Saliency Constrained Arbitrary Image Style Transfer using SIFT and DCNN

Hui-Huang Zhao, Member, IEEE, Yao-nan Wang, and Yu-hua Li

Abstract—This paper develops a new image synthesis approach to transfer an example image (style image) to other images (content images) by using Deep Convolutional Neural Networks (DCNN) model. When common neural style transfer methods are used, the textures and colors in the style image are usually transferred imperfectly to the content image, or some visible errors are generated. This paper proposes a novel saliency constrained method to reduce or avoid such effects. It first evaluates some existing saliency detection methods to select the most suitable one for use in our method. The selected saliency detection method is used to detect the object in the style image, corresponding to the object of the content image with the same saliency. In addition, aim to solve the problem that the size or resolution is different in the style image and content, the scale-invariant feature transform is used to generate a series of style images and content images which can be used to generate more feature maps for patches matching. It then proposes a new loss function combining the saliency loss, style loss and content loss, adding gradient of saliency constraint into style transfer in iterations. Finally the source images and saliency detection results are utilized as multichannel input to an improved deep CNN framework for style transfer. The experiments show that the saliency maps of source images can help find the correct matching and avoid artifacts. Experimental results on different kind of images demonstrate that our method outperforms nine representative methods from recent publications and has good robustness.

Index Terms—Image Style Transfer, Saliency Constraint, Scale-Invariant Feature Transform, Deep Convolutional Neural Networks.

I. INTRODUCTION

Many research works have proved that convolutional neural networks have a strong ability to extract image features, which makes deep learning technology all-pervasive, especially in the field of machine learning and computer vision. Recent literature has shown that deep learning can achieve great results for problems of object detection [11], super-resolution [12], image segmentation [13], image style transfer [43], [45] and so on. Image style transfer is a new technique of synthesising an image in the style of other single image or images. Neural image style transfer is an important application of neural networks. It can produce impressive results and simulate the painter’s image [1]. It also has many successful industrial applications, such as Cartoons [9], makeup [8], and videos [10]. When common neural style transfer methods are applied in image style transfer, however, artefacts are often generated or the textures and colours from the style image are transferred unsuitably to the result image, resulting in unacceptable outputs.

This paper proposes a novel saliency constraint approach that bases on DCNNs for the image style transfer. First, the saliency detection is done for the images (including the style image and content image) to generate saliency maps. Then proposes a new loss function combining the saliency loss, style loss and content loss. In our approach, the saliency maps are utilized to capture style elements and to preserve the structure of the style image and content image. A brief summary of contributions in the paper is given as follows:

1) An amended deep convolutional neural networks model which has N channels of regular filters is augmented by concatenating a saliency detection channel. The feature maps of some special layers in the model are chosen for a loss function. Both the style and content images and their saliency detection results are input into the new convolutional neural networks model to keep the content structure and find a right place for style transferring.
2) In order to solve the problem that the size or resolution is different in the style image and content, and generate more feature maps of the style image and content for patches matching, the Scale-Invariant Feature Transform is used to generate a series of style images and content images. Those image are input to the DCNN model and used to generate local patches that will be used in patches matching. This can help to find a right local patch and generate a better style transfer result.
3) A novel loss function combining saliency loss, style loss and content loss is proposed. Unlike the common approaches, the new loss function consists of three parts: style loss, content loss and saliency loss. The result is updated by adding the gradient of saliency constraint in iterations. The saliency loss can help keep object structures in the content image.
4) Saliency constraint is utilized for three purposes. Firstly, to generate a salient map and it is then used to augment the trained convolutional neural networks. Secondly, saliency constraint is used to improve the loss function. Thirdly, the size ratio of the content salient map and the style salient map are utilized to rescale the style image. The saliency maps of two source images can help find the correct matching during style transfer.

II. RELATED WORK

Image style transfer using deep networks model. In deep learning, a convolutional neural network (CNN, or ConvNet)
is a class of deep neural networks. The success of deep CNNs (DCNNs) in image processing has also raised interest in image style transfer. The process of image neural style transfer can be accomplished in three steps. At first, given two source images (one style image and one content image), a DCNNs model (such as VGG-19) [6], [28] is trained by the style image, style features are obtained from convolution layers of the trained DCNN model. Then, the content image is selected to be input the DCNN too, and content features are also obtained from convolution layers of the trained DCNN model. At last, the third step generates a random noise image in initialization. According to the style and content features of these convolution layers, the random noise image corresponding to these features is restored by an optimization function [7], [26]. VGG-19 deep neural networks model has been widely used in style transfer research to produce promising results. Gatys et.al. [1], [23] applied VGG-19 in unguided environments to generate a new image which is similar with a famous painting. Their work has been adopted in various papers. In [9], the Gram matrix method is extended by adding several controlling perceptual factors, and measures that can control the spatial scale, spatial position and color information were introduced. In [24], an alternative method for training compact feed-forward convolution networks was proposed to significantly improve the speed of image generation. Johnson et.al. [2] presented a new image style transfer method by fusing the advantages of the feed-forward CNN training and a perceptual loss function. Instead of using global representations of styles that were computed by Gram matrices, Li and Wand [4] used the local patches in image style transfer. Based on Gram matrix. Xia et.al. [10] provided a way to transfer styles from one image to a video.

