REGULARIZED LATENT SPACE EXPLORATION FOR DISCRIMINATIVE FACE SUPER-RESOLUTION

Ruixin Shi[1,2], *Junzheng Zhang*[1,2], *Yong Li*[1*], *Shiming Ge*[1,2*]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

ABSTRACT

Learning face super-resolution models is challenged in many practical scenarios where high-resolution and low-resolution face pairs usually are difficult to collect for training examples. Recent self-supervised approach provides a feasible solution by using low-resolution faces to guide the generation of the corresponding high-resolution ones with a pretrained generator. In this paper, we propose a regularized latent space exploration approach to facilitate self-supervised face super-resolution. In the approach, a pretrained generative adversarial network (GAN) is fully used to control the exploration of high-resolution face generation in an iterative optimization manner for a low-resolution face. During the iteration, super-resolution faces are continually generated from a feasible latent space by the generator and evaluated by the discriminator, while the generator is online finetuned. The generation is evaluated by measuring the semantic loss as well as pixel loss between ground-truth low-resolution faces and the corresponding downsampled super-resolution faces. In this way, the generated faces can be appearance natural and semantic discriminative. Experiments validate the effectiveness of our approach in terms of quantitative metrics and visual quality.

Index Terms— Face super-resolution, self-supervised learning, discriminative features

1. INTRODUCTION

Face super-resolution aims at recovering high-resolution faces from low-resolution ones, which has many real-world applications [1, 2]. Many approaches have been proposed and are mainly grouped into supervised or unsupervised category [3]. Supervised approaches usually trained with high-resolution and low-resolution face pairs. Some approaches, *e.g.*, SRCNN [4], FSRCNN [5] and VDSR [6], learned convolutional neural networks with a few layers to perform super-resolution. Later, EDSR [7] adopted residual learning to learn very deep networks. LapSRN [8] progressively predicts the results, and RCAN [9] and SAN [10] used channel attention. FSRNet [11] employs facial prior information to get final results. There are also some approaches based on generative adversarial network (GAN), such as UR-DGN [12] and ESRGAN [13], which rely on discriminator outputs to determine whether the result is real. GLEAN [14] used an additional network encoding the low-resolution faces as inputs embedded in GAN.

In generally, high-resolution and low-resolution face pairs are difficult to collect for training in practical scenarios. To address that, recent unsupervised approaches use only low-resolution faces to guide the generation of the corresponding high-resolution ones, leveraging the rich prior information of a pretrained generator. [16] proved that GAN generator trained on a large-scale image dataset can be used as a generic deep image prior [17]. Thus, PULSE [15] used a pretrained GAN and its latent vector traverses the latent space to

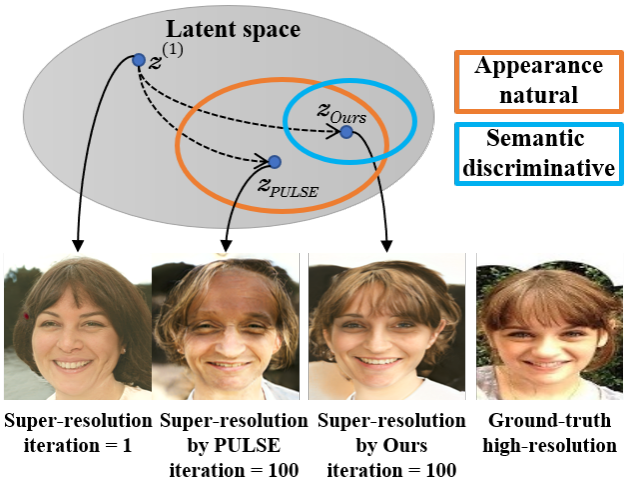

Fig. 1. We take the representative self-supervised approach PULSE [15] as example, which performs face super-resolution by generating high-resolution faces from the latent space. Our approach regularizes the generation by considering both appearance and semantics in latent space exploration thus can generate more discriminative results.

*This work was partially supported by grants from the National Natural Science Foundation of China (61772513), Beijing Natural Science Foundation (19L2040) and National Key Research and Development Plan (2020AAA0140001). Yong Li and Shiming Ge are the corresponding authors. Email: {liyong, geshiming}@iie.ac.cn.

