ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095858

MINIMISING DISTORTION FOR GAN-BASED FACIAL ATTRIBUTE MANIPULATION

Mingyu Shao †, *Li Lu* †, *Ye Ding* †,✉, *Qing Liao* ‡

† School of Cyberspace Security, Dongguan University of Technology, China
‡ School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China

ABSTRACT

Facial Attribute Manipulation (FAM) through GAN-based methods has been an active topic in computer graphics. Existing works show high editing fidelity on randomly generated faces but suffer from distortion on embedded real faces. We alleviate this issue by dividing it into two sub-problems. First, we minimize embedding distortion by introducing a pre-trained Salient Object Detection (SOD) network. Second, we propose a nonlinear transformation network to minimize editing distortion. As a result, our framework, Character Centered Facial Attribute Manipulation (CCFAM), exhibits more disentangled edits on real faces. Moreover, CCFAM is computationally efficient by integrating image complexity into our embedding process. Evaluations demonstrate that our method performs better than the state-of-the-art in terms of both accuracy and fidelity.

Index Terms— Facial Attribute Manipulation, GAN, Salient Object Detection

1. INTRODUCTION

Generative Adversarial Networks (GANs) have made significant progress in image synthesis over the years. In particular, StyleGANs [1, 2, 3] are capable of synthesizing high-resolution photo-realistic images including human faces. Facial Attribute Manipulation (FAM) aims to edit specific facial attributes while keeping others unchanged. However, it remains challenging when applying GANs to FAM. Distortions are found in manipulated images, especially in those complex ones. There are two reasons for this: (i) images must be embedded into the latent space in advance. The embedding methods and the complexity of the images will affect the embedding quality; and (ii) attributes on human faces are entangled with each other in the latent space. Manipulating one attribute may unintentionally affect the others. Furthermore, we demonstrate that entanglements between facial and non-facial attributes will also cause distortion. Thus, we present a novel framework called Character Centered Facial Attribute Manipulation (CCFAM) to minimize the distortion during the embedding and editing process of FAM.

Existing works like [4, 5] are capable of embedding images into the latent space of StyleGANs. But when the complexity of images arises, the embedding quality declines dramatically. As shown in Fig.1, we keep the main characters unchanged and replace them with different backgrounds. Using former embedding techniques, the embedding quality varies with the complexity of images. To further illustrate this issue, we introduce image entropy (H) to evaluate the complexity of an image. As image entropy goes up, image complexity goes up as well. Using [4] with starting latent code $\overline{w}$ and fixed embedding parameters. We randomly embed 500 images from the FFHQ dataset. Then we calculate the *PSNR* between the embedded images and original images. Fig.2 shows the correlation between *PSNR* and H. As we can see, to stabilize the embedding process, it is necessary to control the complexity of the images. We introduce Salient Object Detection (SOD) to the embedding network, which significantly constrains the image complexity and alleviates the embedding distortion.

Fig. 1. Examples of embedding quality varying with image complexity. Upper: Original images; Lower: Embedded results. The background distortions are marked as red boxes. The blurry edges between the foreground and background are marked as green boxes.

By tweaking the latent codes, existing works like [6, 7, 8] can control the editing process with semantic specifications. But it remains challenging to fully disentangle the attributes. To tune the editing direction and achieve disentanglement, [9] introduced pre-trained segmentation networks and localization scores. However, they still assumed there are linear paths among the attributes in the latent space. We further disentangle the latent space with non-linear transformation networks. Benefiting from the localization scores, the proposed CCFAM shows more editing fidelity.