According to the way the image is optimized, there are two main types of approaches in style transfer based on deep neural networks model, global approaches based on the Gram matrix, such as [1], [3], [23] and local approaches based on patch matching, such as [4], [5], [41]. Li and Wand [4] presented a combination of generative Markov Random Field (MRF) models for image style transfer, which can gives effective transfer results and keep the information of the content image. But their approach also generates visible artefacts with uncertain cause and mistake because of matching errors between the feature maps from the style image and the feature maps from the output image. We adapt the method in [4] and attempt to improve it in our method.

According to the speed of style transfer, neural style transfer methods also can be divided into two categories: slow neural networks model image optimization and fast neural networks model image optimization. The first class of methods achieves style transfer and image reconstruction by gradually optimizing images [1], [3], [4], [23]. The second class of methods utilizes the idea of fast image reconstruction technology, optimizes the generation of offline/pre-trained models and uses a single forward transfer to generate stylized images [2], [5], [7].

In particular, Generative Adversarial Networks (GAN) has demonstrated a great power to image and video processing [29], [42]. A GAN consists of two neural networks, which is named discriminator networks and generator networks respectively, competing against each other in a zero-sum game. The generalization of discriminator/generator can be measured by the difference between its performance on training data and the whole data space. Recently several important architectures are proposed for text, image and video synthesis [35], [37], [44]. In [36], an unsupervised representation learning method with deep convolutional network named DCGANs was designed for image generation. In [5], the authors proposed an improved model named MsGAN which is based on Markovian and DCGANs for image style transfer. In [38], a new network model named conditional adversarial networks (cGAN) is investigated for solving problems in image-to-image translation. In [9], the authors proposed a model name CartoonGAN for Cartoon video style transfer.

**Saliency detection.** Human visual system has a strong ability to quickly search and find interested objects in the face of natural scenes. This visual attention mechanism is an important mechanism for processing visual information in computer vision. Saliency detection aims to quickly obtain important information from massive image and video data, and it has become an important topic in the research of computer vision [22], [31].

The advantages of saliency detection mainly lie in two aspects [15]. Firstly, it can allocate more computing resources to more important information in image or video. Secondly, the intuition of introducing visual saliency is more in line with human visual cognitive mechanism. Recently, saliency detection has exhibited important application value in target recognition, object detection [14], [16], semantic segmentation [17], image recognition [21] and so on. With a saliency detection model, we can predict which information in the content is important, and gain more visual attention through computer vision algorithms. Co-saliency detection [32] is used to discover the common saliency on multiple images. A few methods of co-saliency detection were proposed by combining existing saliency detection methods [30], [33]. Zhi et.al. [29] developed a co-saliency detection based on hierarchical segmentation and Cao et. al. [19] developed a co-saliency detection via rank constraint.

### III. Architecture

VGG [23] has shown its special characteristic for style transfer. Many aforementioned style transfer methods are based on VGG 19-layer networks model. Although there are style transfer methods based on other network models, such as ResNet [25], VGG based methods produce superior results than others. Therefore in our proposed method, we chose to use VGG19 but augment it with saliency detection channel and enhance it by introducing a novel loss function. In our augmented networks model, saliency constraint is utilized for three purposes. Like common deep CNNs, the feature maps in the layer \( l \) denoted as \( \Omega(x^l) \) and consist of \( N \) channels. Only some special layers which can mainly represent patterns from the source images are chosen for the loss function optimization. There are three novel features in our framework. Firstly salient maps are resized and inserted to the DCNNs model, and they are utilized to augment the trained model,
and find the correct matching. Secondly saliency constraint is utilized to improve the loss function, and keep the object structures in the content image. Thirdly, the scale-invariant feature transform (SIFT) [8] and the style salient map is utilized to rescale the style image. Our augmented networks model is shown in figure [4].

Figure [4] shows that some special layers (such as Conv5_1, Conv4_1, Conv3_1) which is the same in common method are chose in building content and style loss model, and the augmented networks model (left part in the figure [4]) also takes two saliency detection maps as input. The content saliency detection map and the style saliency detection map are edited as feature maps, then two maps are down-sampled to produce a saliency detection channel \( s_l \) in the 1th layer, let saliency maps have the same resolution as \( x_l \). We augmented the networks model by concatenating those feature maps and saliency maps to form the new output with \( N+ \) the saliency detection channel, defined as \( d_l \) in each selected layer and \( N \) is the number of feature maps in in the 1th layer, and it is shown in figure [1]. In order to constric the weight of salient maps, a parameter \( \beta \) is defined to balance their importance. It is shown in Eq[1] and the processing is shown in figure [2].

\[
d_l = (x_l, \beta s_l).
\]  

We have found experimentally that \( \beta = 30 \) can achieve good results.

IV. STYLE TRANSFER WITH SALIENCY CONSTRAINED

OPTIMIZATION FUNCTION

In our image style transfer method, the proposed loss function combines the saliency 7 constraint and a patch-based matching. The patches are from the feature map in VGG19 networks model and saliency maps. The loss function consists of style matching error \( L_{style} \), content reconstruction error \( L_{content} \) and salient error \( L_{saliency} \), which combines an MRF and a deep CNN model. The goal of the learning process is to minimize the loss function.

