


Article

A Training-Free Latent Diffusion Style Transfer Method

Zhengtao Xiang¹, Xing Wan^{1,2}, Libo Xu^{2,*} , Xin Yu² and Yuhan Mao³

¹ School of Electrical and Information Engineering, Hubei University of Automotive Technology, Shiyan 442002, China; xiangzt_dy@huat.edu.cn (Z.X.); wxhuat@gmail.com (X.W.)

² School of Computer and Data Engineering, Ningbo Tech University, Ningbo 315100, China; yuxin@zju.edu.cn

³ School of Economics and Management, Zhejiang Sci-Tech University, Hangzhou 310018, China; liaohuio_o@163.com

* Correspondence: xlb@nbt.edu.cn; Tel.: +86-0574-88130956

Abstract: Diffusion models have attracted considerable scholarly interest for their outstanding performance in generative tasks. However, current style transfer techniques based on diffusion models still rely on fine-tuning during the inference phase to optimize the generated results. This approach is not merely laborious and resource-demanding but also fails to fully harness the creative potential of expansive diffusion models. To overcome this limitation, this paper introduces an innovative solution that utilizes a pretrained diffusion model, thereby obviating the necessity for additional training steps. The scheme proposes a Feature Normalization Mapping Module with Cross-Attention Mechanism (INN-FMM) based on the dual-path diffusion model. This module employs soft attention to extract style features and integrate them with content features. Additionally, a parameter-free Similarity Attention Mechanism (SimAM) is employed within the image feature space to facilitate the transfer of style image textures and colors, while simultaneously minimizing the loss of structural content information. The fusion of these dual attention mechanisms enables us to achieve style transfer in texture and color without sacrificing content integrity. The experimental results indicate that our approach exceeds existing methods in several evaluation metrics.

Keywords: deep learning; generative model; style transfer; diffusion model; feature fusion; attention mechanism



Citation: Xiang, Z.; Wan, X.; Xu, L.; Yu, X.; Mao, Y. A Training-Free Latent Diffusion Style Transfer Method. *Information* **2024**, *15*, 588. <https://doi.org/10.3390/info15100588>

Academic Editor: Marco Leo

Received: 5 August 2024

Revised: 13 September 2024

Accepted: 18 September 2024

Published: 26 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the wake of the swift progression of artificial intelligence technologies, a plethora of pioneering applications have surfaced within the spheres of computer vision and image processing. Among them, style transfer technology, as a method that can apply artistic styles to images, has attracted widespread attention [1]. Traditional style transfer methods, however, often depend on extensive training datasets and intricate optimization processes, which limit their applicability and increase computational expenses and time. Thus, the development of a training-free style transfer approach that ensures rapid, efficient, and cost-effective style migration is of great research significance and practical value.

In recent years, the burgeoning class of diffusion-based generative models has emerged with significant potential and research value in the field of style transfer. Scholarly works, notably those cited as [2–4], have presented groundbreaking strategies aimed at altering the stylistic presentation of a specified image. These strategies are designed to adeptly shift the aesthetic qualities of the image, all the while safeguarding the inherent features that define its content. The introduction of these methodologies represents a significant advancement in image processing, enabling a more nuanced approach to style modification that preserves the core essence of the visual content. However, these methods still require gradient-based optimization to fine-tune each styled image and invert text, which is highly time-consuming.

Recently, Chung et al. [5] introduced an untrained style transfer method that achieves a certain degree of effective fusion between content and style. Nevertheless, when the

divergence between the intrinsic image and the desired style is pronounced, the resultant images frequently adhere closely to the original stylistic elements, which can lead to discordances in color and deformations in the content. This suggests that the seamless incorporation of stylistic traits remains an unattained goal. This manuscript delves into the expansion of style transfer methodologies that do not require training, harnessing the power of large-scale, pre-trained latent diffusion models. We introduce an innovative latent diffusion style transfer technique, with our key contributions encapsulated as follows:

First, we designed a novel dual-path diffusion model. Unlike previous approaches that collect prior knowledge from forward processes, we align the feature representations of content images, style images, and generated images using prior knowledge from the same step and time step during the diffusion process.

Second, we designed a Feature Normalization Mapping Module with Cross-Attention Mechanism (INN-FMM). Unlike traditional cross-attention mechanisms, it introduces a preprocessing step to align content and style features numerically, thereby achieving effective transfer results.

Third, we introduced a parameter-free attention mechanism (SimAM) [6] that computes attention weights based on local self-similarity of feature maps, helping the model focus on key features in images. SimAM addresses minor content distortions and color mismatches typically encountered during style transfer.

Fourth, the experimental results demonstrate that our method outperforms existing approaches across multiple evaluation metrics, including Art-based Fréchet Inception Distance (ArtFID), Fréchet Inception Distance (FID), Learned Perceptual Image Patch Similarity (LPIPS), Cross-Domain Fréchet Inception Distance (CFSD), time, and the balance between style and content [7–10].

The subsequent sections of this document are structured in the following manner. The second part will provide an overview of the pertinent scholarly endeavors in the field. The third part presents the specific implementation details of our latent diffusion style transfer method. The fourth part describes the experiments and settings, comparing our approach with existing methods. The fifth part concludes the paper.

