information in the network's parameters, allowing the generation of stylized images without the need for multiple iterations of optimization on a noise image. Specifically,

the following feedforward neural network needs to be trained:

$$\begin{aligned} \theta^* &= \arg\min L_{total}\left(I_c, I_s, g_\theta * (I_c)\right), \\ I^* &= g_\theta * (I_c), \end{aligned} \tag{7}$$

where θ^* represents the optimal parameters that minimize the total loss function L_{total}.

The total loss function $L_{total}\left(I_c, I_s, g_\theta * (I_c)\right)$ evaluates the quality of the style transfer. This function has three inputs: the content image I_c, the style image I_s, and the stylized image $g_\theta * (I_c)$. It produces a single output, I^*, which is the final stylized image. The goal of this function is to find the parameters θ^* that minimize L_{total}. These parameters are used to transform the content image l_c into the final stylized image I^*.

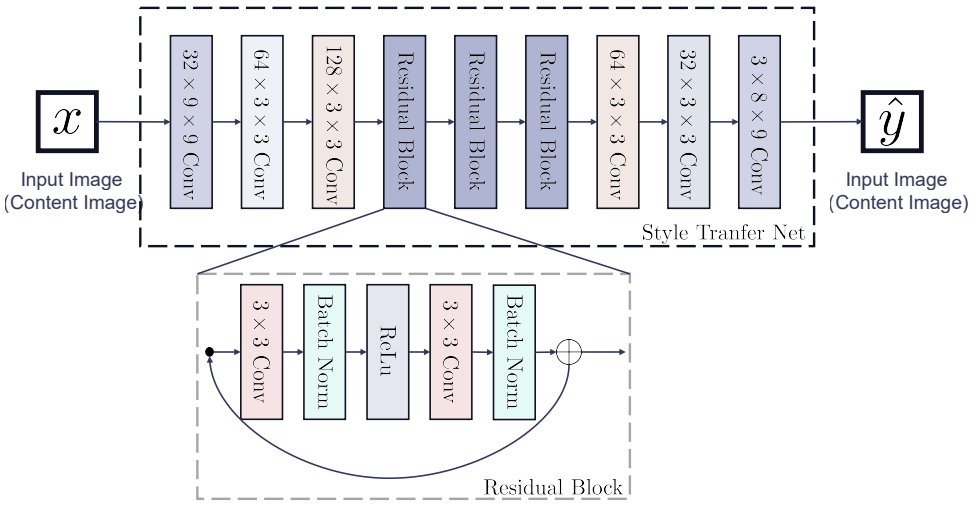

Figure 4: Johnson et al.'s Network Architecture [**?**]

Although Johnson et al. and Ulyanov et al. proposed this method simultaneously, there are differences in their network structures. Johnson[**?**] et al. built upon the method by Radford et al.[**?**], adding residual blocks and step-wise convolution, and introduced an Instance Normalization (IN) layer to accelerate the network's convergence, as shown in Figure 4.

On the other hand, Ulyanov et al.[**?**] used a multi-scale structure as the generative network (as shown in Figure 5), with an objectiveKwo function similar to that of Gatys et al.[**?**] Both Johnson et al. and Ulyanov et al. based their methods on feedforward generative networks, achieving real-time

style transfer. Compared to the method by Gatys et al.[**?**], the speed of style transfer was improved by two orders of magnitude.

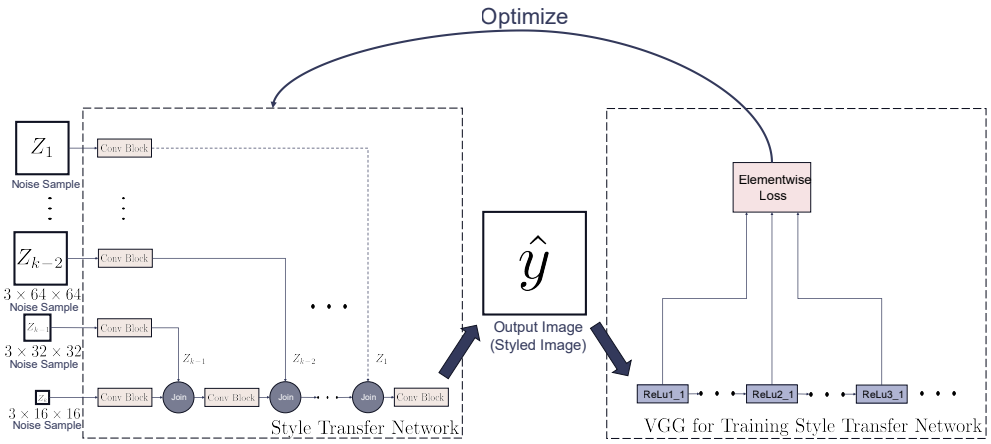

Figure 5: Ulyanov et al.'s Network Architecture[**?**]

However, since both methods during training largely follow the approach proposed by Gatys et al., they encounter similar issues in migration effects, such as unsatisfactory results in image detail and structural consistency.

Using Markov Random Fields for Per Model Per Style. The use of Markov Random Fields can also enhance the speed of style transfer. Li and Wand et al.[**?**] improved their previous work[**?**] by obtaining a Markov feedforward network through adversarial training (Figure 5), thereby addressing the efficiency problem. In principle, this method is similar to their earlier non-parametric style transfer approach based on image patches described in[**?**]. This improvement allows them to achieve better results in object structure consistency, thereby preserving the fine details of the original image more effectively. However, some shortcomings of the original methods remain, such as poor performance when the matching between local patches and style patches is not precise.

Using Variational Autoencoder for Per Model Per Style.

Zhang et al.[**?**] proposed the Caster method to address the limitation of existing cartoon style transfer methods, which struggle to simultaneously learn both specific cartoon styles and domain-specific style information. Caster is based on a variational autoencoder and incorporates a Cartoon Style Projection Module (CSCM) that encodes the style information of a specific cartoon image into a latent code, which is then dynamically projected onto

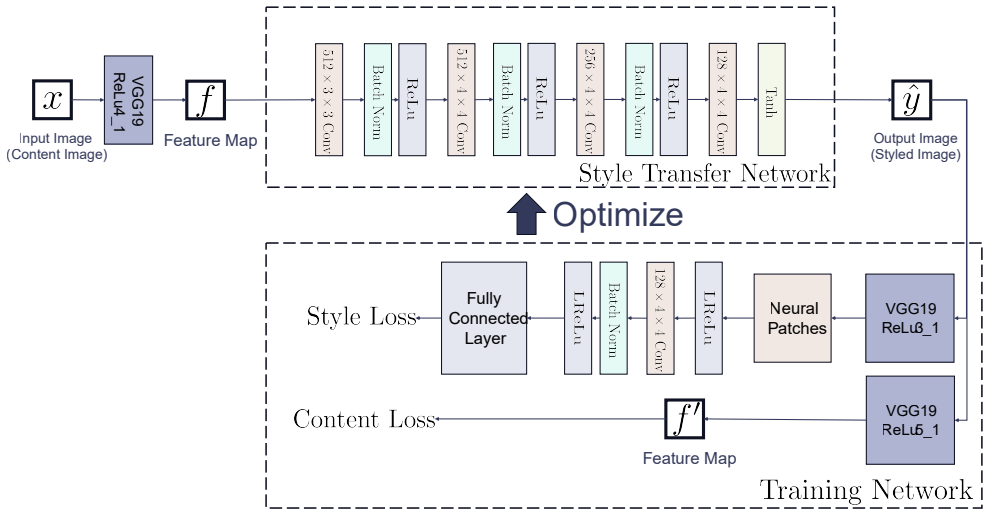

Figure 6: Li et al.'s Network Architecture[**?**]

the content features for more precise style transfer.

Additionally, Caster introduces a Cartoon Contrastive Learning Loss (CCLS), which enhances the quality and color consistency of the style transfer by pulling stylized images with the same style closer together and pushing those with different styles further apart. This further improves the effectiveness of style migration.

Using Generative Adversarial Networks for Per Model Per Style. Using Generative Adversarial Networks (GANs) for style transfer represents another approach to Per Model Per Style transfer.

GANs were introduced by Goodfellow et al. in 2014[**?**]. This model consists of a generator network and a discriminator network that engage in adversarial training. During training, the discriminator network is first trained: given a set of data, the discriminator evaluates whether each data item belongs to the target domain. Then the discriminator is optimized based on the discrepancy between the true values and the network's judgments until it can accurately classify data items in the test set. Subsequently, the generator is trained to produce data, which is then assessed by the discriminator to determine if it belongs to the target domain. The generator adjusts itself based on the feedback from the discriminator. The generator and discriminator continuously compete to balance each other, leading to the training of the network and achieving a similarity in distribution between the

generated data and the real data. The loss function for GANs is as follows:

$$V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log D(x)\right] + \mathbb{E}_{z\sim p_{\text{z}}(z)}\left[\log\left(1 - D(G(z))\right)\right] \tag{8}$$

In this context, G and D represent the generator network and the discriminator network, respectively. $p_{\text{data}}(x)$ denotes the distribution of the data x, $p_{\text{z}}(z)$ denotes the distribution of noise images, $D(x)$ represents the probability that x belongs to the real data rather than the data generated by the generator p_g, and $G(z)$ denotes the generator's output for the noise image z. E represents the expectation. The discriminator is trained to maximize the probability that it can distinguish between images from the dataset and those generated by G. The generator's training aims to minimize $\log\left(1 - D(G(z))\right)$, where $1 - D(G(z))$ represents the probability that the discriminator believes the generated image does not belong to the real data set. Minimizing this function deceives the discriminator into thinking that the generated data comes from the real data set. The excellent network structure of GANs makes them suitable for style transfer tasks.

However, GANs face several issues and drawbacks when applied to real-time style transfer. Firstly, training GANs is often challenging and can be unstable, leading to problems such as mode collapse. Secondly, the loss function of GAN models does not include hyperparameters like α and β from Gatys et al.'s[**?**] loss function, making it difficult to control the similarity between the generated image and the content or style images. Therefore, precise control in style transfer tasks is challenging. While GANs can be used for style transfer tasks, they are not specifically designed for these tasks.

Zhu et al.[**?**] modified GANs[**?**] to better adapt to real-time style transfer tasks. In their work[**?**], they proposed CycleGAN, which achieves unsupervised Per Model Per Style style transfer. CycleGAN's uniqueness lies in its introduction of "cycle consistency loss," which ensures bidirectional transformations by simultaneously training two generators and two discriminators, thus preserving information through conversions from one domain to another and back. Similar to GANs, CycleGAN[**?**]'s generator maps images from one domain to another, while the discriminator tries to distinguish between generated and real images (Figure 7).

One of the advantages of CycleGAN is its ability to handle unpaired data, enabling unsupervised learning. This method can learn how to perform cross-domain image translation without requiring explicit pairings for each sample

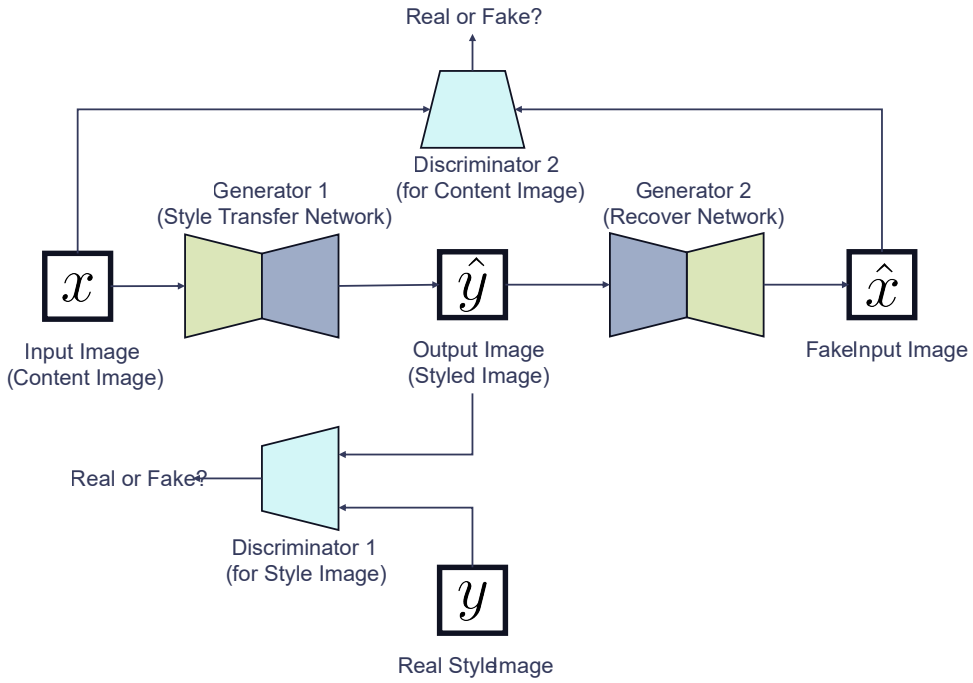

Figure 7: Workflow of CycleGAN [**?**]

during training. However, CycleGAN also has some notable drawbacks in the field of style transfer. The images it generates can sometimes appear blurrier or distorted than real images. Additionally, improper selection of hyperparameters may reduce the stability of the training process or result in poor image quality.

StyleGAN[**?**] is another outstanding achievement in using GANs for style transfer, particularly known for its portrait editing capabilities, which can generate portraits with specific features and high realism. StyleGAN controls various details and the overall style of the generated image by adjusting the "style" parameters in the input latent space. The core innovation of StyleGAN is the introduction of a new style transfer mechanism, which allows for the separate control of image content and style at different levels, thereby improving the diversity and controllability of generated images while maintaining image quality. The advantages of StyleGAN include its ability to generate high-resolution, high-quality images and provide powerful image editing capabilities through style control. It performs exceptionally well in generating faces and other complex images. However, the training process is resource-intensive and time-consuming, requiring significant computational resources. Additionally, the generated images may sometimes contain unpredictable artifacts, necessitating further optimization to improve stability and reliability.

