



IMAGE STYLE TRANSFER MODEL WITH CONTENT PRESERVATION

*Ruixue Liu¹, Shengbei Wang¹, Weitao Yuan¹

¹ Department of Computer Science and Software Engineering, Tianjin Polytechnic University, Binshuixi Road 300387, Tianjin, China.

ABSTRACT

The image rendered in another image style, called style transfer, is one of the most interesting applications of deep learning. Style transfer plays an increasingly important role in basic computer vision research and industrial applications. The method of iterative optimization [1] and training feedforward convolutional neural network [2-4] are used to continuously achieve fast and good effects on style transfer. However, most of these methods ignore an essential problem that the content and contour information of the original images could be lost after the transfer process. Throughout the history of style transfer, although the performance is constantly improving, pursuing higher flexibility and speed, an essential problem is ignored, that is, the main content and contour are unavoidably blurred after style transfer. It is found that in many application fields, the content/contour of the original image has great importance and thus should be kept as clear as possible after style transfer. In this paper, a sub-structure named Important Content Contour Extraction (ICCE) is proposed to generate masks and therefore preserve the clear contour and content after style transfer. The Feature Extraction Module (FEM) for extracting advanced features is added in image style transfer network which is trained by the perceptual loss function. We conducted some experiments to verify the effectiveness of the proposed method. Experiments demonstrated that the proposed framework achieves a good compromise in speed, flexibility and quality. The experiment results showed that the proposed model had good ability for style transfer. Meanwhile, it could preserve clear content and contour of the original images, which was better than [1-4].

KEYWORDS: style transfer, feature extraction, content preservation, feedforward convolution network.

1. INTRODUCTION:

Style transfer is the process of rendering the content of the original image using the style of another image. The work of Gatys et al [1] pioneers a new field called Neural Style Transfer, which uses Convolutional Neural Networks (CNN)[5] to change the style of an image while preserving its content. Style transfer is non-photorealistic rendering [6] and closely related to texture synthesis and transfer[7,8]. Many solutions have been proposed for style transfer which can be divided into two categories. One category is based on the iterative optimization method. For example, the works [9-11] synthesize the stylized image by iteratively updating the pixels in the image through back-propagation. This kind of methods optimizes the generated image to obtain higher quality results. However, there is a disadvantage that the running speed is too slow. Since the large BFGS (L-BFGS)[9] is employed in optimization and used once for each stylization, which leads to high running costs of the algorithm. The other category is implemented based on training the feed-forward convolutional neural networks and then synthesizes the stylized image using a forward pass. In early days, images are transferred by single style feed-forward convolution network such as [12]. The main limitation of these networks is that each network is limited to one style, which will lose the flexibility of the original algorithm. Later, some of these methods are evolved into generation models built by the forward convolutional network to generate multiple styles such as [13-15]. Although the flexibility is improved, the quality of style transfer is unstable and the styles are limited. Moreover, some of these methods output blocky results or a style filter effect, which is not satisfactory.

In many application fields, the content of the original image has great importance. For example, in the animation industry, style transfer can be applied for coloring. In this case, the main content/moving object needs to be separated from the background. In the movie industry, captured images can be rendered directly in real time. It is also necessary to distinguish the protagonist and the background when asked eye after style transfer. In addition, image style transfer has been widely used in many commercial software, such as Prisma, Alter, etc. Therefore, it is now quite necessary to consider how to preserve the content, contour, and image quality after style transfer, under the premise of satisfactory flexibility and transfer efficiency.

2. MODEL:

As shown in Fig. 1, our model consists of three components: an Image Style Transformation (IST) network, Important Content Contour Extraction (ICCE) structure, and the VGG loss network which uses VGG16. The feedforward convolutional IST network is trained by the loss function of the VGG network. The content contour generated by the ICCE structure, the content image, and the style image are then input to the IST network for style transfer. Finally, satisfactory style transfer images are obtained. A detailed description of each module is as follows.

3.1 ICCE structure:

The implementation of ICCE structure is mainly divided into three steps. Firstly, the candidate

Region of Interests (ROIs) are sent to the Region Proposal Network (RPN)[21] for binary classification (foreground or background) and bounding-box regression and the other unrelated ROIs are filtered out. Secondly, the ROIAlign [20] operation is performed on the remaining ROIs. As a result, the original image is corresponded to the pixel of the feature map and then the feature map is corresponded to the fixed feature. Finally, the ROIs are classified into N categories and regressed with bounding-box, and then the Mask for content can be generated.

