Information Sciences 587 (2022) 63–81

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Deep attentive style transfer for images with wavelet decomposition

Hong Ding [a,b], Gang Fu [b], Qinan Yan [c], Caoqing Jiang [a], Tuo Cao [b], Wenjie Li [b], Shenghong Hu [b], Chunxia Xiao [b,*]

[a] School of Information and Statistics, Guangxi University of Finance and Economics, 530003, China
[b] School of Computer Science, Wuhan University, 430072, China
[c] JD.com American Technologies Corporation, CA 94043, United States

ARTICLE INFO

Article history:
Received 6 February 2021
Received in revised form 25 November 2021
Accepted 26 November 2021
Available online 15 December 2021

Keywords:
Deep learning
Wavelet transform
Image style transfer
Photorealistic style
Artistic style

ABSTRACT

To solve the issue of texture preservation in the image style transfer process, this paper presents a novel style transfer method for images that often contain tiny details but are easily noticed by human subjects (*e.g.*, human faces). We aim to achieve content-preserving style transfer via an appropriate trade-off between detail preservation and style transfer. To this end, we utilize wavelet transformation with a deep neural network for decoupled style and detail synthesis. Additionally, style transfer should involve a one-to-one correspondence of semantic structures of scenes and avoid noticeable unnatural-looking style transitions around them. To address the above issue, we leverage an attention mechanism and semantic segmentation for matching and design a novel content loss with local one-to-one correspondence for producing content-preserving stylized results. Finally, we employ wavelet transform to perform feature optimization (FO) to repair some imperfect results. We perform various experiments with Qabf evaluation and a user study to validate our proposed method and show its superiority over state-of-the-art methods for ensemble and texture preservation.

1. Introduction

Style transfer is regarded as a challenging but interesting task in both academia and commerce. To obtain varied and graceful facial images, researchers have proposed many methods to modify image attributes, including color, illumination, shape and semantic segmentation [37,1,9,46,31,12,8,26]. Recent works in artistic style transfer [5,41,13,36,30,10], photorealistic style transfer [18,15,43,49], and makeup style transfer [45,39] have achieved remarkable progress. For artistic style transfer, a model extracts the texture and color information from the referenced artistic image, which is added back into the content image after transformation, to obtain a stylized image (as shown by R_1 in Fig. 1). For photorealistic style transfer, the model applies the reference color style on the scene without removing the details of the content image (as shown by R_2, R_3 in Fig. 1). Despite the success of style transfer, this field remains challenging owing to the requirements of content preservation and semantic consistency. Some methods [41] focus on artistic style transfer, which imposes a weak effect on the transfer of realistic style; some methods [43,15] focusing on photorealistic style transfer exert an undesired effect

* Corresponding author.
E-mail addresses: dhong20123@163.com (H. Ding), xyzgfu@gmail.com (G. Fu), qingan.yan@jd.com (Q. Yan), jcqng@163.com (C. Jiang), maplect@whu.edu.cn (T. Cao), cslwj@whu.edu.cn (W. Li), wuhanhush@126.com (S. Hu), cxxiao@whu.edu.cn (C. Xiao).

https://doi.org/10.1016/j.ins.2021.11.077
0020-0255/

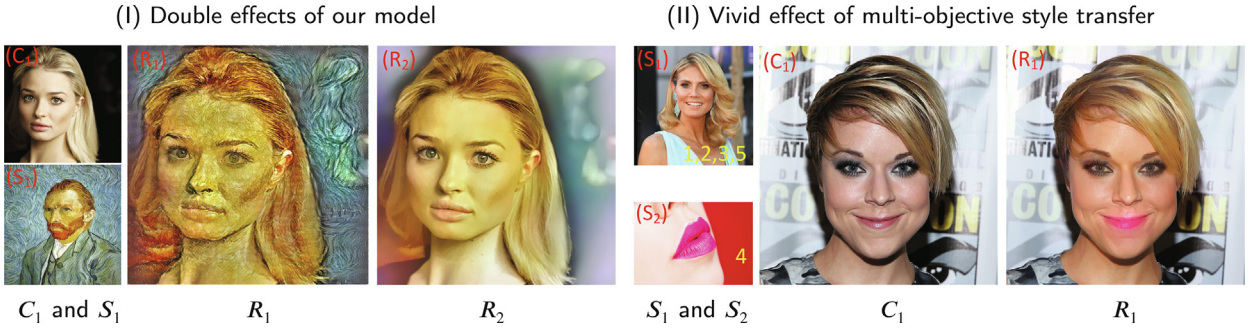

Fig. 1. (I) Double effects of our model. In the left part, our method globally transfers the style of the reference image S_1 to the content image C_1 to synthesize R_1 with artistic style and R_2 with photographic style. (II) Vivid effect of multi-objective style transfer. In the right part, we locally transfer the hair, skin, and background style of S_1 and the mouth style of S_2 to C_1, generating the stylized image R_1. We encourage readers to zoom in on all figures in the paper to observe the tiny details for visual comparison.

on artistic style transfer, *i.e.*, lost details and even structures. Human subjects are often sensitive to structural irregularities and even small distortions, which make the stylized result unrealistic (*e.g.*, human faces and buildings). Hence, we propose a novel content-preserving style transfer method that can perform artistic style and photorealistic style transfer for facial and scene-level images.

Most previous methods [5,18] often perform style transfer by minimizing the difference of a global representation of the Gram matrix. However, this fails to encourage local semantic consistency for facial areas such as eyes, mouth, and some other scene areas. These methods thus do not work well for facial and some scene images since photorealistic style transfer needs to preserve appropriate content details while transferring style. A recent representative work [18] used the Gram matrix with semantic segmentation and achieved impressive results for scene-level images. However, this method also fails to handle facial images since it has no scheme to preserve the smallest details in the images. Li et al. [15] proposed a photorealistic variant of whitening and coloring transform (PhotoWCT) that replaced the upsampling components of the VGG decoder with unpooling. This cannot solve the information loss from the max-pooling of the VGG network, which occasionally blurs artifacts. Yoo et al. [43] propose a wavelet corrected transfer based on whitening and coloring transforms (WCT^2) that substitutes the pooling and unpooling operations in the VGG encoder and decoder with wavelet pooling and unpooling. The decomposed wavelet features provide interpretations on the feature space, such as component-wise stylization. Nevertheless, this method does not handle the semantic segmentation boundary well, and the style transfer results have unnatural boundaries at the point of semantic segmentation. For facial image style transfer, this problem is particularly apparent.

Different from the component-wise stylization proposed by Yoo et al. [43], in this paper, we propose a novel learning-based image style transfer with wavelet feature extraction and repair. Specifically, given a content image and a (multiple) reference image(s), we first estimate its high frequency map (HFM) using wavelet transformation and perform a feature filling operation on HFM to obtain a refined HFM map with clearer details. We then treat this refined map as the actual content image, and combine it with the reference image(s) to generate the final result via an encoder-decoder model. Employing wavelet transform, we can also perform feature optimization (FO) to repair the imperfect results. In addition, combined with the attention mechanism, we design a novel content loss term with semantically meaningful one-to-one correspondences between image parts of content images and output images. In particular, for the facial image pairs with significant differences in face shape and appearance, our method is able to achieve photorealistic and visually pleasing results. The results are shown in Fig. 1. We have evaluated the proposed approach on a variety of test images (including facial images and scene-level images) from the Internet and public datasets, such as the IMDB-WIKI dataset [25], and coco dataset [33], and validate the superiority of our proposed method. An overview of the proposed method is shown in Fig. 2.

In summary, the major contributions of this work are as follows:

- We propose a deep style transfer with wavelet decomposition, which can effectively remove the disturbance of the style of the content image with respect to the output during the style transfer. When the content is an artistic image, we can generate a better and more natural-looking style result and artistic style with more texture information.
- We design a novel content loss term with semantically meaningful one-to-one correspondences between the content image and output stylized image, which can avoid structure distortion in the result.
- We propose a postprocessing step, called feature optimization (FO), which uses the high frequency of the content and the low frequency of the result to repair the missing textures.

2. Related work

Style transfer: Global statistics from one image to another can successfully mimic a visual look for cases such as landscapes [22,47,18]. For example, the methods of [22,38] applied global statistics to transfer the illumination of the source

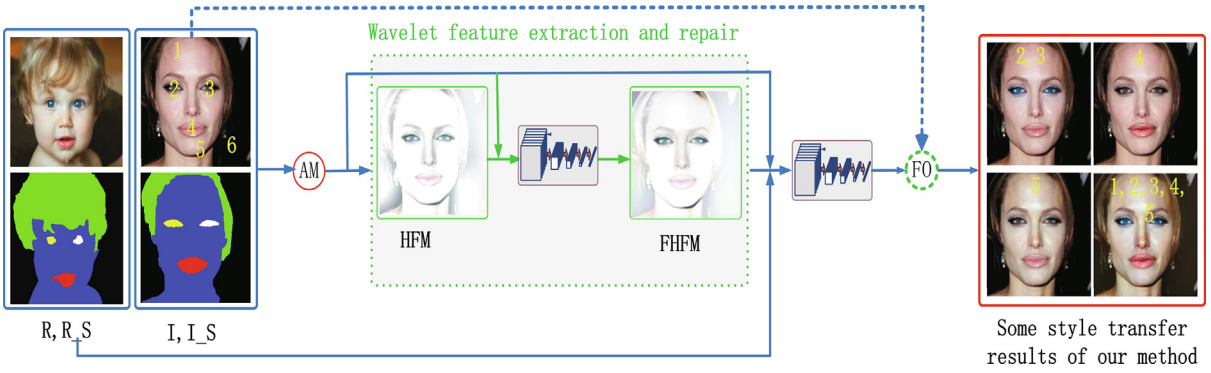

Fig. 2. Schematic illustration of the proposed method: (i) we leverage semantic segmentation for the content (C) and the reference (R) to create their semantic segmentation — C_S and R_S. (ii) We utilize the attention mechanism (AM) to specify the focused areas illustrated in Sections 3.5 and 3.6. Here, 1, 2, 3, 4, 5 and 6 denote hair, left eye, right eye, mouth, skin and background, respectively. (iii) We use wavelet transformation to extract the high frequency map (HFM) from the content image; (iv) we further recover the lost structure details of HFM to produce the feature-filling high frequency map (FHFM); (v) we feed C, C_S, R, R_S, the focused areas, FHFM, and reference image into a convolution neural network to generate the style transfer result. (vi) Using the content image and the model output, we perform feature optimization (FO) as a postprocessing step to repair some outputs with missing texture or detail.

image to the target image. Shi et al. [29] presented local and multiscale techniques to robustly transfer the local statistics of an example portrait onto a new one. While this transfer technique is local and mainly tailored for headshot portraits with very close poses, it cannot work for scene images. Altering the illumination on a face is also a common operation for face recognition and face relighting [20,16]. Zhao et al. [48] proposed automatic semantic style transfer using deep convolutional neural networks and soft masks. However, this method removes some of the texture information of the content image and produces an effect with a more artistic style. Rodriguez et al. [24] proposed a style transfer method via adaptive instance normalization, which was based on a labeled source set for object training and detection. Chen et al. [2] proposed image sentiment transfer using the filtered Visual Sentiment Ontology (VSO) dataset. Hence, the methods of [24,2] limited their application because of the special datasets. Yulun et al. [44] introduced multimodal style transfer via graph cuts that are matched with local content features under a graph cut formulation but cannot realize local style transfer. Meier [30] proposed two-stage peer-regularized feature recombination for image style transfer, which cannot retain the detailed texture of content images.

