IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 15, 2018, accepted December 29, 2018, date of publication January 9, 2019, date of current version April 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2891576

Image Neural Style Transfer With Preserving the Salient Regions

YIJUN LIU[1], ZUOTENG XU[1], WUJIAN YE[1], ZIWEN ZHANG[1], SHAOWEI WENG[1], CHIN-CHEN CHANG[2], AND HUAJIN TANG[3]

[1]School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China
[2]Department of Information Engineering and Computer Science, Feng Chia University, Taichung 4072, Taiwan
[3]Neuromorphic Computing Research Center, Sichuan University, Chengdu 610044, China

Corresponding author: Wujian Ye (yewjian@gdut.edu.cn)

This work was supported in part by the Department of Science and Technology and the Department of Industry and Information Technology of Guangdong Province ([2017] 74) under Grant 2016B090903001, Grant 2016B090904001, Grant 2016B090918126, and Grant 2015B090901060, in part by the Peng Cheng Laboratory, and in part by the Guangdong University of Technology under Grant 220413548.

ABSTRACT Neural style transfer recently has become one of the most popular topics in academic research and industrial application. The existing methods can generate synthetic images by transferring different styles of some images to another given content images, but they mainly focus on learning low-level features of images with losses of content and style, leading to greatly alter the salient information of content images in the semantic level. In this paper, an improved scheme is proposed to keep the salient regions of the transferred image the same with that of content image. By adding the region loss calculated from a localization network, the synthetic image can almost keep the main salient regions consistent with that of original content image, which helps for saliency-based tasks such as object localization and classification. In addition, the transferred effect is more natural and attractive, avoiding simple texture overlay of the style image. Furthermore, our scheme can also extend to remain other semantic information (such as shape, edge, and color) of the image with the corresponding estimation networks.

INDEX TERMS Neural style transfer, image semantic, salient region, localization network, region loss.

I. INTRODUCTION

Image style transfer is the process of applying the style of an image to another content image. By applying style transfer techniques, an image can be re-drawn in a particular style automatically without a well-trained artist. Thus, lots of time also can be saved [1]. Many methods are proposed to automatically turn images into synthetic artworks. The well-developed Non-photorealistic Rendering (NPR) techniques, belonging to computer art category, simulate artistic rendering style by designing algorithms and creating mathematical models with the help of computer techniques [2]–[6]. However, the NPR algorithms highly rely on some particular artistic styles (like ink painting, oil paintings, sketches, animations and so on) they simulate, so that they cannot be easily extended to generate stylized results for other artistic styles [2].

To solve the limitations of the above traditional methods, Gatys *et al.* [7], [8] firstly apply Convolutional Neural Network (CNN) to transfer famous painting styles to some natural images. Their algorithm works based on their observation that the representations of image's content and style can be derived and separable from CNN simply. This method successfully produces fantastic stylized images with the appearance of a given artwork, but the transferred style tends to simple texture overlay [8]. Several approaches are proposed to improve or expand Gatys's model. For example, Johnson *et al.* [9] introduce a fast methods by training a feed-forward CNN with perceptual loss functions, which is orders of magnitude more efficient for stylizing input content image.

Existing methods typically generate transferred images based on pre-trained neural networks with style loss and content loss directly. However, they neglect the semantics of the content image, leading to altering the major informations of salient regions, edges, color and so on. For example, when we consider a good synthetic image, it should preserve its saliency information almost the same with that of original content image, but the salient region of synthetic images generated by existing methods is greatly altered, which is dissimilar to origin content images.

Based on the above observation, an improved neural style transfer scheme is proposed to preserve the salient regions of the transferred image the same as the original content image. By adding a region loss generated by a localization network, the final transferred image can maintain most of the salient region information consistent with that of content image, and its paint effect seems to have more natural and attractive. The proposed scheme can not only keep the content avoiding simple texture overlay of the style image, but also help for the saliency-based tasks of object localization and classification. Furthermore, it can also be extended to maintain other detailed semantic information (like edge, shape, color, depth) of content or style images by applying other estimation networks in the process of image style transfer.

The following is organized with the rest paper. In Section 2, we analyze related works about techniques of image style transfer, and introduce the saliency-based applications. In Section 3, we propose an improved neural style transferring system to generate synthetic images with preserving the original salient regions. In Section 4, we present the performance evaluation of our proposed scheme. Finally, we conclude our research and point out the future works in Section 5.

II. RELATED WORKS

A. IMAGE STYLE TRANSFER

Image style transfer has been an emerging technique during the past decade, which is an important task for transferring the style of a source image to anther target image. This technique is useful for synthesizing or transferring the derivative works of a particular artist, specific painting and photos [10]. There are a lot of research to explore how to combine different styles and contents of images for automatically generating some new artworks [1]. We categorize existing methods into traditional methods and deep learning-based methods, and introduce them as follows [11].

1) TRADITIONAL METHODS

Image style transfer is to migrate a style from an image to another, and is closely related to texture synthesis, which belongs to the NPR techniques [12]. Generally, traditional methods design particular algorithms and create mathematical models with the help of the computer techniques, according to the analyzing and obtaining suitable feature representations of certain styles [11].

Efros and Leung [13] propose a non-parametric methods for texture synthesis, which tries to preserve most of the local structure. Portilla and Simoncelli [3] present a universal statistical model for texture images in the context of an overcomplete complex wavelet transform.

However, traditional methods support only limited artistic styles. For the new styles, plenty of time is needed to analyze their patterns with much human knowledge and experiences [14]. The key limitation of these methods is that they only use low-level features of images and may not be able to capture content and style effectively [11].

2) DEEP LEARNING METHODS

Recently, CNNs are widely applied in plenty of fields due to its strong ability of feature learning, such as computer vision, speech recognition, text processing, medical, finance and advertising [15]–[17]. By visualizing and analyzing the CNN networks, Gatys et al. find that the representations of image's content and style can be separable from different layers simply, and they firstly try to transfer famous painting styles to natural images by obtaining the image representations derived from CNN [7]. Based on above observation, they further propose a iteration-based neural style algorithm to recombine the content of a given photograph and the style of well-known artwork, which successfully generates fantastic given artwork-stylized images [8]; but their algorithm does not work in real-time, and tends to transfer repetitive styles; furthermore, it often fails to transfer the complex pattern [18].

To extend the approaches in [7] and [8], Johnson *et al.* [9] train an equivalent feed-forward generator network for image transformation tasks using perceptual loss functions. Compared with the approach proposed by Gatys *et al.* [8], their network gives similar qualitative results but is three orders of magnitude faster. Chen *et al.* [12] propose a novel explicit representation for style and content, which can be well decoupled by our network. The decoupling allows faster training (for multiple styles, and new styles), and enables new interesting style fusion effects, like linear and region-specific style transfer. And they present a new interpretation to neutral style transfer which may inspire other understandings for image reconstruction and restoration.

