information

Review

Style Transfer Review: Traditional Machine Learning to Deep Learning

Yao Xu [1], Min Xia [2,*], Kai Hu [1], Siyi Zhou [3] and Liguo Weng [1,2]

[1] Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202212490023@nuist.edu.cn (Y.X.); 001600@nuist.edu.cn (K.H.); 002311@nuist.edu.cn (L.W.)
[2] Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China
[3] Department of Computer Science, University of Reading, Whiteknights, Reading RG6 6DH, UK; ao834745@student.reading.ac.uk
* Correspondence: xiamin@nuist.edu.cn

Abstract: Style transfer is a technique that learns style features from different domains and applies these features to other images. It can not only play a role in the field of artistic creation but also has important significance in image processing, video processing, and other fields. However, at present, style transfer still faces some challenges, such as the balance between style and content, the model generalization ability, and diversity. This article first introduces the origin and development process of style transfer and provides a brief overview of existing methods. Next, this article explores research work related to style transfer, introduces some metrics used to evaluate the effect of style transfer, and summarizes datasets. Subsequently, this article focuses on the application of the currently popular deep learning technology for style transfer and also mentions the application of style transfer in video. Finally, the article discusses possible future directions for this field.

Keywords: style transfer; CNN; GAN; video consistency

Academic Editors: Vasco N. G. J. Soares, Robin Haunschild, Giorgio Maria Di Nunzio, Paulo Quaresma, Luigi Laura and Marcin Paprzycki

Received: 14 January 2025
Revised: 15 February 2025
Accepted: 17 February 2025
Published: 19 February 2025

Citation: Xu, Y.; Xia, M.; Hu, K.; Zhou, S.; Weng, L. Style Transfer Review: Traditional Machine Learning to Deep Learning. *Information* **2025**, *16*, 157. https://doi.org/10.3390/info16020157

1. Introduction

Artists and designers have always explored how to blend different styles of artistic elements into their work, creating artworks with unique charm. In the field of computer vision, early research primarily focused on image analysis and recognition, including areas like object detection, image classification, and object segmentation. These two domains naturally converged. Artists employed various techniques to achieve this goal, such as shape and texture techniques. The inspiration for style transfer technology stems from these artistic practices. It seeks to emulate the creative process of human artists by applying different artistic styles to real-world images, thus creating images with an artistic flair.

Style transfer can apply the style of one image to another. Specifically, it can transfer the style features of one image, such as color, texture, composition, etc., onto another image, making the latter appear as if it has been re-rendered in the new style of the former. Gatys [1] defined style transfer as the process of transforming the content of one image into the style of another image while preserving the content information of the original image. They achieved this goal by using neural style transfer methods that leverage deep convolutional neural networks.

The importance of style transfer lies in its ability to bring innovation and novelty not only to artistic creation but also to various practical applications. For instance, in the field

of image editing, style transfer can quickly transform the style of an image into the desired aesthetic, providing users with more creative and imaginative possibilities. In the realm of movie special effects production, style transfer can make movie scenes appear more realistic and vivid. Additionally, style transfer can be applied in areas like virtual reality and augmented reality, transforming videos of real scenes into cartoon or watercolor styles, offering users a richer and more immersive virtual reality experience.

Image style transfer is a research area that spans multiple fields, including computer vision, computer graphics, and machine learning. Its development has evolved from traditional texture-based methods to deep learning-based approaches. The earliest texture-based methods primarily relied on texture matching to achieve style transfer but had some limitations, such as an inability to handle global structure and semantic information.

With the rise of deep learning, researchers began applying convolutional neural networks (CNNs) to style transfer, improving the algorithms' ability to capture global structure and semantic information, leading to more natural and realistic results. Simultaneously, as deep learning algorithms continued to advance and evolve, deep learning-based image style transfer methods also improved and were optimized. Techniques such as adversarial training and multi-scale processing were introduced to generate more realistic and detail-rich images.

The current style transfer techniques can be classified according to the categories shown in Figure 1. In this paper, these classifications will be detailed and discussed.

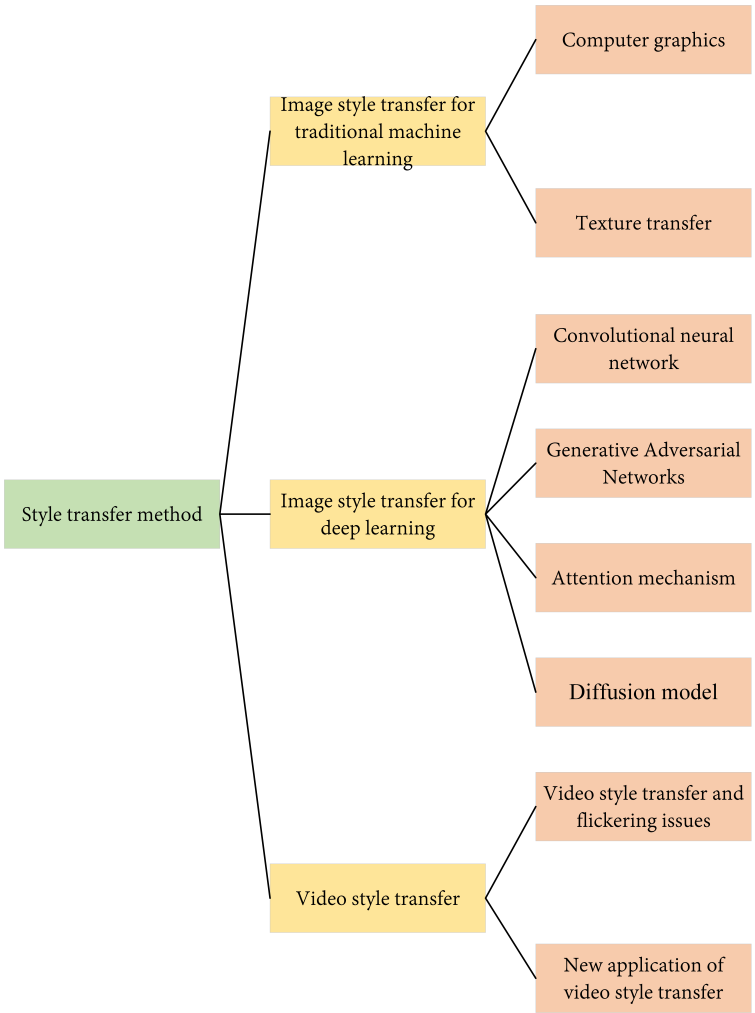

Figure 1. An overview of style transfer methods, divided into traditional machine learning image transfer, deep learning image transfer, and video transfer.

In traditional style transfer methods, early researchers used rules and mathematical models to simulate the creative process of artists. In 1995, Heeger [2] introduced a pyramid-based texture analysis and synthesis method. These methods primarily relied on texture analysis and synthesis techniques. They analyzed the textures in artworks and then synthesized these textures into a new image to achieve style transfer. Texture synthesis laid the foundation for the development of style transfer techniques. However, this method required manual selection of the number of levels in the pyramid and the size of each sub-image at each level, and these parameter choices could impact the synthesis results.

In 1999, Efros [3] proposed a non-parametric sampling-based texture synthesis method. This method generated new texture images by sampling from given texture samples and constructing statistical models. Their approach involved analyzing local statistical information in images and then synthesizing this information into a new image, enabling content synthesis. This content synthesis method provided more inspiration and ideas for the development of style transfer techniques.

In 2004, Lowe [4] introduced an image stitching method based on the SIFT algorithm, providing important ideas for feature extraction and matching in subsequent image style transfer tasks. Then, in 2007, Hays [5] built upon the SIFT algorithm and proposed an image composition algorithm based on local feature matching, capable of automatically composing scenes. Both of these methods involved extracting features from artworks and applying these features to new images to achieve image style transfer.

With the increasing demands for various applications, the limitations of traditional style transfer methods have become more apparent. While traditional methods have their value in certain scenarios, they struggle to handle complex artistic styles, require manual feature extraction, and cannot express high-level semantic information [6]. On the other hand, deep learning techniques have achieved significant success in the fields of image processing and computer vision. They are capable of automatically learning feature representations and semantic information from images. Naturally, deep learning techniques have found their way into the field of style transfer [1]. Deep learning-based style transfer methods offer greater flexibility and effectiveness, as they can automatically learn image features and high-level semantic information. They can also handle more abstract and complex artistic styles. Therefore, the transition from traditional style transfer methods to deep learning-based methods is a natural development trend [7].

Therefore, it is crucial to review the research progress in the field of style transfer, as it not only systematically summarizes and organizes the current methods and technological developments but also guides future research directions and innovations. Through an in-depth analysis of the merits, challenges, and respective solutions of different methods, these reviews facilitate academic and industrial collaboration, while providing essential resources for education and training. This effort has profound significance in advancing and influencing the field of style transfer, as evidenced by the diligent efforts of numerous scholars in this domain.

In 2021, Pang [8] conducted a review of image-to-image translation, presenting various image translation techniques, including methods based on Generative Adversarial Networks (GANs), conditional generative models, image super-resolution, and image semantic segmentation. Their work provided an important overview of research in the field of image-to-image translation. Expanding on the topic of style transfer, Chen [9] categorized style transfer into two main classes: traditional image style transfer and deep learning-based image style transfer. Deep learning-based image style transfer was further subdivided into convolutional neural network-based style transfer and Generative Adversarial Network-based style transfer. Additionally, Zhao [10] introduced significant deep learning-based image style transfer algorithms, including Neural Style Transfer, WCT Style

Transfer, Fast Style Transfer, and CIN Style Transfer, while comparing and analyzing their respective strengths and weaknesses.

However, they did not categorize the deep learning style transfer methods in detail, and as time progresses, more emerging technologies are being applied to the field of style transfer, making a comprehensive review of style transfer techniques increasingly important once again. This paper will classify deep learning-based style transfer algorithms in detail, and sequentially introduce style transfer algorithms under CNN, GAN, attention mechanisms, and the diffusion model.

Jing [11] categorized the architecture of deep learning neural network style transfer. As of 2019, they divided neural network style transfer into two main categories: slow image stylization transfer algorithms based on online image optimization and fast image stylization transfer algorithms based on offline model optimization. Furthermore, they further subdivided these two categories into the following subclasses: slow stylization transfer algorithms based on statistical distribution parametrization, non-parametric slow stylization transfer algorithms based on Markov Random Fields (MRF), and fast stylization transfer algorithms with a single model and style, fast stylization transfer algorithms with a single model and multiple styles, and fast stylization transfer algorithms with a single model and arbitrary styles.

However, their research primarily focuses on neural network-based style transfer, with limited exploration of style transfer using generative adversarial networks (GANs). This paper aims to enhance the discussion by delving into GAN-based style transfer algorithms, attention mechanisms, and the diffusion model in greater detail. It seeks to provide readers with a more comprehensive overview of the development of style transfer techniques.

Chen [12] provided a comprehensive overview of the development and applications of adversarial learning-based image transformation methods in recent years. The introduction of Generative Adversarial Networks (GAN) rapidly found applications in the field of style transfer. On the other hand, Jiao [13] highlighted the challenges that deep learning faces in image processing, including issues such as insufficient data, overfitting, limited model generalization, long runtime, and poor interpretability. Jing [11], proposed future research directions, emphasizing the need to focus on addressing practical issues in style transfer, such as improving image resolution and accelerating algorithm runtime.

The article will continue to delve into key issues facing the current style transfer field, exploring possible solutions and future directions. By systematically analyzing the limitations and challenges of existing methods, the paper aims to reveal the latest advancements in style transfer technology. It will discuss strategies for improving generation quality, achieving more flexible style control, and enhancing widespread practical applications. This effort aims to lead research and technological innovation in the field of style transfer, promoting its further application and development in areas such as computer vision, artistic creation, and image processing.

With the development of image style transfer techniques, the field of video style transfer has also shown immense potential. Video style transfer, as an extension of image style transfer, aims to blend the style of one video segment with another video or image, generating a video with a new style. Initially, video style transfer methods applied image style transfer to each frame, but this approach was computationally intensive, inefficient, and unsuitable for real-time video processing.

With the advancement of deep learning technology, convolutional neural network-based video style transfer methods have emerged. Examples include the Fast Style Transfer Network and Multi-Style Generative Network. These methods leverage convolutional neural networks to extract features from video frames and maintain the continuity and

stability of the video by considering the relationships between consecutive frames and temporal information, and thus achieving video style transfer.

Furthermore, there have been specific video style transfer methods tailored for particular application scenarios, such as 3D model-based video style transfer and semantic segmentation-based video style transfer, among others.

In 2022, Liu [14] provided an overview and summary of color transfer and style transfer in the video domain in their paper, introducing the development and various methods of video style transfer. However, they did not delve specifically into video style transfer in great detail. With the continuous advancement of technology, the application of deep learning in video style transfer has become more widespread. Therefore, there is a need for a more comprehensive summary of style transfer, including both image style transfer and video style transfer, to assist researchers and students working in the field of style transfer and to gain a more comprehensive understanding of the knowledge related to style transfer, providing an overview of this broad field.

Therefore, this article comprehensively updates and supplements the development of style transfer technology. The article sorts out machine learning-related technologies and uses artificial intelligence as a tool to solve application problems in the field of style transfer. In addition, the article classifies these from the perspective of deep learning and explores the progress of CNN, GAN, attention mechanism, the diffusion model, etc., in this field, focusing on improving the quality and effect of style transfer by improving the network architecture, loss function design, training strategy and evaluation method.

The paper divides style transfer into two main sections: image style transfer and video style transfer. Within image style transfer, it further distinguishes between traditional image style transfer and deep learning-based image style transfer. This segmentation aims to help readers better understand the directions of style transfer and grasp future development opportunities in this area.

In this paper, we make the following contributions to the state of the art:

1. This paper takes an application-oriented approach and builds on algorithms to provide a more comprehensive review of the development of style transfer.
2. It places special emphasis on video style transfer in the field of deep learning, offering detailed insights into style transfer in the context of videos.
3. The paper also updates the latest developments in video-style transfer technology.

In the Section 2, this paper introduces the work related to style transfer, including research on datasets, evaluation methods, and Citespace analysis. In the Section 3, traditional machine learning-based style transfer methods are discussed. While traditional style transfer methods perform well in experimental results, they can only modify low-level features; hence, there is a need to introduce deep learning-based image style transfer.

The Section 4 divides deep learning-based image style transfer into convolutional neural network-based style transfer, Generative Adversarial Network-based style transfer, and attention mechanism-based style transfer. These three techniques have played a significant role in advancing style transfer. Following the application of image style transfer techniques, deep learning technology further extends to the realm of videos. This paper focuses on video style transfer, introducing video style transfer technologies.

Finally, the paper summarizes the current research work and highlights future research directions in the field of style transfer.

2. Related Work

2.1. Related Datasets

In deep learning, the scale and quality of the dataset are crucial for the performance and effectiveness of the model. An excellent dataset should reflect the diversity and

variability of the real world and must be carefully curated and labeled to ensure its quality and effectiveness. Large-scale datasets like ImageNet [15], representing one of the world's largest visual recognition application datasets with over one million high-resolution images labeled, have become the standard benchmark dataset in the field of computer vision in recent years, driving research and development in image recognition and classification algorithms.Additionally, Tan [16] presented the WikiArt dataset, which comprises over 150,000 art paintings collected and curated by enthusiasts and volunteers.

Isola [17] used four paired image datasets for supervised training in pix2pix, ensuring that each input image has a corresponding target image. Additionally, Zhu [18] employed four style transfer datasets to validate their method and model's effectiveness in the CycleGAN paper. It is worth noting that the key innovation of CycleGAN is its ability to perform unpaired image translation, so the images in these datasets do not need to be forced to correspond one-to-one.

When performing style transfer, learning the textures of images is crucial, and this requires the use of large-scale image texture datasets. In 2005, Caputo [19] introduced the KTH-TIPS dataset, which consists of 11 material categories, with multiple samples in each category taken in different poses, under different lighting conditions, and at different scales. Subsequently, Sharan [20] used the FMD dataset in their research, which was contributed by users from the Flickr community. It includes various real-world material images, such as plastic, metal, leather, and fabric, with 1000 images in total covering 10 different objects, each having 100 images. This dataset can be used for training visual material classification models. Additionally, in 2009, Quattoni [21] employed the MIT-Indoor dataset to address indoor scene recognition, comprising 67 indoor categories with a total of 15,620 images. The number of images varies among different categories, but each category contains at least 100 images.

In 2014, Cimpoi [22] created a texture database known as the Describable Textures Dataset (DTD). This database consists of 5640 images categorized into 47 types of textures inspired by human perception, including wood, stone, fabric, carpet, glass, metal, and more. Each category contains 120 images. This database provides a rich set of texture samples and offers detailed descriptions of the textures, enabling researchers to conduct various texture-related tasks. However, sometimes the scale of these datasets may still not meet the requirements. Therefore, in 2015, Bell [23] created the Materials in Context (MINC) dataset, a new, large-scale, open material dataset comprising 3 million material samples. MINC is an order of magnitude larger than previous material datasets and is more diverse, making it valuable for material recognition and texture learning tasks.

COCO (Common Objects in Context) is a widely used computer vision dataset designed to advance research in image-understanding tasks. Provided by Microsoft Research, the COCO dataset contains over 330,000 images, with 200,000 of them annotated with object labels. The images in the dataset cover 80 different object categories, including people, animals, vehicles, and everyday items, and depict a wide range of environments and scenes. Its diverse image content, including humans, objects, and various scenes, makes it an ideal source of content images for style transfer. Researchers can combine real-world images from COCO with artistic style images to generate creative and artistic works through style transfer.

LAION-5B dataset [24] is a large-scale open multimodal dataset introduced by the LAION team in 2022, containing over 5 billion pairs of images and text. LAION also provides various subsets, including the LAION-Aesthetics subset, which consists of 120 million aesthetic images focused on high-quality and aesthetically pleasing visuals. By offering rich aesthetic ratings and related information, this dataset supports various image generation and analysis tasks. It provides high-quality and aesthetically pleasing images as references,

allowing trained style transfer models to generate images that not only match the desired style but also exhibit high aesthetic value. This enhances the visual quality and detail representation of generated images.

These datasets provide a wealth of diverse image examples, encompassing various styles, content, and features. They enrich the training data for models, enabling them to learn how to capture and transform information from different styles while preserving the key characteristics of the original images. Additionally, these datasets contribute to the evaluation and improvement of style transfer algorithms. By providing standard test images and target styles, researchers can quantify the performance of models and make comparisons. Therefore, datasets play an indispensable role in style transfer research, laying a solid foundation for creating higher-quality and more diverse style transfer results. The database information is provided in Table 1.

Table 1. Database description and address.

Database	Content	References
ImageNet	Over 1 million tagged high-resolution images.	[25]
WikiArt	Contains more than 150,000 art paintings.	[26]
Pix2pix	4 paired image data sets.	[27]
CycleGAN	4 non-required paired data sets.	[28]
KTH-TIPS	Multiple samples for each of the 11 material categories.	[29]
FMD	1000 pictures of ten objects; 100 pictures of each object.	[30]
MIT-Indoor	67 indoor categories; 15,620 images in total.	[31]
DTD	Composed of 5640 images, organized according to 47 textures inspired by human perception.	[32]
MINC	3 million material samples.	[33]
COCO	Contains more than 330,000 images, of which 200,000 are annotated with objects.	[34]
LAION-Aesthetics	Contains 120 M aesthetic samples.	[35]

2.2. Evaluation Method

The problem of style transfer has been around for several decades. Over the past few decades, many excellent algorithms have emerged, and while their development directions vary, each algorithm has, to varying degrees, propelled the rapid advancement of style transfer technology. Each algorithm should be fairly compared. However, there is currently no universally recognized style transfer evaluation standard. Traditional subjective visual assessment relies on direct human observation of the visual effects of images or videos to determine whether the desired effect has been achieved, but this evaluation lacks objectivity. Therefore, more objective data and methods are needed to assess the quality of style transfer. Various data and methods now exist to provide objective evaluation criteria.

In 2004, Wang [36] introduced the Structural Similarity Index (SSIM), a metric for evaluating image quality. It is designed to measure the structural similarity between two images, thus assessing the level of distortion in images, and it correlates reasonably well with human subjective perception. SSIM provides an accurate measure of image distortion, enabling a more precise evaluation of image processing algorithms. As a commonly used image quality assessment metric, SSIM calculates a score between 0 and 1 by comparing the similarity of two images in terms of brightness, contrast, and structure. A higher score indicates greater similarity between the two images.

Zhang [37], building upon SSIM, introduced a new evaluation metric called Colorization Error. This metric maps the generated images and original images to the Lab color space and calculates the mean squared error of the luminance channel (L) and the weighted sum of SSIM for the chrominance channels (a and b). This metric aims to provide a more comprehensive assessment of image quality, including both brightness and color information, offering a quantitative measure of image color accuracy.

Gatys [38] introduced two evaluation metrics: content loss and style loss. Content loss is used to measure the distance in feature space between the generated image and the original image, while style loss is used to measure the distance in feature space between the generated image and the reference image. These two metrics help quantify the similarity between the generated image and the original image as well as the reference image, providing quantitative criteria for evaluating style transfer algorithms.

Content Loss

$$L_{Content} = \sum (F^l - P^l)^2 \tag{1}$$

The process involves taking the input image and passing it through a VGG-19 network to obtain the feature maps at various layers, denoted as P (the feature map of layer l is represented as P^l). Simultaneously, through the same VGG-19 network, randomly generated white noise images also acquire feature maps at different layers, denoted as F (the feature map of layer l is represented as F^l). Therefore, the content loss for an image can be defined as the mean squared error (MSE) between the feature representations of the image and the target image at a specific layer.

Style Loss

$$L_{Style} = \sum_l w_l E_l \tag{2}$$

Error

$$E_l = sum(G^L - A^L)^2 \tag{3}$$

The feature maps from each layer are further used to calculate the Gram matrix, denoted as A, where the matrix for the l-th layer is represented as A^L, The Gram matrix A captures the style information from the style image. The feature maps F from various layers are also used to calculate the Gram matrix, denoted as G^L. The weights w_l represent the weights of the error E_l for each layer.

$L_{Content}$ and L_{Style} result

$$L_{total} = \alpha L_{Content} + \beta L_{Style} \tag{4}$$

Among them, α and β are used to balance the weight of the content and style. If the similarity places more emphasis on content, α can be set as the larger, and if the similarity places more emphasis on style, β can be set to be larger.

The Inception Score (IS) [39] can serve as an auxiliary evaluation metric in style transfer tasks to measure the clarity and diversity of generated images, especially when the generated images exhibit prominent style features, as it reflects the effectiveness of the transfer. It uses an Inception Network trained on the ImageNet dataset to classify the fake images generated by the evaluated model. Then, the metric measures the average Kullback–Leibler (KL) divergence between the marginal label distribution $p(y)$ and the conditional label distribution $p(y \mid x)$ based on the generated samples.

$$IS = \exp(E_x[D_{KL}(p(y \mid x) \| p(y))]) \tag{5}$$

$p(y \mid x)$ is the predicted label distribution for the image, and E_x represents the expectation over all generated images. $D_{KL}(p(y \mid x) \| p(y))$ is the KL divergence for each image. A lower IS value indicates that the generated data are more similar to real data, meaning higher realism.

Fréchet Inception Distance (FID) [40] is an improved IS evaluation metric, which mainly addresses the issue that IS does not consider real data. Compared to IS, FID has a more reasonable theoretical foundation and can better capture the distributional differences between generated and real images. The core idea of FID is to measure the distributional

difference between real and generated images in feature space using the Fréchet distance (also known as the 2-Wasserstein distance). The formula is as follows:

$$FID = \| \mu_{\text{real}} - \mu_{\text{fake}} \|^2 + \text{Tr}\left(\sum_{\text{real}} + \sum_{fake} -2\left(\sum_{\text{real}} \sum_{fake} \right)^{1/2} \right) \tag{6}$$

where $\| \mu_{\text{real}} - \mu_{\text{fake}} \|^2$ is the squared Euclidean distance between the means of the two distributions; Tr represents the trace of the matrix, which calculates the difference between the covariance matrices; $\left(\sum_{\text{real}} \sum_{fake} \right)^{1/2}$ is the "square root" of the covariance matrices, computed using matrix decomposition methods. The lower the FID value, the more similar the generated data are to the real data, indicating higher realism.