To achieve photorealistic image style transfer, Li et al. [15] designed PhotoWCT. To maximize the stylization effect, they recursively transformed features in a multilevel manner from coarse to fine. Although the strategies are valid, they fail to solve the information loss problem, which occasionally blurs artifacts. Wang et al. [36] improved Li's method to increase the diversity without distinct improvement of the quality of stylization. Yoo et al. [43] address the fundamental problem by introducing a theoretically sound correction to the downsampling and upsampling operations. They propose the WCT^2 that substitutes the pooling and unpooling operations in the VGG encoder and decoder with wavelet pooling and unpooling. This allows WCT^2 to fully reconstruct the signal without any postprocessing steps. Nevertheless, the WCT^2 algorithm generated a clear boundary line, which caused apparent artistic traces. Zheng et al. Roey et al. [19] presented an alternative loss function that does not require alignment. However, the method cannot work well for common scene and face images. [49] proposed a method of image synthesis using masked spatial-channel attention and patch-based self-supervision. However, it changed the content texture. Hence, existing style transfer methods often destroy the content details of the input images and even considerably blur the whole image, producing feature blurring and unrealistic results.

The main difference from these methods is that our work intends to enhance the style transfer effect by removing most of the style information from the content image, aims at retaining more detail of the input face and scene image, and transfers the local and multiobjective style of the reference image to the content image.

Makeup transfer: Different from style transfer, the goal of makeup transfer focuses only on transferring eye shadow, lipstick, skin color, and not background [39,7,45]. Some makeup transfer work does not transfer the hair style [17,14]. These makeup transfer methods do not work well for image pairs without strict dense correspondence. Hence, the makeup transfer methods cannot work well for common scenario images, including whole facial images. Some outputs of the methods [39,45] depend on dense dataset correspondence. Instead, our work aims at the common facial images without dense correspondence and performs style transfer for both local and global regions of the facial images. Moreover, our method can handle general facial image pairs with large differences in background and face shape.

Attention mechanism: The attention mechanism allows the model to focus on the most relevant parts of images or features by incorporating an attention mechanism into a deep learning framework [40]. Shaw et al. [27] presented an alternative approach, and Yao et al. [41] performed a multiscale style swap on content features and style features, which cannot be used for local regions. Because people's attention to the image is relatively fixed, we perform semantic segmentation for several main attention areas of the images. Hence, different from the above works, we can perform style transfer for the attention areas of images which are most focused upon by viewers.

Content loss: Many research works have been devoted to content loss design in style transfer [4,16,29,28]. Gatys et al. [5] presented a flexible iterative optimization approach based on a pretrained VGG19 network. Beyond that, many methods have been proposed to address different aspects of quality [6,23,32,35,11], diversity [42,34], and photorealism [18]. However, the image quality achieved by the costly optimization in [5] still remains an upper bound for the performance of recent methods. Different from the above works, we construct the content loss considering multiple regions instead of the whole image. Our method can be trained on arbitrary unpaired content and style images and is accurate enough to be used for image style transfer.

3. Approach

3.1. Motivation

For both artistic and realistic style transfer, many methods [41,43,15] focus on achieving the effect of style transfer while preserving more texture information of the content image. At the same time, we notice that the original style of the content image has a particular influence on the style transfer. Hence, we first introduce the optimized content loss function and then propose the wavelet feature extraction and feature repair structure to extract the details of the content image before the style transfer, remove the original style, and then conduct the style transfer. These two measures can retain more details of the content image during artistic and photorealistic style transfer.

3.2. Overview

The goal of our method is to transfer the style of a reference image to that of a content image. We employ an attention mechanism (AM) to annotate the only local areas of content on which users are focused. Our model performs style transfer for only these areas. We utilize VGG19 net to calculate the characteristics of each convolution layer of the content image. According to the features of these convolutional layers, we save the original image feature. We restore the image content by content loss and image style by style loss. We use the module of wavelet feature extraction and feature repair to extract the texture of the content and remove its original style to minimize the influence on the style transfer result. We use the feature optimization (FO) module to further refine the style transfer result.

Fig. 2 illustrates the pipeline of the proposed method. Our network takes the content C, reference image R and their semantic segmentation C_S and R_S as inputs and outputs the style transfer result in an end-to-end manner. For clarity, let $\{C, R\}$ with a fixed order denote the input image pair. We first leverage semantic segmentation for the content and the reference. Then, we utilize an attention mechanism (AM) to transfer the style of the focusing areas illustrated in Sections 3.5 and 3.6. We extracted the high-frequency map (HFM) of the content image using wavelet transformation. Then, we fed $\{\mathrm{HFM}, C\}$ into the prediction module, a convolution neural network (CNN), to obtain the feature-filling high frequency map (FHFM). The main reason for this is that excluding the style of the content image significantly benefits the subsequent style transfer. We fed $\{\mathrm{FHFM}, R, C\}$ into the prediction module to generate the style transfer result. Finally, we performed feature optimization (FO) to repair some outputs with lost texture or detail. The whole pipeline can generate photorealistic style transfer results, as shown by R_2 of the left part and R_1 of the right part in Fig. 1. Removing the parts corresponding with wavelet feature extraction and repair, we can obtain the artistic style transfer result, as shown by R_1 of the left part in Fig. 1 and the results in Fig. 9.

In this paper, we adopt the same simplified semantic categories with [18]. We first conduct semantic segmentation on the content and style images and then utilize wavelet analysis and a deep neural network to perform style transfer, as illustrated in Fig. 2.

3.3. Wavelet transfer for feature extraction

Subimages of low frequency present the general characteristics of the original image, while the high frequency subimages reflect the texture details of the original image. We perform a multiple-scale wavelet transform to better capture the details of the content images for style transfer.

Given a content image, multiscale wavelet decomposition of the image is defined as follows:

$$\begin{cases} c_{k;n,m} = \sum_{l,j} h_{l-2n} h_{j-2m} c_{k+1;l,j}, \\ d^1_{k;n,m} = \sum_{l,j} h_{l-2n} g_{j-2m} c_{k+1;l,j}, \\ d^2_{k;n,m} = \sum_{l,j} g_{l-2n} h_{j-2m} c_{k+1;l,j}, \\ d^3_{k;n,m} = \sum_{l,j} g_{l-2n} g_{j-2m} c_{k+1;l,j}. \end{cases} \tag{1}$$

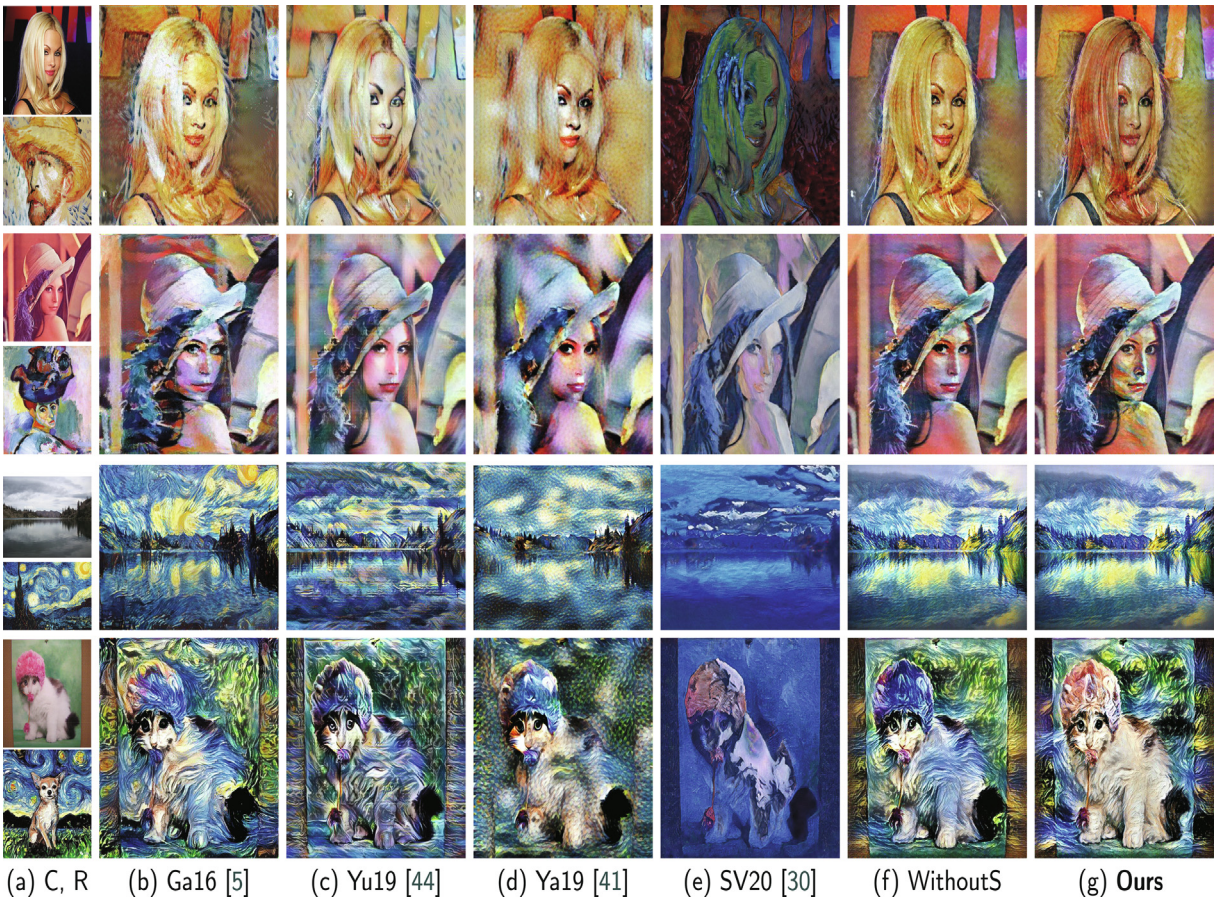

Fig. 9. Artistic style comparison results. Given (a) C and S (top: content, bottom: reference style), we obtain the results of (b) image style transfer [5], where (c) is the results of [44], (d) is the results of [41], (e) is the results of [30], (f) is our method without using semantic segmentation and (g) is our final results.

