Exploiting Human Face Segmentation for Improving Portrait Image Style Transfer

Hongfeng Lai

20215973@stu.neu.edu.cn

Department of Computer Science and Engineering, Northeastern University, Shenyang, Liaoning, 110169, China

Abstract. This research proposes a model to address the common issues of facial distortion and loss of details in style transfer of images containing faces. The model aims to achieve background stylization while preserving facial integrity and improving image clarity by reducing potential blurring during the style transfer process. To achieve this goal, firstly, a pre-trained CycleGAN model is used to perform Monet-style transformation on input images, with a focus on background stylization. Simultaneously, a pre-trained GFPGAN model is employed to enhance the clarity of facial regions in the images. The EasyPortrait dataset is used for model training to find the optimal parameter settings that minimize the loss function, resulting in the lightweight facial segmentation model. and the performance of the pre-trained Segment Anything Model (SAM) model is compared with the self-trained lightweight EasyPortrait model. Based on the evaluation results, the model with superior performance is chosen as the primary facial segmentation model. Subsequently, the high-resolution facial images obtained are fed into the facial segmentation model. Finally, the stylized facial regions are replaced with high-resolution facial images, achieving background stylization and facial enhancement in the original images. In summary, this work aims to provide a novel solution to the challenges of style transfer in images containing faces, enabling visually appealing and faithful transformations while preserving facial integrity. The proposed method holds potential applications in artistic image processing, creative design, and digital content creation

Keywords: Deep learning; Image style transfer; Image segmentation.

1. Introduction

In the field of digital image processing, style transfer has emerged as a captivating technique for transforming images from one domain to another, imbuing them with distinct artistic styles [1,2]. However, applying style transfer to images containing human faces often poses a common challenge, resulting in undesirable distortions and loss of facial details.

This work aims to address this issue by developing a model that achieves style transfer while preserving the fidelity of facial features, with a specific focus on background stylization while maintaining facial integrity. Additionally, efforts are made to enhance image clarity by mitigating the blurriness that may arise during the style transfer process.

By tackling these challenges, the goal of this research is to advance the field of style transfer for human faces, enabling the creation of visually appealing images with preserved facial details and improved overall quality.

ICIDDT 2023, November 03-05, Zhenjiang, People's Republic of China

DOI 10.4108/eai.3-11-2023.2342219

Pretrained large models are typically trained on large-scale datasets, which provide them with a broader range of language and visual knowledge [3,4]. This enables them to have better generalization capabilities and wider applicability across domains. Rich and abstract feature representations are learned by pretrained large models through self-supervised or supervised learning on massive amounts of data. These feature representations can be utilized for various downstream tasks, reducing the need for manual feature engineering.

On the other hand, custom-trained lightweight models can be designed and adjusted according to specific task requirements, offering greater flexibility. Specific architectures, layer configurations, and parameter settings can be chosen to adapt to particular data and tasks. Lightweight models trained from scratch often have smaller model sizes and parameter counts, making them more deployable and efficient in resource-constrained environments. They are suitable for deployment on mobile devices or edge computing scenarios.

In summary, pretrained large models provide extensive language and visual knowledge, superior generalization abilities, and reduced manual feature engineering. Conversely, custom-trained lightweight models offer flexibility, adaptability, and efficient deployment in resource-limited settings.

To achieve the goal of this work, a two-step approach is proposed. Firstly, the CycleGAN pre-trained model is utilized to perform Monet-style transformation on the input images, focusing on background stylization [5]. Concurrently, the GFPGAN pre-trained model is employed to enhance the clarity of facial regions in the images [6]. The resulting high-definition facial images are then subjected to a facial segmentation model, where the performance of the Segment Anything Model (SAM) pre-trained model is compared with the own lightweight model, EasyPortrait [7,8]. Based on the evaluation results, the superior model is selected as the primary facial segmentation model. Finally, the stylized faces are replaced with the high-definition facial images, thereby accomplishing both background stylization and facial enhancement for the original images.

In this paper, the design, implementation, and evaluation of the proposed approach for style transfer with preserved facial details and enhanced image clarity are presented. The characteristics and advantages of the pre-trained models employed in this research are also discussed, highlighting their contributions to the methodology.

