*electronics*

*Article*

# A Compositional Transformer Based Autoencoder for Image Style Transfer

**Jianxin Feng [1,2,*], Geng Zhang [1], Xinhui Li [1], Yuanming Ding [1], Zhiguo Liu [1,2], Chengsheng Pan [3], Siyuan Deng [4] and Hui Fang [4]**

[1] Communication and Network Laboratory, Dalian University, Dalian 116622, China
[2] School of Information Engineering, Dalian University, Dalian 116622, China
[3] School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China
[4] Department of Computer Science, Loughborough University, Loughborough LE11 3TU, UK
* Correspondence: fengjianxin@dlu.edu.cn

**Abstract:** Image style transfer has become a key technique in modern photo-editing applications. Although significant progress has been made to blend content from one image with style from another image, the synthesized image may have a hallucinatory effect when the texture from the style image is rich when processing high-resolution image style transfer tasks. In this paper, we propose a novel attention mechanism, named compositional attention, to design a compositional transformer-based autoencoder (CTA) to solve this above-mentioned issue. With the support from this module, our model is capable of generating high-quality images when transferring from texture-riched style images to content images with semantics. Additionally, we embed region-based consistency terms in our loss function for ensuring internal structure semantic preservation in our synthesized image. Moreover, information theory-based CTA is discussed and Kullback–Leibler divergence loss is introduced to preserve more brightness information for photo-realistic style transfer. Extensive experimental results based on three benchmark datasets, namely Churches, Flickr Landscapes, and Flickr Faces HQ, confirmed excellent performance when compared to several state-of-the-art methods. Based on a user study assessment, the majority number of users, ranging from 61% to 66%, gave high scores on the transfer effects of our method compared to 9% users who supported the second best method. Further, for the questions of realism and style transfer quality, we achieved the best score, i.e., an average of 4.5 out of 5 compared to other style transfer methods.

**Keywords:** image style transfer; autoencoder; transformer; attention; convolutional neural network; Kullback–Leibler divergence

**Citation:** Feng, J.; Zhang, G.; Li, X.; Ding, Y.; Liu, Z.; Pan, C.; Deng, S.; Fang, H. A Compositional Transformer Based Autoencoder for Image Style Transfer. *Electronics* **2023**, *12*, 1184. https://doi.org/10.3390/electronics12051184

Academic Editor: Savvas A. Chatzichristofis

Received: 23 January 2023
Revised: 14 February 2023
Accepted: 20 February 2023
Published: 1 March 2023

## 1. Introduction

Style transfer is a re-rendering technique to blend the content of an image with style from another image [1]. It has become the key to many applications, including artwork generation [2,3], virtual house decoration [4,5], and face transformations with factors such as aging, rejuvenation, and hair transformation [4]. Another promising application is artistic style transfer that could augment artists to create art images. For instance, through an image style transfer model, the virtual house decoration effect can be simulated for designers to facilitate timely and costly interior design. Additionally, style transfer has also been used in camouflage object detection since the camouflage effect can be removed to improve recognition performance by applying style transfer to those camouflage objects [6].

In recent years, many deep learning-based methods have been proposed for the image style transfer task. Traditional, generative adversarial networks (GANs) have been widely applied in image style transfer [7]. However, GAN is theoretically unstable in model training [8], thus hardly converging into a reliable style transfer model. While convolutional neural networks (CNNs) provide alternative models for the applications. In [1], Gatys et al. proposed methods to achieve excellent artistic results by extracting content and style representations from images via CNNs, respectively. However, CNNs may lead to a local bias due to their spatial invariance. Consequently, CNN-based methods are not suitable for high-resolution image style transfer applications. In contrast, transformer-based methods have shown high-quality synthesis performance by exploiting long-range image feature relationships across the entire image spatial domain [9].

Despite the impressive performance achieved by these transformer-based models, it is still difficult to generate high-quality synthesized images when the texture of the style image is complicated [10]. Since the self-attention mechanism exploits feature correlations from all windows across the entire image equally, it is more likely to produce a hallucination effect on those blended images when transferring style from texture-rich images. Additionally, semantic preservation from the content image is weakened when transferring these texture-riched patterns to the newly rendered image.

In this paper, we propose a novel compositional transformer model to solve the above-mentioned challenges for producing high-quality style transfer images when processing style transfer tasks from high-resolution images. Instead of using a self-attention mechanism, we design a novel compositional attention scheme to exploit highly correlated spatial features and weakly correlated blocks separately so that the influence of patterns from texture-rich style images is disjointed to generate more photo-realistic style transfer synthesis. Furthermore, we also embed a region consistency term in our loss function to preserve more semantic information from content images. The main contributions of our work can be summarized as follows:

- We propose a compositional transformer model for image style transfer. The model can generate more photo-realistic synthesized images even when the style image contains complicated texture features.
- A comprehensive loss function, including content loss, style loss, region consistency loss and Kullback–Leibler (KL) divergence loss, is designed to preserve more semantic information from the content image for the style transfer task.
- In order to encourage the similarity between features extracted from any locations of the same label to be the same for the original content image and the style transferred image, a new regional consistency loss is proposed to preserve the internal structure.
- Information theory-based CTA is discussed, and information entropy of the image is obtained to measure how much information the image has. Kullback–Leibler divergence is the way to measure just how much information is lost when we approximate the distribution of the original image with the distribution of the style-transferred image. In order to preserve more brightness information for photo-realistic style transfer, KL divergence loss is introduced.
- A user study is conducted to convincingly prove that our proposed method produces improved results that are consistent with human perception.

Various style transfer models have been applied to many fields. However, each of these methods has its own disadvantages. For an original image with complex textures or enriched backgrounds, the realism of the generated maps after style transfer is typically poor. The traditional image style transfer methods cannot fully extract the high-level semantics of an image, resulting in a biased effect. Most algorithms cannot generate high-quality stylized results efficiently. Moreover, there is a need to improve parameter tuning, generation speed, and producing more realistic details.

## 2. Related Work

### *2.1. Image Style Transfer*

Generative Adversarial Networks (GANs), which were initially proposed by Goodfellow et al. in [11], have been widely used in image style transfer tasks [7,12]. For instance, Pix2pix model [7] learns a mapping function between a large number of image pairs. When taking an image from a source domain, the model has the capability to generate a new image in its target domain. However, the training of the Pix2pix model requires a significant amount of paired image data to build the model. To train a style transfer model with unpaired training data, DiscoGAN [13] discovers cross-domain relations when no training pairs are presented. Their approach works without any explicit pair labels and learns to relate datasets from very different domains. In contrast, CycleGAN [12] proposes a multi-style image transfer system CCGAN, DualGAN [14] use cyclic consistency constraints to learn a style transfer model without image pairs. Despite the impressive performance, these GAN-based methods suffer from poor convergence. Another disadvantage is that it is difficult to control the synthesized outputs by learning these GAN models.

