Pro-PULSE: Learning Progressive Encoders of Latent Semantics in GANs for Photo Upsampling

Yang Zhou, Yangyang Xu, Yong Du, Qiang Wen, and Shengfeng He, *Senior Member, IEEE*

***Abstract*— The state-of-the-art photo upsampling method, PULSE, demonstrates that a sharp, high-resolution (HR) version of a given low-resolution (LR) input can be obtained by exploring the latent space of generative models. However, mapping an extreme LR input (16^2) directly to an HR image (1024^2) is too ambiguous to preserve faithful local facial semantics. In this paper, we propose an enhanced upsampling approach, Pro-PULSE, that addresses the issues of semantic inconsistency and optimization complexity. Our idea is to learn an encoder that progressively constructs the HR latent codes in the extended $\mathcal{W}+$ latent space of StyleGAN. This design divides the complex 64× upsampling problem into several steps, and therefore small-scale facial semantics can be inherited from one end to the other. In particular, we train two encoders, the base encoder maps latent vectors in $\mathcal{W}$ space and serves as a foundation of the HR latent vector, while the second scale-specific encoder performed in $\mathcal{W}+$ space gradually replaces the previous vector produced by the base encoder at each scale. This process produces intermediate side-outputs, which injects deep supervision into the training of encoder. Extensive experiments demonstrate superiorities over the latest latent space exploration methods, in terms of efficiency, quantitative quality metrics, and qualitative visual results.**

***Index Terms*— Photo upsampling, GANs, progressive learning, latent space.**

I. Introduction

RECENT advances in image editing and transformation are driven by the success of latent space exploration in generative models. These methods fully leverage the powerful generation ability of GANs, especially StyleGAN [15], by discovering semantically meaningful directions in the latent space [10], [24], [26], or inverting the generation process by estimating the latent codes of an input real image [1], [2], [25], [38].

These successes are extended to a fundamental, ill-posed super-resolution problem. A photo upsampling approach [21], namely PULSE, is proposed to transform an extreme low-resolution (LR) input into a sharp high-resolution (HR) image. Instead of inverting an HR real image to the latent codes of GANs, PULSE aims to discover the latent code of the LR image that can produce a consistent HR image. In particular, it self-supervisedly optimizes the latent code by enforcing the similarity between input LR and the downscaled output. In this way, although the generated HR images are not exactly the same as the ground truth, it does produce sharp and realistic results. When applying it on face images, it is known as face hallucination [27], [39]. Similar to PULSE, we take face hallucination as an application in this paper.

Notwithstanding the achieved high upsampling quality, PULSE suffers from a severe complexity problem, as it relies on localizing the optimal latent vector for each input LR image. A concurrent work, pSp network [25], is proposed to learn an encoder that maps several types of inputs (LR is one of them) to the latent vectors for specific tasks. This is undisputed faster than optimizing the image-specific latent code as PULSE does. However, both PULSE and pSp share a common limitation that they cannot capture small-scale facial semantics in the LR image. This is because low resolution is akin to myopic vision in that fine visual features are not discernable. Converting such a blurry input to an HR image is extremely ambiguous, as there exists multiple HR images correspond to exactly the same LR input. Direct mapping exhibits spatial shift and semantic inconsistency. As shown in Fig. 1c and 1d, eyeglasses are too blurry to be correctly mapped to the HR images, resulting in either blackened eye regions or failures in eyeglasses recovery.

In this paper, we follow the same spirit of PULSE but propose an alternative solution, namely *Pro-PULSE*, to solve the above notorious and universal problem. Our idea is to progressively predict latent vectors for different scales, such that the complex 64× mapping problem can be divided into several simpler 2× upsampling operations. In this way, fine details of LR can be easily inherited and refined step-by-step.

Manuscript received February 22, 2021; revised September 19, 2021 and December 13, 2021; accepted December 22, 2021. Date of publication January 11, 2022; date of current version January 18, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61972162 and Grant 62102381, in part by the Guangdong International Science and Technology Cooperation Project under Grant 2021A0505030009, in part by the Guangdong Natural Science Foundation under Grant 2021A1515012625, in part by the Shandong Natural Science Foundation under Grant ZR2021QF035, in part by the Guangzhou Basic and Applied Research Project under Grant 202102021074, in part by the Fundamental Research Funds for the Central Universities under Grant 202113035, in part by the China Postdoctoral Science Foundation under Grant 2020M682240 and Grant 2021T140631, and in part by the China Computer Federation (CCF)-Tencent Open Research Fund under Grant CCF-Tencent RAGR20210114. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dong Tian. *(Yang Zhou and Yangyang Xu contributed equally to this work.) (Corresponding author: Shengfeng He.)*

Yang Zhou, Yangyang Xu, Qiang Wen, and Shengfeng He are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: matrixgle19@gmail.com; cnnlstm@gmail.com; csqiangwen@gmail.com; hesfe@scut.edu.cn).

Yong Du is with the School of Computer Science and Technology, Ocean University of China, Qingdao 266005, China (e-mail: csyongdu@ouc.edu.cn).

Digital Object Identifier 10.1109/TIP.2022.3140603

(a) Input LR (b) GT HR (c) pSp [25] (d) PULSE [21] (e) Pro-PULSE (f) Our Side-outputs

Fig. 1. Our Pro-PULSE framework allows capturing small-scale facial semantics (*e.g.*, eyeglasses) in low-resolution input for photo upsampling. We produce inherited semantics and progressively refined upsampled faces as side outputs, with successive intermediate results being more concise.

In particular, we have two encoders that map LR images into the latent vectors of StyleGAN [15]. The first base encoder learns to generate a latent vector in $\mathcal{W}$ space, which serves as a foundation of generating HR results in all scales. We then train the second scale-specific encoder in the $\mathcal{W}+$ space that gradually replaces the previous latent vector obtained from the base encoder, until all the scales are fed with scale-specific latent vectors. Rather than learning all scales simultaneously, our incremental nature allows the training to first discover large-scale structure of the image distribution and then shift attention to increasingly finer scale detail. To ensure semantically correct inheritance during the progressive process, we revise StyleGAN to contain a deeply-supervised generator that produces a side-output with guidance at each scale to aid the training of intermediate latent codes. This not only enables multi-scale upsampling, but also enforces perceptual consistency of side-outputs in all resolutions. With our multi-level analysis and progressive inference, the model is able to discover semantically meaningful small-scale facial components, as shown in Fig. 1e.

Our approach has conceptual similarity to previous progressive [14], [30] or cascaded approaches [5]. We share the same observation that the complex large scaling factor mapping problem can be easier to learn in steps. However, a crucial difference is that we have an emphasis on multi-level analysis in the latent spaces of GANs, instead of learning a direct LR-HR mapping. In addition, we demonstrate the proposed method in the application of face hallucination, but it is general to other context.

Overall, our contributions are three-fold:

- We propose a novel progressive upsampling framework, Pro-PULSE, for mapping an extreme LR input to a latent vector of StyleGAN that produces faithful and semantically correct HR image.
- We tailor a progressive training scheme for exploring latent spaces of GANs, by involving two separate encoders and producing side-outputs, to ease the learning ambiguity and inherit fine details from the lowest level.
- Extensive quantitative and qualitative experiments demonstrate that the favorable characteristics of our underlying techniques manifest in Pro-PULSE being both semantically accurate and computational efficient.