Overall, this research aims to provide a novel solution to the challenge of style transfer on images with human faces, offering a more visually appealing and faithful transformation while preserving facial integrity. The proposed approach holds potential applications in various domains, including artistic image manipulation, creative design, and digital content creation.

2. Method

2.1 Overview

In this work, the overall work flow is demonstrated in Fig. 1. First, CycleGAN is applied to stylize the original image. Subsequently, GFPGAN is utilized for facial enhancement, and then facial segmentation is performed on the enhanced images. A lightweight facial segmentation model is trained using the EasyPortrait dataset and evaluated against the pre-trained SAM model for this purpose. The best-performing model is selected as the facial segmentation model. Finally, the stylized background is combined with the enhanced facial region obtained through high-resolution restoration, resulting in background stylization and facial enhancement of the original image.

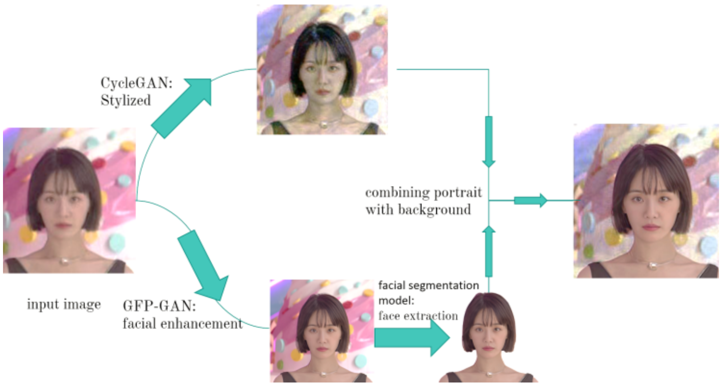

Fig. 1. Work flow of the proposed method (Figure Credits: Original)

2.2 CycleGAN

CycleGAN is a deep learning model used for image translation and transformation tasks, such as style transfer. It aims to learn the mapping between two different domains without the need for paired training data. In this research, CycleGAN is employed as the first step to achieve Monet-style transformation of images.

CycleGAN is characterized by its unsupervised learning capability and the principle of cycle consistency. Unlike traditional image translation methods, it does not require paired training data. Instead, it utilizes two generators and two discriminators in an adversarial training setup to learn the mapping between domains. The generators are responsible for transforming input images from the source domain to the target domain, while the discriminators attempt to differentiate between generated and real images. Through adversarial training, the generators can learn to produce realistic image translations, and the cycle consistency loss ensures bi-directional transformation.

The principle behind CycleGAN is rooted in generative adversarial networks and cycle consistency loss. Generative Neural Network (GANs) facilitate image translation by adversarial learning between generators and discriminators. The generators aim to generate realistic target domain images, while the discriminators strive to distinguish between generated and real images. Simultaneously, cycle consistency loss ensures the consistency and reversibility of the transformation by re-converting the translated images back to the original domain and comparing them to the original images.

CycleGAN offers several advantages. It eliminates the need for paired training data, thereby avoiding the difficulties and costs associated with manual annotation. It can automatically learn the mapping between domains, enabling broader generalization and adaptability. Additionally, the cycle consistency loss helps generate more accurate and consistent transformation results.

2.3 GFPGAN

In the second step, GFPGAN is used to enhance the resolution of the faces in the images. Next, the self-trained image segmentation model is compared with SAM (Spatial Attention Module) to select the best-performing image segmentation model for face segmentation after the enhancement.

GFPGAN is a deep learning model specifically designed for super-resolution tasks, aiming to generate high-resolution versions of low-resolution images. The model leverages a pre-trained facial GAN model to provide rich and diverse prior knowledge (provided by StyleGan2) [9]. The Generative Facial Prior (GFP) is incorporated into the facial restoration process through spatial feature transformation layers, achieving a balance between realism and high fidelity. By applying GFPGAN, the details of facial images can be enhanced, improving their visual quality.

2.4 EasyPortrait

In this work, an image segmentation model is trained on a new image dataset called EasyPortrait for portrait segmentation and facial parsing tasks.

