Image Style Transfer based on DeepLabV3+ Semantic Segmentation

Jixiang Zhang *

Jiyang college of Zhejiang A&F University, Zhejiang, China

* Corresponding author email: 15030440234@xs.hnit.edu.cn

Abstract. Image style transfer has always been a popular research method in the computer vision community, which aims to learn the style of a given image and the content distribution of other images to generate new images with both the above style and content. Thanks to the rapid development of convolutional neural networks, the accuracy, and visualization of image segmentation and migration have been continuously improved, but there is still the problem of local content distortion. Some recent works introduce the segmentation branch to obtain pixel-level content information in order to obtain a more perfect transfer effect. In this paper, we use the VGG19 convolutional neural network model to extract high-level feature maps representing image content information and perform image style transfer based on the DeepLabV3+ semantic segmentation network, which can move images to the same or similar semantic regions during the entire transmission process. In order to prevent content distortion after image migration, we also introduce an affine function to control the content change of the image during image migration. Extensive experimental results show that the method in this paper can further clarify the segmentation boundary and improve the semantic accuracy of the transferred image. In addition, we also evaluated the expectations of different individuals on the degree of style transfer. By distributing the survey, this paper demonstrates that the method described in this study results in better image quality, more in line with the expectations of the majority of respondents.

Keywords: Style Transfer; DeepLabV3+ Semantic Segmentation; Affinef Unction.

1. Introduction

Image style transfer is really an innovative method that individuals have utilized in an attempt to combine the aesthetically pleasing aspects of learning with the more in-depth aspects of the subject matter. The aim of image transfer learning will be to maintain the content information and structure of an original content picture while converting the style of a content image into the style that we wish. This is accomplished by transferring the theme of a content image. As an outcome of the research on image style transfer techniques, people now have a more comprehensive understanding of the potential applications of deep learning. As a consequence of this research, the prospects for the development and application of deep learning algorithms have also been significantly improved.

Gatys et al. [1] successfully combined a convolutional neural network with an artistic flair for the very first time. As the first attempt, Gatys used an image that had random noise. After that, they made use of the Gram matrix[2] to model and extract the style that was present in the raw picture as well as in the style image. After that, they carried out a gradient matrix application. The Gram matrix that represents the starting graph may be brought closer to the Gram matrix that represents the style graph by using the descent method. In an effort to realize the degree of content similarity that has been specified, the subset of features of the noisy base picture is made as close as is practically feasible to the feature space of the target image. The style of the noise picture becomes equivalent to the style of the conveyed style photo, and the noise picture's content becomes comparable to that of the content image when iteration is performed on the noise picture. In 2017, based mostly on project of Gatys, Luan et al. [3] added a DeeplabV2 [4] semantic segmentation network to perform semantic segmentation of style images and content images and affine functions to constrain the change of content structure during the migration process, which solved the problem to some extent. This was accomplished by adding the DeeplabV2 semantic segmentation network to perform semantic segmentation of style images and content images and by adding the affine functions. By including

the DeeplabV2 semantic segmentation network, this goal was successfully realized. However, the semantic segmentation network of Fujun Luan could not obtain a large enough field of view, so multi-scale information would be missing; and the image transferred by the method of Fujun Luan still has the problem of incomplete style transfer and missing content, which we will see in the later experimental results. The method of Gatys applied the problems of content distortion and inaccurate style transfer in realistic style image transfer.

The problem of incorrect style transfer in the transferred picture was one of the reasons why the DeepLabV3+[5] semantic segmentation network was added to the image migration process that was discussed in this study. This was done with the purpose of fixing the issue. For instance, the design of the sky and the grass cannot be interchanged in any way. As a consequence of this, the path that the migration process will take is chosen once the semantically segmented picture of such content imagery and the theme image has been acquired. Each contextual label in the segmented picture is used to transfer the style of the region with the closest semantic information, which mainly respects the scene. All of that is done in order to preserve the integrity of the segmented image. To do this, the theme of such area that contains the most semantically relevant information is copied over to the target region.

As we have seen, the approach developed by Gatys[1] is not appropriate for the transfer of a realistic picture style, and the vast majority of the already available methods for the transfer of style are used for the transfer of an artistic style. In 2021, Xu et al. [6] used GAN to the process of visual style transfer. When applied to the process of image style transfer, numerous understandings may be formed on multiple creative levels of a portfolio, but the end result is still an artistic style transfer. However, the issue of content distortion arises when these technologies are used to the process of photo-realistic style transfer. In response to this issue, we recommend using the affine function described in this work during the migration process in order to directly map the color to the relevant location without modifying its structure. This would allow us to solve the issue of content distortion that has been brought up.