Men et al.[**?**] focused on portrait style transfer, aiming to transform portrait photographs into anime-style avatars. They observed that when using CycleGAN for style transfer involving significant geometric deformations, it is challenging to generate high-quality transfer results while preserving the correct structure. To address this issue, the authors proposed an unsupervised cartoon image generation method based on a Gated Cycle Mapping Network (GCMN). The core of this approach is the introduction of a Gated Style Encoder (Egs), which generates category-specific style codes through domain and group-specific style layers and injects these codes into the generator to achieve precise style control.

Traditional CycleGAN models enforce bidirectional mapping, requiring multiple generators, which perform poorly when dealing with complex geometric structural changes, especially in the conversion from portraits to anime, where high-quality outcomes cannot be assured. To mitigate this limitation, Men et al. designed a gated style encoder that combines gated mapping units (GMUs) with domain-specific and group-specific layers to generate category-specific style codes, thereby achieving precise control over the

transformation process. The domain-specific layers distinguish whether the image originates from the photo domain or the cartoon domain, while the group-specific layers further differentiate between portraits and scenes, ensuring the rationality and accuracy of the style transfer. Additionally, the authors introduced an Adaptive Instance Normalization (AdaIN) mechanism to inject the style codes into the generator via a multi-layer perceptron (MLP), dynamically modulating the stylistic expression within the generated results.

In contrast to traditional multi-generator or multi-encoder methods, the Gated Cycle Mapping Network requires only a single generator to handle image translation tasks in all directions, significantly simplifying the model architecture. Experimental results demonstrated that the gated style encoder not only enhanced the quality of the generated cartoon images but also allowed for flexible migration according to user-specified style requirements, thus achieving higher controllability. With these improvements, the authors' model surpassed existing state-of-the-art methods in terms of visual effects and style control and also exhibited superior performance in video synthesis tasks.Per Model Per Style transfer can mitigate the primary drawbacks of long transfer time and low efficiency of pixel-iteration-based style transfer by shifting the time required for the transfer phase to the training phase. However, the trade-off is that it can only transfer a specific, single style. If multiple styles are to be transferred using this method, corresponding network training is required for each style, resulting in a large number of parameters. In such cases, the large number of parameters makes it difficult to deploy on devices with limited resources (e.g., smartwatches), thus necessitating a balance between the number of styles and the performance constraints.

3.2.2. Per Model Multi Style

Although the Per Model Per Style methods addressed the issue of real-time stylization and improved the efficiency of style transfer, each model can only correspond to a specific style. Transferring to a new style requires significant time for training a new model, and it is challenging to apply these models to the devices with limited resources. To address this, the networks capable of transferring multiple styles while maintaining the ability to perform fast transfers, known as Per Model Multi Style networks, are developed. One of the main approaches to achieving multi-style transfer is to bind a specific style to a small portion of the network parameters to reduce redundant parameters and combine this with the addition of necessary

parameters that reflect style differences in the networks.

Dumoulin et al.[**?**] were the first to bind specific styles to parameters within the network, thereby achieving Per Model Multi Style transfer. In reducing redundant parameters, they noted that certain parts of the computation in the transfer of different styles are similar or identical, as many artistic styles with different names share similar or identical brushstrokes (for example, Impressionist paintings have similar brushstrokes, differing mainly in the colors used). It seems wasteful to treat these paintings with similar brushstrokes as entirely different styles. Traditional one-to-one style transfer models overlooked this, leading to unnecessary time wastage during the training phase when transferring to a new style. In their experiments, Dumoulin et al.[**?**] found that scaling or transforming the normalized parameters is sufficient to adapt to a particular style. For a convolutional neural network, this finding implies that the parameters of all convolutional kernels can be adjusted to facilitate the transfer of different styles. Specifically, style transfer can be achieved by simply adjusting certain parameters of the convolutional kernels after normalization. In implementation, Dumoulin et al.[**?**] built upon the method of Ulyanov et al. [**?**], applying an affine transformation after instance normalization to complete the transfer of different styles. This process is referred to as Conditional Instance Normalization (CIN), which can be represented by the following formula:

$$\mathrm{CIN}\left(\mathcal{F}(I_c, s)\right) = \gamma^s \left(\frac{\mathcal{F}(I_c) - \mu(\mathcal{F}(I_c))}{\sigma(\mathcal{F}(I_c))}\right) + \beta^s \tag{9}$$

The input to the formula consists of the content image I_c and the style index s, with $F(x)$ representing the feature map of image x, and $\mu(x)$ and $\sigma(x)$ representing the mean and standard deviation of the image, respectively. This method achieved the transfer of different styles by scaling or transforming the parameters γ^s and β^s, meaning that each style s can be realized by adjusting the parameters of the affine transformation.

The advantage of this approach is that by using affine transformations on parameters of similar styles, Dumoulin et al. were able to reduce redundant parameters, significantly lowering the number of required parameters compared to Per Model Per Style methods when generating the same number of styles. However, when dealing with images with significant style differences, it is still necessary to retrain the network to obtain the corresponding style parameters and adjustment methods. Therefore, as the number of transferable styles increases, the network parameters in this method also increase.

Chen et al.[?] achieved parameter simplification through a different method. Unlike Dumoulin et al.[?], who used similar parameters to represent similar styles, Chen et al. [?] considered content image processing, proposing that the parts of the network that process content information could be the same. This led to the concept of decoupling the content processing module and the style processing module within the network. By using this decoupling approach, with independent network modules learning the content and style information of an image, the network becomes more flexible in handling style transfer tasks. This method employs mid-level convolutional filters (referred to as the "StyleBank" layer in[?], as shown in Figure 8) specifically designed to learn different styles. The "StyleBank" layer contains multiple sets of parameters, with each set associated with a specific style.

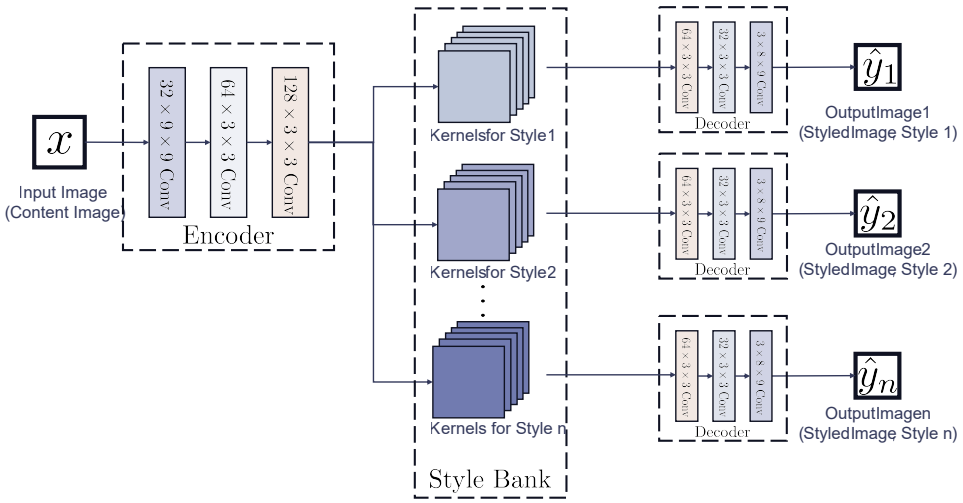

Figure 8: Network Architecture of Chen et al.[?]

The other parts of the network, aside from the "StyleBank" layer, are used to learn content information. Since the content processing module handles different styles in the same way, different styled images can use the same content processing module, thereby improving the network's efficiency in processing various styles. When implementing multi-style transfer, only incremental training is needed. Specifically, when a new style needs to be added, the part of the network responsible for learning content information can be fixed, and only the "StyleBank" layer for the new style is trained. This approach allows the network to effectively learn new styles without affecting

the existing learned ones. In the StyleBank, one or more sets of convolutional kernels represent a specific style. By placing different sets of convolutional kernels into the neural network, it can perform style transfers for different styles, providing good scalability.

Although this approach allows a single network to transfer multiple styles and optimizes the number of parameters in the network, challenges still remain. In terms of parameters, if the method attempts to transfer multiple styles, the number of parameters will gradually increase with the number of transferable styles, making it impossible for this approach to handle arbitrary style transfers. Regarding the quality of generated images, due to the partial parameter sharing, the transfer quality might not reach the excellent results of style transfer methods based on model iteration.

3.3. Arbitrary and Real-Time Style Transfer

The development of style transfer can be summarized in a single phrase: a trade-off between the number of styles, the quality of the transfer, the transfer time, and the resources required. However, in practical applications, users may demand both diverse style transfer and fast processing speeds. Consequently, methods that can perform arbitrary style transfer with high quality and efficiency have emerged.

Currently, mainstream methods for achieving arbitrary and real-time style transfer can be divided into three categories based on whether they use other auxiliary technologies:

1. Using image spatial features or parameter matching for style transfer.
2. Utilizing technologies such as GANs[**?**], attention mechanisms, diffusion models, pre-trained large models, and others as auxiliary tools.
3. Focusing on in-depth research into style transfer, by exploring new processes and network structures to achieve arbitrary and real-time style transfer tasks.

Given the abundance of GAN-based methods, they are separated from the second category. Based on the above descriptions, this section will introduce the achievements in arbitrary and real-time style transfer across the following five categories:

1. Style transfer based on spatial features or parameter matching.
2. Arbitrary and real-time style transfer based on GANs[**?**].
3. Arbitrary and real-time style transfer based on attention mechanism.

4. Arbitrary and real-time style transfer based on pretrained models.
5. Arbitrary and real-time style transfer using self-built networks.

The method of style transfer based on spatial features or parameter matching emerged earlier and, although the quality of the transfer effects is not ideal, it serves as an introduction to the pioneers of arbitrary and real-time style transfer. The development of the latter three categories has progressed in parallel, and the latest advancements in the field of style transfer can generally be classified into these three categories.

3.3.1. Style Transfer Based on Spatial Features or Parameter Matching

The style transfer methods based on spatial features or parameter matching focus on the features in the spatial domain of images, attempting to achieve arbitrary style transfer by adjusting the spatial feature parameters or matching corresponding regions between the content image and the style image.

The first work to achieve arbitrary and simultaneous style transfer was proposed by Huang et al. in 2017[**?**]. Inspired by the CIN method[**?**], Huang et al. introduced the Adaptive Instance Normalization layer (AdaIN). This layer is used to match the variance and mean of content features with the mean and variance of style features, thereby achieving arbitrary style transfer through this matching of variance and mean. The formula for AdaIN can be described as follows:

$$\mathrm{AdaIN}(\mathcal{F}(I_c), \mathcal{F}(I_s)) = \sigma(\mathcal{F}(I_s)) \left(\frac{\mathcal{F}(I_c) - \mu(\mathcal{F}(I_c))}{\sigma(\mathcal{F}(I_c))} \right) + \mu(\mathcal{F}(I_s)) \quad (10)$$

Unlike the method in [**?**], the encoder in Huang et al.'s style transfer network is fixed, containing the first few layers of a pre-trained VGG network. The overall structure of their network is shown in Figure 9.

Considering the common issues in style transfer when using normalization techniques as in previous works [**?** **?**], Jing et al.[**?**] modified AdaIN[**?**] and proposed the Dynamic Instance Normalization layer (DIN). The main advantage of DIN is that it allows for arbitrary style transfer by aligning the mean and variance (the simplest statistical data) between content and style features, without the need for manually defining a formula for calculating affine parameters. Instead, it introduces a more general dynamic convolution transformation, where the parameters adaptively change according to

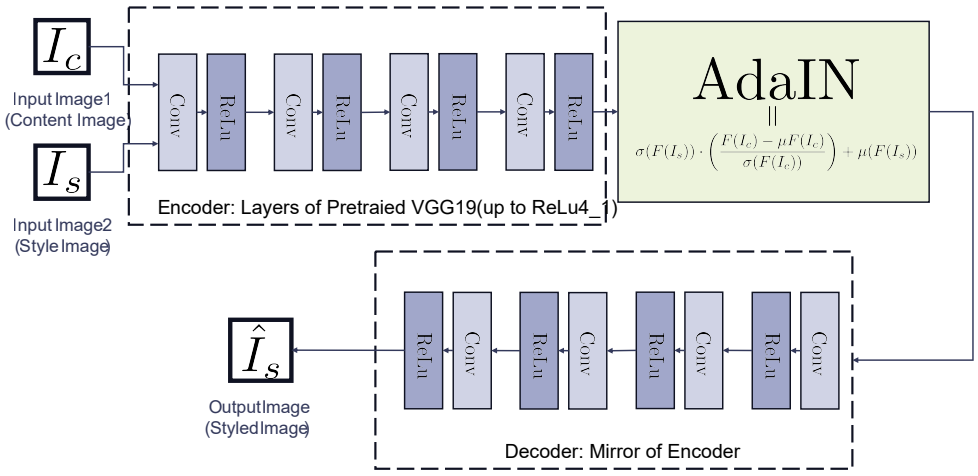

Figure 9: Network Architecture of Huang et al.[**?**]

different styles in a learnable manner, thereby more accurately aligning the complex statistical data of real style features. Given a pair of content image I_c and style image I_s as inputs, the proposed DIN layer can be represented by the following formula:

$$\begin{aligned} \mathrm{DIN}(\mathcal{F}_c, \mathcal{F}_s) &= f[\mathcal{F}_s, \mathrm{IN}(\mathcal{F}_c)], \\ \mathrm{IN}(\mathcal{F}_c) &= \frac{\mathcal{F}_c - \mu(\mathcal{F}_c)}{\sigma(\mathcal{F}_c)} \end{aligned} \tag{11}$$

where $\mathcal{F}_c$ and $\mathcal{F}_s$ are the feature representations of $\mathcal{I}_c$ and $\mathcal{I}_s$, respectively, and f is the dynamic convolution operation[**?**]. Unlike standard convolution, where weights and biases are model parameters, the weights and biases in the dynamic convolution f within DIN are dynamically generated by encoding different input style images. This dynamic adjustment mechanism makes DIN more flexible and precise in capturing and applying fine style features, thereby enriching the expression of style details while preserving the content structure.