II. Related Works

A. Image Super-Resolution

Image super-resolution aims at recovering a realistic HR image based on its blurry LR counterpart. Early works tackle this problem using statistical techniques [4], [12], [28], [32]. The same as other applications, image super-resolution achieves a rapid development with the advance of convolutional neural networks (CNNs). A typical solution is to learn a direct mapping function between paired LR and HR images using a CNN. Dong *et al*. [9] proposed the first CNN for single image super-resolution. Li *et al*. [18] develop an image super-resolution feedback network that refines low-level representations with high-level information. Except for natural images, face super-resolution/hallucination utilize the extra facial knowledge in the upsampling process. Zhu *et al*. [39] proposed a unified framework that estimates the dense correspondence fields to hallucinate face images. Yu *et al*. [35] design a two-branches network, one is for recovering face HR and the other is to predict salient regions for better reconstruction. Chen *et al*. [6] introduced the facial landmark and parsing map as the guidance signal into the upsampling process. Although considerable progress has been made, they still suffer from the following issues: 1) most of those works use MSE loss to train a CNN, which leads to a blurring effect and lost vivid details on the recovered HR face images; 2) state-of-arts super-resolution methods can deal with at most $8\times$ upscaling factors, which limits the practical usages in many scenarios.

B. GAN Inversion

Generative models take random latent codes as input and synthesize various images. To better meet the needs of the real-world applications, GAN inversion is proposed that maps a real image to the latent space of a pre-trained generator that produces a consistent result with the input image [1], [2], [7], [25], [38].

Based on how they search for the latent code, these methods can be categorized in two classes. The first one is optimization-based methods, in which they process a single image by optimizing its latent code directly with pixel-wise reconstruction loss [1], [2], [20]. PULSE adopts this idea on photo upsampling with a fixed StyleGAN. However, it ignores the highly structured semantics that emerges in the generation process of StyleGAN [31]. Besides, optimization-based methods are time-consuming which limits its scalability. The second class is learning-based that learns a mapping function from image space to latent space directly. Particularly, Zhu *et al*. [38] proposed an encoder that ensures the inverted code in the semantic domain of the original latent space. Richardson *et al*. [25] demonstrated various applications can be well-solved by mapping a real image to $\mathcal{W}+$ latent space of StyleGAN. However, all the previous methods neglect the fact that directly mapping a sparse input (either the random code or LR image) to a super-dense high-resolution output is ambiguous and intractable. We overcome this universal problem by introducing the progressive encoder that inherits structured semantics in steps.

The two most related works, PULSE [21] and pSp [25], share the same idea that search an optimal latent code in latent space for image upsampling as ours. However, PULSE is an optimization-based method, which is time-consuming to optimize every input into a target latent code. On the contrary, we explore the StyleGAN latent space from a learning perspective that can significantly accelerate the inference speed. pSp proposes a general framework to learn an encoder that maps the input from domain X to the StyleGAN domain (X can be the LR image). Both PULSE and pSp are one-step mapping methods that neglect a critical fact that fine visual features in the LR input are not discernable. Hence, they may not produce semantically consistent outputs. In contrast, we tailor a progressive encoder and training scheme to remedy the semantic inheritance problem.

III. Approach

A. Preliminary

Before getting into our method, we first briefly introduce the $\mathcal{W}$ and $\mathcal{W}+$ space of StyleGAN, as our method relies heavily on them. Generally, a generative model takes the latent code from the Gaussian distribution $\mathcal{Z}$ as input. This noise input directly control the output of a pretrained generative model, therefore latent space exploration methods aim to figure out the semantic structure within the latent space. StyleGAN [15] maps this latent code from $\mathcal{Z}$ space to another space with a few fully-connected layers, called $\mathcal{W}$ space, before feeding to generator for a better semantic disentanglement. Abdal *et al*. [1] find that using the same latent code in the $\mathcal{W}$ space for all layers limits the representation ability, thus they extend the $\mathcal{W}$ space to $\mathcal{W}+$ space, where a $w+$ latent code consists of a combination of layer-aware w codes. Compared with the W space, $W+$ space improves the representative ability significantly. As a result, many latent space exploration works are applied on the $W+$ space of pre-trained StyleGAN due to its strong expressiveness.

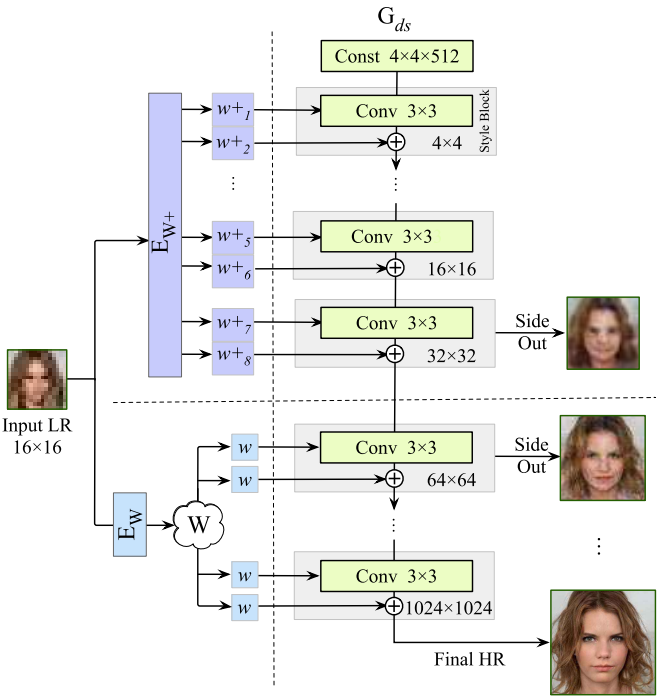


Fig. 2. Overview of our Pro-PULSE. The right part is our modified generator with multi-side-outputs. Encoder E_W learns a latent vector in $\mathcal{W}$ space with LR input, which serves as a foundation of generating HR results in all scales. E_{W+} maps the LR input to the $\mathcal{W}+$ space that gradually replaces the w latent code obtained from E_W. *Note this figure shows the starting status of our progressive process, in which the w latent code for 32×32 is first replaced by an advanced scale-specific $w+$ latent code. At the end of our training, all the w codes will be replaced by the layer-specific $w+$ codes.*

B. Pipeline

Given an LR photo, the proposed Pro-PULSE, which is illustrated in Fig. 2, aims to generate a high-resolution version with more small-scale facial semantics being captured. To achieve this goal, we divide the model within three sub-networks. The first one is a deeply supervised generator, which additionally produces multiple side outputs. These outputs are particularly designed for inheriting facial semantics from each other and thus enables more concise and plausible reconstructions. The second one is a base encoder that maps the LR image to the $\mathcal{W}$ space, while the last is our final target model, *i.e.*, scale-specific encoder, which progressively encodes the input to the $\mathcal{W}+$ space, and is trained under the guidance of the other two modules.

C. Deeply-Supervised Generator

StyleGAN profits from highly structured semantics and has proven to be successful in facial image generation [1], [31]. However, those semantics are presented implicitly, and therefore could be underutilized to translate an LR image to its HR version. To tackle this problem, we transform the generator of a pre-trained StyleGAN into a new one with multiple appended branches, *i.e.*, the deeply-supervised generator G_{ds}, as illustrated in Fig. 2. Specifically, in the proposed generator, each style block is connected with a convolution branch to correspond to a different output resolution

(from 32×32 to 512×512), further producing a side output. Such a design enables to generate images with the same content but different resolutions. In this way, the structured information embedded in those multi-resolution images could be provided explicitly as guidance for recovery of higher resolution versions.