The problem of multi-style transfer in real-time using a single network has been addressed by Zhang *et al.* [19]. They tackle the technical difficulties of existing approaches by introducing a novel Inspiration Layer, which explicitly matches the target styles at run time. They demonstrate that the Inspiration Layer embedded in the proposed Multi-style Generative Network enables 20 styles transfer without losing quality in real-time. However, dealing with unknown style is still an unsolved problem for feed-forward approaches [19].

The existing methods can generate a synthetic image by transferring different styles of images to a given original content image, but they only focus on learning low-level features of images with content and style losses, leading to neglect the semantic detailed information of original content image or style image, such as the salient location, depth space, shape, edge, color and so on. Several recent works are trying to preserve the missing semantic of content images.

Gatys *et al.* [20] and Yin [21] extend the existing method proposed in [9]. Gatys *et al.* [20], [22] introduce control over spatial location, color information and across spatial scale; and they demonstrate that their method allows high-resolution controlled stylization and helps to alleviate common failure cases such as applying ground textures to sky regions. Yin [21] propose a content-aware algorithm to achieve the goal of synthesizing a high resolution painting in more realism.

To address the problem that style representation in [8] is invariant to the space configuration of the style image, Nikulin and Novak [23] propose a new style representation called spatial style, which captures less style details and focuses on more spatial configuration. Then, they also propose several helpful modifications to improving the quality of image stylization including activation shift, augmenting the style representation and geometric weighting scheme [18]. Risser *et al.* [24] improve the generated images in stability and quality by imposing histogram losses, which better constrains the dispersion of the texture statistics. The results show improvements by automating the parameter tuning, and in artistic controls.

Liu *et al.* [11] integrate depth preservation as additional loss, preserving overall image layout while performing style transfer based on the fast neural network proposed by Johnson *et al.* [9]. Luan *et al.* [25] faithfully transfer style from a reference image for a wide variety of image content based on deep learning. They use the Matting Laplacian to constrain the transformation from input to output to be locally affine in color space. Semantic segmentation further drives more meaningful style transfer yielding satisfying photorealistic results in a broad variety of scenarios, including transfer of the time of day, weather, season, and artistic edits.

Furthermore, Castillo *et al.* [10] consider targeted style transfer, in which the style of a template image is used to alter only part of a target image. For example, an artist may wish to alter the style of only one particular object in a target image without altering the object's general morphology or surroundings. They present a method for targeted style transfer that simultaneously segments and stylizes single objects selected by the user. This method uses a Markov random field model to smooth and anti-alias outline pixels near object boundaries, so that stylized objects naturally blend into their surroundings.

B. VISUAL SALIENCY-BASED APPLICATIONS

Visual saliency analysis is about to detect salient regions, which related to uniqueness, rarity and surprise of a scene, characterized by primitive features like color, texture and shape [26]. The salient regions are the areas (may be some objects, pixes, persons and combination of them) that are the most able to indicate the content of the scene and are the most attractive to the human [27]. After Koch and Ullman [28] firstly propose the concept of saliency map to point out the visual dominant regions of images, a variety of saliency maps-based methods are emerged in different applications of computer vision, like object localization, classification, segmentation, tracking and so on [29]–[31]. We introduce them briefly as follows.

1) OBJECT LOCALIZATION

Object localization is an important task for the automatic understanding of images. Simonyan *et al.* [32] compute a specific-class saliency map of image, highlighting the areas of the given image, discriminative with respect to the given classes, and thus can be used for object localization. Oquab *et al.* [33] apply global max pooling to localize a point on objects. However, their localization is limited to a point lying in the boundary of the object rather than determining the full extent of the object. Zhou *et al.* [34] generate Class Activation Maps (CAMs) for CNNs with global average pooling (GAP), which enables classification-trained CNNs to learn to perform object localization without using any bounding box annotations. The generated CAMs visualize the predicted class scores on any given image, highlighting the discriminative object parts detected by the CNNs. As shown in Fig. 1, their method localizes class-specific discriminative image regions in a single forward-pass, such as the toothbrush for brushing teeth and the chain-saw for cutting trees.

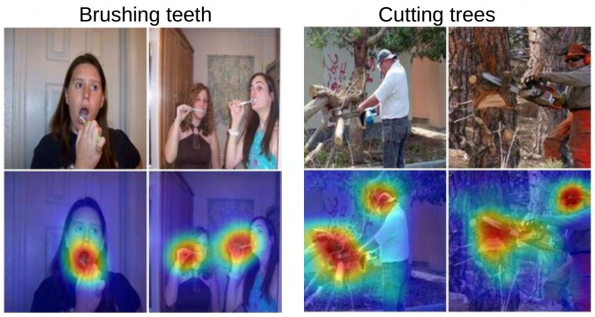


FIGURE 1. CAMs of GAP layer-based CNN [34].

2) RECOGNITION

Object recognition aims to find the existence of a certain object in an image [35]. The idea of using saliency is that not all parts of an image provide useful information, if we attend only to the relevant parts, we can recognize the image more quickly with less resources [29]. Kanan and Cottrell [36] propose an approach based upon two facets of the visual system: sparse visual features that capture the statistical regularities in natural scenes and sequential fixation-based visual attention. In particular, saliency maps are used as interest point operators. Their approach works well since it employs a non-parametric exemplar-based classifier. Ren *et al.* [37] apply saliency maps to better encode image features for object recognition. Since the objects usually correspond to salient regions, and these regions usually play more important roles for object recognition than the background, they incorporate a saliency map into sparse coding-based image representation.

3) SEGMENTATION

Scene segmentation is an important step towards full scene understanding. The saliency is considered as a good cue for salient object segmentation [29]. Simonyan *et al.* [32] use the saliency map to initialize Graph Cut-based object segmentation without the need to train dedicated segmentation or detection models. Cheng *et al.* use the computed saliency map to assist in automatic salient object segmentation. This immediately enables automatic analysis of large internet image repositories [38].

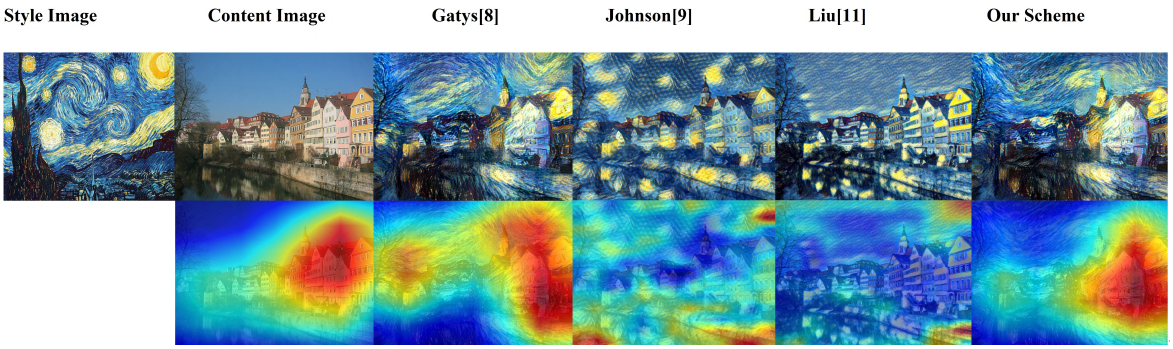


FIGURE 2. A CAM of transferred images generated by different methods.