The work by Chen et al.[**?**] is another early effort in arbitrary style transfer called Style Swap. Unlike the methods[**? ? ?**] that use image spatial feature parameter matching for style transfer, this approach focuses on the differences between the content image and the style image. The method first divides the images into multiple patches and then exchanges the most similar feature patches between the content image and the style image to

achieve real-time and arbitrary style transfer. The specific steps of Step Swap can be list as follows:

1. Extract a set of image patches from the content feature map and the style feature map, denoted as the content feature patch set$\{\phi_i(C)\}_{i\in n_c}$ and the style feature patch set $\{\phi_i(S)\}_{j\in n_s}$, where n_c and n_s represent the number of content feature patches and style feature patches, respectively. These patches should be sampled from all feature maps and should have sufficient overlap;
2. For each content feature patch, determine the closest matching style feature patch based on the normalized cross-correlation metric (Equation 12);
3. Swap each content feature patch $\phi_i(C)$ with its closest matching style image patch $\phi^{ss}(C,S)$;
4. For overlapping areas, if different values are obtained in step 3, average them.

$$\phi_i^{ss}(C,S) := \underset{\phi_j(S),j=1,\ldots,n_s}{\operatorname{argmax}} \frac{\langle \phi_i(C), \phi_j(S)\rangle}{\|\phi_i(C)\| \cdot \|\phi_j(S)\|} \tag{12}$$

Through the above method, the stylized feature map can be obtained. By upsampling and reconstructing this stylized feature map, the final stylized image can be generated. Although this method achieves arbitrary and real-time style transfer with better image quality than the previously mentioned methods based on spatial feature parameter matching[**?** **?** **?**], it is less efficient in terms of time and resource consumption compared to those methods. Additionally, since the method uses image patches as the unit for style transfer, it overlooks global style information and lacks accuracy in measuring the similarity between adjacent image patches.

3.3.2. Arbitrary and Real-Time Style Transfer Based on GAN

Given the advantages of GANs, such as powerful unsupervised learning capabilities, feature learning abilities, and data generalization, researchers have considered further adjusting GANs to enhance their style transfer capabilities as style transfer advances to the arbitrary and real-time stage.

To the best of our knowledge, Xu et al.[**?**] were the first to use GANs to achieve arbitrary and real-time style transfer. They proposed the Dynamic Residual Block Generative Adversarial Network (DRB-GAN) for style transfer. The concept of Dynamic Residual Blocks (DRBs) was inspired by DIN and StyleGAN, modeling the "style code" as shared parameters of dynamic

convolution and AdaINs within the dynamic residual blocks. Multiple DRBs were designed at the bottleneck of the style transfer network. Each DRB consists of a convolutional layer, a dynamic convolutional layer, a ReLU layer, an AdaIN layer, and an instance normalization layer with residual connections. This structure enables the adjustment of the shared parameters of dynamic convolution and adaptively modifies the affine parameters of AdaINs, ensuring statistical matching between the bottleneck feature spaces of the content and style images. The use of dynamic residual blocks is motivated by their ability to provide flexible parameter adjustments, which better achieve statistical feature matching between style and content. This design allows the network to effectively blend various style features while maintaining the basic structure of the content image.

Yang et al.[**?**], building on StyleGAN[**?**], observed that StyleGAN is only capable of fast transfer for specific styles but cannot perform real-time transfer of arbitrary styles or generate truly artistic portraits. To address these challenges, Yang et al. proposed DualStyleGAN to achieve sample-based portrait style transfer. DualStyleGAN retains an internal style path from StyleGAN to control the style of the original domain while adding an external style path to model and control the style of the target extended domain. Additionally, the external style path inherits StyleGAN's layered architecture, modulating structural style in the coarse resolution layers and color style in the fine resolution layers, enabling flexible multi-level style operations. Although DualStyleGAN can achieve flexible portrait style transfer, the network often misidentifies other objects in the background of the content image as faces, resulting in the generation of undesired patterns.

Wu et al.[**?**] approached the style transfer problem from a different angle. They argued that even within the same artist's body of work, there can be a wide variety of artistic styles, so homogenizing different styles from the same artist is not a rigorous approach. Furthermore, existing methods often lack generalization for unseen artists. To address these challenges, Wu et al. proposed a Double-Style Transferring Module (DSTM). This module extracts different artistic styles from various artworks and retains the intrinsic diversity among different artworks by the same artist. Recognizing that learning style from a single artwork could lead to overfitting and the introduction of style image structural features, Wu et al. further introduced an Edge Enhancing Module (EEM). This module extracts edge information from multi-scale and multi-level features to enhance structural consistency. The advantage of this approach is that it can result in stylized artworks where

the content features of the content image are well-preserved, with minimal intrusion of content features from the style image.

3.3.3. Arbitrary and Real-Time Style Transfer Based on Attention Mechanisms

The utilization of attention mechanisms in deep neural networks has gradually become a common choice for style transfer tasks. Previous achievements in style transfer often placed more emphasis on the overall style of the image, neglecting the correspondence of local styles. By incorporating attention mechanisms, which are adept at capturing the spatial layout and semantic relationships within images, style transfer models can effectively identify key regions within an image and apply style features precisely to these areas. This enables the models to achieve a balance between global and local styles to a certain extent.

Furthermore, attention mechanisms permit the exchange of features across different images, which is particularly critical for achieving nuanced style transfer effects. This capability enhances the model's ability to selectively focus on relevant parts of the input data, thereby improving the quality and coherence of the transferred style across the entire image.

Wu et al.[**?**] observed that the multi-head attention mechanism of Transformers is capable of capturing long-range dependencies, a feature that provides Transformers with a significant advantage in capturing global style information and content semantics. Based on this observation, they proposed StyleFormer. The network architecture of StyleFormer can be described as an Encoder-Transformer-based Style Transfer Core-Decoder. In addition to the VGG16-based Encoder and Decoder, the core of the network is refined into three modules: the Style Bank Generation Module, the Transformer-driven Style Composition Module, and the Parametric Content Modulation Module.

The Style Bank Generation Module takes the encoded results of the style image as input and represents them as a finite number of style codes. Each style code consists of a style key and its corresponding style value. The style key is used to represent the style features, while the corresponding style value is an affine transformation matrix, which adjusts the content image based on the style features.

The Transformer-driven Style Composition Module dynamically combines the style codes based on content features. It achieves global consistency between style and content through the multi-head attention mechanism. In

this module, the style key and style value are used as the Transformer's Key and Value, while the content feature is treated as the Query. The multi-head attention mechanism learns the global relationships between style and content, generating content-consistent affine coefficients for different groups. This ensures that the style transfer results preserve the delicate expression of style features globally, while maintaining semantic structure and content consistency.

The Parametric Content Modulation Module applies the affine coefficients generated by the Transformer-driven Style Composition Module to the content features, completing the generation of stylized features. The Decoder then takes these stylized features as input and decodes them into the final stylized image.

Liu et al.[**?**] sought to leverage the characteristics of the aforementioned attention mechanisms to enhance the local quality within the stylized results. They observed that existing solutions either focus solely on integrating deep style features into deep content features without considering feature distributions or adaptively normalize deep content features based on style to match global statistics[**?** **?** **?**]. While effective, these methods concentrate only on deep and global image features, overlooking shallow and local features, which can lead to local distortions. To address this issue, Liu et al. proposed a novel attention and normalization module called Adaptive Attention Normalization (AdaAttN), which performs attention normalization adaptively on a per-pixel basis. The process of AdaAttN involves three steps: 1. Computing attention maps with content and style features from shallow to deep layers; 2. Calculating weighted mean and standard deviation maps of the style features; 3. Adaptively normalizing content features to align per-pixel feature distributions. The output of AdaAttN is then processed by a decoder to complete the style transfer. This method utilizes attention mechanisms to consider local information matching, achieving more detailed and personalized style transformations, thus enhancing the local visual quality of the image.

In contrast to Liu et al.'s approach, which addresses the neglect of local features, Deng et al.[**?**] argue that CNNs have limited receptive fields and can only perceive local information, making it challenging to extract and maintain global information from the input image. To address this issue, Deng et al.[**?**] proposed a Transformer-based method named StyTr2 that considers the long-range dependencies of input images. StyTr2 comprises two distinct Transformer encoders that generate domain-specific sequences

for content and style. After encoding, a multi-layer Transformer decoder is used to stylize the content sequence according to the style sequence. StyTr2 mitigated the content leakage issue inherent in CNN-based models, achieving better style transfer results. However, due to the large parameters of Transformers, the approach by Deng et al.[**?**] is somewhat less efficient in terms of runtime compared to previous methods.

Li et al.[**?**] also noted the difficulty of CNN-based methods in capturing long-range information and, considering the high computational cost of Transformer-based methods, they attempted to balance between long-range information and a large number of parameters while partially addressing the remaining content leakage issue. To resolve the computational and leakage problems, Li et al. designed a compact Transformer named AdaFormer. This design utilizes image patch projection and positional encoding to enhance global interactions and assumes that the differences between content and style features are captured in the higher layers of the encoder. AdaFormer achieves more efficient feature extraction by sharing parameters in the initial Transformer encoding layer and then extracting content and style features through their respective encoding layers. The algorithm also employs a multi-layer Transformer decoder for feature fusion and selects style elements through dynamic weighting. Additionally, adaptive instance normalization (AdaIN)[**?**] is used instead of layer normalization to make the style more consistent with the content. Finally, the upsampling decoder generates diverse stylized outputs. Compared to StyTr2[**?**], Li et al.'s method does reduce memory usage but still requires approximately 35GB of memory[**?**].

Huang et al.[**?**] argued that current style transfer methods often neglect the fidelity of the generated stylized images. To address this, they proposed QuantArt, a vector quantization-based image style transfer approach designed to improve the visual fidelity of the generated images, making them closer to real works of art. In QuantArt, vector quantization and Style-Guided Attention (SGA) are two key techniques that together enable high-visual fidelity image style transfer.

Vector quantization is a technique that transforms continuous feature representations into discrete encodings. It works by learning a codebook that divides the feature space into multiple clusters, mapping each feature to the nearest cluster center, thus achieving feature quantization. SGA is an attention-based style transfer module that learns the relationship between style references and content features, transferring style information onto the content features.

Vector quantization brings the feature representations of the generated image closer to the cluster centers of real art distributions, enhancing visual fidelity. Meanwhile, SGA effectively transfers style information onto content features, ensuring more efficient style transfer. By combining both techniques, QuantArt achieves high-visual fidelity in image style transfer.

Tang et al.[**?**] pointed out that existing style transfer methods typically require an additional style image as a reference, which limits their flexibility and convenience. To address this issue, they proposed Master, a meta-learning-based Transformer model capable of controllable zero-shot and few-shot artistic style transfer.

The core idea of Master is parameter sharing, where different Transformer layers share the same set of parameters. This not only significantly reduces the number of model parameters but also facilitates faster convergence during training. Additionally, Master allows for convenient control over the degree of style transfer by adjusting the number of Transformer layers used during inference.

Master also employs learnable scaling and shifting operations instead of standard residual connections, which helps preserve the structural similarity of the content while transferring style patterns, thereby reducing content distortion. The model is trained using a meta-learning algorithm, enabling it to quickly adapt to new styles with minimal adjustments, requiring only fine-tuning of the Transformer encoder layers' parameters.

Liu et al.[**?**] addressed the lack of flexibility and personalization in existing style transfer methods by proposing the Any-to-Any Style Transfer method. This approach allows users to interactively select specific regions from the style image and apply their corresponding styles to the same regions in the content image, thereby achieving personalized style transfer effects.

The interactive core of this method is the Segment Anything Model (SAM). Users can perform interactive segmentation on both the content and style images using SAM, selecting image regions through clicks, bounding boxes, or drawing contours. The model then converts the selected region information into control signals and merges them with a default attention map to generate a final attention map. Finally, the model computes the stylized features based on the generated attention map and decodes them into an image, producing the final stylized result.

Zhang et al.[**?**] integrated attention mechanisms into the classic Adaptive Instance Normalization (AdaIN)[**?**] method to improve upon it. They identified several shortcomings in current AdaIN-based[**?**] style transfer

works, including mismatches between content and style, image artifacts, and inaccuracies in the extraction of style features during the generation of high-quality stylized images.Zhang et al.[?] analyzed two main causes for these issues: (1) the use of the VGG network for feature extraction, and (2) the reliance on statistical parameter adjustments alone during instance normalization as a means of style transfer. For the first issue, Zhang et al.[?] noted that the VGG network, originally designed for image classification, tends to focus excessively on irrelevant classification information during style transfer. To resolve this, they proposed a Transformer-based style feature extractor called the Perception Encoder (PE). The PE captures long-range dependencies and high-frequency style details in the style image, thereby avoiding the limitations of the VGG network, which focuses predominantly on prominent classification features such as edges or shapes, leading to more accurate extraction of style information.

To address the second issue, they introduced Style Consistency Instance Normalization (SCIN). Unlike AdaIN, which achieves style transfer through simple alignment based on mean and variance, SCIN uses a Transformer to capture long-range, non-local dependencies within the style feature maps, providing richer style information. Additionally, the scaling and shifting parameters generated by SCIN are learned, allowing better adaptation to the distribution of different style images rather than relying solely on fixed statistical features like mean and variance. This improvement makes SCIN more flexible and accurate in aligning style and content features, reducing artifacts and enhancing the quality of stylized images.

To further improve the quality of style transfer results and increase the distinctiveness of stylized images of different styles, the paper also proposed Instance-based Contractive Learning (ICL). ICL helps the model learn the relationships between stylized images, ensuring that embeddings of images with the same content or style are closer together, while those of different styles are further apart, thereby enhancing the quality of the stylized images.