Note that we fix the parameters of original StyleGAN and only optimize those of the additional branches during training of the generator. The adversarial learning cannot ensure a pixel-level consistency, such that the acquired structure semantics would not be used effectively. We instead train the branches with forcing a pixel-wise consistency. Besides, we also utilize a perceptual loss [13] for maintaining the subtle facial details. As a result, the total objective $\mathcal{L}_{G_{ds}}$ to train the generator is given by

$$\mathcal{L}_{G_{ds}} = \mathcal{L}_{pix} + \lambda_{pp}\mathcal{L}_{pp}, \tag{1}$$

where λ_{pp} denotes a balancing factor. Regarding the pixel-wise consistency loss $\mathcal{L}_{pix}$, we calculate the L_2 distance between a side-output image I_o^r (with a resolution of r^2) and the final output face I_o^{1024} (with a resolution of 1024^2) which is downsampled to the same resolution of I_o^r. Then $\mathcal{L}_{pix}$ is formulated as

$$\mathcal{L}_{pix} = \frac{1}{M}\sum_r \frac{1}{S^r}\|I_o^r - DS^r(I_o^{1024})\|_2, \tag{2}$$

where $DS^r(\cdot)$ denotes a downsampling operation to the target resolution r^2, M represents the number of elements in the set of r, $M = 5$ in our case because of $r \in \{32, 64, \ldots, 512\}$, and S^r means the size of image or feature map under resolution r^2.

In regard to the perceptual loss $\mathcal{L}_{pp}$, we compute the L_2 distance between features produced by I_o^r and I_o^{1024} after feeding to a pre-trained VGG-16 network $\Phi_{VGG}(\cdot)$.

$$\mathcal{L}_{pp} = \frac{1}{M}\sum_r \frac{1}{S^r}\|\Phi_{VGG}(I_o^r) - \Phi_{VGG}(DS^r(I_o^{1024}))\|_2. \tag{3}$$

Here we select the features produced by `conv4_2` layer of the VGG-16 network.

D. Base Encoder for $\mathcal{W}$ Space

After training the deeply-supervised generator, we tend to extract and utilize the features from LR images to reconstruct multiple higher resolutions of images. A seemingly good method is to encode LR images into a $\mathcal{W}+$ latent space which is demonstrated more disentangled for semantic editing [1]. However, training such an encoder is an intractable problem due to the complexity of $\mathcal{W}+$ space. On the other hand, we empirically find that learning an encoder to $\mathcal{W}$ space is much easier (latent code in $\mathcal{W}$ is $18\times$ smaller than in $\mathcal{W}+$). As a compromise, we first learn an base encoder E_W that maps the input LR image to its latent code w, which is in $\mathcal{W}$ space and would provide guidance for further training another encoder to $\mathcal{W}+$ space. Note that during training of the base encoder, the parameters of the generator are fixed. Then we formulate the pipeline as

$$\begin{aligned} w &= E_W(I_{LR}), \\ Set(I_o^{r'}) &= G_{ds}(R^q(w)), \end{aligned} \tag{4}$$

where I_{LR} indicates the input LR images and $Set(I_o^{r'})$ denotes the set of generated images with resolutions of r'^2, $r' \in \{32, 64, \ldots, 1024\}$, and $R^q(\cdot)$ denotes a repetitive operation in q times. Following the setting in StyleGAN, we set $q = 18$ so that the generated images can reach the maximum resolution of 1024×1024.

With aid of the multi-branch design, extra structural information could be exploited for retrieving a better latent code. We train this base encoder E_W with three losses: pixel-wise reconstruction loss $\mathcal{L}_{rec}$, identity loss $\mathcal{L}_{id}$, and adversarial loss $\mathcal{L}_{adv}$. The total loss function is then defined as follows:

$$\mathcal{L}_{E_W} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{id}\mathcal{L}_{id} + \lambda_{adv}\mathcal{L}_{adv}, \tag{5}$$

where λ_{rec}, λ_{id} and λ_{adv} are the balancing weights of reconstruction loss, identity loss, and adversarial loss respectively.

The first one is the pixel-wise reconstruction loss that encourages all the generated images to downsample consistently with the input LR image. Bounded by this constraint, the semantic knowledge could be preserved, and we define this loss as follows:

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_{r'} \|I_{LR} - DS^{r'}(I_o^{r'})\|_2, \tag{6}$$

where $N = 6$ is the number of resolution scales.

We also utilize the identity loss for maintaining the identity information between the final HR and the GT faces, that is

$$\mathcal{L}_{id} = 1 - \cos\langle \Phi_{id}(I_o^{1024}), \Phi_{id}(I_{GT})\rangle, \tag{7}$$

where $\Phi_{id}(\cdot)$ denotes a pre-trained ArcFace network for face recognition [8], I_{GT} denotes the ground truth faces, and $\cos\langle\cdot,\cdot\rangle$ denotes the cosine similarity between two inputs.

Besides, adversarial loss between the final HR image and GT face is also involved in our training process to conform real face data distributions, which can be formulated as follows:

$$\mathcal{L}_{adv} = -\mathop{\mathbb{E}}_{I_o^{1024}\sim P_g}[D(I_o^{1024})], \tag{8}$$

$$\mathcal{L}_D = \mathop{\mathbb{E}}_{I_o^{1024}\sim P_g}[D(I_o^{1024})] - \mathop{\mathbb{E}}_{I_{GT}\sim P_r}[D(I_{GT})], \tag{9}$$

where P_g and P_r denotes the distribution of generated data and real data respectively, and $D(\cdot)$ denotes the discriminator. Note that $\mathcal{L}_D$ is the loss of the discriminator.

We do not use the perceptual loss in the training process as we found it cannot boost the performance. This is because the perceptual loss reduces the features differences between target and synthesis results, while this could be compensated by the combination of identity loss and adversarial loss.

E. Scale-Specific Encoder for $\mathcal{W}+$ Space

Once we have trained the base encoder E_W, a 512-dimensional w latent code, which serves as a foundation of generating SR results in all scales, could be obtained. Together with the multi-side-output images, we consequently train a scale-specific encoder E_{W+}. Different from E_W that maps the input LR face image to $\mathcal{W}$ space, E_{W+} maps the input image to a latent code lain in a more complicated $\mathcal{W}+$ space, by treating the latent codes in all scales separately. Previous works [25], [38] learn the whole $w+$ latent codes

directly, which is unstable and with high training ambiguity. In contrast, we learn the $w+$ latent codes in a progressive manner, producing the structural information at low-level scales while refining high-resolution details in the following layers.