4) TRACKING

The main ability of saliency analysis is to tackle different situations when an object appears in different forms and with different background [29]. Li and Ngan propose a facial saliency map(FSM)-based three-stages method for human tracking. At first, they generates a saliency map of the input video frame by using face tracking as the initial step for face segmentation in the subsequent frames. Next, a geometric model and an eye-map built from chrominance components are employed to localize the face region according to the saliency map. The final stage involves the adaptive boundary correction and the final face contour extraction. The experimental results show that their method is able to segment the face area quite effectively [39]. Frintrop and Kessel [40] propose a mobile platform-based cognitive approach for tracking visual object. Based on a biologically motivated attention system, this approach can detect regions of interest in images based on concepts of the human visual system. Specially, the attention system builds a top-down, target-related saliency map by searching for the target features of subsequent images, which enables to concentrate on the most relevant features with this object without knowing anything about a particular object model or scene in advance.

III. PROPOSED SCHEME

A. ANALYSIS OF EXISTING NEURAL STYLE TRANSFER

In the process for neural style transfer, a good synthetic image should keep its saliency map almost consistent with that of original content image. After stylization, it is acceptable to weaken or enhance the saliency map of original images, but its integrity should be retained. In this paper, we used the CAM proposed by Zhou *et al.* [34] as saliency map to highlight the class-specific salient regions of image.

According to our observation, the CAMs of synthetic images are altered greatly as shown in Fig. 2 since most of existing methods just combining the content and style directly in the low-level features, without considering the major saliency semantic of the content image intensively. The last column shows the CAM of transferred image generated by our scheme, the discriminative saliency map of content image can still be preserved mostly .

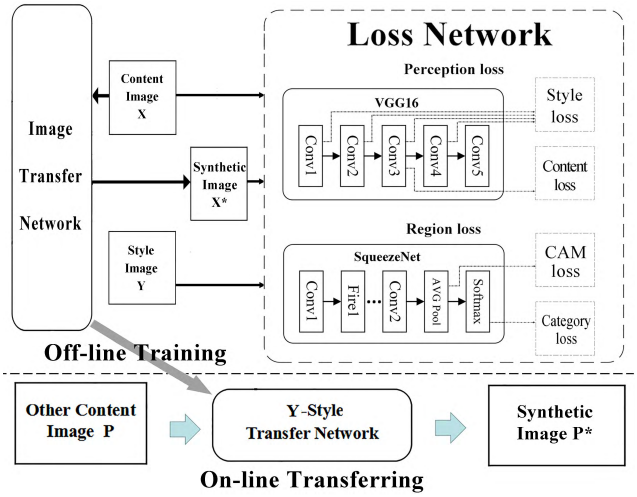


FIGURE 3. Improved neural style transferring system.

B. PROPOSED SYSTEM

In order to preserve the salient regions of transferred image the same as original content image, an improved neural style transferring system is proposed by combining the fast neural style model in [9] and the class activation mapping techniques in [34]. As shown in Fig. 3, the system consists of two stages, off-line training and on-line transferring.

In the off-line training stage, there are mainly an image transfer network and a loss network. The style image Y needs to be specified in advance, while the content image X is constantly obtained from the image datasets for each iteration, which enables the image transfer network to process images with different contents and enhance the robustness of the system. In each iteration, image transfer network is trained for transferring the content image X to the synthetic image X* with Y-style. And the loss network composed of a perception network and a localization network; a perception loss (containing content loss and style loss) is obtained by the perception network, and a region loss (including CAM loss and category loss) is computed by introducing the localization network; then, according to the feedback of losses, the image transfer network adjusts its weights and readies for

the next iteration. Finally, a Y-style transfer network model is generated.

In the on-line transferring stage, when image P needs to integrate Y-style, P is pre-processed into suitable size and input to the Y-style transfer network model, which can quickly generate a transferred image P* with Y-style. The details of our system will be described in below.

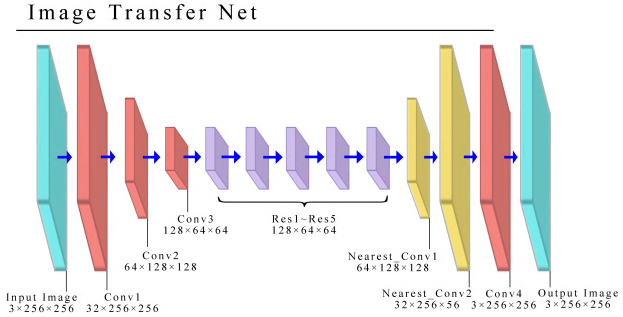


FIGURE 4. Image transfer network.

C. IMAGE TRANSFER NETWORK

The image transfer network is essentially a deep residual neural network, which can be trained to quickly stylize the content image with a specific image style. As shown in Fig. 4, this network is mainly made up of three convolution layers, five residual blocks [41], two nearest neighbor interpolation layers and the last output layer. The last layer uses tanh function to ensure that the output image has a range of 0 to 255 pixels, the related formula is shown in (1), X is the output of tanh function ranging from -1 to 1, Y is the output ranging from 0 to 255; while the other layers uses the Relu as the activation function [9]. The training of this network is the process of iterative optimization by calculating the gradient of network continuously until the total loss tends to be stable and convergent obviously.

$$\mathrm{Y} = (\mathrm{X} + 1) \times 255/2 \tag{1}$$

D. PERCEPTION NETWORK

The perception network is implemented with VGG-16 model proposed by Simonyan and Zisserman [42], which is a pre-trained model using ImageNet [43] for object classification, and its architecture is as shown in Fig. 5. Instead of accurately computing each pixel value of image, this network calculates feature presentation of similar style and content, then generates the losses by calculating the distance between the corresponding features of target image and synthesize image. Content loss is obtained from Relu3_3 layer, and the style loss is got from Relu1_2, Relu2_2, Relu3_3 and Relu4_3 layers of VGG-16.