where n is the row subscript, m is the column subscript, $\{h_k\}_{k\in Z}$ satisfies the wavelet scale equation, $g_k = (-1)^k \overline{h}_{-k+1}$, h and g are called standard filters, $\overline{h}$ is the conjugate h, c is the low frequency coefficient, d is the high frequency coefficient and k is the layer number of the wavelet transform.

The subimages produced by wavelet decomposition with one level include four parts:

$$\begin{pmatrix} c_{k;n,m} & d^1_{k;n,m} \\ d^2_{k;n,m} & d^3_{k;n,m} \end{pmatrix},$$

where each subimage is one-quarter of the size of the original image. The sub-image of low frequency for each level of transformation is recursively decomposed. The reconstruction process is similar. By using this approach, the tower structure of the two-dimensional wavelet transform is constructed. Hence, the number of subimages of the high-frequency parts is $3 \times N$ times that of the low-frequency part, where N is the layer number of wavelet decomposition. Using multiscale wavelet decomposition, we can obtain the high-frequency details and the profile of the low-frequency information of the different frequencies.

Both Gaussian pyramid and wavelet decomposition are widely used for image analysis. The Gaussian pyramid produces multiple sets of signals with different scales through Gaussian blurring and downsampling. However, the Gaussian pyramid algorithm exhibits only one single frequency. The most familiar analogy to wavelet analysis is digital microscopy, as it combines multiscale and multiresolution techniques. In contrast, using wavelet analysis, we can obtain a more sophisticated internal structure of images under different frequencies, which is useful for image fusion.

In Fig. 3, given an image C_1, we use the sym4 wavelet transform to extract its high-frequency map. For the image sized 500×500 pixels, N is set as 6.

3.4. Feature filling for the high frequency map

Although the HFM contains most of the details of the content, some details may still be lost in wavelet decomposition. We perform feature filling for the high frequency map. We design a local style transfer strategy. We take the content as content and HFM as style to transfer the style of regions '1,2,3,4,5,6' in HFM to content to obtain a fine feature image (FHFM) with a rare original style. The feature filling result and the reference image are then imported into the network. The ablation study for the feature filling scheme is shown in Fig. 4.

Effect of wavelet detail extraction. Using wavelet transfer to extract the details of the content image, we can remove most of the color style of the content, thus avoiding the undesired effect of the style of the original image. The results with and without wavelet transfer are shown in Fig. 5. In row 1, the reference is an artistic image. With wavelet transfer, we transfer its color to the content and maintain a better texture in the result. In row 2, the reference is a photorealistic image. We transfer its color to the content, producing a better texture and more uniform background color in the result.

Effect of wavelet scale. Different wavelet decomposition scales will extract different degrees of image detail, which will result in different feature recovery results and different style transfer results, as shown in Fig. 6. In general, the recovery of high-frequency features of different scales will affect the details of style transfer results due to the different degrees of detail retention. The larger the scale value is, the more details of the original content image are retained, while the content image style is added gradually. In this paper, for images sized 500×500 pixels, we typically set the wavelet scale to 8.

3.5. Loss function

Our loss function of the style transfer is defined as:

$$\mathcal{L} = B_c \sum_{l=1}^{L} a_l \mathcal{L}_{content}^{att,l} + B_s \sum_{l=1}^{L} b_l \mathcal{L}_{style}^{att,l} + \lambda \mathcal{L}_p, \tag{2}$$

where L is the total number of convolutional layers, $\mathcal{L}_{content}^{att,l}$ is the content loss of the lth layer of the convolutional layer of the deep neural network, $\mathcal{L}_{style}^{att,l}$ is the style loss, and $\mathcal{L}_p$ is the photorealism expression. B_c is a weight that controls the content loss. B_s is a weight that controls the style loss of the convolutional layer. a_l and b_l are the weights to configure layer preferences. λ is a weight that controls the photorealism regularization. *att* denotes the semantic segmentation regions derived from the attention mechanism, as described in Section 3.4.

Content regularization. Instead of the global content loss in [18], we compute the loss of each corresponding labeled segmented region by using one-to-one correspondence to preserve the content details in the output image, such as textures, illumination, and colors. Furthermore, we also combine the attention mechanism with the original content regularization term. This content regularization term is written as

$$\mathcal{L}_{content}^{att,l} = \sum_{c=1}^{6} \frac{A_{K,c} W_c}{2N_{l,c}^2} \sum_{ij} \left(M_{l,c}[O] - M_{l,c}[C] \right)_{ij}^2. \tag{3}$$

$$\begin{aligned} P_{l,c}[O] &= P_{l,c}[O] H_{l,c}[C], \\ P_{l,c}[C] &= P_{l,c}[C] H_{l,c}[C]. \end{aligned} \tag{4}$$

where $M_{l,c}[\cdot]$ is the Gram matrix corresponding to $P_{l,c}[\cdot]$. $P_{l,c}[\cdot] \in R^{N_{l,c} \times D_{l,c}}$ is the feature matrix, with (i,j) indicating its index. $H_{l,c}[\cdot]$ denotes the channel c of the segmentation mask in layer l of the convolutional layer of the deep neural network.

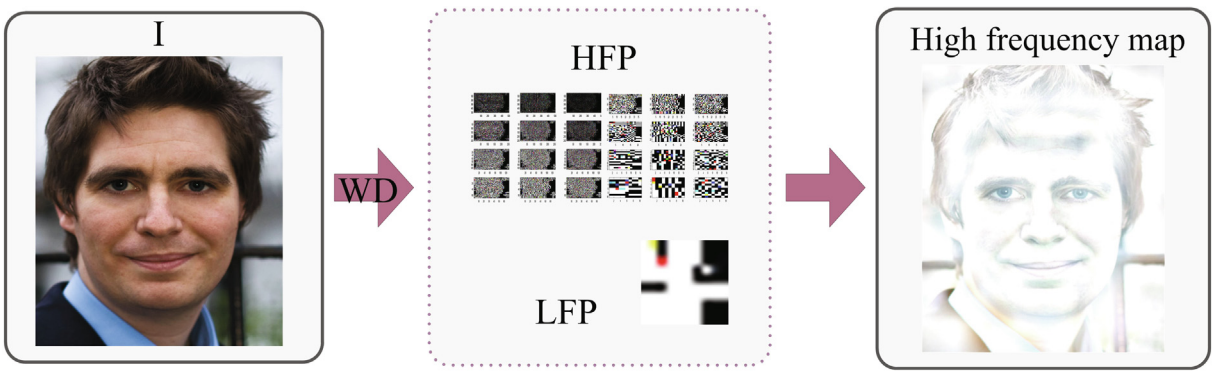

Fig. 3. Overview of extracting the high-frequency map of the wavelet. WD is wavelet decomposition. HFP is the high frequency part. LFP is the low frequency part.

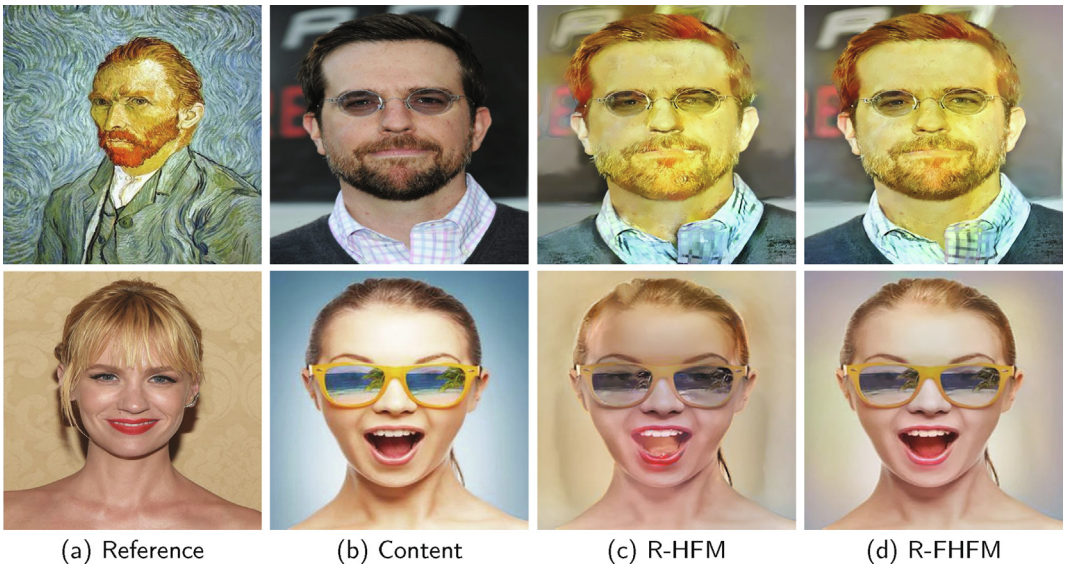

Fig. 4. Ablation study for our feature filling scheme. R-HFM is the style transfer result of HFM. R-FHFM is the style transfer result of FHFM.