2. Related Work

Style transfer is an innovative visual computation technique that merges the intrinsic content of one image with the aesthetic attributes of another. Early research in style transfer was based on rules and handcrafted feature representations, such as texture descriptors and filter combinations [11]. However, these approaches were constrained by the crafted features and struggled to encapsulate the sophisticated semantic nuances within images. They relied on manually defined feature extraction methods, which limited the comprehension and articulation of the deeper meanings embedded in the visual content. With the advent of deep learning, particularly the triumphant application of Convolutional Neural Networks (CNNs), style transfer methods [12] that leverage deep learning have achieved notable advancements. These approaches harness the formidable feature extraction capabilities of deep neural networks to delve into and comprehend the intrinsic structures and stylistic traits of images, thereby attaining unprecedented success in the task of image style transfer [13]. In the work of Gatys [1] and colleagues, Gram matrices in conjunction with the Visual Geometry Group 16-layer network (VGG-16) model were employed to distill features from both stylistic and content-rich images. By refining a composite objective function that integrates content and style losses, they successfully crafted new images. This network can extend to multiple styles and produce images with high perceptual quality, but it suffers from slow training speed and poor algorithm robustness. Gatys et al. [1] discovered that the content and style of convolutional neural networks can be separated and independently manipulated, enabling the generation of a new perceptually meaningful image. Li et al. [14] utilized the Laplacian pyramid to decompose the original image into a series of lower-resolution versions, addressing the issue of missing low-level information. In the research by Risser [15] and colleagues, the histogram loss function was incorporated into the model to represent

the distributional information of image features. This approach effectively enhanced the robustness of the Gram matrix during the style transfer process. However, such models can only perform style transfer on images with specific pre-trained styles and lack generalizability. Therefore, these models are also known as single-model single-style style transfer models.

Single-model multi-style style transfer models aim to improve the model's generalizability by incorporating multiple styles into one model. In the study conducted by Dumoulin [16] and co-authors, they introduced an innovative normalization technique known as Conditional Instance Normalization (CIN). Within this method, the Instance Normalization (IN) layer plays a pivotal role, dynamically selecting parameters that dictate the stylistic attributes of the generated image. Chen et al. [17] matched multiple sets of parameters in the StyleBank layer with multiple different style features. When adding a new style transfer, the model only needs to train one StyleBank layer. However, as the variety of transferrable styles expands, the model's footprint and the count of its parameters escalate. Chen [18] and colleagues were the first to introduce a single-model arbitrary style transfer framework. They devised an optimization objective grounded in local matching, utilizing a pre-trained Visual Geometry Group (VGG) network to extract the content structure and style texture of an image, then aligning each content structure with the most proximate style texture. This technique was dubbed "Style Swap", offering a novel perspective and solution for style transfer tasks. Drawing inspiration from Instance Normalization (IN), Huang [19] and co-authors introduced Adaptive Instance Normalization (AdaIN). With an encoder–decoder architecture, AdaIN facilitates real-time arbitrary style transfer without the need to learn any affine parameters. Zhang et al. [20] proposed an enhanced style transfer mechanism using contrastive learning, which projects features from different levels into separate latent style spaces, thereby capturing both local and global style characteristics more comprehensively. The Any-to-Any [21] Style Transfer strategy enables users to interactively select regions of the style image and apply them to specified content regions. This allows different areas of a single image to be assigned different styles. In the research by Li [22] and colleagues, they introduced a novel transformation technique known as Whitening and Coloring Transforms (WCT). This method aligns the statistical distributions and correlations of intermediate features between content and style images, capturing stylistic characteristics from higher to lower levels. By comprehensively grasping the nuances of style across different levels, WCT offers enhanced generalization capabilities in style transfer tasks. Liu et al. [23] introduced the Adaptive Attention Normalization (AdaAttN) mechanism, which improves style transfer by aligning the attention maps of content and style images. Zhu Zhongxian [24] and colleagues, addressing the challenge of structural consistency and drawing inspiration from Han [25] and co-authors, introduced a bidirectional network model grounded in contrastive learning. Leveraging the foundation of Cycle-Consistent Generative Adversarial Network (CycleGAN), they implemented a bidirectional training approach to thoroughly capture the mappings of corresponding regions. They also introduced a new joint contrastive loss to better utilize image information and enhance the quality of style transfer.

3. Training-Free Approach

In this section, we introduce our proposed training-free approach for latent diffusion style transfer. First, we provide a comprehensive overview of diffusion models and latent diffusion models, which serve as the foundational theories for our method. Following this theoretical introduction, we present the detailed structure of our approach, highlighting key components such as the Normalization Feature Mapping and the SimAM Attention Fusion. These components form the core of our method, enabling effective style transfer without the need for additional training processes.

3.1. Diffusion Models

Diffusion models have demonstrated remarkable success in generating images from text [26–28] and in editing images [29–35]. In this domain, neural style transfer has progressed by harnessing the generative prowess of pre-trained diffusion models. For example,

InST [4] introduced a text inversion-based approach, aiming to map a given style to the corresponding text embeddings. StyleDiffusion [3] calculates the noise distribution adapted to a set of target-style images and then fine-tunes the U-Net to generate images in the corresponding target style. Qi et al. [36] used Quantized Formers (Q-Formers) for paired image training, first extracting decoupled feature representations and then injecting them into subsets of mutually exclusive cross-attention layers to better decouple style and semantics. In contrast, Jeong [37] and colleagues introduced a style transfer method that does not require training, utilizing the h-space to convey stylistic information without the need for direct connections. In this paper, we harness the generative potential of latent diffusion to achieve training-free style transfer.

The Latent Diffusion Model [5] is a model that generates data in a latent space through an iterative process. An image $x \in \mathbb{R}^{H \times W \times 3}$ is encoded by encoder \mathcal{E} into a reduced-dimensional latent space $Z = \mathcal{E}(x)$, where H denotes the height and W denotes the width. In this latent space, the model gradually transforms the latent representation into a simple distribution (usually a Gaussian distribution) by progressively adding noise. Subsequently, the model generates structured data from the simple distribution through a reverse process. Ultimately, the encoded latent space representation is reconstructed back into the original data domain $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(z))$ using a decoder \mathcal{D} .