1) CONTENT LOSS

The content loss is obtained by computing the L2 distance between the feature maps in the VGG-16 of content image and generated image. Here, the feature maps of input image

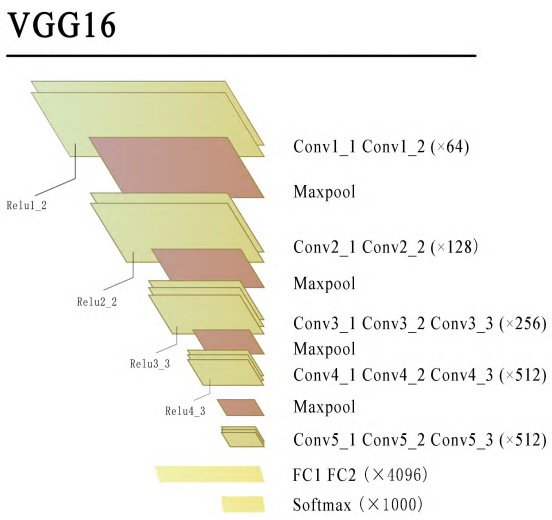


FIGURE 5. VGG-16 network.

are got from the Relu3_3 layer of VGG-16. Let the feature maps in Relu3_3 layer of content image and transferred image be $\varphi_{3\_3}(\mathrm{X})$ and $\varphi_{3\_3}(\mathrm{X}^*)$ respectively, and feature maps can be expanded in N dimensional vector, then the definition of content loss is (2).

$$loss_{cont}(X, X^*) = \frac{1}{N} \left\| \varphi_{3\_3}(\mathrm{X}^*) - \varphi_{3\_3}(\mathrm{X}) \right\|_2^2 \tag{2}$$

2) STYLE LOSS

The style loss is obtained by calculating the distance of the Gram matrices of feature maps in VGG-16 between the style image and the transferred image. Here, the Gram matrix can extract the style information of the original image such as color, texture, edge and so on. Let $\varphi_j(\mathrm{Y})$ be the feature maps in j-th layer, then $\varphi_j(\mathrm{Y})$ is transferred into a shape of $\mathrm{C} \times (\mathrm{H} \times \mathrm{W})$, C is the number of feature maps, H and W are the height and width of each feature map. Then the Gram matrix in j-th layer of VGG-16 for an input image is defined as (3).

$$\mathrm{G}_j(Y) = \frac{\varphi_j(\mathrm{Y}) * \varphi_j(\mathrm{Y})^T}{C \times H \times W} \tag{3}$$

In this paper, the feature maps in layers of Relu1_2, Relu2_2, Relu3_3 and Relu4_3 are chosen to compute the L2 distance of Gram matrices between style image and generated image. Let $\mathrm{G}_j(Y)$ and $\mathrm{G}_j(X^*)$ as the Gram matrices in j^{th} layer of style image and transferred image respectively, then the definition of style loss is (4).

$$loss_{style}(Y, X^*) = \sum_j \left\| \mathrm{G}_j(X^*) - \mathrm{G}_j(Y) \right\|_2^2 \quad \mathrm{j} \in \{\mathrm{Relu1\_2, Relu2\_2, Relu3\_3, Relu4\_3}\} \tag{4}$$

3) PERCEPTION LOSS

The final perception loss is obtained by combining content loss with style loss together. And its formula as shown in (5),

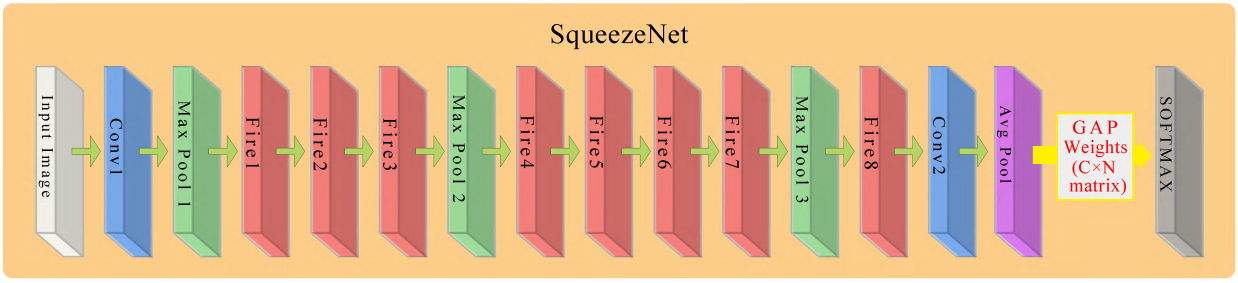


FIGURE 6. Network architecture of SqueezeNet.

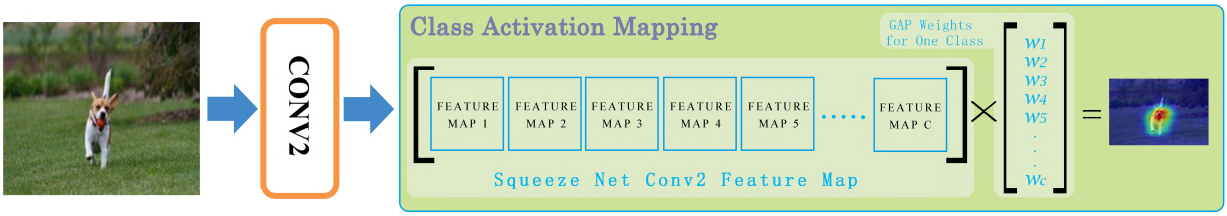


FIGURE 7. The process of generating CAM by using SqueezeNet.

and μ_1 and μ_2 are corresponding factors, generally set into suitable values on experience (default values: μ_1 is 1 and μ_2 is 20).

$$\text{loss}_{perc}\left(X^*, (\text{X}, \text{Y})\right) = \mu_1 \text{loss}_{\text{cont}}\left(X, X^*\right) + \mu_2 \text{loss}_{\text{style}}\left(Y, X^*\right) \quad (5)$$

E. LOCALIZATION NETWORK

The localization network is implemented using Squeeze-Net, which is a lightweight image classification network. We follow the work of Zhou *et al.* [34] to model the process to produce the saliency map by utilizing global average pooling (GAP) in our localization network. We sum the feature maps of last convolutional layer with weights to generate the saliency map for each image. SqueezeNet network is also a pre-trained network for classifying object [44]. By introducing the Fire Module components, its classification accuracy is comparable to AlexNet and the number of model parameters is 461 times less than AlexNet. The Fig. 6 shows its network structure. The proposed CAM loss and category loss are obtained by using the GAP layers and the classification results of SqueezeNet.

1) CAM LOSS

We define a CAM loss for retaining the original salient regions of transferred image. A class activation map (CAM) of an image for a particular category indicates the salient regions of images used by the CNN to identify that category. According to the GAP technique proposed by Zhou *et al.* [34], the common method for generating a CAM map is to do simple modification on CNN by adding a GAP layer, or to directly use some CNN networks already having the GAP layer, such as SqueezeNet, ResNet18. Then let the feature maps of the convolution layer which are most near the GAP layer to do the weighted sum operation with a neuron's weights of Softmax layer. Thus, a CAM of input image can be obtained. We can get a CAM image for each classified result, which highlights the most salient regions, and reflects the related semantic information with corresponding class.

The process for generating a CAM image by SqueezeNet is shown in Fig. 7. When we input the content image X into the SqueezeNet network, the feature maps of X in Conv2 layer is $\varphi_{conv2}(X)$ with size C × (H × W), C is the channel number, H and W is height and width of each feature map.