2. Related Work

As we have discovered, Luan et al. used the DeepLabV2 semantic segmentation network to image style transfer, which carried out semantic segmentation on both the source photos and the style photograph at the same time. Its image semantic segmentation branch is built on top of an extension of the depth wise separable convolution (DeepLab [7]), which employs an encoder-decoder architecture and integrates an Atrous Spatial Pyramid Pooling (ASPP) module into the decoder component of the network. Both of these elements are part of the network. The ASPP module allows the network to operate at different scales, which enables it to capture both fine and coarse features in the images that it analyzes. This is accomplished by the network's ability to capture both fine and coarse characteristics in images. By Liang-Chieh Chen and his colleagues, DeepLabV2[4] was first introduced to the public. The very first model was trained on the VOC[8] dataset, and after training it obtained a result of 79.7% mIOU.

The DeepLabV3+[5] algorithm then was proposed. The input image is sent to a depth-separable convolutional network in DeepLabV3+ so that it may be processed during the phase of the process known as feature extraction. Convolution with depth-separable is a great way for lowering the total amount of computations while yet preserving the ability to represent features accurately via the use of the convolutional operation. After this step, an Atrous Spatial Pyramid Pooling (ASPP) module is applied to the feature space throughout to successfully carry out multi-scale information fusion on the data contained inside it. The ASPP module makes use of a large number of convolution kernels, each of which grows at a different rate than the others in order to carry out the process of convolution. Because of this, the module is able to gather data on a wide variety of scales. Finally, in order to acquire the final result, a classification is performed by using a convolutional layer, and the output features of ASPP are fused with the initial characteristic drawing by utilizing a convolutional layer.

This is done so that the final result may be obtained. Because DeepLabV3+[5] offers such exceptional performance, this specific piece of study is able to take use of it. It obtained results with a mIOU of 89% on the VOC data set, and it is capable of high resolution as well as multi-scale prediction owing to the use of depth separable convolution kernel ASPP.

True beauty transfer and amazingly reconstruction were both accomplished via the use of a method that was introduced in 2016 by Justin Johnson and his colleagues[9]. This method makes advantage of perceptual loss. Its strategy was devised throughout the goal of completing the style transfer. This method is able to solve the problem that the Gatys[1] theory has a migration time that is too long, it is able to accomplish style transfer in a short amount of time, and it has achieved very excellent results in the resolution of the picture after migration. These benefits can be attributed to the fact that this method: Batch Normalization is a process that emphadimensions with normalizing per batch to guarantee that there is consistency with in distribution of the data across each batch. Batch Normalization guarantees that each batch is normalized to ensure that the data distribution within each batch is consistent. Because the outcomes of the model's training rely upon that entire distribution of data in each batch, Batch Normalization ensures that each batch is normalized. Batch Normalization, on the other hand, is not an acceptable solution for the problem of style transfer since the result that is achieved depends a great deal on each individual picture instance. This is due to the fact that batch normalization processes several images in a single pass. As a consequence of this, normalizing a single instance might potentially speed the convergence of the model while still maintaining the independence of each image instance. In the same year, Dmitry Ulyanov and his colleagues[10] made the suggestion that Instance Normalization should be used instead of Batch Normalization in order to make a significant improvement in the quality of the feed-forward style transfer network model. This was done in order to achieve the goal of making the feed-forward style transfer network model more accurate. In 2021, Pei Wang[11] carried out study in order to discover the distinctions between ResNet[12] and VGG[13] in terms of the style transfer work. On the basis of their results, they proposed a new method which they referred to as SWAG (Stylization with Activation smoothingG). The Gram Matrix, which performs a softmax operation on the characteristic drawing, is an essential part of the style transfer that is investigated in this work. This is an illustration of the SWAG technique, the purpose of which is to achieve the goal of "making the small value higher and lowering the peak value." As a direct consequence of this research, the effectiveness of a variety of well-known stylization algorithms, such as ResNet and Inception, has been elevated as a direct result. Second, it improves the performance of stochastic networks for certain designs to the same level as that of pretrained networks. In conclusion, when compared to the performance of stochastic networks, the results of non-VGG models with SWAG may even be superior than those obtained by VGG[13] trained in the normal fashion for pretrained networks.

This study is about image style transfer with realistic style, although the works that are linked to it that were discussed previously are all about image transfer research in terms of artistic style.