Convolutional neural networks (CNNs) are also widely applied for image style transfer [1]. In these methods, CNNs are used to extract both high-level content features and stylized texture features. Starting from a random noise-based generated image, an iterative optimization is applied to align high-level semantic features of a content image based on a pre-trained VGG model [15] and style feature correlations at different channels represented by Gram Matrix. Their method improves the image generation speed to be up to 49% faster and reduces the number of parameters by 20% while maintaining style transferring performance. However, these methods suffer from inefficient computation due to their slow optimization process. To speed up the transformation, Johnson et al. [16] presents a feed-forward model, i.e., deep residual network [17], to minimize a perceptual loss function. This model has a much faster processing speed and presents similar synthesizing results to [1]. It is three orders of magnitude faster. While to preserve more semantic information from the content image, Lin et al. [18] proposes to build a style transfer network based on a mask r-cnn model [19] so that their style transfer model can preserve more high-level semantics. Their algorithm can effectively solve the problem of semantic mismatch in the process of image style transfer.

### *2.2. Transformer Model*

The transformer model is initiated in the field of natural language processing [20]. It uses a series of self-attention modules to compute correlations between long-range blocks to enhance and improve local feature representations. However, due to the nature of the self-attentive mechanism, which requires complex computation and large amounts of storage, various 'X-former' models have been proposed (e.g., Linformer [21], Reformer [22]) to improve both computational efficiency and storage efficiency of the original transformer model.

For vision tasks, Vision Transformer (VIT) [23] is one of the earliest models that are applied to the image classification task. the best model reaches an accuracy of 88.55% on ImageNet. After encoding image patches via a linear projection, self-attention modules are used to replace traditional convolutional layers to exploit long-range dependencies to enhance feature representations for image recognition. Many new transformer models are also proposed for various image processing tasks, such as image recognition [24] introduce DeiT, which are image transformers that do not require a very large amount of data to be trained, image segmentation, Ye et al. [25] have proposed cross-modal self-attention and gated multi-level fusion modules to address two crucial challenges in the referring image segmentation task, image enhancement, Yang et al. [26] have proposed a novel Texture Transformer Network for Image Super-Resolution (TTSR) which transfers high-resolution textures from the reference image to a low-resolution image [27].

Recently, there are studies to propose transformer-based models for image style transfer applications. To address the key issue of the limitations of CNNs, Deng et al. [9] takes long-range dependence of patches from an input image into account for unbiased style transfer by proposing a transformer model, namely StyTr2. They also analyze the deficiency of existing positional encoding methods and propose content-aware positional encoding (CAPE). In contrast to the visual Transformer used for other visual tasks, StyTr2 contains two different Transformer encoders that generate domain-specific sequences for content and style, respectively. To improve the efficiency of the transformer model, Xiao et al. [28]

and Luo et al. [29] ultilize CNNs to extract compact features before applying transformer models. Xiao et al. [28] improves ImageNet top-1 error by 1–2%.

The advantage of the transformer model is that the self-attentive mechanism can grasp the information of high-level semantics, so that it can focus on the contextual information as well as the distant information, instead of the common local information capture, to establish long-distant dependencies. However, the self-attentive mechanism requires a large amount of computation, which limits its potential for processing high-resolution input images. For example, iGPT generates images with only 64 × 64 maximum pixels, while [30] generates images with very low resolution. Further, the self-attention modules treat all blocks equally for feature enhancement. This may generate a hallucinatory effect when transferring patterns from texture-rich images.

## 3. Proposed Method

### *3.1. Overall Network Architecture*

In our proposed style transfer model, we design an autoencoder network architecture for this task which is illustrated in Figure 1. Autoencoder is a data reconstruction method that was initially proposed in [31]. It comprises an encoder for finding latent codes to represent images and a decoder for a decompression that maps to an output space, which could be either reconstruction [30] or style transfer [4].

In our work, we adopt a mixture model of CNN and transformer as the encoder of our style transfer network since both CNN and transformer have their own advantages when applied to this image synthesis task. Due to the weight sharing mechanism of CNN, features extracted by convolutional layers are translation invariant, and they are locally sensitive. Thus, convolution layers are more suitable for decoupling image content and styles. Similar to [4], we use a Resnet architecture to decompose an input image into a tensor $F_c$ and a vector $F_s$. Here, $F_c$ represents content information extracted from an input content image while $F_s$ represents style information of an input style image. After passing through four sequential down-sampling residual blocks to generate a feature map $F_g$. This feature map is fed into a following convolution layer to produce the content representation $F_c$ that will be positionally encoded to prepare the input of its following transformer model. In contrast, $F_g$ is further processed with two other convolution layers followed by average pooling and dense layer to extract its style code $F_s$. Regarding the transformer, we propose a novel attention module to further exploit context information when swapping style features for this style transfer task. The detail of this transformer model will be explained in the following subsection.

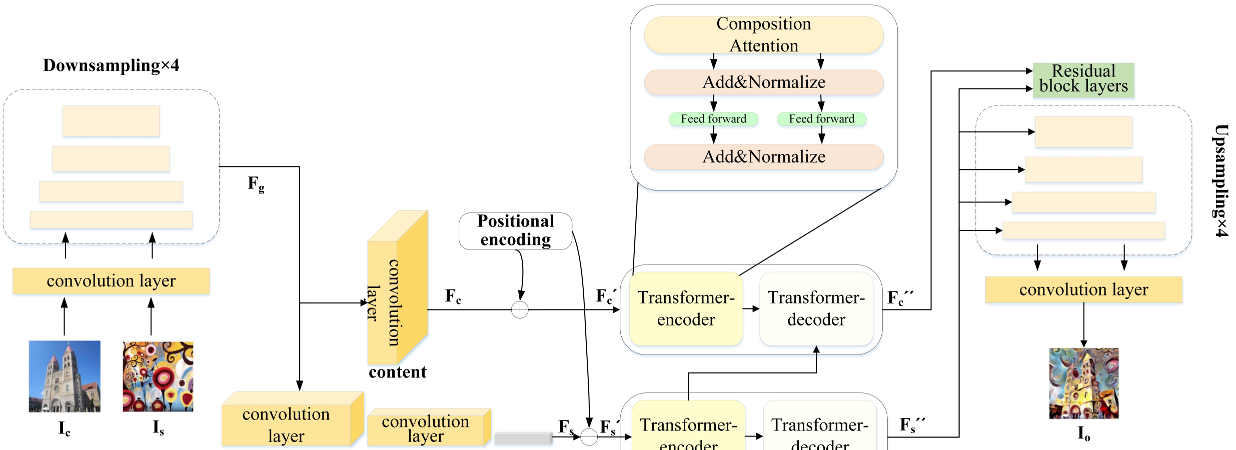


**Figure 1.** The network architecture of our proposed compositional transformer model which is composed of an encoder combining CNN and compositional attention blocks, and a StyleGAN2-based decoder.