The reverse process is accomplished by a noise estimation network, which infers the true distribution $q(x_t | x_{t-1})$. Starting from random noise, the network progressively denoises the input, eventually generating a realistic sample. The corresponding equation is presented in Equation (1).

$$p(x_{t-1} | x_t) = N[x_{t-1}; \mu_\theta(x_t, t), \sigma^2 I] \quad (1)$$

In the equation, $p(x_{t-1} | x_t)$ represents the posterior distribution, N denotes a normal distribution, x_{t-1} represents the state variable at time $t - 1$, x_t represents the state variable at time t , $\mu_\theta(x_t, t)$ is the parameterized estimated mean, which is indirectly obtained by predicting the noise $\varepsilon_\theta(X_t, t)$ through the noise estimation network, σ^2 denotes the variance and I denotes the identity matrix for the variance.

The network training formula is shown in Equation (2). The process involves iteratively repeating the steps from time $t - 1$ to 0, transforming x_t back into x_{t-1} (computing x_{t-1} and using the resulting x_{t-1} as the new x_t).

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left[x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \varepsilon_\theta(x_t, t) \right] \quad (2)$$

In the equation, α_t denotes an element of the state transition matrix at time t , $\bar{\alpha}_t$ represents an average or summation of the state transition matrix elements at time t , and $\varepsilon_\theta(X_t, t)$ denotes a parameterized error term that depends on the state X_t and time t .

Employing a set of denoising autoencoders $\varepsilon_\theta(x_t, t); t = 1 \dots T$ trained to denoise the input z_t , where z_t is a noisy version derived from Z , and t is randomly drawn from $\{1, \dots, T\}$. The associated training objective is:

$$L_{\text{LDM}} = \mathbb{E}_{z, \varepsilon, t} [\| \varepsilon - \varepsilon_\theta(z_t, t, y) \|_2^2]_{\tau_\theta(y)} \quad (3)$$

In the equation, L_{LDM} represents the loss function for the Latent Diffusion Model. The expectation $\mathbb{E}_{z, \varepsilon, t}$ is taken over the latent variable z , noise ε , and time step t . The term $\varepsilon_\theta(z_t, t, y)$ refers to the predicted noise at time step t , parameterized by θ , conditioned on the latent variable z_t and an auxiliary variable y . The squared L_2 -norm $\| \cdot \|_2^2$ measures the discrepancy between the true noise ε and the predicted noise $\varepsilon_\theta(z_t, t, y)$. Additionally, $\tau_\theta(y)$ denotes a weighting function that adjusts the contribution of different noise levels based on y .

In our work, we employ latent diffusion and introduce cross-attention layers in the model architecture, where y represents a text and t denotes a U-Net architecture. The pro-

cess involves projecting y into an intermediate representation and subsequently mapping it onto the intermediate layers of the U-Net through cross-attention mechanisms.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \quad (4)$$

In the equation, Q, K and V denote the Query matrix, Key matrix, and Value matrix, respectively, while d represents the scaling factor. $Q = W_Q^{(i)} \cdot \varphi_i(z_t)$, $K = W_K^{(i)} \cdot \tau_\theta(y)$, $V = W_V^{(i)} \cdot \tau_\theta(y)$, $W_V^{(i)} \in \mathbb{R}^{d \times d_\xi^i}$, $W_Q^{(i)} \in \mathbb{R}^{d \times d_\tau}$ and $W_K^{(i)} \in \mathbb{R}^{d \times d_\tau}$ is a learnable projection matrix. $\varphi_i(z_t) \in \mathbb{R}^{N \times d_\xi^i}$ represents the intermediate representation of ϵ_θ achieved by U-Net. We do not utilize any textual conditions in this study, and therefore y is always a null value.

3.2. Detailed Work

Figure 1 presents a schematic overview of the entire methodology being proposed. First, the content and style images are mapped to the latent space through an encoder and transformed into Gaussian noise Z_c and Z_s via a diffusion process [28]. Next, the AdaIN technique is applied to convert Z_c and Z_s into the initial latent noise Z_{cs} for the stylized image, guiding the generation of the stylized image. During the reverse diffusion process of Z_{cs} , we utilize an attention mechanism to inject style and content information. Specifically, at each time step, we replace the keys K_{cs} and values V_{cs} of the stylized image with the style features K_s and V_s of t at the same time step. Simultaneously, the query Q_{cs} is replaced with the query Q_c of the content features to maintain the integrity of the content information. To tackle the possible issue of magnitude diminution due to feature substitution, we incorporated a scaling parameter T for the attention map. Ultimately, by harnessing the SimAM attention mechanism for the amalgamation of content and style features, we ensure the preservation of the content’s structural coherence and effectively mitigate the discordance in coloration. The INN-FMM is a feature mapping module composed of a cross-attention mechanism [38] and instance normalization [39], enabling the injection of style into the content image. The SimAM module, which draws inspiration from neuroscientific principles in its three-dimensional attention mechanism, adaptively modulates the distribution of attention. This ensures that the integrity of the content is preserved during the transfer of the stylistic image’s texture and color.

