Can Semantics Boost Style Transfer? An Enhanced Context-aware Cycle-GAN

Chen Ze*
School of the Gifted Young
chenze_ustc@mail.ustc.edu.cn

Chen Zeyuan
School of the Gifted Young
nnice1216@mail.ustc.edu.cn

Wang Yu
School of the Gifted Young
wy2001@mail.ustc.edu.cn

Wu Yinxin
School of Data Science
wuyxin@mail.ustc.edu.cn

Wu Yuefan
School of the Gifted Young
wyf8899@mail.ustc.edu.cn

Abstract

Style transfer is one of the most well-investigated tasks in computer vision that the semantic content of an image can be rendered into different styles. Various methods have been proposed to accomplish the tasks and boost the visual performance. However, there are several inherent problems in the existing style-transfer models: the objects in the images could interfere with each other, thus causing undesired transfers on some unrelated parts of the image. In this work, we aim to design a framework that first segment the image into different parts, and then transfer the styles of these parts independently. Then, we can combine the transferred image segments together to obtain the final output. The experimental results have shown that our goal has been accomplished ideally. The codes are available at https://github.com/wangyu-ustc/ECCG/

1 Introduction

Is there a way that artificial intelligence can produce works of art? With the rapid advance of computer vision technologies, more researches begin to pay attention on the intersection of computer vision and art. Style transfer is a task that has taken the computer vision community by storm, once after the pioneer [4] become one of the top trends in 2016. People are fascinated while seeing the masterpieces of Claude Monet's be expressed in a Van Goph style.

One state-of-the art style transfer method is CycleGAN[15], which utilizes Generativa Adversarial Network (GAN) together with cycle consistency constraint, to transfer one image from its source domain to a specific target domain. However, it still has some inherent problems. As shown in figure 1, the example demonstrate that although the original intention of training the model is to make the model transfer the horses in the picture into zebras, in real applications, it will not recognize which part of the picture should be transferred. As a result, the person in figure 1(b) is also covered with white and black stripes. More details can be found in the next Section.

*The authors are ranked by alphabetic order. All authors come from University of Science and Technology of China. For every member's contribution, please refer to Section 6.

Preprint. Under review.

(a) Origin (b) Cycle-GAN [15] (c) ECCG (d) ECCG-Apply Twice

Figure 1: The style transfer results from horse to zebra. ECCG-background transferred is the result that we apply the transfer model twice, first for the foreground(the man and the horse) and second for the background. One can see that the second transferring process does not have any impact on the foreground.

In contrast, as shown in figure 1(c) the output of our model is much better than the output of Cycle-GAN. After segmentation, the horse are separated from the background, and then with transferring the horse alone, the other elements of the picture could remain unaffected. Besides, with this improvement ahead, we can transfer the animal and the background simultaneously, with different style-transfer models. As shown in figure 1(d), after transferring the horse and the background with different models, we get the picture of zebra walking in the winter, from the original picture depicting a horse walking in the summer.

Our contributions can be summarized as follows:

- A novel framework for the sytle transfer task is presented: *E*nhanced *C*ontext-aware *C*ycle-*G*an (**ECCG**). We separate the image into regions with different context information and apply style transfer on them separately, which avoids different elements in one image interfering with each other.
- An L0-smooth regularization is imposed to make the segmentation process more robust and boost the performance of style transfer.
- With the segmentation results, we are able to transfer different elements in a single image to different target domains. In addition, segmentation information could help tackle with the out-of-distribution problems, meaning that we can better process the images that have never showed up in the training set.
- We apply the Poisson Editing technology as the post-processing to obtain the seamless combination of different transferred parts, which have boost the fidelity of our outputs.