Through the implementation of these three approaches, the paper addresses the limitations of AdaIN[?], which considers only the unification of global features, reducing artifacts and ultimately improving the quality of stylized images from both a global and local perspective.

Wang et al.[?] took a different approach, asserting that the textures produced by current style transfer methods are unpredictable and do not align with artistic creation logic. They emphasized the importance of interactive participation in the style transfer process. To generate predictable

textures, Wang et al. proposed an Interactive Image Style Transfer Network (IIST-Net). This network produces stylized results of brushstrokes guided by doodle curves, ensuring that the style distribution of the stylized image is closer to real-world artworks. Specifically, IIST-Net consists of two encoders, an Interactive Brush-texture Generation (IBG) module, a Multilayer Style Attention (MSA) module, and a decoder. The two encoders encode content style images and brush texture images to produce multi-layer style features and fused content features, respectively. The IBG module generates controllable brush textures based on user input, while the MSA module further refines multi-layer style features and integrates them with the fused content features. Although this method partially addresses the interactivity gap in neural style transfer, it overly emphasizes texture, making it challenging to highlight content features.

Zhang et al.[**?**] recognized the importance of balancing content and style features. They noted that excessive style patterns in image stylization could obscure content details, sometimes making it difficult to distinguish objects in the image. To address the balance between content and style, Zhang et al. proposed a Transformer-based style transfer method called STT (Style Transfer via Transformers). In this method, Transformers are primarily used for encoding content and style features and for decoding. To maintain the clarity of the content structure in the stylized image, Zhang et al. designed a novel edge loss to enhance object edges in the output image. Unlike edge detection or contour extraction tasks, the content details in style transfer outputs may differ from those in the content image, especially as the background may adopt artistic patterns from the style image. Therefore, using the similarity between edge maps of the content and stylized images directly as an optimization objective could result in blurry outcomes. One of the challenges is filtering out edges not present in the content image's main structure. Zhang et al. introduced a masking operation to address this issue: edges in the stylized image's edge map that do not correspond to positions in the content image's edge map are masked out. Additionally, Zhang et al. set a threshold to exclude weak responses in the edge map to prevent potential noise. Zhang et al.'s method achieves a balance between style and content patterns in the stylized image, resulting in a better visual experience, and somewhat alleviates the content leakage problem. However, similar to other methods using Transformers, it requires more time and resources compared to arbitrary and real-time style transfer methods that do not use Transformers.

Hong et al. [54] expressed concerns about mismatched training data in

style transfer. They argued that the low semantic correspondence between arbitrary content and style images causes attention mechanisms to focus on limited regions of the style image. This impedes attention-based methods from accurately capturing and expressing the entire style of the reference image, leading to discordant patterns. To overcome these limitations, Hong et al. focused on enhancing the attention mechanism and capturing the rhythm of textures in the style image. They designed a pattern repetition rate to measure how well different style features represent the entire style image. By selecting the style with the highest pattern repetition rate as the primary style for stylization, they effectively avoided generating discordant patterns in the output image.

Zhu et al.[**?**], almost simultaneously with Hong et al.[**?**], observed the issue of repeated style features in style images, but achieved the selection of primary style features through a different approach than Hong[**?**]. Zhu et al.'s solution is a novel All-to-Key (A2K) mechanism, which matches each query with stable keys. A2K consists of two main components. First, the distributed attention mechanism (Figure 10). To address the issues of All-to-All attention, distributed attention initially learns distributed key points to describe local regions of style features. Each query of content features is then matched with these representative key points. Since the matched key points represent regional styles rather than isolated locations, distributed attention can tolerate matching errors better. Moreover, because these key points represent several local regions, the matching performance of distributed attention is more stable across different query locations. Second, the progressive attention mechanism. Unlike traditional All-to-All attention, which directly focuses on specific locations, the progressive attention mechanism initially attends to coarse-grained regions and then gradually focuses on fine-grained positions. This approach helps match style patterns at a larger scale, finding more similar semantics within coarse-grained style patterns. On this basis, point-to-point attention further refines fine-grained positions within coarse-grained regions. Additionally, since queries within the same local region match the same key points, the transformed features also exhibit regional stability.

Through the innovative attention mechanisms described, Zhu et al. similarly improved upon the issues caused by full attention mechanisms, avoiding discordant patterns in the generated images. However, their exploration in enhancing style expressiveness[**?**] remains somewhat similar to previous works in terms of expressiveness.

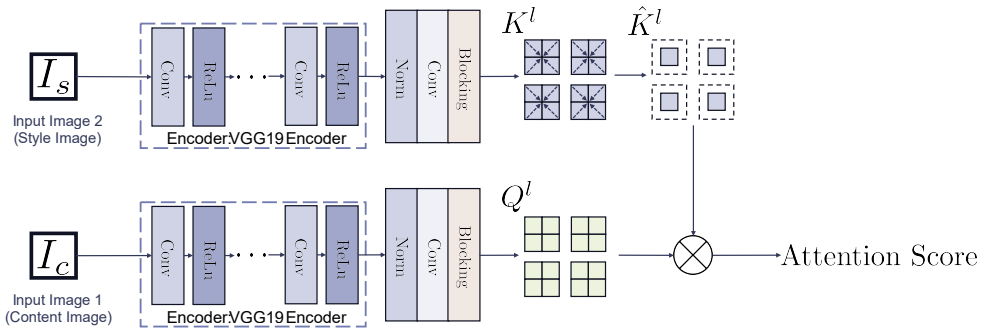

Figure 10: Zhu et al.'s Distributed Attention[**?**]

Zhang et al.[**?**] observed that when Transformer models based on window attention mechanisms are used for style transfer tasks, they tend to generate grid-like patterns, which lead to poor style transfer results. They attributed this issue to the powerful local capabilities of the window attention mechanism, which lacks long-range information extraction capabilities. To address this problem, they proposed an image style transfer method called S2WAT, which introduces a Strips Window Attention (SpW Attention) mechanism and an Attn Merge technique to improve the network's long-range information extraction capabilities.

Specifically, S2WAT adopts a hierarchical architecture similar to the Swin Transformer, progressively reducing the image resolution and extracting features at different scales. This enables the model to effectively capture both local and global information of the image. The SpW Attention is the core module of S2WAT, which combines different types of window attention, including horizontal strip-shaped windows, vertical strip-shaped windows, and square windows. Horizontal and vertical strip-shaped window attention is used to capture non-local features, while square window attention focuses on capturing local features. This design strikes a balance between modeling short-range and long-range dependencies, thus avoiding the locality issue inherent in traditional window attention.

The Attn Merge technique is used to fuse the results from different window attention outputs. It computes the spatial correlation between the input features and the outputs of the various window attention mechanisms, and then performs a weighted sum based on the correlation scores. This dynamic weighting of the attention outputs helps determine the importance of different window attention mechanisms, thereby enhancing the style transfer

performance.

3.3.4. Arbitrary and Real-Time Style Transfer Based on Pretrained Models

Pretrained models are deep learning models that have been trained on extensive datasets, possessing broad knowledge and substantial learning capability. These models can be fine-tuned for specific tasks to improve performance and efficiency. Recently, numerous studies have sought to leverage large models to assist in style transfer tasks. Here we focus on the achievements in style transfer using large models.

CLIP. Utilizing the CLIP[**?**] large model for style transfer assistance has become a current research hotspot. Kwon et al.[**?**] observed that in many practical situations, users may lack reference style images but still wish to experience the results of style transfer. To address such applications, Kwon proposed a new framework aimed at transferring the semantic style of target text to content images using a pretrained CLIP model. During the process of obtaining semantically transformed images solely through CLIP supervision, Kwon found that traditional pixel optimization methods could not reflect the desired texture. To resolve this issue, Kwon et al. introduced a CNN encoder-decoder model to capture hierarchical visual features of the content image while simultaneously adding style in the deep feature space to achieve realistic stylization results. The advantage of Kwon et al.'s method lies in achieving realistic style transfer results through changes in text conditions alone, without requiring any style images. However, due to the use of large models, this method involves longer inference time compared to other methods.

Liu et al.[**?**] pointed out that existing style transfer methods typically require an additional style image as a reference, which limits their flexibility and convenience. To address this issue, they proposed TxST, a text-based artistic style transfer model. TxST utilizes an image-text encoder (such as CLIP) to understand the style described in textual descriptions and applies it to the content image. Through a contrastive training strategy and cross-attention mechanisms, TxST is able to learn the styles of specific artists, such as Picasso, oil painting, or sketches, and transfer them onto the target image. This text-driven style transfer approach eliminates the need for an extra style image, providing users with a more flexible and convenient personalized style customization experience.

Koley et al.[**?**] attempted to combine hand-drawn sketches, text guidance, and style transfer in their proposed image retrieval framework named

"You'll Never Walk Alone." This framework enables precise generation of fine-grained images by integrating hand-drawn sketches and textual descriptions. The authors argue that traditional methods typically rely on either sketches or textual descriptions, but they posit that sketches and text have a complementary role in fine-grained representation.

By converting sketches into pseudo-token representations and combining them with textual descriptions, the framework effectively captures fine-grained image features such as color, texture, and contextual information. To address the issue of lacking fine-grained textual descriptions during training, the authors introduce a novel compositional constraint. This constraint leverages the discrepancy between sketches and photographs to simulate the missing textual descriptions. Additionally, to enhance the model's generalization capability, the authors introduce a text-to-text generalization loss, which ensures that the learned prompt vectors closely resemble actual textual prompts. This combination of sketch and text guidance improves the precision and flexibility of image generation, particularly in fine-grained tasks.

Yu et al.[**?**] combined object detection with style transfer and proposed an image style transfer method called SA2-CS, which achieves significant object style transfer based on textual descriptions. Notably, this method does not require a style image as input; instead, it uses a text description to specify the desired style. The approach distinguishes between the foreground and background of the content image, applying different degrees of style transfer to these regions.

Specifically, the authors use the U2-Net salient object detection network to separate the foreground and background of the image. They then apply different levels of style transfer to each region, thereby preventing overstylization of the foreground and enhancing the contrast between the foreground and background. To further prevent the style transfer of the foreground from interfering with the background, they introduce a global background loss function. This loss function utilizes masking to avoid damaging the foreground content during the background style transfer process, while also enhancing the style transfer effect for the background.

To address the limitations of traditional global CLIP loss, the authors propose a semantic-aware PatchCLIP loss. This loss function employs several mechanisms, such as semantic-aware random cropping, semantic-aware weight penalties, semantic-aware threshold regularization, and semantic-aware adjustable image patches, to precisely control the degree of style transfer in different semantic regions. These innovations enable more fine-grained and

targeted style transfer for specific parts of the image.

Diffusion Models. Diffusion models are generative models that were first detailed with mathematical proofs, derivations, and runnable code by Ho et al.[**?**]. These models can generate target data samples from noise and consist of two processes: the forward process and the reverse process, where the forward process is also known as the diffusion process. The forward process is a noising process, where an image x_t only depends on the previous $x_(t-1)$. This process can be viewed as a Markov process:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})q(x_t|x_{t-1}) = N(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{13}$$

whereβ_t is a set of predefined hyperparameters, meeting the demand of $\beta_1 < \beta_2 < ... < \beta_T$. The reverse process is a denoising process; if the reverse process $q(x_{t-1}|x_t)$ is obtained, an image can be progressively restored from random noise x_T.

Hamazaspyan and Navasardyan[**?**] combined diffusion models with style transfer tasks, proposing a Diffusion-Enhanced PatchMatch (DEPM) model. This model utilizes Stable Diffusion to capture high-level style features while preserving fine-grained texture details of the original image. DEPM allows for the transfer of arbitrary styles during inference without any fine-tuning or pre-training, making the process more flexible and efficient.

Zhang et al.[**?**] considered style transfer from pretrained Stable Diffusion models[**?**]. They noted that methods based on small models can preserve content structure but fail to generate highly realistic stylized images, introducing artifacts and discordant textures. Conversely, pretrained model methods can generate highly realistic stylized images but struggle to maintain content structure. To address these issues, Zhang et al. proposed ArtBank, which can generate highly realistic stylized images while preserving the content structure of the content image. Specifically, to fully exploit the knowledge within pretrained models, they designed an implicit style prompt library consisting of a set of trainable parameter matrices to learn and store knowledge from an art collection, serving as visual prompts to guide the pretrained model in generating highly realistic stylized images while maintaining content structure. During the training phase, Zhang et al. introduced a novel spatial-statistics-based self-attention module to accelerate the convergence of the implicit style prompt library. By training the implicit style

prompt library, Zhang et al. effectively extracted relevant knowledge from Stable Diffusion (Ver 1.4) to accomplish style transfer. Zhang et al.'s method pioneered a new approach to style transfer, focusing on how to quickly and effectively extract the necessary knowledge from pretrained models.

Building on diffusion models, Zhang et al.[**?**] proposed a novel approach to style transfer, focusing on learning implicit textual labels of various artistic styles as the core of style transfer. The primary concept of this method is to treat the style in artwork as a learnable textual description of a painting and to guide the Diffusion model in generating images based on style labels. In terms of implementation, Zhang et al. proposed Inversion-Based Style Transfer Method (InST) and employed it to efficiently and accurately learn image-related information. Specifically, a conditional generative model is used to learn the correspondence between images and text, thereby obtaining image embeddings. Based on these image embeddings, an attention-guided inversion module receives the embeddings and utilizes an attention mechanism to generate corresponding text embeddings. This module focuses on various features in the image embeddings, such as semantics, texture, object shape, brushstrokes, and color, ultimately resulting in the corresponding text embeddings and text labels. These text labels guide the Diffusion model in style transfer. The text labels need not be describable in natural language but are a sequence of characters or a token (token) that only the Diffusion model can interpret to describe the style. Once learned, the Diffusion model can fix the style corresponding to this label, making this method a model-based iterative approach to style transfer. The distinctive feature of this approach is its ability to alter the shape of the image during the stylization process, which was not possible with previous style transfer models.