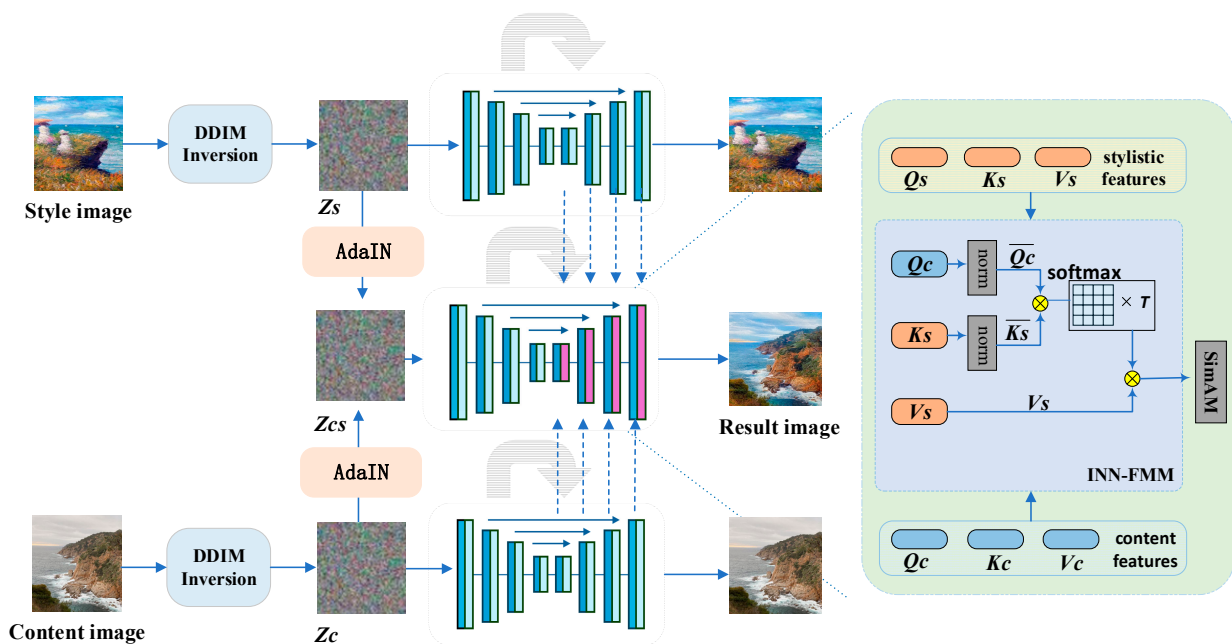


Figure 1. The structural flowchart.

3.2.1. Normalization Feature Mapping

In the realm of image synthesis, the diffusion conditioning mechanism [29] is celebrated for its capacity to direct the generation of images, facilitated by specific inputs y , such as textual descriptions, semantic maps, or other conditional parameters. The core of this process lies in first preprocessing the input y through an encoder specific to the domain, transforming it into an intermediate representation compatible with the model's internal features. Specifically, we employ a cross-attention mechanism that adeptly substitutes the keys and values from the conditions with those present in the original feature map. This substitution operation essentially injects the semantic information of the conditions y into the intermediate layers of the U-Net, achieving precise control over the image synthesis process [29].

Inspired by this mechanism, we further explored the application of self-attention layers in style transfer. We regard the features in the style image as a condition and inject these conditions into the reverse process of denoising generation through the attention layer. Utilizing this approach, the stylistic characteristics of color and texture from the style image are transferred to the content image, thus merging content with style. To this end, this paper proposes a feature normalization mapping module based on the cross-attention mechanism (INN-FMM), which consists of a cross-attention module and an instance normalization module.

As depicted in Figure 1, the latent representations for both the content and style images are initially derived through the Denoising Diffusion Implicit Models (DDIM) inversion process [40]. Subsequently, features from these images are extracted during the DDIM inversion. Specifically, at a pre-specified time step $t = \{0, \dots, 50\}$, both the style and content images are progressively reverted from the original image ($t = 0$) towards the Gaussian noise ($t = 50$). During this process, the content queries Q_c as well as the style key and value features (K_s, V_s) at each time step are collected. AdaIN is used to initialize the style and content potential noise. Finally, during the entire reverse process of executing the programmatic potential noise Z_{cs} , the K_s, V_s , and Q_c collected at each time step are injected into the attention layer to achieve the transfer of the target style.

(1) Instance Normalization

In style transfer tasks, relying solely on the cross-attention mechanism may be insufficient to produce high-quality stylized images. Based on experimental findings, the direct application of cross-attention may excessively emphasize the color and texture features of the original content image, thereby impacting the effectiveness of style transfer. To mitigate this issue, we introduced a preprocessing step to normalize the queries from the content image and the keys from the style image. This step aims to make the two features more consistent in numerical scale, thereby facilitating their integration. The transformation process of the instance normalization submodule is as follows:

$$\overline{Q_c} = \text{norm}(Q_c) \quad (5)$$

$$\overline{K_s} = \text{norm}(K_s) \quad (6)$$

In the above two equations, Q_c denotes the content Query, and K_s denotes the style Key. Through this normalization process, the numerical differences between content and style features are reduced, which helps to balance the weights between the two and prevents excessive residue of content features.

(2) Cross-Attention

In the domain of image synthesis, the cross-attention mechanism is pivotal for adeptly capturing semantic information between content feature maps and style feature maps, facilitating the generation of stylized content feature maps. This mechanism takes two input sequences and maps them into query, key, and value matrices through different linear transformations. The similarity matrix between the Q_c and K_s matrices is then computed, and the V_s matrix is multiplied by the attention weight matrix, which is processed by the

softmax function, to obtain the weighted output matrix. The calculation formula can be expressed as follows:

$$A = \text{Softmax}(\overline{Qc} \otimes \overline{Ks}) \times T \quad (7)$$

$$M = V_s \otimes A^T \quad (8)$$

In the above two equations, the symbol ' \otimes ' denotes matrix multiplication, ' T ' represents the attention scaling parameter, and V_s denotes the style Value. Utilizing the cross-attention mechanism, the color and texture information encoded in the style image can be imparted onto the content image, thereby accomplishing the fusion of content and style.