2 Related Work

Style Transfer. Style transfer algorithms aim at migrating the style from an example image to a content image. The origin of this problem is from non-photo-realistic rendering [6]. It is also closely related to texture synthesis and transfer [3, 2]. Gatys [4] firstly utilizes convolutional neural network to generate impressive style transfer results. After their work, various methods have been developed to address different aspects of the transfer problem, including visual quality [5], head portrait [10], and so on. Cycle-GAN [15] is the most famous style transfer method, in which the authors utilize Generative Adversarial Networks (GANS) to generate a transferred image in the target domain, by inputting an image from a source domain. Compared with previous methods that learn to transfer between two specific images, Cycle-GAN, on the other hand, aims to learn the mapping between two image collections, which is achieved by capturing correspondences between higher-level appearance structures. Hence, it could be applied to other tasks such as painting $\Longrightarrow$ photo. However, even with powerful transfer ability, it still suffers from some problems. Cycle-GAN takes the whole picture as

the input, and then output the transferred picture, in which case the transfer may achieve undesired results. For instance, the pretrained model that can transfer the horse to the zebra might mistake the background also as the object, thus it will add some black and white stripes on the grass or the sky. In our work, we use Image Segmentation before Cycle-GAN to avoid the problems mentioned above.

Semantic Segmentation. Semantic segmentation is a key topic in image processing and computer vision with applications such as scene understanding, medical image analysis, image compression, etc. Recently, with the rapid advance of deep learning, there has been a substantial amount of works aimed at developing semantic segmentation approaches using neural networks. The goal of semantic segmentation is to label *each pixel* of an image with a corresponding class of what is being represented. Since we are predicting for every pixel in the image, the task is commonly referred to as dense prediction. Long [7] is one of the first deep learning works for semantic segmentation, that uses a fully convolutional network (FCN). A FCN consists of only convolutional layers, which enables it to take images with arbitrary resolution as input, and produce a segmentation map of the same size. In order to integrate more context information, several methods incorporate probabilistic graphical models into DL architectures. For example, Chen [1] propose an algorithm based on the combination of CNNs and fully connected Conditional Random Fields (CRFs). There also exists some other similar approaches [14, 9] in semantic segmentation, which integrates CRF with CNN. In [11], Wu propose Fast-FCN, in which a computationally efficient joint upsampling module, named JPU, is introduced to replace the time and memory consuming dilated convolutions in the backbone FCN.

3 Method

In this section, we detail several components in our framework. Specifically, we introduce the image segmentation model in Section 3.1. In Section 3.2, we transfer the styles of image segments indivisually via CycleGAN [16] to fulfill the constraint of object independence. In Section 3.3, we use Poission Image Editing [8] for better visual perception of the transferred images.

3.1 Image Segmentation

We build our segmentation model as FastFCN [12], which adopts a widely used encoder-decoder structure where a joint pyramid sampling (JPU) is attached after the encoder to upsample the low-resolution final feature map. Specifically, Joint upsampling attempts to generate a high-resolution target picture from a low-resolution target image and a high-resolution guiding image by transferring features and structures from the guidance image. Briefly, the assumption underlying JPU is that (i) $y_l = f(x_l)$, where x_l and y_l are the low-resolution guidance image and the low-resolution target image, respectively — the low-resolution target image is generated from the high-resolution guidance image. (2) Regardless its resolution, the mapping from guidance images to target images are shared. Thus, with x_l and y_l, we are required to obtain a transformation $\hat{f}(\cdot)$ to approximate $f(\cdot)$ while the computation complexity of $\hat{f}(\cdot)$ is much lower than $f(\cdot)$. Formally, given x_l, y_l and x_h, we have

$$y_h = \hat{f}(x_h), \text{ where } \hat{f}(\cdot) = \underset{h(\cdot)\in\mathcal{H}}{\operatorname{argmin}} \|y_l - h(x_l)\| \tag{1}$$

As shown in Figure 3.4, the proposed JPU exploits multiscale context across multi-level feature maps, generating a high-resolution feature map for each image. Further, we can obtain the image segmentation mask via Encoding Head, whose effectiveness is well validated. We suggest readers refer to [12] for details.