Similarly, Namhyuk et al.[**?**] observed that style information is difficult to describe accurately in language; therefore, they considered encoding style images into the text space to provide textual constraints for Stable Diffusion. Specifically, Namhyuk et al.[**?**] combined style transfer with text-to-image generation tasks based on Stable Diffusion, proposing the DreamStyler framework. This framework is capable of extracting style information from images into the CLIP text space. To integrate textual descriptions with Stable Diffusion, the authors proposed the concept of extended text embedding space based on Textual Inversion (TI). This idea involves dividing the time steps of the diffusion model into multiple groups, each referred to as a Chunk of Timesteps. A combination of a Chunk of Timesteps with corresponding textual descriptions is termed a TI Stage. By combining multiple TI Stages,

Stable Diffusion can understand similar yet distinct style description embeddings at different time step chunks during image synthesis. This method is referred to by the authors as Multi-Stage Textual Inversion. Additionally, Namhyuk et al.[**?**] introduced context-aware prompt enhancement, which can decouple style and contextual information from style images. After decoupling, the style information can be encoded as special text embeddings, providing more accurate style descriptors for use with Multi-Stage Textual Inversion.

Wang et al.[**?**] developed a framework called StyleDiffusion, which achieves the separation of image style and content. This framework is based on the Diffusion model and uses the diffusion process to separately remove style information and content information from the image. It then utilizes a CLIP-based style disentanglement loss, coordinated with style reconstruction priors, to achieve complete disentanglement of content and style. The framework comprises three key components: a diffusion-based style removal module, a diffusion-based style transfer module, and a CLIP-based style disentanglement loss coordinated with style reconstruction priors. Experiments demonstrate that this framework can produce high-quality style transfer results and better considers the relationship between content and style compared to other methods. In contrast to previous methods, this approach completely decouples content (C) and style (S) through the diffusion model, thereby providing a more nuanced understanding of their relationship. Consequently, the style transfer results are more natural and harmonious, especially for challenging styles such as Cubism and oil painting. Wang et al.'s method achieves precise control over the style transfer process through the diffusion model and CLIP-based style disentanglement loss. By adjusting parameters, one can flexibly control the degree of style removal and the extent of content-style disentanglement, leading to more desirable style transfer outcomes. Moreover, this method has high interpretability and scalability. The introduction of the diffusion model and CLIP-based style disentanglement loss makes the style transfer process more interpretable. Additionally, this method can be applied to other image transformation or manipulation tasks, demonstrating significant scalability.

Lu et al.[**?**] attempted to address the challenge of fine-tuning a pre-trained diffusion model with a minimal number of image samples to learn any unseen style. They proposed a method called "Specialist Diffusion," which enables the learning of any unseen style by fine-tuning a pre-trained diffusion model with only a small number of images (e.g., fewer than 10). This method

allows the fine-tuned diffusion model to generate high-quality images of arbitrary objects in a specified style. To achieve such low-sample fine-tuning, Lu et al. introduced a novel set of fine-tuning techniques, including custom data augmentation for text-to-image tasks, content loss to facilitate the disentanglement of content and style, and sparse updates focusing on only a few time steps.The "Specialist Diffusion" method can seamlessly integrate with existing diffusion models and other personalization techniques, achieving superior fine-tuning performance compared to state-of-the-art few-shot personalized diffusion models, particularly in learning highly complex styles. Moreover, "Specialist Diffusion" can be combined with inversion methods to further enhance performance, even achieving success with very unusual styles. However, the method does have certain limitations. Firstly, while it effectively fine-tunes with a small number of images to learn unseen styles, there may still be cases where learning is insufficient for highly specific and unusual styles. Secondly, although the method demonstrates good sample efficiency, the generated results may be suboptimal for some complex or content close to the training data distribution. Lastly, the performance of this method is constrained by the quality of the pre-trained model; if the pre-trained model is of poor quality, it may adversely affect the results of the fine-tuning.

Chung et al.[**?**] aimed to address the issue of excessive inference times when using diffusion models for style transfer tasks. Although there were already training-free approaches available, previous research had not effectively applied these methods to large diffusion models such as Stable Diffusion. By reviewing existing literature, Chung et al. identified two key characteristics of using large diffusion models for image translation: first, attention maps determine the spatial layout of the generated images; second, adjusting the queries and keys in the cross-attention mechanism can influence the content of the generated images.

Based on these findings, Chung et al. proposed a strategy for style transfer without retraining the diffusion model. The core idea of this strategy is to replace the queries and keys in the self-attention maps of the content image with those derived from the style image's self-attention maps. When implementing this core idea, Chung et al. encountered two primary issues: content disruption and color errors. To address content disruption, they introduced a "query preservation" strategy. For color errors, they employed Initial Latent Adaptive Instance Normalization (Initial Latent AdaIN) technology.

By combining these three elements, their method achieved a training-free style transfer approach based on large diffusion models. The advantage of

this method lies in its ability to maintain the relationship between queries in the content image if they share similar semantics, as they will utilize similar keys after style transfer. Moreover, the high similarity between queries of the content image and keys with similar textures and semantics in the transfer results leads to a more natural and harmonious effect in the style transfer outcome.

Unlike Namhyuk et al. [**?**], who encoded style information into the text space, Deng et al.[**?**] believed that using an encoder to convert style images into text features to guide the Stable Diffusion model could lead to significant losses because such textual feature descriptions are imprecise, resulting in outputs that often fail to capture the detailed style features of the images adequately. They pointed out that fundamental diffusion models can directly extract style information without textual constraints, given that the commonly used U-Net architecture in such models possesses this capability, thereby achieving the goal of avoiding dependence on text embeddings.

Based on this insight, Deng et al. proposed a dual-path denoising model based on Stable Diffusion to achieve style transfer. This model consists of two independent yet identical diffusion models, one handling content images and the other handling style images, both employing a U-Net as the core network. For ease of description, these two diffusion models are referred to as the Style Diffusion Model and the Content Diffusion Model. The Style Diffusion Model aims to reconstruct the original style image progressively through a T-step diffusion process from a noise image. Similarly, the Content Diffusion Model aims to reconstruct the content image.

When the diffusion models operate at any time step $t \in [0, T]$, the U-Net extracts content feature maps X_t^c from the content diffusion path and style feature maps X_t^s from the style diffusion path. These feature maps are then combined through a cross-attention mechanism to generate a stylized latent feature map $\hat{f}_c$. Subsequently, these latent feature maps pass through a reverse diffusion process to produce the final stylized image. Compared to traditional methods that compute style features based on Gram matrices, this approach can better preserve the structure and details of the content image.

The integration mechanism described herein is termed "Cross-Attention Reconstruction," with the central idea being to treat different pixels within the content image as Queries, which are correlated with the features (Keys) of the style image. Given that the relevance between content pixels and style information may vary, these differences can influence the effect of style

transfer. Some content pixels might contribute disproportionately to the style information, leading to regions within the stylized result that appear unnatural or compromise the fidelity of the content. Therefore, the authors propose applying a weighted suppression to these less relevant content pixels to diminish their impact on the stylized latent feature map $\hat{f}_c$.

However, due to the properties of the Softmax function, pixels with lower relevance (i.e., those with smaller QK^T values) might end up receiving relatively larger attention weights after the Softmax operation, leading to suboptimal style transfer outcomes. To address this issue, we introduce a reweighted cross-attention mechanism that incorporates an adjustable weighting parameter λ into the Softmax function, dynamically modulating the attention weights of different pixels. This approach not only mitigates the influence of low-relevance pixels but also effectively enhances the representation of pixels strongly correlated with the style image.

Based on the aforementioned methodology, Deng et al. have realized a novel style transfer method grounded in diffusion models that does not require textual embeddings. This method ensures that the stylized results retain the details of the content while accurately reflecting the stylistic characteristics.

Chen et al.[**?**] approached style transfer from a text-based perspective and proposed a framework called ArtAdapter for Text-to-Image (T2I) style transfer. The core of this framework lies in the combination of a multi-layer style encoder and an explicit adaptive mechanism, which improves the fidelity of style transfer while ensuring consistency with the textual description.

The multi-layer style encoder extracts features from different layers of a pre-trained VGG network (low, medium, high levels), and each of these features is encoded using MLPs to generate rich style embeddings. This process is akin to "describing" the style reference using pseudo-words. The explicit adaptive mechanism primarily operates within the cross-attention layers of a diffusion model, adapting the style encoding while maintaining the integrity of the text encoding. This ensures that the model accurately captures the style details while preserving the generalization ability of the text. Furthermore, during training, an Auxiliary Content Adapter (ACA) provides weak content guidance, helping to separate the content structure and style features of the style reference. This prevents the generated image from being dominated by the content semantics of the style reference, allowing for more precise style transfer.

3.3.5. Arbitrary and Real-Time Style Transfer Based on Self-Built Networks

Frequency Domain-Based Approach. Frequency domain processing holds significant importance in the field of computer imaging. It involves transforming images from the spatial domain to the frequency domain to analyze and modify the frequency characteristics of images. This approach is instrumental in achieving various critical tasks such as filtering, compression, and feature extraction. Frequency domain processing effectively reduces noise, enhances image quality, detects edges and texture details, and provides powerful tools for applications such as image compression and recognition, offering a wealth of technical means for research and application in the field of image processing.

The approach proposed by Li et al.[**?**] differs from most previous methods [**? ? ? ? ? ?**] that focus on the spatial domain, as it considers the content and style features of images from the frequency domain perspective. Li et al. argue that effective disentanglement of content and style is crucial for synthesizing images in arbitrary styles. They note that existing methods tend to focus on disentangling content and style feature representations in the spatial domain, where content and style features are inherently entangled. This entanglement leads to issues such as local distortions, weaker generalization capability, and inflexibility in previous methods. Li et al.'s approach is based on the observation that images or feature maps can be transformed into the frequency domain, where the low-frequency components describe smooth variations, and the high-frequency components are associated with rapid changes[**?**]. This characteristic is referred to by them as the Frequency Separable Property (FSP).

Based on the Frequency Separable Property, Li et al. proposed the FreMixer module, which is capable of disentangling and re-entangling the frequency spectra of content and style components in the frequency domain. Since content and style components exhibit distinct frequency domain characteristics (such as frequency bands and patterns), FreMixer effectively decomposes these two components. The procedure is as follows: Firstly, a two-dimensional fast Fourier transform (2-D FFT) is performed along the spatial dimensions to convert spatial feature maps into the frequency domain; next, frequency kernels learned through the network are introduced, serving a role similar to that of global depth convolutional layers, to disentangle the content and style frequency parts in the spectral map; thirdly, after disentangling the content and style frequency patterns, the two frequency

spectra are recombined through element-wise addition; and finally, the frequency spectra are converted back into spatial domain stylized features using a two-dimensional inverse fast Fourier transform (2-D IFFT). By following these steps, content and style can be separated in the frequency domain and then recombined in the spatial domain to achieve image style transfer. Li et al.'s work[**?**], starting from the frequency domain, achieves the decoupling of style and content features, pointing out a new direction for neural style transfer.

Wen et al.[**?**] proposed a universal style transfer framework named CAP-VSTNet, designed to address the shortcomings of existing methods in terms of fidelity and consistency. The core of CAP-VSTNet consists of an invertible residual network and an unbiased linear transformation module, trained with the addition of a masked Laplacian loss.

The invertible residual network effectively preserves content affinity and, through a channel refinement module, prevents the accumulation of redundant information, leading to improved style transfer results. The masked Laplacian loss is employed to mitigate the loss of pixel affinity caused by the linear transformation, ensuring the clarity and consistency of the generated image.

Kwon et al.[**?**], although approaching the problem from a different perspective than Li et al.[**?**], reached similar conclusions. Kwon posits that current style transfer methods struggle to effectively transfer aesthetically pleasing artistic information and face high computational costs and poor feature disentanglement due to the use of pre-trained models. To address this, they propose a lightweight yet effective model called the Aesthetic Feature-Aware (AesFA) model. Similar to Li et al.[**?**], Kwon et al.'s primary idea is also to decompose images through their frequency to better separate aesthetic styles from reference images. During training, the entire model is trained end-to-end to completely eliminate pre-trained models during inference. Additionally, to enhance the network's ability to extract more distinctive representations and further improve style transfer quality, Kwon et al. introduce a new loss function: the Aesthetic Feature Contrast Loss. This method generates stylized images with superior features, and AesFA also shows reduced time consumption when transferring high-resolution images, demonstrating good style transfer performance.

Constructing Novel Feature Extractors. Wang et al.[**?**] address the efficiency of style transfer algorithms by noting that existing arbitrary style transfer methods struggle or are unable to handle ultra-high-resolution

images (e.g., 4K), which significantly hinders their further application. They reconsidered the entire style transfer process and identified that the slow transfer speeds are primarily due to the widespread use of VGG[**?**] as a feature extractor in previous neural style transfer methods. This is because the fully connected layers of large pre-trained deep convolutional neural networks (DCNNs) require substantial computational resources.

To develop a style transfer model that can efficiently perform tasks on high-resolution images, Wang et al. completely abandoned the traditional use of VGG[**?**] as the content and style feature extractor. Instead, they designed an entirely new feature extractor, naming the overall network model MicroAST. Additionally, Wang et al. proposed a new loss function called "Style Signal Contrastive Loss" to assist MicroAST in style transfer tasks. MicroAST consists of two main components: a micro encoder and a micro decoder. The micro encoder is further divided into a micro content encoder and a micro style encoder, which share the same structure, including one standard stride-1 convolutional layer, two stride-2 depthwise separable convolution (DS Conv) layers, and two stride-1 residual blocks (ResBlocks). The structure of the micro decoder is nearly symmetrical to that of the micro encoder. By discarding the VGG[**?**] feature extractor, Wang et al. significantly reduced the time and memory required for high-resolution style transfer while achieving satisfactory style transfer results.

It is worth noting that the study by Kwon et al.[**?**] also discarded the traditional VGG[**?**] feature extractor to achieve higher speed and lower memory usage. However, the core of that study lies in the other methods employed, so it was not classified under this category.