The EasyPortrait dataset has a size of approximately 26GB and includes 20,000 RGB images (~17.5K full HD images) with high-quality annotated masks [10]. The dataset is divided into training, validation, and testing sets based on themes. The training set consists of 14,000 images, the validation set consists of 2,000 images, and the testing set consists of 4,000 images. It contains 20,000 primarily indoor photos from 8,377 distinct users, along with fine-grained segmentation masks divided into 9 classes. The subjects vary in age, race, gender, and emotion, mostly capturing photos at home using laptops or smartphones, sometimes in offices or on the street. Each image is annotated with keypoints transformed into segmentation masks for portrait segmentation and facial parsing tasks (as cited in the paper).

The presented semantic segmentation network model is based on the UNet architecture and is designed for pixel-level segmentation tasks on images. Prior to utilization, the model undergoes a training process to obtain optimized weights. Fig. 2 shows the change of loss with the training process.

Initially, a dataset class named EasyPortraitInferDataset is defined to facilitate the loading of test images and their corresponding segmentation masks. This class utilizes the glob function to retrieve file paths for images and masks and provides the getitem() and len() methods for accessing samples within the dataset. Subsequently, the inference_model is created, implementing the UNet architecture and specifying the encoder and the number of classes. The encoder part of the model is trained and adjusted during the training process to adapt to the characteristics of the dataset. The pretrained weights are loaded into the inference_model using the torch.load() function, ensuring their appropriate application.

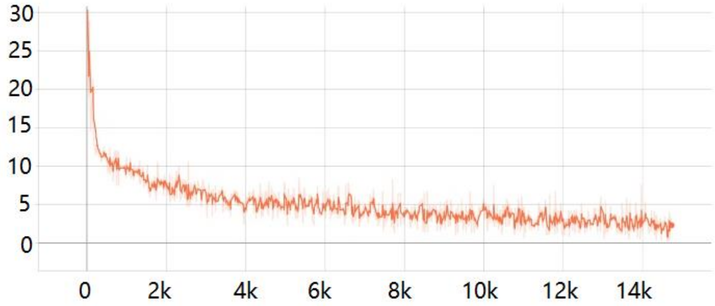

Fig. 2. Change of loss with the training process (Figure Credits: Original)

To ensure the preservation of model parameters during the inference process, the inference_model is set to evaluation mode.The code then iterates through the test dataset, performing inference on each sample. For each sample, the image and its normalized counterpart are obtained. The inference_model predicts the corresponding segmentation mask based on the normalized image. Finally, the original image, predicted segmentation mask, and ground truth segmentation mask are visualized, and the predicted results are saved as images.

This model has a parameter count of 10.05M, while SAM has approximately 635M parameters, which is 60 times larger than this model. As a lightweight alternative, this model can run on relatively low hardware configurations, reducing computational resource consumption.

2.5 SAM

SAM, on the other hand, is a widely used image segmentation method that combines spatial attention mechanisms to improve segmentation accuracy. It focuses on capturing spatial relationships and contextual information within the image to enhance segmentation performance.

By comparing the performance of the self-trained model and SAM, the segmentation accuracy, robustness, and ability to handle complex facial features and variations are evaluated. The model that demonstrates the best performance in terms of lightweight and accurate facial segmentation will be selected for further analysis and processing.

Overall, in the second step, GFPGAN is utilized to enhance the resolution of facial images, and the self-trained image segmentation model is compared with SAM to select the most effective model for facial segmentation. This comprehensive approach aims to improve the quality and accuracy of facial segmentation, which is crucial for subsequent analysis and applications.

3. Result

In the result section, the outcomes of the proposed approach for style transfer on images containing human faces are presented. The model aims to address the issue of facial distortion commonly encountered when applying specific stylization to images with human subjects.

The focus is on achieving stylization of the background while preserving the integrity of the facial region. Additionally, a high-resolution enhancement process is incorporated for blurry facial images.

To begin, a pre-trained CycleGAN model is utilized for Monet-style transformation of the images. Simultaneously, the GFPGAN model is employed to enhance the resolution of facial regions. The high-resolution facial images are then passed through a facial segmentation model, where the effectiveness of the pre-trained SAM model and the lightweight model, EasyPortrait, are compared.

Fig. 3 is a comparison of the segmentation results between SAM and the self-trained EasyPortrait model in this work.