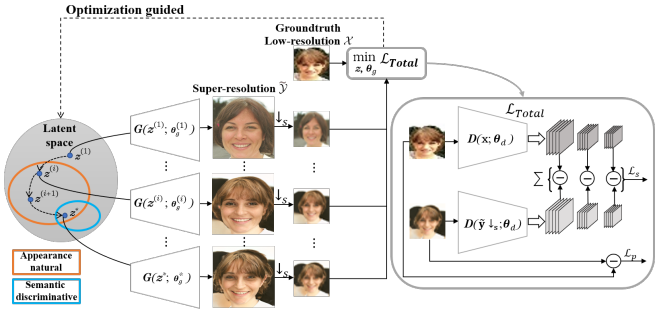

Fig. 2. The framework of our approach. It fully uses a pretrained GAN in an online latent space exploration manner. During iteration, the generator $\boldsymbol{G}$ continually generates super-resolution faces $\tilde{\mathbf{y}}$ from a random initialized latent code $\boldsymbol{z}^{(1)}$, while the generation is evaluated by measuring the discriminator semantic loss as well as pixel loss between $\mathbf{x}$ and $\tilde{\mathbf{y}}\downarrow_s$. The exploration is regularized by the total loss to get the discriminative result $\tilde{\mathbf{y}}^*$

search super-resolution images that downscale correctly in a self-supervised manner. mGANprior [18] employed multiple latent codes and adaptive channel importance to reconstruct the input image.

However, self-supervised approaches usually have poor control over appearance and the super-resolved faces may look unnatural. Fig. 1 shows an example, where the generated super-resolution faces by PULSE [15] look increasingly unnatural during learning. We intuitively analyze this problem from the perspective of image semantic. Instead of introducing extra network like VGGNet [19], we focus on using the corresponding GAN discriminator to get semantic features. According to GAN, the generator simulates real faces through the feedback of the discriminator during the training process. So the discriminator can retain the original structural parameters and provide correct guidance of generation. In this way, the acquired features can accurately represent the location of faces in the semantic space. Through empirical experiments, we find that these unnatural results are far from the ground-truth faces in the semantic space.

Inspired by this, we propose a regularized latent space exploration approach to facilitate self-supervised face super-resolution. The approach fully uses the pretrained GAN to control the exploration of face generation in an online optimization manner for low-resolution faces. During the iteration, super-resolution faces are continually generated from a feasible latent space by the generator. The generation is evaluated by measuring the discriminator semantic loss and pixel loss between ground-truth low-resolution faces and the corresponding downsampled super-resolution counterparts, while the generator is online finetuned. In this way, the latent vector is gradually converged to the optimal solution.

Our main contributions are three folds: 1) We study the control ability of generative models over face appearance and propose a regularized latent space exploration approach by fully using the pretrained GAN to control the exploration of face generation in an iterative optimization manner, 2) We introduce a semantic loss measured by the discriminator feature differences between the input low-resolution face and the downsampled super-resolution one to achieve appearance natural and semantic discriminative super-resolution results, and 3) We conduct extensive experiments to validate the effectiveness of our approach in terms of quantitative metric and visual quality, especially on few-sample scenario.

2. PROPOSED APPROACH

Our main idea is fully using a pretrained GAN including the generator and its corresponding discriminator to control the exploration of face generation in a self-supervised iterative optimization manner, in the case of only low-resolution faces $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{X}|}$ are available. We utilize a pretrained generator $\boldsymbol{G}(\boldsymbol{z};\boldsymbol{\theta}_g)$ to capture the rich prior information of natural faces, where $\boldsymbol{z}$ is the input latent vector, $\boldsymbol{\theta}_g$ is model parameter. We sample $\boldsymbol{z}$ from the hypersphere latent space to generate intermediate super-resolution faces $\tilde{\mathcal{Y}} = \{\tilde{\mathbf{y}}_i\}_{i=1}^{|\tilde{\mathcal{Y}}|}$. Meanwhile, the corresponding discriminator $\boldsymbol{D}(\tilde{\mathcal{Y}}|\mathcal{X};\boldsymbol{\theta}_d)$ is used to extract semantic features with fixed parameters $\boldsymbol{\theta}_d$. Our goal is optimizing the latent vector $\boldsymbol{z}$ while online finetuning $\boldsymbol{\theta}_g$ in order to restrict $\boldsymbol{z}$ to the optimal location where photo-realistic super-resolution faces can be generated. Therefore, the super-resolution problem can be formulated as:

$$\tilde{\mathbf{y}} = \boldsymbol{G}(\boldsymbol{z};\boldsymbol{\theta}_g), \tag{1.1}$$

$$\tilde{\mathbf{y}}\downarrow_s \doteq \mathbf{x}, \tag{1.2}$$

$$\boldsymbol{D}(\tilde{\mathbf{y}}\downarrow_s;\boldsymbol{\theta}_d) \doteq \boldsymbol{D}(\mathbf{x};\boldsymbol{\theta}_d), \tag{1.3}$$

where $\doteq$ means "equivalence" in some metric, $\downarrow_s$ is a downscaling operation, and s is the scale factor. Eq.(1.1) tries to generate super-resolution $\tilde{\mathbf{y}}$ satisfying two rules by making the ground-truth low-resolution $\mathbf{x}$ and the corresponding $\tilde{\mathbf{y}}\downarrow_s$ as close as possible, in the view of appearance perception Eq.(1.2) and semantic perception Eq.(1.3). Thus, our final goal is optimizing the following objective in a unified way:

$$\boldsymbol{G}(\boldsymbol{z}^*;\boldsymbol{\theta}_g^*) = \min_{\boldsymbol{z},\boldsymbol{\theta}_g}\{\lambda_p\mathcal{L}_p + \lambda_s\mathcal{L}_s + \lambda_g\mathcal{L}_g\}, \tag{2}$$

where $\mathcal{L}_p$ measures the pixel distance between $\tilde{\mathbf{y}}\downarrow_s$ and $\mathbf{x}$ to meet rule one Eq.(1.2). $\mathcal{L}_s$ constraints these two faces to be close in the semantic feature space to meet rule two Eq.(1.3). $\mathcal{L}_g$ is the geodesic cross loss adopted in PULSE [15] to prevent $\boldsymbol{z}$ from deviating too far from the latent space. λ_p, λ_s and λ_g are weights to balance different losses and set to be 100, 0.5 and 500, respectively.

Next, we introduce how to conduct regularized latent space exploration for discriminative face super-resolution in

Table 1. Comparison with other super-resolution approaches based on unsupervised (left) and supervised learning (right).

Scale	Dataset	Metric	Bilinear	mGANprior [18]	PULSE [15]	**Ours**	VDSR [6]	ESRGAN [13]	FSRNet [11]
8×	CelebA	PSNR↑	25.84	21.29	22.54	23.52	23.18	23.74	25.08
		SSIM↑	0.73	0.53	0.54	0.56	0.76	0.63	0.56
		LPIPS↓	0.57	0.32	0.28	0.25	0.28	0.30	0.23
	TinyFace	NIQE↓	15.01	13.28	9.81	8.94	15.12	16.84	16.53
16×	CelebA	PSNR↑	22.73	20.53	21.43	21.74	22.42	21.83	23.04
		SSIM↑	0.56	0.50	0.48	0.49	0.59	0.46	0.62
		LPIPS↓	0.65	0.36	0.30	0.27	0.33	0.31	0.28
	TinyFace	NIQE↓	18.44	14.55	11.98	10.55	16.95	15.64	15.90