4. Ebaluation Metrics

Due to the relatively recent development of the style transfer field[**?**] and the difficulty of objectively evaluating the quality of stylized images, there are currently numerous evaluation standards in the style transfer domain. Researchers often struggle to find widely accepted objective metrics when attempting to evaluate a model's style transfer capability. Therefore, this paper has analyzed some evaluation standards used in style transfer-related literature published over the past three years in conferences such as CVPR, ICCV, ECCV, and AAAI, for reference.

Content Loss and **Style Loss** are the loss functions used in the first work on neural style transfer[**?**], and some scholars still consider these

two metrics as standards for image generation. The calculation method for Content Loss is shown in Equation 2, and the calculation method for Style Loss is shown in Equation 5.

Time and Memory Usage are used to measure the efficiency and practicality of an algorithm in style transfer. Time usage reflects the length of time required for an algorithm to complete a task, which is especially important for applications that require quick responses. Memory usage indicates the demand for computational resources when processing images, which is particularly critical for resource-limited devices, such as mobile or embedded systems. Therefore, these two metrics are of significant importance in evaluating and selecting style transfer algorithms that meet specific needs.

The concept of **Deception Rate** was first introduced by[**?**] in 2018, and is described as follows: Train a VGG16 network to distinguish between different artworks from 624 artists in the WikiArt dataset. On this basis, the Deception Rate is defined as:

$$\text{Deception Rate} = \frac{F_{cs}^{t}}{F_{cs}} \tag{14}$$

where F_{cs} represents all stylized images, and F_{cs}^{t}t denotes the number of stylized images that are recognized by VGG16 as artworks created by an artist. Under this definition, Deception Rate represents the proportion of stylized images that are perceived as artist-created rather than computer-generated.

FID (Fréchet Inception Distance) is a commonly used metric for evaluating the output quality of generative models, especially in the field of image generation. This metric was first introduced by Heusel et al.[**?**]. FID evaluates the quality of generated images by comparing the distribution of generated images to that of real images in feature space. Specifically, FID calculates the Fréchet distance (also known as Wasserstein-2 distance) between the distribution of generated data and the distribution of real data in the feature space output by a layer of the Inception network. The FID calculation process is as follows:

1. Feature Extraction: Use a specific layer of the Inception-v3 network (typically a layer after pooling) to extract features from both generated images and real images.
2. Compute Statistics: For the two sets of features (generated image features and real image features), calculate the mean and covariance matrix for each.

3. Calculate Fréchet Distance: Use the following formula 12 to compute the Fréchet distance between the two multivariate Gaussian distributions:

$$\text{FID}(x, g) = \|\mu_x - \mu_g\|^2 + Tr\left((\Sigma_x + \Sigma_g - 2(\Sigma_x\Sigma_g)^{\frac{1}{2}})\right) \tag{15}$$

where μ_x and Σ_x are the mean and covariance matrix of the real image features, and μ_g and Σ_g are the mean and covariance matrix of the generated image features.

A lower FID value indicates higher quality of generated images, meaning that the distribution of generated images is closer to the distribution of real images. The advantage of FID is that it considers not only the quality of individual images but also the overall diversity of the generated image set. Compared to other evaluation metrics such as Inception Score, FID more accurately reflects the differences perceived by human vision, which is why it has gained widespread use in the field of image generation. However, the FID score also has certain limitations, such as its sensitivity to different datasets and model architectures, and it requires significant computational resources for calculation.

Learned Perceptual Image Patch Similarity (LPIPS) is primarily used to measure the visual similarity between two images and was first introduced by Zhang et al. in 2018[**?**]. Unlike traditional pixel-level comparison methods, LPIPS focuses more on perceptual differences between images. It utilizes deep learning models, especially pre-trained neural networks, to simulate the human visual system's perception of image differences. Specifically, the calculation method of LPIPS can be described through the following steps:

1. Feature Extraction: For two images x and y, use a pre-trained deep neural network (such as AlexNet, VGG, etc.) to extract their features at multiple levels l, denoted as $F^l(x)$ and $F^l(y)$.
2. Compute Normalized Feature Differences: For each spatial location i at each layer, compute the feature difference between x and y at that location, and normalize the features. The difference is calculated using the following formula:

$$d_i^l = \frac{1}{N_l}\|_i^l(x) - F_i^l(y)|Vert_2^2 \tag{16}$$

where N_l is the number of channels in layer l, and $||\cdot||_2$ denotes the L2 norm.

3. Apply Learned Weights and Aggregate: The feature differences at each layer are weighted by a learned weight w_l, and then the weighted differences across all layers and spatial locations are summed to obtain the final LPIPS distance:

$$LPIPS(x, y) = \sum_l \sum_i w_l \cdot d_i^l \tag{17}$$

where w_l is the learned weight for layer l through the training process.

A lower LPIPS score indicates that the two images are more similar on a perceptual level. This metric better simulates the human visual system's perceptual characteristics by considering features from multiple layers of the network and adjusting the importance of features at different layers using learned weights.

Structural Similarity Index Measure (SSIM) is a metric used to assess image quality by quantifying the visual similarity between two images. Unlike traditional pixel-based error measures, SSIM considers the structural information of images, better simulating the human visual perception system. The core idea of SSIM is based on the observation that the human visual system is highly adapted to extract structural information from visual scenes. Therefore, SSIM evaluates image quality by considering three comparison dimensions: luminance, contrast, and structure. Specifically, for two images x and y, SSIM is defined as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{18}$$

where μ_x and μ_y are the mean luminances of images x and y, σ_x^2 and σ_y^2 are the variances of images x and y, and σ_{xy} is the covariance between images x and y. Constants c_1 and c_2 are small constants added to avoid division by zero, typically defined as $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$, where L is the dynamic range of pixel values, and k_1 and k_2 are constants chosen to be small.

The SSIM value ranges from 0 to 1, where 1 indicates that the two images are identical. SSIM is more accurate and aligns better with human visual perception when evaluating image quality than metrics based solely on pixel differences, such as MSE (Mean Squared Error), as it considers luminance, contrast, and structural aspects.

Pixel Distance (PD) was first proposed by Wang et al.[?]. This evaluation metric measures diversity by calculating the average distance between sample pairs in both pixel space and deep feature space. Specifically, the average pixel distance between generated image pairs is calculated. For distances in pixel space, the average pixel distance between two images is directly computed across the RGB channels, with the calculation formula as follows:

$$d_{pixel}(\tilde{x}_1, \tilde{x}_2) = \frac{||\tilde{x}_1 - \tilde{x}_2||_1}{W \times H \times 255 \times 3} \tag{19}$$

where $\tilde{x}_1$ and $\tilde{x}_2$ represent the two images for which the pixel distance is to be calculated, and W and H are the width and height of the images. This method quantifies the degree of visual difference between image pairs.

User Surveys are crucial in evaluating style transfer methods because they provide subjective assessments of the style transfer results. While technical metrics can objectively measure certain aspects, the effectiveness of artistic style transfer often relies more on human vision and perception. Through questionnaires, researchers can gather feedback on user satisfaction, visual appeal, and realism of the style transfer results, allowing for a more comprehensive evaluation and improvement of style transfer methods. Based on our knowledge, most neural style transfer papers published after 2019 involved user surveys to demonstrate the superiority of their method in generating stylized images.

Table 1 in the appendix shows the evaluation metrics used in the papers mentioned in this article.

5. From 2D to Multimodal Exploration

Style transfer tasks were initially based on 2D images, with the deep neural network-based style transfer method proposed by Gatys et al. in 2015 marking the inception of this field. Since then, style transfer techniques have primarily focused on transferring the style of images, achieving high-quality artistic style image generation through the extraction and combination of content and style features from the input image. However, as research has progressed, the application scope of style transfer has gradually expanded. It now encompasses various domains, including video style transfer, 3D style transfer, and multimodal style transfer, and has sparked research into maintaining style consistency and cross-modal coordination in these new areas.

Paper	Year	Content& Sytle Loss	Time	Memory Cost	Deception Rate	FID	LPIPS	SSIM	Content Fidelity	User Study	Others
Gatys et al.[?]	2016	1	-	-	-	-	-	-	-	-	-
Li et al.[?]	2023	7.91/2.40	10.7ms(256)	-	-	-	-	0.51	-	1	-
Huang et al.[?]	2017	2.2/8.76	18ms(256)	-	-	-	-	-	-	-	-
Ke et al.[?]	2023	-	-	-	-	-	-	-	-	1	-
Johnson et al.[?]	2016	-	15ms(256)	-	-	-	-	0.61	-	-	-
Ulyanov et al.[?]	2016	-	2-ms(256)	17-MB	-	-	-	-	-	-	-
Risser et al.[?]	2017	-	568s(512)	-	-	-	-	-	-	-	-
Li et al.[?]	2017	-	-	-	-	-	-	-	-	-	-
Li et al.[?]	2016	-	-	-	-	-	-	-	-	1	-
Li et al.[?]	2016	-	10ms(256)/40ms(512)/160ms(1024)	298MB	-	-	-	-	-	-	-
Zhu et al.[?]	2017	-	-	-	-	-	-	-	-	-	1
Men et al.[?]	2022	-	-	-	-	79.74	-	-	-	-	1
Chen et al.[?]	2017	-	-	-	-	-	-	-	-	-	-
Jing et al.[?]	2020	-	-	-	-	-	-	-	-	-	1
Chen et al.[?]	2016	-	45ms(256)	-	-	-	-	-	-	-	-
Xu et al.[?]	2021	-	80ms(768)	1262MB	0.573	-	-	-	-	1	-
Yang et al.[?]	2022	-	-	-	-	-	-	-	-	-	1
Wu et al.[?]	2023	-	-	-	0.672	42.83.	0.223	-	-	1	1
Liu et al.[?]	2021	-	-	-	-	-	-	-	-	1	1
Deng et al.[?]	2022	1.91/1.47	116ms(256)	-	-	-	-	-	-	1	-
Li et al.[?]	2023	1.54/1.06	-	34.87GB	-	-	-	-	-	1	-
Zhang et al.[?]	2024	-	-	-	0.573	-	-	-	0.432	-	-
Wang et al.[?]	2023	-	-	-	-	-	-	-	-	1	-
Zhang et al.[?]	2023	2.18/1.35	270ms(512)	-	-	-	-	-	-	-	-
Hong et al.[?]	2023	-/0.258	-	-	-	-	-	-	0.69	1	1
Zhu et al.[?]	2023	0.55/1.04	14ms(512)	-	-	-	0.52	-	-	1	-
Kwon et al.[?]	2022	-	-	-	-	-	-	-	-	-	1
Hamazaspyan et al.[?]	2023	-	-	-	-	-	0.596	-	-	-	1
Zhang et al.[?]	2023	-	3.5725s(512)	-	-	-	-	-	-	1	1
Zhang et al.[?]	2023	-	-	-	-	-	-	-	-	-	1
AHN et al.[?]	2024	-	-	-	-	-	-	-	-	1	-
Wang et al.[?]	2023	-/0.837	5.612s(512)	-	-	-	-	0.672	-	1	1
Lu et al.[?]	2023	-/0.094	-	-	-	375.206	-	-	-	1	-
Chung et al.[?]	2024	-	-	-	-	18.131	0.506	-	-	-	1
Deng et al.[?]	2024	-	-	-	-	-	-	-	-	1	-
Chen et al.[?]	2019	0	11ms(256)	21885MB	-	-	-	-	-	-	1
Kwon et al.[?]	2023	-/0.216	16ms(256)/20ms(1024) 20m(2048)/20ms(4096)	-	-	-	0.372	-/0.216	-	1	-
Wang et al.[?]	2023	-/2.342	522ms(4096)	1.857MB	-	-	-	0.531	-	1	1
Lin et al.[?]	2023	-	-	-	-	-	0.178	0.7478	-	-	-
Chen et al.[?]	2023	-	-	-	0.474	-	0.375	-	-	1	-

Table 1: Evaluation Metrics Used in Papers. The first column of the table represents the papers and their respective authors, while the second column indicates the year of publication. The remaining columns, starting from the third, represent the evaluation metrics used for style transfer. The third column (Content & Style Loss) refers to the loss function used by Gatys et al., so it cannot be explicitly provided. The fourth column (Time) indicates the amount of time required for the style transfer model to generate an image at the corresponding resolution during inference. In the final columns, "User Study" and "Others," the number "1" indicates that the paper used user studies or other metrics as evaluation criteria, while the symbol "-" indicates that the standard was not used.

Although this paper focuses on presenting the latest advances in 2D image style transfer, a brief overview of the current state of research on style transfer in video, 3D, and multimodal scenarios is provided for researchers to gain a preliminary understanding of its applications in these fields. Notably, with the improvement in computational power and the continuous evolution of technology, style transfer has expanded from 2D images to video, 3D scenes, and multimodal data processing. This not only broadens its application scope but also offers more flexible and creative solutions to practical problems.

5.1. Video Style Transfer

In fact, most real-time style transfer techniques can be applied to some extent to video style transfer. However, since these methods are not specif-

ically optimized for video stylization tasks, they may exhibit deficiencies in temporal consistency during video style transfer. This is manifested as flickering and erroneous discontinuities between video frames. Additionally, generating high-resolution stylized long videos may require substantial computational time.

Existing video style transfer methods can be broadly classified into two categories based on the use of optical flow. To the best of our knowledge, Ruder et al.[**?**] were the pioneers in integrating optical flow information into the video style transfer task. They proposed a method for transferring artistic styles to video sequences. This method utilizes deep neural networks and incorporates both optical flow information and a temporal consistency loss function, generating videos that are stylized with an artistic appearance while maintaining temporal consistency. To address issues of flickering and discontinuities in video style transfer, the authors introduced short-term and long-term consistency losses, and further enhanced the visual quality of the video through a multi-pass algorithm.