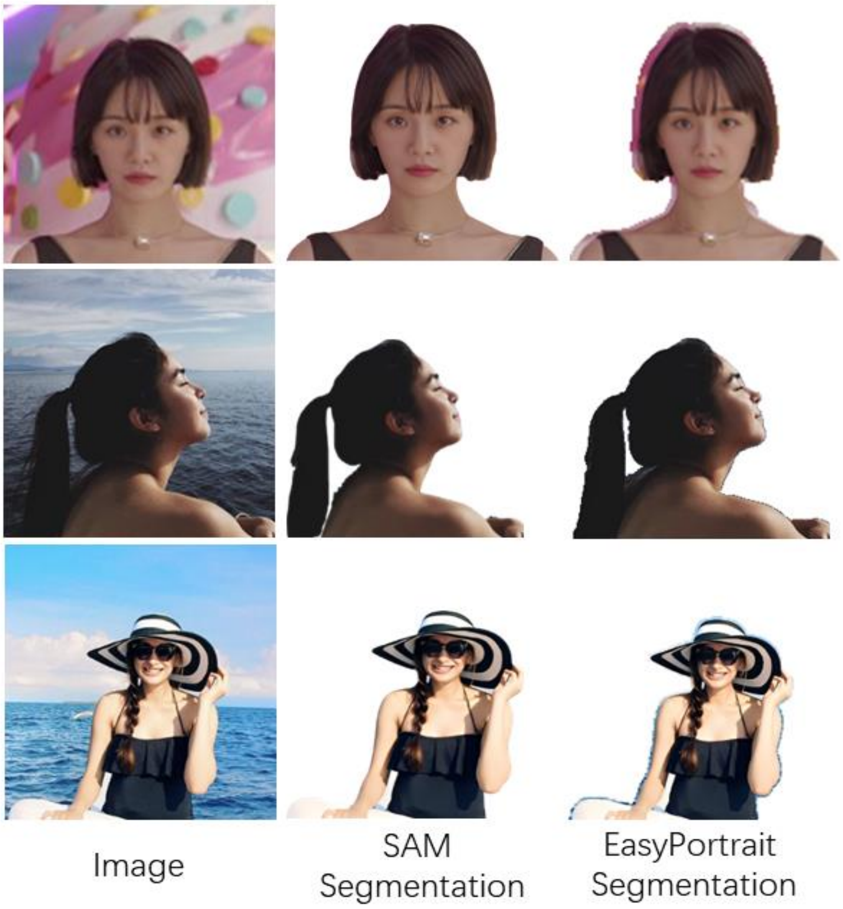

Fig. 3. Visual comparison of SAM and EasyPortrait segmentation results (Figure Credits: Original)

Fig. 4 is the final model comparison result. From left to right are the original image, direct stylization and enhancement, segmentation using SAM model, and segmentation using self-trained EasyPortrait model, respectively.

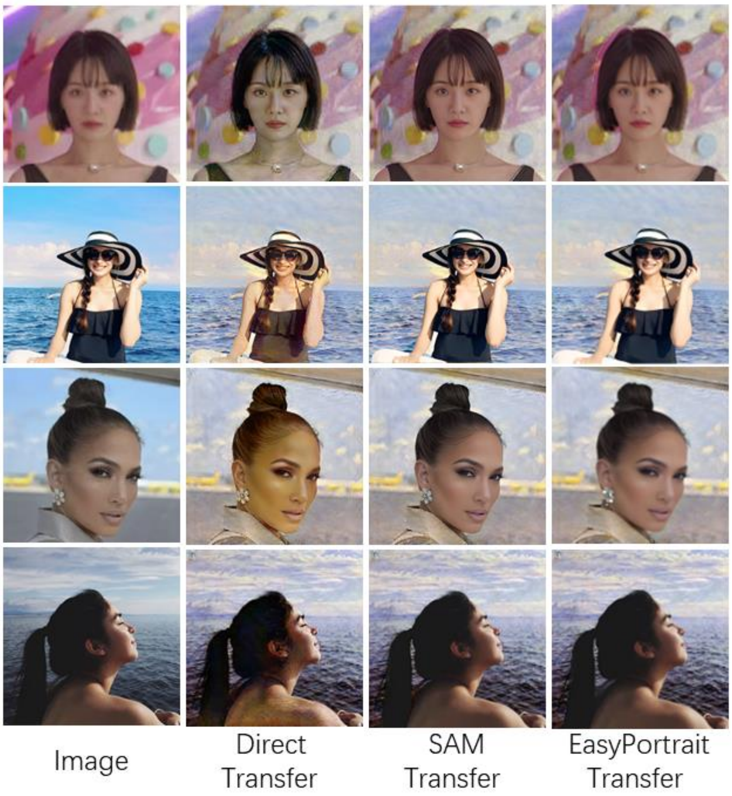

Fig. 4. Results comparison (Figure Credits: Original)