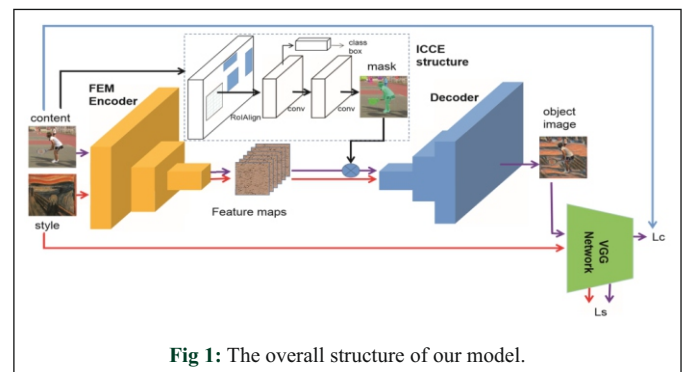


Fig 1: The overall structure of our model.

3.2 IST Network:

The IST network consists of two subnets: advanced feature extraction subnet and improved generation subnet. As to effectively extract the features of image, the Feature Extraction Module (FEM) is designed in the advanced feature extraction subnet. The FEM is composed of a set of two-dimensional convolution filters, two-dimensional convolution normalization, and a highway network [24]. It is worthwhile to note that our FEM is inspired but different from CBHG [25]. This is because the original image feature (H-embedding) and the low-dimensional embedding (L-embedding) are synthesized by considering the characteristics of the input image to produce a hybrid enhanced feature, which is more suitable for IST network.

The images for training are input into the IST network. The output images are then processed with the content loss and the style loss networks to calculate their losses, and the total loss is a weighted combination of the style loss and the content loss. The above processes is used to train the style transfer IST network. Similar to the loss definition in the previous quick style transfer work [2-4], the content loss and the style loss are referred to Gatys et al [1].

3.3 Complex Perceptual Loss Functions:

We define two complex perceptual loss functions that measure the high-level perceptual and semantic differences between images.

Starting from some initialization of r , the algorithm adapts to r to minimize the loss function. $L_{content}$ is the content loss and L_{style} is the style loss. The scaling

hyperparameters are a and b . Given a set of content layers C and a set of style layers S , $\theta_l(x)$ ($l \in C, S$) are the classifier activations at layer l , U_l is the total number of units at layer l and $G(\theta_l(x))$ is the Gram matrix associated with the layer l activations.

The content loss is the (squared, normalized) Euclidean distance between content features, where c is the content image,

$$L_{content}(r, c) = \sum_{i \in C} \frac{1}{U_i} \|\theta_i(r) - \theta_i(c)\|_2^2 \quad (1)$$

The style loss is the squared Frobenius norm of the difference between the Gram matrices of the output and the style image, where s is the style image,

$$L_{style}(r, s) = \sum_{j \in S} \frac{1}{U_j} \|G(\theta_j(r)) - G(\theta_j(s))\|_F^2 \quad (2)$$

In order to speed up the procedure outlined above, a feed-forward convolutional network, termed style transfer network IST Networks, is introduced to learn the transformation. The process of learning transformation through IST network is called T transformation. The network is trained on many content images using the weighted combination of loss functions listed above, i.e

$$L(s, c) = aL_{content}(T(c), c) + bL_{style}(T(c), s) \quad (3)$$

4. EXPERIMENTS SETTING AND COMPARISONS:

In the training part, we used the Microsoft COCO dataset to train our IST network. In the testing part, we used the content and style images from the Internet and those from papers [1–4] for comparison. We trained the network with a batch size of 4 for iterations. The neural network weights were updated using the Adam optimization method. In the experiments, we computed the style loss at layer $relu_1$, $relu_2$, $relu_3$, and $relu_4$ of the pretrained VGG-16 network and the content loss at layer $relu_2$ of the pre-trained VGG-16 network.

(i) Comparison with iterative optimization method:

In Fig. 2, we compared our results with iterative optimization method [1]. Our method used $a/b = 1/100$ (in Eq. (3)) in the comparison. The Gatys et al [1] apparently could not work well in Lenna image and even had blocky results. The face of the Lenna image was not well transferred. The transfer of the second image also showed a blocky effect and the contour information was blurred. Our approach was thousands of times faster than the iterative optimization based method [1].