At first, two source images (a style image \( x_s \), a content image \( x_c \) ) and their salient maps \( f_{style} \) and \( f_{content} \) and are input into the augmented networks model, where \( x_s, x_c \in \mathbb{R}^{3\times w_s\times h_s}, f_{style} \in \mathbb{R}^{w_s\times h_s}, f_{content} \in \mathbb{R}^{w_c\times h_c} \), respectively. \( x \) is a output image after style transfer, and is denoted by \( x \in \mathbb{R}^{3\times w_c\times h_c} \).

In our method, we make the neural feature maps encoding of \( x \) similar to \( x_c \) and use the local patches similar to patches in feature maps of \( x_s \). The same number of new neural feature maps are generated and replace the previous neural feature maps in the networks models. Meanwhile, we penalise patch matches with two salient maps. After a backwards operation, the output is \( x \) and the style of \( x_c \) is transferred onto the layout of \( x_c \). An energy loss function is defined as follows.

\[
L(x) = a_1 L_{style}(\Omega(x), \Omega(x_s), \Omega(f_{content}), \Omega(f_{style}))+a_2 L_{saliency}(\Omega(f_{content}), \Omega(f_s))+a_3 L_{content}(\Omega(x), \Omega(x_c)).
\]  

where \( \Omega(x_s) \) is denoted as the feature map (activation) of the style image \( x_s \) in the same layer, and \( \Omega(f_{content}) \) and \( \Omega(f_{style}) \) are the salient maps of the content and style images down-sampled to the same resolution as \( \Omega(x) \) and \( \Omega(x_s) \). As shown in Eq[2] the energy loss function consists of three parts, \( L_{style} \), \( L_{saliency} \) and \( L_{content} \) are defined as the style loss function, saliency loss function and content loss function respectively, and \( a_1, a_2, a_3 \) are their coefficients, respectively. We generate \( x \) by minimizing the loss function.

Specifically, the loss function \( L_{style} \) incorporates salient maps as follows.

\[
L_{style}(\Omega(x), \Omega(x_s), \Omega(f_{content}), \Omega(f_{style})) = \sum_{i=1}^{P1} \left\| \Gamma_i(\Omega(x)) - \Gamma_i(\Omega(x_s)) \right\|^2 + \sum_{i=1}^{P2} \sum_{k=1}^{2} \left\| \tilde{\Gamma}_i(\Omega(f_{content})) - \tilde{\Gamma}_i(\Omega(f_{style})) \right\|^2,
\]  

where \( x \) is the generated image. \( \Gamma_i(\Omega(x)) \) is the local patches from \( \Omega(x) \) and there is \( P1 \) patches with each of size \( t \times t \). Similarly, \( \tilde{\Gamma}_i(\Omega(f_{content})) \) and \( \tilde{\Gamma}_i(\Omega(f_{style})) \) are the down-sampled salient maps, and there are \( P2 \) patches with each of size \( t \times t \).

\( \phi(i) \) is defined as an operation using normalized cross-correlation which can find the best matching patch. It is denoted as follows.

\[
\phi(i) := \arg\max_{j=1,...,P_s} \frac{\Gamma_j(\Omega(x_s)) \cdot \Gamma_j(\Omega(x_c))}{|\Gamma_j(\Omega(x_s))|}.
\]  

where \( \Gamma_j(\Omega(x_s)) = (\Gamma_j(\Omega(x_s)), \beta \Gamma_j(\Omega(f_{content}))) \), \( \Gamma_j(\Omega(x_c)) = (\Gamma_j(\Omega(x_c)), \beta \Gamma_j(\Omega(f_{style}))) \).

For each patch \( \Gamma_j(\Omega(x_s)) \) with salient maps \( \Gamma_j(\Omega(f_{content})) \) we find its best matching patch \( \Gamma_{\phi(i)}(\Omega(f_{style})) \) or \( \Gamma_{\phi(i)}(\Omega(x_s)) \). The best matching patch thus have both style similarity and saliency consistency.

The squared euclidean distance (SED) formula [1], [2], [4] is a common method for calculating the similarity of two images. \( L_{saliency} \) and \( L_{content} \) are denoted as a loss of SED between \( \Omega(f_s) \) and \( \Omega(f_{content}) \), between \( \Omega(x) \) and \( \Omega(x_c) \) respectively.

\[
L_{saliency}(\Omega(f_{content}), \Omega(f_s)) = \left\| (\Omega(f_{content}) - \Omega(f_s)) \right\|^2.
\]  

\[
L_{content}(\Omega(x), \Omega(x_c)) = \left\| (\Omega(x) - \Omega(x_c)) \right\|^2.
\]

V. SALIENCY DETECTION AND IMAGE RESCALE

A. Saliency detection

In this paper we aim at generating salient maps automatically. At present, the saliency detection method can be divided into two categories. The common methods which are based on local information and co-saliency methods which combine the local information and global information. We evaluate three popular common methods [22], [27], [31] and two classic co-saliency methods [29], [30] with different content images to select the one for using in our study. Saliency detection
results of these five methods in our evaluation are shown as in figure [3].

As shown in figure [3], the method in [27] generates some inaccurate results, such as those in rows 2, 3, and 5. Because some images are very abstract, the co-saliency methods [29], [30] cannot generate a satisfactory result for some style images such as in rows 2, 3, and 5 of figure [3](e, f). Comparing the results in figure [3](b-f), the method in [22] which is based on a Fully Convolutional Neural Networks (FCNN) model generate better results, so it is selected for use in our automatic salient map extraction.