For the decoder part, we utilize StyleGAN2 architecture [32] as the network structure. As illustrated in Figure 1, the decoder takes transformer-enhanced style codes for the modulation process to apply style on style-transferred images layer by layer. At the same time, another transformer block exploits correlations between content representations with its transferred style to generate the content input of the decoder. The decoder consists of four residual blocks followed by four upsampling layers.

### *3.2. Compositional Attention Module and Compositional Transformer*

We present a novel transformer model to explore long-range dependencies across an image as well as correlations between content and style representations for the transfer task. To improve the processing speed of our model, we flatten the feature maps produced by our CNN encoder into a 1D sequence. To avoid losing spatial information, positional encoding is applied to each feature representation at each position. Sine and cosine function with different frequencies [22,33] are computed to encode the positional coordinates which are expressed as following equations:

$$\begin{cases} PE_{\#}(pos, 2t) = \sin(pos \cdot h) \\ PE_{\#}(pos, 2t+1) = \cos(pos \cdot h) \end{cases} \tag{1}$$

where # $\in \{H, W\}$ indicates each of two dimensions, $h = 1/10,000^{2t/128}$. For width dimension, pos represents the row of data in the image. For the height dimension, pos represents the column of data in the image. We concatenate $PE_H$ and $PE_W$ as the positional encoding. In this manner, we can get the $F_c^{'}$ and $F_s^{'}$ with the location information added as the input of the transformer encoder.

The structure of our transformer also consists of an encoder and a decoder. Instead of using a self-attention module in the transformer encoder, we propose a novel compositional attention module to encode context information. Due to the uniqueness of style transfer, stylization, which is applied to the entire picture, will generate a visually unreal output image since self-attention features are calculated with relatively scattered data that leads to poor artifact effect. To tackle this issue, we design composition attention to replace the self-attention modules. The core idea is intuitive as learning feature representations via grouping both the neighboring and similar data and other data can disjoin the non-existing intertwining relations between these patterns.

After taking a current input F to get its key $K$ and value $V$, queries are generated from all other positions to calculate attention scores. It determines how much attention needs to be focused on each data at other locations in the image when encoding the current F. The equation of attention is expressed in the following equation, where $Q \times K^T$ represents the score value:

$$Softmax(\frac{Q \times K^T}{\sqrt{d_k}}) \cdot V \tag{2}$$

The larger the score value, the higher the correlation between the input and this corresponding query block. The original self-attention module treats these blocks equally to produce the output-enhanced features. While in our work, we threshold these similarity values and treat them into two groups for producing the output features shown in Figure 2. Through the experiments which will be presented in our ablation study, we set the value of c to 0.8 heuristically. For each $Q_i$, when comparing its score to the c, it filters m data with blocks whose similarity scores are larger than c, n > m $\geq$ 1. These blocks as considered as context which has strong correlations to the current position. The features from these blocks are combined as filtered data ($Z_1, \ldots, Z_i, \ldots, Z_m$) by a Concat&Normalize unit. For other blocks, their features are input into the attention module which process is similar to a normal self-attention module. The proposed attention module makes the learning processing pays more focus on enhancing features with those blocks with higher correlations. After the transformer processing, we get the $F_c^{''}$ and $F_s^{''}$.

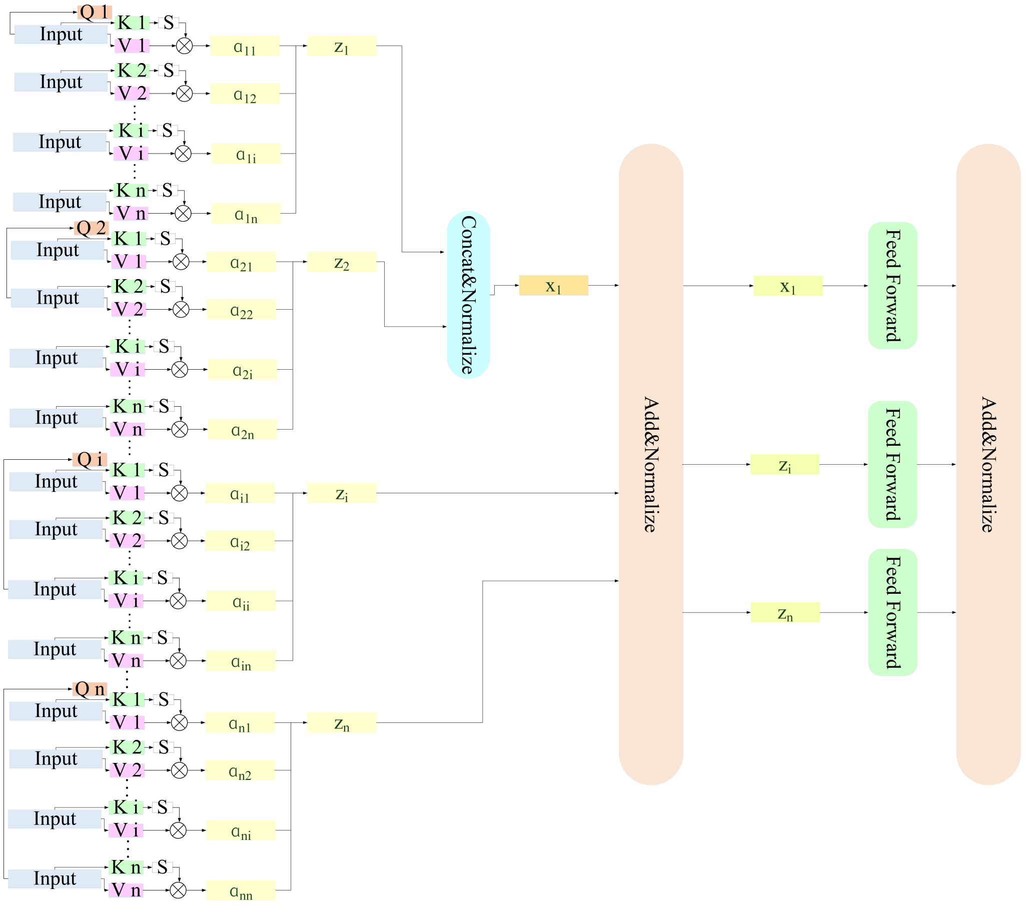


**Figure 2.** The detail of composition attention calculation. The context features are grouped into two parts where blocks with high correlations and other blocks are processed separately to disjoin style patterns.

