VR-FAM: VARIANCE-REDUCED ENCODER WITH NONLINEAR TRANSFORMATION FOR FACIAL ATTRIBUTE MANIPULATION

Yifan Yuan, Siteng Ma, Junping Zhang†

Shanghai Key Lab of Intelligent Information Processing
School of Computer Science, Fudan University, Shanghai, China
{yfyuan21, stma21}@m.fudan.edu.cn, jpzhang@fudan.edu.cn

ABSTRACT

Facial attribute manipulation (FAM) aims to infer desired facial images by modifying specific attributes while keeping others unchanged. Existing works suffer from the entanglement of facial attributes, leading to unexpected artifacts and the loss of facial identity information after editing. To alleviate these issues, we propose a novel FAM framework based on StyleGAN, termed VR-FAM, which can meet the requirements of FAM—editing ability, distortion, and fidelity. First, we propose a variance-reduced encoder to make the latent space close to the one of StyleGAN. Second, we present a nonlinear latent transformation network, which can convert the source latent code to target latent code in line with the nonlinear latent space of StyleGAN. Experimentally, we evaluate the proposed FAM framework on the benchmark FFHQ dataset and demonstrate the improvement gain over the recently published models in terms of edit accuracy and fidelity.

Index Terms— Facial attribute manipulation, disentangle learning, StyleGAN, GANs, style transfer

1. INTRODUCTION

Facial attribute manipulation (FAM) is regarded as an image-to-image translation task, which has broad applications in facial beautification, photo retouching, and film post-production industry. Generally speaking, it can be divided into three separate tasks: *GAN (Generative Adversarial Network) inversion*, *latent code manipulation*, and *image synthesis*. Among them, GAN inversion is the inversion operation of GAN, mapping the facial image into a latent space. Then latent code manipulation task is to transform the inversion representation to a new latent code depending on the specific editing requirements. Finally, a new facial image with desired attribute(s) is generated through generative models such as Variational AutoEncoders (VAEs) [1] and GANs [2].

†: Corresponding author.

This paper is supported by National Natural Science Foundation of China (No. 62176059), National Key Research and Development Program of China (No. 2018YFB1305104), Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01), ZJ Lab, and Shanghai Center for Brain Science and Brain-Inspired Technology.

Related work. Among different variants of GANs, StyleGAN [3, 4] achieves promising results, especially in quality, naturalness and diversity. A crucial reason is that StyleGAN utilizes a mapping network to map a random sampled code to an intermediate latent space $\mathcal{W}$, capturing more disentangled properties of learned distribution [5, 6, 7, 8, 9]. However, StyleGAN's intermediate latent space is different from the one formed by the real facial images [10]. Therefore, in order to accomplish the real images editing, Abdal *et al*. [10] presented Image2StyleGAN, which can invert real images to StyleGAN's $\mathcal{W}+$ space by optimizing the random sampled latent code directly. Nevertheless, the optimization process requires a long time. Therefore, Richardson *et al*. [11] proposed an encoder which can directly invert real images to the extended $\mathcal{W}+$ space based on Feature Pyramid Network [12]. However, Tov *et al*. showed that although the reconstruction ability is promoted, the editability and perceptual quality are worse when editing in $\mathcal{W}+$ compared with $\mathcal{W}$. Hence, they designed an e4e encoder [5] to map real facial images to a latent space close to $\mathcal{W}$ space from $\mathcal{W}+$ space. As for the latent manipulation algorithm, Shen *et al*. [7] assumed that there exists a hyperplane in the latent space such that all samples from the same side are with the same attribute. Thus, one can edit images along the direction that is orthogonal to the normal vectors of other unrelated attributes' separating hyperplanes with linear transformations. However, the generated results are still attribute-entangled to some extent.

Despite recent advances in using GANs for FAM, it is desirable to meet the following three requirements in one FAM framework: 1) maximizing editing ability—maximizing the capability of editing images—guaranteeing the correct change of desired attributes, *i.e*. "change as you wish" [13]; 2) minimizing distortion—minimizing the difference between input images (original one) and output images (edited one)—being as small as possible while keeping others unchanged, *i.e*. "only change what you want"; and 3) maximizing fidelity—making the edited images indistinguishable from real images.