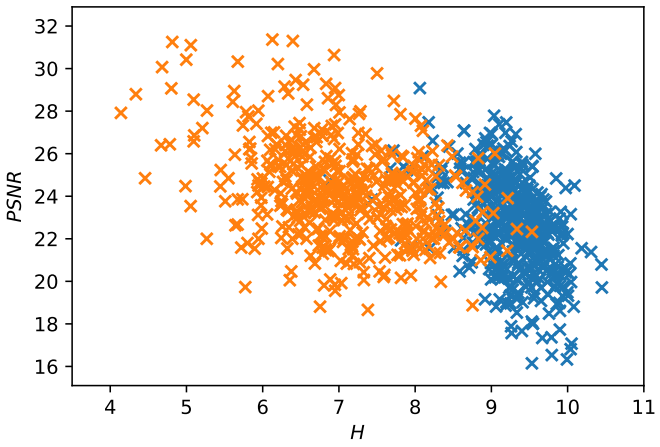

Fig. 2. The correlation between embedding quality and image complexity. As the image complexity increases, the embedding quality decreases. Blue: Image2StyleGAN [4]; Orange: Image2StyleGAN + SOD.

We also propose a new dataset that contains 2500 face images of 500 celebrities. The backgrounds of these images vary from solid colours to in-the-wild scenarios. Based on SOD and our new dataset, we not only disentangle the facial and non-facial attributes for the first time, but also expand FAM to more application scenarios, such as portrait drawing manipulation and portfolio portrait generation.

2. RELATED WORKS

2.1. FAM and GAN Inversion

Many former FAM approaches are based on Conditional GANs (CGANs) [10]. By incorporating additional information as input, CGANs can reduce the generation uncertainty, making the output more in line with our expectations. CGANs can translate an image from one domain to another. In FAM, these domains could be gender, age, facial expressions, etc. One representative approach CycleGAN [11] can translate images across two domains. Furthermore, StarGAN [12] realized translations across multiple domains.

Since StyleGANs [1, 2, 3] provided us powerful generators with scale-specific control of high-level attributes. Many FAM approaches were proposed to explore the latent space of StyleGANs. To reconstruct an image through the generators, one typical way [13] is to train an encoder network to map a given image into a latent code. Some approaches [2, 4] also realized embedding by directly optimizing the latent code. These optimization models usually achieve low distortion, but their embedding results are less editable because of the entanglements.

There are linear transformations [6, 7] and non-linear transformations [14, 8, 15] to map the latent code from $\mathcal{W}$ to $\mathcal{W}^+$. Normally, non-linear transformations perform better than linear ones. One representative non-linear approach is StyleFlow [8], which requires a commercial Face API for classifying the attributes. By contrast, GuidedStyle [15] introduced a pre-trained classification model, which still has state-of-the-art performance.

2.2. Salient Object Detection

Salient Object Detection (SOD) aims to detect and segment the visually distinctive object in an image, which, in our CC-FAM, is the main character. The performance of SOD has improved significantly over the years due to the development of Convolutional Neural Networks (CNNs). Many SOD approaches employ backbones like AlexNet [16], VGG [17], and ResNet [18] for feature extraction, which are effective but computationally expensive. We, instead, adopt U^2-Net [19], a simpler yet powerful state-of-the-art SOD approach to constrain the complexity of our embedding images. Unlike image matting, which requires extra trimaps as input, SOD requires only one image at a time, which is ideal for our embedding process.

The features of SOD can provide us with some interesting applications such as portrait drawing generation. We will also make use of these features to broaden the application scenarios of FAM in Sec.4.

3. METHOD

Mapping images into latent codes is a lossy compression process. Therefore, the encoder-decoder structure of GAN inversion will hit its bottleneck when the complexity of images arises. This will lead to embedding distortion and less editability. We observed that images with a deeper depth of field usually perform better embedding fidelity. The depth of field highlights the salient objects and decreases the image complexity. Thus, we will improve the GAN Inversion editability from a new perspective by actively constraining the complexity of images with SOD.

Fig. 3 shows the overview of our method. The original image I can be formulated as:

$$I = \lambda C + (1 - \lambda)B, \tag{1}$$

where C stands for the main character, B stands for the background and λ controls the transparency of pixels. Our embedding target can be formulated as:

$$I' = \lambda C + (1 - \lambda)\mathcal{B}(B), \tag{2}$$

where $\mathcal{B}$ stands for blur algorithms. Here we choose Gaussian blur with a kernel size of 99×99. By applying SOD and Gaussian blur, the average image entropy of our test samples drops from 9.27 to 7.03, which significantly relieves the embedding pressure.