After getting the composition attention outputs, they are added, normalized and transformed via a feed-forward network whose process is the same as a classical transformer model. Note that another additional advantage of the proposed compositional attention is its computational efficiency. Since the model reduces processing data by focusing on highly correlated blocks, its operation speed is much faster when compared to a traditional self-attention-based model.

### 3.3. Loss Function

Our loss function comprises a content perceptual loss, style perceptual loss and a region consistency loss. We follow [9] to define the content perceptual loss and style perceptual loss. The content perceptual loss $L_c$ is expressed as

$$L_c = \frac{1}{N_l} \sum_{i=0}^{N_l} \parallel \phi_i(I_o) - \phi_i(I_c) \|_2 \tag{3}$$

where $\phi_i(\bullet)$ is the content feature extracted from the $i$ layer of a pre-training network, and $N_l$ is the number of layers.

The style perceptual loss $L_s$ is expressed as follows:

$$L_s = \frac{1}{N_l} \sum_{i=0}^{N_l} \parallel \mu(\phi_i(I_o)) - \mu(\phi_i(I_s)) \|_2 + \parallel \sigma(\phi_i(I_o)) - \sigma\phi_i(I_s) \|_2 \tag{4}$$

where $\mu(\bullet)$ and $\sigma(\bullet)$ denote the mean and variance of extracted features, respectively.

In addition to the perceptual similarity of images in terms of content and style, the internal structure and brightness of images should also be considered in the training process since it helps to preserve more semantic information. The new region consistency loss and Kullback–Leibler divergence loss are proposed.

In order to encourage the similarity between features extracted from any locations of the same label to be same for the original content image and the style transferred image, the new region consistency loss $L_r$ is defined as below:

$$L_r = \frac{1}{N_l^2} \sum_{r=1}^{R} \sum_{i,j \in m_r} \| d_{ij}^o - d_{ij}^c \|_2 \tag{5}$$

where $d_{ij} = w_l \odot (F_i - F_j)$ is the distance between $i$ and $j$ positions of the same label $r$ in area $m_r$. $w_l = 1 \forall l$, where $l \in \{1, \ldots, l, \ldots, N_l\}$. $o$ is the generated image and $c$ is the initial content image. This term facilitates the region consistent across both the original content image and the style transferred image.

Information entropy is the basic concept of information theory, which is used to measure the information in data. The definition of information entropy for a probability distribution is:

$$H = -\sum_{i=1}^{N} p(x_i) \cdot \log p(x_i) \tag{6}$$

where $p(x_i)$ represents the probability of random event $x_i$. So if we have the probability distribution of the image, the information entropy of the image can be obtained.

After transferring color to a grayscale image, the entropy of the image is gotten by gray value statistics. We want to quantify how much information is lost when the original image is transferred to the style-transferred image. Relative entropy, also known as KL divergence or information divergence is the way.

KL divergence is an asymmetric measure of the difference between two probability distributions. When two random distributions are the same, their relative entropy is 0. When the difference between two random distributions increases, their relative entropy also increases. In image style transfer, KL divergence can calculate exactly how much information is lost when we approximate the distribution of the original image with the distribution of the style-transferred image.

Comparing the distribution of the original image with that of the style transferred image, the more similar they are, the better the effect of image style transferring will be in the style transfer of the image.

Then we introduced Kullback–Leibler divergence loss into the loss function to preserve the brightness semantic information with relative entropy:

$$L_{KL} = KL[E(c) \parallel E(o)] \tag{7}$$

Overall, the complete learning objective of our model is formulated as:

$$L_{total} = \lambda_c L_c + \lambda_s L_s + \lambda_r L_r + \lambda_k L_{KL} \tag{8}$$

The four terms are balanced by four weights which are empirically set to 1, 0.7, 1, and 0.3.

## 4. Experiments

### 4.1. Experimental Setting

To evaluate the performance of our proposed method, we conducted experiments based on three standard datasets, including the Churches dataset from LSUN [34], the Flickr Landscapes dataset and the Flickr Faces HQ (FFHQ) dataset [35]. The images in these datasets are mainly buildings, faces and landscapes, which are cropped and normalized into a standard size.

Our model was built based on Pytorch and trained by using 8 NVIDIA Ge Force GTX 2080Ti GPU cards. Random pairing of content and style images was performed during training. The data batch size was set to 8, the initial value of the learning rate was 0.005, and the Adam optimizer was adopted.

### *4.2. Visual Comparison*

To demonstrate our excellent performance, we compared out method with several state-of-the-art methods, including WCT2 [36], AdaIN [37], Johnson [16], StyTr2 [9], Style-TUNE [38] and Gatys [1]. As shown in Figure 3, it is observed that our method improves the quality of style-transferred images significantly. The generated images retained high semantics of their corresponding original images with consistent high-semantic regions. Further, the style has been successfully transferred to the synthesized images to express the artistic effect as well as preserve photo-realistic quality. For example, the sky shown in the transferred image on row 2 looks more like the sky in a painting without those distracting patterns transferred by other methods. Similarly, the reflection of the boat hull in the water is also preserved when compared to other methods.