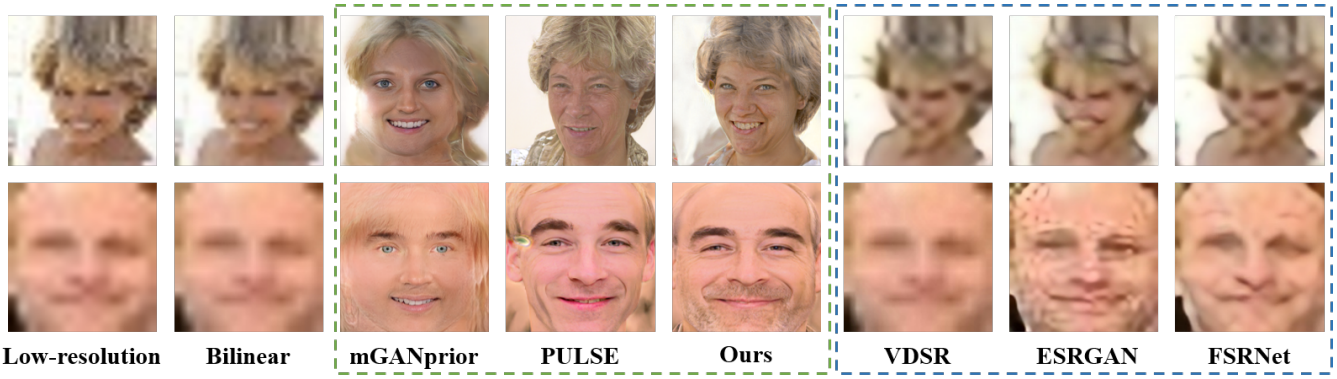

Fig. 3. Comparison with self-supervised and supervised approaches on CelebA (top) and TinyFace (bottom).

detail. Begin with randomly initializing $\boldsymbol{z}^{(1)}$ and inputting it to the pretrained generator. First, the pixel loss is measured:

$$\mathcal{L}_p = \|\tilde{\mathbf{y}} \downarrow_s - \mathbf{x}\|_2 , \tag{3}$$

where $\|*\|_2$ is l_2 norm operator. In the meantime, the semantic loss $\mathcal{L}_s$ is measured to enforce the features obtained closer during inference by l_1 norm operator and is defined as:

$$\mathcal{L}_s = \sum_{i=1}^{b} \left\| \boldsymbol{D}^{(i)}(\tilde{\mathbf{y}} \downarrow_s ; \boldsymbol{\theta}_d) - \boldsymbol{D}^{(i)}(\mathbf{x}; \boldsymbol{\theta}_d) \right\|_1 , \tag{4}$$

where b is the total number of blocks we use in $\boldsymbol{D}$, $\boldsymbol{D}^{(i)}$ returns the feature maps extracted from the ith-block of $\boldsymbol{D}$. It is noted that we use the corresponding $\boldsymbol{D}$ of the pretrained GAN which is optimized in an adversarial way instead of introducing other models. Thus, the acquired features can accurately represent the location of faces in the semantic space.

We optimize the latent vector $\boldsymbol{z}$ based on the total loss by projecting gradient descent on the latent space to guide the search. Meanwhile, $\boldsymbol{\theta}_g$ is online finetuned to make better use of the GAN prior because fixing $\boldsymbol{\theta}_g$ may suffer from generator capability limitations. During the iteration, $\boldsymbol{z}$ is dragged and restricted in order to make the downsampled $\tilde{\mathbf{y}}$ and the ground-truth $\mathbf{x}$ closer in both pixel and semantic space jointly. Also super-resolution faces are continually generated from the feasible latent space and evaluated until reaching the optimal solution $\tilde{\mathbf{y}}^* = \boldsymbol{G}(\boldsymbol{z}^*; \boldsymbol{\theta}_g^*)$.

3. EXPERIMENTS

To validate the effectiveness of our approach, we conduct experiments on three datasets (CelebA-HQ [20], CelebA [21] and TinyFace [22]) and comparisons with six approaches. In the experiments, we use StyleGAN pretrained on Flicker Face HQ (FFHQ) dataset [23] as GAN prior, and the last three convolution blocks of discriminator is employed to calculate the semantic loss. For a given low-resolution face, we iterate 100 times for optimization, starting with a random initialization. We use a spherical optimizer with learning rate 0.4, and online finetune the $\boldsymbol{\theta}_g$ adopting Adam optimizer with learning rate 5e-5. Three reference-based metrics (PSNR, SSIM [24] and LPIPS [25]) and no-reference NIQE [26] are used for evaluation. LPIPS measures the deep feature distance with pretrained AlexNet [27].

3.1. Comparison with Other Approaches

We conduct comparisons on CelebA and TinyFace with six super-resolution approaches, including traditional **Bilinear** interpolation, two unsupervised approaches (**PULSE** [15] and **mGANprior** [18]), and three supervised approaches (**VDSR** [6] based on CNN, **ESRGAN** [13] based on GAN and **FSRNet** [11] employing facial information). We use their default settings for all approaches and test with scale factor of $8\times$, $16\times$. For CelebA, each face is aligned and resized to 128×128 to simulate high-resolution faces [2].