To that end, we present a novel FAM framework to achieve better realistic facial image manipulation through

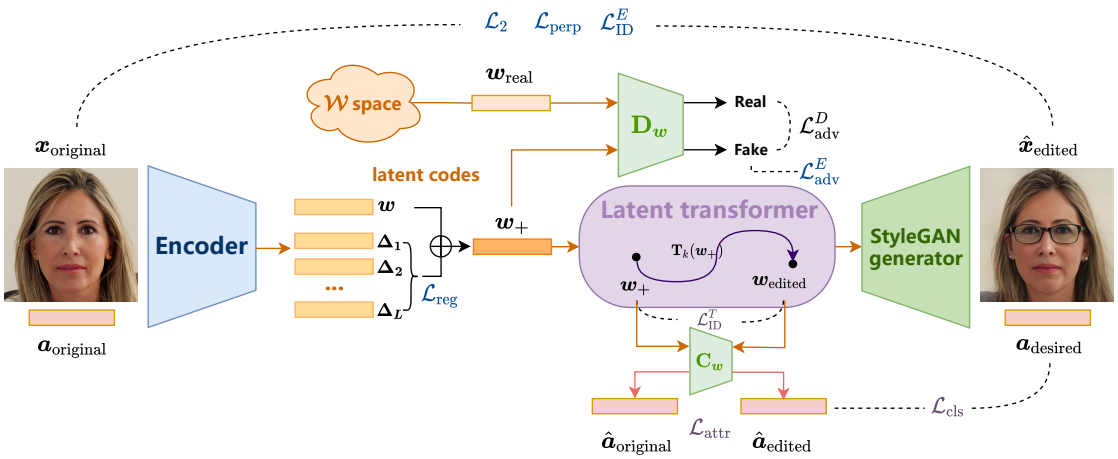

Fig. 1. Framework of our proposed algorithm

obtaining a trade-off among editing ability, distortion, and fidelity. Specifically, this framework is made up of three components: a variance-reduced encoder, a nonlinear latent transformation network, and a StyleGAN2 generator. The model first maps facial images into a latent space that is close to the one of StyleGAN. Then the latent transformation network edits latent code to guarantee the desired attribute manipulation while keeping the other unrelated regions unchanged. Finally, a StyleGAN2 generator is used to generate high-quality edited facial images.

Contribution. Our contributions can be summarized as follows. *First*, we propose a framework consisting of an encoder, a nonlinear latent transformation network and a StyleGAN2 generator to achieve high-quality and high-fidelity real facial image manipulation. *Second*, we propose a variance-reduced encoder whose generated latent space is close to StyleGAN's latent space. *Third*, we present a nonlinear latent transformation network that is designed to move the original latent code to the latent subspace with desired attributes. *At last*, the proposed framework achieves better facial editing performances in quantitative criterion and visual effects compared to the state-of-the-art algorithms on FFHQ dataset.

2. PROPOSED METHOD

This section presents the proposed VR-FAM framework consisting of a StyleGAN2 generator, a variance-reduced encoder, and a nonlinear latent transformation network, which is illustrated in Fig. 1.

2.1. Variance-reduced GAN Inversion

Within our framework, the encoder is used to invert an original real facial image to StyleGAN's latent space.

Note that although an extended latent space $\mathcal{W}+ \subseteq \mathbb{R}^{512}$ through inputting different style codes instead of one style code can enhance the reconstruction and generalization ability of StyleGAN, the editing ability and the fidelity may be worse than the original $\mathcal{W}$ space [5]. Therefore, it is necessary to make the inferred $\mathcal{W}+$ get closer to $\mathcal{W}$ so that an explicit trade-off between the perceptual quality and distortion can be reached in the latent space [14]. To find a better trade-off among editing ability, fidelity and distortion, we propose an encoder based on e4e encoder [5] detailed as follows.