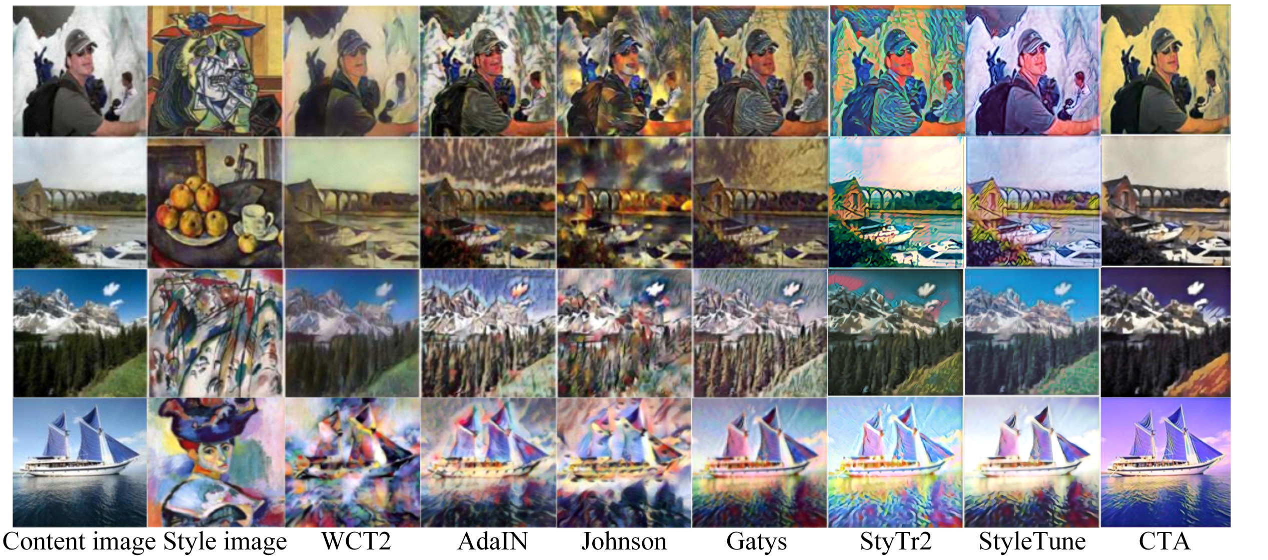


**Figure 3.** Comparison of the results between ours and others.

Since our model proposed composition attention to replace traditional self-attention-based transformers. We also generated transferred images by using these two methods to prove the effectiveness of our new composition attention model. As illustrated in Figure 4, (b) and (c) were produced by using CTA with composition attention, while (e) and (f) were generated by using the traditional transformer model. It is observed that the face of humans and the contour of humans are not fully transferred well due to the influence of holistic contextual features in CTA with self-attention. CTA with composition attention disjoins less relevant context to generate a more photo-realistic high-resolution style transferred image.

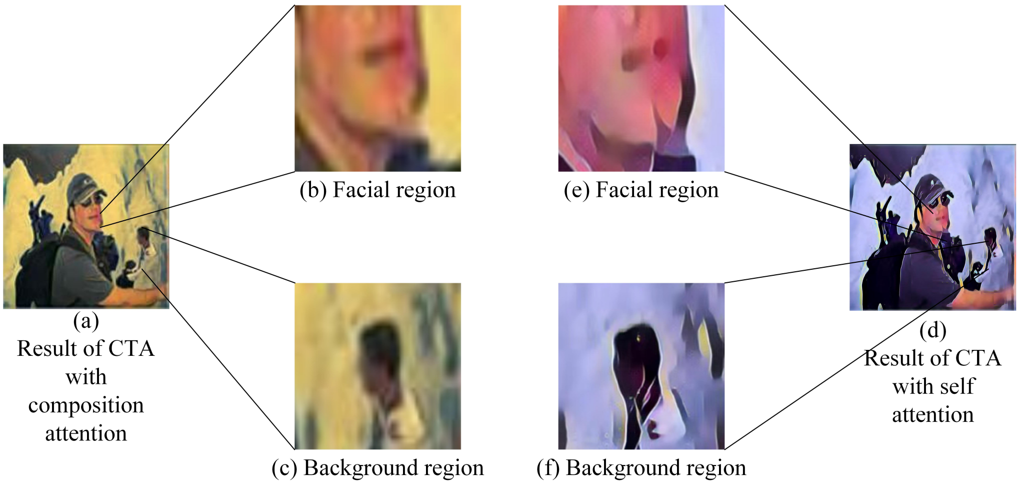


**Figure 4.** Comparison of the results between CTA with composition attention and CTA with self attention.

At present, there is no quantitative index to measure the quality of style transfer results. While to achieve an objective comparison, we use two evaluation metrics, including Structural Similarity (SSIM) and Peak Signal Noise Ratio (PSNR), to compare our model with other style transfer methods. Experiments compared to these methods were conducted with three image datasets, namely, Churches, Flickr Landscapes, and Flickr Faces HQ, and the results are shown in Table 1. Our method achieved the best performance in both metrics compared to other methods, indicating its advantages in the maintenance of content image details and style transfer effects.

**Table 1.** Performance indicators of different models (results are retained to two decimal places).

| | **Dataset** | **WCT2** | **AdaIN** | **Johnson** | **Gatys** | **StyTr2** | **StyleTune** | **CTA** |
|---|---|---|---|---|---|---|---|---|
| PSNR (dB) | Churches | 20.03 | 14.24 | 12.67 | 10.44 | 16.8 | 15.6 | 20.92 |
| | Flickr-Landscopes | 18.29 | 12.35 | 11.61 | 9.26 | 15.9 | 14.5 | 20.14 |
| | Flickr Faces HQ | 19.06 | 13.56 | 11.82 | 9.64 | 16.2 | 16.7 | 21.02 |
| SSIM | Churches | 0.69 | 0.35 | 0.36 | 0.43 | 0.44 | 0.43 | 0.53 |
| | Flickr-Landscopes | 0.61 | 0.31 | 0.29 | 0.37 | 0.38 | 0.35 | 0.42 |
| | Flickr Faces HQ | 0.65 | 0.33 | 0.31 | 0.41 | 0.45 | 0.45 | 0.48 |

To further demonstrate the advantages of our approach, we apply three different style images to the Qingming Shanghetu, one of the most famous ancient paintings in China. The Qingming Shanghetu depicts the conditions of the capital city (present-day Kaifeng, Henan Province) during the Northern Song Dynasty, mainly Bianjing and Bianhe the natural scenery, and prosperous scenes on both sides of the river. The various parts of the picture are correlated and have more details, which require high-resolution image processing. By applying three style images, the transfer results are illustrated in Figure 5. In terms of synthesis results, the method in this paper can handle the details clearly when processing high-resolution content images. Thanks to the convolutional neural network part in our encoder, the details of the image are fully processed for decoupling the content and style of the image, which lays a good foundation for style transfer. Additionally, with the help of Transformer the overall style can be grasped well and the visual effect is excellent. For example, the comparison between the generated image and the original image is shown in Figure 5, which shows that the facial details, text, and stripes are well-migrated and retained.