(ii) Comparison with arbitrary content image & single style method:

In Fig. 3, we compared our results with feed-forward network method [2]. Our method used $a/b = 1/75$ (in Eq. (3)) in the comparison. Compared to [2] method, our results were obviously more vivid after the transfer and the image style was embedded in the content image to make the content image more abstract. For instance, in the facial expression in Fig. 3, our results showed that the style texture was assigned to the facial expression and the original information was not lost. However, the result of the method [2] was like the repetition of style texture on the content image.

(iii) Comparison with arbitrary content image & arbitrary style method:

In Fig. 4, we compared our results with [4] which is an Adaptive Instance Normalization (AdaIN) improved based on the Conditional Instance Normalization (CIN). We used $a/b = 1/50$ (in Eq. (3)) for the comparison. According to the results obtained by [4], the information of the content image was lost and the contour of the castle and the Golden Gate Bridge was almost indistinguishable. In contrast, our results had stylized features and the content were very clear.

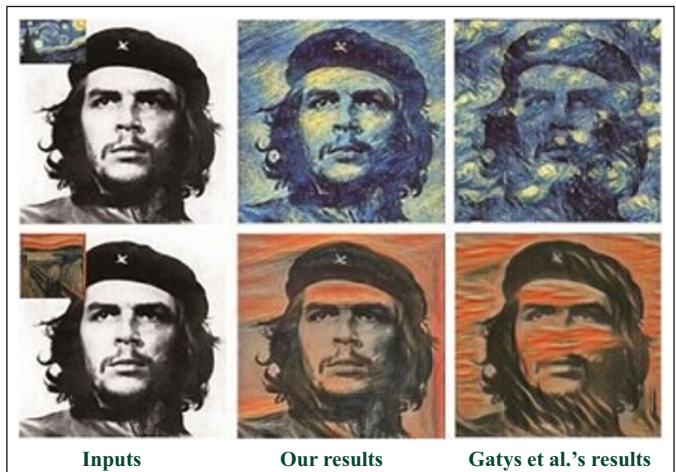
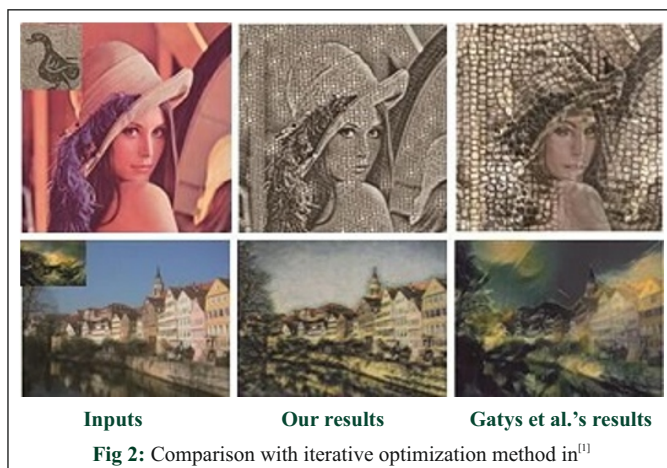


Fig 3: Comparison with arbitrary content image & single style method in [2]



Fig 4: Comparison with arbitrary content image & arbitrary style method in [4]

4.2.2 Subjective image quality score:

Ten male subjects and ten female subjects with normal vision participated in the scoring test. In order to ensure the fairness of the score, we showed the experiment results in Figs. 2, 3, 4 side by side and distributed them to the subjects randomly. The subjects then reported scores for the above images. The score range was set from "1" to "10". "1" meant that the content and contour information of the stylized image were completely indistinguishable and "10" meant that the content and contour information of the stylized image was very clear. We showed the averaged subjective scores of 20 subjects in the Tab. 1. These results showed that the [1,2,4] methods had lower scores and it was difficult for subjects to distinguish important content and contour information in stylized image. In contrast, our method could provide clear content and contour information to the subjects according to their scores.

Table 1: Average scores for subjective evaluations

Method	1	2	4	Ours
Average score	5.53	7.64	6.84	8.13
Standard Deviation	1.94	2.03	3.14	0.93

5. CONCLUSIONS:

This paper proposed a novel image style transfer framework. Different from the existing methods, a structure for extracting the important content from images was added. Then, based on the deep learning strategy, the image style transfer network and the VGG loss network were used to achieve fast transfer and obtain high quality stylized images. The experiment results showed that our model can perform image style transfer well while retaining important target content and contour information.