Generally, the distortion will worsen while the editability and perceptual quality improve as we approach $\mathcal{W}$ from $\mathcal{W}+$. Note that the difference between $\mathcal{W}$ and $\mathcal{W}+$ is that $\mathcal{W}+$ has multiple different style codes while $\mathcal{W}$ has the same style code be fed into the synthesis network in all layers. We intend to minimize the variance of different style codes for obtaining a better quality of facial editing manipulation. Denote the output style codes as $\mathbf{w}_+ = (w_0, w_1, \ldots, w_{L-1})$, where L is the number of style-modulation layers. We convert the output form to $\mathbf{w}_+ = (w, w + \Delta_1, \ldots, w + \Delta_{L-1})$ and let the encoder learns the single style code w at the beginning of training process and learns the residual terms Δ_i *sequentially* in the subsequent process. To minimize the variance between different style codes produced by this encoder, we add a $\mathcal{L}_2$-regularization loss:

$$\mathcal{L}_{\text{reg}}(\mathbf{w}_+) = \sum_{i=1}^{L-1} \|\Delta_i\|_2 \,. \tag{1}$$

Besides, it is expected that these style codes lie within the actual distribution of $\mathcal{W}$. We thus utilize a latent discriminator [15] to let latent codes learned by the encoder get closer to the real latent codes sampled from the StyleGAN's $\mathcal{W}$ space in an adversarial manner. We use the vanilla GAN loss with R_1 regularization [16] as follows:

$$\begin{aligned}\mathcal{L}_{\text{adv}}^D = &- \mathop{\mathbb{E}}_{\mathbf{w}\sim\mathcal{W}} \left[\log D_{\mathcal{W}}(\mathbf{w})\right] - \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{\mathbf{x}}} \left[\log\left(1 - D_{\mathcal{W}}\left(E(\mathbf{x})_i\right)\right)\right] \\ &+ \frac{\gamma}{2} \mathop{\mathbb{E}}_{\mathbf{w}\sim\mathcal{W}} \left[\|\nabla_{\mathbf{w}} D_{\mathcal{W}}(\mathbf{w})\|_2^2\right], \end{aligned} \tag{2}$$

$$\mathcal{L}_{\text{adv}}^{E} = - \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{\mathbf{x}}} \left[\log D_{\mathcal{W}}\left(E(\mathbf{x})_i\right)\right], \tag{3}$$

where $\mathcal{W}$ and $p_{\mathbf{x}}$ denote the distributions of $\mathcal{W}$ and the training data respectively. $D_{\mathcal{W}}$ is a latent discriminator and $E(\mathbf{x})_i$ denotes the i-th latent code generated by the encoder.

2.2. Nonlinear Latent Transformation Network

In order to exploit the latent subspace with the desired attribute, we require to obtain the correspondence between latent code and facial attributes. Therefore, we first train a latent classifier C to predict the latent code's binary attribute $\{0,1\}^N$ with the training data, *i.e.* "latent code $E(\mathbf{x})$ - attribute label $\mathbf{a}$" pairs. Subsequently, a nonlinear latent transformation network T based on [17] is trained to manipulate them to the latent subspace with desired attributes. More specifically,

$$T(\mathbf{w}_+, \alpha) = \mathbf{w}_+ + \alpha \cdot f(\mathbf{w}_+), \tag{4}$$

where $f(\cdot)$ denotes a nonlinear function and α is an editing scaling factor ranging from [-1.5, 1.5] during the test phase. Multiple T_ks are trained for each attribute $\mathbf{a}_k \in \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N\}$. Compared to the previous linear mapping method [7], the latent transformation network learns a nonlinear mapping to alleviate the attribute-entangled issue.

2.3. Network Architecture

We adopt the same architecture as pSp [11] as the encoder. Specifically, multiple feature maps are extracted through a standard feature pyramid [12] over a ResNet backbone. Then different style vectors are generated from the corresponding feature map. As for the latent classifier and the transformation network, they are all consist of three fully connected networks. Finally, each dimension of the edited latent code is split out and fed to each layer of the StyleGAN2 [4] generator's synthesis network with a pyramid structure.