3. Methods

The visual style transfer approach that was employed in this work may be broken down into two different components. The first module is a semantic segmentation network called DeepLabV3+[5]. It uses both the raw picture and the style image to conduct semantic segmentation. The second part of the system is called the style transfer module. In order to transfer actual photographs, we use the two different types of semantic segmentation maps that were created by the first module and use them in the process of style transfer.

3.1 DeepLabV3+ Semantic Segmentation Module

3.1.1 Backbone

The backbone adopted in DeepLabV3+ semantic segmentation module is MobileNetV2 [14], which is a type of a tiny deep neural network that Google designed primarily for use in embed phones

and other mobile devices. The first version of MobileNet has been upgraded to become what is known as MobileNetV2 at this point. One of its most important characteristics is that it makes use of inverted resblock, which is a technique. As the Figure 1 shown, the MobileNetV2 architecture is composed entirely of the inverted resblock in its current iteration. Inverted resblock may be divided into two sections: the major portion on the left can have its dimension enhanced by starting with convolution, then moving on to profundity dissociable convolution for feature extraction, and then returning to convolution to reduce its dimension. The leftover edge component is located on the right side, and both the input and the output are directly linked. When the process of extracting features for MobileNetV2[14] has been completed, we will have the ability to acquire two feature layers that are efficient. Compressing the shape and dimension of the input photograph twice yields one effective feature layer while compressing the altitude and breadth of the input photograph four times yields the second effective feature layer. Both of these effective feature layers are the result of effective compression.

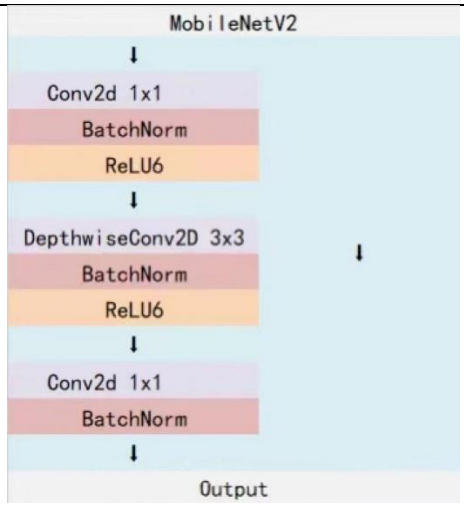


Fig 1. The structure of MobileNetV2

3.1.2 Enhanced Framework for Feature Extraction

As shown in Figure 2, the improved feature extraction network seen in DeepLabV3+[5] may be broken down into two distinct components: In the encoder, we are going to use parallel Atrous Convolution for the preliminary effective feature layer, and then we are going to use various rates of Atrous Convolution for feature extraction, and then we are going to combine those two sets of results together. The 1×1 convolution compression feature should then be performed. In the Encoder section of DeepLabV3+, a significant number of expansive convolutions have been added. This has the effect of expanding the field of view without sacrificing any details, thereby increasing the output per each convolution including a wider variety of data.

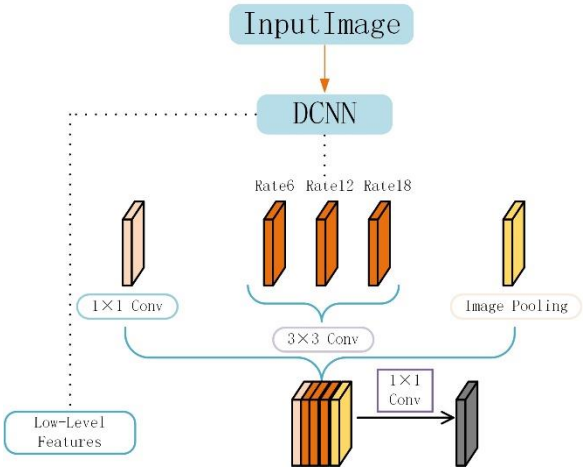


Fig 2. The encoder and decoder architecture in DeepLabV3+

In Decoder, we will use 1×1 convolution to adjust the number of channels on the preliminary effective feature layer compressed twice, and then stack the result of upsampling the effective feature layer after hole convolution. After the stacking is completed, two depth-variable Separate convolutional blocks. At this time, we have obtained a final effective feature layer, which is the feature concentration of the entire picture. At this point, we need to use features to obtain prediction results. The process of using features to obtain prediction results can be divided into 2 steps: (1) Use a 1×1 convolution to adjust the channel and adjust it to Num_Classes. (2) Use redimension to up sample so that the final output layer has the same width and height as the input image.