Fig. 2 depicts the beginning status when training the E_{W+} encoder. Specifically, we first train E_{W+} that maps the input LR image to immediate resolutions $(r^i)^2$ such as 32×32, and produce a part of latent codes $w+^{r^i}$ with the maximum resolution of r^i. Then to obtain the final HR faces, we supplement the rest $w+^{(r''\backslash r^i)}$ codes with the duplicates of the w vector, where $r'' \in \{4, 8, \ldots, 1024\}$. In other words, we concatenate the latent code $w+^{r^i}$ with the repetitive w code and feed the concatenated code into G_{ds}, and the pipeline is given by

$$\begin{aligned} w+ &= E_{W+}(I_{LR}), \\ Set(I_o^r) &= G_{ds}(C(w+^{r^i}, R^{2q}(w))), \\ & r^i \in \{4, 8, \ldots, 1024\}, \quad q = \#\{r''\backslash r^i\}, \end{aligned} \tag{10}$$

where $C(\cdot)$ represents concatenate operation, $R(\cdot)$ is the repetitive operation, $\{r''\backslash r^i\}$ denotes the difference set between r'' and r^i, $2q$ denotes the repetitions and is the output of the counting operator #. In this way, the training is stabilized by using the pre-trained w, while the encoder can focus on a simpler training task of $2\times$ mapping. We gradually replace the previous latent w vector obtained from E_W until all the style blocks are fed with scale-specific latent vectors $w+$. Note that we train the scale-specific encoder E_{W+} by using the same losses as E_W.

F. Implementation Details

We revise the StyleGAN2 [2] pre-trained on the FFHQ [15] dataset to equip our deeply-supervised generator. Each side-output branch consists of 3 convolutional layers, and we set the kernel size as 1×1 with stride $= 1$. The numbers of channels are set as 128, 64 and 3 respectively to transform the feature maps into an RGB image gradually. Meanwhile, our base encoder E_W maps an input LR image to $\mathcal{W}$ space and serves as a basic encoder. The structure of E_W is shown in Fig. 3. It consists of several convolutional layers and the residual channel attention blocks (RCAB) [37]. We set the first convolutional layer with a kernel size $= 3 \times 3$ and stride $= 1$. The following transposed convolutional layer has the same kernel size and stride. Then we add 4 groups of two RCABs and its detailed architecture can be seen in [37]. We also follow [37] that use a residual connection after the RCABs. At the end of the model, there are 3 convolutional layers converting the feature map into a $1 \times 1 \times 512$ output, reshaped as a 1×512 code. The (kernel size, stride) are set as $(4, 4)$, $(4, 4)$ and $(2, 1)$, respectively. We set their channel numbers as 512. Fig. 4 shows the architecture of the scale-specific encoder E_{W+}. It is formed by repeating two RCABs and the subsequent three convolutional layers of E_W.

IV. Experiments

We conduct our experiments on the PyTorch platform with an Nvidia GeForce Titan Xp GPU. Adam [16] optimizer with

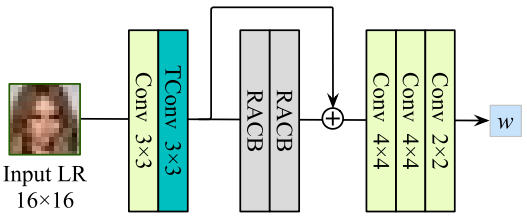


Fig. 3. Architecture of E_W.

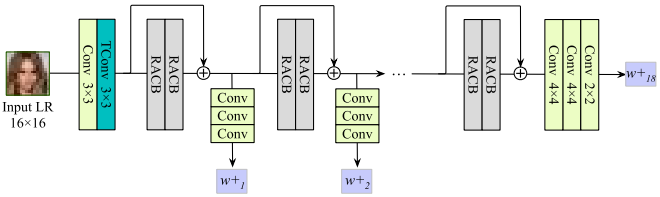


Fig. 4. Architecture of E_{W+}.

the learning rate of $\alpha = 1e-4$ is adopted for training the G_{ds}, E_W, and E_{W+}. We set $\lambda_{pp} = 1$, $\lambda_{rec} = 1$, $\lambda_{id} = 1$, and $\lambda_{adv} = 0.01$ empirically. We train the G_{ds} and E_W with 50,000 iterations. Besides, We train the E_{W+} with 10,000 iterations in each resolution scale, resulting in 60,000 iterations totally.

A. Experimental Settings

1) Dataset: FFHQ [15] dataset contains 70,000 face images with resolution of 1024×1024, which is the key to generating high-quality faces of StyleGAN. We train the proposed progressive encoders on the CelebA-HQ dataset [17] using the same setting as the pSp network [25]. It consists of 30,000 images with the resolution of 1024×1024. We follow the standard train-test split, resulting in 24,183 and 5,817 images for training and testing respectively. We use an scaling factor of $64\times$ to convert an 16×16 LR image to 1024×1024 HR for comparisons.

2) Evaluation Metrics: Face hallucination is not suited to be evaluated by pixel-wise measurements like PSNR. We use two metrics, Fréchet Inception Distance (FID) [11] and Naturalness Image Quality Evaluator (NIQE) [23], for evaluating the human-perceptual quality of upsampling results. In particular, FID computes the Wasserstein-2 distance between the features of GT and recovered HR faces. An InceptionV3 model [29] pre-trained for image classification is used as a feature generator in here. NIQE is a completely blind assessment with no request for the GT image, which measures the image quality using perceptual features extracted from the recovered SR results. We also report the runtime and FLOPs of these methods. The runtime is estimated by averaging the inference time-cost of the entire testing set using a single Nvidia GeForce Titan Xp GPU.

3) Competitors: We mainly compare to two latent space exploration methods pSp network [25] and PULSE [21] quantitatively and qualitatively.

Note that although the test set used in PULSE is the same as ours, the actual testing samples are different as it filters

TABLE I

COMPARISONS WITH STATE-OF-THE-ART METHODS ON FID AND NIQE METRICS. ↓ DENOTES THE LOWER THE BETTER AND ↑ DENOTES THE OPPOSITE. THE BEST RESULTS ARE MARKED IN **BOLD**

Methods \ Metrics	FID↓	NIQE↓	Runtime↓	FLOPs↓
pSp [25]	39.34	2.95	0.09	3.3×10^{11}
PULSE [21]	32.08	2.53	18.37	2.5×10^{13}
Ours	**25.78**	**2.49**	**0.06**	$\mathbf{2.1\times10^{11}}$

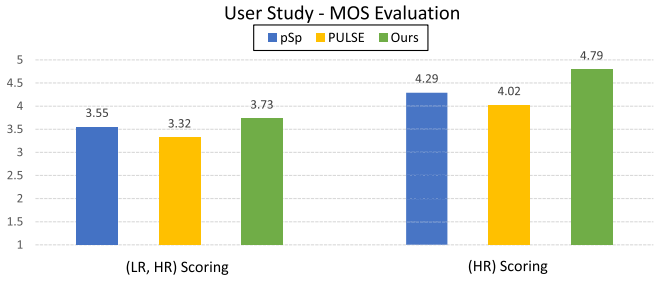


Fig. 5. User study results on mean-opinion-score. Two studies, a pairwise evaluation on (LR, HR) and a no-reference evaluation on HR only, are conducted.

out images with low self-supervised losses. We record the performances on the entire test set for all the methods.

B. Quantitative Comparison

In this section, we first show the quantitative comparison with PULSE [21] and pSp [25] on the recovered face with resolution of 1024×1024 upsampled from the 16×16 LR input.