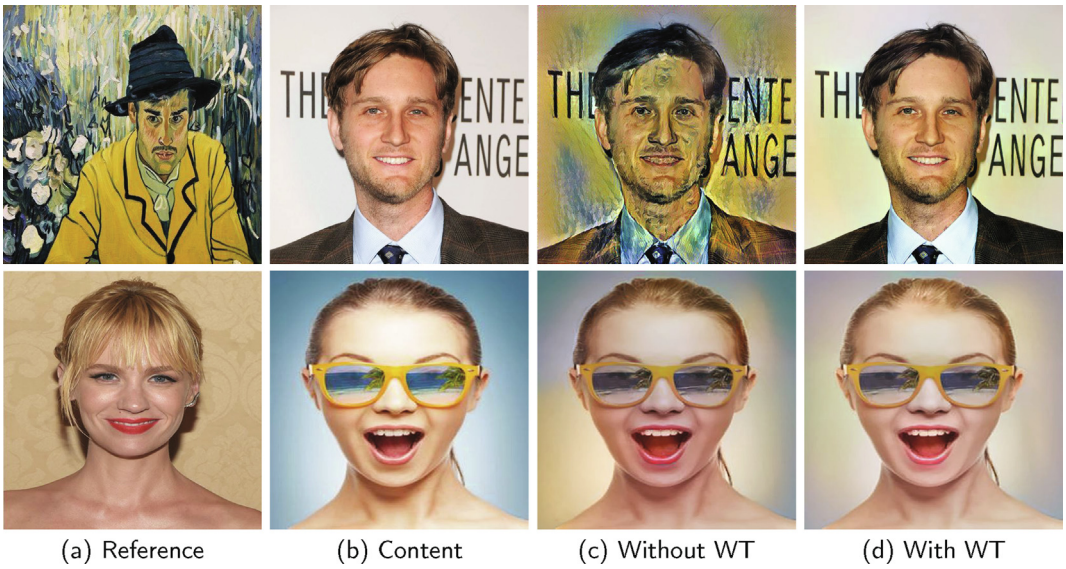

Fig. 5. Ablation study for wavelet transfer (WT). We can clearly see that the results in the last column preserve more minute details from the content images than in the third column.

$A_{K,c}$ is the cth element of A_K (for the user-specified attention area, $U_{K,c}$ is used here), O is the output, C is the content facial image, and c denotes the *cth* channels in the semantic segmentation mask. Consider facial images, for example. The total number of semantic segmentation masks of the facial image is 6. $N_{l,c}$ represents the total filters of the *lth* layer and *cth* semantic segmentation mask. $D_{l,c}$ denotes the size of a vectorized feature map. W_c is the weight of the *cth* semantic segmentation mask, which is a weight empirically specified by users and typically set to 1.

Note that instead of the feature matrix, we adopt the Gram matrix to construct this regularization term. The reason for this is that the results obtained by the inner product of the feature matrix have richer texture information than those obtained by the feature matrix, which results in better results with clearer details/features. Here, we adopt the method of [3] to perform the semantic segmentation to generate the semantic masks (including hair, eyes, mouth, skin, and background areas) for both content image and reference style image(s). The hair, left eye, right eye, mouth, skin, and background are labeled with 1, 2, 3, 4, 5, and 6, respectively, as illustrated in the attention target in Fig. 2.

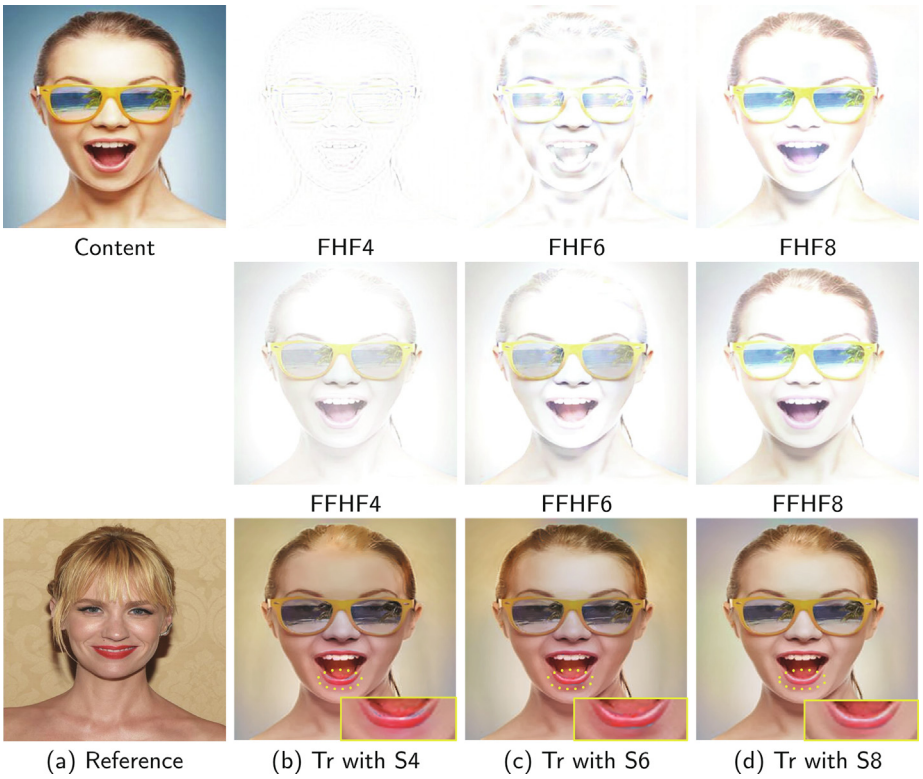

Fig. 6. Comparison results with different wavelet scales. Tr means transfer. FHF4, FHF6 and FHF8 denote the high-frequency feature maps using wavelet scales 4, 6 and 8, respectively. FFHF4, FFHF6 and FFHF8 denote the filling feature results from FHF4, FHF6 and FHF8, respectively.

Style regularization. We combine the attention areas with the original loss in [18] for transferring the styles of reference images to the FHFM shown in Fig. 2 to produce the output image. This regularization term is defined as

$$\mathcal{L}_{style}^{att,l} = \sum_{c=1}^{6} \frac{A_{K,c}}{2N_{l,c}^2} \sum_{ij} \left(M_{l,c}[O] - M_{l,c}[S]\right)_{ij}^2, \tag{5}$$

$$\begin{aligned} P_{l,c}[O] &= P_{l,c}[O]H_{l,c}[S], \\ P_{l,c}[S] &= P_{l,c}[S]H_{l,c}[S], \end{aligned} \tag{6}$$

where S is the facial style image.

Photorealism regularization. To preserve the structure of the content image and generate photorealistic results, as used in [18], we include a photorealism regularization term:

$$\mathcal{L}_p = \sum_{ch=1}^{3} Q_{ch}[O]^T J_C Q_{ch}[O]. \tag{7}$$

Here, $Q_{ch}[O]$ is the vectorized version $(N \times 1)$ of the output image O in channel ch. J_C is a matrix that only depends on the content image C.

3.6. Attention mechanism for focusing areas

When the contents are facial images, we use semantic segmentation to partition the images into 6 areas (hair, mouth, skin, background, left, and right eye areas). We utilize an attention mechanism to specify focused areas. Then, we perform style transfer for these focused areas by combining the attention mechanism and semantic segmentation. We introduce the content loss corresponding to the semantic segmentation task and obtain more precise content for the results by concatenating the segmentation channels.

Automatically specified attention area. With the result of semantic segmentation, specific areas (including hair, left eye and right eye, mouth, skin, background and the whole facial image) are automatically used to perform style transfer. We use *att_A* to express the area set as follows:

$$\begin{aligned} &att\_A = \{A_K \,|\, K = 1, 2, \cdots, 6\}, \\ &T_K = \{t_i | i = 1, 2, \cdots, \} i \leq 6, \\ &A_K = [A_{K,c} \,|\, c = 1, 2, \cdots, 6], A_{K,c} = 0, 1. \end{aligned} \tag{8}$$

where K is the serial number of the reference facial images, and K ranges within [1,6] since the maximum number of semantic facial areas is equal to 6. c is the serial number of the six semantic regions in a facial image. A_K is the Kth set of attA. T_K is the temporary variable to represent the style transfer regions of A_K. $A_{K,c}$ is the cth semantic segmentation label of A_K that will be style transferred, which is the binary matrix of A_K and is directly used in our loss function. Each element $A_{K,c}$ of the A_K matrix is 1 or 0: when the region labeled by c needs to be transferred, it is 1; otherwise, it is 0. For example, for reference style facial images, we transfer the style of the six areas from the reference image to the content and transfer the style of the whole reference image to the content, as shown in Fig. 2; then, $K = 6, att\_A = \{A_1, A_2, \cdots, A_6\}, T_1 = \{1\}, T_2 = \{2, 3\}$, $T_3 = \{4\}, T_4 = \{5\}, T_5 = \{6\}, T_6 = \{1, 2, 3, 4, 5, 6\}$. For $A_{1,c}$, we have $A_{1,1} = 1, A_{1,2} = A_{1,3} = A_{1,4} = A_{1,5} = A_{1,6} = 0, A_1 = [100000]$. For $A_{2,c}$, we have $A_{2,2} = A_{2,3} = 1, A_{2,1} = A_{2,4} = A_{2,5} = A_{2,6} = 0, A_2 = [011000]$, and so on.

User-specified attention area. In addition to the automatic style transfer mentioned above, we can also establish an interaction operation to perform style transfer. The users can provide the labeled numbers corresponding to the parts of reference style images they want to transfer. Then, we construct the new attention matrix *att_U* for style transfer to generate the desired results. For example, if there are 2 reference style facial images, we transfer the left and right eye styles from the first image to the content, transfer the skin style from the second image to the content, and then $K = 1, 2, att\_U = \{U_1, U_2\}, T_1 = \{2, 3\}, T_2 = \{5\}$. For $U_{1,c}$, we have $U_{1,2} = U_{1,3} = 1, U_{1,1} = U_{1,4} = U_{1,5} = U_{1,6} = 0, U_1 = [011000]$; and for $U_{2,c}$, we have $U_{2,5} = 1, U_{2,1} = U_{2,2} = U_{2,3} = U_{2,4} = U_{2,6} = 0, U_2 = [000010]$.