3.2 CycleGAN

After the image segmentation, we group the segments into groups by their semantics, e.g. animals, people, backgrounds. In such fashion, we can change the styles for each group (or objects) individually without generating counterintuitive images. Then, we need to transform an object from a source domain X to a target domain Y in the absence of paired examples. Formally, we would like to learn the mapping $G : X \rightarrow Y$. For backgrounds, for example, X may denotes *Sunny* while Y denotes *Sunset*. Besides, we introduce a discriminative model D_Y in addition, to criticize the generated images $G(X)$. The discriminative model D_Y outputs a real values, indicating the probability from

Figure 2: Illustration of Poission Image Editing in seamless combination of image segments, where the sources represents X, destinations represents Y, cloning shows the failure combination while the seamless cloning by Poission Image Editing shows the successful one [8].

the input image to be "real" in the distribution Y. Ideally, $G(X)$ should be indistinguishable from Y. Thus, we use the the adversarial loss:

$$\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p(x)}[\log(1 - D_Y(G(x))] \tag{2}$$

Following CycleGAN, we couple the mapping G with an inverse mapping $F : Y \to X$ and introduce a cycle consistency loss to enforce $F(G(X)) \approx X$.

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p(x)}[||F(G(x)) - x||_1] + \mathbb{E}_{y \sim p(y)}[||G(F(y)) - y||_1] \tag{3}$$

Overall, the full objective is

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F) \tag{4}$$

where λ controls the relative importance of the two kinds of loss. We aim to solve

$$G^*, F^* = arg \min_{G,F} \max_{D_x, D_y} \mathcal{L}(G, F, D_X, D_Y), \tag{5}$$

By optimizing the Equation 4, we can obtain G^* and F^*, i.e. the style transfer models.

3.3 Poission Image Editing

To obtain seamless combination of the outputs segments from the transferring models, we adopt the Poisson Image Editing at the last pass of our model. Take I_A as the result of the transferring object and I_B as the result of the transferring background, the blend image I is the solution to the partial differential equations:

$$\nabla^2 I = \nabla^2 I_A \tag{6}$$

$$I|_{\partial\Omega} = I_B|_{\partial\Omega} \tag{7}$$

The examples shown in Figure 2 illustrates the effectiveness of the smoothing technique.

3.4 Segmentation Enhanced Style Transfer

Overall, we summarized our proposed model, as shown in figure 3.4. We conclude the process as the following steps,

- Given the input X, we first use L_0-smooth to sharpening the major edges while eliminating a manageable degree of low-amplitude structure, then use Image Segmentation to get the mask M for the object(s) to be transferred.

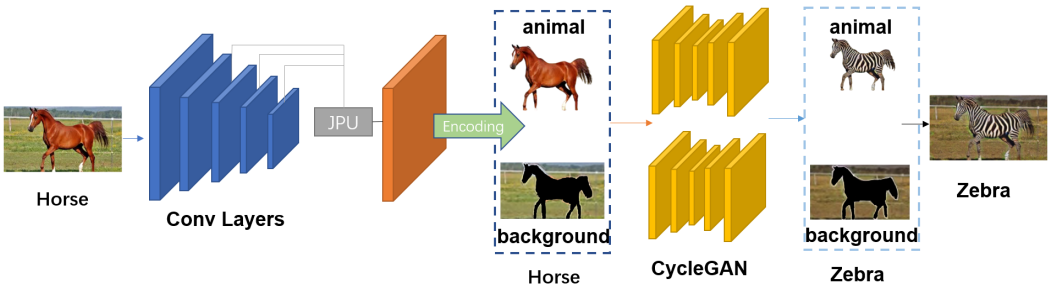

Figure 3: **Framework Overview of Our Method.** Our method applies the FastFCN model serving as the segmentation model. After separating pictures into foreground and background, we utilize CycleGAN to do style transfer.

- Split X into X_1 and X_2 under the mask M.
- Take the animal-background image for example, we process the image X_1 with the style transfer model for animals, which results in the output Y_1. Then, we generate the transferred background output Y_2 with the background-model with the input X_2.
- With the mask M and Poisson Image Editing, we can merge Y_1 and Y_2 to obtain the image of high quality with changed style.

4 Experiments

In this section, we have provided the details and results of our experiments on Pascal Context dataset, which have shown the validity of our methods compared to the baselines.

4.1 Experimental Settings

We use ResNet101 as encoder and Encoding head as decoder. We use L-0 smooth with the default parameters in [13]. For the Semantic Segmentation, we use FastFCN as the segmentaion model, with the default parameters. The pretrained model can be obtained from our open-sourced codes. Specifically, two pretrained domain mappings "*Horse→Zebra*" and "*Summer→Winter*" are also included in our repository. All the experiments are conducted on a single NVIDIA Tesla V100 (32GB).