The comparison results are shown in Table I. We can see that our Pro-PULSE outperforms PULSE and pSp on both FID and NIQE metrics. In particular, our Pro-PULSE is superior over competitors on the FID metric by a large margin, which shows that our model has a better hallucination ability and produces consistent face identities. We believe this improvement attributes to our progressive training strategy, which allows our model to inherit faithful characteristics from the input LR. As for the no-reference evaluation metric NIQE, our model outperforms the other two competitors as well, especially for the image-specific optimization-based method PULSE. This reveals that our recovered HR faces receive favorable visual quality by the human perceptual.

Besides the quality metrics, our model can process images with a faster speed and lower complexity than others. This is because our progressive encoders focus on the training stages, once they are properly trained, our encoder is lightweight and the generator is almost the same as the original StyleGAN. Compared with pSp network, which takes a feature pyramid as input, our solution directly injects deep guidance in either the training of encoder or generator, yielding better accuracy yet lower complexity. On other hand, PULSE has a complex optimization process. The proposed Pro-PULSE largely remedies this disadvantage while achieving better performances. Note that we test PULSE and pSp using their official source codes on the same computer as ours. Additionally, PULSE is an optimization method that has an additional backward pass during inference. Therefore, the FLOPs of the backward pass is estimated as 2 times of that of the forward pass, following the evaluation strategy of DeepSpeed [22].

C. User Study

Except using objective metrics, we also conduct user studies using mean-opinion-score (MOS) with 40 raters. The panel is composed of 20 undergraduates and 20 graduates, with an equal proportion of male and female. In addition, to help the raters to be more familiar with the test, we provide 20 ground truth examples of full marks for them in advance. We designed two questions for different purposes. The first one evaluates the consistency between the input LR faces and the recovered HR ones. Raters are asked to score each pair of (LR, HR) with a question "How well the HR image is restored from the LR image?" They rate between 1 (worst) and 5 (best). Another study cares about the subjective perceptual quality of HR faces. For each method, raters provide a score from 1 to 5, evaluating the visual quality solely on the HR images. For both studies, each method provides 30 output images, and we mixed all the results (90 in total) in random order for evaluation.

We show the comparison result in Fig. 5. In terms of pairwise evaluation, the proposed method outperforms both competitors by around 5% and 11% respectively, showing that our produced results are more consistent with the LR input. On the other hand, our method performs favorably against state-of-the-arts on no-reference subjective evaluation by a large margin. It indicates that the generated results are mostly of high-quality and without visually noticeable artifacts.

D. Qualitative Comparison

Fig. 6 gives the visual comparison of Pro-PULSE, pSp and PULSE on the resolution of 1024×1024. We can see that both pSp and PULSE failed to maintain the semantically consistent facial components like sunglasses in the first row of Fig. 6. In particular, PULSE focuses on finding a HR that can be downsampled to the input LR. The deviations of exploration space from the true natural image manifold lead to the poor result when there is ambiguous semantic in the input LR. Also, pSp cannot recover the fine details since it does not consider the locality. In contrast, our recovered faces could preserve small-scale semantic regions well from the blurry LR input. On the other hand, when the input LR images contain rare semantics, like hats are rarely appeared in the FFHQ dataset, all the methods cannot recover the missing semantics that not existed in the latent space. However, unlike the competitors affected by the irregular shapes of unknown LR semantics, our method recovers more visually plausible results. These interesting findings attribute to the reduced learning ambiguity of our progressive mechanism, such that rare and difficult semantics can be correctly inherited to HR results.

Due to the incremental nature of our progressive model, we can produce multi-resolution upsampling results. Fig. 7 shows how our produced results evolve. Lower resolution

(a) Input LR (b) GT HR (c) pSp [25] (d) PULSE [21] (e) Pro-PULSE

Fig. 6. Qualitative comparison of latent space exploration methods with a scaling factor of 64×.

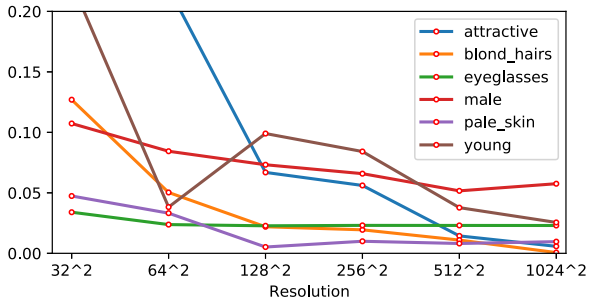
(a) Input LR (b) 32×32 (c) 64×64 (d) 128×128 (e) 256×256 (f) 512×512 (g) 1024×1024 (h) GT HR

Fig. 7. Multi-resolution images produced by our deeply-supervised generator.

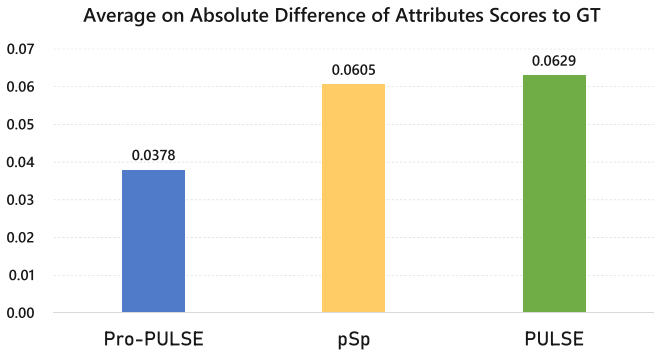
images (see Fig. 7b and Fig. 7c) mainly focus on the structure and shape of faces, while the higher resolution images fill up detailed textures (see Fig. 7f and Fig. 7g). This implies that our scale-specific encoder E_{W+} learns knowledge that consistent with the inherent multi-scale properties of StyleGAN.

E. Analysis on Semantic Preservation

In this section, we first analyze the semantics inheritance in the progressive learning process. We use the semantic score as the indicator, by taking the attribute classifier pre-trained on the CelebA-HQ dataset [17] to predict 40 attributes scores. We aim to examine whether the results from different resolutions are semantically-consistent with the ground truths. The absolute differences for all attribute scores between inputs and GTs are reported. We present 6 representative attributes in Fig. 8a, and we can see that our method can maintain correct semantics from the very beginning of our progressive process, except for 'attractive' and 'young' that are relied heavily on image resolution.



(a) Absolute difference of attributes scores to GT on different resolutions.



(b) Average absolute difference of attributes scores to GT with different methods.

Fig. 8. (a) Our method can inherit correct semantics from the very beginning of our progressive process. (b) Our method can obtain a closer semantic score (averaged across all 40 attributes) compared to state-of-the-art methods.

We also present the average absolute difference scores among 40 attributes with 3 up-sampling methods in Fig 8b, we can see our Pro-PULSE has a lower value on score differences, which shows it achieves better semantic-preservation than other two competitors.

Furthermore, we performs semantic editing based on the latent space interpretation approach, AdvStyle [33], which is an interpretable direction exploration method to find the latent semantics of GANs. In this experiment, we move the $w+$ codes from Pro-PULSE along the 'Gender' and 'Smile' directions found by AdvStyle, then feed the resulting codes into generator G_{ds} to get edited results. As shown in Fig. 9, semantics are edited successfully and images of multiple resolutions still maintain the same semantics after editing. This indicates that the obtained latent codes are editable, and the edited semantics can be propagated correctly to the HR images.