With the attention mechanism, we perform local style transfer and multiobjective transfer for facial images. The results are shown in Figs. 1, 2, 7 and 8.

3.7. Feature optimization (FO)

We find that there is more or less loss of texture in some style transfer results. To solve this problem, we propose a post-processing step, named feature optimization (FO). It should be noted that FO is different from the feature filling for the high-frequency map in Section 3.4. The latter is a repair of extracted features of the content image after removing the original style (because we find that the original style of the content image thwarts the style transfer).

Here, we utilize the process to optimize the outputs of style transfer models, whose texture or detail is lost. The high-frequency information of the content image and the low-frequency information of the output are extracted, where the wavelet scale is typically set to 6. Then, using the wavelet transform, we reconstruct the extracted wavelet coefficients to obtain the optimization result. The visual results are shown in Fig. 11. With FO, the results of all methods have clearer texture details.

3.8. Implementation details

The computer configuration used in this paper is as follows: processor: Intel ® Core™ i7-6800 k, CPU @ 3.40 GHz x 12, memory (RAM): 64.0 GB, system type: a 64-bit operating system, graphics card: GeForce GTX 1080/PCie/SSE2, and operating system: Ubuntu 16.04 LTS.

We employed the pretrained VGG-19 [18] as the feature extractor. We chose $conv11, conv21, conv31, conv41$ and $conv51$ ($a_l = 1/5$ for those layers and $a_l = 0$ for all other layers) as the content representation and $conv12, conv22, conv32, conv42$ and $conv52$ ($b_l = 1/5$ for those layers and $b_l = 0$ for all other layers) as the style representation. In the process of artistic style transfer, the parameters are set to $B_c = 9, B_s = 10^2, \lambda = 10^4$. In the process of artistic style transfer, the parameters are set to $B_c = 90, B_s = 10^2, \lambda = 10^4$.

Fig. 7. Local transfer results. 1, 2, 3, 4, 5 and 6 denote hair, left eye, right eye, mouth, skin and background, respectively. *att* denotes the style transfer area.

Fig. 8. Multi-objective style transfer. (C) Content images. (R_1) and (R_2) are the multiobjective style transfer results produced by the method in [13] and our method, respectively. (a) and (b) are the reference style images. In the first row, the areas of each referenced image need to be transferred with style according to the numbers. In the second row, we transfer the tree style of (a) and the ground style of (b) to (c) and generate (R_1) and (R_2).

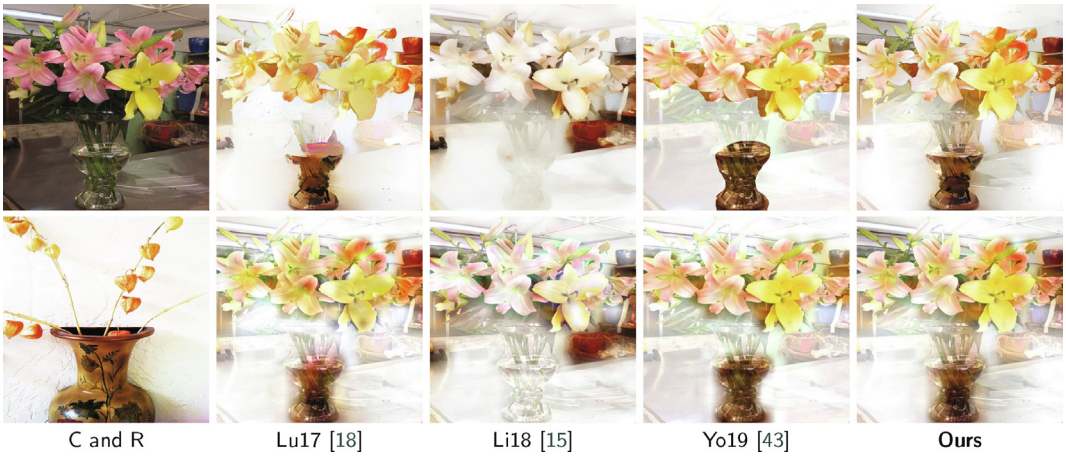

Fig. 11. Feature optimization (FO). From the left, the first column includes the content and the reference images. The first row includes the results of the comparison methods and our method without FO. The second row includes the results of the comparison methods and our method with FO.

4. Experiments

In this section, we present experiments to evaluate the performance of our framework. We first compare our method with recent state-of-the-art methods both qualitatively and quantitatively. A user study is conducted. Finally, we provide the ablation study of our key parameter B_c and some failure cases of our method.

4.1. Comparison with other methods

4.1.1. Comparison with style transfer methods.

Multiobjective-transfer comparison. We compare our multiobjective method with LEON's method [13], which can transfer the style of several reference images to the content image. Fig. 8 shows the comparison results. We use the code [13] (https://github.com/jcjohnson/neural-style) provided by the author, which is the default optimal code with multiple

Fig. 10. Photorealistic comparison results. Given (a) C and R (top: content, bottom: reference style), we obtain the results of (b) deep photo style transfer [18], where (c) is the results of [15], (d) is the results of [19], (e) is the results of [43] and (f) is our results.

Table 1
Qabf and SSIM evaluation of Fig. 9. Ga16 is [5], Yu19 means [44], Ya19 is [41], SV20 is [30] and WithoutS is our method without semantic segmentation.

Group	Ga16		Yu19		Ya19		SV20		WithoutS		Our method	
	Qabf	SSIM	Qabf	SSIM	Qabf	SSIM	Qabf	SSIM	Qabf	SSIM	Qabf	SSIM
1	0.1490	0.6808	0.1404	0.6874	0.1119	0.6227	0.1095	0.6649	0.2014	**0.8057**	**0.2077**	0.7907
2	0.1484	0.7315	0.1689	0.7623	0.1366	0.6884	0.1768	0.8379	0.2127	0.8784	**0.2161**	**0.8806**
3	0.0941	0.5604	0.1088	0.5037	0.0835	0.5401	0.0518	0.6057	0.1120	0.6826	**0.1120**	**0.6826**
4	0.0760	0.4069	0.0766	0.4122	0.0677	0.4616	0.0484	**0.5249**	0.0766	0.5148	**0.0780**	0.5088

Table 2
Qabf and SSIM evaluation of Fig. 10. Lu17 means [18], Li18 is [15], Me18 is [19] and Yo19 means [43].

Group	Lu17		Li18		Me18		Yo19		Our method	
	Qabf	SSIM	Qabf	SSIM	Qabf	SSIM	Qabf	SSIM	Qabf	SSIM
1	0.1606	0.6273	0.2235	0.8885	0.1013	0.4684	0.2428	0.8951	**0.3720**	**0.9425**
2	0.1460	0.6811	0.2490	0.9226	0.1508	0.8085	0.2483	0.8113	**0.3114**	**0.9354**
3	0.1127	0.7234	0.1892	0.9086	0.0979	0.5401	0.1498	0.8008	**0.1961**	**0.9125**
4	01376.	0.8778	0.1736	0.9471	0.1557	**0.9109**	0.1478	0.8883	**0.1947**	0.8903
5	0.1460	0.8429	0.1245	0.9154	0.1086	0.8669	**0.2439**	**0.9492**	0.2169	0.8997

style images but without masks and transfers two reference styles by mixing ruleless methods. In our results, several styles of the reference style images are harmoniously transferred to the content image to generate the output. The multiple styles from multiple images can be flexibly transferred according to the needs of users.

Artistic style comparison. We select the four recent state-of-the-art methods [5,41,44,30], which are adept at artistic style transfer of images, for artistic style transfer comparison of facial and scene images. We obtain the implementation code of each compared method from the author's homepage and adopt the default optimal parameters. Our test images with a variety of backgrounds are selected from the Internet and other papers. Several visual comparison results are shown in Fig. 9.

As presented in the figure, both the WithoutS and our results preserve more texture details of the content image for both face and scene images than other methods. They traced out more of the texture details of the content image while performing the artistic style transfer. However, in column (g), semantic segmentations endow the style transfer result with the meaning of the corresponding semantics.

When the content of the content image was largely different from that of the reference image, we performed our experiments without using semantic segmentation, such as the third row results.

Photorealistic style comparison. We selected three state-of-the-art photorealistic style transfer methods [18,15,19,43] for vivid style transfer comparison of facial and scene images. For the methods of [18,43] and our method, we used the same semantic segmentations for style transfer. Yoo's method [43] was not great at manipulating the boundaries of many objects, which left evident artificial traces. Because the method of [15] does not utilize semantic segmentation, color overflow and distortion occur in the results. [19] Luan's method [18] also includes visible defects in the facial image style transfer results. However, our work produces more pleasing results with more precise details, natural-looking styles, and vivid colors.

Fig. 10 shows the comparison results of photorealistic style transfer. In photorealistic style transfer, when the reference image is a real image, we transfer the reference style to the content image to generate a realistic style transfer result. When the reference image is artistic, we transfer style information such as color to the content image to produce the style transfer result with aesthetic flavor. Whether the reference image is an artistic or real image, our method can perform style transfer while better maintaining the texture information of the content image. The result of [18] has visible distortion in the local area. Li's method [15] has little effect on the corresponding semantic region transfer. For the method of [19] which includes a landmark algorithm, we chose the first group as the results which have better style transfer and retain more original texture than other groups. However, landmark-based methods rely on the quality of content images and the accuracy of characteristic detection; they also warp shape when transferring style images. These problems make the method perform poorly for some scene examples. Yoo's method [43] produces an unnatural boundary texture.