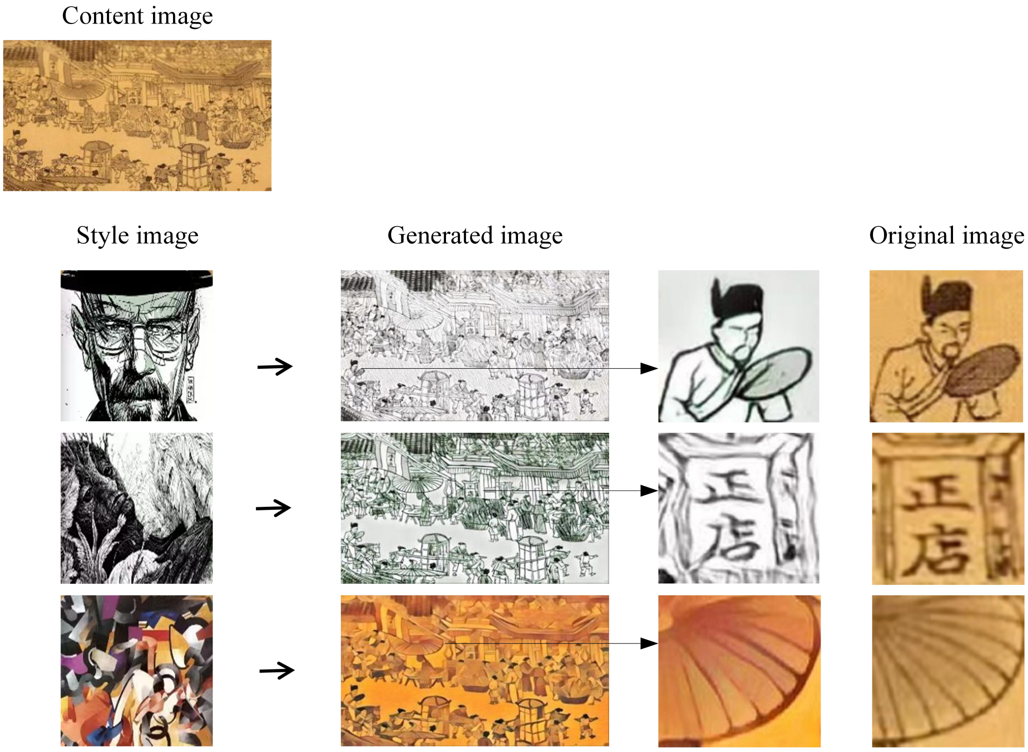
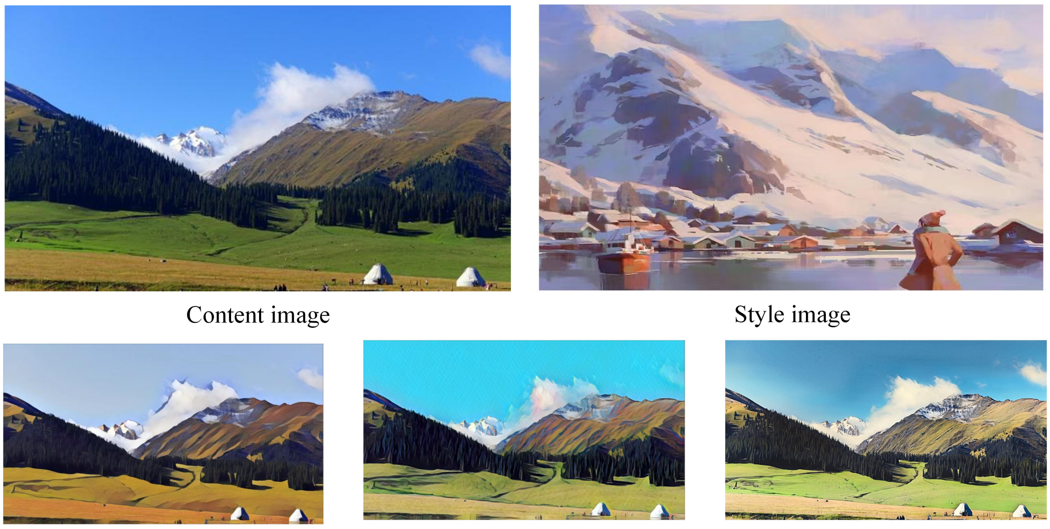


**Figure 5.** Results of style transfer (Qingming Shanghetu).

### *4.3. Ablation Study*

Since our model proposed a new loss function with region consistency loss and KL divergence loss to replace the traditional loss function in style transfer. The traditional loss function only includes content loss and style loss. We also generated transferred images by using these three loss functions to prove the effectiveness of our loss function. As illustrated in Figure 6, output result 1 was generated by using the traditional loss function, output result 1 was the result generated using the loss function with only region consistency loss added, while output result 3 was generated by using our overall loss function. It is observed that the grass, in result 1 forest and mountain regions are not fully transferred well due to being short of the internal structure of the image. In result 2, both of sky part and the vegetation part appeared to color disharmony problem. In contrast, our loss function increases the internal structure constraint of the image to generate a more photo-realistic high-resolution style transferred image in result 2. Further, compared with result 2, result 3 balances the partial brightness information better.

Since the grouping threshold is a crucial parameter of the proposed transformation model, we evaluate its impact on three datasets by using the SSIM and PSNR metrics as these two metrics provide objective measurement of content preservation in the transferred images. It can be seen from Table 2 that the content can be best preserved when this value is set to 0.8. Thus, we use this setting in all our presented results.



Content image

Style image

Output result 1

Output result 2

Output result 3

**Figure 6.** Landscape style transfer results.

**Table 2.** SSIM with different c values.

| | **Dataset** | **c = 0.5** | **c = 0.6** | **c = 0.7** | **c = 0.8** | **c = 0.9** |
|---|---|---|---|---|---|---|
| SSIM | Churches | 0.47 | 0.49 | 0.49 | 0.53 | 0.48 |
| | Flickr Landscapes | 0.31 | 0.35 | 0.37 | 0.42 | 0.36 |
| | Flickr Faces HQ | 0.42 | 0.45 | 0.45 | 0.48 | 0.46 |

Another ablation conducted is to determine whether the modulation process is used in upsampling layers only or in both residual and upsampling layers. The SSIM and PSNR results are presented in Table 3. It can be found that it is better to use modulation in both residual and upsampling layers in order to preserve more semantic features from a content image. With Churches, Flickr Landscapes, and Flickr Faces HQ, the values of PSNR and SSIM in the modulation structure of both residual and upsampling layers are larger than the modulation structure of upsampling layers only. The larger the PSNR, the higher the Signal-to-Noise Ratio, and the larger the SSIM, the more similar the structure.

**Table 3.** Comparison of different modulation structure.

| | **Dataset** | **Separate Upper Sampling Layer** | **Residual Layer + Upsampling Layer** |
|---|---|---|---|
| PSNR | Churches | 18.2 | 19.6 |
| | Flickr Landscapes | 17.6 | 17.8 |
| | Flickr Faces HQ | 16.8 | 17.2 |
| SSIM | Churches | 0.39 | 0.48 |
| | Flickr Landscapes | 0.29 | 0.31 |
| | Flickr Faces HQ | 0.27 | 0.32 |

### *4.4. User Study*

We further perform a user study to quantitatively demonstrate that the proposed method has the best style transfer performance compared to the state-of-the-art algorithms. Our user study is based on the validation dataset that consists of 56 content images and 37 style images. We obtain the style transfer results of WCT2, AdaIN, Johnson, Gatys, StyTr2, StyleTune CTA, and the proposed method on every content-style pair, respectively. We finally obtain 2072 style transfer results for each method. For questions 1–3, we asked the user to choose the most suitable one, and for questions 4–5, we asked the user to rate the partition between 1 and 5. We eventually collected 1088 votes. The male and female ratio is 51:49, and the average age is 28 years old, with a fluctuation of fewer than 5 years. There is no color blindness among the participants.