F. Comparison With Hallucination Method

In Fig. 10 we also show the comparison results with the state-of-the-art face hallucination method, DIC and DICGAN [19]. Due to the maximum $8\times$ unsampling factor of these compared methods, we also give the comparison results on the resolution of 128×128. Note that our 128×128 results are produced by our side-outputs. We also show our final results with a resolution of 1024×1024. Comparing with DIC and DICGAN, we can see those hallucination methods produce similar structures of the ground truth faces, but the faces recovered by our Pro-PULSE contains richer details than the hallucination methods, such as teeth or eyes. This is due to the different mechanisms between GAN inversion-based methods and hallucination methods. The former projects an input image to a low-dimensional code, this step discards the original spatial information which is difficult to be fully recovered by exploring the latent space. Despite the spatial discrepancy between GT and the output, it can produce vivid details. The latter focuses on expanding the original coarse input. Although hallucination methods can preserve the plausible structure of the ground truth identity, their results lack details and are limited to a small factor of upscaling.

(a) Reverse gender

(b) Add Smile

Fig. 9. Semantic editing by the latent space interpretation approach. In (a) and (b), the top-left image is 1024×1024 input from Pro-PULSE, and the following images are edited outputs from resolution 16×16 to 1024×1024.

G. Ablation Study

In this section, we conduct an ablation study on our Pro-PULSE and its variants. We analyze the efficiency of our progressive learning strategy, the effectiveness of using the base encoder E_W for guidance or deeply-supervised generator G_{ds}, and finally the loss functions. These result in 11 variants:

1) E_W^{direct}: Learning a direct encoder for $\mathcal{W}$ space.
2) E_{W+}^{direct}: learning a direct encoder for $\mathcal{W}+$ space. Both previous two variants are learned without progressive strategy.
3) $E2E$: Training encoder and generator in an end-to-end manner.
4) w/o E_W: Pro-PULSE without the guidance of the pre-trained E_W.
5) w/o G_{ds}: Pro-PULSE using the original StyleGAN generator rather than the deeply supervised generator G_{ds}.

(a) Input LR (b) BICUBIC (8×) (c) DIC (8×) (d) DICGAN (8×) (e) Pro-PULSE (8×) (f) Pro-PULSE (64×) (g) GT HR

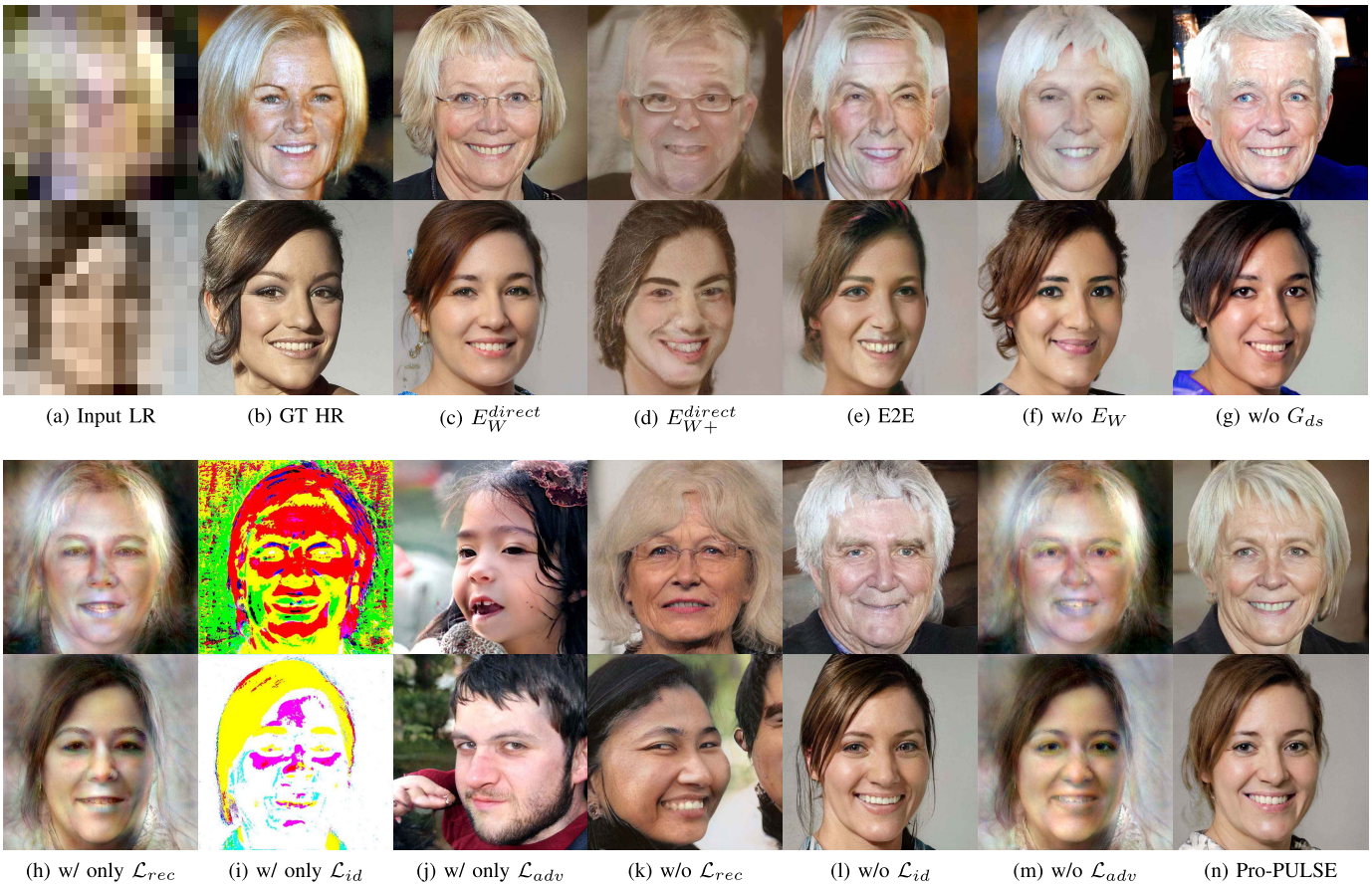
Fig. 10. Comparison of Pro-PULSE with bicubic upsampling, DIC, and DICGAN. Pro-PULSE (8×) denotes the side-output of resolution 128 × 128. Our recovered results contain richer details.

6) w/ only $\mathcal{L}_{rec}$: Pro-PULSE with only the reconstruction loss.
7) w/ only $\mathcal{L}_{id}$: Pro-PULSE with only the identity loss.
8) w/ only $\mathcal{L}_{adv}$: Pro-PULSE with only the adversarial loss.
9) w/o $\mathcal{L}_{rec}$: Pro-PULSE without the reconstruction loss.
10) w/o $\mathcal{L}_{id}$: Pro-PULSE without the identity loss.
11) w/o $\mathcal{L}_{adv}$: Pro-PULSE without the adversarial loss.

Note that variants from 6 to 11 are conducted when training the scale-specific encoder E_{W+}. We show both quantitative and qualitative results of above variants.

1) Quantitative Ablation Results: Quantitative results are shown in Table II. E_W^{direct} and E_{W+}^{direct} are two baselines that directly learn the encoders for mapping latent codes in different spaces. Embedding into the $\mathcal{W}$ space means that we have to find the original space the GAN was trained on, which applies the same latent vector to all the layers of the network. It is demonstrated [2] that different layers correspond to different levels of semantics, and therefore the enlagred $\mathcal{W}+$ space leads to diverse and accurate editing. However, we can see that E_{W+}^{direct} performs much worst than E_W^{direct}. It evidenced our observation that learning the $\omega+$ latent code with the dimension of the 18×512 is a challenging task. Instead, our tailored progressive training scheme could ease the learning ambiguity and achieve even better performance than solely learning on either spaces.