We use Qabf [21], and structural similarity (SSIM) [50] metrics to quantitatively evaluate style transfer results. Qabf is an objective metric for evaluating image style transfer result quality. The higher the Qabf value is, the better the transfer quality the image exhibits. SSIM is an alternative complementary framework for quality assessment based on the degradation of structural information. Compared with the content image, the higher the SSIM value is, the more structural similarity the

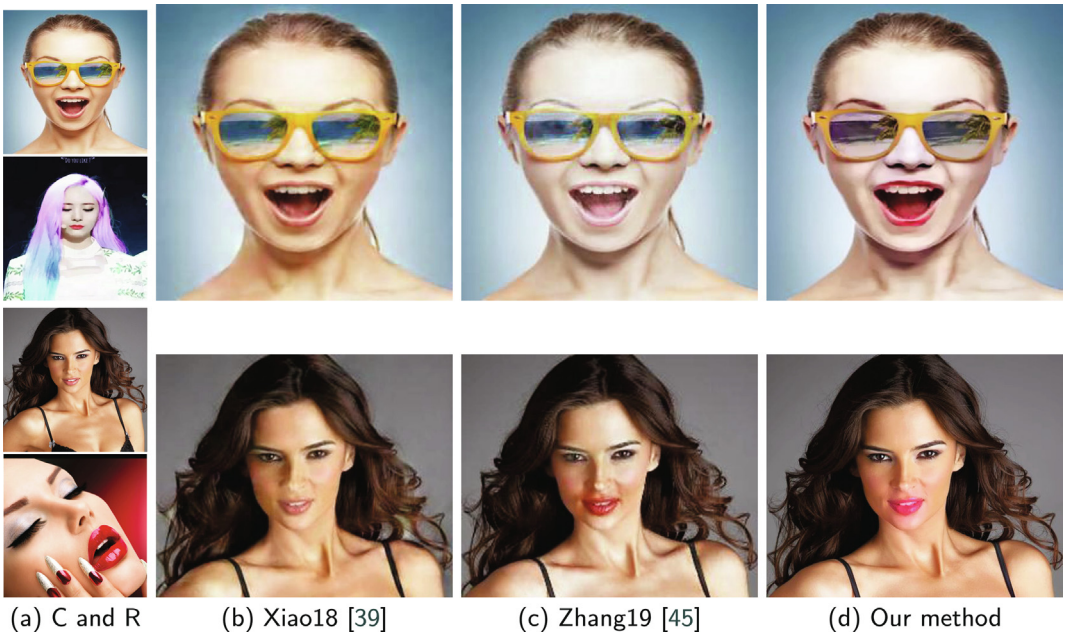

Fig. 12. Comparison results of local makeup style transfer. We transferred the style of the mouth, skin and eyes in the reference (R) to the content (C) for visual comparison: (b) shows the results of Xiao [39], and (c) shows those of Zhang[45]. In the second row, we transferred the style of the mouth in (a) to that in (b) for visual comparison.

Table 3
Vote results obtained by different methods of Fig. 9.

Methods	Qu.1	Qu.2	Qu.3	Overall
Ga16 [5]	14.33%	12.80%	12.53%	13.22%
Yu19 [44]	17.28%	16.95%	16.43%	16.88%
Ya19 [41]	13.03%	15.03%	14.00%	14.02%
SV20 [30]	13.93%	14.68%	13.48%	14.03%
WithoutS	19.88%	18.28%	21.05%	19.73%
Ours	**21.58**%	**22.28**%	**22.53**%	**22.13**%

Table 4
Vote results obtained by different methods of Fig. 10.

Methods	Qu.1	Qu.2	Qu.3	Overall
Lu17 [18]	16.43%	17.43 %	16.18%	16.68%
Li18 [15]	15.90%	17.25 %	15.45%	16.20%
Me18 [19]	15.08%	14.98%	15.63%	15.23%
Yo19 [43]	22.13%	20.38%	20.03%	20.84%
Ours	**30.48**%	**29.98**%	**32.73**%	**31.06**%

output image exhibits. To calculate the SSIM of the style transfer results more accurately, we used the sym4 wavelet transform to extract the details of the style transfer results and then calculated the SSIM value between the details and the content. The feature extraction results of Fig. 9 are shown in our supplementary material. Artistic style comparison results are shown in Table 1. It can be observed that the Qabf and SSIM values of our method with or without wavelets are higher than the others in the first three rows. In the last row, the SV20 value of SSIM is the largest because it maintains most of the texture of the content. However, it cannot transfer the reference style to the content as well as other methods. The feature extraction results of Fig. 10 are shown in our supplementary material. The photorealistic comparison results are shown in Table 2. We notice that in the fourth row, SSIM values of [19] are higher than our values because only minimal style is transferred from the reference to the content. In the fifth row, the Qabf and SSIM values of [43] are higher than our values. Although these methods maintain more texture information of the content image, their style transfer effect is not significant, and the artificial trace of the image contour is noticeable. Hence, the proposed method can produce better style transfer results than existing state-of-the-art methods both qualitatively and quantitatively.

4.1.2. Comparison with makeup transfer methods.

Makeup transfer methods [17,14,7] do not work well for image pairs (one content image and one or multiple reference images) without dense correspondence. Zhang et al. [45], and Xiao et al. [39] transfer makeup using the generative adversarial network to produce more facial image pairs. Their output is dependent on the trained dataset. When the content face image is substantially different from the reference image, the results are not compelling, as shown in Fig. 12.

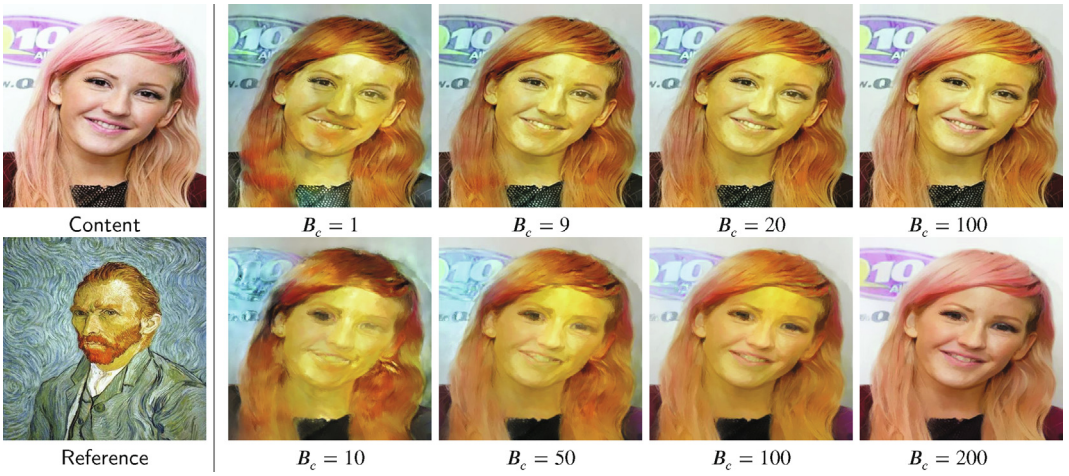

Fig. 13. Transferred result comparison with varying B_c. **1st row:** Our method; **2nd row:** results of [18].

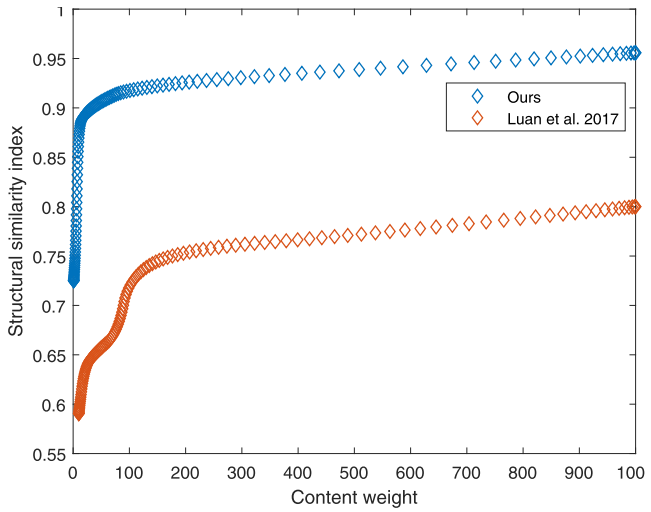

Fig. 14. Effect of varying B_c on the detail similarity between the output and input images.

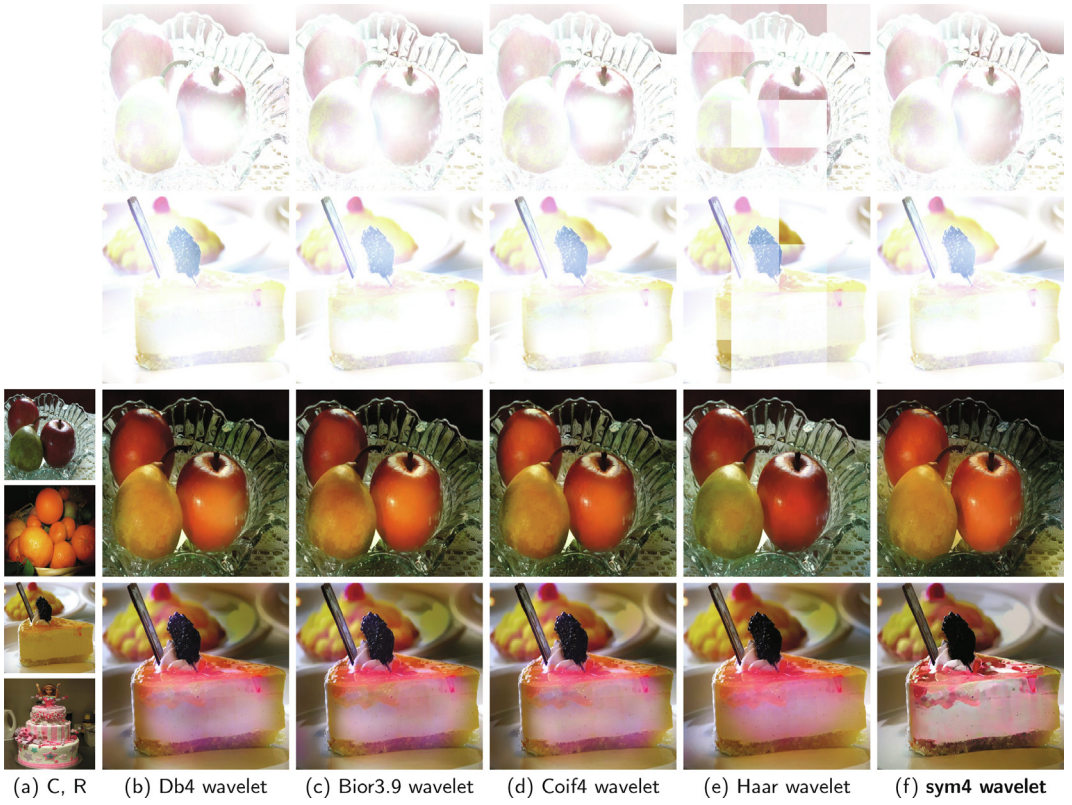

Fig. 15. Comparison results with different wavelets. Given (a) C and R (top: content, bottom: reference style), we obtain the results of the (b) Db4 wavelet, (c) Bior3.9 wavelet, (d) Coif4 wavelet, (e) Haar wavelet and (f) Haar wavelet. The last two rows are the style transfer results with different wavelets, and the first two rows are the feature extraction results of the last two rows.