B. Image Rescale with Scale-Invariant Feature Transform

During image style transfer, when the sizes of two input images (the style image and the content image) are different, or the resolutions are not the same, the result of style transfer will be affected greatly. In [2], Johnson et al. proposed a method based on super-resolution to improve the results. We can resize the two images and make the two input images' scales consistent. But this still cannot solve the size difference problem completely, because even though the size of style image and content image are the same, the size of target in two source images may be different. Scale-invariant feature transform (SIFT) is which is a feature extraction algorithm in computer vision [39]. SIFT can help to locate local features in the image, which is usually called "key points" of the image [40].

In order to solve the problem that the size or resolution of the style image and content is different, and generate more local patches for patch matching, the scale-invariant feature transform is used to generate a series of new style images and content images which have different sizes and resolutions.
Fig. 3: Saliency detection results with different methods.
After that, those new images are input to the DCNN model and used to more feature maps. These feature maps are segmented to local patches according to Eq.\(n\). This can supply more choices in for patch matching and help to find right matching patches. During our method, the input image scale-invariant feature transform is based on Eq.\(m\) and \(n\).

\[
G(w,h,\sigma) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{w^2+h^2}{2\sigma^2}\right]. \quad (7)
\]

\[
D(w,h,\sigma) = (G(w,h,k\sigma) - G(w,h,\sigma)) \ast I(w,h) = I_g(w,h,k\sigma) - L(w,h,\sigma), \quad (8)
\]

where \(G\) is a Gaussian filter, \(I_g\) is an image generated by convolution of a Gaussian filter with the original image. \(I(w,h)\) denotes an input image which is \(w \times h\), and \(k\) denotes a scale coefficient of an adjacent scale-space factor. \(\sigma\) denotes a Scale kernel factor. The process of image rascal with scale-invariant feature transform is show in figure [4].

Different scale images are obtained by changing \(\sigma\). Then the images are adjacent in the same resolution to get a Difference of Gaussian (DOG) pyramid. We have found experimentally that and \(\sigma = 1.6, k = 3\) can achieve good results.

### VI. RESULTS AND COMPARISON

We use the pre-trained 19-layer VGG-Network with the augmented network model. The stride, iteration and local patch size are set as 1, 100 and 3\(\times3\) respectively. Specially, layers Conv4\_1, Conv3\_1 are chosen as content layers, and Conv5\_1 is chosen as style layers for feature maps extraction. In order to improve the results, we run the iterations in three loops with re-sizing the image at three increasing resolutions. On a GTX Titan with 12Gb of GPU RAM, a 256\(\times256\) image takes about 15 minutes in the style transfer.

According to the experiments, the table [1] shows that when set scales size \(n = 6\), the proposed method can achieve a balance between the runtime and good results.

In our experiments, nine famous methods which represent different categories mentioned above are compared with the proposed method: local patch matching [4], gram matrix [1], [2], fast methods [2], [6], [7], and slow methods [1], [4]. We also experiment with the replacement of VGG19 by ResNet50 [25] in our evaluation.

#### A. Comparison with No Saliency Maps Using

At first, in order to show the saliency map is useful, we compare the proposed approach with the method without using saliency maps. Figure [3] shows results with different examples.

As clearly shown in figure [3] the method which doesn’t use any saliency maps generates errors, such as the facial parts (eyes, nose) that are transferred to wrong places (in rows 2 and 3 of figure [5]c), and the style of wall is transferred into the car (in rows 1 of figure [5]c). The method using saliency maps can achieve effective results and preserve the content of the source images as shown in figure [5]d). It is evident that the saliency maps of source images can help find the correct matching. They can achieve background texture and object texture style transfer separately, and prevent them from contaminating each other.

#### B. Single Object Style Transfer with Different Methods

The next set of experiments is to compare the proposed method with nine representative methods. Each of content and style images used in the experiments contains only one object (Note each pair of the content and style images are merged as shown in the first column of figure [6] to save space). The input style images like boat, train, bus and car are shown in figure [3]. And the style man and women are shown in figure [5]. The comparison results are shown in figure [6]. It is observed from figure [6] that our approach achieves effective results, and preserves the content of the images. Most results of [4] have many errors in which styles are mismatched. Method in [1] cannot transfer enough style to content images, and the results don’t contain much style information. Methods in [2], [7], [6], [6] and ResNet50 output some imperfect results too. It seems that due to the complexity of ResNet50, the feature maps do not work as well as those of VGG19 networks model. The texture style cannot be transformed properly. Because MGANS need huge training dataset, the methods based on MGANs [5] cannot generate good results when the datasets only has a limited images. The content information is destroyed badly. Although the methods of cGANS [25](iteration=100) and cGANS(iteration=200) can obtain a strong style results in different iteration, they keep very little content information. It is evident that the proposed method can generate better results in specific parts of background area and face area in the image.

#### C. Multi Objects Style Transfer with Different Methods

Next we demonstrate style transfer experiments when multiple objects are present. Figure [7] shows man and dog in the source images and the corresponding saliency detection results.

The style transfer results from images of figure [7]a) (as content) and figure [7]e) (as style) are shown in are shown in figure [8].