Learned Perceptual Image Patch Similarity (LPIPS) [41] is used to measure the difference between two images. LPIPS measures the content fidelity between a stylized image and its corresponding content image. The core idea is to assess the perceptual similarity of images by comparing their deep features rather than pixel-level differences. It uses deep convolutional neural networks, such as VGG, to extract features from the images and calculates the differences in these features at different layers. The formula is as follows:

$$\text{LPIPS}(x, y) = \sum_{l} \omega_l \cdot \| \varnothing_l(x) - \varnothing_l(y) \|_2^2 \tag{7}$$

where $\varnothing_l(x)$ and $\varnothing_l(y)$ are the feature representations of images x and y at layer l (extracted through a convolutional neural network), $\| \varnothing_l(x) - \varnothing_l(y) \|_2^2$ represents the L2 norm, i.e., the Euclidean distance, used to calculate the difference between feature vectors, and ω_l is the weighting coefficient for each layer's feature. These weights are learned during training to adjust the relative importance of features at different layers. A lower LPIPS value indicates that the two images are more similar, while a higher value indicates greater difference.

In style transfer evaluation, the style information of an image may influence the LPIPS score as it tends to focus on texture features. To reduce the impact of style information on the evaluation, Content Feature Structural Distance (CFSD) was proposed, which is a new distance metric that only considers the spatial correlation between image patches. The formula is as follows:

$$CFSD = \frac{1}{\hbar w} \sum_{i=1}^{hw} D_{KL}(S_i^c \| S_i^{cs}) \tag{8}$$

In this context, S_i^c represents the i-th element of the correlation matrix for the content image, and S_i^{cs} represents the i-th element of the correlation matrix for the style image. D_{KL} denotes the Kullback–Leibler divergence. By capturing the structural and detail differences between style-transferred images, CFSD can effectively assess whether the image successfully conveys the target style while maintaining a balance between style and content and preserving the structural features of the original image. This approach focuses on spatial correlations, making it more robust to variations in texture and style, which are common in style transfer tasks.

In recent years, text-to-image generation techniques have garnered increasing attention. The CLIP model (Contrastive Language-Image Pretraining) [42], introduced by OpenAI in 2021, is a multimodal model trained through contrastive learning on a large corpus of image–text pairs. It can understand and associate images with textual descriptions by encoding both into a unified high-dimensional space. The similarity between images and text can be measured by the distance in this space.

The CLIP model is applicable to style transfer tasks, assessing the effectiveness of a style transfer or image edit. It achieves this by calculating the distance between CLIP embeddings of the target style or edited image and the CLIP embedding of the original image. This evaluation method effectively measures how well the generated images match the description of the target style, facilitating optimization and helping to assess the accuracy and style consistency of style transfer results.

When it comes to video style transfer or dynamic image style transfer, two types of errors, temporal errors and warping errors, are commonly encountered. These are mainly used to measure visual quality issues caused by different factors during the time sequence or image transformation process. Temporal errors primarily involve consistency issues in the time dimension, especially in video style transfer tasks. When style transfer is applied to each frame in a video, differences in the style transfer between frames may occur. To measure the temporal consistency within a video sequence, temporal errors are usually calculated based on the similarity between adjacent frames.

$$L_{\text{temporal}} = \sum_{t} \| I_t - I_{t-1} \|_2^2 \tag{9}$$

I_t and I_{t-1} represent the style-transferred images at time points t and $t-1$, respectively. Temporal errors measure the difference between adjacent frames, with the goal of minimizing the visual differences between consecutive frames to ensure temporal consistency in the video.

Warping errors typically affect the transformation differences between images, especially during the process of image content alignment or spatial mapping. The formula is as follows:

$$L_{\text{warping}} = \sum_{t} \left\| T_t(I_t) - \hat{I}_{t+1} \right\|_2^2 \tag{10}$$

T_t represents the spatial transformation of image I_t at time t, and $\hat{I}_{t+1}$ is the target image of I_t at time $t+1$ after spatial transformation. Warping errors focus on the spatial transformation consistency of the image content. By minimizing warping errors, spatial consistency can be maintained during the image content transfer process, preventing visual distortion caused by inconsistent transformations.

Currently, there are various evaluation methods for style transfer, each focusing on different aspects. While significant progress has been made, there are still certain limitations. Pixel-based evaluation methods, such as Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR), are typically used for content preservation but fail to effectively assess the visual quality and artistic perception of images. They struggle to capture the high-level features of style transfer. Perceptual difference evaluation methods (such as LPIPS) measure the perceptual differences between images by extracting features through deep learning networks, providing an evaluation standard that aligns better with human visual perception. These methods have become mainstream for style transfer evaluations. However, perceptual evaluation methods are still dependent on specific network models and may be limited by the network architecture and training data. For video style transfer, temporal consistency and spatial consistency errors, as new evaluation dimensions, ensure image stability during dynamic changes. However, these evaluation methods are still in the exploratory stage and require further optimization and research. Additionally, human evaluation relies on subjective standards, providing intuitive feedback from the perspective of artistry and visual effects, but lacks repeatability and standardization. Overall, the current style transfer evaluation methods still have certain shortcomings, especially in terms of comprehensiveness and consistency. In the future, further integration of automated evaluations and human assessments will be needed to develop a more complete and

systematic evaluation framework.The summary of style transfer evaluation indicators is shown in Table 2.

Table 2. Summary of style transfer evaluation methods.

Evaluation Methods	Algorithms	Focus
SSIM	Measuring the structural similarity between two images to assess the degree of image distortion.	Evaluating the visual quality of generated images.
Colorization Error	Map the generated image and the original image to the Lab color space, calculate the mean square error of the brightness channel L and the weighted sum of the SSIM of the color channels a and b.	Measure the color difference between the generated image and the original image.
Content loss, Style loss	The content loss is used to measure the distance between the generated image and the original image in the feature space, while the style loss is used to measure the distance between the generated image and the reference image in the feature space.	Measuring content and style characteristics of generated images.
IS	The generated fake images are classified and the average KL divergence between the marginal label distribution p(y) and the conditional label distribution based on the generated samples p(y | x) is measured.	Measuring the clarity and diversity of generated images.
FID	The Frechet distance is used to measure the distribution difference between the real image and the generated image in the feature space.	Better capture of the distribution difference between generated images and real images
CLIP	Calculate the distance between the CLIP embedding of the target style or edited image and the CLIP embedding of the original image.	The model's semantic understanding and accuracy of the visual content of the generated image or video.
LPIPS	Evaluate the perceptual similarity of images by comparing their deep features rather than pixel-level differences	Content fidelity between stylized images and corresponding content images
CFSD	Distance metrics that take into account spatial correlation between image patches	Conveys the target style and preserves the structural features of the original image
Temporal error	Based on the similarity between adjacent frames	Temporal consistency in video sequences
Warping error	Differences in transformations between images, especially during image content alignment or spatial mapping	Spatial consistency in video sequences

2.3. CiteSpace Research

Style transfer has been used in extensive applications in the field of computer science. As shown in Figure 2, in the Web of Science Core Collection, this study conducted a keyword search using "style transfer" and found nearly a thousand related papers. Based on these data, a keyword heat map was generated. In the map, the size of the circles represents the frequency of keyword occurrences, while the layers of circles, from inner to outer, represent the progression of time from the past to the present. Lighter colors indicate more recent years, and the connecting lines represent the correlations between different keywords.

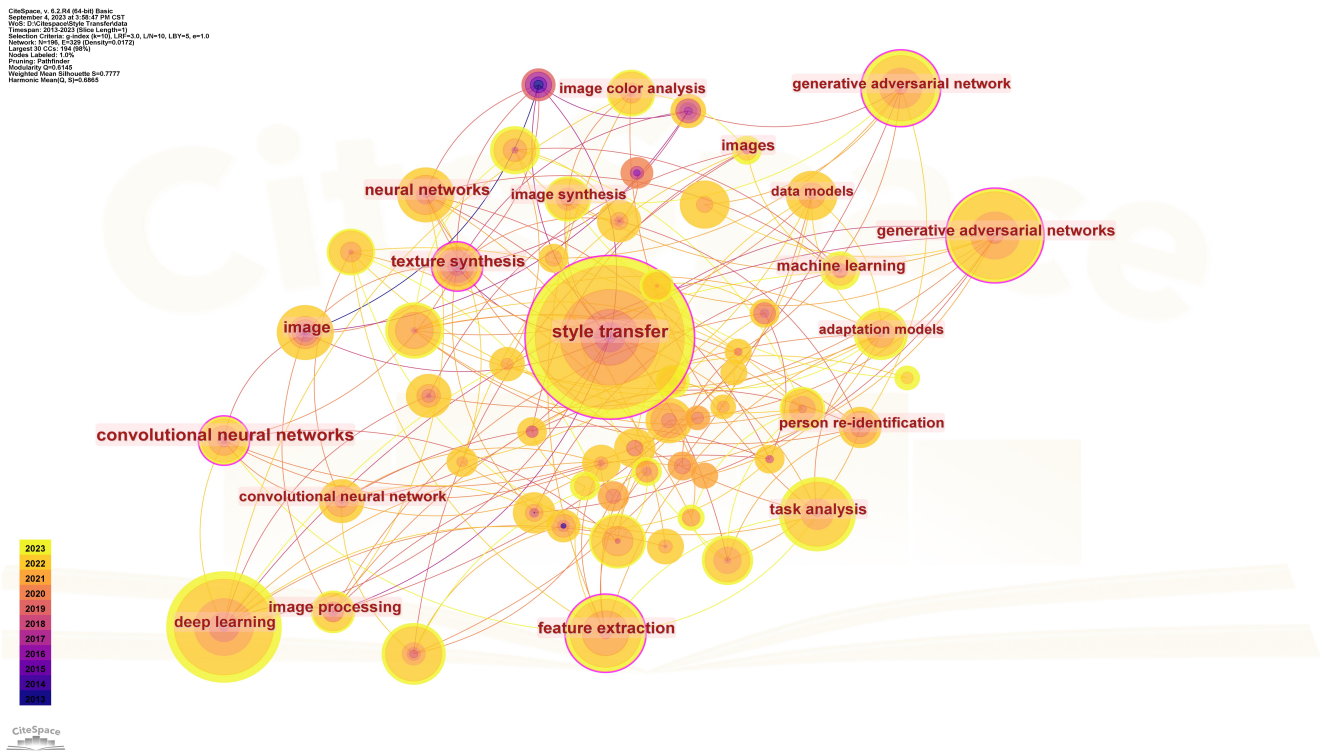

Figure 2. Hot words in the field of style transfer.

From Figure 2, it can be observed that machine learning played a significant role in the field of style transfer from 2013 to 2017. Machine learning methods such as texture analysis laid a solid foundation for early implementations of style transfer. Since 2016, with the advancement of deep learning, deep learning methods have gradually dominated the field of style transfer. Simultaneously, the field of style transfer has become increasingly popular, with convolutional neural networks and generative adversarial network algorithms playing increasingly important roles.

In addition to deep learning technologies, a variety of diverse techniques and methods have emerged in the style transfer field. This richness in methodologies enhances the options for implementation strategies in style transfer, not only advancing model performance and effectiveness but also promoting the widespread application of style transfer technology across multiple domains, such as artistic creation, image processing, and visual effects.

From the above figure, it is clear that there is a significant connection between the development of style transfer and deep learning. This connection is also evident in the trends of publication years observed in academic papers. The study investigated two databases, Web of Science and CPCI, which encompass over 10,000 journals from four major citation indexes (SCIE, SSCI, A&HCI, ESCI) and two chemical databases (CCR, IC). It also included records from over 225,000 international conferences spanning various disciplines such as the natural sciences, social sciences, and humanities.

The analysis counted the number of papers related to style transfer published in relevant journals and conferences. As depicted in Figure 3, the quantity of papers related to style transfer is rapidly increasing alongside the continuous development of the deep learning field. The reason for the decline in 2024 is that statistics for December are not available yet.

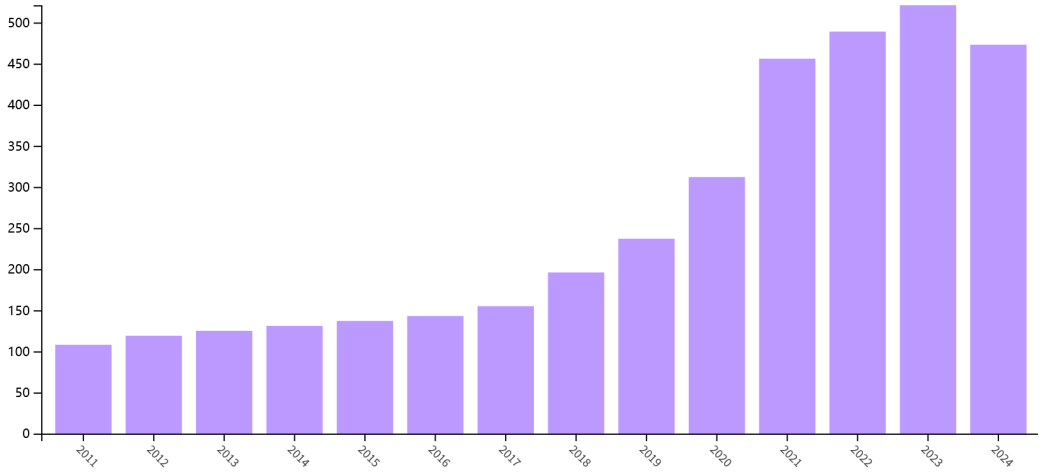

Figure 3. The number of papers on style transfer published in relevant journals and conferences.

Based on the aforementioned data, it is clear that with the advancement of deep learning, the field of style transfer has made unprecedented progress, reflecting its increasing importance and growth in both academic research and applications. This trend not only demonstrates the widespread attention to and active exploration of style transfer technology by researchers but also underscores its potential applications in fields such as image processing, artistic creation, and visual effects.

The increased number of publications also indicates that researchers are exploring new algorithms, improving existing methods, and applying innovations, thereby driving advancements in the field and maturing its technologies. The combination of deep learning and style transfer has become a highly regarded research area, and it is likely that it will continue to be a popular topic of interest.

The increase in the style transfer literature reflects the outstanding contributions of numerous scholars to the fields of style transfer and deep learning, driving continuous development in this area. This paper categorizes the progress of style transfer into different domains and showcases the contributions of various scholars in each domain. These contributions have transformed image style transfer from purely an area of research into one with practical application prospects, providing vast possibilities for creative and innovative image processing.

This section of the paper aims to enable interested readers to track the latest developments in this field by following the work of these leading scientists. Additionally, the paper draws on previous work to highlight current challenges in the field of style transfer and potential future directions.

3. Image Style Transfer Algorithm Based on Traditional Machine Learning

Before the rise of deep learning, traditional machine learning methods for image style transfer primarily relied on the construction of physical models and the synthesis of textures. These traditional methods conducted in-depth research on physical models and image textures. While traditional style transfer methods performed well in experiments, they had limitations in that they could only achieve changes in style at a low-level feature level. Furthermore, the manual effort required to construct these methods and the need for labeled datasets in supervised methods were incalculable.

3.1. Computer Graphics

In 1962, E. Sutherland at the Lincoln Laboratory of MIT coined the term "computer graphics" and demonstrated in his paper that interactive computer graphics was a feasible and useful research area [43,44]. This marked the recognition of computer graphics as an

independent branch of science. In the field of computer graphics, graphic style transfer falls under the category of non-photorealistic graphics (NPR), which can be roughly divided into three main methods: stroke-based rendering, image analogy-based methods, and image filtering-based methods.

3.1.1. Stroke Rendering

Stroke rendering is a technique used to render images or videos in a hand-drawn artistic style and has widespread applications in the field of image style transfer. This technique can closely mimic the characteristics of hand-drawn art, including the shape, color, and width of the strokes, to produce images that resemble hand-drawn artwork. While this method has relatively few computational resource requirements, the resulting image effects are often relatively simple and may struggle to achieve high-quality artistic style effects.

In 1998, Hertzmann [45] introduced an early stroke-based rendering algorithm, which was a method for simulating the effect of oil brushstrokes. This algorithm employed curved brushes of various sizes to achieve its effect. The key steps of this algorithm involved converting the input image into two channels representing color and brightness and then using brushes of different sizes along with their corresponding stroke directions to render the image.

Later, in 2002, he proposed a fast oil brushstroke rendering algorithm [46]. This algorithm involved preprocessing and storing the original image, allowing for the quick extraction and application of brushstroke textures during the rendering process. This approach combined fast multi-resolution image pyramids with local texture mapping techniques, resulting in an efficient oil painting rendering effect.

In 2003, DeCarlo [47] introduced a novel stroke-based rendering technique known as "Implicit Contours". This method simulated hand-drawn effects by extracting and enhancing suggestive contours within the image. This technique significantly improved the visual quality and expressiveness of images while also possessing robust real-time rendering capabilities.

In 2012, Lu [48] introduced a pencil sketch generation technique that combined line art and grayscale tones. This technology could automatically transform a color photograph into a pencil sketch while preserving the original image's tones and details. To further enhance the accuracy of style transfer, in 2018, Jing [49] demonstrated that the aspect ratio of the style image indeed affects the effectiveness of style transfer. To achieve a style transfer effect with mixed brushstrokes, they designed the network structure depicted in Figure 4.

The network architecture mainly consists of three modules. The stroke pyramid module aims to change the size of convolutional kernels, allowing each stroke branch to have a larger receptive field than the previous branch. It encourages consistency between adjacent stroke branches. The pre-encoder module provides layers that share weights between different stroke style branches, enabling the learning of semantic information from the content image and the basic representation of the style image. The stroke-decoder module decodes the stroke results into style results based on the corresponding stroke size. Through these structural strategies, the network in the paper can achieve continuous variations in stroke sizes and generate different stroke sizes in different spatial regions within the same output image.

To control different stroke sizes, the network is divided into multiple branches with a pyramid structure. Each branch is trained with a corresponding stroke size. Subsequently, a gating function ($\oplus$) is used for interpolation between the feature maps output by the different branches in the StrokePyramid. This allows for feature interpolation and the continuous control of stroke sizes, enabling smooth transitions between different stroke

scales in the generated styled images.By combining these modules, the network is able to generate fine-grained, adjustable brush strokes that are responsive to both the content and style information, facilitating a more flexible and detailed style transfer.

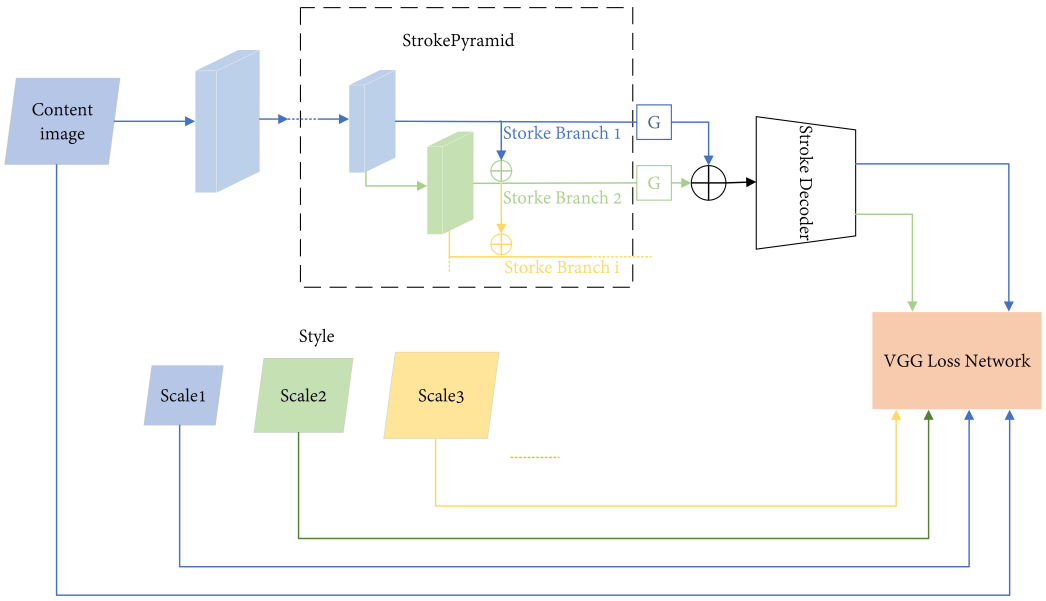

Figure 4. Mixed stroke network structure. Includes three modules: StrokePyramid module, pre-encoder module and stroke-decoder module.

Stroke-based rendering methods have several main drawbacks in style transfer. Firstly, they often have high computational costs and long running times, resulting in low efficiency. Secondly, since these methods rely on modeling specific painting styles, they may be limited by the model's constraints, reducing their versatility. Additionally, the output results of stroke-based rendering methods can be influenced by the quality of the input image, potentially leading to loss of details or inaccurate rendering. Finally, these methods typically require manual adjustments and parameter tuning, which may not be suitable for the needs of the average user. A summary of relevant articles is provided in Table 3.

Table 3. Summary of brushstroke rendering methods.

Author	Algorithm	Contribution
Hertzmann [45]	Simulate the stroke effect of oil painting based on a combination of curve brushes and brushes of various sizes.	Early stroke rendering algorithms.
Hertzmann [46]	Preprocess and store raw images to quickly extract and apply stroke textures at the render time.	Fast oil painting stroke rendering algorithm to achieve efficient rendering.
Decarlo [47]	Simulate a hand-drawn effect by extracting and enhancing suggestive contours in images.	Improve the visual quality and expressiveness of images, and have strong real-time rendering capabilities.
Lu [48]	Pencil drawing production technology that combines line drawing and grayscale tones.	Automatically convert a color photo into a pencil drawing while retaining the tones and details of the original image.
Jing [49]	The proposed network structure mainly has three parts, StrokePyramid module, pre-encoder module, and stroke-decoder module.	Achieve continuous changes in stroke size and produce different stroke sizes in different spatial regions within the same output image.

3.1.2. Image Analogy

Image analogy is a concept in image processing that involves applying the structural and texture characteristics of one image to another. In the context of style transfer, image analogy algorithms can be used to apply the style of a reference image to a target image, creating a new image that incorporates the stylistic elements of the reference image.

In the field of style transfer, the Image Analogies algorithm proposed by Hertzmann in 2001 [50] is a typical approach based on the concept of image analogy. This method involves comparing a pair of reference images, typically an input image and an output image, to find the similarities between them. Then, using this similarity information, it generates a new target image that is similar to the reference input image in both style and content. Furthermore, in 2001, Ashikhmin [51] introduced a texture synthesis method based on the Image Analogies algorithm. This approach involves aligning, comparing, transforming, and synthesizing input and target textures to achieve high-quality texture synthesis.

In 2001, Efros [52] introduced an image synthesis method based on the Image Analogies algorithm, known as "image quilting". This technique involves partitioning, selecting, and stitching operations on input and target images, resulting in high-quality image synthesis and style transfer. In 2004, Agarwala [53] presented an interactive digital montage technique that allows users to create high-quality image montages by selecting and arranging image elements. However, this method is primarily suitable for static images. In 2012, Darabi [54] introduced a novel image fusion method called "Image Melding" based on Digital Photomontage. Unlike traditional techniques, Image Melding applies not only to static images but also to video images. It can combine two or more incompatible images, producing incredible results.

In 2017, to keep pace with the advancements in deep learning, Liao [55] introduced a style transfer method based on the deep image analogy, which combines image analogy with feature extraction in deep convolutional neural networks. The core algorithm of this paper is the NNF algorithm (Nearest Neighbor Field). It starts by extracting features using VGG19, initializing A and B'. Then, a forward NNF is computed to establish mappings between them, including current layer upsampling as the initialization for the next layer. These mappings and extracted features are used to reconstruct the next layer's A' and B, eventually generating the images A' and B. The network structure is depicted in Figure 5.

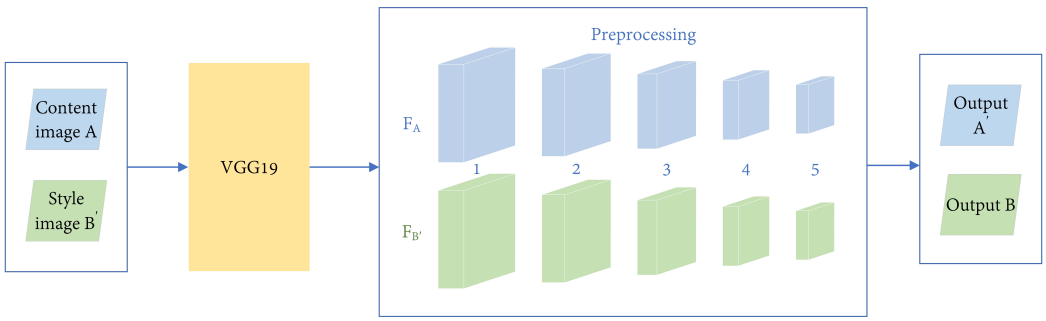

Figure 5. Deep image analogy style transfer using NNF.