4.2 Overall Results

As the quantitative metric of style transfer task is less explored, we report the qualitative results as shown in Figure 4. Several conclusions can be drawn from the results.

- **The baselines are susceptible to confusion:** From Column II of the left block, we can find that the transferred images are defective where the sky and grass are mixed with stripes. As for Column III, the horses are also negatively affected by the background transfer model, where the blue shadow appears on the horse object.
- **Context-aware Cycle-GAN generates high-quality images with changed styles:** In constrast, the results generated by our framework is much better. By Observing Column IV and Column V, we can see the background and the animals are well transferred and are consistent with our intuition. The animals and the backgrounds are individually independent w.r.t the style transfer process.

5 Conclusion and Future Work

In this work, we investigate the question that "How can the semantic information assist style transfer task". We proposed a framework that consists of the semantic segmentation and the style transfer model, which is easy to plug-and-play on various transfer models and can generate images of high

Figure 4: Overall Experimental Results. In the left part, the first column is original images. The second and third column are the results of the vanilla Cycle-GAN from two domain mappings: "horse→zebra" and "summer→winter", respectively. The fourth and fifth columns are the corresponding results of our model ECCG. The sixth column is the outputs after poisson image editting.

quality. The Possion Image Editing technique can further guarantee the seamlessness merge of the individually transferred images. Empirically, the enhanced Context-aware Cycle-GAN outperforms the state-of-the-art methods. As for future work, we are considering utilize more complicated semantic segmentation models to empower our methods. Also, the flexibility of our framework can be improved via more fine-grained grouping of image segments.

6 Every member's contribution

In this section, we will summarize the contribution of each member of our group.

- **Wu Yinxin**: (1) Clone the code from the open-sourced FastFCN project and then use the project to generate reasonable semantic segmentaion results. (2) Write the Semantic Segmentaion part of the report.
- **Chen Zeyuan**: (1) Add poisson image editting after the combination of independently processed animals and the background, in order to smooth the boundaries of different parts.(2) Write the Poisson Image Editting of the report.
- **Wu Yuefan**: (1) Collect the data of horses, dogs, cows; separate the classes in FastFCN into animals and backgrounds; (2) Draw the model structure and the whole process of style transfer.
- **Wang Yu**: (1) Clone the code from the open-sourced CycleGAN and successfully ran the project, then use CycleGAN to generate the final results taking the output picture from semantic segmentation as the input. (2) Write the CycleGAN part of the report.
- **Chen Ze**: (1) Add L-0 smooth before the segmentation, in order to make the segmentation easier and clearer; (2) Write the framework of our report, including the introduction.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[2] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346, 2001.

[3] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1033–1038. IEEE, 1999.

[4] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[5] Shuyang Gu, Congliang Chen, Jing Liao, and Lu Yuan. Arbitrary style transfer with deep feature reshuffle. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8222–8231, 2018.

[6] Jan Eric Kyprianidis, John Collomosse, Tinghuai Wang, and Tobias Isenberg. State of the" art": A taxonomy of artistic stylization techniques for images and video. *IEEE transactions on visualization and computer graphics*, 19(5):866–885, 2012.

[7] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[8] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Trans. Graph.*, 22(3):313–318, 2003.

[9] Alexander G Schwing and Raquel Urtasun. Fully connected deep structured networks. *arXiv preprint arXiv:1503.02351*, 2015.

[10] Ahmed Selim, Mohamed Elgharib, and Linda Doyle. Painting style transfer for head portraits using convolutional neural networks. *ACM Transactions on Graphics (ToG)*, 35(4):1–18, 2016.

[11] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yu Yizhou. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. In *arXiv preprint arXiv:1903.11816*, 2019.

[12] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yizhou Yu. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. *CoRR*, abs/1903.11816, 2019.

[13] Li Xu, Cewu Lu, Yi Xu, and Jiaya Jia. Image smoothing via l 0 gradient minimization. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pages 1–12, 2011.

[14] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015.

[15] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[16] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society, 2017.