Figure [8] shows that methods in [2], [1], [6] and ResNet50 generate many artifacts, e.g., the style in the cloth (in style image) is transferred into to human face, see figure [8]a,b,e). The results that are generated by methods in [1], [4] are imperfect. Figure [8]d) shows mismatched style in nose and mouth. Especially, we can find that, figure [8]h) and (i) are very different from others. The two images are very similar to the style image. It is because the method in cGANS generated strong style results and they keep very little content information. The proposed approach can achieve better results in style details and keep the background clean without artifacts.

Comparing with patch-based approach method in [4], our method can achieve background texture and object texture style transfer separately, and avoid errors in transferring (mouth in figure [8]c)). A further example for images in figure [7]c) (as content) and figure [7]e) (as style) shown in figure [9].

Figures [9]c), (d) and (g) show that the cloth color in style image is transferred into human’s face. There are artifacts in...
Fig. 4: Image rescale with scale-invariant feature transform.

TABLE I: The runtime with different scales.

| scales | 3    | 6    | 9    | more than 1 h | more than 5 h |
|--------|------|------|------|---------------|---------------|
| run time | around 5 mins | around 15 mins | more than 1 hour | more than 5 hours |

Fig. 5: Style transfer comparison without/with saliency constrained. (a) content, (b) style, (c) the results without saliency map, (d) the results with saliency constrained.

D. Quantitative Evaluation.

Since there is no standard metrics for the quantitative comparisons of style transfer, the user evaluation is a popular method in style transfer quantitative comparisons [28], [34]. We performed a user study in which the users were presented with a style image, a content image, and stylised output images from the following nine methods: I (Johnson et al. [2]), II (Gatys et al. [1]), III (Li and Wand [4]), IV, which is our method, V (Ulyanov et al. [7]), VI (LFW [6]), VII (ResNet50), VIII (MGANs [5]), IX (cGANS [38] (iteration=200)).

Two-alternative forced choice (2AFC) is a subjective experience measurement approach of human through their pattern of choices [18], and it is simplistic and reliable. Our user study is developed by using the 2AFC.

In the first part of the user evaluation, a user answers the following question:

Task: Given two result images, which image do you prefer (for example which image combines content image and style image better)?

Four images (two rows and two columns) will be shown in each trial, one style image and one content image in the first row, and two stylized images in the second row, respectively. The order of the two stylized images in a trial are randomized. The user chooses the image (that she/he believes that it combines the style and content images in the first row better) from the two result images in the second row. The full image set which were shown in figures 6, and 5 include 7 sets (4 sets of human face and 6 sets other objects), 9 methods, so $10 \times 9 = 90$ test images. 60 persons were asked to participate the study. Their ages range from 18 to 50. So for each method, its result will be compared $10 \times 8 = 80$ with other results. The total click number of participants for each method is counted and analyzed. The results of different methods are shown in figure 10.

In figure 10 red lines are denoted as the mean. Blues lines are denoted as the quartiles, and black lines are denoted as extremes of the distributions. Next, a statistical method named
Fig. 6: Different objects style transfer comparison. (a) input, (b) Johnson et al. [2], (c) Gatys et al. [1], (d) Li and Wand [4], (e) Ulyanov et al. [7], (f) LFYW [6], (g) ResNet50 [25], (h) MGANs [5], (i) cGANS [38](iteration=100), (j) cGANS [38](iteration=200), (k) Ours.
Fig. 7: Saliency detection results.

Fig. 8: Style transfer comparison, (a) Johnson et al. [2], (b) Gatys et al. [1], (c) Li and Wand [4], (d) Ulyanov et al. [7], (e) LFYW [6], (f) ResNet50 [25], (g) MGANs [5], (h) cGANS [38] (iteration=100), (i) cGANS [38] (iteration=200), (j) Ours.

Fig. 9: Style transfer comparison, (a) Johnson et al. [2], (b) Gatys et al. [1], (c) Li and Wand [4], (d) Ulyanov et al. [7], (e) LFYW [6], (f) ResNet50 [25], (g) MGANs [5], (h) cGANS [38] (iteration=100), (i) cGANS [38] (iteration=200), (j) Ours.
analysis of variance (ANOVA) was used to test differences of the participants’ clicks in this user study. In order to better understand the results, the p-values which is used to show the level of marginal significance within a statistical hypothesis test are shown in table II. Usually the values less than 0.05 show statistically significant.

We can see from figure 10 and table II that the proposed approach has the highest mean score, our stylised images are preferred by the majority of the participants.

Then we also evaluate the stylization perception of several techniques by asking the participants to score images on a scale of 1 to 9 for each of the nine results, ranging from "bad" to "excellent" which is used in [34]. There are 10 different cases for each of the nine methods used for the following question:

- Task: Given the nine result images, which image better matches the target style? Then give a scale of 1 to 9 for each image.

Figure 11 summarizes the average scores of the study by different methods based on the evaluation of users.

In the second study, the average score of the test reflects the style similarity of the generated image to the style of the reference style image. Our proposed algorithm obtains the highest score, which demonstrates that the proposed method can transfer style effectively and preserve the information of the content image. The methods based on ResNet50 and cGANS [38](iteration=200) score lowest.

E. Failure cases and Limitations.

Inaccurate saliency detection may occur, which may generate new errors in patch matching. If objects in the input images are inaccurately detected, the faulty saliency detection parts will lead to wrong style transfer. Some examples of the effect of an inaccurate saliency detection are shown in figure 12. The red color belonging to the style train is transferred onto the hill in content image because of inaccurate saliency detection of style image. Figure 12(h) is more accurate than figure 12(e), so figure 12(i) which has less red color in the hill is better than figure 12(f). Despite this inaccuracy, the result remains considerably better than the figure 12(c) which is without saliency constrained.