3.2 Style Transfer Module

As can be seen in Figure 3, the VGG19[13] convolutional neural network is the one that we use in the picture migration module for the purpose of characteristic drawing extraction. The convolutional neural network represented by the VGG19 may be broken down into 5 separate Blocks, each of which incorporates convolution operations but varies in the amount of convolution cores it uses. The picture being processed as input travels through each Block, which then extracts information from the image via a convolutional technique. Poor convolutional layers are responsible for extracting the look of the picture, whereas strong convolutional layers are responsible for extracting the photo's actual content. Different from VGG16 in that a convolutional layer is added to Blocks 3, 4, and 5 to enable the extraction of additional feature information. The sum of convolution cores inside each block will be increased, but their size will remain same. Following convolution, greatest pooling will be carried out. The objective of greatest pooling is to retain the initial properties of a neural network while minimizing its training parameters, hence saving training time. At the conclusion of the VGG19 structure are three layers with complete connections and a softmax algorithm. The initial two completely linked layers consist of 4096 neurons, whereas the last fully connected layer has 1000 neurons. The fully connected layer is accountable for transforming the two-dimensional characteristic drawing generated by the convolution into a one-dimensional tensor, so achieving the end-to-end learning process. The last fully linked layer produces one thousand probabilities, the total of which is one. The Softmax function selects the highest likelihood among these 1000 possibilities and uses it as the full VGG19's prediction result.

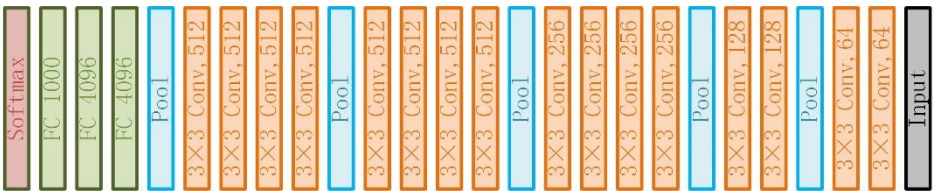


Fig 3. The VGG19 structural diagram

3.3 Loss Function

The affine function in this paper is to map the color of a certain area of the style image to the original image, avoiding the distortion of the image content after migration, as shown in Figure 9. The affine function in this paper is similar to the method of Fujun Luan[3]. This strategy is not expressed by adding constraints directly on the output image, but on the transformation applied to the input image. Our strategy is to ensure that its properties are not lost during style transfer by adding a penalty term related to image warping. Our solution is to find a local affine transformation of an image in a color space. For each output block, there is an affine function that maps the RGB values of the input image to the corresponding output positions. The affine function is different for each region and varies spatially. We give the affine function as:

$$\mathscr{L}_m = \sum_{c=1}^{3} V_c[O]^T M_l V_c[O] \tag{1}$$

As the Figure 4 shown, by using the affine function, we are able to assign the RGB value of the style picture to the region of the original image that corresponds to it. The affine function is not applied directly toward the input picture; rather, it is applied in the context of the photo migration

project using the affine function. due to the constraints placed on it, the initial image's local structure has been retained quite well.

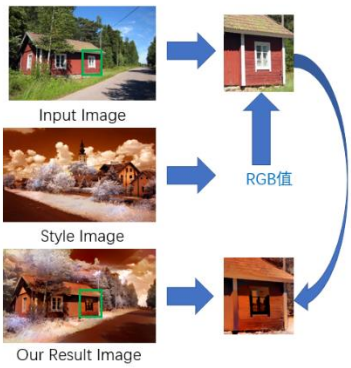


Fig 4. The assignment of RGB values from style to the content based on an affine transformation

Administer semantic style on both the raw photo and the theme photo at the same time using the DeepLabV3+[5] semantic segmentation approach, and then color-label the various semantic areas of both images. The number of colors that are utilized is determined by the value of such mask, which should be added as another channel to the source images. The following method should be used to update the style loss:

$$\mathcal{L}_{s+}^{\ell} = \sum_{C=1}^{C} \frac{1}{2N_{\ell,C}^2} \sum_{ij} (G_{\ell,C}[O] - G_{\ell,C}[S])_{ij}^2 \tag{2}$$

$$F_{\ell,C}[O] = F_{\ell}[O] M_{\ell,C}[I] \tag{3}$$

$$F_{\ell,C}[S] = F_{\ell}[S] M_{\ell,C}[S] \tag{4}$$

The following is how to calculate the total loss using the approach described above:

$$\mathcal{L}_{total} = \sum_{l=1}^{L} \alpha_{\ell} \mathcal{L}_{C}^{\ell} + \Gamma \sum_{\ell=1}^{L} \beta_{\ell} \mathcal{L}_{S+}^{\ell} + \lambda \mathcal{L}_{m} \tag{5}$$