4.2. User study

We performed a user study with 80 random volunteers to validate the effectiveness of the proposed method. For artistic style transfer, we randomly showed 80 groups of style transfer results of our approach, WithoutS and four other compared

methods [5,44,41] to each volunteer. Each volunteer browsed the labeled images shown in Fig. 9, and a survey was conducted to collect feedback on the following questions: (1) Which one do you think preserves the most content image detail information? (2) Which one do you think transfers the style best? (3) Which one do you prefer most? For each question, the volunteer was asked to vote for only one result. For the sake of fairness, the six methods are labeled by R_1, R_2, R_3, R_4, and R_5, while our results are in R_6. Here, the facial image dataset used is the IMDB-WIKI dataset [25]; the scene image data used are from the coco dataset [33] and Luan's paper [18]. For fair comparison, the results of each compared method are obtained via leave-one-out cross-validation. Final results are shown in Table 3.

For photorealistic style transfer, we randomly showed 80 groups of style transfer results of our approach and three other compared methods [18,15,19,43] to each volunteer. Each volunteer browsed the labeled images shown in Fig. 10, and a survey was conducted to collect feedback on the following questions: (1) Which one do you think has the fewest artifacts? (2) Which one do you think transfers the style best? (3) Which one do you prefer most? For the sake of fairness, the results generated by the five methods are labeled by R_1, R_2, R_3 and R_4, while our results are in R_5. Final results are shown in Table 4.

Let V_{ij} denote the total votes of R_i on the *jth* question. To evaluate each method of the individual question, we compute the percentage of votes (PoV) obtained by R_i for the jth question by $PoV = (V_{ij}/4000) * 100\%$. To provide an overall evaluation of different methods, we further calculate the percentage of votes obtained by R_i in all by $\overline{PoV} = \left(\sum_{j=1}^{3} V_{ij}\right)/12000 * 100\%$. In Tables 3 and 4, we provide the percentages of votes obtained by different methods of the survey, where Qu.x denotes the *x*th

Fig. 16. Photorealistic style comparison of our method against state-of-the-art style transfer methods without using semantic segmentation results as guidance. (a) is C and S (top: content, bottom: reference style), (b) is [18], (c) is [43] and (d) is our method.

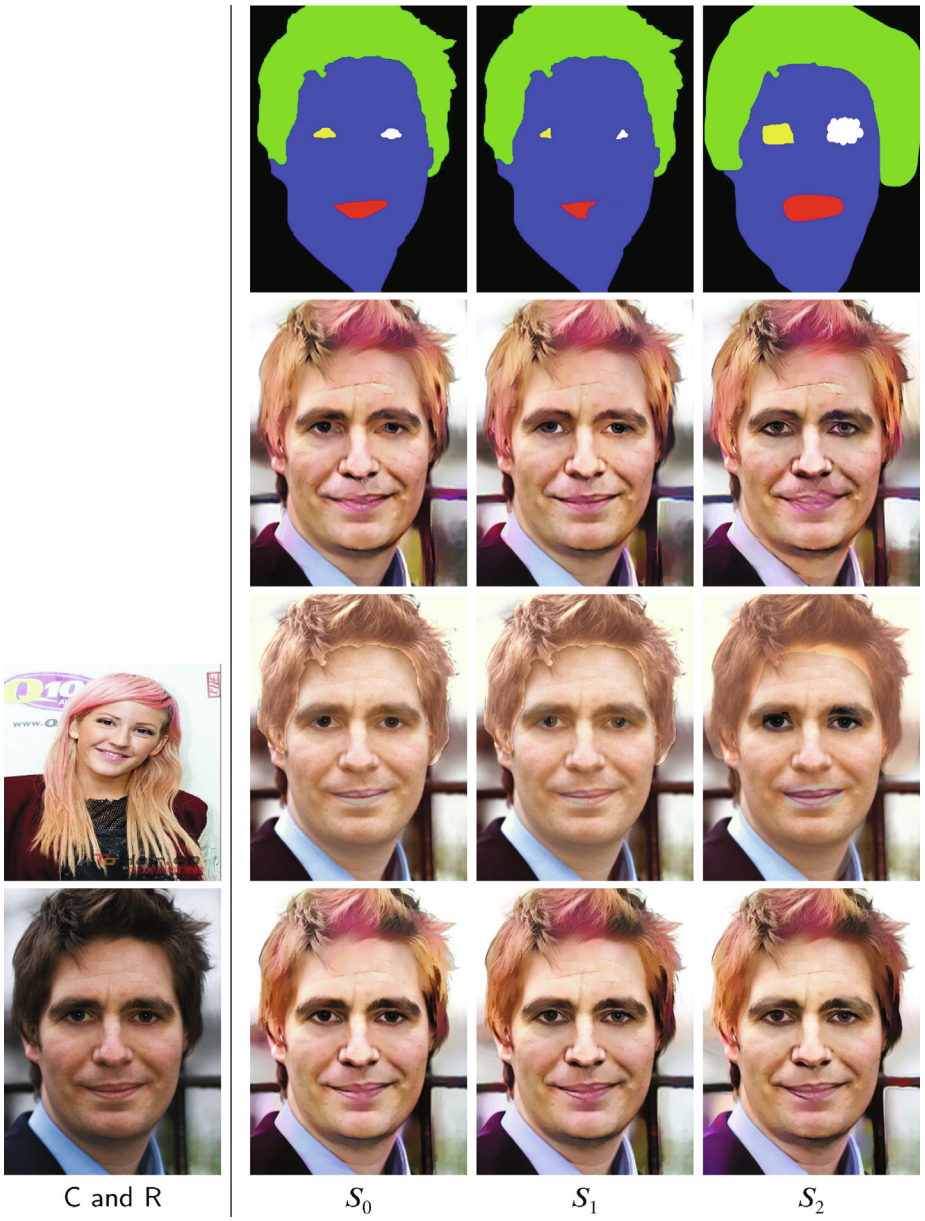

Fig. 17. Transferred result comparison with varying semantic segmentation. **1st row:** varying semantic segmentation of the content; **2nd row:** results of [18]; **3rd row:** results of [43]; **4th row:** Our method.

question. From the tables, we can see that our method has achieved the highest scores. This means that our results are preferred by human subjects.

4.3. Discussion

Influence of the parameter B_c. We change B_c while keeping the other parameter unchanged to see how the transfer result varies, as shown in Fig. 13. We can see that our method produces better results with clearer details and more natural-looking appearance than that of [18]. For example, with a small B_c, our method clearly generates the pupils of the eyes, whereas the method of [18] cannot work well even with a large B_c value and stabilized output. Since our results can further reduce the loss of facial features, postprocessing can be carried out, such as iteration in multiobjective transfer. Moreover, to further observe the ability to preserve the details/features while transferring style, we also plot the curve of structure similarity index (SSIM) values between the detail components of the content image and the output image with varying B_c, as shown in Fig. 14. We can see that our SSIM values are larger than those of [18], which means that our method

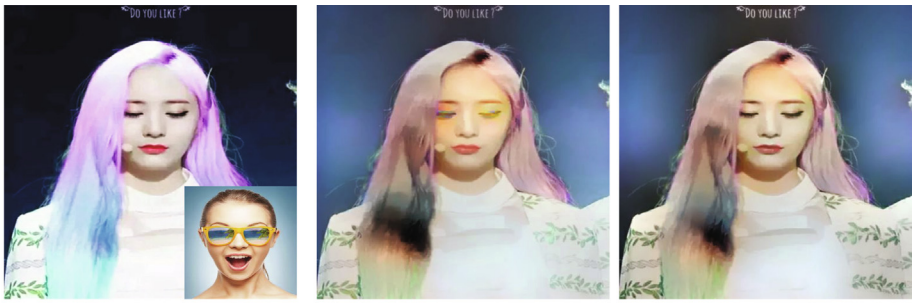

Fig. 18. Failure case of our method. TS denotes transferring the style of eyes.

can preserve the details/features as much as possible while transferring style to obtain better results. In other words, our method is more robust to yield a trade-off between preserving details/features and transferring style than is the method of [18]. This ability of our method benefits from our special design of the content regularization term with semantic one-to-one correspondence.

Influence of wavelet. We choose four wavelets, including Haar, DbN, Coifn, Symm, and Biornr.nd, to illustrate the effects of the wavelets on style transfer results.

The Haar wavelet is discontinuous, which will cause a block effect in the restored image. The DbN wavelet is a tightly supported orthonormal-normalized wavelet, most of which have no symmetry and poor smoothness. The CoifN wavelet has better symmetry than DbN. The SymN wavelet is an approximately symmetric wavelet function, which is an improvement upon the DbN wavelet. Biornr.nd is a biorthogonal wavelet in which two functions are used to decompose and reconstruct the image. It includes a linear phase and can avoid distortion. We compared the feature extraction and style transfer results of the Sym4 wavelet with those of the Db4, Bior3.9, Coif4 and Haar wavelets. Results are shown in Fig. 15. In the first two rows, the feature extraction of the Haar wavelet has block effects. In the third row, the style transfer results of both the fruits and the plate with the Sym4 wavelet are closer to the reference than other wavelets. In the last row, both the style transfer result and the texture of the cake with the Sym4 wavelet are better than those of the others.