3.2.2. SimAM Attention Fusion

In the domain of image style transfer, diffusion models are renowned for their exceptional generative capabilities. Nevertheless, maintaining a high level of consistency with the original content image in terms of content remains challenging when transferring one style to another image. Our objective is to accurately capture and preserve the core structure and essential features of the content image during the style transfer process, thereby achieving a seamless integration of style and content.

SimAM is a lightweight, parameter-free convolutional neural network attention mechanism [6] designed to provide a more computationally efficient way to enhance feature representation in content regions for deep learning models. The core concept of SimAM is based on the local self-similarity within images. In images, adjacent pixels typically exhibit higher similarity, whereas similarity decreases between pixels that are farther apart. SimAM calculates attention weights by assessing the similarity between each pixel and its neighboring pixels in the feature map. The attention mechanism of SimAM can be defined through the following energy function:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{u})^2 + 2\hat{\sigma}^2 + 2\lambda} \quad (9)$$

where e_t^* is the optimal estimate of the prediction error, $\hat{\sigma}^2$ the variance of the estimate, λ the regularization parameter, t the current time point, and \hat{u} the control input of the estimate.

According to the formula, lower energy values indicate greater differentiation between the neuron t and its surrounding neurons, thereby signifying a higher importance of the neuron. Therefore, neuron importance can be determined based on $1/e_t^*$. Furthermore, according to the definition of the attention mechanism, we need to enhance the following features:

$$\tilde{X} = \text{sigmoid}\left(\frac{1}{E}\right) \odot X \quad (10)$$

where \tilde{X} denotes the transformed input data after the application of the activation function; E denotes a constant; X denotes the original input data.

After the injection of the attention style, we faced slight distortions in the image content structure and some disharmonies in color. To address these challenges, we immediately introduced SimAM after the cross-attention module to dynamically adjust the distribution of attention. This adjustment strategy ensures that the integrity of the content remains undamaged while harmonizing the colors to achieve a visually cohesive unity.

4. Experiments and Results Analysis

In this section, we conduct a series of experiments to evaluate the performance of our training-free latent diffusion style transfer method. We begin by describing the datasets and experimental setup used for evaluation. Next, we present the Style Arbitrariness Experiment to verify the generalization capability of our approach. This is followed by a Comparative Experiment, where we qualitatively and quantitatively assess the effectiveness and advancement of our method against existing techniques. Additionally, we perform an

ablation study to validate the contribution of individual components, and a balance test between style and content to explore trade-offs. We examine the impact of the attention injection step through a dedicated experiment. Lastly, we discuss the selection of datasets and the details of the model choices for the inference time comparison experiments to provide a comprehensive understanding of our methodology and results.

4.1. Dataset and Experimental Setup

This study conducted experiments using the Microsoft Common Objects in Context (MS-COCO) [41] dataset for content images and the WikiArt Dataset [42] for style images. We employed a latent diffusion model, utilized a pre-trained text-to-image model, and conducted 50-step DDIM sampling.

4.2. Style Arbitrariness Experiment

In this study, a set of images was randomly selected from the MS-COCO and WikiArt datasets to serve as content and style images. Specifically, five style images and three content images were selected to evaluate the method's generalization performance. As depicted in Figure 2, the approach successfully mimics arbitrarily chosen styles, effectively capturing and reproducing the color distribution and texture characteristics of the style images.



Figure 2. The method in this paper produces images of arbitrary style transfer.

4.3. Comparative Experiment

To comprehensively validate the effectiveness and advancement of our method, we conducted a series of comparative experiments evaluating our approach against several popular style transfer methods. Below are the methods involved in the comparative experiments and their brief characteristics:

AdaIN [19]: Adjusts the mean and variance of the content image to match the statistical properties of the style image, thereby transferring the texture and color distribution of the style image to the content image.

StyTR² [43]: Integrates the self-attention mechanism of transformers with multi-scale processing strategies to achieve style transfer.

Diffusion [44]: Based on diffusion models, it introduces a Contrastive Language-Image Pretraining (CLIP)-based style separation loss to decouple style and content.

StyleID [5]: Utilizes a unique attention injection approach within a diffusion model to perform style transfer.

InST [4]: Also based on a diffusion model, this method converts style into learnable textual descriptions and uses text inversion for image style transfer.

DiffStyle [28]: Utilizes a diffusion model to leverage hidden space and adjusts skip connections to convey style and content information separately.

MAST [45]: The core of the paper is the introduction of a multi-adaptation network that achieves a seamless fusion of content and style through multi-level adaptive adjustments, enabling arbitrary style transfer.

CAST [20]: The core of the paper is the introduction of a domain-enhanced arbitrary image style transfer method using contrastive learning, which improves style representation and content preservation by incorporating domain features.

4.3.1. Quantitative Comparison

(1) Comparison of ArtFID and other metrics: To ensure the fairness of the comparison when evaluating style transfer methods, we adopted a series of recently proposed metrics, including ArtFID, FID, LPIPS, and CFSD. FID measures the similarity between generated images and real images. Lower FID values mean the generated images are closer to real ones. ArtFID is a version of FID that focuses on evaluating the artistic quality of generated images, particularly in style transfer tasks. LPIPS measures how similar two images are from a perceptual perspective. Lower values mean the images look more alike to humans. CFSD evaluates how well a generated image balances content retention and style transformation. Lower values indicate a better balance. As shown in Table 1, our method demonstrates superiority over traditional style transfer methods in terms of the ArtFID evaluation metric, closely matching human visual preferences. Additionally, our approach exhibits lower FID values, indicating a high level of consistency between the generated images and the target style. Our method scores significantly lower on CFSD compared to existing techniques, highlighting our advantage in maintaining spatial consistency of the content image. Furthermore, the LPIPS scores further demonstrate our method's ability to preserve content integrity during style transfer. The experimental results indicate that the method performs exceptionally well in comparison.