Our method is applicable to any embedding approach. Here we choose latent code optimization as an illustration.

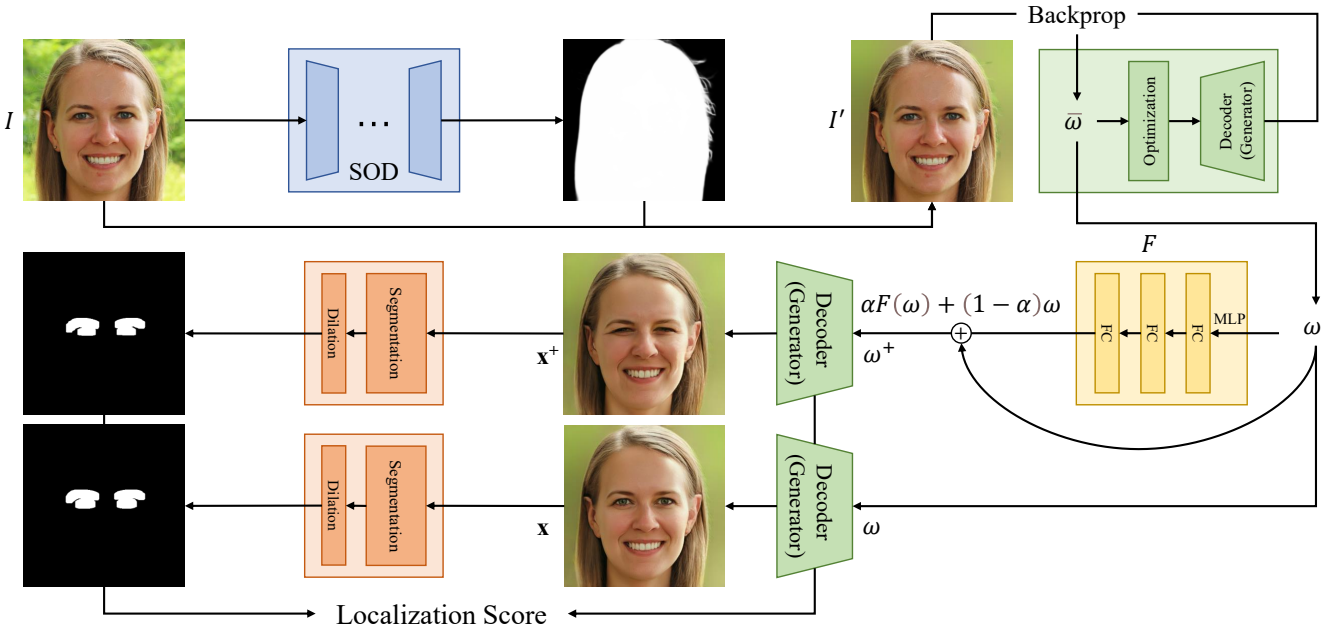

Fig. 3. The overview of our framework. The blue box indicates the pre-trained SOD network. The green box shows the embedding process using the latent code optimization strategy. The yellow box indicates the RA-MLP block, which performs the non-linear latent transformation. The semantic segmentation model is indicated by orange boxes. Here we extend the segmentation model with an extra dilation operation to cover more area around the eyes.

Starting from the mean latent code $\bar{\omega}$, the iteration of ω and embedding loss $\mathcal{L}$ can be formulated as:

$$\mathcal{L} \leftarrow \mathcal{L}(G(\omega), I'), \tag{3}$$

$$\omega \leftarrow \omega - \eta \mathcal{F}(\nabla_\omega \mathcal{L}). \tag{4}$$

$G(\cdot)$ is the generator network. The embedded latent code ω and the embedded image $G(\omega)$ are optimized via $\mathcal{F}$.

We can perform FAM by moving the latent code ω along a proper direction in the latent space. The Localization Score from [9] is an effective method to tune an existing direction. To take this method one step further, we combine it with a Residual Attention MLP block (RA-MLP) [15]. Unlike the original Localization Score, which was only used for optimizing linear directions, we realize non-linear optimization to further reduce the editing distortion.