Gao et al.[**?**] proposed a novel real-time video style transfer network called ReCoNet, which is capable of generating stylized video sequences with temporal consistency in near real-time. ReCoNet achieves short-term consistency by integrating a flow sub-network and a mask sub-network into the intermediate layers of a pre-trained style transfer network. Additionally, the temporal consistency is propagated over longer periods through the use of a recurrent neural network (RNN) structure, ensuring consistency over extended durations.

Chen et al.[**?**] combined knowledge distillation with optical flow information to train an efficient and stable video style transfer network. The method leverages knowledge distillation to transfer the knowledge of a stable style transfer network with an optical flow module (the teacher network) to a lightweight network that lacks the optical flow module (the student network). By introducing residual distillation loss and low-rank distillation loss, the student network learns the temporal consistency information brought by the optical flow from the teacher network, enabling it to generate stylized videos with similar low-rank characteristics. This approach enhances the stability and efficiency of video style transfer.

Optical flow-based video style transfer often struggles to achieve an optimal balance between stability and efficiency. Gao et al.[**?**] only employed optical flow during the training phase, aiming to improve the efficiency of the inference process. However, compared to Ruder et al.[**?**] who also used op-

tical flow during the inference phase, methods that rely solely on optical flow in the training phase tend to produce less stable results. On the other hand, the optical flow distillation method proposed by Chen et al.[**?**] partially balances efficiency and stability. Nevertheless, it requires training a separate model for each style, thus limiting its ability to perform style transfer for arbitrary styles.

In recent years, video style transfer methods that do not rely on optical flow constraints to ensure temporal consistency have emerged as a promising approach. Li et al.[**?**] introduced a method that achieves video style transfer by learning a linear transformation matrix. This approach effectively preserves content information, enabling the generation of stable results without the need for optical flow-based temporal consistency guarantees.

Wang et al.[**?**] proposed a novel video style transfer framework that generates stylized video sequences with consistency both temporally and visually. To address the issues of flickering and discontinuity in video style transfer, the paper introduces a new composite regularization term that better captures the intrinsic temporal variations and effectively balances the temporal and stylization effects. Additionally, Wang et al. designed a sequence-level global feature sharing strategy to achieve long-term temporal consistency and proposed a dynamic inter-channel filter module to enhance the style transfer performance while supporting end-to-end training.

Deng et al.[**?**] proposed a video style transfer network called MCCNet, which achieves high-quality arbitrary style transfer by cross-domain feature alignment while preserving the structural content of the video. The network does not require optical flow computation; instead, it directly rearranges and fuses features in the content and style feature spaces, ensuring that the style patterns adapt to the content structure. Furthermore, to enhance the model's stability under complex lighting conditions, the paper introduces an illumination loss that simulates lighting changes, thereby improving the temporal consistency of the generated video.

Xia et al.[**?**] took a novel approach by employing bilateral learning to achieve real-time, locally light-realistic video style transfer. They proposed a deep neural network-based method capable of locally transferring artistic styles to the semantic regions of a video while maintaining photorealism and temporal consistency. By introducing a spatiotemporal feature transfer layer (ST-AdaIN), enhancing segmentation masks, and performing color transformation fusion in a low-resolution grid space, their method significantly improves both the quality and efficiency of the results.

5.2. 3D Style Transfer

Compared to the 2D image style transfer proposed by Gatys et al., 3D style transfer is a more recent concept.[**?**] 3D style transfer can be categorized into geometry stylization and model texture stylization, depending on the target.

Liu et al.[**?**] are pioneers in the field of 3D style transfer, where they transform the input shape into a fixed cubic style, achieving geometric 3D style transfer. Wang et al.[**?**] built upon this foundation, presenting a more mature approach to 3D geometric style transfer. Their work is capable of distorting the source geometry to match the target geometry while preserving the surface details of the source data.

In recent years, 3D texture transfer technology has gradually gained prominence. The core challenge lies in accurately mapping the target style onto the surface texture and details while preserving the geometric shape of the 3D object. Compared to 2D image style transfer, 3D texture style transfer involves additional complexities, including multi-view consistency, geometric integrity of the 3D shape, and high-resolution texture mapping. These characteristics make 3D style transfer highly applicable in fields such as virtual reality (VR), game development, and architectural rendering.

Huang et al.[**?**] proposed a point cloud-based 3D scene stylization method. This approach constructs point clouds by back-projecting image features into 3D space and uses a point cloud transformation module to transfer the style information from the reference image onto the point cloud. The final result is a new viewpoint image with consistent style.

Wu et al.[**?**] aimed to achieve arbitrary and real-time style transfer. To this end, they proposed a feedforward 3D style transfer method called StyleFormer. StyleFormer consists of three modules: the Style Bank Generation Module, the Transformer-driven Style Composition Module, and the Parametric Content Modulation Module. These modules work collaboratively to generate 3D transfer results with coherent content structure.

Yin et al.[**?**] proposed a neural style transfer method called 3DStyleNet, which simultaneously transfers both the geometric shape and texture style of 3D objects. 3DStyleNet first defines the style by learning the global geometric relationships between semantic components, then predicts a component-aware affine transformation field to distort the source shape, mimicking the overall geometric style of the target shape. Additionally, 3DStyleNet leverages a differentiable renderer and a pre-trained image style transfer network

to transfer the texture style of the target shape onto the distorted source object. By jointly optimizing the geometry and texture, 3DStyleNet generates 3D objects with realistic style details and coherent content structure, making it applicable to tasks such as 3D content creation and style-based data augmentation.

Chiang et al.[**?**] addressed the issue of viewpoint inconsistency in 3D style transfer by proposing a NeRF-based 3D scene style transfer method. This approach learns the implicit representation of the 3D scene and utilizes a hypernetwork to transfer the style information from the reference image into the scene representation. The method ultimately generates new viewpoint images with consistent style.

Huang et al.[**?**] proposed a 3D scene stylization method called Stylized-NeRF, which integrates a 2D-3D co-learning framework that combines a 2D image stylization network with NeRF to achieve high-quality 3D scene stylization while maintaining geometric consistency. The method first trains a standard NeRF to represent the 3D scene and then replaces its color prediction module with the stylization network to obtain a stylized NeRF. Next, by introducing a consistency loss, the spatial consistency prior knowledge from NeRF is distilled into the 2D stylization network. Additionally, the method incorporates learnable latent codes to address the ambiguity of 2D stylization results and enable conditional stylization.

Hollein et al.[**?**] focused on 3D indoor scene reconstruction. They scanned and constructed RGB-D maps of indoor scenes to generate 3D meshes and, by combining depth and surface normal data, achieved consistent and viewpoint-dependent style transfer in 3D space. However, this method requires separate optimization for each scene, which results in relatively low efficiency.

Thu et al.[**?**] aimed to generate clear and detailed stylized 3D scenes and proposed the StyleMesh model. This method optimizes the display textures of reconstructed meshes and incorporates style information from all available images, achieving the generation of 3D consistent stylized scenes.

Liu et al.[**?**] proposed the StyleRF model, which consists of three core components: Explicit Feature Grid 3D Representation, Sampling-invariant Content Transformation (SICT), and Deferred Style Transformation (DST). The Explicit Feature Grid 3D Representation utilizes high-level feature grids to represent 3D scenes, achieving high-fidelity geometric reconstruction through volume rendering. The Sampling-invariant Content Transformation (SICT) eliminates dependence on the statistics of sampling points through volume-

adaptive normalization and channel self-attention mechanisms, ensuring multi-view consistency. The Deferred Style Transformation (DST) delays the style transformation to the 2D feature maps after volume rendering. By employing the Explicit Feature Grid 3D Representation, the model enhances the accuracy of geometric reconstruction. SICT ensures consistent styling across multiple views, while DST reduces computational complexity by postponing the style transfer process to the rendered 2D feature maps. The combination of these components improves both geometric reconstruction accuracy and computational efficiency.

NeRF-Art, proposed by Wang et al.[**?**] is a text-driven NeRF stylization method that allows for style modification of a pre-trained NeRF model through simple text prompts. This approach combines the CLIP model with NeRF and introduces a global-local contrastive learning strategy and directional constraints, simultaneously controlling the direction and intensity of the target style. It achieves joint stylization of both appearance and geometry. Furthermore, the method employs weight regularization techniques to effectively reduce cloud-like artifacts and geometric noise generated during the stylization process.

Similar to Wang et al.[**?**] Haque et al.[**?**] also aimed to use natural language for 3D scene stylization. To this end, they proposed Instruct-NeRF2NeRF, a novel 3D scene editing method that allows for the editing of pre-captured NeRF scenes using natural language instructions. The method first uses the InstructPix2Pix diffusion model to edit the images rendered by NeRF based on the instructions. The edited images are then used to update the NeRF training dataset, followed by re-training NeRF to integrate the editing results into the 3D scene. This approach supports a variety of local and global scene edits, such as changing textures, replacing objects, and altering global scene attributes, while maintaining good consistency.

Chen et al.[**?**] proposed a universal, realistic style transfer method called UPST-NeRF, which transfers the style of any given image to a 3D scene while ensuring consistency across images rendered from different viewpoints. The method first reconstructs the scene's geometric structure using a voxel grid, then employs a hypernetwork to incorporate features from the style image as latent codes for scene stylization. It further constrains the style of the rendered images using a pre-trained 2D realistic style transfer network, thus achieving high-quality 3D scene stylization.

He et al.[**?**] tackled the issue of viewpoint consistency in 3D style transfer from the perspective of frequency domain decomposition. They proposed

Freditor, a 3D style transfer method based on frequency decomposition techniques. The method is built on two key observations: (1) the low-frequency components of an image exhibit better multi-view consistency after editing than the high-frequency components, and (2) the appearance style is mainly represented in the low-frequency part, while content details are concentrated in the high-frequency part. Based on these observations, they introduced a frequency decomposition and feature space editing approach. The frequency domain decomposition step primarily splits the NeRF-rendered images into low-frequency and high-frequency parts, with the low-frequency components used for style editing and the high-frequency components preserving content details. Style editing is then performed in the feature space, rather than directly in the RGB space, enabling more stable control over style intensity and migration. Through these two methods, Freditor alleviates the multi-view inconsistency problem in 3D style transfer to some extent.

From the above studies, it is evident that current 3D style transfer faces a triple dilemma involving geometric reconstruction accuracy, high-quality stylization, and generalization to arbitrary new styles. Increasing geometric reconstruction accuracy often requires significant computational resources, whereas enhancing the quality of stylization may come at the cost of the ability to transfer arbitrary styles. Balancing these three aspects—accuracy, quality, and generalization—represents a core challenge in the field of 3D style transfer.

To address this, future approaches might need to explore more efficient optimization techniques, such as leveraging multi-scale representations or novel neural architectures that balance these competing goals. For example, introducing adaptive strategies for geometric reconstruction and style transfer at different levels of detail could help manage the trade-offs between computational cost and output quality. Additionally, incorporating domain-specific priors or style-specific models might enhance the flexibility and efficiency of 3D style transfer methods, allowing for the consistent transfer of various styles without excessively compromising on any one aspect.

6. Frontiers and Challenges

By combining the artistic style of one image with the content of another, style transfer technology has had a profound impact on areas such as image editing, film production, and computer art (e.g., [**? ? ?**]). However, despite impressive progress, the field of style transfer still faces a series of

challenges. These issues not only involve improvements in image quality and visual realism but also include challenges related to algorithm efficiency, generalization, and interpretability. This section focuses on the unresolved problems in the field, aiming to provide valuable guidance and inspiration for future research.

6.1. Evaluation Criteria

Currently, there is a wide variety of evaluation criteria in the field of style transfer, with no unified objective standard. For example, in 2023 papers, Wu et al.[**?**] used Deception Rate, FID, and LPIPS as evaluation metrics; Wang et al.[**?**] employed time and memory consumption, SSIM, and Style Loss; Li et al.[**?**] utilized SSIM, LPIPS, Content Loss, and Style Loss; Cheng et al.[**?**] adopted Pixel Distance, LPIPS, and Deception Rate. This diversity and lack of standardization in evaluation metrics make it difficult for researchers to compare the performance of different models across various aspects.

Additionally, since style transfer often generates images with artistic qualities, researchers have tried to obtain more objective evaluations through user surveys. According to our statistics, almost all publications in the field of style transfer since 2019 involved questionnaires to compare current methods with previous ones.

To discuss the effectiveness of this approach, Jing et al.[**?**] investigated the impact of age and profession on aesthetic judgment. They asked eight participants with the same profession but different ages (four males and four females) to rate each stylized image. The experimental results revealed that even with the same stylized output, different observers of the same profession and age still gave significantly different ratings.

As a result, there is currently no widely accepted evaluation standard in the field of style transfer. Considering that the evaluation of artistic images is influenced by aesthetic ability, it may be necessary to establish a convincing evaluation standard with the assistance of art professionals.

6.2. Interpretability

Current research on style transfer develops limited research on explicit interpretability [**?** **?** **?** **?**]. While Transformer-based models have demonstrated significant success in capturing global context and fine-grained details through self-attention mechanisms, the interpretability of their multi-head attention layers remains a major challenge. Specifically, it is often unclear

how individual attention heads contribute to the disentanglement of content and style features, making the style transfer process opaque to users [? ? ? ?].

Attention maps are a promising avenue for enhancing the interpretability of style transfer models, as they provide visual insights into how the model attends to different regions of the input during the style transfer process. In Transformer-based style transfer, attention maps generated by multi-head self-attention layers can reveal the relationship between content and style features, helping to understand which regions of the content image are most influenced by the style image. However, current research rarely incorporates attention map visualizations, leaving the decision-making process of these models largely opaque.

Some diffusion models use visual programming-based approaches to achieve controllability [? ?]; however, the transferring process is not transparent. Diffusion models typically rely on iterative noise removal processes that involve multiple stochastic steps, which are difficult to interpret. It is challenging to determine how intermediate representations evolve towards the final style-transferred output, and this "black-box" nature limits their usability in applications requiring high transparency and user control.