As for the end-to-end training scheme, the generator cannot propagate accurate multi-scale information back to the encoder since its side-outputs are always changing during training. These unstable branches bring wrong feedback that influence the training of the encoder. We also show that without the guidance from the pre-trained E_W, our Pro-PULSE decreases its performance both on FID and NIQE. This, on one hand,



(a) Input LR (b) GT HR (c) E_W^{direct} (d) E_{W+}^{direct} (e) E2E (f) w/o E_W (g) w/o G_{ds}

(h) w/ only $\mathcal{L}_{rec}$ (i) w/ only $\mathcal{L}_{id}$ (j) w/ only $\mathcal{L}_{adv}$ (k) w/o $\mathcal{L}_{rec}$ (l) w/o $\mathcal{L}_{id}$ (m) w/o $\mathcal{L}_{adv}$ (n) Pro-PULSE

Fig. 11. Qualitative ablation comparison on different variants of the proposed method.

TABLE II

QUANTITATIVE ABLATION COMPARISONS. ↓ DENOTES THE LOWER THE BETTER AND ↑ DENOTES THE OPPOSITE. THE BEST RESULTS ARE MARKED IN **BOLD**

Variants \ Metrics	FID↓	NIQE↓
E_W^{direct}	32.65	2.53
E_{W+}^{direct}	50.23	3.40
$E2E$	26.81	2.97
w/o E_W	26.75	2.69
w/o G_{ds}	28.05	2.75
w/ only $\mathcal{L}_{rec}$	86.69	3.15
w/ only $\mathcal{L}_{id}$	217.54	5.95
w/ only $\mathcal{L}_{adv}$	51.81	**2.44**
w/o $\mathcal{L}_{rec}$	40.25	2.81
w/o $\mathcal{L}_{id}$	29.52	2.52
w/o $\mathcal{L}_{adv}$	82.33	3.25
Pro-PULSE	**25.78**	2.49

implicitly reveals that the effectiveness of our progressive learning, as we enables the learning on $\mathcal{W}+$ space, and largely improves the performance over E_{W+}^{direct}. On the other hand, it also demonstrates that the base encoder E_W provides extra knowledge and stabilizes the training process. Besides, without the deep supervision from G_{ds}, the performance of our model also degrades, which shows that the intermediate guidance ensures correct information propagation in our progressive training strategy.

Regarding the loss functions, we find $\mathcal{L}_{adv}$ serves as an essential loss to our model, as our model decreases dramatically without adversarial loss. This is because LR-HR mapping is highly ambiguous, using pixel-wise constraints cannot guarantee plausible images, instead they tend to produce middle values to conform pixel-wise measurement. On the other hand, using only the $\mathcal{L}_{adv}$ loss can produce realistic results, but they are not related to the input LR images (see Fig. 11j). Besides, $\mathcal{L}_{rec}$ and $\mathcal{L}_{id}$ influence the FID a lot. As FID is a reference-related metric, the $\mathcal{L}_{rec}$ and $\mathcal{L}_{id}$ could ensure semantic consistency between the generated faces and the identity characteristics. However, using only the $\mathcal{L}_{rec}$ or $\mathcal{L}_{id}$ is insufficient, especially using $\mathcal{L}_{id}$ only may lead to the failure of training due to the deficiency of pixel-level appearance information.

2) Qualitative Ablation Results: We further show the visual results of various variants in Fig. 11. We can see that learning an encoder to the $\mathcal{W}+$ space directly always produce abnormal results (See Fig. 11d) due to the complexity of $\mathcal{W}+$ space. However, using our progressive learning can largely ease this difficulty (Fig. 11f). On the other hand, learning an encoder to the $\mathcal{W}$ space directly will produce plausible results except for small scale objects (see the glasses in the top row of Fig. 11c), which implies that learning in $\mathcal{W}$ space is not sufficiently disentangled from various semantics [1].

Comparing Fig. 11e and Fig. 11n, we can know that the stable side-output branches of generator are helpful for the

(a) Input LR (b) GT HR (c) $\lambda_{rec} = 1$, $\lambda_{id} = 1$, $\lambda_{adv} = 0.01$ (d) $\lambda_{rec} = 5$, $\lambda_{id} = 1$, $\lambda_{adv} = 0.01$ (e) $\lambda_{rec} = 1$, $\lambda_{id} = 5$, $\lambda_{adv} = 0.01$ (f) $\lambda_{rec} = 1$, $\lambda_{id} = 1$, $\lambda_{adv} = 0.05$

Fig. 12. Qualitative ablation comparison on loss weights. (d), (e), (f) separately increase the weight of $\mathcal{L}_{rec}$, $\mathcal{L}_{id}$, and $\mathcal{L}_{adv}$, compared to our final weight setting (c).

encoder to capture multi-scale features during progressive learning. The inaccurate side-outputs lead to poor results of end-to-end training scheme. We also find that without the guidance of G_{ds} or E_W, the recovered HR faces contain artifacts that lack of detailed texture. Fig. 11h demonstrates that training with only the pixel-wise $\mathcal{L}_{rec}$ loss brings blurry artifacts. The poor results shown in Fig. 11i indicate that only the supervision of $\mathcal{L}_{id}$ is insufficiently informative for training, as $\mathcal{L}_{id}$ is adopted on the feature level, but cannot confine the generated RGB outputs. The adversarial loss shows its importance in Fig. 11m, we cannot recover the realistic faces when deactivating $\mathcal{L}_{adv}$, and using pixel-wise constraints only typically produce blurry results for a highly ambiguous problem. When removing the reconstruction loss, the recovered images are inconsistent with the GT faces, and the identity loss is essential for maintaining the facial details of faces.

3) Loss Weights Analysis: Here we examine the effectiveness of loss weights. As shown in the Fig. 12, we increase the λ_{rec}, λ_{id}, and λ_{adv} separately, compared to our final setting Fig. 12c. When enhancing the reconstruction loss, it weakens the semantic-consistency producing incorrect semantics like eyeglasses. Also, by comparing Fig. 12c and Fig. 12e, we can see that the model tries to match the identity but introduces implausible details. As for the λ_{adv}, Fig. 12f demonstrates that $\mathcal{L}_{adv}$ encourages more realistic faces, but sacrifices the consistency between LR and HR. Overall, our final setting can have a optimal balance between reconstruction, semantic and identity preservations.

H. Real Image Upsampling

In addition, to verify the generalization of Pro-PULSE, we provide results of real images which are unseen in the training. From Fig. 13 we see that Pro-PULSE is capable to synthesize meaningful semantics for real images, such as beard and glasses. It demonstrates that Pro-PULSE is generalized enough to handle images in reality.