Given a pre-trained RA-MLP block, the extended latent code can be formulated as:

$$\omega^+ = H(\omega) = \alpha F(\omega) + (1-\alpha)\omega, \tag{5}$$

where $F(\cdot)$ stands for the MLP layer to be tuned and α controls the magnitude of changes.

Using $r_i(\omega)$ to represent the i-th layer activation of $G(\omega)$ and $s_i(\mathbf{x}, \mathbf{x}^+)$ is the average of segmentation masks downsampled to the resolution of the i-th layer. Similar to [9], our objective function can be formulated as:

$$LS(H(\omega)) = \frac{\sum_i s_i(\mathbf{x}, \mathbf{x}^+) \odot |r_i(\omega) - r_i(H(\omega))|^2}{\sum_i |r_i(\omega) - r_i(H(\omega))|^2}. \tag{6}$$

	Image2StyleGAN	pSp	IDInvert	e4e
Original	22.66	22.81	22.89	22.92
+ SOD	**24.20**	**24.27**	**24.38**	**24.35**

Table 1. The average value of PSNR between the embedded images and original images (higher is better). The embedding images are randomly selected from the FFHQ dataset.

Instead of setting a latent direction as the optimization target, our optimization target is a non-linear transformation function that can move the latent code more precisely.

4. EXPERIMENTS

For embedding parts, we apply SOD to multiple embedding approaches to verify the effectiveness of our method. For editing parts, we take [20] as the segmentation model and train our model on randomly generated face images using SGD with momentum weight of 0.9, learning rate of 0.001, and batch size of 10. We use the FFHQ dataset and our proposed dataset for evaluation. We compare our CCFAM with other competing methods using evaluation metrics of Fréchet Inception Distance (*FID*) and Sliced Wasserstein Distance (*SWD*) for measuring the similarity between the edited faces and the originals, Cosine Similarity (*CS*) and Euclidean Distance (*ED*) for quantifying the identity preservation.

Metric	InterFaceGAN	StyleFlow	Ours
FID ↓	62.13 85.07	52.97 82.78	**50.32** **80.37**
SWD ↓	361.82 425.79	300.14 410.02	**286.58** **397.30**
CS ↑	0.87 0.61	0.84 0.63	**0.91** **0.78**
ED ↓	0.83 0.85	0.76 0.72	**0.59** **0.57**

Table 2. Quantitative comparison with latent space manipulation models measured by different metrics. Here we choose two attributes as illustrations. First row: Manipulation on the smile. Second row: Manipulation on the eyeglasses.

Background Colour	CS ↑	ED ↓
White	**0.97**	**0.38**
Red	0.96	0.42
Blue	0.94	0.43

Table 3. The influence of background colour on FAM. We choose three colours that are commonly used in portfolio portraits: white (r: 255, g: 255, b: 255), red (r: 210, g: 10, b: 50), and blue (r: 30, g: 170, b: 230).

4.1. Embedding with SOD

SOD enforces the encoders to focus on the salient objects by reducing unimportant information in an image. We calculate the PSNR improvements with and without SOD on Image2StyleGAN [4], IDInvert [5], pSp [13], and e4e [14]. As shown in Table 1, when the complexity of embedding images is constrained, all the methods show higher embedding fidelity. We also use the image entropy to dynamically control the embedding process, rather than the commonly used early stopping strategy, which saves 7% of the embedding iterations on average.

4.2. Semantic Editing

Table 2 shows the quantitative comparison with InterFaceGAN [7] and StyleFlow [8]. As we can see, our CCFAM outperforms other competitors with less editing distortion. Based on the proposed dataset, we explore the influence of background colour on FAM. As shown in Table 3, among the three colours we choose, embedded images with solid white backgrounds have better identity preservation capability.