A better image is selected from them according to the following question:

Question 1: Which image better preserves content information (shape, semantics, etc.)?

Question 2: Which image better demonstrates the effect of style transfer in terms of texture and color?

Question 3: Which image is more like the one in the target domain?

Question 4: Is the image (generated image) a real image or a computer-generated image? We set the score 1–5. A score of 1 means a completely computer-generated image, and a score of 2 means a nearly computer-generated image. A score of 3 means it is difficult to differentiate. A score of 4 is close to the real picture. A score of 5 means a completely real picture.

Question 5: Are style and content images adequately integrated? We set the score 1–5. A score of 1 indicates no combination at all, and a score of 2 indicates almost no combination. A score of 3 indicates an ordinary combination. A score of 4 means that most of that is combined. A score of 5 is a perfect combination.

The results of the user study are presented in Figures 7 and 8. Based on the results from Q1 to Q3, In Q1, 80% of the respondents considered that the generated images displayed the content images better, which indicated that our method retained the content information of the original images better. In Q2, 71% of respondents agreed that the images generated by our method can embody the effect of style transfer better. In Q3, 72% of respondents chose the images generated by our method, which proved that our method has completed the task of image style transfer well. It is observed that the majority votes supported that the proposed method generated higher performance than other methods.

We also scored the best 4.5 for Q4 and Q5 with very low standard deviations, which are 0.13 and 0.12, respectively. In Q4, our method scored 4.5 on average, higher than other methods. In addition, our standard deviation is 0.13, which indicates that the calculated average score has very high precision. So our results are much more like real images than other methods. In Q5, our method scored an average of 4.5, higher than the other methods. This shows that the respondents think our result is the best combination of content and style in these methods. All the results confirmed that our proposed method achieved the best style transfer results when compared to our benchmark methods.

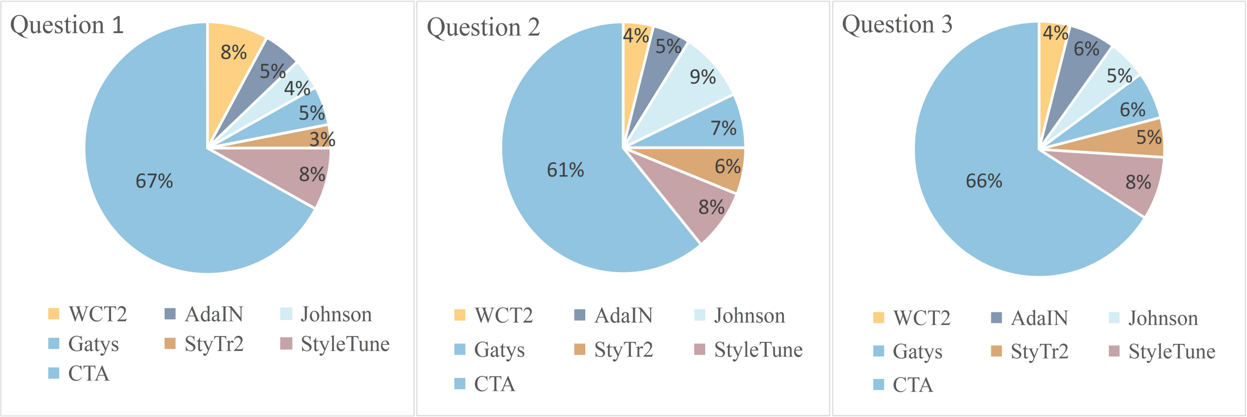


**Figure 7.** Questionnaire results of question 1, 2 and 3.

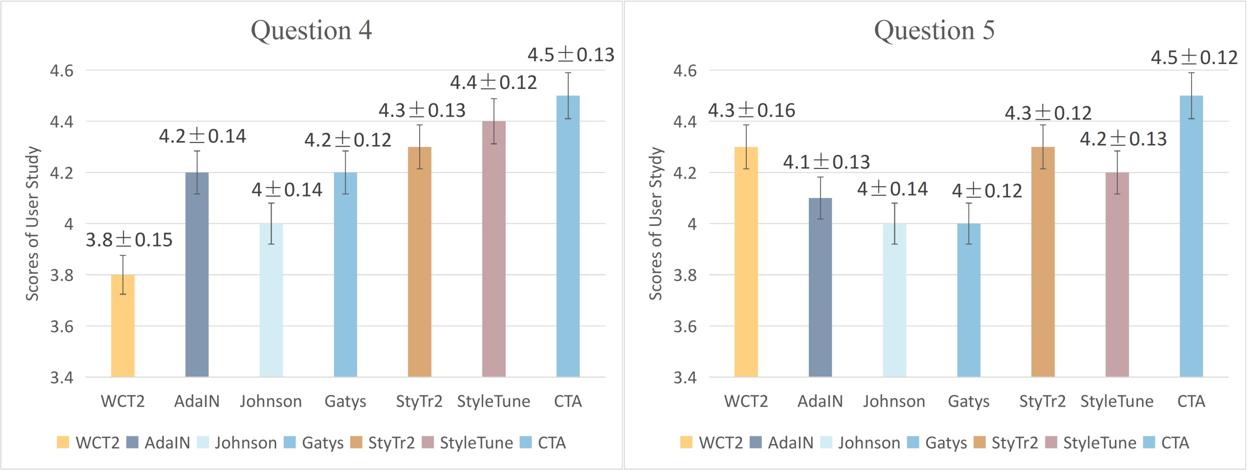


**Figure 8.** Questionnaire results of question 4 and 5.

## 5. Conclusions

In our paper, we proposed a novel compositional transformer-based style transfer model. We designed a CNN-transformer-based autoencoder to generate photo-realistic style transferred images when processing high-resolution images. Instead of using self-attention modules in our transformer model, we process highly correlated blocks and other blocks separately to enhance feature representations for style transfer tasks. With this design, we alleviate the artifact effect for better synthesis with a more efficient processing pipeline. In addition to the loss of content and style, it also strengthens the constraints of the internal structure of the image with the region consistency loss $L_r$ proposed in this paper. Moreover, inspired by information entropy, KL divergence loss $L_{KL}$ is introduced to preserve the brightness semantic information of the image. Compared with WCT2, AdaIN, Johnson, Gatys, StyTr2 and StyleTune, the value of PSNR is 20.92 for Churches, 20.14 for Flickr Landscapes, 21.02 for Flickr Faces HQ, and the value of SSIM is 0.53 for Churches, 0.42 for Flickr Landscapes, 0.48 for Flickr Faces HQ. Our scores of user study are the highest in these methods. Based on a user study assessment, the majority number of users, ranging from 61% to 66%, gave high scores on the transfer effects of our method compared to 9% users who supported the second best method. Further, for the questions of realism and style transfer quality, we achieved the best score, i.e., an average of 4.5 out of 5 compared to other style transfer methods. Our experimental results demonstrate that our method outperforms several state-of-the-art style transfer models when processing high-resolution images. At present, the test-time speed of our method can be improved further. Our future work will focus on further improvement of the model efficiency and design of an adaptive grouping threshold in our transformation model for different real-world applications.