Then, we apply bicubic downscaling to get low-resolution ones. We use part of high-resolution and low-resolution face pairs for supervised training and select the other 500 faces to test. TinyFace only consists of low-resolution faces with average resolution of 20×16. We resize the images to 16×16 to simulate real-world low-resolution faces. The results are shown in Tab. 1 and Fig. 3.

Comparison with Unsupervised Approaches. On CelebA, we can see that our approach consistently outperforms unsupervised approaches in both quantitative and qualitative metrics. In spite of higher PSNR achieved by Bilinear interpolation, the results are very blurry. On TinyFace, our approach remarkably surpasses other approaches in NIQE measure which is based on the quality-aware features derived from the natural scene statistical model, *e.g.*, at least reduction of 0.87 in NIQE score, implying the effectiveness of our approach in real-world scenarios. Moreover, our approach can provide clearer results compared with PULSE, seeing in Fig. 3. The main reason comes from the semantic loss measured by the discriminator in regularizing latent space exploration. In addition, our approach can perform inference efficiently. It costs less inference time while delivering better super-resolution performance, as shown in Fig. 4.

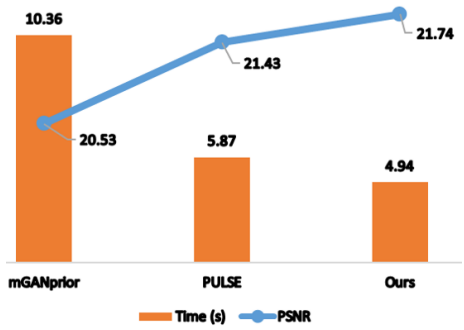

Fig. 4. Inference time of $16\times$ super-resolution on CelebA

Comparison with Supervised Approaches. We use the models pretrained on CelebA for evaluation. Generally, supervised approaches achieve higher metrics in PSNR and SSIM and more realistic results, since they rely on the availability of low- and high-resolution face pairs for training. By contrast, our approach inherits ill-posed super-resolution process due to the absence of high-resolution faces for output reference and may lead the super-resolved faces looking different from ground-truth high-resolution ones. However, our approach still has considerable advantages over supervised approaches, *e.g.*, achieving better LPIPS which indicates clearer results. Most importantly, our approach can learn to super-resolve low-resolution faces naturally without high-resolution ones (*e.g.*, TinyFace), which is helpful in practical scenarios where such face pairs usually are unavailable.

Table 2. Effects of discriminator semantic loss.

Scale	Metric	Without $\mathcal{L}_s$	With $\mathcal{L}_s$	Improvement
$16\times$	PSNR↑	22.66	23.04	0.38
	SSIM↑	0.54	0.55	0.01
	LPIPS↓	0.24	0.20	0.04
$32\times$	PSNR↑	19.93	20.78	0.85
	SSIM↑	0.42	0.44	0.02
	LPIPS↓	0.27	0.24	0.03
$64\times$	PSNR↑	18.93	19.15	0.22
	SSIM↑	0.33	0.34	0.01
	LPIPS↓	0.30	0.29	0.01

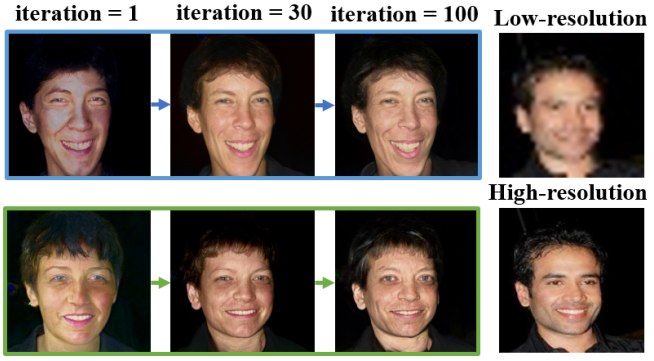

Fig. 5. An example generated without (top) and with (bottom) discriminator semantic loss.