4.3. More Application Scenarios

Benefiting from the features of SOD, we can apply FAM to more interesting application scenarios. As shown in Fig. 4, our CCFAM can be transferred to portrait drawing manipulation. With our new proposed dataset, we also realize portfolio portrait generation (Fig. 5). With fully disentangled foreground and background, we can manipulate the background colour directly in the latent space. Unlike image matting, which can only change the background colour, we can also manipulate other attributes at the same time.

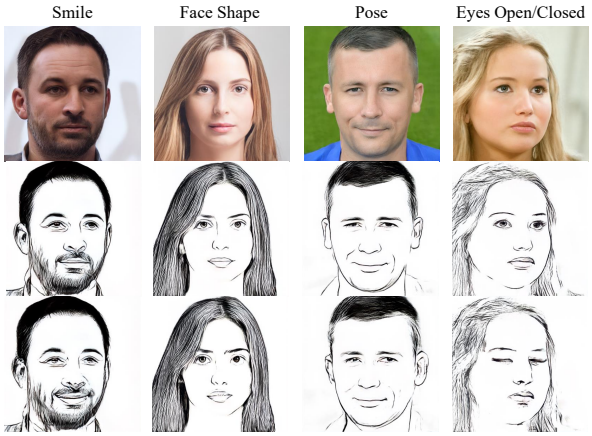

Fig. 4. Visualization of portrait drawing manipulation.

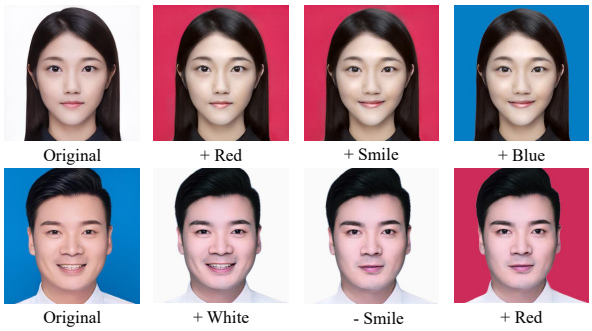

Fig. 5. Visualization of portfolio portrait generation.

5. CONCLUSION

In this work, we present CCFAM to minimize the distortions of FAM in two aspects. First, we introduce a pre-trained SOD network to minimize embedding distortion. Second, we combine RA-MLP with Localization Score to minimize editing distortion. We also proposed a new dataset to further broaden the application scenarios of FAM. Experiments show the superior embedding and editing fidelity of our method over previous works.

6. ACKNOWLEDGEMENTS

This work was supported by National Key R & D Program of China (No. 2022YFF0606303) and National Natural Science Foundation of China (No. U19A2067).

7. REFERENCES

[1] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[2] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[3] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[4] Rameen Abdal, Yipeng Qin, and Peter Wonka, "Image2stylegan: How to embed images into the stylegan latent space?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4432–4441.

[5] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou, "In-domain gan inversion for real image editing," in *European Conference on Computer Vision*. Springer, 2020, pp. 592–608.

[6] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris, "Ganspace: Discovering interpretable gan controls," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9841–9850, 2020.

[7] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou, "Interfacegan: Interpreting the disentangled face representation learned by gans," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[8] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka, "Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows," *ACM Transactions on Graphics*, vol. 40, no. 3, pp. 1–21, 2021.

[9] Ehsan Pajouheshgar, Tong Zhang, and Sabine Süsstrunk, "Optimizing latent space directions for gan-based local image editing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 1740–1744.

[10] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[11] 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*, 2017, pp. 2223–2232.

[12] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[13] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2287–2296.

[14] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or, "Designing an encoder for stylegan image manipulation," *ACM Transactions on Graphics*, vol. 40, no. 4, pp. 1–14, 2021.

[15] Xianxu Hou, Xiaokang Zhang, Hanbang Liang, Linlin Shen, Zhihui Lai, and Jun Wan, "Guidedstyle: Attribute knowledge guided style manipulation for semantic face editing," *Neural Networks*, vol. 145, pp. 209–220, 2022.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[17] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[19] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, pp. 107404, 2020.

[20] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 325–341.