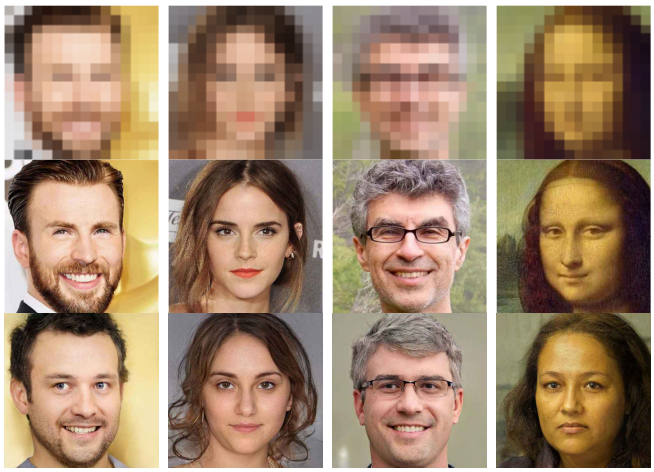


Fig. 13. Upsampling real images using our method. The rows from top to bottom are Input LR, GT HR and Pro-PULSE, respectively. It shows that Pro-PULSE is generalized enough to handle images in reality.

I. Generalization on Non-Human Domains

We also carry out an experiment on the non-human datasets to validate that our method is general to other domains, including Anime [3], Cat [36], and Church [34]. We use the StyleGAN generators pre-trained on these datasets with the dataset resolutions of 512×512, 256×256, and 256×256 respectively. Note that we deactivate the $\mathcal{L}_{id}$ in

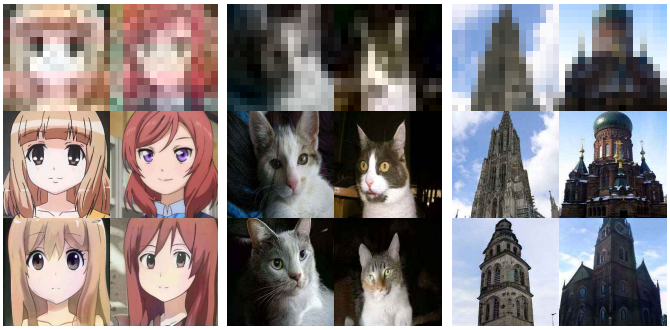


Fig. 14. Upsampling results for non-human domains. From top to bottom rows are input LR, GT HR, and our results, respectively.

the training of Pro-PULSE on these non-human dataset, and the other settings remain the same. The results on these dataset are shown in Fig. 14. Although without identity information, we can see that Pro-PULSE is still qualified for producing plausible HR on various non-human domains.

V. Conclusion and Limitations

In this paper, we propose a novel progressive upsampling framework, *i.e.*, Pro-PULSE, that mapping an extreme LR input to a latent vector of StyleGAN for producing faithful and semantically correct HR image. Our progressive training strategy allows the encoder inherit the semantic information from the LR input and ease the learning ambiguity. Both quantitative and qualitative experiment results demonstrate that our Pro-PULSE could recover the SR faces both semantical accurately and computational efficiently. Although, our Pro-PULSE also work well on non-human data, it requires a specifically pre-trained StyleGAN generator of the corresponding domain. This is a common limitation for the latent space exploration works, as we cannot synthesize out-of-distribution results. We may explore a one-shot or few-shot cross-domain solution to resolve this limitation in the future.

References

[1] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the StyleGAN latent space?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4432–4441.

[2] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN++: How to edit the embedded images?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8296–8305.

[3] Anonymous, D. community, and G. Branwen. (2021). *Danbooru2020: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset*. [Online]. Available: https://www.gwern.net/Danbooru2020

[4] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun./Jul. 2004, pp. I–I.

[5] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1511–1520.

[6] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2492–2501.

[7] A. Creswell and A. A. Bharath, "Inverting the generator of a generative adversarial network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 1967–1974, Jul. 2019.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[9] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. ECCV*, 2014, pp. 184–199.

[10] L. Goetschalckx, A. Andonian, A. Oliva, and P. Isola, "GANalyze: Toward visual definitions of cognitive image properties," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5744–5753.

[11] 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," in *Proc. NeurIPS*, 2017, pp. 6626–6637.

[12] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.

[13] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.

[14] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. ICLR*, 2018.

[15] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[17] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5549–5558.

[18] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3867–3876.

[19] C. Ma, Z. Jiang, Y. Rao, J. Lu, and J. Zhou, "Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5569–5578.

[20] F. Ma, U. Ayaz, and S. Karaman, "Invertibility of convolutional generative networks from partial measurements," in *Proc. NeurIPS*, 2018, pp. 9628–9637.

[21] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: Self-supervised photo upsampling via latent space exploration of generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2437–2445.

[22] Microsoft. (2021). *Deepspeed*. [Online]. Available: https://www.deepspeed.ai

[23] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[24] A. Plumerault, H. L. Borgne, and C. Hudelot, "Controlling generative models with continuous factors of variations," in *Proc. ICLR*, 2020.

[25] E. Richardson *et al.*, "Encoding in style: A styleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2287–2296.

[26] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9243–9252.

[27] Y. Song *et al.*, "Joint face hallucination and deblurring via structure generation and detail enhancement," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 785–800, 2018.

[28] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[30] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers, "A fully progressive approach to single-image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 864–873.

[31] C. Yang, Y. Shen, and B. Zhou, "Semantic hierarchy emerges in deep generative representations for scene synthesis," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1451–1466, May 2021.

[32] C.-Y. Yang and M.-H. Yang, "Fast direct super-resolution by simple functions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 561–568.

[33] H. Yang, L. Chai, Q. Wen, S. Zhao, Z. Sun, and S. He, "Discovering interpretable latent space directions of GANs beyond binary attributes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12177–12185.

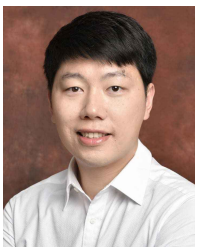
[34] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop," 2015, *arXiv:1506.03365*.
[35] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Proc. ECCV*, 2018, pp. 217–233.
[36] W. Zhang, J. Sun, and X. Tang, "Cat head detection–how to effectively exploit shape and texture features," in *Proc. ECCV*, 2008, pp. 802–816.
[37] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. ECCV*, 2018, pp. 286–301.
[38] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-domain GAN inversion for real image editing," in *Proc. ECCV*, 2020, pp. 592–608.
[39] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded Bi-network for face hallucination," in *Proc. ECCV*, 2016, pp. 614–630.

Yang Zhou received the B.Sc. degree from the School of Computer Science and Engineering, South China University of Technology, in 2020, where he is currently pursuing the master's degree. His research interests include computer vision, image processing, and deep learning.

Yangyang Xu received the Ph.D. degree from the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His current research interests include computer vision, image processing, computer graphics, and deep learning.

Yong Du received the B.Sc. and M.Sc. degrees from Jiangnan University and the Ph.D. degree from the South China University of Technology. He is currently an Assistant Professor with the Department of Computer Science and Technology, Ocean University of China. His research interests include computer vision and image processing.

Qiang Wen received the B.Eng. degree from the School of Information Science and Engineering, Central South University, in 2018. He is currently pursuing the master's degree with the School of Computer Science and Engineering, South China University of Technology. His research interests include computer vision, image processing, and deep learning.



Shengfeng He (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Macau University of Science and Technology in 2009 and 2011, respectively, and the Ph.D. degree from the City University of Hong Kong in 2015. He is currently an Associate Professor with the School of Computer Science and Engineering, South China University of Technology. His research interests include computer vision, image processing, and computer graphics. He serves on the editorial board of *Neurocomputing*.