4. Experiments and Results Analysis

4.1 Dataset

We provide the results of our experiments utilizing the DeepLabV2[4] and DeepLabV3+[5] algorithms on the semantic segmentation dataset VOC[8] in the Experiments section of this paper. The VOC Dataset is a big data set that includes 21 different target components. The training set of this data set has 17,125 high-quality photographs, the verification set has 500 photos, and the test set has 1,449 photos. By comparing the estimated of the two approaches, our findings demonstrate that DeepLabV3+ is more effective than the other way for semantic segmentation. In our reports, we not only provide the outcomes of our experiments pertaining to semantic segmentation, but we also discuss the outcomes of our studies pertaining to style transfer. We demonstrate that the good results obtained using DeepLabV3+ semantic segmentation assist our style transfer in producing more realistic pictures by comparing them to the results obtained by Fujun Luan and Gatys. We instruct and analyze every one of our models on Ubuntu Server 18.04 LTS 64-bit using NVIDIA TeslaT4 graphics processing units.

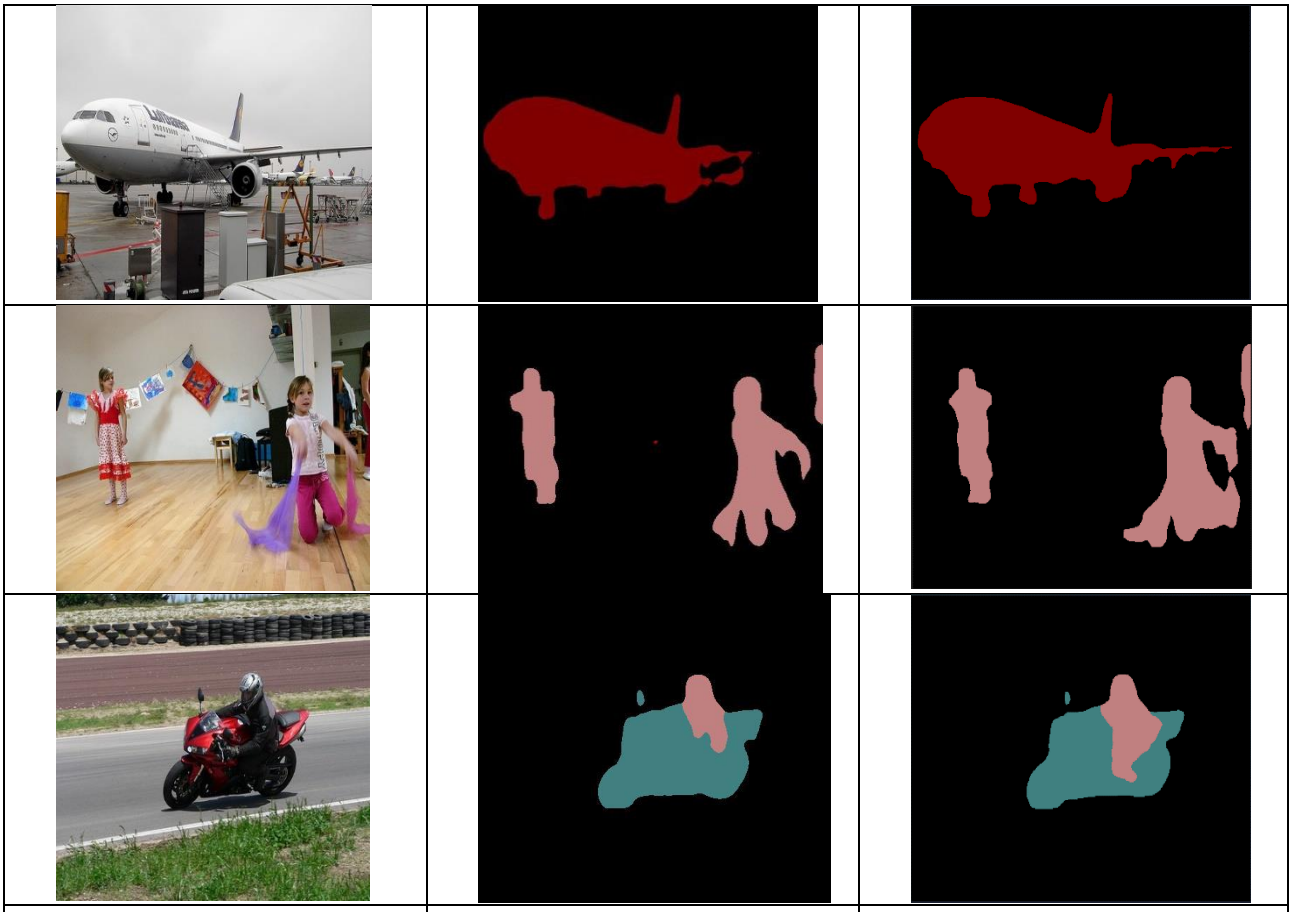
4.2 Experiment Settings

In the semantic segmentation experiment that was carried out using DeepLabV3+, the weights of the MobileNetV2[14] model that we used were pre-trained weights. The value of 8 that we choose for the down sampling factor, and the training phase is split into two parts: the freezing phase and the unfreezing phase. During the freezing period, the primary structure of something like the prototype is frozen, the feature extraction network does not change, the amount of video card memory that is used is minimal, and the network receives only fine-tuning at this point. During the stage known as "unfreezing," the backbone of the model is not frozen; the feature extraction network will change; it