2.4. Loss Function

Reconstruction loss. To guarantee a low distortion of the synthesized image, we adopt the pixel-wise and perceptual loss in pSp [11].

Identity loss. Moreover, an identity loss is proposed to preserve one's identity information:

$$\mathcal{L}_{\text{ID}}^{E} = 1 - \langle R(\mathbf{x}), R(G(E(\mathbf{x})))\rangle, \tag{5}$$

where $R(\cdot)$ is the pretrained ArcFace network [18] and $G(\cdot)$ denotes the StyleGAN2 generator.

Attribute classification loss. In order to guarantee the correct manipulation of the target attribute, a cross-entropy loss is utilized:

$$\mathcal{L}_{\text{cls}} = -\mathbf{a}_k \log(\mathbf{p}_k) - (1-\mathbf{a}_k)\log(1-\mathbf{p}_k), \tag{6}$$

where $\mathbf{a}_k \in \{0,1\}$ is the desired attribute and the probability that $T_k(\mathbf{w}_+)$ is classified as a_k is denoted as $\mathbf{p}_k = C(T_k(\mathbf{w}_+))[k]$.

Meanwhile, the other unrelated attributes (*i.e.*, $\mathbf{a}_i$, $i \neq k$) should be unchangeable:

$$\mathcal{L}_{\text{attr}} = \sum_{i\neq k} (1-\gamma_{ik}) \mathbb{E}_{\mathbf{w}_+, i}[\|\mathbf{p}_i - C(\mathbf{w}_+)[i]\|_2], \tag{7}$$

where γ_{ik} measures the relevance between $\mathbf{a}_i$ and $\mathbf{a}_k$, which is obtained by calculating their correlation coefficient matrix. Moreover, we present an identity loss to preserve one's identity:

$$\mathcal{L}_{\text{ID}}^{T} = 1 - \langle R(G(\mathbf{w}_+)), R(G(T(\mathbf{w}_+)))\rangle. \tag{8}$$

Total loss. In summary, the encoder's objective function is defined as:

$$\begin{aligned}\mathcal{L}_E = &\lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{adv}}^{E}\mathcal{L}_{\text{adv}}^{E} + \lambda_{\ell_2}\mathcal{L}_2 + \\ &\lambda_{\text{perp}}\mathcal{L}_{\text{perp}} + \lambda_{\text{ID}}^{E}\mathcal{L}_{\text{ID}}^{E}.\end{aligned} \tag{9}$$

The total loss of latent transformation network is defined as:

$$\mathcal{L}_T = \mathcal{L}_{\text{cls}} + \lambda_{\text{attr}}\mathcal{L}_{\text{attr}} + \lambda_{\text{ID}}^{T}\mathcal{L}_{\text{ID}}^{T}. \tag{10}$$

3. EXPERIMENTS

In this section, we evaluate our proposed method in Flickr-Faces-HQ (FFHQ) [3] dataset while use CelebA-HQ [19] for training. Further, we compare it with several recently published methods including InterFaceGAN [7] and e4e [5].

3.1. Experimental Setup

Datasets. CelebA-HQ and FFHQ contain 30K and 70K images respectively with a resolution of 1024×1024.

Implementation details. For the generator, we use the pretrained StyleGAN2 generator [4]. As for the encoder, we train the first $\mathbf{w}_+$ style vector in the first 20K training steps and train each Δ_i, $i \in [1, L-1]$ for 2K steps in the following steps, with a mini-batch size of 4. We use Adam optimizer [20] with a learning rate of 1.0×10^{-4}, $\beta_1 = 0.95$ and $\beta_2 = 0.999$. The loss weights are set to $\lambda_{\ell_2} = 1$, $\lambda_{\text{perp}} = 0.8$, $\lambda_{\text{ID}}^{E} = 0.1$ in reconstruction loss, $\lambda_{\text{reg}} = 2.0 \times 10^{-4}$ and $\lambda_{\text{adv}}^{E} = 0.1$ in editing loss. For the latent transformation network, we use a fixed learning rate of 1.0×10^{-3}, minibatch size of 32, Adam optimizer, and training steps of 100K. In loss functions, λ_{cls}, λ_{attr} and λ_{ID} are all set to 1. Experiments are conducted on a NVIDIA RTX 2080 Ti GPU.