Based on the experimental results, it can be observed that using a face segmentation model effectively addresses the issue of distortion in facial stylization. Furthermore, when comparing SAM with the self-trained EasyPortrait model in this work, the performance difference between the two is not significant. However, the EasyPortrait model is more lightweight, requiring fewer computational resources and offering greater flexibility. Therefore, the EasyPortrait model is chosen as the face segmentation model. Through visualization, it is evident that the model proposed in this work successfully achieves background stylization and facial enhancement while preserving facial integrity.

4. Discussion

In the study, two separate models were employed for different processing tasks: the CycleGAN model for image style transfer and the GFPGAN model for facial enhancement. This separate processing approach offers the following advantages:

Preservation of facial accuracy and clarity: By processing the background and facial regions separately, it ensures that the accuracy and clarity of the facial features are preserved during the background style transfer. The GFPGAN model, dedicated to facial enhancement, can provide clear details of the face without distorting the facial characteristics.

Improved control over style transfer effects: By conducting style transfer and facial enhancement as distinct processes, better control over the parameters and optimization objectives of each process can be achieved. This enables a better balance between background style transfer and facial enhancement, resulting in more desirable outcomes.

On the other hand, employing a self-trained lightweight model (such as EasyPortrait) as a substitute for the SAM model offers the following advantages:

Reduced computational resource consumption: Lightweight models typically have smaller model sizes and lower computational requirements compared to larger models like SAM. This means that the model can be run on relatively low hardware configurations, reducing the consumption of computational resources. This is particularly important for deploying models in resource-constrained environments or real-time applications.

Good performance: Although lightweight models may entail some performance trade-offs compared to larger models, the reduced computational resource consumption allows maintaining relatively high performance levels. This enhances the practical usability of the model, striking a balance between resource efficiency and performance.

5. Conclusion

In conclusion, the proposed approach effectively addresses the challenge of stylizing images with human faces. By combining CycleGAN for background stylization, GFPGAN for facial enhancement, and a suitable facial segmentation model, the desired results of Monet-style backgrounds with preserved facial details are achieved. The evaluation of different combinations and methodologies underscores the importance of selecting appropriate models based on performance, efficiency, and application requirements.

Looking ahead, there is room for further improvements to enhance the robustness and accuracy of the facial segmentation process. Additionally, exploring alternative style transfer methods and incorporating user preferences for personalized stylization could be interesting directions for future research. Overall, this work contributes to the field of image stylization with human faces and opens up possibilities for various practical applications in art, design, and entertainment domains.

References

[1] Yongcheng, J., et al. Neural style transfer: A review. IEEE transactions on visualization and computer graphics, vol. 26.11, pp. 3365-3385 (2019).

[2] Di, J., et al. Deep learning for text style transfer: A survey. Computational Linguistics, vol. 48.1, pp. 155-205 (2022).

[3] Xu, H., et al. Pre-trained models: Past, present and future. AI Open, vol. 2, pp. 225-250 (2021).

[4] Ce, Z., et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. arXiv preprint arXiv:2302.09419 (2023).

[5] Jun-Yan, Z., et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision, pp. 2223-2232 (2017).

[6] Xintao, W., et al. Towards real-world blind face restoration with generative facial prior. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9168-9178, (2021).

[7] Alexander, K., et al. Segment anything. arXiv preprint arXiv:2304.02643 (2023).

[8] Alexander K., Karina, K., and Sofia, K. EasyPortrait--Face Parsing and Portrait Segmentation Dataset. arXiv preprint arXiv:2304.13509 (2023).

[9] Tero, K., et al. Analyzing and improving the image quality of stylegan. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8110-8119 (2020).

[10] EasyPortrait homepage. URL: https://github.com/hukenovs/easyportrait. Last accessed 2023/08/25.