3.2. Effects of Discriminator Semantic Loss

We use higher resolution CelebA-HQ to validate the effectiveness of semantic loss, in order to show the ability of our approach on handling with much larger scale factors. We remove the semantic loss $\mathcal{L}_s$ and compare it with the original one. We select the first 500 faces and resize them to 16×16 to simulate low-resolution faces using bicubic downscaling. As shown in Tab. 2, the model with discriminator semantic loss achieves better PSNR, SSIM and LPIPS. LPIPS focuses on visual quality of faces. This means our discriminator semantic loss has improved quantitative metrics in both pixel and semantic indicators, implying the discriminative ability of our approach. Some examples can be seen in Fig. 5

4. CONCLUSION

In this work, we discover that existing self-supervised face super-resolution approaches usually have poor control over appearance. To address that, we propose a regularized latent space exploration approach to facilitate self-supervised face super-resolution, in which a pretrained GAN is fully used by applying discriminator to regularize the latent space exploration in an iterative optimization way. Extensive experiments have proven the effectiveness of our approach. Future work is extending the approach for more generation applications.

5. REFERENCES

[1] Shiming Ge, Shengwei Zhao, Chenyu Li, and Jia Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE TIP*, vol. 28, no. 4, pp. 2051–2062, 2019.

[2] Junjun Jiang, Chenyang Wang, Xianming Liu, and Jiayi Ma, "Deep learning-based face super-resolution: A survey," *ACM Computing Surveys*, vol. 55, no. 1, pp. 13:1–13:36, 2023.

[3] Zhihao Wang, Jian Chen, and Steven CH Hoi, "Deep learning for image super-resolution: A survey," *IEEE TPAMI*, vol. 43, no. 10, pp. 3365–3387, 2021.

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE TPAMI*, vol. 38, no. 2, pp. 295–307, 2015.

[5] Chao Dong, Chen Change Loy, and Xiaoou Tang, "Accelerating the super-resolution convolutional neural network," in *ECCV*, 2016, pp. 391–407.

[6] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *IEEE CVPR*, 2016, pp. 1646–1654.

[7] Bee Lim, Sanghyun Son, and Heewon Kim, "Enhanced deep residual networks for single image super-resolution," in *IEEE CVPRW*, 2017, pp. 136–144.

[8] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *IEEE CVPR*, 2017, pp. 624–632.

[9] Yulun Zhang, Kunpeng Li, Kai Li, and *et al.*, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018, pp. 286–301.

[10] Tao Dai, Jianrui Cai, Yongbing Zhang, and *et al.*, "Second-order attention network for single image super-resolution," in *IEEE/CVF CVPR*, 2019, pp. 11065–11074.

[11] Yu Chen, Ying Tai, Xiaoming Liu, and *et al.*, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *IEEE/CVF CVPR*, 2018, pp. 2492–2501.

[12] Xin Yu and Fatih Porikli, "Ultra-resolving face images by discriminative generative networks," in *ECCV*, 2016, pp. 318–333.

[13] Xintao Wang, Ke Yu, Shixiang Wu, and *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks," in *ECCV Workshop*, 2018, pp. 63–79.

[14] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, and *et al.*, "Glean: Generative latent bank for large-factor image super-resolution," in *IEEE/CVF CVPR*, 2021, pp. 14245–14254.

[15] Sachit Menon, Alexandru Damian, Shijia Hu, and *et al.*, "Pulse: Self-supervised photo upsampling via latent space exploration of generative models," in *IEEE/CVF CVPR*, 2020, pp. 2437–2445.

[16] Xingang Pan, Xiaohang Zhan, Bo Dai, and *et al.*, "Exploiting deep generative prior for versatile image restoration and manipulation," in *ECCV*, 2020, pp. 262–277.

[17] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Deep image prior," in *IEEE/CVF CVPR*, 2018, pp. 9446–9454.

[18] Jinjin Gu, Yujun Shen, and Bolei Zhou, "Image processing using multi-code gan prior," in *IEEE/CVF CVPR*, 2020, pp. 3012–3021.

[19] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[20] Tero Karras, Timo Aila, and Samuli Laine, "Progressive growing of gans for improved quality, stability, and variation," *ICLR*, 2018.

[21] Ziwei Liu, Ping Luo, and Xiaogang Wang, "Deep learning face attributes in the wild," in *IEEE ICCV*, 2015, pp. 3730–3738.

[22] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong, "Low-resolution face recognition," in *ACCV*, 2018, pp. 605–621.

[23] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE/CVF CVPR*, 2019, pp. 4401–4410.

[24] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.

[25] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE/CVF CVPR*, 2018, pp. 586–595.

[26] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, "Making a "completely blind" image quality analyzer," *IEEE SPL*, vol. 20, no. 3, pp. 209–212, 2012.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.