will take a significant amount of video memory; and all of the parameters of the network will change. We use a batch dimension value of 8 during the training process of the freezing phase, but during the unfreezing phase, we use a batch dimension value of 4, which is half of the freezing phase. This is due to the fact that training during the unfreezing phase consumes more video memory, and all network parameters will change during this phase. We optimized the model parameters by using the SGD as we were going through the training process. The gradient descent approach presented in this study only makes use of a single training sample throughout order to modify the model parameters throughout the iterative process. For the purpose of the style transfer experiment, we make use of the VGG19[13] model to abstract characteristic drawing from the photographs that we have taken. The weights that are used by the VGG19 network model are also pre-trained weights.

4.3 Performance Analysis for Semantic Segmentation

In Figure 5, we compare our semantic segmentation results with those of previous work[3]. During the semantic segmentation process, we made use of DeepLabV3+'s [5] encoder-decoder technique, and thereafter, we compared our findings to those obtained by DeepLabV2[4] during the same process. The findings of DeepLabV2's semantic segmentation do have significant issue with improper semantic partition, and they also reveal an issue with semantic information that is missing from the results. When it comes to the level of detail, DeepLabV2 is nowhere near as good as DeepLabV3+. DeepLabV3+ is able to not only comprehend the nuances but also more effectively derive the general information.



(a) Input Image	(b) DeepLabV2 Results	(c) DeepLabV3+ Results

Fig 5. Comparison of segmentation result of different methods

4.4 Performance Analysis for Style Transfer

Figure 6 depicts a contrast between the pictures that were shifted using the approach described in this research and such photos that were shifted using the methodology described by Gatys[1] and Fujun Luan[3].