And we can get a C-dimension vector by inputting the feature maps into GAP layer. Then we operate full connection between the C-dimension vector and the neurons of Softmax layer, the number of whole weights is C × N, N is the number of classification categories. Here, we take Top-1 classification category to generate the CAM of input image, which can ensure the CAM of transferred images as accurately as possible reflects the original content image. Let W_n be the weights of the Top-1 classification category neuron, which is a C-dimension vector, then the calculation of the CAM is shown in formula (6).

$$\text{M}(\text{X}) = W_n * \varphi_{conv2}(X) \quad (6)$$

Let CAMs of generated image and content images be $M(X^*)$ and $M(X)$ respectively, we can define a CAM loss as (7), H and W is the height and width of a CAM map.

$$\text{loss}_{cam}\left(\text{X}, X^*\right) = \frac{1}{HW}\left\|M\left(X^*\right) - M(X)\right\|_2^2 \quad (7)$$

2) CATEGORY LOSS

A Category Loss is defined for keeping the Top-1 classification results of transferred images the same with that of content image. The CAMs of given images are different

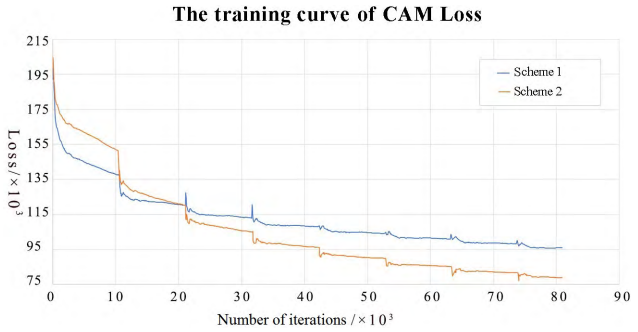


FIGURE 8. Comparing the CAM losses of two schemes in training stage.

for the different classification categories, leading to having diverse salient regions. In this paper, we generate the CAM according to the Top-1 classification category. During the training process, we observe that the Top-1 classification category of transferred image changes largely and the CAM loss is difficult to converge when we only use CAM loss to operate the gradient calculation. Thus, the CAM of final transferred images cannot correctly emphasize the salient regions of the original content image.

To overcome the above problem, we introduce a category loss to restrain the classification results of the transferred images. After adding category loss, we not only can stabilize the Top-1 classification category of transferred images and make CAM loss tend to convergence, but also can maintain the original salient semantic information of content image.

For a transferred image, let L(X) be the Top-1 class of the transferred image predicated by the SqueezeNet, and the L(X*) be the Top-1 class of corresponding content image classified by SqueezeNet; then, the category loss is defined as formula (8), which is the cross entropy loss between the predicted value L(X*) and true value L(X) of the transferred image category. For a transferred image, the predicted value is the Top-1 class when the transferred image send to the SqueezeNet, and the true value is the Top-1 class when correspond content image send to the SqueezeNet.

$$loss_{cate}\left(X, X^*\right) = -L\left(X\right) lnL\left(X^*\right) - \left(1 - L\left(X\right)\right) \ln\left(1 - L\left(X^*\right)\right) \quad (8)$$

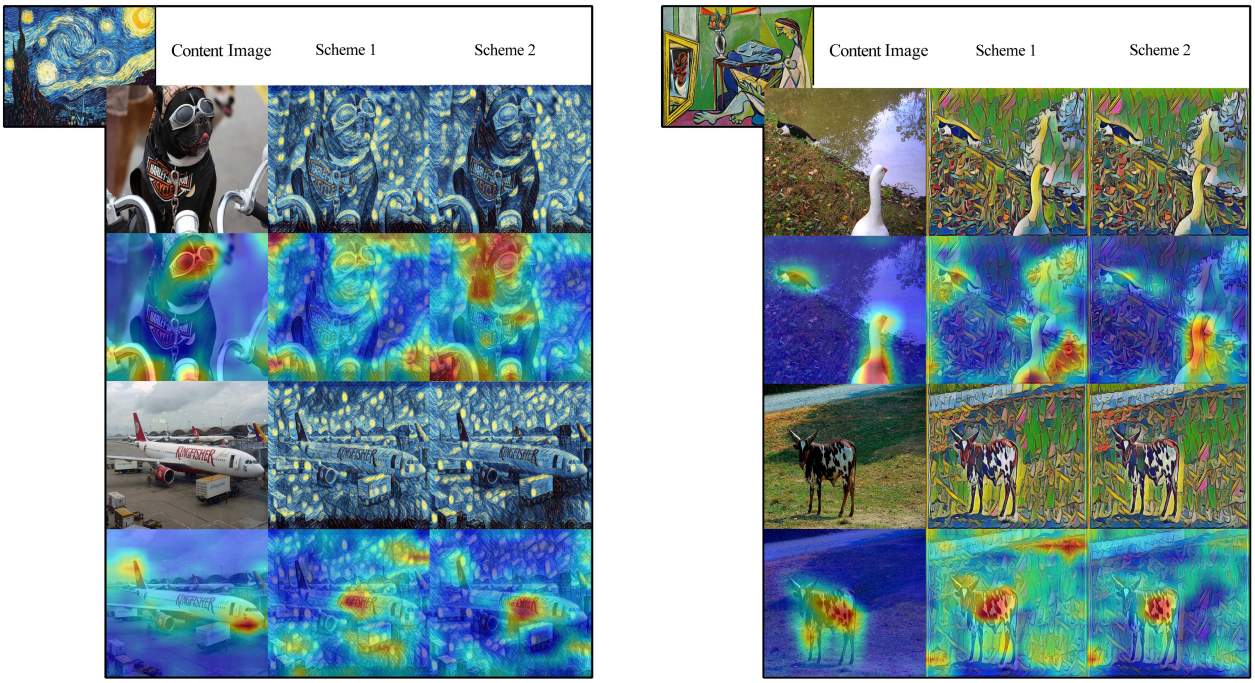
3) REGION LOSS

The final region loss is obtained by multiplying category loss with a weighted factor, and adding cam loss together. And its formula as shown in (9), and ε generally is set into a suitable value (default value: 40) on experience.

$$loss_{region}\left(X, X^*\right) = loss_{cam}\left(X, X^*\right) + \varepsilon loss_{cate}\left(X, X^*\right) \quad (9)$$

F. DEFINITION OF TOTAL LOSS

The total loss function of our model is proposed as (10), which combines following two parts of perception and region losses into a linear function. X*, X and Y refer to the generated Image, content image, and style image respectively; λ_1 and λ_2 are the weighted factors of every losses, which is



(a) (b)

FIGURE 9. Performance of proposed scheme with different losses. (a) Stylized by painting of starry night. (b) Stylized by painting of contemplative model.