F. Effect of Style Loss Weight

As shown in Eq 2, the energy loss function contains three weights $\alpha_1$, $\alpha_2$, and $\alpha_3$ for style loss function, saliency loss function and content loss function respectively. Based on the empirical results of published studies (e.g., [1], [2] and [4]) and our experiments, we found that the weights $\alpha_1 \in [10^{-5}, 2 \times 10^{-4}]$, $\alpha_2 \in [0, 200], \alpha_3 \in [0, 100]$ can produce satisfactory results.

In this investigation of the effect of the style loss weight, experiments are carried out by changing style loss weight $\alpha_1$ while keeping $\alpha_3 = 20$ and $\alpha_2 = 50$ fixed. Figure 13 shows the results with modifying $\alpha_1$ from $\alpha_1 = 10^{-5}$ to $\alpha_1 = 2 \times 10^{-4}$.

When $\alpha_1$ is smaller, the results in matched patches have stronger content consistency. On the other hand, setting $\alpha_3$ to a larger value, the results have greater style structure information.
TABLE II: The $p$-value of the ANOVA test.

| id | Method          | $p$-value |
|----|-----------------|-----------|
| I  | Johnson et al.  | 0.0122    |
| II | Gatys et al.    | 0.0019    |
| III| Li and Wand     | 0.0036    |
| IV | Ulyanov et al.  | 0.0866    |
| V  | LFYW            | 0.0047    |
| VI | ResNet50        | 0.0030    |
| VII| MGANs           | 0.0050    |
| VIII| cGANS     | 4.7341e-8|

G. Effect of Saliency Weight

Next more experiments are carried out by changing salient loss weight $\alpha_2$ while keeping $\alpha_3 = 20$ and $\alpha_1 = 10^{-4}$ fixed. The results by modifying $\alpha_2$ from $\alpha_2 = 10$ to $\alpha_2 = 150$ are shown as figure 14.

It is observed from figure 14 that, when $\alpha_2$ is smaller, more green color in style image is transferred into human’s face. On the other hand, setting $\alpha_2$ to a larger value, the results have more errors, for example, human’s right eye is disappeared.

H. Effect of Content Loss Weight

At last, more experiments are carried out where the content loss weight $\alpha_3$ is modified while $\alpha_1 = 10^{-4}$ and $\alpha_2 = 50$ are fixed. The results with modifying $\alpha_3$ from $\alpha_3 = 10$ to $\alpha_3 = 60$ are demonstrated in figure 15. When $\alpha_3$ is too small, the results in matched patches have strong style consistency. On the other hand, setting $\alpha_3$ to a larger value, the results have sufficient content structure information.

It is easily observed from figure 15 that, when $\alpha_3$ is smaller, more red color in style image is transferred into human’s face and dog body. And the results have more errors. On the other hand, setting $\alpha_3$ to a larger value, the results in matched patches have stronger content consistency, and very fewer color in style image is transferred.

VII. Conclusions

This paper investigates the advantage of incorporating saliency extraction in image style transfer. In the proposed method, we modify the traditional DCNN model by concatenating a saliency detection channel with regular filters, and then propose a novel loss function combining saliency constraint. The new loss function consists of three parts: style loss, content loss and saliency loss. The result is updated by adding the gradient of saliency constraint in iterations. The experiments results on different kind of images show that the proposed method can generate better results than recent approaches considered in the paper. The saliency maps of two source images (content and style image) can help find the correct matching. The correctness and accuracy of the salient maps are critical. They can achieve background texture and object texture style transfer separately, and prevent them from contaminating each other. The proposed method remains having some limitations, for example, if the saliency detection results are inaccurate, it would generate new errors in the patches matching and lead to wrong style transfer. Further study will investigate more accurate saliency constraint method and systematically fine tune the weight of each part of the loss function to improve results.
This work was supported by National Natural Science Foundation of China (61503218, 61772179), Science and Technology Plan Project of Hunan Province (2016TP1020). We would like to thank NVIDIA for the GPU donation.

REFERENCES

[1] Gatys Leon A., Ecker, Alexander S and Bethge Matthias,Image style transfer using convolutional neural networks[C], Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016.pp:2414-2423.

[2] Johnson Justin, Alahi Alexandre and Fei-Fei Li,Perceptual losses for real-time style transfer and super-resolution[C],European Conference on Computer Vision,2016.pp:694-711.

[3] Gatys Leon A., Ecker Alexander S., Bethge Matthias, Hertzmann Aaron and Shechtman Eli,Controlling Perceptual Factors in Neural Style Transfer[C],2017 IEEE Conference on Computer Vision and Pattern Recognition,2017.pp:3985-3993.

[4] Li Chuan and Wand Michael,Combining markov random fields and convolutional neural networks for image synthesis[C],Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016.pp:2479-2486.

[5] Li Chuan and Wand Michael,Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks[C],European Conference on Computer Vision,2016.pp:702-716.

[6] Li Yijun, Fang Chen, Yang Jimei, Wang Zhaowen, Lu Xin and Yang Ming-Hsuan,Universal style transfer via feature transforms[C],Advances in Neural Information Processing Systems,2017.pp:386-396.