Table 1. Comparison of Metric Results.

Metrics	Ours	AdaIN	StyTR ²	MAST	CAST	DiffuseIT	StyleID
ArtFID ↓	28.124	30.933	30.720	31.282	34.685	40.721	28.801
FID ↓	17.891	18.242	18.890	18.199	20.395	23.065	18.131
LPIPS ↓	0.4705	0.6076	0.5445	0.6293	0.6212	0.6921	0.5055
CFSD ↓	0.2156	0.3155	0.3011	0.3043	0.2918	0.3428	0.2281

In the figure, arrow ↓ indicates that the lower the metric, the better the style transfer performance.

(2) Time Comparison: Due to their complex network structures and multiple iterative steps, diffusion models generally take longer for synthesis inference compared to traditional methods. Using a single NVIDIA GeForce RTX 4090 GPU, which is manufactured by NVIDIA Corporation based in Santa Clara, California, United States, we conducted inference time measurements for a pair of content and style images. In these experiments, we examined a total of 20 images with a resolution of 640×480 . We performed the inference measurements five times for each image pair and recorded the minimal, maximal, average times, and standard deviation from these measurements. As indicated in Table 2, our method achieves an average total inference time of 60.59 s, which is notably faster compared to other diffusion-based approaches. These detailed statistics provide a comprehensive understanding of the inference time performance.

Table 2. Comparison of Time Results.

Metrics	DiffuseIT	InST	DiffStyle	Ours	Unit
Average time	149.02	153.55	66.89	60.59	second
Minimum time	145.00	150.00	64.00	58.00	second
Maximum time	153.00	157.00	70.00	63.00	second
Standard deviation	2.50	2.80	2.00	1.80	second

Based on the data, we can provide an analysis of the performance of our model as follows:

1. **Minimum Time:** Our model achieves a minimum inference time of 58.00 s, which is faster than DiffuseIT (145.00 s) and InST (150.00 s), and faster than DiffStyle (64.00 s). This indicates that our model performs efficiently in the best-case scenarios, achieving the fastest minimum time among all models.
2. **Maximum Time:** The maximum inference time for our model is 63.00 s. This is lower than the maximum times recorded for DiffuseIT (153.00 s) and InST (157.00 s), and also lower than DiffStyle (70.00 s). This suggests that our model generally maintains efficient performance and does not experience the longest inference times compared to all models.
3. **Standard Deviation:** Our model's standard deviation of 1.80 s is the lowest among the models compared. This indicates less variability in the inference times compared to DiffuseIT (2.50 s), InST (2.80 s), and DiffStyle (2.00 s). A smaller standard deviation reflects greater consistency and stability in performance, which is advantageous for reliable and predictable inference times.

In summary, while our model shows competitive performance in both minimum and maximum inference times, it particularly excels in consistency, as evidenced by the lowest standard deviation. This highlights the efficiency and reliability of our model.

4.3.2. Qualitative Comparison

Figure 3 illustrates results from various style transfer methods applied to content and style images. Each output shows the artistic transformation of the content image using different algorithms and techniques to capture and transfer elements such as color tones, textures, and brushstrokes from the style image. Comparing these outputs reveals how each method integrates the style of the style image while preserving the content image's distinctive features. Observations indicate that the AdaIN method is overly conservative in preserving the content color, resulting in a less pronounced style transfer effect. Although the MAST method achieves impressive style transfer, the presence of numerous color blocks in the image significantly compromises the integrity of the content structure. Although the CAST method performs well in transferring content and color, the texture transfer is either subtle or shows discontinuities. While the StyTR² method exhibits strong style color penetration, it lacks sufficient harmony with the original style colors, leading to a visual mismatch. The DiffuseIT method performs well in terms of stylization but falls short in the clarity of image contours, causing an overall slightly blurred visual effect. Although the StyleID method demonstrates good structural clarity, there is still room for improvement in preserving content colors. Compared to these methods, our approach not only captures and transfers artistic features from the style image, like color tones and textures, but also preserves the fundamental structure and characteristics of the content image. This balance enables us to generate artistic effects while retaining critical details and forms of the original content, resulting in outputs that are more natural and aligned with expectations.



Figure 3. Comparison Experiment Chart.

4.4. Ablation Study

In this paper, we conducted an ablation study to evaluate the effectiveness of the SimAM module. As shown in Figure 4, we found that integrating the SimAM module significantly enhances the detail expression capability during style transfer. The generated images closely approximate the artistic effects of the original style images in terms of color and texture. Moreover, the SimAM module outperforms the standalone INN-FMM module in preserving the integrity of the content structure.



Figure 4. Ablation Study.

4.5. Achieving a Harmony between Style and Content

Our method allows for flexible adjustment of the balance between content and style by tuning parameter γ . In the INN-FMM module, the content query and the AdaIN-stylized query can be allocated and combined into a new query based on weights, using the following formula:

$$Q = Q_{cs} \times (1 - \gamma) + Q_c \quad (11)$$

The parameter γ represents the proportion of the content query in the total query. As illustrated in Figure 5, a smaller value tends to emphasize style features but may sacrifice some original content details, while a larger γ value helps preserve content but may reduce the prominence of style features. This design grants users greater autonomy, allowing them to adjust the intensity of style transfer according to their personal aesthetic preferences. By adjusting the parameter, users can balance between stylization effects and content fidelity, achieving a personalized visual experience. In this paper, to maximize content structure preservation, we adopted a parameter value $\gamma = 1$, using the content image query entirely.