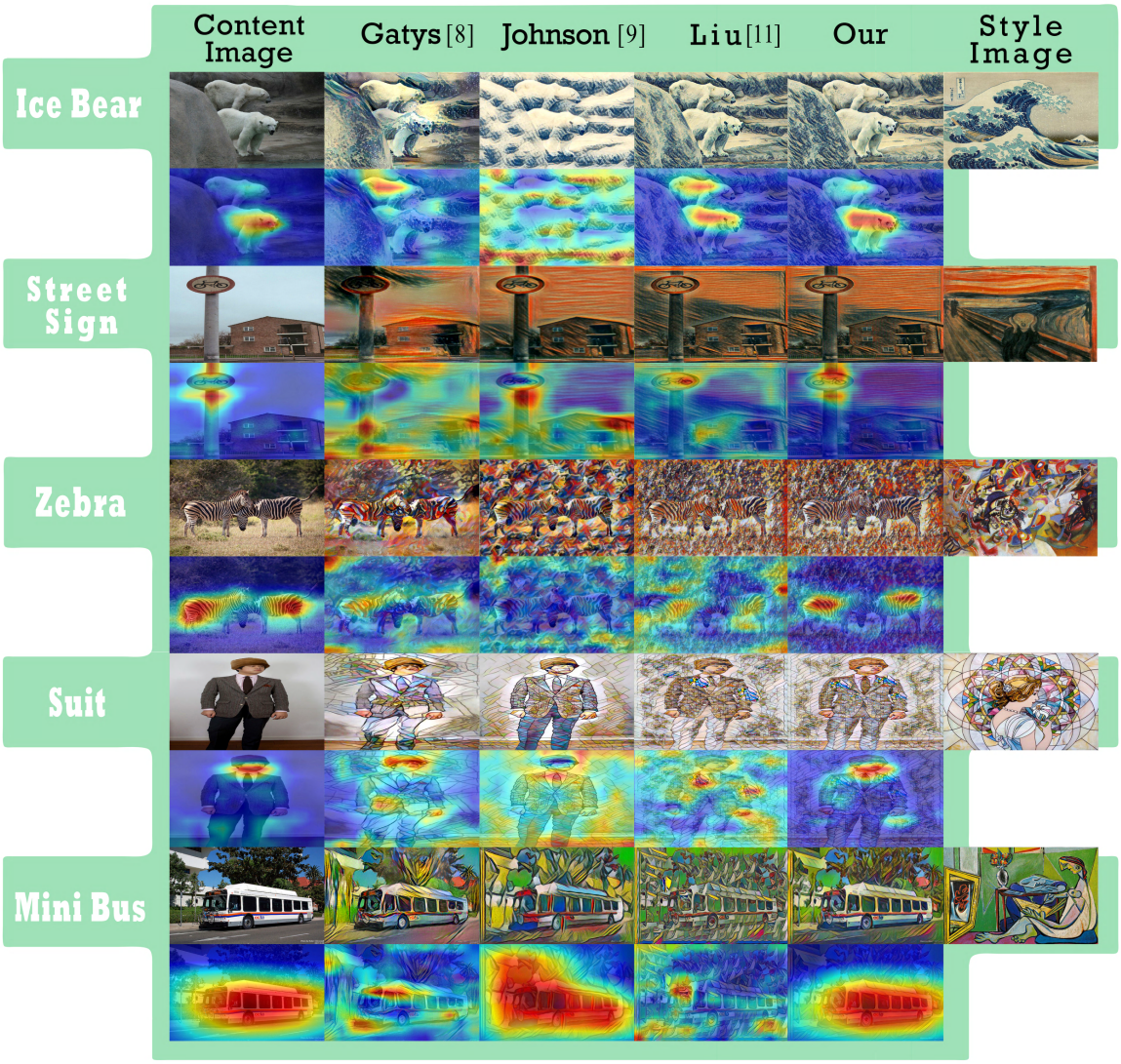


FIGURE 10. Performance evaluation of proposed scheme with other schemes.

set according to experience.

$$\mathrm{TL}\left(X^*, (\mathrm{X}, \mathrm{Y})\right) = \lambda_1 \mathrm{loss}_{perc}\left(X^*, (\mathrm{X}, \mathrm{Y})\right) + \lambda_2 \mathrm{loss}_{\mathrm{region}}\left(X, X^*\right) \quad (10)$$

According to the value of total loss, the image transfer network will be optimized iteratively. Finally, the transferred image generated by Y-style transfer network not only achieve more natural and attractive visual effects, but also remain the original salient regions of content image, which helps in the tasks of object localization and classification.

IV. VERIFICATION

A. EXPERIMENTAL PLATFORM

The proposed scheme not only needs to compute a lot of matrix operations, but also requires to process forward and backward propagations on the image transfer network and forward propagation on loss network iteratively, which will consume large amounts of computing resources [31]. Thus, we use the NVIDIA P40 GPU memory with 24G DDR III for training our proposed model, and deep learning framework is the PyTorch with Linux version of Python 2.7.

B. DATASETS

In this paper, Microsoft COCO 2014 datasets [46] are used for evaluating our proposed scheme. Since the image transfer network belongs to unsupervised learning areas, it does not need any label information. We mainly use the original image data of the COCO datasets as the input content images by adjusting the original image into 224 × 224 pixels in the training phase.

C. EVALUATING PROPOSED SCHEME WITH DIFFERENT LOSSES

In this subsection, we evaluate our proposed schemes combining with different losses, the CAM losses of two schemes

TABLE 1. Comparison in transferring time and model size.

	Gatys[8]	Johnson[9]	Liu[11]	Our scheme
Transferring time	280.32s	83.94ms	83.57ms	82.96ms
Model size	574673361KB	6725021KB	6723222KB	6723186KB

in the training stage are shown in Fig. 8, where the red line and blue line stand for the CAM loss of proposed scheme 1 with CAM loss and the CAM loss of proposed scheme 2 with CAM loss and category loss.

As shown in Fig. 9 (a) and (b), both of two schemes maintain image information of their original salient regions. The first column is content image and its CAM image, the second column is the effect of proposed scheme 1 with only cam loss, the third column is the performance of proposed scheme 2 with CAM and category losses. The CAMs of images transferred by scheme 1 can keep consistent with that of original content images, but their salient regions seem to diverge. After combining category loss, the CAMs of images transferred by scheme 2 reduce tend to converge, which highlight that the key actions or areas of the image. Thus, our experimental results show that after joined the cam and category losses, the generated stylized images not only have the natural blend of structure of content images and texture of style image, but also can maintain the most of salient regions and category information of original content images. We choose the scheme 2 as our proposed scheme.

D. COMPARING WITH OTHER METHODS

In this subsection, we compare proposed scheme with other existing methods. As shown in Fig. 10, the second column is the effects of scheme proposed by Gatys *et al.* [8], the third column is the performance of scheme proposed by Johnson *et al.* [9], the fourth column is the performance of scheme proposed by Rosin *et al.* [11], and the last column is performance of our proposed scheme. It is found that our scheme can generate stylized images keeping the salient regions of the image, which does not distort the focused action or area of the content image, and maintains the original semantic information is maintained. Furthermore, the stylized images is nature and attractive, which avoiding the tendency of simple texture overlay.

Structural similarity (SSIM) is used for measuring the similarity between two images. The Fig. 11 shows the SSIMs of CAMs between content images with stylized images generated by different schemes. In different cases of content and style images, the SSIMs of CAMs are higher than the other schemes. That is, the CAMs of stylized images generated by our scheme is most similar with that of original content images, which can keep the semantic information, which are help for the task of object localization and image classification.