3.2. Experimental Results

As shown in Fig. 2, our method achieves an appropriate and effective manipulation while maintaining one's identity information during the sequential editing process. Furthermore,

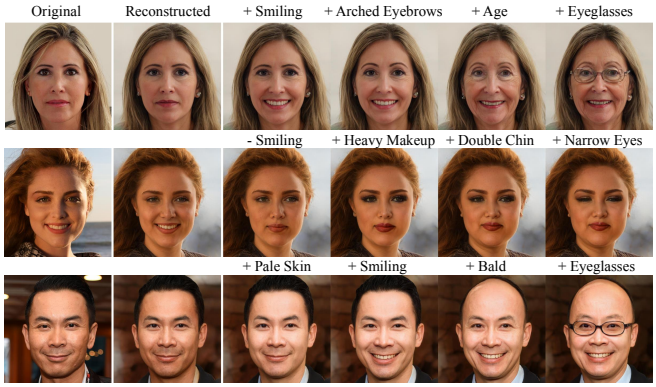

Fig. 2. Sequential facial attribute editing on real images. We sequentially manipulate the latent code to edit the real facial images with different attributes.

we compare our proposed framework with the state-of-the-art FAM algorithms including InterFaceGAN [7] and e4e [5], as shown in Fig. 3. Specifically, InterFaceGAN usually changes the unrelated attributes, *e.g.* put on eyeglasses when letting one gets older (Row 1, Col 2) and add hair when changing one's gender (Row 1, Col 5). It is mainly for the reason that InterFaceGAN cannot fully decouple the attributes through attribute separation and linear editing. In the second row, we use the e4e encoder to accomplish GAN inversion while use the linear transformation to conduct latent code manipulation. As shown in row 2, the person's mouth is not completely closed and its shape is changed when changing *'Smiling'* attribute. In addition, the area around his eyes becomes darker when putting the eyeglasses on him (Row 2, Col 4). Besides, the quantitative results in Table 1 show that our method is remarkablely better than InterfaceGAN, and has a subtle difference with e4e according to perception and editablity indices. Note that the extensive qualitative results demonstrate that FID [21] does not necessarily reflect human judgment.

3.3. Ablation Study

We also perform ablation study as in Fig. 4. It can be seen that if we remove the $\mathcal{L}_{reg}$ (Config. (a)) or $\mathcal{L}_{adv}$ (Config. (b)), which means the style vectors in latent space different from each other and get close to $\mathcal{W}+$ space, this will result in the decline of reconstruction ability in some aspects. For example, extra hair is added to the first person around her neck (see Row 1, Col 2). Other configurations such as removing $\mathcal{L}_2$ or $\mathcal{L}_{perp}$ will lead to the blurry of images. In addition, $\mathcal{L}_{ID}$ plays a significant role in preserving the facial identity as shown in config. (f) compared to config. (e) (w/o $\mathcal{L}_{ID}$).

4. CONCLUSIONS

In this paper, we present a real facial editing framework consisting of a variance-reduced encoder, a nonlinear latent transformation network and a StyleGAN generator through focusing on GAN inversion, latent code manipulation and image synthesis separately. Specifically, considering the distortion-editability balance when editing images in StyleGAN latent space, we utilize an encoder to let the generated latent space approach to a space with a better editing ability from another space with a better reconstruction ability. Furthermore, a latent transformation network is proposed to edit latent code non-linearly as attribute embeddings are not linearly decoupled in latent space. Experiments conducted on FFHQ dataset demonstrate that our framework can generate realistic editing results with better disentanglement and identity preservation.