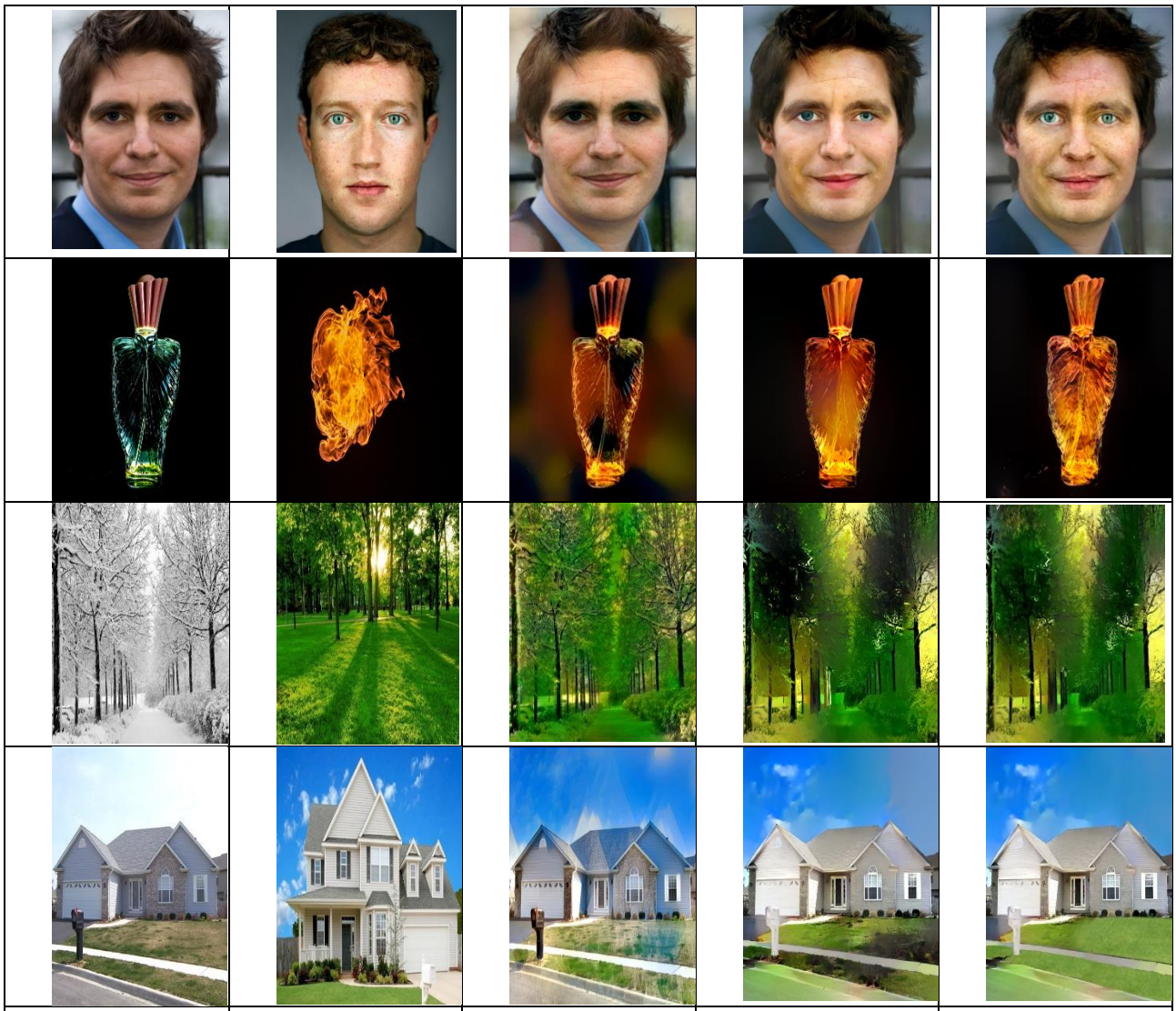


Fig 6. The style transfer results of different methods

Following the completion of the prior semantic segmentation, we now have the images of the semantic segmentation for both the raw photo and the design aesthetic image. Our goal was accomplished by applying semantic segmentation to a picture in order to do the style transfer, and it was successful. We contrast our findings to those obtained by Gatys and Fujun Luan, and we show that the outcomes obtained using our methodology are superior to those obtained using the other two methodologies.

In addition, we released the results of all these methods. After that, we compiled the results of a poll about our anticipations regarding the standard of migration. Figure 7 presents our findings, which were derived from an investigation in which forty individuals were polled on their opinions on the credibility of the migration data:

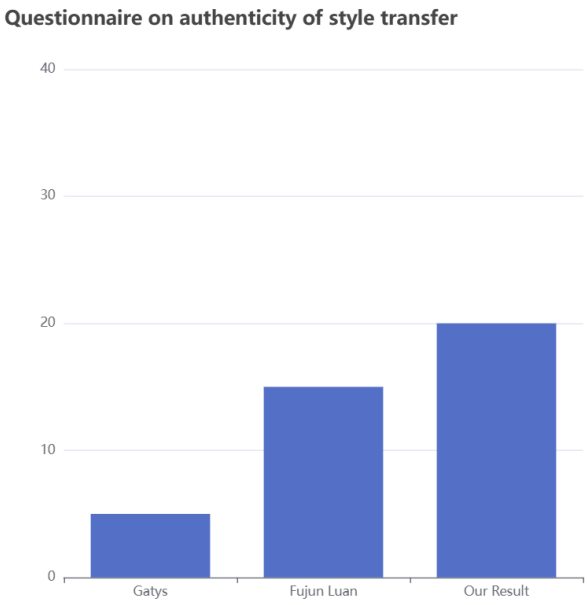


Fig 7. Perception of the trustworthiness of the migration data

The number of persons who are happy with the three techniques is represented by the ordinate, and the abscissa describes the author of each of the three approaches. The total number of respondents was forty. Those who are satisfied with Gatys'[1] images are the consequences of artistic style transfer, individuals who are comfortable with Fujun Luan's[3] images are the consequences of realistic style. However, when people compare with the transfer method that utilizes the same realistic way, more people will be thrilled with the results of just this essay.