[7] Ulyanov Dmitry, Vedaldi Andrea and Lempsky Victor,Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis[C],Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017.pp:1-6.

[8] Deng Hao, Han Chu, Cai Hongmin, Han Guoqiang and He Shengfeng,Spatially-Invariant Style-Codes Controlled Makeup Transfer[C],Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2021.pp:6549-6557.

[9] Chen Yang, Lai Yu-Kun and Liu Yong-Jing,CaroonGAN: Generative Adversarial Networks for Photo Cartoonization[C],Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018.pp:9465-9474.

[10] Xia Xide, Xue Tianfan, Lai Wei-Sheng, Sun Zheng, Chang Abby, Kulis Brian and Chen Jiawen,Real-Time Localized Photorealistic Video Style Transfer[C],Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,2021.1089-1098.

[11] Chen Chaqi, Zheng Zheng, Huang Yang, Ding Xinghao and Yu Yizhou,LI3Net: Implicit Instance-Invariant Network for Adapting One-Stage Object Detectors,Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C],IEEE Conference on Computer Vision and Pattern Recognition,2021.pp:12576-12585.

[12] Deng Xin, Wang Hao, Xu Mai, Guo Yichen, Song Yanghang and Yang L.L.,LAU-Net: Latitudinal Adaptive Upscaling Network for Omnidirectional Image Super-Resolution[J], The IEEE Conference on Computer Vision and Pattern Recognition,2021.pp:9189-9198.

[13] Cheng Guang, Hu Zhiwei, Zhang Lihe and Lu Huchuan, Encoder Fusion Network With Co-Attention Embedding for Referring Image Segmentation[C],Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2021.pp:15506-15515.

[14] Gorji Siavash and Clark James J.,Going From Image to Video Saliency: Augmenting Image Salience With Dynamic Attentional Push[C],The IEEE Conference on Computer Vision and Pattern Recognition,2018.pp:7501-7511.

[15] Jalwana Mohammad A. K., Akhtar Naveed, Bennamoun Mohammed and Mian Ajmal, CAMERAS: Enhanced Resolution and Sanity Preserving Class Activation Mapping for Image Salience[C],The IEEE Conference on Computer Vision and Pattern Recognition,2021.pp:16327-16336.

[16] Petriuk Vitali, Jain Rajiv, Manjunatha Varun, Morariu Vlad L, Mehra Ashutosh, Ordzone Vicente and Saenko Kate,Black-Box Explanation of Object Detectors via Saliency Maps[C],The IEEE Conference on Computer Vision and Pattern Recognition,2021.pp:11443-11452.

[17] Lee Seungho, Lee Minhyun, Lee Jonguk and Shim Hyunjung, Road Is Not a Train: Saliency As Pseudo-Pixel Supervision for Weakened Supervised Semantic Segmentation[C],The IEEE Conference on Computer Vision and Pattern Recognition,2021.pp:5495-5505.

[18] Arzoumanian Dorothée, de Kergang Mathilde and de Cheveigné, Alain,A sliding two-alternative forced-choice paradigm for pitch discrimination[J],The Journal of the Acoustical Society of America,2017,142(1):167-172.

[19] Xiaochun Cao, Zhiqiang Wang, Bao Zhaohua and Wei Feng,Self-adaptively weighted co-saliency detection via rank constraint[J],IEEE Transactions on Image Processing,2014,23(9):4175-4186.

[20] Goodfellow Ian, Pouget-Abadie Jean, Mirza Mehdi, Xu Bing, Wardefarley David, Sherjil, Courville Aaron and Bengio Yoshua,Generative Adversarial Nets[C],Neural Information Processing Systems,2014.pp:2672-2680.

[21] Ren Zhiyuan, Gao Shenghua, Chia Liang Tien and Tsang Wai Hung,Region-Based Saliency Detection and Its Application in Object Recognition[J],IEEE Transactions on Circuits and Systems for Video Technology,2014,24(5):761-771.

[22] Li Xi, ZhaoLiming, Wei Lina, Yang Ming-Hsuan, Wu Fei, Zhuang Yueting, Ling Haibin, Wang Jingdong, Deepsaliency: Multi-task deep neural network model for salient object detection[J],IEEE Transactions on Multimedia,2016,25(8):3919-3930.

[23] Gatys Leon, Ecker Alexander S and Bethge Matthias,Texture synthesis using convolutional neural networks[C],Advances in Neural Information Processing Systems,2015,28:321-329.

[24] Ulyanov Dmitry, Lededev Vadim, Vedaldi Andrea and Lempskiy Victor,Texture networks: Feed-forward synthesis of textures and stylized images[C],Int. Conf. on Machine Learning,2016.pp:1394-1395.

[25] He Kaiming, Zhang Xiangyu, Ren Shaoqing and Sun Jian, Deep Residual Learning for Image Recognition[C], The IEEE Conference on Computer Vision and Pattern Recognition,2016.pp:770-778.

[26] Chandran Prashanth, Zoss Gaspard, Gotardo Paulo, Gross Markus and Bradley Derek,Adaptive Convolutions for Structure-Aware Style Transfer[C],The IEEE Conference on Computer Vision and Pattern Recognition,2021.pp:7972-7981.

[27] Hou Qibin, Cheng Ming-Ming, Hu Xiaowei, Borji Ali, Tu Zhouwen and Torr Philip H. S., Deeply Supervised Salient Object Detection With Short Connections[C],The IEEE Conference on Computer Vision and Pattern Recognition,2017,36(4):1063-691.