REFERENCES:

1. L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," CoRR abs/1508.06576 (2015).
2. D. Ulyanov, V. Lebedev, A. Vedaldi, et al., "Texture networks: Feed-forward synthesis of textures and stylized images," in ICML 2016, New York City, NY, USA, June 19-24, 2016, 1349–1357 (2016).
3. D. Chen, L. Yuan, J. Liao, et al., "Stylebank: An explicit representation for neural

- image style transfer,” in CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2770–2779 (2017).
- 4 X. Huang and S. J. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in ICCV 2017, Venice, Italy, October 22-29, 2017, 1510–1519 (2017).
 - 5 A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States., 1106–1114 (2012).
 - 6 J. E. Kyprianidis, J. P. Collomosse, T. Wang, et al., “State of the art”: A taxonomy of artistic stylization techniques for images and video,” IEEE Trans. Vis. Comput. Graph. 19(5), 866–885 (2013).
 - 7 A. A. Efros and W. T. Freeman, “Image quilting for texture synthesis and transfer,” in Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2001, Los Angeles, California, USA, August 12-17, 2001, 341–346 (2001).
 - 8 M. Ashikhmin, “Synthesizing natural textures,” in Proceedings of the 2001 Symposium on Interactive 3D Graphics, SI3D 2001, Chapel Hill, NC, USA, March 26-29, 2001, 217–226 (2001).
 - 9 L. A. Gatys, A. S. Ecker, and M. Bethge, “Texture synthesis using convolutional neural networks,” in Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, 262–270 (2015).
 - 10 F. Luan, S. Paris, E. Shechtman, et al., “Deep photo style transfer,” in CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 6997–7005 (2017).
 - 11 R. Mechrez, E. Shechtman, and L. Zelnik-Manor, “Photorealistic style transfer with screened poisson equation,” CoRR abs/1709.09828 (2017).
 - 12 J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and superresolution,” in ECCV 2016, Proceedings, Part II, 694–711 (2016).
 - 13 V. Dumoulin, J. Shlens, and M. Kudlur, “A learned representation for artistic style,” CoRR abs/1610.07629 (2016).
 - 14 A. Mahendran and A. Vedaldi, “Understanding deep image representations by inverting them,” in CVPR 2015, Boston, MA, USA, June 7-12, 2015, 5188–5196 (2015).
 - 15 H. Zhang and K. J. Dana, “Multi-style generative network for real-time transfer,” CoRR abs/1703.06953 (2017).
 - 16 Y. Li, H. Qi, J. Dai, et al., “Fully convolutional instance-aware semantic segmentation,” in CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 4438–4446 (2017).
 - 17 Y. Li, H. Qi, J. Dai, et al., “Fully convolutional instance-aware semantic segmentation,” in CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 4438–4446 (2017).
 - 18 S. Liu, J. Jia, S. Fidler, et al., “SGN: sequential grouping networks for instance segmentation,” in ICCV 2017, Venice, Italy, October 22-29, 2017, 3516–3524 (2017).
 - 19 M. Bai and R. Urtasun, “Deep watershed transform for instance segmentation,” in CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2858–2866 (2017).
 - 20 K. He, G. Gkioxari, P. Dollár, et al., “Mask R-CNN,” in ICCV 2017, Venice, Italy, October 22-29, 2017, 2980–2988 (2017).
 - 21 S. Ren, K. He, R. B. Girshick, et al., “Faster R-CNN: towards real-time object detection with region proposal networks,” in Advances in Neural Information Processing Systems 28, December 7-12, 2015, Montreal, Quebec, Canada, 91–99 (2015).
 - 22 J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in CVPR 2015, Boston, MA, USA, June 7-12, 2015, 3431–3440 (2015).
 - 23 R. B. Girshick, “Fast R-CNN,” in ICCV 2015, Santiago, Chile, December 7-13, 2015, 1440–1448 (2015).
 - 24 R. K. Srivastava, K. Greff, and J. Schmidhuber, “Training very deep networks,” in Advances in Neural Information Processing Systems 28, December 7-12, 2015, Montreal, Quebec, Canada, 2377–2385 (2015).
 - 25 Y. Wang, R. J. Skerry-Ryan, D. Stanton, et al., “Tacotron: Towards end-to-end speech synthesis,” in Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017, 4006–4010 (2017).