Since the artistic style transfer mechanism is indeed not acceptable for realistic style transfer, the Gatys method has the lowest number of people who are satisfied with it. On the other hand, the Fujun Luan methodology employs the DeepLabV2 semantic segmentation algorithm, which means that the migrated image still has style images in some content that have style transfer issues that are incomplete and inaccurate color transfer issues. In this study, the much more detailed semantic segmentation method as well as a more specific semantic region division are used to achieve more specific color coverage of the style image as well as more complete style transfer. This helps to avoid the issue of some regions being successfully migrated while other regions are unable to be successfully migrated. The approach presented in this research has significantly increased the authenticity of picture transfer when done in a realistic manner, and it has garnered the majority of positive feedback.

5. Discussion

Throughout this essay, by contrasting the findings of semantic segmentation and the migration effect, it is proven that DeepLabV3+[5] is capable of bringing a more precise and reliable semantic segmentation effect, that the perception of segmentation between different semantics is stronger, that the perception of demarcation line is clearer, and that it can bring realism style photo type image migration to more realistic results. Convolutional neural networks like the VGG19[13] have a stronger network structure, which allows them to retrieve richer aspect details from the images they process. It is able to recreate the style of the target picture more accurately while keeping the content structure of the photograph intact. It is able to handle the nuances of the theme transfer process better than the standard approaches that have been used in the past. In conclusion, with respect to the various people's anticipations of the outcomes of the image style transfer process, we have shown that the impact of our transfer is more in accordance with the anticipations of the majority of people than are those of Fujun Luan[3]. The majority of the research that have been done on image style transfer so far have focused on artistic style image style transfer. This work presents a novel concept for the

research of image style transfer, specifically focusing on the image style transfer of realistic style photographs.

When considering practical applications, our method can be used in advertising design, artwork restoration, video style transfer, etc. Since the image style transfer method in this paper is for image style transfer of photo-realistic style, it can be conveniently helped Photographers post-process photos to meet photographers' different needs for photo styles. Our method not only improves image quality, but also meets style-specific requirements in many applications. The DeepLabV3+ semantic segmentation method[5] in this paper can also be used in automatic driving systems to help vehicles recognize surrounding roads, traffic lights, obstacles, etc. In addition, DeepLabV3+ can also be used for medical image analysis to help doctors identify lesions or abnormalities in the body.

In future research, we plan to continue to explore more possibilities of realistic image style transfer methods, and try to apply realistic image style transfer in more fields. We believe that through our next exploration we can transfer the realism of the after image meets the expectations of more people. Through the research in this paper, we have also seen the excellent segmentation results of DeepLabV3+, and we also expect to explore more exciting methods in the field of image semantic segmentation, improve the accuracy of semantic segmentation, and make more contributions.

6. Conclusion

In this paper, we suggest using a convolutional neural network model developed with DeepLabV3+ and VGG19 to the task of realistic image style transfer. In order to acquire the semantic segmentation map of the raw picture and the style image, we make use of DeepLabV3+. Additionally, we make use of the VGG19 network model in order to acquire image characteristic information. We show that the results obtained with our technique are more realistic than those obtained with the existing methods. Image style transfer research has recently been moving in such a fresh methodological and theoretical direction, which is consolidated in this report's technique.

References

[1] Gatys L A, Ecker A S, Bethge M. A neural algorithm of artistic style[J]. arXiv preprint arXiv:1508.06576, 2015.

[2] Levin A, Lischinski D, Weiss Y. A closed-form solution to natural image matting[J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 30(2): 228-242.

[3] Luan F, Paris S, Shechtman E, et al. Deep image style transfer[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4990-4998.

[4] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[5] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[6] Xu W, Long C, Wang R, et al. Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 6383-6392.

[7] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.

[8] Everingham M, Eslami S M, Van Gool L, et al. The pascal visual object classes challenge: A retrospective[J]. International journal of computer vision, 2015, 111(1): 98-136.

[9] Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution[C]// European conference on computer vision. Springer, Cham, 2016: 694-711.

[10] Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: The missing ingredient for fast stylization [J]. arXiv preprint arXiv:1607.08022, 2016.

[11] Wang P, Li Y, Vasconcelos N. Rethinking and improving the robustness of image style transfer [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 124-133.

[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[14] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

[15] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation [J]. arXiv preprint arXiv:1706.05587, 2017.