**Author Contributions:** Conceptualization, J.F.; methodology, G.Z.; software, X.L.; validation, G.Z.; formal analysis, J.F.; investigation, X.L.; resource, X.L.; writing—original draft preparation, G.Z.; writing—review and editing, J.F., S.D. and H.F.; supervision, J.F., Y.D., Z.L., C.P. and H.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by The National Natural Science Foundation of China grant number 61931004.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image Style Transfer Using Convolutional Neural Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2414–2423.
2. Kim, M.; Choi, H.; Paik, J. Style Transfer Using Convolutional Neural Network and Image Segmentation. *TECHART J. Arts Imaging Sci.* **2021**, *8*, 5–8. [CrossRef]

3. Liao, Y.M.; Huang, Y.F. Deep Learning-Based Application of Image Style Transfer. *Math. Probl. Eng.* **2022**, *2022*, 1693892. [CrossRef]
4. Park, T.; Zhu, J.Y.; Wang, O.; Lu, J.; Shechtman, E.; Efros, A.; Zhang, R. Swapping Autoencoder for Deep Image Manipulation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 7198–7211.
5. Liu, S.; Bo, Y.; Huang, L. Application of Image Style Transfer Technology in Interior Decoration Design Based on Ecological Environment. *J. Sens.* **2021**, *2*, 1–7. [CrossRef]
6. Mao, Y.; Zhang, J.; Wan, Z.; Dai, Y.; Barnes, N. Transformer Transforms Salient Object Detection and Camouflaged Object Detection. *arXiv* **2021**, arXiv:2104.10127.
7. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 12–17 July 2017; pp. 5967–5976. [CrossRef]
8. Saxena, D.; Cao, J. Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions. *ACM Comput. Surv.* **2022**, *54*, 63.
9. Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; Xu, C. StyTr2: Image Style Transfer with Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 14–19 June 2022; pp. 11326–11336. [CrossRef] [PubMed]
10. Ma, Z.; Lin, T.; Li, X.; Li, F.; He, D.; Ding, E.; Wang, N.; Gao, X. Dual-Affinity Style Embedding Network for Semantic-Aligned Image Style Transfer. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–14. [CrossRef]
11. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. *Commun. ACM* **2020**, *11*, 139–144. [CrossRef]
12. Tu, C.T.; Lin, H.J.; Tsia, Y. Multi-style image transfer system using conditional cycleGAN. *Imaging Sci. J.* **2020**, *12*, 1–14.
13. 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 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 2533–2538. [CrossRef]
14. Yuan, Q.L.; Zhang, H.L. RAMT-GAN: Realistic and accurate makeup transfer with generative adversarial network. *Image Vis. Comput.* **2022**, *120*, 104400–104415. [CrossRef] [PubMed]
15. Kim, M.; Choi, H.C. Compact Image-Style Transfer: Channel Pruning on the Single Training of a Network. *Sensors* **2022**, *22*, 8427.
16. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-time Style Transfer and Super-resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 18–16 October 2016; pp. 694–711.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 24–30 June 2016; pp. 770–778. [CrossRef]
18. Lin, Z.; Wang, Z.; Chen, H.; Ma, X.; Xie, C.; Xing, W.; Zhao, L.; Song, W. Image Style Transfer Algorithm Based on Semantic Segmentation. *IEEE Access* **2021**, *9*, 54518–54529.
19. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–28 October 2017; pp. 2980–2988.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 3568–3570.
21. Wang, S.; Li, B.Z.; Khabsa, M.; Fang, H.; Ma, H. Linformer: Self-Attention with Linear Complexity. *arXiv* **2020**, arXiv:2006.04768.
22. Kitaev, N.; Kaiser, L.; Levskaya, A. Reformer: The Efficient Transformer. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020; pp. 526–534.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Online, 28–30 April 2020; pp. 1–27.
24. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning (ICML), Vienna, Austria, 2–5 March 2021; pp. 7358–7367.
25. Ye, L.; Rochan, M.; Liu, Z.; Wang, Y. Cross-Modal Self-Attention Network for Referring Image Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10502–10511.
26. Yang, F.; Yang, H.; Fu, J.; Lu, H.; Guo, B. Learning Texture Transformer Network for Image Super-Resolution. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2020; pp. 5791–5800.
27. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Gao, W. Pre-Trained Image Processing Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Kuala Lumpur, Malaysia, 19–25 June 2021; pp. 12299–12310.
28. Xiao, T.; Singh, M.; Mintun, E.; Darrell, T.; Dollár, P.; Girshick, R. Early Convolutions Help Transformers See Better. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia, 6–14 December 2021; pp. 1–16.
29. Luo, X.; Hu, M.; Song, T.; Wang, G.; Zhang, S. Semi-Supervised Medical Image Segmentation via Cross Teaching between CNN and Transformer. *arXiv* **2022**, arXiv:2112.04894. [CrossRef]

30. Sun, W.; Shao, S.; Yan, R. Induction Motor Fault Diagnosis Based on Deep Neural Network of Sparse Auto-encoder. *J. Mech. Eng.* **2016**, *52*, 65–71. [CrossRef]
31. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Representations by Back-propagating Errors. *Nature* **1986**, *323*, 533–536.
32. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of StyleGAN. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2020; pp. 8110–8119.
33. Xie, Y.; Zhang, J.; Shen, C.; Xia, Y. CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021, Strasbourg, France, 27 September–1 October 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 171–180.
34. Yu, F.; Zhang, Y.; Song, S.; Seff, A.; Xiao, J. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *Comput. Sci.* **2015**, *6*, 56–59.
35. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4396–4405.
36. Yoo, J.; Uh, Y.; Chun, S.; Kang, B.; Ha, J.W. Photorealistic Style Transfer via Wavelet Transforms. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9035–9044.
37. Huang, X.; Belongie, S. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1510–1519. [CrossRef]
38. Reimann, M.; Buchheim, B.; Semmo, A.; Döllner, J.; Trapp, M. Controlling strokes in fast neural style transfer using content transforms. *Vis. Comput.* **2022**, *38*, 4019–4033.

**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.