While image analogy methods have many advantages in style transfer, such as controllability, flexibility, and efficiency, they also come with some drawbacks. Firstly, they require matched pairs of images and manually annotated matching points, which can be time-consuming and requires expertise. Secondly, this method is typically limited to specific application scenarios and may be challenging to apply to large-scale image datasets. Finally, because the approach relies on local matching, it may encounter issues when attempting to achieve global consistency in style transfer. A summary of relevant articles is given in Table 4.

Table 4. Summary of image analogy methods.

Author	Algorithm	Contribution
Hertzmann [50]	Compare a pair of reference images to find similarities between them. Use this similarity information to generate a new target image.	Proposed Image Analogies algorithm.
Ashikhmin [51]	High-quality texture synthesis is achieved by aligning, comparing, transforming, and synthesizing the input texture and the target texture.	Proposed texture synthesis method based on the Image Analogies algorithm.
Efros [52]	Perform operations such as the blocking, selection, and splicing of input images and target images to achieve high-quality image synthesis and style transfer.	An image synthesis method based on the Image Analogies algorithm is proposed, called "image collage".
Agarwala [53]	Select, delete, modify, and recombine multiple images to create a composite image.	An interactive digital collage technology is proposed that allows users to create high-quality image collages by selecting and arranging picture elements.
Darabi [54]	By rearranging patches to minimize discontinuities in the fusion result.	Not just for still images, but for video images as well.
Liao [55]	NNF algorithm first uses VGG19 to extract features in advance, initialize A' B, and use forward NNF to calculate the mapping between them.	Match image analogies with features extracted from deep convolutional neural networks.

3.1.3. Image Filtering

In style transfer, image filtering is a commonly used technique. It can be employed to smooth and blur images, reducing noise and details, resulting in a smoother style. In certain cases, image filtering can also be used to enhance features such as edges and textures in an image, leading to a more pronounced and clear style.

In 1987, Burt [56] introduced an image compression method based on Gaussian and Laplacian pyramids, where the Gaussian pyramid is constructed through successive downsampling and Gaussian filtering. This pyramid structure is effective in extracting features from images and can be applied in style transfer applications. However, the decomposition of the image into different scales using Gaussian pyramid filters may result in the loss of some fine details, leading to potential blurriness or distortion in the style transfer results.

Tomasi [57] introduced the method of bilateral filtering. Bilateral filtering is a nonlinear filter that takes into account both the spatial distance and grayscale differences between pixels. As a result, it effectively balances preserving image details while smoothing the image, making it widely applied in the field of image style transfer. In this paper, the author applied bilateral filtering to grayscale and color images and validated its effectiveness through image-denoising experiments.

The advantage of bilateral filtering is its ability to smooth while preserving edge information, avoiding the blurring that traditional filters may cause during the smoothing process. However, its disadvantage is that it can lead to higher computational complexity when dealing with large images. In 2006, Winnemöller [58] attempted to use bilateral filtering in combination with Gaussian difference filters to extract image contours, thus achieving the rendering of image contours.

Following the application of deep learning in the field of style transfer, image filtering was actively combined with it. In 2017, Semmo [59] integrated neural networks with image filtering to achieve style transfer functionality. This combination leverages the strengths of both artistic rendering paradigms: deep convolutional neural networks can transmit style features globally, while image filtering can simulate artistic media phenomena locally.

To address the issue of customizing filter styles, in 2020, Yim [60] introduced Filter Style Transfer (FST), with a network structure depicted in Figure 6.

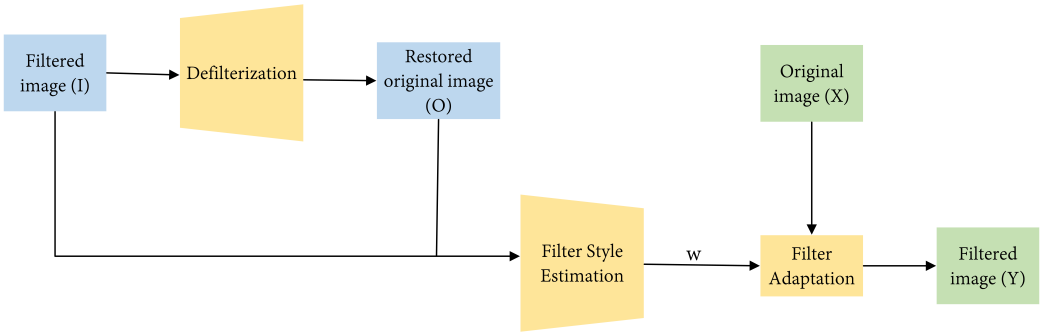

Figure 6. FST network structure. Extract custom filter style from filter style image and apply it to content image.

FST has the ability to extract custom filter styles from a reference style image and apply them to a content image. First, a custom photo filter is extracted from a single reference image (I) and applied to a new reference image (X). FST recovers the original image (O) from I, which is referred to as defiltering in the paper. Then, using both images, the filter parameters w are obtained, which is referred to as filter style estimation. Finally, the filter function designed using w can be used to filter the user's original image (X) into a newly stylized image (Y).

In style transfer, image filtering is often used to remove high-frequency noise and details from images, reducing artifacts and textures to obtain cleaner and smoother images. This helps improve the stability and effectiveness of style transfer algorithms, making the transferred images appear more realistic and natural while preserving the original image's structure and semantic information. However, traditional image filtering algorithms are typically based on fixed filter kernels, making it challenging to find a universal filter kernel that works well for all style transfer tasks. Today, image filtering is more frequently combined with deep learning methods to develop new image-filtering techniques and applications. A summary of relevant articles is given in Table 5.

Table 5. Summary of image filtering methods.

Author	Algorithm	Contribution
Burt [56]	An image compression method based on Gaussian and Laplacian pyramids is proposed, where the Gaussian pyramid is implemented by stepwise downsampling and Gaussian filtering.	The pyramid structure can effectively extract features in images and can be used in style transfer applications.
Tomasi [57]	Proposed the method of Bilateral Filtering.	Avoids the shortcomings of traditional filters that blur the image during the smoothing process.
Winnemöller [58]	Extract image contours using a bilateral filter combined with a Gaussian difference filter.	Render image outlines.
Semmo [59]	Combining neural networks with image filtering to achieve style transfer.	Has the advantages of both artistic rendering paradigms.
Yim [60]	Extract custom filter styles from filter style images and migrate them to content images.	Solve the problem of custom filtering style.

3.2. Texture Transfer

Texture transfer is the process of applying the texture from one image to another image. Texture can be thought of as the local visual patterns within an image, typically composed

of repeating elements such as patterns, shapes, and lines. Texture transfer can further be categorized into single-style texture transfer and multi-style texture transfer.

3.2.1. Single-Style Texture Transfer

Single-style texture transfer refers to the process of applying the texture from one image to another to make the output image have the same texture style as the input image.

In 1995, Heeger [2] introduced a pyramid-based texture analysis/synthesis method for generating new texture images from a single texture sample. In 2005, Kwatra [61] proposed an example-based synthesis algorithm designed to generate new textures by optimizing input example textures. This method decomposes the input example textures into disjoint patches and then reorganizes them to create new textures.

The aforementioned methods were based on traditional signal processing techniques, requiring manual parameter tuning and feature computation. While capable of generating texture images, they were computationally intensive, operationally complex, and challenging to apply in practical scenarios. With the advancement of deep learning technology, texture transfer methods based on deep neural networks have gradually gained popularity.

In 2017, Elad [62] introduced a new method based on texture synthesis called STTS (Style Transfer via Texture Synthesis). This approach leverages texture synthesis techniques to combine the style of a given style image with the content information of a target image, resulting in a new image that possesses both style and content characteristics. However, since it relies on texture synthesis, it may not perform well when dealing with non-textured images, leading to the generation of images with pronounced textures. It might also suffer from overfitting, where the generated images overly depend on the style image and lack unique features.

While single-style texture transfer can produce excellent results in certain contexts, it also comes with several significant limitations. The most prominent drawback is its lack of diversity and flexibility, as it cannot simultaneously incorporate multiple styles or offer a higher degree of customization. Additionally, this method demands high standards in the selection and quality of style images, and some complex textures and styles may be challenging to accurately capture and synthesize. Therefore, in certain specialized application scenarios, more advanced texture transfer methods may be required. A summary of relevant articles is given in Table 6.

Table 6. Summary of single-style texture transfer methods.

Author	Algorithm	Contribution
Heeger [2]	Use the Gaussian pyramid to decompose the input image, perform texture analysis, and synthesize a new texture image.	A pyramid-based texture analysis/synthesis method is proposed.
Kwatra [61]	Break the input sample texture into small, disjoint tiles and rearrange them to create a new texture.	An example-based synthesis algorithm is proposed that aims to generate new textures by optimizing input sample textures.
Elad [62]	Generate a new image with style and content information using the given style image and the content information of the image to be processed.	Proposed a new method based on texture synthesis called STTS.

3.2.2. Multi-Style Texture Transfer

Multi-style texture transfer refers to the process of applying the texture features from multiple style images to a single target image, generating a new image with diverse texture styles. Unlike single-style style transfer, multi-style texture transfer combines texture features from different styles into one target image, resulting in a texture effect that is more

diverse and rich. The significance of multi-style texture transfer lies in its extension of traditional single-style style transfer methods, allowing users greater freedom to choose multiple styles for generating texture effects.

In 2001, Efros [52] introduced a method that decomposes the texture of an input image into multiple small texture patches and then finds the best-matching patches on an output image, applying them to create the output image. By assembling texture patches, this approach achieves natural texture synthesis and transfer, and is capable of handling various scales and orientations of texture variations. However, during the process of patch assembly, unnatural seam artifacts can occur, resulting in less realistic texture synthesis and transfer effects.

In 2009, Barnes [63] introduced a randomized algorithm called PatchMatch. The core idea behind PatchMatch is to optimize global matching using local matching. It divides an image into multiple blocks and then matches each block with other blocks in the image, ultimately establishing correspondences between the blocks. When multiple texture styles are required for texture synthesis, the PatchMatch algorithm can be used to find the best-matching style for each pixel and perform cross-synthesis between matching styles, resulting in a new image with multiple texture styles.

In 2017, Song [64] introduced a comprehensive image appearance transfer method that combines color and texture. The workflow of this method involves the following steps: first, initializing the image using color information, then performing texture synthesis using a texture-based approach to obtain an image with synthesized textures. Finally, the synthesized textures are combined with the color information from the original image to achieve the ultimate appearance transfer result.

As time has progressed and technology continued to advance, an increasing number of researchers have turned their attention to multi-scale texture transfer methods based on deep learning. Currently, deep learning-based multi-scale texture transfer algorithms have become one of the mainstream approaches in this field.

In 2017, Huang [65] introduced a multi-style texture transfer algorithm capable of achieving real-time texture transfer in scenarios with high-speed requirements. Then, in 2018, Men [66] proposed a general solution for interactive texture transfer problems, which is more suitable for transfer environments with various textures. The core idea of the universal framework in this paper is to use multiple customized channels to dynamically guide the synthesis process, as depicted in Figure 7.

In the paper, the spatial distribution of stylized textures is controlled through semantic channels. The automatic extraction and propagation of structural guidance information obtained through two stages provide prior knowledge for initialization. This information is used to search for the nearest-neighbor fields (NNF) with structural consistency, preserving significant structural features. Additionally, texture consistency is utilized to maintain styles similar to the source image. Ultimately, the method presented in the paper demonstrates efficiency and an outstanding performance in various scenarios.

The evolution of multi-style texture transfer has gone through various stages, including those based on texture patches and neural networks, gradually improving the quality and efficiency of texture synthesis. While existing algorithms have achieved significant success, there are still some challenges and issues to address, such as the inability to handle large-scale images and the naturalness of the synthesis results. Further research and exploration are needed to address these challenges and continue advancing the field. A summary of relevant articles is given in Table 7.

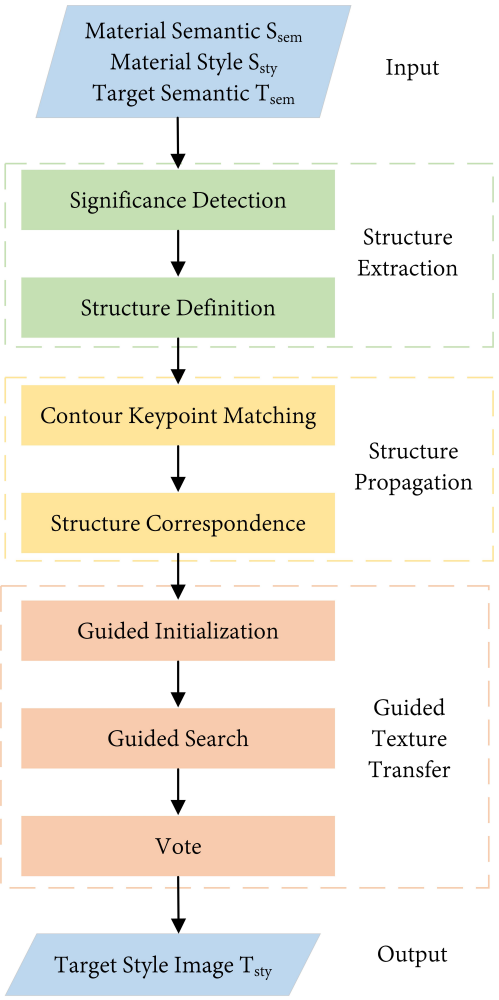

Figure 7. Multi-channel implementation of arbitrary texture migration flow chart.

Table 7. Summary of multi-style texture transfer methods.

Author	Algorithm	Contribution
Efros [52]	The texture of the input image is broken down into several small texture blocks, and then the best matching block is found on the output image.	Texture synthesis and transfer are achieved by assembling texture patches.
Barnes [63]	The local matching is used to optimize the global matching, the image is divided into multiple blocks, and the corresponding relationship between the blocks is finally established.	Introduced a randomized algorithm called PatchMatch.
Song [64]	The color information is used to initialize the image and then texture synthesis is performed using a texture-based approach.	A comprehensive image appearance transfer method is introduced that combines color and texture.
Huang [65]	Generative Adversarial Network (GAN)-based multi-style texture transfer algorithm.	Achieves real-time texture transfer in scenarios with high-speed requirements.
Men [66]	Utilize multiple customized channels to dynamically guide the synthesis process.	A general solution for interactive texture transfer problems.

4. Image Style Transfer Algorithm Based on Deep Learning

In recent years, deep learning has found widespread application in image processing due to its powerful performance and learning capabilities. Deep learning-based style transfer is a technique that transforms the content and style of an image using deep neural networks. It leverages pre-trained convolutional neural networks (CNN) or generative

adversarial networks (GAN) to combine the content of one image with the style of another, generating synthetic images with unique styles. Depending on the specific deep learning network model, these methods can be categorized into convolutional neural networks (CNN), generative adversarial networks (GAN), and attention-based methods.

4.1. Convolutional Neural Network Method

The pioneering work of deep learning in style transfer was introduced by Gatys in 2016 [38]. This work is considered a significant breakthrough in the field of deep learning for style transfer and laid the foundation for subsequent research and methods. This approach uses convolutional neural networks (CNN) to achieve style transfer in images. It combines features representing both image content and style and generates synthetic images through optimization algorithms. In neural network-based style transfer, there are distinctions between single-network neural style transfer and multi-network neural style transfer depending on the network structure.

4.1.1. Single-Network Neural Style Transfer

The term "single-network neural architecture" refers to the use of only one neural network for the entire style transfer process without involving additional networks. This network plays the dual role of both a content feature extractor and style feature extractor and is responsible for combining content and style features to generate the final style transfer result. The single-network neural architecture consists of a convolutional neural network (CNN), typically composed of multiple convolutional layers, pooling layers, and fully connected layers, for learning the feature representation of the input image. In style transfer, this single network generates an output image with the target style based on the content and style of the input image.

1. VGG network

In 2014, Simonyan [67] introduced a deep convolutional neural network architecture known as the VGG network, which was designed for image classification and recognition tasks on large-scale image datasets. The VGG network achieves an increased depth by stacking multiple convolutional layers and pooling layers while using small-sized convolution kernels and smaller strides to maintain high-resolution feature representations. Due to its outstanding performance in image processing, the VGG network was quickly adopted for style transfer applications.

The primary reason for using the VGG network lies in its powerful feature expression capabilities, allowing it to extract multi-level feature representations that capture both content and style information from images. Therefore, the VGG network has become a commonly used feature extractor and foundational network structure in the field of style transfer.

Gatys [38] took a significant step in style transfer applications by utilizing the VGG network and introducing the network structure depicted in Figure 8, which has been widely disseminated and employed. In this network, the input consists of a base image composed of noise, while the constraints are Style Loss and Content Loss, used to optimize the style of the style image and the content of the target image, respectively. The workflow of the network involves starting with a base image made up of random noise and then iteratively updating this base image by calculating Style Loss and Content Loss. This process aims to make the base image increasingly similar to the Style Image in terms of style texture while maintaining its similarity with the original photograph in terms of content.

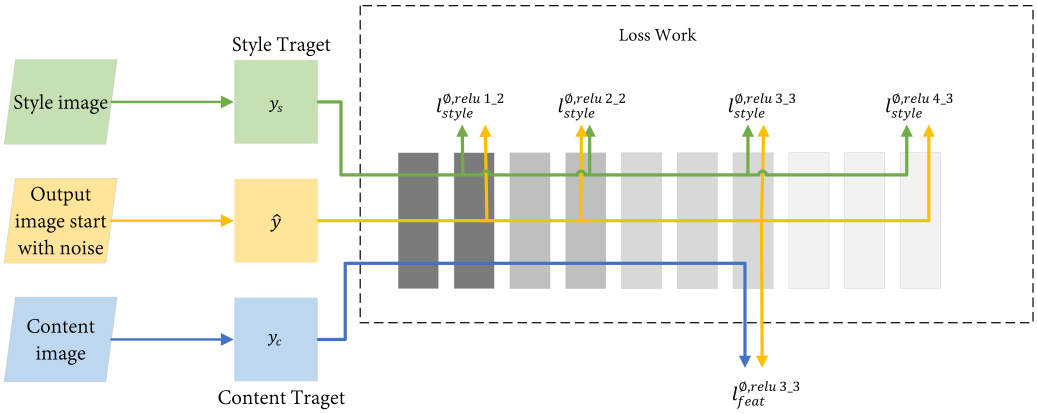

Figure 8. Style transfer through VGG. Input a base map composed of random noise, and continuously update the base map iteratively by calculating Style Loss and Content Loss to make it similar to the Style Image in style and texture, and similar to the original photo in content.

After Gatys [38] introduced the VGG network into style transfer, the application of the VGG network gradually expanded. Initially, the VGG network was used as a feature extractor in style transfer methods to extract content and style features from images. In 2016, Gatys [68] further improved the style transfer method based on the VGG network, addressing the issue of color distortion in the original method. Subsequently, Li [6] combined the traditional Markov Random Field with the VGG style transfer method, proposing an image synthesis method based on Markov Random Field. This method not only preserves the content and style of the image but also considers global consistency, thereby enhancing the quality of the synthesized image.

In the same year, researchers continued to make improvements to the VGG network. Novak [69] enhanced traditional style transfer methods to improve the quality and efficiency of synthesized images. Traditional methods often lose some detailed information while maintaining global consistency during style transfer. This paper modified the style representation to capture more information, imposed stricter constraints on style transfer results, and generated more detailed and realistic synthesized images.

In 2022, Tan [70] introduced the Faster-CNN model, which aimed to reduce storage space and computational costs. By compressing the CNN model and employing matrix factorization methods, including Singular Value Decomposition (SVD) algorithms, it not only significantly reduced the time required for image feature recognition, improving efficiency, but also optimized the model's recognition performance, better supporting the artistic style transfer of painted images.

With the continuous advancement of technology, the scope of application of VGG networks has expanded. In 2017, Luan [71] modified the VGG network and successfully achieved high-quality photo style transfer by combining deep learning techniques with local adjustments. In 2020, Kolkin [72] argued that explicit content loss was unnecessary and proposed Neural Neighbor Style Transfer (NNST)—a direct method based on nearest neighbors that achieved higher-quality stylization than previous works without sacrificing content preservation. Good transfer results were produced without using content loss, as shown in Figure 9.

Furthermore, in 2022, Wang [73] presented the application of VGG networks in cartoon-style transfer. VGG networks play a crucial role in cartoon-style transfer due to their image feature extraction capabilities. These applications demonstrate the versatility and applicability of VGG networks in various domains.

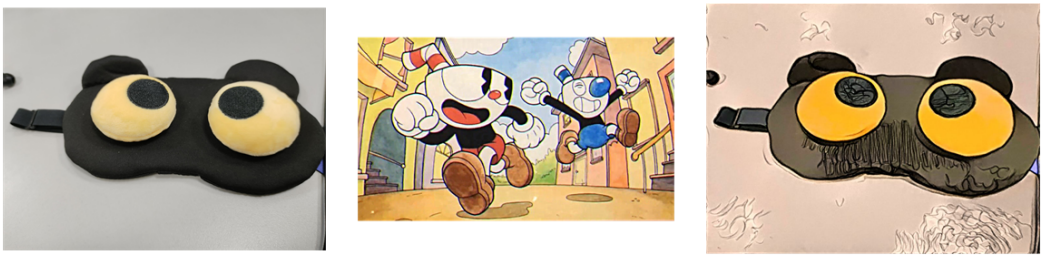

Figure 9. NNST effect diagram.

As mentioned earlier, the VGG single-network architecture is a commonly used convolutional neural network architecture in style transfer, employed for extracting feature representations from images. The straightforward structure and powerful feature extraction capabilities of the VGG single-network have established it as a classic model in the field of style transfer, providing an effective approach for image style transformation. A summary of relevant articles is shown in Table 8.

Table 8. Summary of VGG network style transfer method.

Author	Algorithm	Contribution
Gatys [38]	Applying the VGG network to style transfer.	A breakthrough in style transfer in the field of deep learning.
Gatys [68]	Based on the VGG network, the style transfer method is further improved.	Solved the problem of color distortion in the original method.
Li [6]	Combining the traditional Markov random field and VGG style transfer method.	Retain the content and style of the image while maintaining overall consistency.
Novak [69]	Modify the style representation to impose tighter constraints on the results of style migration.	Generate more detailed and realistic composite images.
Tan [70]	Proposed Faster-CNN model.	The process of image feature recognition is greatly reduced and the efficiency is improved.
Luan [71]	Deep learning technology combined with local adjustment.	High-quality photo style transfer achieved.
Kolkin [72]	a direct method based on nearest neighbors	achieved higher-quality stylization than previous works without sacrificing content preservation
Wang [73]	VGG networks play a crucial role in cartoon-style transfer due to their image feature extraction capabilities.	Introducing the application of the VGG network in cartoon-style transfer.

2. ResNet network

The ResNet network, introduced by Kaiming He [74] in 2015, is a deep residual network designed to address the issues of gradient vanishing and network depth. It introduces residual connections, allowing for the training of much deeper networks. In style transfer, ResNet networks are often used as feature extractors to capture the content features of images. However, directly applying the ResNet architecture to style transfer often results in the style of the original image being prominently preserved with little style transfer. Is there a method to adapt the ResNet architecture for style transfer with appropriate adjustments?

In 2019, Ilyas A [75] demonstrated that adversarial examples can be directly attributed to the presence of non-robust features—these features exhibit high predictability but are fragile and not interpretable by humans. It was speculated that VGG, which is unable to capture non-robust features in images as ResNet does, may perform worse in image classification accuracy but better in style transfer tasks. Subsequently, Nakano R [76]

conducted experiments and found that, in cases where the network architecture is identical, the code for performing style transfer is identical, and only the weights differ; robust ResNet models show significant improvements compared to regular ResNet models. In 2021, Wang [77] delved further into this issue and proposed a simple yet effective solution to enhance the robustness of the ResNet architecture: a softmax transformation based on feature activations to increase feature entropy. The network structure is illustrated in Figure 10.

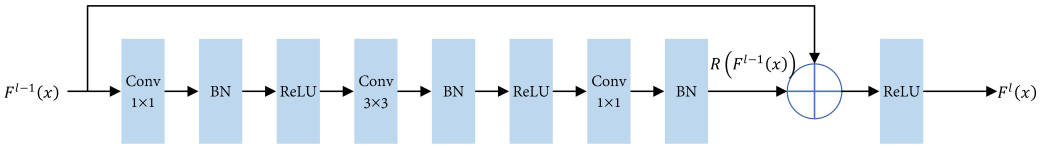

Figure 10. Stylization via ResNet.

The experimental results demonstrate that such a minor modification can significantly enhance the quality of stylized results, even for networks with randomly initialized weights.

Therefore, ResNet has found widespread applications in the field of style transfer. In 2019, Li [78] introduced a method for fusing infrared and visible light images based on ResNet and ZCA. This approach combines the feature extraction capabilities of deep learning with image processing techniques, improving image quality and enhancing object recognition performance. In 2023, Wang [79] applied ResNet to the development of willow pattern recognition, conducting image recognition experiments using Convolutional Neural Networks (CNN) ResNet and applying it to a Funan willow pattern.

The advantage of using a single ResNet network for style transfer lies in its powerful feature extraction capabilities and excellent transfer results. It can handle larger-sized images and, to some extent, mitigate the vanishing gradient problem. However, the limitations of the single ResNet network are revealed when dealing with complex styles and multi-style scenarios, where its performance may not be as strong as methods employing multiple network structures. Additionally, it requires a substantial amount of training data and computational resources to ensure high-quality transfer results. Therefore, in the field of style transfer, the single ResNet network can be considered a simple and effective approach, but for more complex tasks, it may benefit from being combined with other techniques or network structures to enhance transfer performance. A summary of relevant articles is shown in Table 9.

Table 9. Summary of ResNet network style transfer method.

Author	Algorithm	Contribution
He K [74]	Introducing residual connections to solve the problems of gradient disappearance and network depth.	Allows the network to be trained in more depth.
Ilyas A [75]	Exploring non-robust characteristics.	Demonstrate that adversarial examples can be directly attributed to the presence of non-robust features.
Wang P [77]	The softmax transformation based on feature activation enhances the entropy of Resnet.	Greatly improve the quality of Resnet stylized results.
Wang T [79]	ResNet image recognition experiment applied to Funan wicker pattern.	Provides a more comprehensive and rich set of new materials for the creation of wicker patterns.
Nakano R [76]	Robust ResNet model compared to conventional ResNet model.	The robust ResNet model shows a huge improvement.
Li H [78]	Infrared and visible light image fusion method of ResNet and ZCA.	Improve image quality and enhance target recognition performance.

4.1.2. Fast Style Transfer from Pre-Trained Networks

Single networks typically suffice for simple style transfers. However, as the demand for real-time and diverse styles grows, pre-trained networks have emerged. Fast-style transfer with pre-trained networks is a specific approach. It involves training a forward network so that an image can be reconstructed in a single forward pass, followed by optimization under various constraints. The effectiveness of this approach depends on how many different styles the pre-trained network has learned as classification criteria. Therefore, based on the number of styles a pre-trained network can handle, these algorithms can be further categorized into single-style, multi-style, and arbitrary-style fast-style transfer algorithms [11].

1. Single-style fast transfer

The single-style fast-style transfer method aims to quickly transform an input image into a specific single style. The core idea is to train a separate forward network for each style image. In this way, during testing, you only need to input the image into the corresponding model for a single forward pass to obtain the output result.

In 2016, Johnson [80] introduced a loss function called "Perceptual Losses" and a pre-trained forward network to achieve real-time style transfer and super-resolution reconstruction. This network uses a pre-trained VGG network as the feature extractor and optimizes the generated image by minimizing the perceptual loss, making it closer to the target image in terms of perception. The network's performance is shown in Figure 11.

Figure 11. Fast style transfer effect diagram.

Compared to the model proposed by Gatys, the key difference in this method is that the input is no longer a noisy base image but the target image itself. Additionally, it includes a forward network in the form of an autoencoder to fit the style transfer process. Other core ideas remain consistent: the loss function is still extracted by another neural network (usually VGG), including the content loss and style loss, collectively referred to as perceptual loss, and texture modeling is represented using Gram matrices. The network structure is shown in Figure 12.

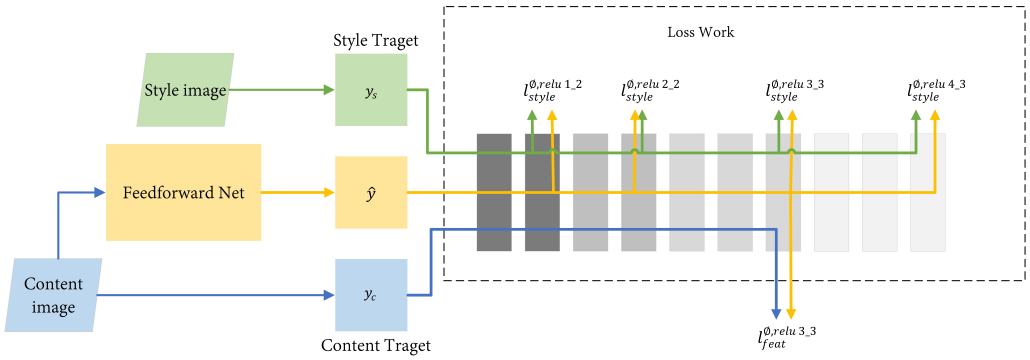

Figure 12. Fast style transfer network. A forward network in the form of an autoencoder is added to the VGG network to fit the style transfer process and achieve real-time purposes.

In 2016, Ulyanov [81] adopted a feedforward neural network approach that could directly synthesize textures and stylized images. Similar to Johnson's work, the core idea of these two methods is that they both use a feedforward network to learn styles. However, the difference lies in the fact that this method employs a multi-layer convolutional neural network (CNN) as the generator. It generates synthesized texture images by learning the feature representations of input texture images.

In 2016, Li [82] proposed the method of Adaptive Batch Normalization (AdaBN), which reduces interference from the source domain and better adapts to the data in the target domain, thus improving the effectiveness of style transfer. In 2017, Ulyanov [83] made two key improvements to Texture Networks: Instance Normalization and Spatially Adaptive Denormalization. Instance Normalization is applied to each channel of the generator network to normalize the statistical distribution of generated features. Spatially Adaptive Denormalization, on the other hand, introduces a parameterized denormalization operation that allows the generator network to dynamically denormalize at each pixel position based on the input image's features. As a result, Improved Texture Networks can generate more realistic, diverse, and detailed synthetic images.

With the advancement of technology, single-style transfer methods have also become more diverse. In 2019, Li [82] introduced a fast style transfer method based on linear transformations. This method achieves faster transfer speeds and a reduced computational complexity by learning linear transformations to convert the style transfer for images and videos into linear operations. While maintaining transfer quality, this approach significantly improves speed and efficiency.

The advantages of single-style fast transfer include its speed, simplicity of implementation, and ability to achieve real-time style transfer effects. However, it also has some limitations, such as a lack of diversity (limited to a single style transfer), potential constraints imposed by pre-trained models making it challenging to achieve fine-grained control, and issues like texture distortion and content loss that still need to be addressed. Therefore, future research directions are focused on enhancing the diversity, controllability, and quality of style transfer to better meet the needs of various application scenarios. A summary of relevant articles is shown in Table 10.

Table 10. Summary of single style fast transfer methods.

Author	Algorithm	Contribution
Johnson J [80]	A forward network is added to fit the style transfer process.	Real-time style transfer.
Ulyanov [81]	A multi-layer convolutional neural network (CNN) is used as the generator.	Directly composite textured and stylized images.
Li [82]	Proposed the method of Adaptive Batch Normalization (AdaBN).	Reduce interference in the source domain and better adapt to data in the target domain, improving the effect of transfer learning.
Ulyanov [83]	Two key improvements are introduced to address the shortcomings of Texture networks.	Ability to generate more realistic, diverse, and detailed synthetic images.
Li [82]	A fast style transfer method based on linear transformation is proposed.	Dramatically improve speed and efficiency while maintaining migration quality.

2. Multiple-style fast transfer

Single-style transfer involves applying a single, specific style to an image. In contrast, multi-style transfer, as compared to single-style transfer, is capable of handling features from multiple styles. Multi-style transfer refers to the ability to simultaneously process

various styles within a single model. This means that the model, after being trained, can handle features from multiple styles and apply these styles to style transfer for images. Multi-style transfer allows you to choose different styles for transfer or blend multiple styles to generate images with elements from various styles. The underlying concept behind multi-style transfer is to use a single network to accomplish the transfer of multiple styles.

Initially, people attempted to achieve multi-style fast transfer by using different parts of a single network. In 2016, Dumoulin [84] proposed an approach where multiple styles share one model. Dumoulin's method drew inspiration from previous research on fast-style transfer network architectures [80,81]. By studying the normalization operations within style transfer networks, they discovered that all convolutional weights of a style transfer network could be shared across many different styles. After normalization, only affine transformation parameters needed to be adjusted for each style. They named this method Conditional Instance Normalization (CIN). This means that by associating a small set of affine transformation parameters in the CIN layer with each style, and training only these parameters for each new style, it is possible to achieve multi-style transfer within a single network.

The implementation formula for CIN is as follows: building on Ulyanov's earlier work, selectrows in the γ and β matrices that correspond to a specific style 's' to obtain γ and β parameters, making them N × C matrices, where N is the number of styles to be modeled, and C is the number of output feature maps.

$$z = \gamma_s((x - \mu)/\sigma) + \beta_s \tag{11}$$

In 2017, Chen [85] achieved multi-style fast transfer by utilizing different parts of a single network. They introduced a new StyleBank module to explicitly model styles. In the StyleBank layer, they represented and classified input styles using StyleBank. Each StyleBank represents a specific style. For example, if there are 50 different styles, there would be 50 StyleBanks, and each StyleBank contains several convolutional kernels, with each representing a texture element. These texture elements perform convolution operations with feature maps. During the style transfer process, by selecting appropriate convolutional kernels within the StyleBank module, they could transform the style of the input image into the desired target style, enabling multi-style transfer. The network architecture is illustrated in Figure 13.

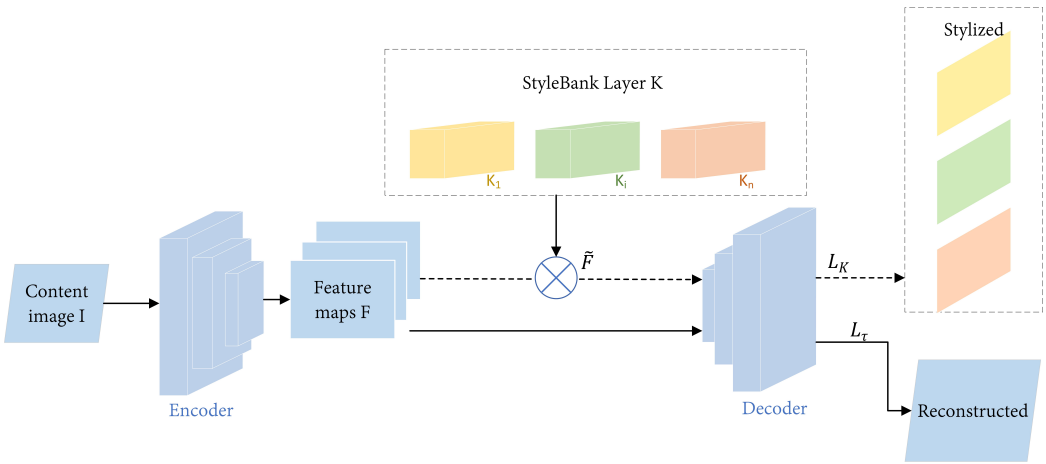

Figure 13. StyleBank fast multi-style migration network. Each StyleBank represents a style. Selecting the appropriate convolution kernel in the StyleBank module can convert the style of the input image into the target style.

As shown in Figure 13, the network consists of two branches: an autoencoder and a stylization branch, which are trained alternately. $\otimes$ represents the StyleBank Layer. In this layer, StyleBank convolves with the feature maps obtained from the Encoder applied to the input image to generate style-transformed features, which are then fed into the decoder to obtain the stylized result. Therefore, it is necessary to define two separate loss functions for each branch.

For the autoencoder branch, we measure the loss using the Mean Squared Error (MSE) between the input image I and the output image O:

$$L_\tau(I, O) = \|O - I\|^2 \tag{12}$$

In the stylization branch, we use the perceptual loss L_K, which consists of the content loss L_C, style loss L_S, and total variation regularization loss $L_{tv}(O_i)$:

$$L_K(I, S_i, O_i) = \alpha L_C(O_i, I) + \beta L_S(O_i, S_i) + \gamma L_{tv}(O_i) \tag{13}$$

$$L_C(O_i, I) = \sum_{l \in \{l_c\}} \left\| F^l(O_i) - F^l(I) \right\|^2 \tag{14}$$

$$L_S(O_i, S_i) = \sum_{l \in \{l_s\}} \left\| G_i\left(F^l(O_i)\right) - G_i\left(F^l(S_i)\right) \right\|^2 \tag{15}$$

where I, S_i, and O_i are the input content image, style image, and stylized result (for the i-th style), respectively. $L_{tv}(O_i)$ denotes the total variation regularization loss used. F^l and G_i represent the feature maps and Gram matrices computed from the VGG-16 network. l_c and l_s are used to compute content loss and style loss, respectively, at VGG-16 layers.

In 2017, Wang [86] introduced a multi-modal training strategy. This strategy involves joint training of multiple subnetworks and utilizes various style images to enhance the network's generalization ability and style transfer performance. In 2019, Huang [87] introduced the application of semantic awareness in multi-style transfer. They proposed a semantic-aware multi-style transfer network called Style Mixer, which is used to achieve multi-style transformations of images. These methods are based on unsupervised learning. Then, in 2021, Kim [88] introduced a pseudo-supervised learning framework for semantic multi-style transfer. By incorporating the idea of pseudo-supervised learning, their goal was to achieve supervised control over image styles.

As time has passed, an increasing number of techniques have been applied to the field of multi-style transfer. In 2022, Alexandru [89] employed geometric deformation techniques to achieve multi-style transfer. Their architecture takes multiple style images and one content image as input and then applies geometric deformations to the style images from different artists. This paper extended the research in the direction of Deformable Style Transfer (DST), providing an intriguing approach to multi-style transfer.

Multi-style fast transfer offers the advantages of speed, convenience, and user-friendliness, allowing for the quick transformation of images into multiple styles. It is suitable for various image processing and artistic creation applications. However, it still has certain limitations and drawbacks, including limited style choices, constraints on creating new styles, style consistency issues, and a lack of personalized customization. In certain situations, users may require more flexibility and personalized style transfer options.The summary of relevant articles is shown in Table 11.

Table 11. Summary of multi-style fast transfer methods.

Author	Algorithm	Contribution
Dumoulin [84]	Binding the relatively few parameters in the CIN layer to each style allows for training only these parameters for each new style, enabling multi-style transfer within a single network.	The approach of having N styles share a single model, along with the application of CIN, was proposed.
Chen [85]	In the StyleBank layer, StyleBank is used to classify and represent the input styles, with each bank representing a specific style.	Achieving multi-style fast transfer using different StyleBanks within a single network.
Wang [86]	Enhancing the network's generalization ability and style transfer effectiveness by jointly training multiple subnetworks and utilizing various style images.	Proposing a multi-modal training strategy.
Huang [87]	Applying the semantic-aware multi-style transfer network called Style Mixer.	Introducing the application of semantic awareness in multi-style transfer.
Kim [88]	(i) The pseudo-ground truth (pGT) generation stage and (ii) the Semantic Multi-Style Transfer (SMST) learning stage, two distinct stages.	Introducing a pseudo-supervised learning framework for Semantic Multi-Style Transfer (SMST).
Alexandru [89]	An architecture that takes multiple style images and one content image as input, and applies geometric deformations to images with different styles from various artists.	Deformable Style Transfer (DST) extends the direction of style transfer, offering an approach for multi-style transfer.

4.1.3. Arbitrary Style Fast Transfer

The demand for more flexible and personalized style transfer gave rise to Arbitrary Style Fast Transfer. Arbitrary Style Fast Transfer allows for the transformation of input images or videos into outputs with any desired style while maintaining the consistency of the image content. The key to Arbitrary Style Fast Transfer lies in learning the representation of styles through pre-trained neural network models and applying these representations to input images or videos.

How can one match when using arbitrary styles in a network? In 2016, Chen [90] first introduced his idea. He proposed a patch-based arbitrary style transfer method that matches the features of the target image with those of the style image, before exchanging the features of the target image and the style image to achieve arbitrary style transfer. However, this method still comes with a certain time cost and has not yet met the requirements for real-time processing.

In 2017, Huang [91] introduced a real-time arbitrary style transfer network that achieved a faster style transfer. The author demonstrated in the paper that normalization (IN) achieves style normalization by standardizing feature statistics. Therefore, the author innovatively proposed the AdaIN layer for style transformation. In Conditional Instance Normalization (CIN), the network learns affine transformation parameters β and γ. However, the AdaIN layer proposed by the author does not need to learn these two parameters and directly uses the feature mean and standard deviation of the style image. Its formula is as follows:

$$AdaIN(x, y) = \sigma(y)((x - \mu(x))/\sigma(x)) + \mu(y) \tag{16}$$

In the formula, $\mu(x)$ and $\sigma(x)$ represent the mean and standard deviation of the features of the content image, while $\mu(y)$ and $\sigma(y)$ represent the mean and standard deviation of the features of the style image. This formula can be understood as first normalizing the content image (subtracting its mean and dividing by its standard deviation) and then stylizing it to match the style of the style image (multiplying by the standard deviation

of the style image and adding its mean). The network architecture in the paper is as following Figure 14.

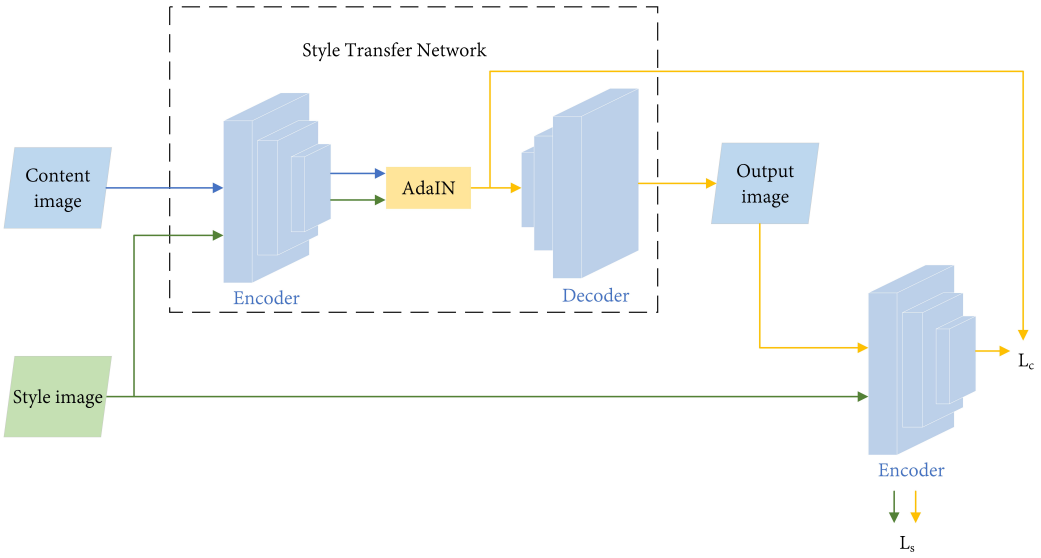

Figure 14. AdaIN arbitrary style migration network.

During the training process, features are first extracted from the content image and style image using the VGG network. These features are then manipulated within the AdaIN module. Subsequently, the processed features are input into a symmetric Decoder network to reconstruct the image. These reconstructed images are then fed back into the VGG network to calculate both content loss and style loss.

The total loss is defined as:

$$L = L_C + \alpha L_S \tag{17}$$

L_C represents the content loss and L_S represents the style loss, weighted by the style loss coefficient α. The content loss L_C is the Euclidean distance between the target features and the output image features. The article uses the AdaIN (Adaptive Instance Normalization) output t as the content target, rather than the conventional feature responses of the content image:

$$L_C = \|f(g(t)) - t\|_2 \tag{18}$$

As the AdaIN (Adaptive Instance Normalization) layer in the article only transfers the mean and standard deviation of style features, the style loss only matches these statistical data:

$$\begin{aligned} L_S = &\sum_{i=1}^{L} \|\mu(\varnothing_i(g(t))) - \mu(\varnothing_i(s))\|_2 + \\ &\sum_{i=1}^{L} \|\sigma(\varnothing_i(g(t))) - \sigma(\varnothing_i(s))\|_2 \end{aligned} \tag{19}$$

Each $\varnothing_i$ represents a layer in VGG-19 used to compute the style loss.

It is worth noting that the parameters of the VGG network are not updated during the training process. The primary goal of training is to obtain an excellent decoder network. The key advantages of this method are its ability to achieve real-time performance and flexibility. The introduction of AdaIN successfully enabled fast transfers to arbitrary styles, making a significant contribution to the field of real-time arbitrary style transfer.

To achieve arbitrary style fast transfers, other researchers proposed various methods. In 2017, Ghiasi [92] introduced an approach that trains a style prediction network to generate feature vectors for each style image. They proposed the Style Prediction Network, which is trained on a large number of content and style images. In the same year, Li [93] presented an arbitrary style fast transfer method based on feature transformation. This method utilizes Whitening and Coloring Transforms (WCT) to operate between the encoder and decoder. This approach offers broad applicability and flexibility but tends to be slower.

In 2018, Gu [94] introduced a method called feature rearrangement to achieve arbitrary style transfer, which involves reordering image features in the spatial domain. The research showed that the Gram matrix of the rearranged image features remains unchanged, ensuring overall consistency between the generated image and the style image. In 2020, Jing [95] proposed a novel normalization module called Dynamic Instance Normalization (DIN) to effectively achieve arbitrary style transfer. In 2022, building on the use of AdaIN, Zhang [96] introduced Adaptive Style Modulation (AdaSM) to provide more precise control over global style. In the same year, Wang [97] incorporated contrastive learning into arbitrary style transfer. They used contrastive learning to model both style and content, constraining contrastive style loss and contrastive content loss and, thus developing a style transfer framework suitable for arbitrary styles.

Compared to single-style and multi-style transfer, arbitrary style transfer offers greater flexibility and artistic creativity, making it a promising field for future development. However, the current techniques for arbitrary style transfer still require further research and development to address the challenges related to algorithm real-time performance, accurate transfer results, and detail preservation. This will enable wider applications and higher-quality style transfer effects in the future. A summary of relevant articles is shown in Table 12.

Table 12. Summary of arbitrary fast style transfer methods.

Author	Algorithm	Contribution
Chen [90]	Any style transfer is achieved by matching the characteristics of the target image with the characteristics of the style image and then exchanging the target image and the style image.	Any style migration is achieved, but the speed still does not meet the real-time requirements.
Huang [91]	The AdaIN layer is proposed to achieve style conversion.	Realized real-time migration of any style.
Ghiasi [92]	Proposed a style prediction network, which is trained by inputting a large number of content maps and style maps.	Train a style prediction network and let it generate feature vectors for each style image.
Li [93]	Perform WCT operations between Encoder and Decoder to achieve arbitrary style migration through WCT.	Quick transfer of any style through feature transformation.
Gu [94]	Use feature reshuffle to rearrange image features in the spatial domain.	Proposed a feature reshooting method for arbitrary style transfer.
Jing [95]	Dynamic Instance Normalization (DIN).	A new general normalization module is proposed to achieve arbitrary style migration more efficiently.
Zhang [96]	Adaptive Style Modulation (AdaSM) is proposed.	Achieved more precise control over global style.
Wang [97]	Modeling style and content using contrastive learning, constrained by contrastive style loss and contrastive content loss.	Introducing contrastive learning to arbitrary style transfer.

4.2. Generative Adversarial Network Method

With the continuous advancement of technology, Generative Adversarial Networks (GANs) have made significant progress in the field of style transfer. GANs offer multiple advantages in style transfer, including the ability to work without paired data, diverse style transformations, high-quality image generation, real-time performance, and efficiency. These advantages have significantly enhanced the application of style transfer techniques in the fields of image processing and computer vision.

The emergence of GANs has not only improved the quality and efficiency of style transfer but has also expanded its application areas, making image style transformation more flexible and diverse. Particularly noteworthy is Liang's 2020 review of GAN networks [98], where GAN and its important branch networks are listed, with CycleGAN being highly favored in the field of style transfer due to its ability to achieve high-quality style transfer without the need for paired data. Next, we will analyze the development of GANs in the field of style transfer from two aspects: variant networks of GANs and CycleGAN networks.

4.2.1. GAN Network Variants

In 2014, Goodfellow [99] introduced the Generative Adversarial Network (GAN) model, which marked a significant breakthrough in the field of generative models. The goal of style transfer is to transfer the style features of one image to another image to create new images or change the style of images. GANs provided a powerful framework for style transfer by training generative models adversarially to produce highly realistic style transfer results. As a result, GAN networks naturally became important models in the field of style transfer.

The core idea of GAN is to employ two adversarial networks: the Generator and the Discriminator. The Generator's task is to generate realistic data samples from random noise, while the Discriminator is responsible for distinguishing whether a given sample is a real one or a generated one. In the process of these two networks competing and playing against each other, the Generator gradually learns to produce more realistic samples, and the Discriminator continuously improves its ability to discriminate. This adversarial training process drives the Generator to approximate the real data distribution, resulting in the generation of increasingly realistic samples. The structure of a GAN network is as following Figure 15.

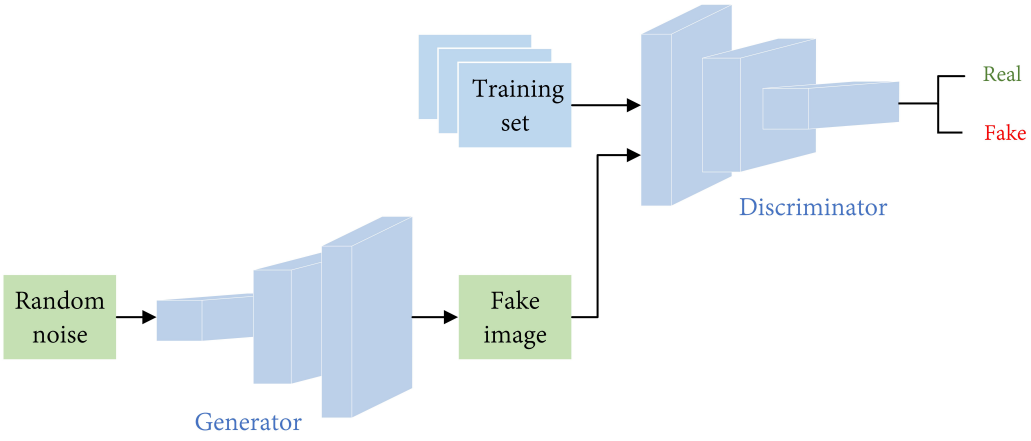

Figure 15. GAN network.

Variant networks of GANs have extended the application of GANs in the field of style transfer by introducing methods such as image synthesis, attention mechanisms, and image

translation. These variants offer more flexibility and controllability in style transfer, making image transitions between different styles appear more natural and realistic.

In 2016, Wang [100] introduced the S^2-GAN model, which decomposes the image generation process into structure and style, with two separate GAN networks responsible for the different aspects. This approach leads to more realistic image generation. In 2019, Karras [101] introduced StyleGAN, which utilized a multi-level generator network, with each level corresponding to a specific image resolution. StyleGAN achieved high-quality, diverse, and controllable image generation.

To further enhance the quality of generated images, In 2020, Li [102] introduced an approach called SDP-GAN, aiming to achieve high-perceptual-quality style transfer while emphasizing the retention of significant details. In 2023, Han [103] designed the Deep Extraction Generative Adversarial Network (DE-GAN) specifically for artistic style transfer. DE-GAN proposed a multi-feature extractor to extract color features, texture features, depth features, and shape masks from style images, resulting in higher-quality style images that better conform to aesthetic characteristics.

In 2014, alongside GAN networks, Conditional Generative Adversarial Networks (cGAN) [104] were introduced. In cGANs, both the generator and discriminator receive conditional information as input and use this information to guide the generation and discrimination processes during training. By introducing conditional information, cGANs can generate specific types of samples based on the given conditions, offering more control and flexibility.

In 2019, Mao [105] utilized cGAN models for image synthesis tasks, addressing issues with model collapse in cGAN networks and ensuring the quality of image synthesis. In 2021, Lv [106] introduced a novel multi-style unsupervised image synthesis model using a Generative Adversarial Network (MSU-GAN). This model addressed the issue of training data pairing and was capable of generating high-quality images with multiple styles.

As style transfer extended to the field of image translation, GANs found broader applications. In 2017, Dong [107] introduced a general approach based on deep convolution and conditional generative adversarial networks to address image translation problems. They developed a two-step (unsupervised) learning method that transformed images between different domains using unlabeled images, successfully making the model versatile.

Similarly, in 2018, starGAN [108] was introduced to tackle multi-domain image translation problems. It used a shared generator and a shared discriminator to achieve image translation across multiple domains, offering a unified solution for multi-domain image translation. In 2020, Hicsonmez [109] proposed GANILLA, a method based on cGANs for automated image-to-illustration transformation, achieving highly realistic photo-to-illustration conversion. In 2021, Roy [110] presented Trigan, a Generative Adversarial Network-based Multi-Source Domain Adaptation (MSDA) approach capable of producing good image translation results across multiple sources.

GANs achieve optimization by introducing competition between a generator and a discriminator. This competitive mechanism enables GANs to learn the mapping relationship between input images and target styles and generate high-quality, realistic style transfer results. However, GANs may face challenges such as training instability and mode collapse during the training process. Future directions in this field include improving the stability and quality of GANs, increasing the application scenarios and domains, and considering integration with other technologies to further enhance the effectiveness and practicality of style transfer. A summary of relevant articles is shown in Table 13.

Table 13. Summary of GAN-related style transfer methods.

Author	Algorithm	Contribution
Goodfellow [99]	The generator and the discriminator compete with each other.	Proposed the GAN model.
Wang [100]	S^2-GAN model, there are two different parts of GAN network the model.	Generate more realistic images.
Karras [101]	StyleGAN uses a multi-level generator network to a specific image resolution.	Achieve high-quality, diverse image generation.
Li [102]	The method of SDP-GAN is proposed, which introduces a saliency network.	Solve the problem of significant detail loss in traditional style transfer methods.
Han [103]	Designed a deep extraction generative adversarial network (DE-GAN) and proposed a multi-feature extractor.	Generate higher-quality, more aesthetically pleasing style images.
Mirza [104]	In conditional GAN, both the generator and the discriminator receive condition information as input.	Conditional GAN has higher control and flexibility.
Mao [105]	Image synthesis tasks through the cGAN model.	Solve the problem of cGAN network model collapse and ensure the quality of image synthesis.
Lv [106]	Multi-style unsupervised image synthesis model based on MSU-GAN.	Solve training data matching problem.
Dong [107]	A general method for solving image translation problems.	Successfully made the model universal.
Choi [108]	Image conversion between multiple domains is realized through shared generator discriminator.	Provides a unified solution for multi-domain image translation.
Hicsonmez [109]	GANILLA, an automatic image-to-illustration conversion method based on GAN.	Achieve very realistic conversion of photos into illustrations.
Roy [110]	Proposed Trigan, a multi-source domain adaptation (MSDA) method.	Enables better image translation results.

4.2.2. CycleGAN

Due to the excellent scalability of Generative Adversarial Networks (GANs), many outstanding GAN extension networks have emerged, and CycleGAN is one of them. CycleGAN is an unsupervised image translation method based on Generative Adversarial Networks (GANs), used to perform image translation between two different domains, such as turning horses into zebras or apples into oranges, and so on. In 2017, Isola [17] proposed the pix2pix network. Pix2pix is based on the GAN architecture and uses paired images for image translation, where the input consists of two different styles of the same image, making it suitable for style transfer. A facades style transfer of the pix2pix network is shown in the following Figure 16.

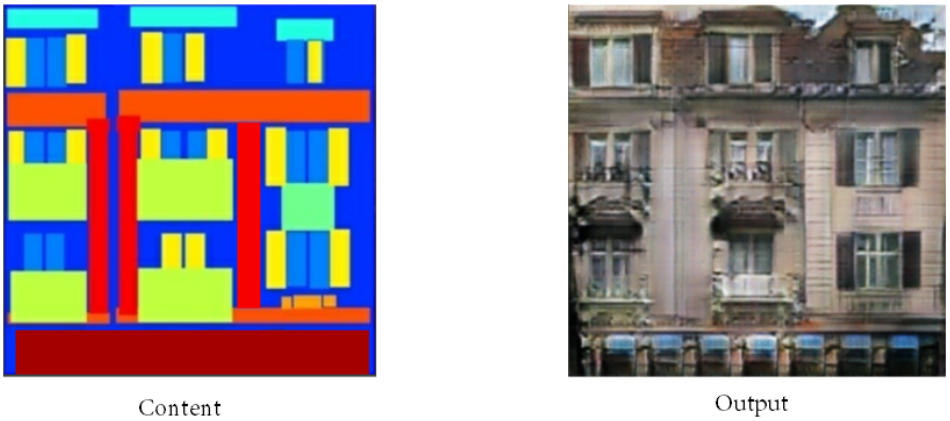

Figure 16. pix2pix sample.

However, pix2pix requires a paired dataset, which is not very convenient to use. Therefore, Zhu [18] proposed the CycleGAN framework, which enables style transfer between unpaired images. The structure of CycleGAN is shown in Figure 17.

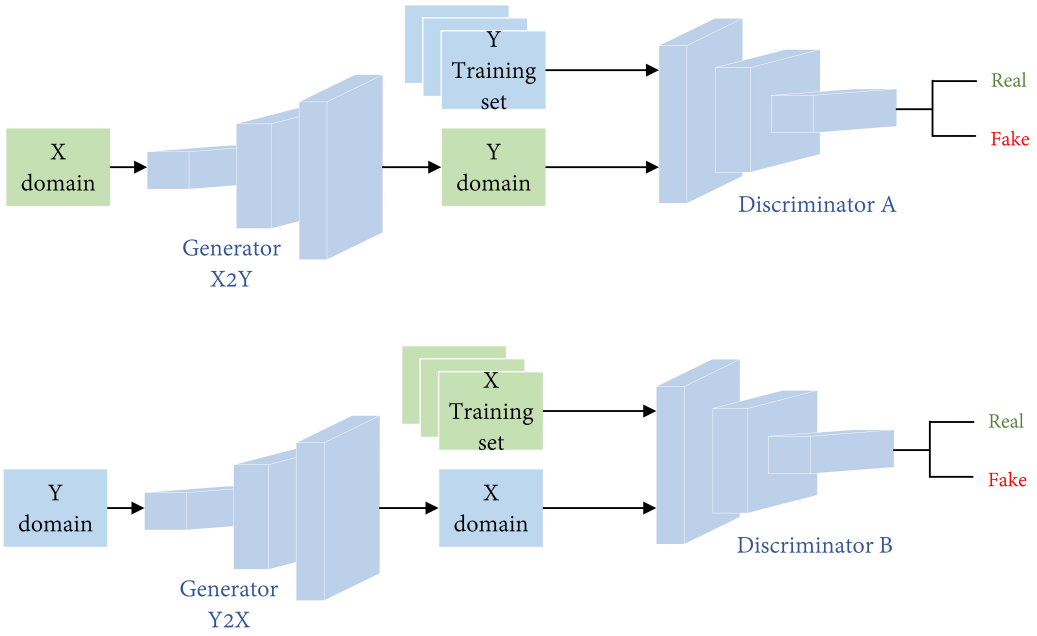

Figure 17. CycleGAN network.

CycleGAN consists of two generators and two discriminators. In the CycleGAN framework, one generator transforms images from one domain to another, while the other generator reverses this process. The discriminators were used to assess the authenticity of the generated images and provide feedback signals for training the generators. One of the core ideas of CycleGAN is cycle consistency loss, which ensures that the transformed images can be mapped back to the original domain while preserving the content and style as closely as possible. In the same year, DiscoGAN [111], DualGAN [112], and CycleGAN all shared the same network architecture, but their implementation details differ significantly, demonstrating that the use of a dual-network architecture is a viable approach for solving problems in unsupervised networks.

Based on CycleGAN, researchers have made various improvements and extensions. In 2020, Tu [113] proposed a multi-style CycleGAN image transmission system by introducing conditional constraints on the target style feature map, achieving more efficient image transmission and processing. In 2021, Zhang [114] introduced spectral normalization in each convolution kernel of the discriminator, so that the parameter matrix of the convolutional neural network satisfies the Lipschitz constraint. This method helps to better preserve the facial features and was therefore successfully used in human cartoonization applications. In addition, Liu [115] introduced noise constraints and the LBP texture features of the prototype into the loss function of CycleGAN, which improved the edge and texture intensity of the image after the oil painting style was transferred.

In 2022, Liao [116] used the UNET neural network to replace the original RESNET neural network to train the generator model and then built a discriminator network model based on standard convolution and depth-separable convolution. In addition, Wang [117] proposed a cycle-consistent generative adversarial network (ESA-CycleGAN) based on edge features and self-attention. In this network, a self-attention mechanism module is added to obtain more correlations in the image, thereby improving the image quality

and better retaining the details of the image. A summary of relevant articles is shown in Table 14.

Table 14. Summary of CycleGAN-related style transfer methods.

Author	Algorithm	Contribution
Isola [17]	Pix2pix is based on GAN architecture and uses paired images for image conversion.	Proposed pix2pix network.
Zhu [18]	CycleGAN uses a bidirectional transformation process that can self-correct and learn mapping relationships between domains.	The CycleGAN framework is proposed to achieve style transfer between unpaired images.
Tu [113]	Add conditional constraints to the target style feature map.	Implemented a multi-style CycleGAN image transmission system.
Zhang [114]	Spectral normalization is introduced in each convolution kernel of the discriminator.	Better retain facial features and successfully cartoonize humans.
Liu [115]	Introducing the LBP texture features and total variation of the prototype into the CycleGAN loss noise constraint.	Better image oil painting style transfer and reconstruction performance, image oil painting style transfer and reconstruction effect is better.
Liao [116]	The discriminator network model is established by using UNET neural network in generator model.	Improves the efficiency of the network and effectively completes the task of style migration.
Wang [117]	Propose a cycle-consistent GAN (ESA-CycleGAN) based on edge features and self-attention.	Improved image quality and better-preserved image details.

4.2.3. New Applications of GAN

With the continuous advancement of technology, GAN networks are continuously exploring and pioneering new application scenarios and creative approaches in the field of style transfer. These new applications have expanded the scope of GAN networks in style transfer, providing people with fresh ideas and perspectives. Among them, cartoonization has become an important branch of style transfer. Transforming images into a cartoon style brings people incredibly creative and visually appealing experiences.

In 2017, Zhang [118] proposed a style transfer method specifically for anime sketches. This approach combined an enhanced residual U-Net with an Auxiliary Classifier GAN (ACGAN) to achieve the transformation of anime-style rendering. In 2018, Chen [119] introduced a generative adversarial network called "Cartoongan", designed specifically for converting photos into a cartoonized style. Cartoongan adopted the cGAN form, using input photo images as conditional labels to guide the generator in creating images with the desired cartoon style. However, due to the characteristic smooth surfaces, lack of pronounced color variations, and sharp edges often found in cartoon-style images, it may sometimes be challenging to generate entirely satisfactory cartoon images.

To address this issue, in 2021, Dong [120] introduced a Generative Adversarial Network (GAN) called CartoonLossGAN, which is based on a cartoon loss. This novel approach utilizes a unique cartoon loss function that mimics the sketching and coloring process when generating cartoon-style images. This method results in smoother surfaces and finer coloring in cartoon-style images, leading to the creation of more exquisite cartoon-style images. In the same year, Shu [121] proposed an innovative Multi-Style CartoonGAN architecture, MS-CartoonGAN. This can transform input photos into various cartoon styles, offering greater flexibility and expressiveness compared to other methods.

In 2016, Li [122] introduced the Markovian Generative Adversarial Network (MGAN). By combining precomputation with GAN networks, this method efficiently achieves texture synthesis, allowing for the real-time generation of high-quality texture images. The

emergence of MGAN brought new solutions to the field of texture synthesis. GANs have also found application in the realm of ink painting. In 2018, He [123] proposed ChipGAN, an architecture based on an end-to-end generative adversarial network designed for transferring photos into the style of Chinese ink paintings. The application of this architecture provided new perspectives and methods for the transfer of Chinese ink painting styles.

GANs have also found applications in the realm of human facial expression portraiture. In 2020, Yi [124] introduced APDrawingGAN++, a GAN designed to transform facial photos into artistic portrait drawings (APDrawings). This method addressed several significant challenges, including highly abstract style conversion, variations in artistic techniques for different facial features, and sensitivity to perceptual artifacts. To further encompass various portrait styles, in 2021, Song [125] proposed AgileGAN, a framework that can generate high-quality stylized portraits through reverse-consistent transfer learning. AgileGAN achieved higher portrait stylization quality.

The development of GANs in the field of style transfer has indeed seen continuous evolution and innovation. From the earliest GAN models to the present day, many researchers have delved deep into various aspects of GANs, leading to constant improvements in their style transfer performance. Looking forward, as GAN technology continues to advance and further research is conducted, it is believed that style transfer techniques will find applications in even more domains, continuously enhancing the quality and efficiency of image generation. A summary of relevant articles is shown in Table 15.

Table 15. Summary of new applications of GAN in style transfer.

Author	Algorithm	Contribution
Zhang [118]	Use enhanced residual U-Net combined with auxiliary classifier GAN to achieve animation-style conversion.	Proposed a style transfer method for animation sketches.
Chen [119]	Cartoongan uses cGAN to guide the generator to generate images with the style of the target cartoon.	Turn real-world photos into cartoon-style illustrated images for photo cartoonization.
Dong [120]	CartoonLossGAN is proposed for cartoonization, using a novel cartoon loss function.	Mimic the coloring process to produce a more elegant cartoon-style image.
Shu [121]	Proposed a novel multi-style generative adversarial network (GAN) architecture called MS-CartoonGAN.	Photos can be converted into a variety of cartoon styles.
He [123]	Proposed ChipGAN, an architecture based on end-to-end generative adversarial networks.	Realize the transfer of photos to Chinese ink painting style.
Yi [124]	APDrawingGAN++ is proposed for transforming facial photos into artistic portraits.	Addresses highly abstract styles, painting techniques of different facial features, and sensitivity to artefacts.
Song [125]	A framework for generating high-quality style portraits through reverse consistent transfer learning.	Better encoding of different levels of detail, resulting in higher-quality portrait stylization.

4.3. Attention Mechanism

The Attention Mechanism is a method that emulates human visual attention to enhance a model's focus on important parts of input data, thereby improving its ability to handle complex information. It allows models to prioritize important information relevant to the task at hand, improving model performance and efficiency. In style transfer, attention mechanisms can be employed to enhance a model's focus on different regions of an image, thereby facilitating more effective style transfer. The application of attention mechanisms in style transfer enables models to flexibly transform different regions of an image with varying styles, ultimately enhancing the quality and expressiveness of style transfer. It leverages both local and global features in images, enabling models to better handle intricate style transformation tasks.

4.3.1. Neural Network Applications

The application of attention mechanisms in neural networks has become highly prevalent, with these serving as a powerful tool for optimizing and improving models. This has significantly advanced the field of neural networks, particularly in the context of style transfer.

In 2019, Park [126] proposed an intriguing Style-Attention Network (SANet). This method introduced a unique attention mechanism called the "style-attention module", allowing the network to adaptively select different styles at each pixel level. The innovation of this approach lies in the incorporation of attention mechanisms to enhance the modeling and control of styles, thereby making style transfer more precise and flexible for arbitrary styles. The network architecture is shown in Figure 18.

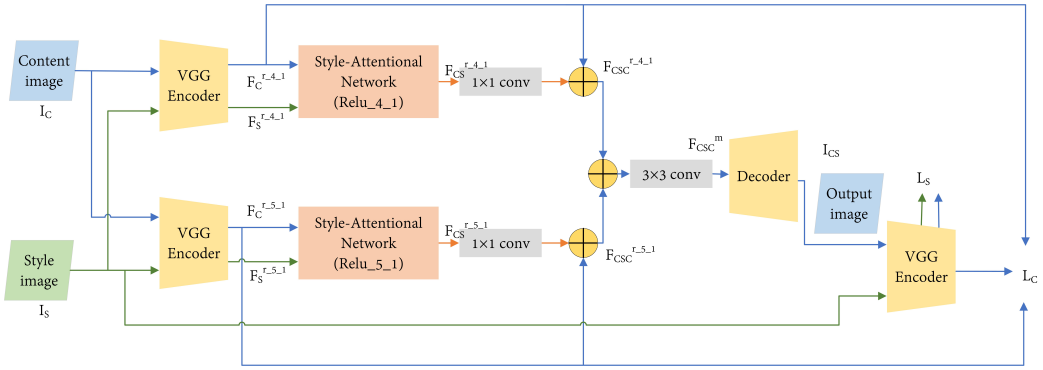

Figure 18. SANet.

The network primarily consists of an encoder, SANet (Style-Attention Network), and a decoder. The encoder utilizes a pre-trained VGG19 model to extract features from the content and style images. SANet is responsible for combining the Style feature map and Content feature map from VGG19 based on their respective attention. The decoder's network structure mirrors the encoder's, allowing it to reconstruct images using the combined feature maps to obtain stylized images. The final SANet architecture effect is shown in Figure 19.

Figure 19. SANet effect.

This algorithm leverages the features output from two layers of the VGG19 model: Relu4-1 and Relu5-1. The reason for this choice is that using only Relu4-1 effectively preserves the global structure of both style and content images but may not perform well when representing local style effects. On the other hand, Relu5-1 has a larger receptive field and is better at expressing local style effects. $\oplus$ is used to combine features from multiple different sources, which are then passed as input to the subsequent layers of the network. Therefore, by using the features from Relu4-1 and Relu5-1, and combining them through two SANet modules before integration, the final stylized image is obtained.

In SANet, its inputs Fc and Fs are the feature maps of the content image and style image output by VGG19. The formula is as follows:

$$F_{cs}^{i} = \frac{1}{C(F)} \sum_{\forall j} exp(f(\bar{F}_{C}^{i})^{T} g(\bar{F}_{S}^{j})) h(F_{S}^{j}) \tag{20}$$

$$C(F) = \sum_{\forall j} exp(f(\bar{F}_{C}^{i})^{T} g(\bar{F}_{S}^{j})) \tag{21}$$

$C(F)$ serves as a normalization factor. SANet can learn to map the relationship between content and style feature maps (e.g., similarity) to embed the appropriate style features at each position of the content feature map. The loss function in the paper consists of three parts: traditional style and content losses, as well as the newly proposed identity loss.

$$L = \gamma_{C} L_{C} + \gamma_{S} L_{S} + L_{identity} \tag{22}$$

The role of identity loss is to calculate the differences between the generated stylized image and the original image at both the pixel level and perceptual level. The formula is as follows:

$$\begin{aligned} L_{identity} = \lambda_{identity1}(||I_{CC} - I_{C}||_{2} + ||I_{SS} - I_{S}||_{2}) + \\ \lambda_{identity2} \sum_{i=1}^{L} (||\delta_{i}(I_{CC}) - \delta_{i}(I_{C})||_{2} + \\ ||\delta_{i}(I_{SS}) - \delta_{i}(I_{S})||_{2}) \end{aligned} \tag{23}$$

where I_C (I_S) represents the original content (style) image, and I_{CC} (I_{SS}) represents the output image synthesized from the image pair (content or style). However, SANet still faces distortion issues in local style transfer. Therefore, in 2021, Liu [127] proposed a new AdaAttN module for arbitrary style transfer, which focuses on both shallow and deep features to better balance global and local style transfer effects. In 2020, Deng [128] introduced a multi-adaptive network that includes two self-adaptive (SA) modules and one cooperative self-adaptive (CA) module. This approach further enhances the accuracy and flexibility of style transfer.

In 2022, Zhu [129] introduced a Lightly Progressive Adaptive Instance Normalization (LPAdaIN) model that combines Adaptive Instance Normalization (AdaIN) layers with Convolutional Block Attention Modules (CBAM). This model alleviates the structural distortion in stylized images. In the same year, Li [130] proposed a style transfer attention network based on Unbalanced Optimal Transport, which increases the global distribution similarity and enhances the robustness of the results. In 2023, Ye [131] utilized Conditional Instance Normalization (CIN) layers and Convolutional Block Attention Modules (CBAM) in the stylization network. This approach not only achieves multi-style transfer but also successfully preserves the semantic information of the original image, enabling natural transitions between stylized regions in the artwork.

By introducing attention mechanisms, neural networks can dynamically adjust their focus on different parts of the data, thereby handling complex data and tasks more effectively and ultimately improving model performance and generalization. In the field of style transfer, attention mechanisms have demonstrated extensive potential applications, providing powerful tools for model optimization and improvement. Overall, the integration of attention mechanisms has led to significant advancements in neural networks across various tasks, further driving the continuous development and innovation of neural network technology. A summary of relevant articles is shown in Table 16.

Table 16. Summary of the application of neural network attention mechanisms.

Author	Algorithm	Contribution
Park [126]	Introducing an attention mechanism called "Style Attention Module".	Introducing attention mechanism to enhance style modeling and control.
Liu [127]	Arbitrary style transport using a new AdaAttN module.	AdaAttN better balances global and local style transfer effects.
Deng [128]	Two adaptive (SA) modules and one cooperative adaptive (CA) module are used.	Arbitrary style transfer via multiple adaptive networks.
Zhu [129]	LPAdaIN model combining adaptive instance normalization (AdaIN) layers and convolutional block attention modules (CBAM).	The LPAdaIN model alleviates the problem of image structure distortion.
Li [130]	Style transfer attention network based on imbalanced optimal transfer.	It better ensures the similarity of global distribution and improves the robustness of the results.
Ye [131]	CIN layers and convolutional block attention modules (CBAM) are used to achieve multi-style transfer while retaining the original semantic information.	The generated works maintain the original salient semantic and visual characteristics of the content image and achieve regional stylization.

4.3.2. Applications of GAN

In the field of GAN, the incorporation of attention mechanisms makes GANs more flexible and intelligent in processing images. This mechanism allows models to dynamically focus on the crucial information within an image, thereby enhancing the quality and diversity of image generation. Additionally, attention mechanisms can boost the model's focus on input images, playing a vital role in applications like style transfer.

In 2019, Zhang [132] proposed a method called "Self-Attention Generative Adversarial Network" (Self-Attention GAN), which combines self-attention mechanisms with Generative Adversarial Networks (GANs). In Self-Attention GAN, both the generator and discriminator include self-attention modules that can learn the global dependencies in images. This approach is better equipped to capture the global structure and details of images.

To extract key semantic information, Tang [133,134] introduced a novel Attention-Guided Generative Adversarial Network (AGGAN and AttentionGAN), which can detect the most discriminative semantic objects and minimize unnecessary variations in semantic operations. However, at times, attention mechanism network models can become overly large. Therefore, in 2022, Zhao [135] proposed a lightweight Domain Attention Generative Adversarial Network (LDA-GAN). This model achieves fewer parameters and reduces memory usage by introducing an improved Domain Attention Module (DAM) to establish remote dependencies between two domains, resulting in a lightweight attention-based generative adversarial network.

In 2022, Zhang [136] introduced a new Generative Adversarial Network that embeds channel attention mechanisms between the upsampling and downsampling layers of the generator network. This approach aims to avoid increasing model complexity while preserving the complete details of low-level information. This method prevents the model from becoming overly complex. In 2023, Zhang [137] combined Convolutional Block Attention (CBA) modules with a Generative Adversarial Network, proposing a new model called Convolutional Block Attention Generative Adversarial Network (CBA-GAN). This model is capable of transforming real photos into cartoon images.

The application of attention mechanisms in GANs enhances the focus of both the generator and discriminator on crucial information within images, thus improving GAN performance and generation quality. By introducing channel and spatial attention, GANs can dynamically adjust the feature maps of the generator and discriminator, allowing the generator to prioritize important features during image generation and the discriminator to focus on critical regions during image recognition. The integration of attention mechanisms provides powerful tools for optimizing and improving GANs, driving the development and application of GAN technology, and achieving significant progress in fields like style transfer. A summary of relevant articles is shown in Table 17.

Table 17. Summary of GAN attention mechanism application.

Author	Algorithm	Contribution
Zhang [132]	The method of Self-Attention GAN combines the self-attention mechanism and GAN.	Has better visual quality and realism, able to capture the global structure and details of the image.
Tang [133,134]	Proposed a novel attention-guided generative adversarial network (AGGAN and AttentionGAN).	Detect the most discriminating semantic objects and minimize unwanted changes.
Zhao [135]	A lightweight domain attention generative adversarial network (LDA-GAN).	Fewer parameters and lower memory usage.
Zhang [136]	The channel concern mechanism is embedded between the upper and lower sampling layers of the generator network.	Successfully avoided model complexity and found a balance between content and style.
Zhang [137]	The CBA is combined with GAN to propose a new model called CBA-GAN.	The ratio of edge, texture and smoothness in image effect can be adjusted flexibly, which has good performance and robustness.

4.4. Diffusion Model

To address the issues of training instability and poor quality when generating high-quality data using Generative Adversarial Networks (GANs), diffusion models have attracted attention. Diffusion models define a forward process of gradually adding noise and a reverse process of gradually removing noise to generate images. This approach not only stabilizes the generation process but also better preserves image details. The key characteristics of diffusion models include step-by-step generation processes, strong interpretability, high-quality generation, and training stability. Moreover, diffusion models excel in style transfer tasks by operating in latent spaces, enabling more flexible and efficient style transfer processes, thereby producing images with enhanced stylized effects.

In 2015, Sohl-Dickstein [138] introduced diffusion models, which generate high-quality data by defining forward diffusion and reverse denoising processes. This seminal paper laid the foundation for subsequent diffusion models and related research, significantly advancing the field of generative models. Building upon this work, in 2020, Ho [139] proposed the Denoising Diffusion Probabilistic Models (DDPM). DDPM employs a step-by-step denoising approach to generate high-quality and detail-rich images, achieving impressive generation results across multiple datasets. Ho's work introduced the framework of forward noise addition and reverse denoising, contributing to a more stable and interpretable generation process within the field of generative models.

Diffusion models are closely related to style transfer tasks through their iterative generation and denoising processes, operations in latent space, and flexible generation and editing capabilities. In style transfer, diffusion models can progressively transform the style features of images while preserving their content features, thereby achieving high-quality style transfer effects. In 2023, Wang [140] first introduced diffusion models into the field of

neural style transfer. He proposed a new C-S disentangled style transfer framework that leverages the powerful style removal and generation capabilities of diffusion models. This framework explicitly extracts content information and implicitly learns complementary style information, enabling interpretable and controllable C-S disentanglement and style transfer. The network architecture is shown in Figure 20.

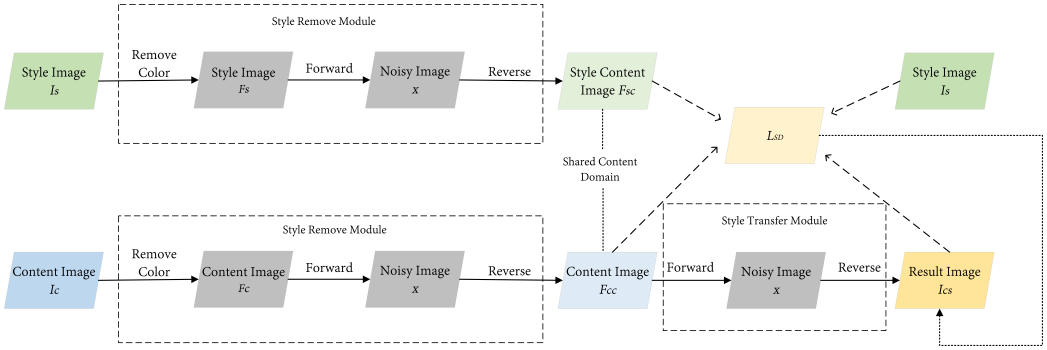

Figure 20. C-S disentangled style transfer framework.

The network consists of a style removal module and a style transfer module. The content image I_C and style image I_S are first fed into the diffusion-based style removal module to explicitly extract domain-aligned content information, resulting in decolorized images F_{SC} and F_{CC}. Then, F_{CC} content is input into the diffusion-based style transfer module to obtain the stylized result I_{CS}. During training, we fine-tuned the style transfer module by coordinating CLIP-based style disentanglement loss L_{SD} with pre-reconstruction style.

To appropriately transfer disentangled style information to other content, the paper defines an L_1 loss:

$$L_1 = \|D_{CS} - D_S\| \tag{24}$$

D_{CS} represents the disentangled style information of I_{CS}, and D_S represents the disentangled style information of I_S. Therefore, L_1 achieves minimal absolute pixel difference but does not guarantee the generation of stylized images, potentially leading to stylistic deviations. To address this issue, the paper further constrains the disentanglement direction:

$$L_2 = 1 - \frac{D_{CS} \cdot D_S}{\|D_{CS}\|\|D_S\|} \tag{25}$$

L_2 aligns the direction of transferring content from the content image to its stylized counterpart (i.e., stylized result) with the direction of transferring content from the style image to its stylized counterpart (i.e., the style image itself). Together, L_1 and L_2 achieve an exact one-to-one mapping from content in the content domain to stylization in the style domain.

Finally, the style disentanglement loss is defined as the composition of L_1 and L_2:

$$L_{SD} = \alpha L_1 + \beta L_2 \tag{26}$$

where α and β are the hyperparameters set in the experiments. Because style information arises from differences between the content and its stylized result, it can be deeply understood through learning the relationship between C-S. The paper suggests that image styles can naturally and harmoniously transfer into content, thereby producing better stylized images. Ruta [141] similarly combines neural style transfer networks with diffusion models, intertwining pre-extracted content noise for shape and composition with style attention values from style images and extracting content and style information from interleaved data to generate the final stylized image. However, these methods depend on the quality

and diversity of the training data. Insufficient or imbalanced training data may affect the quality of the generated images.

In many applications, obtaining thousands of required style images may not be feasible. Everaert [142] proposed a style adaptation method in 2023 that modifies the initial latent distribution. The method adjusts the diffusion model using noise and sampled initial latent tensors, requiring only a relatively small number of target images, typically between 50 and 200, to complete style transfer tasks. This fine-tuning approach improves the diffusion model's performance in style transfer but still demands significant computational resources.

In the same year, Yang [143] introduced a zero-shot contrastive loss diffusion model that eliminates the need for additional fine-tuning or auxiliary networks. This method utilizes block-wise contrastive losses between generated samples from a pretrained diffusion model and original images to generate images with identical semantic content to the source image in a zero-shot manner.

In 2024, Chung [144] proposed a new artistic style transfer method based on a pretrained large-scale diffusion model. This method employs the feature of self-attention layers as a cross-attention mechanism, substituting the value of style imagery for the value of content during the generation process. Both methods reduce computational costs, shorten training times, and achieve better style transfer results through pretrained models.

Arbitrary style transfer is an important research direction within the field of style transfer, garnering attention from many researchers. In 2023, Hamazaspyan [145], based on the characteristics of stable diffusion algorithms, derived a new diffusion-enhanced template matching algorithm called DEPM (Diffusion-Enhanced Pattern Matching). DEPM captures high-level style features while preserving fine-grained texture details of the original image. By supporting arbitrary style transformation during the inference process, DEPM makes the style transfer process more flexible and efficient.

In addition, LDMs (Latent Diffusion Models) significantly reduce computational complexity and achieve better results by employing diffusion processes in latent representation space, generating more detailed images. Therefore, they have attracted attention in the field of arbitrary style transfer. In 2023, Chen [146] proposed a new method called ArtFusion, which provides a flexible balance between content and style. It utilizes a dual conditional latent diffusion probability model (Dual-cLDM) to reduce repetitive patterns and enhance subtle artistic aspects such as brush strokes and specific types of features.

Furthermore, in 2024, Wang [147] introduced a new semantic-based stylization method called HiCAST. This method allows for the explicit customization of stylized results based on different semantic clues. A notable feature of this method is the introduction of style adapters, which match multi-level style information with intrinsic knowledge in LDM, enabling the flexible manipulation of output results.

The introduction of diffusion models into the field of style transfer enables stable incremental generation processes and flexible operations in latent space, achieving high-quality and controllable style transformations. This enhances detail preservation and style consistency in image generation and editing. Future developments in style transfer under diffusion models aim to improve computational efficiency, enhance the decoupling of style and content features, and expand the model's applications in more complex and diverse style transfer tasks. This will further drive innovation in visual creation and artificial intelligence. A summary of relevant articles is shown in Table 18.

Table 18. Summary of Diffusion Model Style Transfer.

Author	Algorithm	Contribution
Wang [140]	A new C-S disentanglement style transfer framework is proposed.	The diffusion model was first introduced into the field of neural style transfer.
Ruta [141]	The final stylized image is generated from the content and style information extracted from the interleaved data.	Combining Neural Style Transfer Networks with Diffusion Models.
Everaert [142]	Fine-tuning the diffusion model using the noisy and sampled initial latent tensor.	Only relatively few target images are needed to complete the style transfer task.
Yang [143]	A diffusion model with zero-shock contrast loss is proposed.	No additional fine-tuning or auxiliary networks are required, reducing computational costs.
Chung [144]	A new artistic style transfer method based on pre-trained large-scale diffusion model is proposed.	Shortened training time and achieved better style transfer results through pre-trained models.
Hamazaspyan [145]	A new Diffusion-Enhanced Pattern Matching algorithm (DEPM) is derived.	Supports arbitrary style conversion during inference, making the style transfer process more flexible and efficient.
Chen [146]	A new method ArtFusion is proposed.	Repetitive patterns are reduced and subtle artistic aspects such as brushstrokes and type-specific features are enhanced.
Wang [147]	Propose a new semantic-based stylization method, HiCAST.	Allows users to flexibly manipulate output results.

5. Video Style Transfer

5.1. Video Transfer and Flickering Issues

With the continuous advancement of technology, style transfer on images has gradually extended to the realm of videos, giving rise to video style transfer. Video style transfer is a process of applying artistic styles or appearances to videos, thereby altering the visual effects in the original video.

In the early days of technology development, people tried to use traditional machine learning algorithms to complete the task of video stylization. In 2011, Cao [148] proposed an automated algorithm to achieve video stylization tasks through the MRF model and optical flow estimation method. However, due to the difficulty in accurately capturing style features, the resulting stylization was not obvious. In 2012, Kyprianidis [149] reviewed the style transfer task in the field of non-photorealistic rendering (NPR). The article pointed out that in ensuring the stability of vision-based video stylization is still a largely unsolved problem.

These articles show that due to the complex spatiotemporal features and highly nonlinear mapping relationships in video data, traditional machine learning algorithms often find it difficult to effectively capture and process these features. In addition, video style transfer requires large-scale data and efficient computing power to learn complex model parameters. Traditional machine learning algorithms have limitations in their ability to process large-scale data and complex computing tasks, and it is difficult to meet the requirements of a real-time and stable style transfer. Therefore, traditional machine learning algorithms are not very suitable for video style transfer.

In contrast, deep learning excels in video style transfer because it can effectively capture and express the complex spatiotemporal features in videos through hierarchical feature learning and nonlinear modeling. Deep learning not only enables end-to-end learning and the optimization of models but also adapts well to large-scale data and complex computational tasks. This capability allows for high-quality video transformations with artistic styles.

Video style transfer is a process that applies artistic styles or appearances to videos, resulting in visual effects distinct from the original footage. However, video style transfer faced certain issues in its early stages, with one of the most significant problems being the potential for unstable or discontinuous results in the transformed videos, commonly referred to as the "flickering problem" [14].

In 2016, Ruder [150] introduced style transfer techniques to the domain of videos to apply an artistic style from a single image to an entire video sequence. However, independently processing each frame of the video could lead to flickering and erroneous inconsistencies. To address this issue and ensure smooth style transitions, the paper introduced temporal constraints, penalizing differences between adjacent frames. This approach combined image style transfer and optical flow estimation techniques, resulting in continuous artistic style transformations within videos.

To achieve stronger consistency between adjacent frames, the paper introduced an explicit consistency penalty into the loss function. The temporal consistency loss function penalizes deviations in distorted images within regions of consistent and high-confidence optical flow estimates:

$$L_{temporal}(x, \omega, c) = \frac{1}{D} \sum_{k=1}^{D} c_k \cdot (x_k - \omega_k)^2 \tag{27}$$

$c \in [0,1]^D$ represents the weight of each pixel for the loss, where $D = W \times H \times C$ represents the dimension of the image, x is the stylized frame, and ω is the optical flow in the forward direction.

The overall loss is computed using the following losses:

$$\begin{aligned} L_{shortterm}(p^{(i)}, a, x^{(i)}) & \\ &= \alpha L_{Content}(p^{(i)}, x^{(i)}) + \beta L_{Style}(a, x^{(i)}) \\ &\quad + \gamma L_{temporal}(x^{(i)}, \omega_{i-1}^{i}(x^{(i-1)}), c^{(i-1,i)}) \end{aligned} \tag{28}$$

$p^{(i)}$ represents the i-th frame of the original video. This approach ensures temporal consistency by having the style transfer be influenced by the previous frame, helping to control the occurrence of flickering issues in style transfer.

In 2017, Huang [151] explored the possibility of using a feedforward network for video-style transfer. They employed a feedforward network and trained it to ensure temporal consistency between consecutive frames. This method enables real-time processing and produces high-quality style transfer results. The network architecture is depicted in Figure 21.

The style transfer model consists of two main components: a stylization network and a loss network, as shown in the above figure. Firstly, the stylization network takes one frame as thr input and generates the corresponding stylized output. Then, features of the stylized output frame are extracted from a loss network pre-trained on the ImageNet classification task, and loss values used for training the stylization network are computed. These loss values include spatial loss, which assesses the quality of style transfer in the spatial domain and is a weighted sum of content loss and style loss. Content loss measures the high-level content similarity between the input image and stylized output, while style loss quantifies the similarity of style features between the given style image and stylized output. Additionally, the model introduces a new temporal sequence loss to enhance the temporal consistency between consecutive outputs.

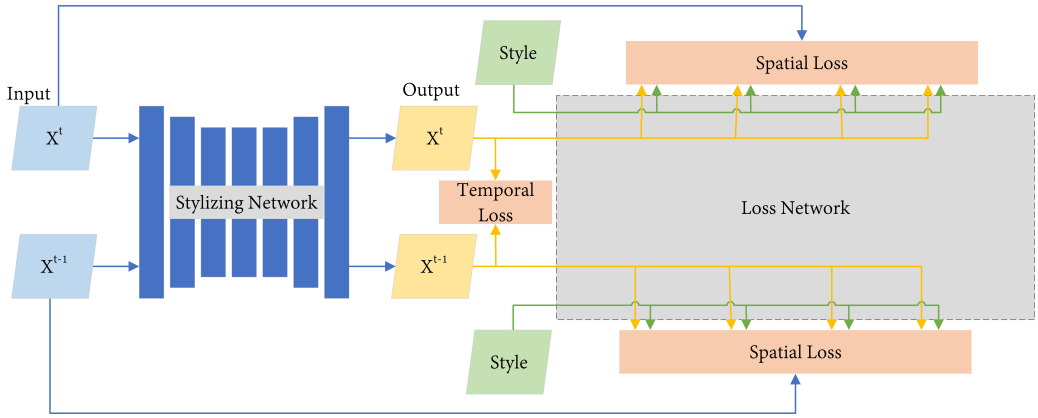

Figure 21. Fast video-style migration network.

As one of the most critical perceptual factors in videos, temporal consistency needs to be taken into account. Therefore, the paper defines a combined loss:

$$L_{hybrid} = \sum_{i\in(t,t-1)} L_{spatial}(X_i, X'_t, S) + \lambda L_{temporal}(X'_t, X'_{t-1}) \tag{29}$$

X_t is the video input frame at time t, X'_t is the corresponding output video frame, and S is the given style image. The paper introduces a new temporal loss to ensure temporal consistency between adjacent stylized output frames, and the temporal loss formula is as follows:

$$L_{temporal}(X'_t, X'_{t-1}) = \frac{1}{D}\sum_{k=1}^{D} c_k \cdot (X'_{tk} - f(X'_{t-1k}))^2 \tag{30}$$

$f(X'_{t-1k})$ is a function that distorts the stylized output based on precomputed optical flow from time $t-1$ to t, $c \in [0,1]^D$ represents the optical flow confidence for each pixel, $D = W \times H \times C$ represents the dimension of the image.

Other researchers have also proposed methods for achieving real-time video-style transfer. Chen [152] introduced a coherent and real-time video-style transfer method. This approach uses a feedforward network and offers a speed improvement of thousands of times compared to optimization-based style transfer methods. Additionally, in 2021, Xia [153] proposed a novel real-time algorithm based on deep neural network architecture. This algorithm transfers the artistic style of semantically meaningful local regions of an image to corresponding local regions of the target video while preserving their realism.

Similar to image style transfer, video style transfer has also rapidly developed methods for multi-style and arbitrary-style transfer. In 2020, Gao [154] introduced a framework called Video Multi-Style Transfer (VMST) that enables fast and versatile video transfer for multiple styles within a single network. This paper designed a Multi-Instance Normalization block (MIN-Block) to learn examples of different styles and used two ConvLSTM modules to address flickering issues.

In 2021, Liu [127] introduced a novel Adaptive Attention Normalization (AdaAttN) module that can perform attention normalization adaptively based on the conditions of each point, enabling arbitrary video style transfer. In the same year, Deng [155] utilized multi-channel techniques for arbitrary video style transfer. Their paper proposed a Multi-Channel Correlation Network (MCCNet) that can train the network to fuse example style features and input content features efficiently, achieving effective video style transfer. In 2023, Kong [156] made improvements to the MCCNet network and introduced an intra-channel similarity loss during training. Qualitative and quantitative evaluations showed that MCCNet performed excellently in arbitrary video and image style transfer tasks.

To overcome the flickering issue in video style transfer, more and more researchers have been working on improving video style transfer algorithms. In 2020, Wang [157,158] introduced a novel time consistency regularization strategy and designed a zero-shot video style transfer framework to better accomplish the task of video style transfer. In the same year, Chen [159] proposed a method to train a lightweight video style transfer network using the knowledge distillation paradigm. This approach enhances the performance of the student network through a teacher network, allowing it to generate stable videos without the need for optical flow modules and with faster processing speeds.

In 2021, Liu [160] introduced global content loss and local region structure loss to train models, aiming to improve their robustness and preserve image structures. Additionally, Xu [161] proposed an online video style transfer method called VTNet, which is based on self-supervised spatiotemporal convolutional networks. VTNet utilizes coherent loss to address the flickering issue. In 2023, Lin [162] made improvements to the base transfer module of MCCNet to preserve semantic structures. They introduced skip connections based on AdaIN and self-similarity loss to further alleviate the flickering problem.

While there have been significant advancements in the field of video style transfer, the challenging flickering issue still exists. Future research directions could include further exploring more accurate optical flow estimation methods, designing more effective temporal consistency constraints, and integrating other computer vision tasks like object detection and motion segmentation with video style transfer to enhance its quality and stability. A summary of relevant articles is shown in Table 19.

Table 19. Summary of video style transfer algorithm.

Author	Algorithm	Contribution
Ruder [150]	A temporal constraint is introduced to penalize the deviation between two frames.	Apply style transfer to the video.
Chen [152]	Ensures consistency over longer periods by incorporating short-term consistency and propagating short-term consistency.	Coherent and real-time video-style transfer.
Huang [151]	It is trained by forcing the output of successive frames to be consistent in time.	It can run in real time.
Xia [153]	A real-time algorithm based on deep neural network architecture is proposed.	Enable real-time video migration while maintaining a sense of realism.
Gao [154]	MIN-Block learns different styles of examples and two ConvLSTM modules to solve flicker problems.	Enables fast and multi-style video transfer over a single network.
Liu [127]	A novel adaptive attention normalization (AdaAttN) module is proposed.	Achieve arbitrary style video transfer.
Deng [155]	A multi-channel correlation network (MCCNet) is proposed.	Use multiple channels for arbitrary video-style transfer.
Kong [156]	MCCNet introduces intra-channel similarity loss.	Improvements were made to the MCCNet network.
Wang [157,158]	Proposed a new regularization strategy and designed a zero-shot video style transfer framework.	Better realize the task of video-style transfer.
Chen [159]	This paper proposes to learn lightweight video transmission network through knowledge sublimation paradigm.	Achieve stable video generation without an optical flow module, and run faster.
Liu [160]	Introducing global content loss and local region structure loss to train the model.	Improve the robustness of the model and preserve the image structure.
Xu [161]	Design coherence loss to make the migration result closer to the original video.	VTNet is proposed.
Lin [162]	Maintain semantic structure during stylization using AdaIN-based skip connections and self-similarity loss.	Further, reduces flickering issues.

5.2. New Applications for Video Transfer

The application of artificial intelligence is increasingly playing a crucial role in creative production [163]. Video style transfer, as a creative and artistic technology, will continue to attract the attention of researchers and application developers. Through continuous innovation and improvement, people actively seek new approaches to address the challenges of video style transfer and strive to expand its applications into new domains.

Video style transfer is currently most commonly achieved using neural networks [164], but researchers are continually exploring various new methods for video transfer. In 2020, Aberman [165] introduced an innovative unpaired method aimed at transferring motion styles from videos to animations. This unpaired approach provides a new solution for applying motion styles from videos to the animation domain. As early as 2013, Bonnel [166] proposed an example-based video color adjustment method that automatically adjusts the color of an entire video based on given sample images. In 2019, Jamriška [167] went further by introducing a new example-based video stylization method that selects key frames from the input video and learns style information from example images for stylizing videos. In 2022, Way [168] used semantic segmentation for video style transfer, enabling the separation of foreground and background to achieve different stylization effects.

Video style transfer has evolved from its initial basic applications to explore new domains. In 2018, Ruder [169] introduced a spherical convolution operation specifically designed for handling style transfer on spherical images. This paper provided valuable methods and insights for further research and applications of artistic style transfer on spherical images. In 2022, Yang [170] introduced a new VToonify framework for exploring challenging high-resolution portrait video style transfer. The framework leveraged StyleGAN's mid-to-high-resolution layers to maintain the integrity of the portrait region in the output transferred video, thus achieving portrait video style transfer.

The research directions in video style transfer are focused on improving the quality and computational efficiency of algorithms, addressing issues related to temporal continuity and consistency, and expanding application scenarios. Additionally, integrating video style transfer with other computer vision tasks such as object detection and semantic segmentation will also drive the development of video style transfer further. A summary of relevant articles is shown in Table 20.

Table 20. Summary of video style transfer applications.

Author	Algorithm	Contribution
Aberman [165]	An innovative pairing-free approach based on GAN designed to transfer motion styles from videos to animations.	Applying the motion style in the video to the field of animation provides a new solution.
Bonnel [166]	Automatically adjust the color of the entire video based on a given sample image.	An example-based video color adjustment method is proposed.
Jamriška [167]	Select key frames from the input video and learn style information from sample images to stylize the video.	Introducing a new example-based approach to video stylization.
Way [168]	Semantic segmentation for video style transfer.	The foreground and background can be separated.
Ruder [169]	A spherical convolution operation suitable for the special properties of spherical images.	It provides valuable methods and ideas for studying and applying style transfer of spherical image.
Yang [170]	A new VToonify framework to study challenging controllable high-resolution portrait video style transfer.	Complete portrait video style migration.

6. Style Transfer Experimental Performance

In the field of style transfer research, experimental performance evaluation is crucial for understanding and comparing the effectiveness of different methods. Through a comprehensive analysis of metrics such as image quality, style accuracy, diversity in generated outputs, and efficiency, trends in technological advancements and innovative directions can be revealed.

The evaluation of style transfer can be categorized into qualitative and quantitative approaches, with both complementing each other to provide a comprehensive understanding of the algorithm's performance. Qualitative evaluation typically focuses on assessing the quality of generated images through human visual perception. In the context of style transfer, qualitative evaluation often relies on intuitive observation of the generated images, emphasizing whether the resulting images or videos retain the stylistic features of the target style while preserving the content and structure of the source image. This evaluation method is concerned with the visual effects of the images or videos, such as whether the generated images appear natural and smooth, whether the style transfer aligns with the expected artistic style, and whether there are any noticeable visual discomforts, such as color distortion or texture fragmentation. Furthermore, qualitative evaluation is often conducted through user surveys or visual rating methods, where human observers score multiple style transfer results to gather subjective feedback. This approach is highly flexible and can reflect human perception of image quality; however, due to its reliance on subjective judgment, its accuracy and consistency may be influenced by individual differences among observers.

In contrast, quantitative evaluation focuses on measuring the effectiveness of style transfer using a set of objective numerical metrics. Common quantitative evaluation metrics include perceptual loss, structural similarity index (SSIM), and style loss, which are primarily used to quantify the quality of the transferred image and the extent to which the style is preserved. For example, perceptual loss measures the difference between the generated image and the target image based on the features extracted by convolutional neural networks, while SSIM evaluates the similarity between the two images in terms of luminance, contrast, and structure. Through these metrics, quantitative evaluation provides more objective and reproducible results, avoiding the biases that may arise from subjective assessments. However, quantitative evaluation also has its limitations, particularly in tasks like style transfer, which is highly artistic and visual in nature. Some traditional evaluation standards, such as MSE or SSIM, may fail to capture the differences in style and artistic expression adequately. Therefore, it is often necessary to combine multiple evaluation metrics to achieve a more comprehensive assessment.

Currently, the field of style transfer has not established a unified and universally accepted quantitative evaluation standard, primarily due to the inherent diversity and complexity of style transfer itself. Style transfer is not merely a simple image transformation task; it involves balancing multiple aspects such as content, style, and details, and these aspects are often assessed based on varying domain-specific and application-driven requirements. For example, the goal of style transfer might be to generate an image visually resembling the style of a particular artist, or to create a video that is stylistically consistent while still maintaining coherent motion. In such cases, different tasks and objectives result in variations in evaluation standards. Therefore, existing evaluation methods are largely optimized for specific needs, such as image style transfer, video style transfer, or image detail preservation, which may require different quantitative evaluation metrics. This chapter will explore the objective experimental performance of various methods in the field and discuss their significant role in advancing both academic research and practical applications.

6.1. Image Style Transfer Experimental Performance

In the field of image style transfer, evaluating experimental performance is crucial for assessing and comparing the effectiveness of different methods. Through a comprehensive analysis of metrics such as the visual quality of generated images, accuracy of style transfer, preservation of content, and generation speed, we can delve into the practical effects of various algorithms in applications like artistic creation and image editing. This section will explore the experimental performance of the current image style transfer methods, providing a more intuitive comparison of the strengths and weaknesses of each algorithm.

In the field of deep learning, key metrics used to evaluate style transfer include image quality, style accuracy, content preservation, and generation speed. Image quality is typically assessed using traditional metrics such as PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index), as well as emerging perceptual quality evaluation metrics like FID (Fréchet Inception Distance) and IS (Inception Score).

The "Perceptual Losses" method proposed by Johnson [80] yields a PSNR of 22.54 and an SSIM of 0.5526 on the BSD100 dataset. The relatively low PSNR indicates a loss in image quality, while the low SSIM suggests weaker structural similarity between the generated image and the original. Liao [116] introduced the Unet-CycleGAN model, which achieves a PSNR of 19.50 and an SSIM of 0.6893. While the higher SSIM indicates good structural preservation, the low PSNR suggests poor retention of image details. In comparison, Zhu [129] proposed the LPAdaIN model, which performs significantly better with a PSNR of 26.5 and an SSIM of 0.787, showcasing clear advantages in both image quality and structural fidelity.

Li [102] introduced the SDP-GAN model, which achieved an Inception score (IS) of 5.77 ± 0.78 and a Fréchet Inception Distance (FID) of 112.42. While the higher IS indicates good diversity of the generated images, the higher FID suggests that there is a considerable distribution gap between the generated and real images. Wang [117] presented the ESA-CycleGAN model, which achieved an IS of 5.2392 and a lower FID of 101.6921. Despite the slightly lower IS, the lower FID shows that the generated images are closer to the real ones, indicating better similarity to real image distributions.

Liu [127] introduced the AdaAttN module, which achieved a motion flow error of 0.0391 in arbitrary style transfer, suggesting that the method can effectively preserve the motion information in images, ensuring higher transfer accuracy. Li [93] proposed a feature transformation-based rapid style transfer method, achieving a covariance matrix difference (Ls) of 6.3, indicating excellent performance in maintaining style accuracy during the transformation. Lastly, Tan [70] developed the Faster-CNN model, which achieved a style recognition transfer rate between 77% and 79%, demonstrating high stability and accuracy in style recognition tasks.These results illustrate the various strengths and trade-offs of different methods in style transfer, highlighting the importance of selecting appropriate evaluation metrics based on the task and application needs.

With the application of diffusion models in style transfer, evaluation methods specific to these models have emerged, such as CLIP and LPIPS. CLIP uses joint understanding of images and text to assess the semantic coherence and visual quality of generated images. LPIPS, based on perceptual image similarity learning, measures the visual similarity between the generated image and the target style, capturing finer perceptual differences. Lower LPIPS values typically indicate higher image quality, with better retention of the target style and details, while higher LPIPS values may suggest significant deviations in quality, with style or details not being preserved well.

In terms of specific results, the evaluation results of the C-S disentangled style transfer framework proposed by Wang [140] on 512 × 512 pixel images show an SSIM of 0.672 and a CLIP score of 0.741. This indicates that the method performs well in structural fidelity

and semantic consistency. The higher CLIP score (0.741) suggests that the generated image aligns well with the textual description semantically, meaning the image effectively reflects the semantic content described by the text in terms of style and content, while the moderate SSIM value indicates that the structural similarity is acceptable, though there may be some loss of details. Chung [144] defined the KL-divergence (CFSD) between content and stylized images and proposed an art style transfer method based on a pre-trained large-scale diffusion model. Its CFSD value is 0.2281, and the LPIPS value is 0.5055, indicating high stability in style transfer. The lower CFSD value suggests that the style transfer method performs well in maintaining a balance between content and style, while the LPIPS value of 0.5055 shows a relatively high perceptual similarity in the generated image.

The DEPM algorithm proposed by Hamazaspyan [145] has an average LPIPS loss of 0.596 over 100 images, indicating high visual consistency in the style transfer task. The higher LPIPS value suggests some loss in visual quality of the generated image, but compared to other methods, it performs better in perceptual consistency, effectively preserving style features in the image. However, there may be some trade-off in terms of detail, resulting in slightly lower perceptual similarity.The HiCAST model proposed by Wang [147] has an LPIPS loss of 0.658, indicating lower visual quality in style transfer. Although this method performs well in generating large-scale images, it may need further optimization in terms of style consistency and perceptual quality.

Image style transfer, despite having various objective evaluation methods, still heavily relies on subjective observation and human perception for evaluation. With technological advancements, researchers and developers are looking forward to introducing more effective objective evaluation methods to accurately measure the visual quality of generated images and the fidelity of artistic styles.The experimental results of image style transfer are shown in Table 21.

Table 21. Summary of image style transfer experimental results.

Author	Method	Database	Result
Tan [70]	Faster-CNN	MNIST, URLs, CIFAR-10, ImageNet	Identify mobility (%): 77–79
Johnson [80]	Perceptual Losses	ImageNet, BSD100, COCO	PSNR: 22.54 SSIM: 0.5526
Li [93]	Arbitrary style fast transfer method based on feature transformation	DTD, COCO	Log(L_S): 6.3
Li [102]	SDP-GAN	CycleGAN	IS: 5.77 ± 0.78 FID: 112.42
Liao [116]	Unet-CycleGAN	monet2photo	PSNR: 19.50 SSIM: 0.6893
Wang [117]	ESA-CycleGAN	orange2apple	IS: 5.2392 FID: 101.6921
Liu [127]	Arbitrary style transfer with the AdaAttN module	MSCOCO, WikiArt	Optical flow error: 0.0391
Zhu [129]	LPAdaIN	MSCOCO, WikiArt	IS: 5.2392 FID: 101.6921
Wang [140]	C-S Disentanglement Style Transfer Framework	ImageNet	PSNR: 26.50 SSIM: 0.787
Chung [144]	Artistic style transfer method based on pre-trained large-scale diffusion model	LAION	CFSD: 0.2281 LPIPS: 0.5055
Hamazaspyan [145]	DEPM	PascalVOC, WikiArt	LPIPS: 0.596
Wang [147]	HiCAST	MSCOCO, WikiArt	LPIPS: 0.658

6.2. Video Style Transfer Experimental Performance

Compared to image style transfer, evaluating algorithms for video style transfer involves greater complexity and challenges. This is because it not only considers image quality and style fidelity but also factors like temporal continuity and motion consistency. Research in video style transfer aims not only to produce visually appealing results but also to maintain the dynamic characteristics and content consistency of the original video.

Video style transfer evaluation criteria typically include Temporal error and Warping error. Temporal error measures the frame rate consistency and motion smoothness between the generated video and the original video, ensuring that the temporal continuity of the video is not broken or inconsistent during the style transformation. This reflects the smoothness and consistency of the generated video in the time dimension, preventing issues like stuttering or unnatural motion due to style transfer. Warping error, on the other hand, focuses on the accuracy of geometric deformation and motion paths in the generated video, ensuring that style transfer does not introduce excessive distortion or deformation. A lower warping error indicates that the shapes, textures, and movements in the video maintain good fidelity during the style transfer process.

In terms of specific evaluation results, the feedforward network style transfer model proposed by Huang [151] was tested on the Sintel dataset, and the average Temporal error across five videos was 0.04968. This indicates that the model performs well in ensuring temporal consistency, effectively avoiding significant temporal discontinuities or motion stuttering in the generated videos. The warping error of the video multi-style transfer (VMST) framework proposed by Gao [154] was 0.0370, showing that the method preserves video geometry well during style transfer with minimal distortion, maintaining low levels of geometric distortion throughout the process.The novel real-time algorithm based on deep neural network architecture proposed by Xia [153] achieved warping error and Temporal Consistency Change (TCC) scores of 0.0009 and 0.688, respectively. These results demonstrate strong performance in both accuracy and temporal consistency, with the model especially excelling in maintaining motion paths and temporal continuity. The MCCNet method proposed by Deng [155] calculated the mean and variance between adjacent frames after rendering, with values of 0.0297 and 0.0054, respectively. This shows that the model performs well in avoiding geometric deformation and maintaining frame-to-frame continuity, further proving its stability and effectiveness in practical applications.

In the future, the evaluation standards for video style transfer may evolve towards more comprehensive and multidimensional directions. In addition to the current focus on temporal error and warping error, there may be introductions of finer-grained metrics for temporal continuity, such as quantitative assessments of motion smoothness, and more detailed analyses of motion consistency. Furthermore, with advancements in deep learning technology, there could be the development of more sophisticated perceptual quality metrics, such as deep neural network-based assessments of video realism, to better capture perceptual differences in generated videos. The experimental results of image style transfer are shown in Table 22.

Table 22. Summary of video style transfer experimental results.

Author	Method	Database	Result
Huang [151]	Feedforward Network Style Transfer Model	Video.net	Temporal error: 0.04968
Gao [154]	VMST	COCO	Warping error: 0.0370
Xia [153]	New real-time algorithm for deep neural network architectures	DAVIS 2017	Warping error: 0.0009 TCC: 0.688
Deng [155]	MCCNet	MSCOCO, WikiArt	Mean: 0.0297 Variance: 0.0054

7. Summary and Future Outlook

Style transfer is a technique that transfers the style of one image or video to another. It holds significant importance in the fields of computer vision and image processing and finds applications in creating artistic effects, image enhancement, image editing, and more. This paper is application-oriented and algorithm-based. It more comprehensively analyzes the development of style transfer and pays attention to the latest progress in style transfer. Moreover, this article pays special attention to the field of video style transfer and introduces the development and new applications of video style transfer in more detail to help people better understand the importance of video style transfer and its broader future development prospects.

This article divides style transfer into image and video directions, providing a detailed overview of image style transfer. It analyzes the development of image style transfer based on applications and algorithms. The article also focuses on the development of video style transfer in the deep learning field, serving as a foundation to better understand the evolution of video consistency issues and guide new directions in style transfer. Throughout the article, it is observed that style transfer methods in various domains have their limitations.

Style transfer methods in machine learning were primarily based on rules and feature engineering, often requiring expertise and complex algorithm design. These methods need to be adjusted and adapted for different styles and tasks, constrained by domain-specific rules and assumptions. Traditional machine learning-based style transfer algorithms, such as stroke rendering, image analogy, image filtering, and texture transfer, each have their strengths and weaknesses. Stroke rendering creates a highly artistic style but tends to distort when handling complex images. Image analogy preserves the details and structure of the content but requires significant computational resources. Image filtering is simple and efficient but fails to express complex styles, resulting in a more limited effect. Texture transfer enhances texture details but may lose the original image's structure, leading to unnatural results. Overall, traditional methods excel in efficiency and specific style representation, but often lack the flexibility and intricacy of deep learning approaches when dealing with complex artistic styles.

However, with the advancement of deep learning, style transfer has gradually shifted from traditional rule-based and feature engineering approaches to methods based on deep neural networks, achieving significant progress.

Deep learning methods have overcome many of the limitations of traditional approaches, offering greater flexibility and generalization capabilities, leading to significant achievements in image and video style transfer tasks. Deep learning-based style transfer algorithms, such as single-network style transfer, fast style transfer, GAN-based style transfer, and attention mechanism-based style transfer, each have distinct advantages and disadvantages. Single-network style transfer learns both content and style features simultaneously by training a single network, which helps preserve the structure and details of the image during the style transfer process. However, it requires longer training times and higher hardware demands. Fast style transfer accelerates the style transfer process by utilizing a pre-trained network, enabling real-time style conversion and making it suitable for efficiency-demanding applications. However, its transfer results may not be as refined as traditional methods, and the style representation is limited by the network architecture. GAN-based style transfer generates more natural and detailed style images using generative adversarial networks, producing more realistic and intricate effects, especially in generating complex styles. Nevertheless, it suffers from training instability and is prone to mode collapse issues. Attention mechanism-based style transfer focuses on key regions of the image, allowing for more precise style transfer and the better handling of local details and overall style balance. However, it comes with high computational costs and

its effectiveness depends on the design of the attention mechanism and the stability of the training process. Overall, deep learning methods excel in achieving superior style transfer results, capable of handling more complex and diverse styles, but they face challenges in computational cost and training stability.

This shows that deep learning style transfer methods still have some shortcomings. Firstly, these methods often require large amounts of training data and computational resources, especially for complex style transformations or diverse style transfer tasks. Secondly, the training process of deep learning methods can be unstable, and prone to issues like mode collapse or non-convergence, necessitating careful hyperparameter tuning and training techniques. Additionally, the results generated by deep learning methods may lack consistency and controllability, making it challenging to achieve fine-grained control over the generated images.

In summary, deep learning-based style transfer methods still face challenges related to training difficulties, instability, consistency, and controllability. Further research and improvements are needed to address these issues. Therefore, the development of style transfer technology can be further strengthened in the following directions:

1. Orthogonality between style transfer algorithms: In the future development of style transfer technologies, considering the orthogonality of different methods is crucial. This means that various techniques should be combined in a way that allows each to leverage its strengths for different objectives while avoiding conflicts and mutual exclusions. For example, generative adversarial networks (such as CycleGAN) could be combined with traditional convolutional neural networks (such as VGG) to separately handle style transfer and content preservation. On the other hand, there is a contradiction between content preservation optimization based on deep convolutional neural networks (DCNN) and content loss (such as VGG feature loss) and extreme style transfer. When the weight of content preservation is too high, the style transformation is suppressed, leading to a lack of sufficient style features in the image. Conversely, if the weight of style is too dominant, the content information of the original image might be lost, making it difficult to achieve a balance between the two. It is essential to properly manage the orthogonality between different methods to maintain a balance between style and content, thus enhancing the diversity and accuracy of the generated results.
2. Data sets and training strategies increase the diversity of training data and enhance the generalization and adaptability of the model. In addition, more effective training strategies are designed to improve the stability and convergence of the model.
3. Model architecture and loss function design: Developing a more powerful and flexible network architecture and exploring new loss functions and optimization objectives can improve the quality, consistency, and controllability of the generated images.
4. Adversarial training and generator optimization: It is critical to improve the stability and convergence of adversarial training to mitigate problems such as mode collapse and training instability. Design a more efficient generator optimization algorithm to improve the quality of the generated images.
5. Cross-domain and multi-modal style transfer: Researching cross-domain and multi-modal style transfer can enable models to flexibly transform between different domains and styles. For example, exploring domain adaptation and transfer learning methods can help to achieve effective style transfer between different domains.

Author Contributions: Conceptualization, Y.X. and M.X.; methodology, Y.X. and M.X.; formal analysis, K.H.; investigation, S.Z.; resources, M.X. and L.W.; data curation, Y.X.; writing—original draft preparation, Y.X.; writing—review and editing, K.H. and L.W.; visualization, M.X.; supervision, M.X.; project administration, M.X.; and funding acquisition, M.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of PR China of grant number 42075130.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

List of Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	**Meaning**
ACGAN	Auxiliary Classifier Generative Adversarial Network
AdaBN	Adaptive Batch Normalization
AdaIN	Adaptive Instance Normalization
AdaSM	Adaptive Style Modulation
AGGAN	Attention-Guided Generative Adversarial Network
BSD	Berkeley Segmentation Dataset
CA	Co-Adaption
CBAM	Convolutional Block Attention Module
CGAN	Conditional Generative Adversarial Network
CIN	Conditional Instance Normalization
CLIP	Contrastive Language-Image Pretraining
DAM	Domain-Attention Module
DE-GAN	Depth Extraction Generative Adversarial Network
DIN	Dynamic Instance Normalization
DST	Deformable Style Transfer
ESA-CycleGAN	Self-Attention Based Cycle-Consistent Generative Adversarial Network
FFHQ	Flickr-Faces-HQ
FID	Fréchet Inception Distance
FMD	Flickr Material Dataset
FST	Filter Style Transfer
GAN	Generative Adversarial Network
IS	Inception Score
LAS	Laser absorption spectrometer
LDA-GAN	Lightweight Domain-Attention Generative Adversarial Network
LPIPS	Learned Perceptual Image Patch Similarity
LSTM	Long Short-Term Memory
MCCNet	Multi-Content Complementation Network
MGAN	Markov Generative Adversarial Network
MINC	Materials in Context Database
MRF	Markov Random Field
MSDA	Multi-Source Domain Adaptation
MSU-GAN	Multi-Style Unsupervised Generative Adversarial Network
NNF	Nearest Neighbor Field
SA	Self-Adaption
SANet	Self-Attention Generative Adversarial Networks
SDP-GAN	Saliency Detail Preservation Generative Adversarial Networks
SSIM	Structural Similarity Index
STTS	Style Transfer via Texture Synthesis

SVD	Singular Value Decomposition
VAE	Variational Auto-Encoders
VMST	Video Multi-Style Transfer
VTNet	Visual Transformer Network
WCT	Whitening and Coloring Transforms

References

1. Gatys, L.A.; Ecker, A.S.; Bethge, M. A neural algorithm of artistic style. *arXiv* **2015**, arXiv:1508.06576. [CrossRef]
2. Heeger, D.J.; Bergen, J.R. Pyramid-based texture analysis/synthesis. In Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 6–11 August 1995; pp. 229–238.
3. Efros, A.A.; Leung, T.K. Texture synthesis by non-parametric sampling. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Corfu, Greece, 20–27 September 1999; Volume 2; pp. 1033–1038.
4. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
5. Hays, J.; Efros, A.A. Scene completion using millions of photographs. *ACM Trans. Graph. (ToG)* **2007**, *26*, 4-es. [CrossRef]
6. Li, C.; Wand, M. Combining markov random fields and convolutional neural networks for image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2479–2486.
7. Raghu, M.; Schmidt, E. A survey of deep learning for scientific discovery. *arXiv* **2020**, arXiv:2003.11755.
8. Pang, Y.; Lin, J.; Qin, T.; Chen, Z. Image-to-image translation: Methods and applications. *IEEE Trans. Multimed.* **2021**, *24*, 3859–3881. [CrossRef]
9. Chen, H.; Zhang, G.; Chen, G.; Zhou, Q. Research progress of image style transfer based on deep learning. *Comput. Eng. Appl.* **2021**, *57*, 37–45.
10. Zhao, C. A survey on image style transfer approaches using deep learning. In *Proceedings of the Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2020; Volume 1453, p. 012129.
11. Jing, Y.; Yang, Y.; Feng, Z.; Ye, J.; Yu, Y.; Song, M. Neural style transfer: A review. *IEEE Trans. Vis. Comput. Graph.* **2019**, *26*, 3365–3385. [CrossRef] [PubMed]
12. Chen, Y.; Zhao, Y.; Jia, W.; Cao, L.; Liu, X. Adversarial-learning-based image-to-image transformation: A survey. *Neurocomputing* **2020**, *411*, 468–486. [CrossRef]
13. Jiao, L.; Zhao, J. A survey on the new generation of deep learning in image processing. *IEEE Access* **2019**, *7*, 172231–172263. [CrossRef]
14. Liu, S. An Overview of Color Transfer and Style Transfer for Images and Videos. *arXiv* **2022**, arXiv:2204.13339.
15. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
16. Tan, W.R.; Chan, C.S.; Aguirre, H.E.; Tanaka, K. Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3703–3707.
17. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
18. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 2223–2232.
19. Caputo, B.; Hayman, E.; Mallikarjuna, P. Class-specific material categorisation. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, Beijing, China, 17–21 October 2005; Volume 2, pp. 1597–1604.
20. Sharan, L.; Rosenholtz, R.; Adelson, E. Material perception: What can you see in a brief glance? *J. Vis.* **2009**, *9*, 784. [CrossRef]
21. Quattoni, A.; Torralba, A. Recognizing indoor scenes. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 413–420.
22. Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; Vedaldi, A. Describing textures in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3606–3613.
23. Bell, S.; Upchurch, P.; Snavely, N.; Bala, K. Material recognition in the wild with the materials in context database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3479–3487.
24. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 25278–25294.
25. ImageNet. Available online: https://www.image-net.org/ (accessed on 25 April 2022).
26. WikiArt. Available online: https://github.com/cs-chan/ArtGAN/tree/master/WikiArt%20Dataset (accessed on 25 April 2022).

27. Pix2pix. Available online: https://www.kaggle.com/datasets/vikramtiwari/pix2pix-dataset (accessed on 25 April 2022).
28. CycleGAN. Available online: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix/blob/master/docs/datasets.md (accessed on 25 April 2022).
29. KTH-TIPS. Available online: https://www.csc.kth.se/cvap/databases/kth-tips/index.html (accessed on 25 April 2022).
30. FMD. Available online: https://drive.google.com/drive/folders/1aygMzSDdoq63IqSk-ly8cMq0_owup8UM (accessed on 25 April 2022).
31. MIT-Indoor. Available online: https://web.mit.edu/torralba/www/indoor.html (accessed on 25 April 2022).
32. DTD. Available online: https://www.robots.ox.ac.uk/~vgg/data/dtd/ (accessed on 25 April 2022).
33. MINC. Available online: http://opensurfaces.cs.cornell.edu/publications/minc/ (accessed on 25 April 2022).
34. COCO. Available online: https://cocodataset.org/#home (accessed on 25 April 2022).
35. LAION-Aesthetics. Available online: https://laion.ai/blog/laion-aesthetics/ (accessed on 25 April 2022).
36. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
37. Zhang, R.; Isola, P.; Efros, A.A. Colorful image colorization. In *Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part III 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 649–666.
38. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
39. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, 5–10 December 2016.
40. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
41. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
42. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (PMLR), Online, 18–24 July 2021; pp. 8748–8763.
43. Foley, J.D.; Van Dam, A.; Feiner, S.K.; Hughes, J.F.; Phillips, R.L. *Introduction to Computer Graphics*; Addison-Wesley: Reading, MA, USA, 1994; Volume 55.
44. Foley, J.D. *Computer Graphics: Principles and Practice*; Addison-Wesley Professional: Reading, MA, USA, 1996; Volume 12110.
45. Hertzmann, A. Painterly rendering with curved brush strokes of multiple sizes. In Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, Orlando, FL, USA, 19–24 July 1998; pp. 453–460.
46. Hertzmann, A. Fast paint texture. In Proceedings of the 2nd International Symposium on Non-Photorealistic Animation and Rendering, Annecy, France, 3–5 June 2002; p. 91-ff.
47. DeCarlo, D.; Finkelstein, A.; Rusinkiewicz, S.; Santella, A. Suggestive contours for conveying shape. In *Seminal Graphics Papers: Pushing the Boundaries*; Association for Computing Machinery: New York, NY, USA, 2023; Volume 2, pp. 401–408.
48. Lu, C.; Xu, L.; Jia, J. Combining sketch and tone for pencil drawing production. In Proceedings of the Symposium on Non-Photorealistic Animation and Rendering, Annecy, France, 4–6 June 2012; pp. 65–73.
49. Jing, Y.; Liu, Y.; Yang, Y.; Feng, Z.; Yu, Y.; Tao, D.; Song, M. Stroke controllable fast style transfer with adaptive receptive fields. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 238–254.
50. Jacobs, C.; Salesin, D.; Oliver, N.; Hertzmann, A.; Curless, A. Image analogies. In Proceedings of the Siggraph, Los Angeles, CA, USA, 12–17 August 2001; pp. 327–340.
51. Hertzmann, A.; Jacobs, C.E.; Oliver, N.; Curless, B.; Salesin, D.H. Image analogies. In *Seminal Graphics Papers: Pushing the Boundaries*; Association for Computing Machinery: New York, NY, USA, 2023; Volume 2, pp. 557–570.
52. Efros, A.A.; Freeman, W.T. Image quilting for texture synthesis and transfer. In *Seminal Graphics Papers: Pushing the Boundaries*; Association for Computing Machinery: New York, NY, USA, 2023; Volume 2, pp. 571–576.
53. Agarwala, A.; Dontcheva, M.; Agrawala, M.; Drucker, S.; Colburn, A.; Curless, B.; Salesin, D.; Cohen, M. Interactive digital photomontage. In *ACM SIGGRAPH 2004 Papers*; Association for Computing Machinery: New York, NY, USA, 2004; pp. 294–302.
54. Darabi, S.; Shechtman, E.; Barnes, C.; Goldman, D.B.; Sen, P. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Trans. Graph. (TOG)* **2012**, *31*, 82. [CrossRef]
55. Liao, J.; Yao, Y.; Yuan, L.; Hua, G.; Kang, S.B. Visual attribute transfer through deep image analogy. *arXiv* **2017**, arXiv:1705.01088. [CrossRef]

56. Burt, P.J.; Adelson, E.H. The Laplacian pyramid as a compact image code. In *Readings in Computer Vision*; Elsevier: Amsterdam, The Netherlands, 1987; pp. 671–679.
57. Tomasi, C.; Manduchi, R. Bilateral filtering for gray and color images. In Proceedings of the Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), Bombay, India, 7 January 1998; pp. 839–846.
58. Winnemöller, H.; Olsen, S.C.; Gooch, B. Real-time video abstraction. *ACM Trans. Graph. (TOG)* **2006**, *25*, 1221–1226. [CrossRef]
59. Semmo, A.; Trapp, M.; Döllner, J.; Klingbeil, M. Pictory: Combining neural style transfer and image filtering. In *ACM SIGGRAPH 2017 Appy Hour*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 1–2.
60. Yim, J.; Yoo, J.; Do, W.j.; Kim, B.; Choe, J. Filter style transfer between photos. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part VI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 103–119.
61. Kwatra, V.; Essa, I.; Bobick, A.; Kwatra, N. Texture optimization for example-based synthesis. In *ACM SIGGRAPH 2005 Papers*; Association for Computing Machinery: New York, NY, USA, 2005; pp. 795–802.
62. Elad, M.; Milanfar, P. Style transfer via texture synthesis. *IEEE Trans. Image Process.* **2017**, *26*, 2338–2351. [CrossRef] [PubMed]
63. Barnes, C.; Shechtman, E.; Finkelstein, A.; Goldman, D.B. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* **2009**, *28*, 24. [CrossRef]
64. Song, Z.C.; Liu, S.G. Sufficient image appearance transfer combining color and texture. *IEEE Trans. Multimed.* **2016**, *19*, 702–711. [CrossRef]
65. Zhang, H.; Dana, K. Multi-style generative network for real-time transfer. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
66. Men, Y.; Lian, Z.; Tang, Y.; Xiao, J. A common framework for interactive texture transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6353–6362.
67. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
68. Gatys, L.A.; Bethge, M.; Hertzmann, A.; Shechtman, E. Preserving color in neural artistic style transfer. *arXiv* **2016**, arXiv:1606.05897.
69. Novak, R.; Nikulin, Y. Improving the neural algorithm of artistic style. *arXiv* **2016**, arXiv:1605.04603.
70. Tan, Y. Feature Recognition and Style Transfer of Painting Image Using Lightweight Deep Learning. *Comput. Intell. Neurosci.* **2022**, *2022*, 1478371. [CrossRef] [PubMed]
71. Luan, F.; Paris, S.; Shechtman, E.; Bala, K. Deep photo style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4990–4998.
72. Kolkin, N.; Kucera, M.; Paris, S.; Sykora, D.; Shechtman, E.; Shakhnarovich, G. Neural Neighbor Style Transfer. *arXiv* **2022**, arXiv:2203.13215.
73. Wang, L. Cartoon-Style Image Rendering Transfer Based on Neural Networks. *Comput. Intell. Neurosci.* **2022**, *2022*, 2958338. [CrossRef] [PubMed]
74. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
75. Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; Madry, A. Adversarial examples are not bugs, they are features. In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.
76. Nakkiran, P. A discussion of'adversarial examples are not bugs, they are features': Adversarial examples are just bugs, too. *Distill* **2019**, *4*, e00019-5. [CrossRef]
77. Wang, P.; Li, Y.; Vasconcelos, N. Rethinking and improving the robustness of image style transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 124–133.
78. Li, H.; Wu, X.j.; Durrani, T.S. Infrared and visible image fusion with ResNet and zero-phase component analysis. *Infrared Phys. Technol.* **2019**, *102*, 103039. [CrossRef]
79. Wang, T.; Ma, Z.; Zhang, F.; Yang, L. Research on Wickerwork Patterns Creative Design and Development Based on Style Transfer Technology. *Appl. Sci.* **2023**, *13*, 1553. [CrossRef]
80. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 694–711.
81. Ulyanov, D.; Lebedev, V.; Vedaldi, A.; Lempitsky, V. Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv* **2016**, arXiv:1603.03417.
82. Li, X.; Liu, S.; Kautz, J.; Yang, M.H. Learning linear transformations for fast image and video style transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3809–3817.

83. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6924–6932.
84. Dumoulin, V.; Shlens, J.; Kudlur, M. A learned representation for artistic style. *arXiv* **2016**, arXiv:1610.07629.
85. Chen, D.; Yuan, L.; Liao, J.; Yu, N.; Hua, G. Stylebank: An explicit representation for neural image style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1897–1906.
86. Wang, X.; Oxholm, G.; Zhang, D.; Wang, Y.F. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5239–5247.
87. Huang, Z.; Zhang, J.; Liao, J. Style Mixer: Semantic-aware Multi-Style Transfer Network. In *Computer Graphics Forum*; Wiley Online Library: New York, NY, USA, 2019; Volume 38, pp. 469–480.
88. Kim, S.; Do, J.; Kim, M. Pseudo-supervised learning for semantic multi-style transfer. *IEEE Access* **2021**, *9*, 7930–7942. [CrossRef]
89. Alexandru, I.; Nicula, C.; Prodan, C.; Rotaru, R.P.; Voncilă, M.L.; Tarbă, N.; Boiangiu, C.A. Image Style Transfer via Multi-Style Geometry Warping. *Appl. Sci.* **2022**, *12*, 6055. [CrossRef]
90. Chen, T.Q.; Schmidt, M. Fast patch-based style transfer of arbitrary style. *arXiv* **2016**, arXiv:1612.04337.
91. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 1501–1510.
92. Ghiasi, G.; Lee, H.; Kudlur, M.; Dumoulin, V.; Shlens, J. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv* **2017**, arXiv:1705.06830.
93. Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; Yang, M.H. Universal style transfer via feature transforms. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
94. Gu, S.; Chen, C.; Liao, J.; Yuan, L. Arbitrary style transfer with deep feature reshuffle. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8222–8231.
95. Jing, Y.; Liu, X.; Ding, Y.; Wang, X.; Ding, E.; Song, M.; Wen, S. Dynamic instance normalization for arbitrary style transfer. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 4369–4376.
96. Zhang, Y.; Hu, B.; Huang, Y.; Gao, C.; Wang, Q. Adaptive Style Modulation for Artistic Style Transfer. *Neural Process. Lett.* **2023**, *55*, 6213–6230. [CrossRef]
97. Wang, X.; Wang, W.; Yang, S.; Liu, J. CLAST: Contrastive Learning for Arbitrary Style Transfer. *IEEE Trans. Image Process.* **2022**, *31*, 6761–6772. [CrossRef]
98. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
99. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014.
100. Wang, X.; Gupta, A. Generative image modeling using style and structure adversarial networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 318–335.
101. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
102. Li, R.; Wu, C.H.; Liu, S.; Wang, J.; Wang, G.; Liu, G.; Zeng, B. SDP-GAN: Saliency detail preservation generative adversarial networks for high perceptual quality style transfer. *IEEE Trans. Image Process.* **2020**, *30*, 374–385. [CrossRef] [PubMed]
103. Han, X.; Wu, Y.; Wan, R. A Method for Style Transfer from Artistic Images Based on Depth Extraction Generative Adversarial Network. *Appl. Sci.* **2023**, *13*, 867. [CrossRef]
104. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
105. Mao, Q.; Lee, H.Y.; Tseng, H.Y.; Ma, S.; Yang, M.H. Mode seeking generative adversarial networks for diverse image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1429–1437.
106. Lv, G.; Israr, S.M.; Qi, S. Multi-style unsupervised image synthesis using generative adversarial nets. *IEEE Access* **2021**, *9*, 86025–86036. [CrossRef]
107. Dong, H.; Neekhara, P.; Wu, C.; Guo, Y. Unsupervised image-to-image translation with generative adversarial networks. *arXiv* **2017**, arXiv:1701.02676.
108. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.

109. Hicsonmez, S.; Samet, N.; Akbas, E.; Duygulu, P. GANILLA: Generative adversarial networks for image to illustration translation. *Image Vis. Comput.* **2020**, *95*, 103886. [CrossRef]
110. Roy, S.; Siarohin, A.; Sangineto, E.; Sebe, N.; Ricci, E. Trigan: Image-to-image translation for multi-source domain adaptation. *Mach. Vis. Appl.* **2021**, *32*, 41. [CrossRef]
111. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In Proceedings of the International Conference on Machine Learning (PMLR), Sydney, Australia, 6–11 August 2017; pp. 1857–1865.
112. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 2849–2857.
113. Tu, C.T.; Lin, H.J.; Tsia, Y. Multi-style image transfer system using conditional cycleGAN. *Imaging Sci. J.* **2021**, *69*, 1–14. [CrossRef]
114. Zhang, T.; Zhang, Z.; Jia, W.; He, X.; Yang, J. Generating cartoon images from face photos with cycle-consistent adversarial networks. *Comput. Mater. Contin.* **2021**, *69*, 2733–2747. [CrossRef]
115. Liu, Y. Improved generative adversarial network and its application in image oil painting style transfer. *Image Vis. Comput.* **2021**, *105*, 104087. [CrossRef]
116. Liao, Y.; Huang, Y. Deep Learning-Based Application of Image Style Transfer. *Math. Probl. Eng.* **2022**, *2022*, 1693892. [CrossRef]
117. Wang, L.; Wang, L.; Chen, S. ESA-CycleGAN: Edge feature and self-attention based cycle-consistent generative adversarial network for style transfer. *IET Image Process.* **2022**, *16*, 176–190. [CrossRef]
118. Zhang, L.; Ji, Y.; Lin, X.; Liu, C. Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier gan. In Proceedings of the 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR), Nanjing, China, 26–29 November 2017; pp. 506–511.
119. Chen, Y.; Lai, Y.K.; Liu, Y.J. Cartoongan: Generative adversarial networks for photo cartoonization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9465–9474.
120. Dong, Y.; Tan, W.; Tao, D.; Zheng, L.; Li, X. CartoonLossGAN: Learning surface and coloring of images for cartoonization. *IEEE Trans. Image Process.* **2021**, *31*, 485–498. [CrossRef] [PubMed]
121. Shu, Y.; Yi, R.; Xia, M.; Ye, Z.; Zhao, W.; Chen, Y.; Lai, Y.K.; Liu, Y.J. Gan-based multi-style photo cartoonization. *IEEE Trans. Vis. Comput. Graph.* **2021**, *28*, 3376–3390. [CrossRef]
122. Li, C.; Wand, M. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part III 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 702–716.
123. He, B.; Gao, F.; Ma, D.; Shi, B.; Duan, L.Y. Chipgan: A generative adversarial network for chinese ink wash painting style transfer. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1172–1180.
124. Yi, R.; Xia, M.; Liu, Y.J.; Lai, Y.K.; Rosin, P.L. Line Drawings for Face Portraits From Photos Using Global and Local Structure Based GANs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3462–3475. [CrossRef] [PubMed]
125. Song, G.; Luo, L.; Liu, J.; Ma, W.C.; Lai, C.; Zheng, C.; Cham, T.J. Agilegan: Stylizing portraits by inversion-consistent transfer learning. *ACM Trans. Graph. (TOG)* **2021**, *40*, 117. [CrossRef]
126. Park, D.Y.; Lee, K.H. Arbitrary style transfer with style-attentional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5880–5888.
127. Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; Ding, E. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6649–6658.
128. Deng, Y.; Tang, F.; Dong, W.; Sun, W.; Huang, F.; Xu, C. Arbitrary style transfer via multi-adaptation network. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2719–2727.
129. Zhu, Q.; Bai, H.; Sun, J.; Cheng, C.; Li, X. LPAdaIN: Light Progressive Attention Adaptive Instance Normalization Model for Style Transfer. *Electronics* **2022**, *11*, 2929. [CrossRef]
130. Li, J.; Wu, L.; Xu, D.; Yao, S. Arbitrary style transfer with attentional networks via unbalanced optimal transport. *IET Image Process.* **2022**, *16*, 1778–1792. [CrossRef]
131. Ye, W.; Liu, C.; Chen, Y.; Liu, Y.; Liu, C.; Zhou, H. Multi-style transfer and fusion of image's regions based on attention mechanism and instance segmentation. *Signal Process. Image Commun.* **2023**, *110*, 116871. [CrossRef]
132. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning (PMLR), Long Beach, CA, USA, 9–15 June 2019; pp. 7354–7363.
133. Tang, H.; Xu, D.; Sebe, N.; Yan, Y. Attention-guided generative adversarial networks for unsupervised image-to-image translation. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
134. Tang, H.; Liu, H.; Xu, D.; Torr, P.H.S.; Sebe, N. AttentionGAN: Unpaired Image-to-Image Translation Using Attention-Guided Generative Adversarial Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 1972–1987. [CrossRef] [PubMed]

135. Zhao, J.; Lee, F.; Hu, C.; Yu, H.; Chen, Q. LDA-GAN: Lightweight domain-attention GAN for unpaired image-to-image translation. *Neurocomputing* **2022**, *506*, 355–368. [CrossRef]
136. Zhang, T.; Yu, L.; Tian, S. CAMGAN: Combining attention mechanism generative adversarial networks for cartoon face style transfer. *J. Intell. Fuzzy Syst.* **2022**, *42*, 1803–1811. [CrossRef]
137. Zhang, F.; Zhao, H.; Li, Y.; Wu, Y.; Sun, X. CBA-GAN: Cartoonization style transformation based on the convolutional attention module. *Comput. Electr. Eng.* **2023**, *106*, 108575. [CrossRef]
138. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning (PMLR), Lille, France, 6–11 July 2015; pp. 2256–2265.
139. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Virtual, 6–12 December 2020; Volume 33, pp. 6840–6851.
140. Wang, Z.; Zhao, L.; Xing, W. Stylediffusion: Controllable disentangled style transfer via diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 7677–7689.
141. Ruta, D.; Tarrés, G.C.; Gilbert, A.; Shechtman, E.; Kolkin, N.; Collomosse, J. Diff-nst: Diffusion interleaving for deformable neural style transfer. *arXiv* **2023**, arXiv:2307.04157.
142. Everaert, M.N.; Bocchio, M.; Arpa, S.; Süsstrunk, S.; Achanta, R. Diffusion in style. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 2251–2261.
143. Yang, S.; Hwang, H.; Ye, J.C. Zero-shot contrastive loss for text-guided diffusion image style transfer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 22873–22882.
144. Chung, J.; Hyun, S.; Heo, J.P. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 8795–8805.
145. Hamazaspyan, M.; Navasardyan, S. Diffusion-enhanced patchmatch: A framework for arbitrary style transfer with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 797–805.
146. Chen, D.Y. Artfusion: Arbitrary style transfer using dual conditional latent diffusion models. *arXiv* **2023**, arXiv:2306.09330.
147. Wang, H.; Wang, H.; Yang, J.; Yu, Z.; Xie, Z.; Tian, L.; Xiao, X.; Jiang, J.; Liu, X.; Sun, M. HiCAST: Highly Customized Arbitrary Style Transfer with Adapter Enhanced Diffusion Models. *arXiv* **2024**, arXiv:2401.05870.
148. Cao, C.; Chen, S.; Zhang, W.; Tang, X. Automatic motion-guided video stylization and personalization. In Proceedings of the 19th ACM International Conference on Multimedia, Scottsdale, AZ, USA, 28 November–1 December 2011; pp. 1041–1044.
149. Kyprianidis, J.E.; Collomosse, J.; Wang, T.; Isenberg, T. State of the "art": A taxonomy of artistic stylization techniques for images and video. *IEEE Trans. Vis. Comput. Graph.* **2012**, *19*, 866–885. [CrossRef]
150. Ruder, M.; Dosovitskiy, A.; Brox, T. Artistic style transfer for videos. In *Proceedings of the Pattern Recognition: 38th German Conference, GCPR 2016, Hannover, Germany, 12–15 September 2016*; Proceedings 38; Springer: Berlin/Heidelberg, Germany, 2016; pp. 26–36.
151. Huang, H.; Wang, H.; Luo, W.; Ma, L.; Jiang, W.; Zhu, X.; Li, Z.; Liu, W. Real-time neural style transfer for videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 783–791.
152. Chen, D.; Liao, J.; Yuan, L.; Yu, N.; Hua, G. Coherent online video style transfer. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1105–1114.
153. Xia, X.; Xue, T.; Lai, W.-S.; Sun, Z.; Chang, A.; Kulis, B.; Chen, J. Real-time localized photorealistic video style transfer. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 1089–1098.
154. Gao, W.; Li, Y.; Yin, Y.; Yang, M.H. Fast video multi-style transfer. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 3222–3230.
155. Deng, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Xu, C. Arbitrary video style transfer via multi-channel correlation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 11–15 October 2021; Volume 35, pp. 1210–1217.
156. Kong, X.; Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Chen, Y.; He, Z.; Xu, C. Exploring the temporal consistency of arbitrary style transfer: A channelwise perspective. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 8482–8496. [CrossRef] [PubMed]
157. Wang, W.; Yang, S.; Xu, J.; Liu, J. Consistent video style transfer via relaxation and regularization. *IEEE Trans. Image Process.* **2020**, *29*, 9125–9139. [CrossRef] [PubMed]
158. Wang, W.; Xu, J.; Zhang, L.; Wang, Y.; Liu, J. Consistent video style transfer via compound regularization. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12233–12240.
159. Chen, X.; Zhang, Y.; Wang, Y.; Shu, H.; Xu, C.; Xu, C. Optical flow distillation: Towards efficient and stable video style transfer. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part VI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 614–630.
160. Liu, S.; Zhu, T. Structure-guided arbitrary style transfer for artistic image and video. *IEEE Trans. Multimed.* **2021**, *24*, 1299–1312. [CrossRef]

161. Xu, K.; Wen, L.; Li, G.; Qi, H.; Bo, L.; Huang, Q. Learning self-supervised space-time CNN for fast video style transfer. *IEEE Trans. Image Process.* **2021**, *30*, 2501–2512. [CrossRef] [PubMed]
162. Lin, H.; Wang, M.; Liu, Y.; Kou, J. Correlation-based and content-enhanced network for video style transfer. *Pattern Anal. Appl.* **2023**, *26*, 343–355. [CrossRef]
163. Rebelo, A.D.P.; Inês, G.D.O.; Damion, D.V. The impact of artificial intelligence on the creativity of videos. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2022**, *18*, 9. [CrossRef]
164. Dong, S.; Ding, Y.; Qian, Y.; Li, M. Video Style Transfer based on Convolutional Neural Networks. *Math. Probl. Eng.* **2022**, *2022*, 8918722. [CrossRef]
165. Aberman, K.; Weng, Y.; Lischinski, D.; Cohen-Or, D.; Chen, B. Unpaired motion style transfer from video to animation. *ACM Trans. Graph. (TOG)* **2020**, *39*, 64:1–64:12. [CrossRef]
166. Bonneel, N.; Sunkavalli, K.; Paris, S.; Pfister, H. Example-based video color grading. *ACM Trans. Graph.* **2013**, *32*, 39:1–39:12. [CrossRef]
167. Jamriška, O.; Sochorová, Š.; Texler, O.; Lukáč, M.; Fišer, J.; Lu, J.; Shechtman, E.; Sỳkora, D. Stylizing video by example. *ACM Trans. Graph. (TOG)* **2019**, *38*, 107. [CrossRef]
168. Way, D.L.; Chang, R.J.; Chang, C.C.; Shih, Z.C. A video painterly stylization using semantic segmentation. *J. Chin. Inst. Eng.* **2022**, *45*, 357–367. [CrossRef]
169. Ruder, M.; Dosovitskiy, A.; Brox, T. Artistic style transfer for videos and spherical images. *Int. J. Comput. Vis.* **2018**, *126*, 1199–1219. [CrossRef]
170. Yang, S.; Jiang, L.; Liu, Z.; Loy, C.C. Vtoonify: Controllable high-resolution portrait video style transfer. *ACM Trans. Graph. (TOG)* **2022**, *41*, 203. [CrossRef]