[28] Zhang Dingwen, Fu Huazhu, Han Junwei, Borji Ali, Li Xuejian, Yu Tao, Yuan, Ren Jian, Xie Xinsong, Liu Weidong, Liu Yong-Jin and Wang Jun,Attention-aware Multi-style Tone Transfer[C],Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2019.pp:1467-1475.

[29] Zhi Liu, Zou Wenbin, Li Lina, Shen Lifan and Meur, Olivier Le,Co-Saliency Detection Based on Hierarchical Segmentation[J],IEEE Signal Processing Letters,2014,21(1):88-92.

[30] H.Fu, X.Cao and Z.Tu,Cluster-Based Co-Saliency Detection[J], IEEE Transactions on Image Processing,2013,22(10):3766-3778.

[31] Cheng Ming-Ming, Mitra Niloy J, Huang Xiaolei, Torr Philip HS and Hu Shi-Min, Global contrast based salient region detection[J],IEEE transactions on pattern analysis and machine intelligence,2014,37(3):569-582.

[32] Zhang Dingwen, Fu Huazhu, Han Junwei, Borji Ali and Li Xuejian,A Review of Co-saliency Detection Technique: Fundamentals, Applications, and Challenges[J], ACM Transactions on Intelligent Systems and Technology,2018,9(4):38-41.

[33] Li Bo, Sun Zhengxing, Tang Lv, Sun Yunhan and Shi Jinlong,Detecting robust co-saliency with recurrent co-attention neural network[C],In Proceedings of the 28th International Joint Conference on Artificial Intelligence,2018,pp:818-825.

[34] Jiachuan Sheng, Caifeng Song, Jun Wang and Yanghuan Convolutional Neural Network Style Transfer Towards Chinese Painting[J],IEEE Access,2019,7:163719-163728.

[35] Cheng Jiaxin, Jaiswal Ayush, Wu Yue, Natarajan Pradeep and Nataraajan Prem,Style-Aware Normalized Loss for Improving Arbitrary Style Transfer[C],The IEEE Conference on Computer Vision and Pattern Recognition,2021.pp:134-143.

[36] Radford Alec, Metz Luke and Chintala Soumith,Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[C], International Conference on Learning Representations,2016.pp:97-108.

[37] Kotoenko Dmitry, Wright Matthias, Heimbrecht Arthur and Ommer Bjorn, Rethinking Style Transfer: From Pixels to Parameterized International[C], IEEE Conference on Computer Vision,2021.pp:12196-12205.

[38] Isola Phillip, Zhu Jun-Yan, Zhou Tinghui and Efros Alexei A,Image-To-Image Translation With Conditional Adversarial Networks[C],The IEEE Conference on Computer Vision and Pattern Recognition,2017.pp:1125-1134.

[39] Lowe David G,Distinctive image features from scale-invariant keypoints[J],International journal of computer vision,2004,60(2):91-110.
[40] Etezadifar Pouriya and Farsi Hassan, A new sample consensus based on sparse coding for improved matching of SIFT features on remote sensing images[J], IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(8): 5254-5263.

[41] Zhao Hui-Huang, Rosin Paul L, Lai Yu-Kun, Lin Mu-Gang and Liu Qin-Yun, Image neural style transfer with global and local optimization fusion[J], IEEE Access, 2019, 7: 85573-85580.

[42] Liu Yuan, Improved generative adversarial network and its application in image oil painting style transfer[J], Image and Vision Computing, 2021, 105: 104087.

[43] Virtusio John Jethro, Ople Jose Jaena Marie, Tan Daniel Stanley, Tanveer M., Kumar Neeraj and Hua Kai-Lung, Neural Style Palette: a multimodal and interactive style transfer from a single style image[J], IEEE Transactions on Multimedia, 2021, doi:10.1109/TMM.2021.3087026

[44] Qiao Yingxu, Cui Jiabao, Huang Fuxian, Liu Hongmin, Bao Cuizhu and Li Xi, Efficient Style-Corpus Constrained Learning for Photorealistic Style Transfer[J], IEEE Transactions on Image Processing, 2021, 30(8): 3154-3166.

[45] Wang Pei, Li Yijun and Vasconcelos Nuno, Rethinking and Improving the Robustness of Image Style Transfer[C], The IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp: 124-133.

Hui-Huang Zhao received his Ph.D. degree in 2010 from XiDian University. He was a Sponsored Researcher in the School of Computer Science and Informatics, Cardiff University. Now he is a full professor in the College of Computer Science and Technology, Hengyang Normal University. His main research interests include, Compressive Sensing, Machine Learning, and Image Processing.

Yao-Nan Wang received his Ph.D. degree in 1995 from Hunan University. Now he is a Professor, Academician of Chinese Academy of Engineering, Ph.D. supervisor and the dean of College of Electrical and Information Engineering of Hunan University. His main research interests include intelligent control and pattern recognition.

Yu-hua Li received the Ph.D. degree in general engineering from the University of Leicester, U.K. He successively worked with the University of Manchester, Ulster University and the University of Salford, from 2003 to 2017. He is currently a Senior Lecturer with the School of Computer Science and Informatics, Cardiff University, U.K. His current research interests include machine learning, data science, neural networks, and condition monitoring.