Influence of semantic segmentation. In Fig. 9, we show the influence of semantic segmentation for artificial style transfer. We also compare the photorealistic style with existing methods without using semantic segmentation. The comparison results are shown in Fig. 16, in which we compare our work with the methods of [18,43] without using semantic segmentation. All methods transfer style without changes, and our results include more texture information and style transfer effects.

Influence of semantic segmentation accuracy. Our method improves the content loss function and enhances the accuracy of the content after image style transfer. When there is some error in semantic segmentation, our model can still generate good image style transfer results. Nevertheless, other methods are more significantly influenced by the semantic segmentation results. Some comparison results are shown in Fig. 17.

Limitations. Our method also has some limitations. When the face is covered by something such as large glasses, our method, as well as existing methods, produces unsatisfactory results with unnatural-looking style transitions, as shown in Fig. 18. As a workaround to alleviate this issue, we perform a local style transfer only for the areas of '1, 4, 5, 6' (excluding the eyes) to achieve the better result (see (c)).

5. Conclusion and future work

In this paper, combining an attention mechanism and semantic segmentation, we have proposed a novel style transfer approach combined with wavelet decomposition for facial images. Our approach successfully suppresses distortion and yields natural-looking results. The excellent performance of our method benefits from the novel wavelet decomposition algorithm, which reduces the influence of the content image style, and the content regularization term, which considers the one-to-one corresponding correlation of semantic parts between image pairs. We evaluate our method on a variety of scene and facial images to show its superiority over state-of-the-art methods. In the future, we will extend our method to style transfer for facial videos, making it applicable to more tasks.

CRediT authorship contribution statement

Hong Ding: Methodology, Formal analysis, Writing – original draft. **Gang Fu:** Writing – review & editing. **Qinan Yan:** Writing – review & editing. **Caoqing Jiang:** Investigation, Funding acquisition. **Tuo Cao:** Software. **Wenjie Li:** Software. **Shenghong Hu:** Investigation. **Chunxia Xiao:** Writing – review & editing, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is partially supported by the Key Technological Innovation Projects of Hubei Province (2018AAA062), NSFC (No. 61972298), Wuhan University Huawei GeoInformatics Innovation Lab, NSFC (No. 61662003), and the Humanities and Social Science Project of Ministry of Education (18YJCZH050).

References

[1] S. Ayas, M. Ekinci, Single image super resolution using dictionary learning and sparse coding with multi-scale and multi-directional gabor feature representation, Inf. Sci. 512 (2020) 1264–1278.
[2] T. Chen, W. Xiong, H. Zheng, J. Luo, Image sentiment transfer, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4407–4415.
[3] L.C. Chen, G. Papandreou, K.I. Deeplab, Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2016) 834–848.
[4] O. Frigo, N. Sabater, J. Delon, P. Hellier, Split and match: Example-based adaptive patch sampling for unsupervised style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 553–561.
[5] L.A. Gatys, Ecker, Image style transfer using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2414–2423.
[6] L.A. Gatys, A.S. Ecker, M. Bethge, A. Hertzmann, E. Shechtman, Controlling perceptual factors in neural style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3730–3738.
[7] S. Gu, J. Bao, Yang, Mask-guided portrait editing with conditional gans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3436–3445.
[8] H.T. Hu, L.Y. Hsu, H.H. Chou, An improved svd-based blind color image watermarking algorithm with mixed modulation incorporated, Inf. Sci. 519 (2020) 161–182.
[9] C.H. Hua, T. Huynh-The, S.H. Bae, S. Lee, Cross-attentional bracket-shaped convolutional network for semantic image segmentation, Inf. Sci. 539 (2020) 277–294.
[10] H. Huang, X. Liu, R. Yang, Image style transfer for autonomous multi-robot systems, Inf. Sci. 576 (2021) 274–287.
[11] Y. Jing, L. Yang, Y. Yang, Z. Feng, Y. Yu, M. Song, Stroke controllable fast style transfer with adaptive receptive fields, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 244–260.
[12] M. Kaur, D. Singh, V. Kumar, K. Sun, Color image dehazing using gradient channel prior and guided l0 filter, Inf. Sci. 521 (2020) 326–342.
[13] A. Leon, A.S.E. Gatys, Controlling perceptual factors in neural style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3985–3993..
[14] T. Li, R. Qian, Dong, Beautygan: Instance-level facial makeup transfer with deep generative adversarial network, in: Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 645–653.
[15] Y. Li, M.Y. Liu, X. Li, M.H. Yang, J. Kautz, A closed-form solution to photorealistic image stylizations, in: Proceedings of the European conference on computer vision, 2018, pp. 1–16.
[16] J. Liao, Y. Yao, L. Yuan, G. Hua, S.B. Kang, Visual attribute transfer through deep image analogy, ACM Trans. Graphics 36 (2017) 120–133.
[17] S. Liu, X. Ou, Qian, Makeup like a superstar: Deep localized makeup transfer network, 2016. arXiv preprint arXiv:1604.07102..
[18] F. Luan, S. Paris, E. Shechtman, K. Bala, Deep photo style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6997–7005.
[19] R. Mechrez, I. Talmi, L. Zelnik-Manor, The contextual loss for image transformation with non-aligned data, in: Proceedings of the European conference on computer vision, 2018, pp. 768–783..
[20] P. Peers, N. Tamura, W. Matusik, P. Debevec, Post-production facial performance relighting using reflectance transfer, ACM Trans. Graphics 26 (2000) 52–63.
[21] G. Piella, H. Heijmans, A new quality metric for image fusion, in: Proceedings of the IEEE International Conference on Image Processing, 2003, pp. 173–176.
[22] E. Reinhard, M. Ashikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Comput. Graphics Appl. 21 (2002) 34–41.
[23] E. Risser, P. Wilmot, C. Barnes, Stable and controllable neural texture synthesis and style transfer using histogram losses, 2017. arXiv preprint arXiv:1701.08893..
[24] A.L. Rodriguez, K. Mikolajczyk, Domain adaptation for object detection via style consistency, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 740–755.
[25] R. Rothe, R. Timofte, L.V. Gool, Dex: Deep expectation of apparent age from a single image, in: Proceedings of the IEEE International Conference on Computer Vision Workshop, 2016, pp. 252–257.
[26] L.F. Scabini, L.C. Ribas, O.M. Bruno, Spatio-spectral networks for color-texture analysis, Inf. Sci. 515 (2020) 64–79.
[27] P. Shaw, J. Uszkoreit, A. Vaswani, Self-attention with relative position representations, 2018. arXiv preprint arXiv:1803.02155..
[28] Y. Shih, S. Paris, F. Durand, W.T. Freeman, Data-driven hallucination of different times of day from a single outdoor photo, ACM Trans. Graphics 32 (2013) 1–11.
[29] Y.C. Shih, S. Paris, C. Barnes, W.T. Freeman, F. Durand, Style transfer for headshot portraits, ACM Trans. Graphics 33 (2014) 1–14.
[30] J. Svoboda, A. Anoosheh, C. Osendorfer, J. Masci, Two-stage peer-regularized feature recombination for arbitrary image style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 2761–2776.
[31] C. Tang, K. Xu, Z. He, J. Lv, Exaggerated portrait caricatures synthesis, Inf. Sci. 502 (2019) 363–375.
[32] Q.C. Tian, M. Schmidt, Fast patch-based style transfer of arbitrary style, CoRR (2016), abs/1612.04337.
[33] M.M. Tsung-Yi Lin, S. Belongie, Microsoft coco: Common objects in context, in: Proceedings of the European conference on computer vision, 2014, p. 740–755..
[34] D. Ulyanov, Vedaldi, Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4105–4113.
[35] X. Wang, G. Oxholm, D. Zhang, Y.F. Wang, Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7178–7186.
[36] Z. Wang, L. Zhao, H. Chen, L. Qiu, Q. Mo, S. Lin, W. Xing, D. Lu, Diversified arbitrary style transfer via deep feature perturbation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 7789–7798.

[37] X. Wei, Y. Guo, B. Li, Black-box adversarial attacks by manipulating image attributes, Inf. Sci. 550 (2021) 285–296.
[38] C. Xiao, R. She, D. Xiao, K.L. Ma, Fast shadow removal using adaptive multi-scale illumination transfer, Comput. Graphics Forum 32 (2013) 207–218.
[39] T. Xiao, J. Hong, J. Ma, Elegant: Exchanging latent encodings with gan for transferring multiple face attributes, in: Proceedings of the European conference on computer vision, 2018, pp. 168–184.
[40] Z. Yang, X. He, J. Gao, D. Li, A. Smola, Stacked attention networks for image question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21–29.
[41] Y. Yao, J. Ren, X. Xie, W. Liu, Y. Liu, J. Wang, Attention-aware multi-stroke style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1467–1475.
[42] Yijun Li, J.Y. Chen Fang, Diversified texture synthesis with feed-forward networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2732–2738.
[43] J. Yoo, Y. Uh, S. Chun, B. Kang, J.W. Ha, Photorealistic style transfer via wavelet transforms, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9036–9045.
[44] Yulun Zhang, Y.W. Chen Fang, Multimodal style transfer via graph cuts, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5943–5951.
[45] H. Zhang, Chen, Disentangled makeup transfer with generative adversarial network, 2019. arXiv preprint arXiv:1907.01144..
[46] T. Zhang, X. Yang, X. Wang, R. Wang, Deep joint neural model for single image haze removal and color correction, Inf. Sci. 541 (2020) 16–35.
[47] C. Zhang, Y. Zhu, Z.S.C., Metastyle: Three-way trade-off among speed, flexibility, and quality in neural style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 785–794..
[48] H.H. Zhao, P.L. Rosin, Y.K. Lai, Y.N. Wang, Automatic semantic style transfer using deep convolutional neural networks and soft masks, Visual Comput. 36 (2020) 1307–1324.
[49] H. Zheng, H. Liao, L. Chen, Example-guided scene image synthesis using masked spatial-channel attention and patch-based self-supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 1–17..
[50] W. Zhou, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process 13 (2004).