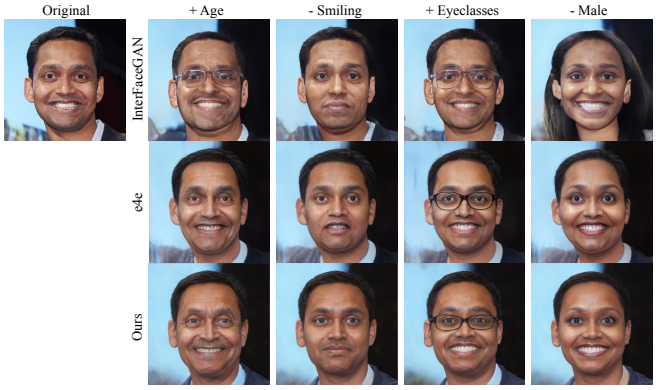

Fig. 3. Comparison with InterFaceGAN [7] and e4e [5].

Table 1. Quantitative comparisons between the recent methods. We use Fréchet Inception Distance (FID) [21] to calculate the mean and covariance matrix of the two distributions before and after editing the facial attributes. *Perception quality* is measured between the original and reconstructed images while *editability* is measured between the original and edited images. For all metrics, the lower is better.

Method	**Distortion** L_2	**Perception** FID	**Editability** FID
InterFaceGAN [7]	0.049	74.92	130.23
e4e [5]	0.053	**30.96**	**81.08**
Ours	**0.041**	32.21	83.75

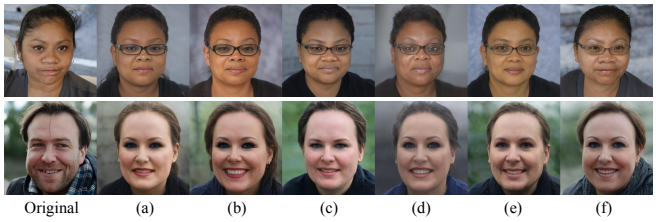

Fig. 4. Ablation study on the loss composition in Eq. (9). From left to right, they are: 1) Ground-truth. 2) Configuration (a)-(e) represent without $\mathcal{L}_{reg}$, $\mathcal{L}_{adv}$, $\mathcal{L}_2$, $\mathcal{L}_{perp}$ or $\mathcal{L}_{ID}$, respectively. 3) Configuration (f) represents our method involving all losses. In row 1 we put *eyeglasses* on the person while in row 2 we change his *gender* to female.

5. REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[3] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8110–8119.

[5] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for StyleGAN image manipulation," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–14, 2021.

[6] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "GANspace: Discovering interpretable GAN controls," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[7] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9243–9252.

[8] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "StyleRIG: Rigging StyleGAN for 3D control over portrait images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6142–6151.

[9] Z. Wu, D. Lischinski, and E. Shechtman, "Stylespace analysis: Disentangled controls for StyleGAN image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12863–12872.

[10] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the StyleGAN latent space?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4432–4441.

[11] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. 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 (CVPR)*, 2021, pp. 2287–2296.

[12] Lin T., Dollár P., and Girshick R. B.and He K.and Hariharan B.and Belongie S. J., "Feature pyramid networks for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.

[13] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Transactions on Image Processing (TIP)*, vol. 28, no. 11, pp. 5464–5478, 2019.

[14] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6228–6237.

[15] Y. Nitzan, A. Bermano, Y. Li, and D. Cohen-Or, "Face identity disentanglement via latent space mapping," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–14, 2020.

[16] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge?," in *International Conference on Machine Learning (ICML)*, 2018, pp. 3481–3490.

[17] X. Yao, A. Newson, Y. Gousseau, and P. Hellier, "A latent transformer for disentangled face editing in images and videos," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2021, pp. 13789–13798.

[18] J. Deng, J. Guo, and S. Xue, N.and Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019.

[19] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6626–6637, 2017.