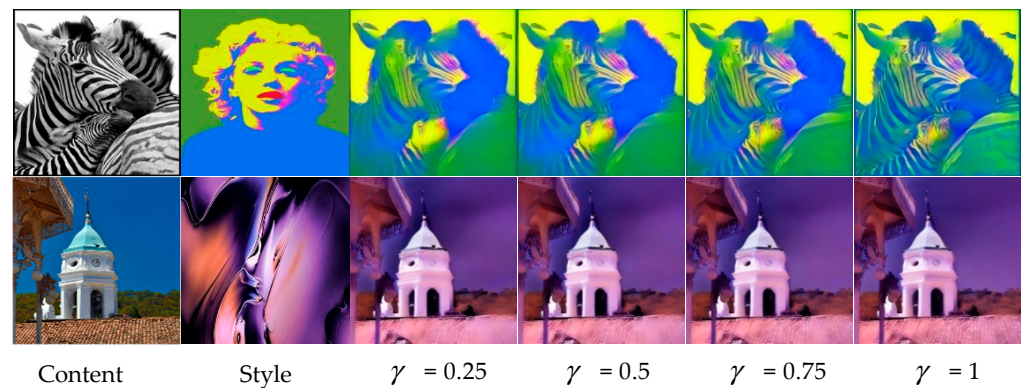


Figure 5. Trade-off Between Style and Content.

4.6. Impact of Attention Injection Step

As shown in Figure 6, we conducted an analysis and exploration of the impact of the attention injection step on the generated results. By injecting attention at both the early and later stages of the denoising process, we observed that later injection yields better results at the same time step (this is clearly seen in the comparison between the fourth and sixth columns). On the other hand, increasing the overall denoising steps makes the style patterns more prominent without affecting the content structure (as evident in the comparison between the fifth, sixth, and seventh columns). Therefore, we chose to inject attention throughout the entire denoising process to achieve optimal stylization results.



Figure 6. Injecting Attention at Different Denoising Steps.

4.7. Discussion and Explain

In this study, the proposed method demonstrates outstanding performance on the MSCOCO and WikiArt datasets, particularly in handling complex content and diverse styles. Compared to previous works, our method exhibits significant advantages in both style transfer quality and inference time. We chose the MSCOCO and WikiArt datasets because they are widely used in current style transfer research with diffusion models. To ensure fairness in comparative experiments, we selected images from these datasets. Additionally, these datasets are highly diverse, covering a wide range of image types, which we believe ensures the method's generalization across different styles and content. Lastly, our training-free style transfer method directly utilizes pre-trained models for inference, and the generalization capability of these models has already been validated in previous studies. Therefore, we assume the generalization during the inference process does not require further consideration, although its reliability may require further experimental verification.

Diffusion models have gained widespread recognition due to their powerful generative capabilities. However, this powerful capability is accompanied by a large model scale, hundreds to thousands of iterative steps, and complex reverse inference processes. These factors inherently lead to longer time consumption compared to traditional models. Therefore, in our inference time experiments, we focus primarily on comparisons with other diffusion models and do not consider non-diffusion models. Today, diffusion models are considered the mainstream approach for image generation tasks. However, the increase in training and inference time significantly impacts experimental efficiency. One of our research focuses is to reduce the time cost of diffusion models while maintaining high image

quality. We employ pre-trained models to avoid lengthy training processes and use prior knowledge from forward inference to guide reverse inference, thus saving inference time. Finally, we validate the effectiveness of our time-saving approach through comparisons with other diffusion models.

5. Conclusions and Future Work

This study addresses the challenges faced by diffusion-based methods in style transfer, where diffusion models possess significant potential for generating images. However, traditional approaches often encounter limitations such as lengthy optimization processes or underutilization of these models' capabilities. To overcome these issues, we propose a novel method introducing a dual-attention dual-chain diffusion model, which requires no additional training and achieves style transfer by adjusting a pre-trained model. Our approach manipulates self-attention features, simulating cross-attention mechanisms by injecting the style image's keys and values into the content generation process to integrate content and style characteristics. This method not only preserves the integrity of content structures but also imparts a novel visual style to the images. To further enhance style transfer effectiveness, we employ SimAM technology, significantly boosting vibrancy and adaptability in style transfer, ensuring that transferred images exhibit visual harmony and naturalness. In summary, our method, through innovative dual-attention mechanisms, improves both efficiency and visual quality in style transfer processes.

Our future work will focus on the following:

1. Expanding Application Areas: We will investigate and illustrate how our method can be adapted and utilized in various fields such as medical imaging, video processing, and other domains where style transfer or latent diffusion techniques may be beneficial.
2. Case Studies: We will conduct case studies and provide concrete examples to showcase the practical utility and versatility of our approach in different real-world scenarios.
3. Benchmarking: We plan to benchmark our method against existing solutions in these new application areas to evaluate its effectiveness and advantages in practical settings.