We also evaluate the computational efficiency of proposed scheme by comparing with the existing methods in transferring time and memory size. The results are as following table. The memory size are the trained model size. The transferring

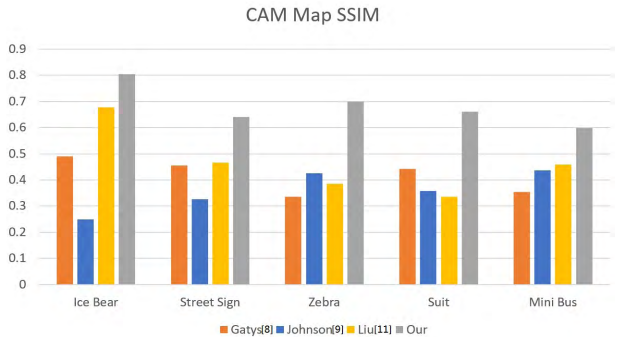


FIGURE 11. SSIM of CAMs between content images with stylized images.

time of each model are the average time of 15 times' style transferring by inputting images of size 480×640 pixels into each model. As shown in Table 1, the transferring time and model size of [9] and [11] and our scheme are almost the same since they all work based on the fast neural network proposed by Johnson *et al.* [9], and the transferring time and model size of [8] are relatively large as it works on pre-trained VGG model and transfers image's style to content images by iterative computation.

Thus, by comparing with other existing methods, our scheme shows some advantages. It focuses on analyzing the salient regions of synthetic images. By adding the region loss including two sub-losses of cam loss and category loss, our scheme can not only maintaining the salient regions of synthetic images alike to that of original content images, meanwhile, but also can transfer style well and avoid the effects of simply texture overlay, which provides more attractive visual effects since the salient regions are preserved. And it can still work fast and its size is small since it performs based on fast neural network. Furthermore, our scheme can assist in the saliency-based tasks, such as classification and localization, image compression and encoding, enhancement of images' edges and regions, object segmentation and so on. The main limitations of our scheme is that it only focuses on salient regions of synthetic images, neglecting the other semantics such as edge, depth, and colors.

V. CONCLUSION

In the process of neural style transfer, the existing methods can obtain a certain stylized image with content and style directly, but they also ignore some semantics of the content images. For example, these methods greatly change the salient regions or key action areas of transferred images, which is not the same with content images. In order to solve the above problem, this paper proposed region loss containing CAM loss and category loss calculated from SqueezeNet network. The generated image can not only preserve the key

semantics of original salient regions, but also tend to more attractive and natural, avoiding the effects of simply texture overlay and helping for the tasks of saliency-based tasks of generated images, such as object localization, classification, segmentation, and image compression.

Our scheme still have some limitations. First, it only considers one semantics (like edge, color), how to preserve multiple semantics is still a problem, which can be solved by extending our proposed framework to retain other structural or semantics of content or style images by using other corresponding estimation networks and losses. Second, to achieve controlling multi-styles (such as texture and color) while maintaining semantics well is also needed for us to research, which can make the content richer, the style more diverse and stereo. Third, it is also important to study transferring different styles to other paintings (such as comics, ink painting, photos, etc.) and achieve a balance between semantics and style. At last, since our scheme is based on the fast neural network proposed by Johnson *et al.* [9], it is necessary to train a specific model for each transferred style. Thus, we hope to realize multi-stylization in a single model with preserving main semantics of images in the future.

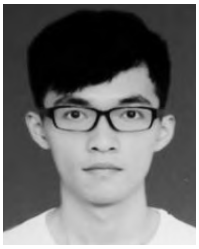
REFERENCES

[1] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song. (2017). "Neural style transfer: A review." [Online]. Available: https://arxiv.org/abs/1705.04058
[2] H. Huang, Y. Zang, and L. Zhang, "Survey on image and video painterly rendering," *Comput. Sci.*, vol. 38, no. 6, pp. 1–7, 2011.
[3] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. J. Comput. Vis.*, vol. 40, no. 1, pp. 49–70, Oct. 2000.
[4] S. Greenberg, "Why non-photorealistic rendering?" *ACM SIGGRAPH Comput. Graph.*, vol. 33, no. 1, pp. 56–57, Feb. 1999.
[5] T. Strothotte and S. Schlechtweg, *Non-Photorealistic Computer Graphics: Modeling, Rendering, and Animation*. San Francisco, CA, USA: Morgan Kaufmann, 2002, pp. 431–457.
[6] J. Lopez-Moreno, "Non-photorealistic rendering," in *Encyclopedia of Color Science and Technology*, R. Luo, Eds. Berlin, Germany: Springer, 2015.
[7] L. A. Gatys, A. S. Ecker, and M. Bethge. (2015). "A neural algorithm of artistic style." [Online]. Available: https://arxiv.org/abs/1508.06576
[8] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2414–2423.
[9] J. Johnson, A. Alahi, and L. Fei-Fei. (2016). "Perceptual losses for real-time style transfer and super-resolution." [Online]. Available: https://arxiv.org/abs/1603.08155
[10] C. Castillo, S. De, X. Han, B. Singh, A. K. Yadav, and T. Goldstein. (2017). "Son of zorn's lemma: Targeted style transfer using instance-aware semantic segmentation." [Online]. Available: https://arxiv.org/abs/1701.02357
[11] X.-C. Liu, M.-M. Cheng, Y.-K. Lai, and P. L. Rosin, "Depth-aware neural style transfer," in *Proc. Symp. Non-Photorealistic Animation Rendering, NPAR*, Los Angeles, CA, USA, 2017, pp. 1–10, doi: 10.1145/3092919.3092924.
[12] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua. (2017). "StyleBank: An explicit representation for neural image style transfer." [Online]. Available: https://arxiv.org/abs/1703.09210
[13] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1–6.
[14] W. Zhang, C. Cao, S. Chen, J. Liu, and X. Tang, "Style transfer via image component analysis," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1594–1601, Jul. 2013.
[15] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017.
[16] L. Zhao, Q. Sun, and Z. Zhang, "Single image super-resolution based on deep learning features and dictionary model," *IEEE Access*, vol. 5, pp. 17126–17135, 2017.
[17] G. Wang, "A perspective on deep imaging," *IEEE Access*, vol. 4, pp. 8914–8924, 2017.
[18] R. Novak and Y. Nikulin. (2016). "Improving the neural algorithm of artistic style." [Online]. Available: https://arxiv.org/abs/1605.04603
[19] H. Zhang and K. Dana. (2017). "Multi-style generative network for real-time transfer." [Online]. Available: https://arxiv.org/abs/1703.06953
[20] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. (2016). "Controlling perceptual factors in neural style transfer." [Online]. Available: https://arxiv.org/abs/1611.07865
[21] R. Yin. (2016). "Content aware neural style transfer." [Online]. Available: https://arxiv.org/abs/1601.04568
[22] L. A. Gatys, M. Bethge, A. Hertzmann, and E. Shechtman. (2016). "Preserving color in neural artistic style transfer." [Online]. Available: https://arxiv.org/abs/1606.05897
[23] Y. Nikulin and R. Novak. (2016). "Exploring the neural algorithm of artistic style." [Online]. Available: https://arxiv.org/abs/1602.07188
[24] E. Risser, P. Wilmot, and C. Barnes. (2017). "Stable and controllable neural texture synthesis and style transfer using histogram losses." [Online]. Available: https://arxiv.org/abs/1701.08893
[25] F. Luan, S. Paris, E. Shechtman, and K. Bala. (2017). "Deep photo style transfer." [Online]. Available: https://arxiv.org/abs/1703.07511
[26] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng, "Salient object detection: A discriminative regional feature integration approach," *Int. J. Comput. Vis.*, vol. 123, no. 2, pp. 251–268, 2017.
[27] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.
[28] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.
[29] T. V. Nguyen, Q. Zhao, and S. Yan, "Attentive systems: A survey," *Int. J. Comput. Vis.*, vol. 126, no. 1, pp. 86–110, 2018.
[30] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016.
[31] J. Luo, C.-M. Vong, and P. K. Wong, "Sparse Bayesian extreme learning machine for multi-classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 4, pp. 836–843, Apr. 2014.
[32] K. Simonyan, A. Vedaldi, and A. Zisserman. (2013). "Deep inside convolutional networks: Visualising image classification models and saliency maps." [Online]. Available: https://arxiv.org/abs/1312.6034
[33] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-Weakly-supervised learning with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 685–694.
[34] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. (2016). "Learning deep features for discriminative localization." [Online]. Available: https://arxiv.org/abs/1512.04150
[35] M. Tzelepi and A. Tefas, "Deep convolutional learning for content based image retrieval," *Neurocomputing*, vol. 275, pp. 2467–2478, Jan. 2018.
[36] C. Kanan and G. Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics," in *Proc. Comput. Soc. Conf. Vis. Pattern Recognit.*, Jun. 2010, vol. 119, no. 5, pp. 2472–2479.
[37] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 769–779, May 2013.
[38] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 409–416.
[39] H. Li and K. N. Ngan, "Saliency model-based face segmentation and tracking in head-and-shoulder video sequences," *J. Vis. Commun. Image Represent.*, vol. 19, no. 5, pp. 320–333, 2008.
[40] S. Frintrop and M. Kessel, "Most salient region tracking," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2009, pp. 1869–1874.
[41] K. He, X. Zhang, S. Ren, and J. Sun. (2015). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1512.03385
[42] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[44] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. (2016). "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size." [Online]. Available: https://arxiv.org/abs/1602.07360
[45] Y. Zhang *et al.*, "Multi-kernel extreme learning machine for EEG classification in brain-computer interfaces," *Expert Syst. Appl.*, vol. 96, pp. 302–310, Apr. 2018.
[46] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, vol. 8693. Springer, 2014, pp. 740–755.