Similarly, GAN-based style transfer models, while capable of generating high-quality stylized outputs, often lack interpretability due to their adversarial training paradigm. The generator network focuses solely on minimizing the loss function provided by the discriminator, without providing insights into how style features are applied or modified during the generation process. Additionally, GANs are prone to mode collapse and instability during training, further complicating their interpretability and making it difficult to diagnose errors or optimize style transfer results [? ?].

6.3. Deformation

Current style transfer algorithms primarily focus on converting the texture and color of a content image to match a style image. However, some styles are abstractions and simplifications of the real world (e.g., animation styles and abstract art). Therefore, transferring texture merely is insufficient. It is necessary to explore the target style during the transfer process and design methods to account for image deformation during style conversion.

6.4. Transferring Texture and Color

Sometimes, people wish to retain the original colors of an image while only transferring the texture from the style image to the content image. However, current algorithms often transfer both texture and color simultaneously. Therefore, the ability to transfer only texture or only color remains a problem that needs to be addressed.

6.5. Applications

The current development of style transfer largely focuses on achieving arbitrary style transfer with a single model. However, achieving arbitrary style transfer does not necessarily mean that the model can be used in production activities. In production processes, customization is often required. Therefore, the ability to intervene in the generation process to create images with the desired style is an important issue.

6.5.1. Controllability Issues

A major challenge hindering the widespread application of style transfer in fields such as medicine, 3D scene design, and assistive design is its lack of controllability. Professionals in these fields typically do not have a computer science background, and when attempting to apply style transfer methods, they often seek a high level of customization tailored to their specific needs. However, existing methods are often black-box operations, offering little intuitive control over the generated results.

In the medical field[**?**], doctors may need to apply differentiated style transfer to different regions of pathology images during diagnosis. For example, they might want to highlight key areas, such as lesion sites, with enhanced stylization to emphasize structural features, while leaving other areas, such as the background, with minimal or no style transfer. However, current style transfer methods typically apply a uniform style across the entire image, lacking the ability to fine-tune localized areas. This limitation increases the complexity of image interpretation in practice and may impact diagnostic accuracy and efficiency.

In the AR and VR domains of 3D scene style transfer[**? ? ? ? ? ? ? ?**], designers often need to apply diverse styles to different parts of the virtual environment. For instance, they may want to render objects like buildings with a realistic style while applying an abstract style to other elements, such as natural landscapes. Additionally, the natural transition and blending between different styles within the same scene remain a weak point in current

technology. The inability to achieve smooth stylistic transitions limits the practical application of style transfer in high-fidelity virtual environments and game design, causing the generated scenes to lack consistency and artistic expression in finer details.

In the assistive design field[? ? ? ? ?], the controllability of style transfer is particularly crucial for meeting users' personalized creative demands. Designers may need to integrate multiple styles within a single project, such as applying different styles to buttons, backgrounds, and fonts in user interface design. However, current style transfer technologies struggle to provide independent style control for specific elements. Moreover, designers may want to dynamically adjust the intensity of style transfer for real-time previews and design optimization, a need that has yet to be adequately addressed within current technological frameworks.

In fact, some existing works [? ? ? ? ? ?] have explored the controllability issues of style transfer to some extent. Several studies[? ?] have attempted to integrate object detection and image segmentation technologies into style transfer. The incorporation of these techniques provides style transfer with more fine-grained and target-oriented capabilities, especially demonstrating significant potential in localized style transfer and multi-target scenarios.

In terms of image segmentation[? ?], segmentation techniques allow the input image to be decomposed into semantically meaningful regions, such as foreground, background, or specific object categories. This region-based segmentation provides a precise operational foundation for style transfer, enabling not only global image processing but also the ability to apply differentiated stylization to specific regions. For example, in medical image style transfer, by accurately segmenting the lesion area, doctors can apply higher contrast or specific styles to the targeted regions, enhancing the visualization for diagnosis. In artistic image style transfer, segmentation techniques help achieve independent style transfers for different objects in complex scenes, such as applying different artistic styles to buildings, skies, and people, thereby enhancing the diversity of overall artistic expression.

Object detection technology[? ?] further strengthens the target-oriented nature of style transfer tasks. With object detection models, style transfer can identify and select specific object categories for stylization, free from the interference of irrelevant background elements. For instance, in virtual reality (VR) design, object detection can automatically recognize specific elements in the scene, such as trees, buildings, or vehicles, and apply the

desired style to those elements, instead of performing a broad, uniform style transfer across the entire scene. This fine-grained control not only enhances the flexibility of style transfer but also improves the practical usability of the generated results.

Moreover, image segmentation and object detection technologies can also be used to achieve multi-style fusion. By using segmentation techniques to identify different regions and combining object detection to determine their semantic categories, style transfer models can facilitate seamless transitions of styles across different regions within the same image. This approach holds great potential in fields such as game design and advertising. For example, designers can apply an impressionist style to the sky portion of a scene, while using an abstract art style for the ground, ensuring a smooth and natural transition between the two styles.

In the context of domain adaptation[**? ? ? ?**], style transfer can be viewed as a specialized form of domain adaptation. Both tasks involve the process of transferring knowledge from one domain to another, with domain adaptation focusing on generalizing a model trained on a source domain to a target domain, and style transfer aiming to blend the content of one domain with the style of another. More specifically, style transfer can be seen as an adaptation of content features from a source domain (such as a photo or real-world image) to a target style domain (such as an artwork or painting).

Both domains share similar challenges, such as the need for handling domain discrepancies and ensuring the consistency of features across domains. However, style transfer often operates on a more specific level, dealing with the fusion of content and style, whereas domain adaptation encompasses a broader range of tasks, including but not limited to, classification, detection, and segmentation in cross-domain scenarios. Despite these differences, the techniques used in domain adaptation—such as feature alignment, adversarial training, and transfer learning—are highly relevant to style transfer, especially in tasks that involve transferring across vastly different domains (e.g., from real-world imagery to artistic representations).

In this sense, style transfer can be viewed as a form of domain adaptation that focuses specifically on transferring the "style" of one domain to the "content" of another. For example, in a style transfer task, the domain adaptation process may involve learning how to adapt the feature distributions of content and style from two separate domains in such a way that both content and style are preserved while minimizing discrepancies. Thus, advancements in domain adaptation, such as improved feature alignment and

adversarial loss functions, hold significant promise for enhancing the realism and effectiveness of style transfer, particularly in challenging scenarios where the content and style domains differ significantly.

6.5.2. Efficiency Issues

Efficiency becomes a significant challenge when dealing with high-resolution images or high-precision scenes in style transfer. Current research primarily focuses on improving the quality and diversity of style transfer models, with less attention given to the computational resource consumption involved in processing high-resolution images. However, in practical applications, the demand for high-resolution images is ubiquitous. For instance, stylizing images at 4K resolution and beyond has become a standard in industries such as film production, virtual reality (VR) rendering, and advertising design. Yet, existing methods often struggle to meet the high-resource demands required for such tasks.

As the resolution of images increases, the number of pixels the model must process grows exponentially, resulting in a sharp increase in computational time and memory requirements. While some studies have attempted to address the efficiency issues in style transfer for high-resolution images, there is still limited research on high-precision video and 3D scene style transfer.

To address these challenges, some research has started exploring methods to improve the efficiency of high-resolution style transfer. One approach is patch-based processing, which divides the image into smaller regions, applies style transfer to each region separately, and then seamlessly stitches the stylized regions back together to form the complete stylized image. However, this method may introduce style inconsistencies between different regions. Another effective approach is the use of lightweight network architectures, which significantly reduce resource consumption by minimizing network parameters and optimizing the computational pathways.

In the future, optimizing model structures, enhancing hardware utilization, and developing algorithms specifically tailored for high-resolution style transfer will be key directions for advancing the application of style transfer in high-precision scenarios.

6.5.3. Cross-Model Style Transfer

Current style transfer research mainly focuses on specific modalities, such as static images, videos, 3D scenes, or audio. However, in real-world applications, data from different modalities often appear simultaneously and are

closely interrelated.

For instance, in film production, virtual reality (VR), and multimedia creation, video and audio frequently need to work in tandem to present a unified artistic style and emotional expression. However, most existing style transfer methods are confined to single modalities and struggle to meet the consistency requirements for cross-modal style transfer. Typically, style transfer is performed separately on each modality, neglecting the semantic relationships and style alignments between modalities. For example, after applying visual style transfer to a video, the audio may be completely disregarded, resulting in a mismatch between the visual aesthetics and the auditory atmosphere. This inconsistency diminishes the immersive experience for users.

Moreover, cross-modal style transfer requires maintaining stylistic coherence across different modalities, an aspect that has received insufficient attention in existing research. For instance, in 3D scene style transfer, the style of the accompanying sound effects or dynamic video must be coordinated with the visual style. However, current methods have not sufficiently explored how to establish consistent styles across modalities.

In cross-modal scenarios, the sheer volume and complexity of the data present significant challenges. For example, the collaborative style transfer of 4K video and high-quality audio demands immense computational resources. Yet, existing methods often face efficiency bottlenecks when dealing with high-resolution images in a single modality, making it difficult to scale up to multi-modal tasks. Furthermore, users often require flexible control over the style transfer effects in multi-modal contexts, a need that current approaches fail to fully address.

The limitations of cross-modal style transfer directly hinder the potential applications of style transfer technologies in numerous practical scenarios, particularly in fields such as virtual reality, film production, and multimedia art creation. This fragmentation not only weakens the overall experience but also burdens creators with the need to manually coordinate styles across different modalities, reducing work efficiency and increasing complexity.

6.6. Frontiers

Although this article primarily introduces the field of image style transfer in chronological order of development, to help researchers better understand the current frontier of the field, we will now present the common themes in image style transfer research. Current research in style transfer can be

broadly classified into the following major directions based on the task objectives: improving style transfer quality, enhancing style transfer speed, and strengthening human-computer interaction.

Improving Style Transfer Quality. In recent years, style transfer technology has made significant progress in generating high-quality images[**? ? ? ? ? ? ? ?**], particularly in its ability to preserve details in high-resolution scenes. However, there is still room for improvement in terms of visual fidelity, especially regarding the intuitive similarity between generated images and artistic works. At times, existing methods may overly focus on content loss and style loss, neglecting the overall visual consistency.

Enhancing Style Transfer Speed The demand for real-time style transfer is increasing, especially in applications like live streaming and interactive creative tools, where users expect style transfer operations to be completed in milliseconds. Although lightweight networks and efficient inference techniques are gradually gaining traction[**? ? ? ?**], speed bottlenecks in high-resolution and multi-modal tasks remain a prominent research challenge.

Strengthening Human-computer Interaction The controllability of style transfer[**? ? ? ? ? ? **] is directly related to user experience. In many application scenarios, users need to fine-tune the intensity, regions, and style blending ratio of the transfer. However, existing methods offer limited support for user interaction, and their interactivity and customization capabilities need further improvement.

Since 2016, style transfer technology has undergone several paradigm shifts, with various methods demonstrating different strengths and limitations. The classic method proposed by Gatys et al., based on iterative optimization, utilizes convolutional neural networks and iterative optimization to precisely match content and style losses. However, this method often comes with high computational costs, making real-time applications difficult, especially for high-resolution images.

Generative Adversarial Networks (GANs), through adversarial training, generate high-quality stylized images, becoming a mainstream method for style transfer after iterative-based approaches. They achieve an excellent balance between real-time performance and image quality in many scenarios. However, such methods are highly dependent on training data and are prone to mode collapse, which limits their stability and robustness.

Transformer-based methods have become one of the mainstream approaches in style transfer. By leveraging the attention mechanism, Transformer mod-

els can capture global contextual relationships, providing new possibilities for preserving fine details during style transfer. However, due to the focus of window-based attention primarily on local features within a window, these methods tend to produce grid-like artifacts when handling large images, which can negatively impact the quality of the style transfer.

Large model-based methods, which combine multimodal pretraining, offer cross-task generalization and bring more innovative possibilities to style transfer tasks. However, the images generated by these methods often exhibit a high degree of stylistic homogeneity, which is particularly evident when transferring the style of portrait photos.

In summary, beyond traditional efforts to improve quality and transfer speed, style transfer technology is gradually evolving from single-modal applications (such as image style transfer) to multimodal development, positioning itself as an entry point for human-computer interaction. Works like (insert reference) have attempted to integrate doodles, textual descriptions, and style transfer, not only realizing multimodal style transfer but also enhancing interactivity to some extent. In addition to text and images, future style transfer technologies need to support multiple modalities simultaneously (e.g., video and audio, 3D and text) while maintaining stylistic consistency across modalities. This trend toward multimodality is not only an expansion of technical capabilities but also opens new avenues for the application of style transfer in fields such as virtual reality and multimedia creation.

7. Conclusion

Since its rise in 2016[**?**] , neural style transfer has seen significant growth. Historically, style transfer has evolved from slow to real-time in terms of speed, from arbitrary style transfer to specific style transfer, and then to arbitrary transfer in terms of style realization, and from simple to complex in terms of structure. Differences in researchers' understanding of the concept of style have led to significant differences in implementation methods. However, current approaches to improving style transfer mainly fall into three categories: enhancing the quality of style transfer, optimizing the runtime, and increasing the interactivity of style transfer. Of these, interactivity is relatively less studied.

Meanwhile, as style transfer has developed, researchers in other fields have also adopted the ideas of style transfer. Some researchers have focused on applying style transfer in everyday life (e.g., [**? ?**]); others have applied

it in the field of artistic design (e.g., [? ? ? ? ?]); still others have applied style transfer methods to other research areas, such as adversarial example research[? ?], image generation[?], domain adaptation[?], font generation[?], and font recognition[?].

It seems that style transfer is no longer a field for mere self-amusement but has become a research direction that can provide new perspectives for other areas. Although the field of style transfer has made significant progress, challenges remain. These challenges will continue to drive researchers to develop deeper, propose more advanced models and algorithms, and further develop style transfer technology. In the future, research in style transfer may focus more on integrating with other technologies, such as combining self-supervised learning and multimodal learning, to open up more innovative application scenarios.

References