Author Contributions: Conceptualization, Z.X.; Methodology, L.X. and X.W.; Formal analysis, X.W.; Writing—original draft preparation X.W.; Experiment X.W.; Project administration, L.X. and Z.X.; Validation Z.X. and L.X.; Writing—review and editing, L.X. and Y.M.; Resources L.X. and X.Y.; Data curation X.Y. and Y.M.; Supervision, L.X. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Philosophy and Social Science Planning Cross-disciplinary Key Support Subjects of Zhejiang Province (No. 22JCXK08Z), Ningbo Natural Science Foundation (No. 2022J162), Ningbo Philosophy and Social Science Research Base Project (No. JD6-228).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Please contact author for data requests.

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

References

1. Gatys, L.A.; Ecker, A.S.; Bethge, M. A Neural Algorithm of Artistic Style. *arXiv* **2015**, arXiv:1508.06576. [[CrossRef](#)]
2. 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.
3. 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.
4. Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; Xu, C. Inversion-based style transfer with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10146–10156.

5. 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.
6. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; pp. 11863–11874.
7. Wright, M.; Ommer, B. Artfid: Quantitative evaluation of neural style transfer. In Proceedings of the DAGM German Conference on Pattern Recognition, Konstanz, Germany, 27–30 September 2022; pp. 560–576.
8. 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, Long Beach, CA, USA, 4–9 December 2017; p. 30.
9. 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.
10. Naem, M.F.; Oh, S.J.; Uh, Y.; Choi, Y.; Yoo, J. Reliable fidelity and diversity metrics for generative models. In Proceedings of the 37th International Conference on Machine Learning, Online, 13–18 July 2020; pp. 7176–7185.
11. Banar, N.; Sabatelli, M.; Geurts, P.; Daelemans, W.; Kestemont, M. Transfer learning with style transfer between the photorealistic and artistic domain. In Proceedings of the IS&T International Symposium on Electronic Imaging 2021, Computer Vision and Image Analysis of Art 2021, Online, 11–28 January 2021.
12. Li, H.; Wan, X.X. Image style transfer algorithm under deep convolutional neural network. In Proceedings of the Computer Engineering and Applications, Guangzhou, China, 18–20 March 2020; pp. 176–183.
13. Chen, C.J. *Chinese Painting Style Transfer Based on Convolutional Neural Network*; Hangzhou Dianzi University: Hangzhou, China, 2021. [[CrossRef](#)]
14. Li, S.; Xu, X.; Nie, L.; Chua, T.S. Laplacian-steered neural style transfer. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1716–1724.
15. Risser, E.; Wilmot, P.; Barnes, C. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv* **2017**, arXiv:1701.08893.
16. Dumoulin, V.; Shlens, J.; Kudlur, M. A learned representation for artistic style. *arXiv* **2016**, arXiv:1610.07629.
17. 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.
18. Chen, T.Q.; Schmidt, M. Fast patch-based style transfer of arbitrary style. *arXiv* **2016**, arXiv:1612.04337.
19. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
20. Zhang, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Lee, T.Y.; Xu, C. Domain enhanced arbitrary image style transfer via contrastive learning. In Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–8.
21. Liu, S.; Ye, J.; Wang, X. Any-to-any style transfer: Making picasso and da vinci collaborate. *arXiv* **2023**, arXiv:2304.09728.
22. 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, Long Beach, CA, USA, 4–9 December 2017; p. 30.
23. 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.
24. Zhu, Z.X.; Mao, Y.S.; Cai, K.W. Image style transfer method for industrial inspection. In Proceedings of the Computer Engineering and Applications, Hangzhou, China, 7–9 April 2023; pp. 234–241.
25. Han, J.; Shoeiby, M.; Petersson, L.; Armin, M.A. Dual contrastive learning for unsupervised image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 746–755.
26. Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv* **2021**, arXiv:2112.10741.
27. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv* **2022**, arXiv:2204.06125.
28. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
29. Avrahami, O.; Lischinski, D.; Fried, O. Blended diffusion for text-driven editing of natural images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18208–18218.
30. Brooks, T.; Holynski, A.; Efros, A.A. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18392–18402.

31. Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; Zheng, Y. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 22560–22570.
32. Couairon, G.; Verbeek, J.; Schwenk, H.; Cord, M. Dffedit: Diffusion-based semantic image editing with mask guidance. *arXiv* **2022**, arXiv:2210.11427.
33. Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv* **2022**, arXiv:2208.01626.
34. Wu, C.H.; De la Torre, F. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 7378–7387.
35. Zhang, Z.; Han, L.; Ghosh, A.; Metaxas, D.N.; Ren, J. Sine: Single image editing with text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6027–6037.
36. Qi, T.; Fang, S.; Wu, Y.; Xie, H.; Liu, J.; Chen, L.; Zhang, Y. DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 8693–8702.
37. Jeong, J.; Kwon, M.; Uh, Y. Training-free style transfer emerges from h-space in diffusion models. *arXiv* **2023**, arXiv:2303.15403.
38. Lin, H.; Cheng, X.; Wu, X.; Shen, D.; Wang, Z.; Song, Q.; Yuan, W. Cat: Cross attention in vision transformer. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, 18–22 July 2022; pp. 1–6.
39. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv* **2016**, arXiv:1607.08022.
40. Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. *arXiv* **2020**, arXiv:2010.02502.
41. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Lawrence Zitnick, C.; Dollár, P. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
42. Tan, W.R.; Chan, C.S.; Aguirre, H.E.; Tanaka, K. Improved ArtGAN for conditional synthesis of natural image and artwork. *IEEE Trans. Image Process.* **2018**, *28*, 394–409. [[CrossRef](#)] [[PubMed](#)]
43. Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; Xu, C. Stytr2: Image style transfer with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11326–11336.
44. Kwon, G.; Ye, J.C. Diffusion-based image translation using disentangled style and content representation. *arXiv* **2022**, arXiv:2209.15264.
45. 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.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.