YIJUN LIU received the B.S. degree from Beijing Normal University, in 1999, the M.Sc. degree from the Guangdong University of Technology, China, in 2002, the M.Phil. degree in 2003, and the Ph.D. degree from The University of Manchester, U.K., in 2005, all in computer science.

He is currently a Full Professor with the School of Information Engineering, Guangdong University of Technology, Guangzhou, China. His current research interests include neuromophic computing, deep learning, computer architecture, and GPS/Beidou Navigation.



ZUOTENG XU received the B.S. degree in information engineering from the School of Information Engineering, Guangdong University of Technology, Guangzhou, China, in 2017.

He is currently pursuing the master's degree with the Guangdong University of Technology. His current research interests include computer vision, including image generation, deep learning, spectral clustering in machine learning, and their applications in images.

WUJIAN YE received the B.S. degree in computer science and technology from the School of Computers, Guangdong University of Technology, Guangzhou, China, in 2010, and the M.S and Ph.D. degrees in computer science from Dankook University, South Korea, in 2012 and 2015, respectively.

Since 2016, he has been a Lecturer with the School of Information Engineering, Guangdong University of Technology. His current research interests include machine learning application, computer networks and security analysis, deep learning and computer vision, and voice recognition.

ZIWEN ZHANG received the B.S. degree from the Department of Surveying and Mapping Engineering, East China Jiaotong University, in 2007, and the M.S. and Ph.D. degrees in geodesy and surveying engineering from the Liaoning University of Technology, China, in 2013 and 2017 respectively.

He is currently a Post-Doctoral Researcher with the Guangdong University of Technology. His current research interests include geotechnical application of differential interferometry for spaceborne radar and deep learning.

SHAOWEI WENG received the B.S. degree from North China Electric Power University and the Ph.D. degree from Beijing Jiaotong University, in 2009. From 2016 to 2017, she was a Visiting Scholar with the New Jersey Institute of Technology, USA.

She is currently an Associate Professor with the School of Information Engineering, Guangdong University of Technology. Her current research interests include image processing, data hiding and digital watermarking, pattern recognition, and computer vision.

CHIN-CHEN CHANG received the Ph.D. degree in computer science from National Tsing Hua University, in 1982. Since 1982, he has been an Associate Professor with National Chiao Tung University, a Professor with National Chung Hsing University, the Chair and a Professor with the Computer Science Department, National Chung Cheng University, the Director of the Automation Research Center, the Dean of the College of Engineering, a Provost, and the Acting President of National Chung Cheng University. He was the Director of the Advisory Office, Ministry of Education, Taiwan, and was a Visiting Scholar/Researcher with Tokyo University and Kyoto University. He is currently a Chair Professor with Feng Chia University and an Honorary Professor with National Chung Cheng University He holds a joint appointment at National Chiao Tung University.

He has worked on many different topics in information security, cryptography, and multimedia image processing, and has published several hundreds of papers in international conferences and journals and over 30 books. He was cited over 27 668 times and has an h-factor of 80 according to Google Scholar. Several well-known concepts and algorithms were adopted in textbooks.

HUAJIN TANG received the B.S. degree from Zhejiang University, in 1998, the M.S. degree from Shanghai Jiao Tong University, in 2001, and the Ph.D. degree in computer engineering from the National University of Singapore, in 2004.

From 2008 to 2014, he was the Director of the Cognitive Computing and Robotics Laboratory, Institute of Information and Communication, Research Bureau of Science and Technology, Singapore. Since 2014, he has been the Director of the Brain Computing Research Center, College of Computer Science, Sichuan University, China.

In 2013 and 2015, he was selected into the programs of national "Thousand Talent Youth Plan" and sichuan "Thousand Talent Plan." His current research interests include brain computing, neuromorphic computing and cognitive systems, neural circuits, intelligent hardware, and intelligent robot. He is currently the Chairman of the 2016 Internal Workshop on Neuromorphic Computing and Cyborg Intelligence and the Education Sub-Committee of the IEEE Institute of Computational Intelligence. He is also a Deputy Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and the IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS.