frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 24 November 2022
DOI 10.3389/fcomp.2022.1037435

OPEN ACCESS

EDITED BY
Chaohui Wang,
Gustave Eiffel University, France

REVIEWED BY
Changbo Wang,
East China Normal University, China
Haibo Qiu,
The University of Sydney, Australia

*CORRESPONDENCE
Ahmed Cheikh Sidiya
ac0004@mix.wvu.edu

SPECIALTY SECTION
This article was submitted to Computer Vision, a section of the journal Frontiers in Computer Science

RECEIVED 05 September 2022
ACCEPTED 02 November 2022
PUBLISHED 24 November 2022

CITATION
Cheikh Sidiya A and Li X (2022) Toward extreme face super-resolution in the wild: A self-supervised learning approach.
Front. Comput. Sci. 4:1037435.
doi: 10.3389/fcomp.2022.1037435

Toward extreme face super-resolution in the wild: A self-supervised learning approach

Ahmed Cheikh Sidiya* and Xin Li

Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV, United States

Extreme face super-resolution (FSR), that is, improving the resolution of face images by an extreme scaling factor (often greater than ×8) has remained underexplored in the literature of low-level vision. Extreme FSR in the wild must address the challenges of both unpaired training data and unknown degradation factors. Inspired by the latest advances in image super-resolution (SR) and self-supervised learning (SSL), we propose a novel two-step approach to FSR by introducing a mid-resolution (MR) image as the stepping stone. In the first step, we leverage ideas from SSL-based SR reconstruction of medical images (e.g., MRI and ultrasound) to modeling the realistic degradation process of face images in the real world; in the second step, we extract the latent codes from MR images and interpolate them in a self-supervised manner to facilitate artifact-suppressed image reconstruction. Our two-step extreme FSR can be interpreted as the combination of existing self-supervised CycleGAN (step 1) and StyleGAN (step 2) that overcomes the barrier of critical resolution in face recognition. Extensive experimental results have shown that our two-step approach can significantly outperform existing state-of-the-art FSR techniques, including FSRGAN, Bulat's method, and PULSE, especially for large scaling factors such as 64.

KEYWORDS

extreme face super-resolution, self-supervised learning, degradation learning, latent space interpolation, face in the wild

1. Introduction

Face recognition at long range (e.g., from hundreds to thousands of meters) or high altitude (e.g., from aerial platforms such as UAVs) has received increasingly more attention in recent years. At this distance, the resolution of facial regions in the acquired images can be as low as 16 pixels. In the literature, the smallest spatial resolution for a human operator to discern facial identity information has been reported to be around 18 × 24 pixels (Bachmann, 1991). Despite rapid advances in face recognition and super-resolution (SR), reliable extraction of face identity information from extreme low-resolution (LR) face images such as Widerface (Yang et al., 2016) has remained beyond the capability of current technologies. As shown in Figure 1, there exists a critical

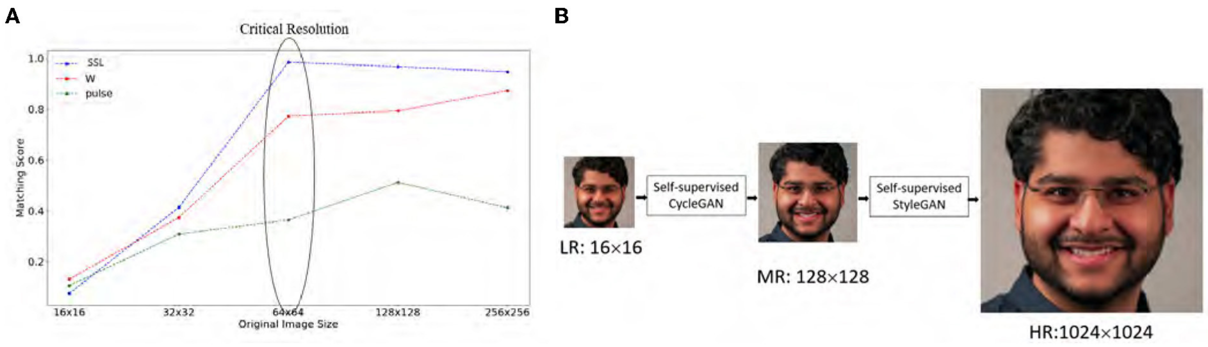

FIGURE 1
Motivation to study extreme FSR. **(A)** Preserving identity information using the Openface face matching tool (Amos et al., 2016) and StyleGAN FSR is infeasible below the critical resolution of 64 × 64. Note that our two-step SSL-based approach (marked blue) significantly outperforms PULSE (Menon et al., 2020) (marked green). **(B)** This work presents a two-step SSL-based approach to extreme FSR–MR serves as the stepping stone to facilitate both degradation learning and image reconstruction.

resolution of 64 × 64 below which even the latest StyleGAN-based face SR (FSR) (Menon et al., 2020) cannot preserve important facial identity information.

Another fundamental limitation of existing FSR methods is that they mostly assume a synthetic degradation model; that is, the LR image is a down-sampled version of the high-resolution (HR) image. This assumption is not valid for LR images in the wild due to the notorious difficulty in modeling complex acquisition conditions related to varying illumination, poses, and camera distances (the so-called "simulated-to-real gap", Köhler et al., 2019). It takes a lot of effort to capture paired LR/HR images by adjusting the focal length of a camera (still requiring image registration) (Cai et al., 2019). More importantly, the generalization property of such a supervised approach remains questionable. So far, the issue of modeling real-world degradation has been tackled using a GAN-based approach in Bulat et al. (2018), but with limited success; An unsupervised learning approach to real-world SR has recently been studied in Lugmayr et al. (2019) and Wei et al. (2020), but it is not specifically tailored to faces.

Self-supervised learning (SSL) (Liu et al., 2021b) allows two birds to be killed (large magnification ratio and unknown degradation process) with one stone. Two lines of research findings have been reported at the intersection of SSL and SR. On the one hand, several recent works, such as Synthetic Multi-Orientation Resolution Enhancement (SMORE) (Zhao et al., 2020) and Self-supervised CycleGAN (Liu et al., 2021a), have shown promising performance in modeling real-world degradation in MRI and ultrasound images. They share the principle of generating the required LR-HR pairs in a self-supervised manner. On the other hand, SSL has demonstrated initial success to extreme FSR—e.g., self-supervised photo-upsampling *via* latent space extrapolation (PULSE) (Menon et al., 2020) has shown promising SR reconstruction results for the magnification ratio as large as 64. However, the LR images used in PULSE are assumed to be synthetic, making its generalization property to realistic LR images unknown.

In this paper, we propose a novel two-step FSR method that combines the two prior lines of research (self-supervised CycleGAN+StyleGAN). A new insight brought about by our approach is to introduce a mid-resolution (MR) image as the stepping stone connecting LR with HR (see Figure 2). Our two-step FSR strategy simultaneously tackles challenges with unknown degradation process and large magnification ratio using the SSL strategy. In the first step, SSL allows us to bridge the simulated-to-real gap by developing an embedded face hallucination method. Similarly to self-supervised CycleGAN (Liu et al., 2021b), ours enforces perceptual consistency by self-supervised CycleGAN; unlike self-supervised CycleGAN (Liu et al., 2021b), our network design is specially tailored for face images (e.g., style-based generator). In the second step, SSL allows us to achieve a large upsampling ratio through latent space exploration, similar to StyleGAN-based PULSE (Menon et al., 2020). Unlike PULSE, the MR image in our two-step approach has a resolution of 128 × 128, which guarantees the preservation of information on the identity of the face in the latent space. Our technical contributions are summarized below.

(1) Two-step SSL-based FSR method. Previous work such as PULSE (Menon et al., 2020) treats ×8 and ×64 as two independent problems and trains separate networks. Our approach introduces the MR image as a stepping stone connecting LR with HR (the ×8 solution is embedded in the ×64 one); the proposed two-step FSR method is based on a hybrid of self-supervised CycleGAN (Liu et al., 2021b) and StyleGAN (Karras et al., 2019), which unifies SSL-based degradation learning and image upsampling.

(2) Self-supervised CycleGAN for degradation learning of face images in the wild. Unlike Bulat's open-loop approach

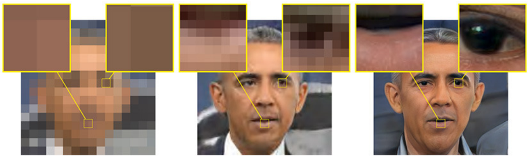

FIGURE 2
Left: Simulated LR input (16 × 16); **middle**: MR image (step-1 output, 128 × 128); **right**: HR final (step-2 output, 1024 × 1024).

(Bulat et al., 2018), we aim to exploit the dual-cycle losses of forward and inverse processes (i.e., close-loop optimization), as well as the adversarial characteristics of self-supervised discriminator to promote a style-based generator for improved perceptual consistency.

(3) Self-supervised StyleGAN for Semantic Interpolation in Latent Space. Inspired by InterFaceGAN (Shen et al., 2020a), we propose to cast FSR as a semantic interpolation problem in latent space, which allows us to identify a principal direction in HR face images and generate SR images by interpolating latent codes. Our approach can be interpreted as imposing a Laplacian prior on the latent code to regularize the inversion process. Our FSR results dramatically outperform those of PULSE for the upsampling ratio of 64 (see Figure 7).

2. Related works

2.1. Face super-resolution (FSR)

FSR (Bulat et al., 2018; Chen et al., 2018; Menon et al., 2020), also known as face hallucination (FH) (Liu et al., 2007; Jia and Gong, 2008; Wang et al., 2014), is one of the most studied low-level vision problems due to the wide application of face recognition systems in the real world. Model-based approaches toward FSR/FH can be classified into Bayesian inference-based, subspace learning-based, and sparse representation-based approaches. Rapid advances in generative adversarial networks (GANs) (Goodfellow et al., 2014) have made a splash in the field of FSR in recent years. For example, designing a facial geometry prior in terms of facial landmark heatmap and parsing map has led to the construction of FSRNet and FSRGAN (Chen et al., 2018) with end-to-end optimization. Most recently, an attention map of facial components was implicitly imposed to improve the performance of SR for face images in Kalarot et al. (2020); StyleGAN's style-based encoder (Karras et al., 2020) was leveraged in the problem of FSR that leads to photo upsampling through Latent Space Extrapolation (PULSE) (Menon et al., 2020). Unfortunately, the performance of PULSE on real-world low-resolution face images remains unsatisfactory, which partially inspired this research.

2.2. Self-supervised learning

SSL has recently been studied for the SR of biomedical and satellite images. SMORE (Zhao et al., 2020) is a self-supervised anti-aliasing and SR Algorithm designed specifically for MRI image reconstruction. A similar idea was developed in ultrasound Image SR with perception consistency through self-supervised CycleGAN (Liu et al., 2021b). Self-supervised multi-image SR, namely DSA-Self (Nguyen et al., 2021), was proposed for push-frame satellite images. Most recently, SSL for real-world SR from dual-zoomed observations was proposed in Zhang et al. (2022). The first step of our two-stage FSR is closely related to these existing works; our focus is to develop an SSL-based approach to degradation learning.

2.3. Semantic face image manipulation

Rapid advances in GAN-based face image synthesis (e.g., Karras et al., 2017, 2019, 2020) have inspired a flurry of work on semantic face image manipulation. Image2StyleGAN (Abdal et al., 2019) presented an efficient algorithm for embedding a given image in the latent space of StyleGAN, allowing semantic image editing operations. The extended latent space (often known as "W+") works better than the original latent space (often known as "W"). Following this line of research, InterFaceGAN (Shen et al., 2020a), StyleRig (Tewari et al., 2020), and in-domain GAN inversion (Zhu et al., 2020) have been developed for various semantic face image manipulation tasks such as morphing attack, pose normalization, aging simulation, expression transfer, and illumination compensation. The second step of our two-stage FSR is built upon semantic manipulation in the latent space and our method differs from the existing approaches in terms of regularization strategy.

3. Problem formulation and system overview

Given a set of LR face images $X^{LR} \in R^{n\times n}$ and a set of HR face images $X^{HR} \in R^{N\times N}$, FSR deals with the problem of learning a non-linear mapping $f : X^{LR} \rightarrow X^{HR}$. Depending on the modeling of the relationship between LR and HR, there are two ways of formulating the FSR problem: (1) *Simulated*—LR images are artificially downsampled versions of HR images. Such paired LR/HR training data correspond to supervised learning approaches [e.g., FSRNet/FSRGAN (Chen et al., 2018; Kalarot et al., 2020; Menon et al., 2020)]; (2) *Realistic*—LR images are acquired in the wild/real world. Such unpaired HR/LR training data give rise to unsupervised learning approaches (e.g., Cai et al., 2019; Lugmayr et al., 2019; Wei et al., 2020). The question of how to model realistic degradation with large scaling factors has remained an open problem in low-level vision.

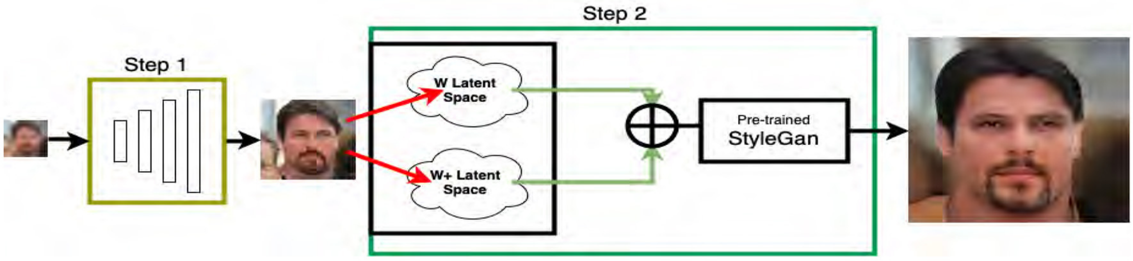

FIGURE 3
System overview: self-supervised CycleGAN (step-1, **left**) and self-supervised StyleGAN (step-2, **right**). Note that the MR image sized by 128 × 128 serves as a bridge connecting the 16 × 16 LR and 1024 × 1024 HR images. The detail of Step 1 network architecture is shown in Figure 4. The trained Low to High generator is used to get the MR image. Red arrows indicate the projection to the generative model latent space. Green arrows represent the interpolation in the latent space as defined in Equation (6). The interpolated latent vectors are combined using Equation (10).

Inspired by rapid advances in SR, we propose to tackle both the problems of unknown degradation learning and the large scaling factor using a two-stage SSL-based approach. In the first stage, $f_1 : X^{LR} \rightarrow X^{MR}$, we propose to tackle the problem of modeling the unknown degradation process in LR images by an *SSL* approach (refer to Figure 4). That is, to overcome the challenge of acquiring paired LR/HR faces, it is plausible to embed the FSR solution to simulated LR into the more general solution to realistic LR. In other words, any of existing works based on simulated LR faces can be used to pre-train the generator network by paired LR/HR data; while such pretrained generator network can be fine-tuned by unsupervised learning using unpaired LR/HR data. Since both supervised and unsupervised settings share the same generator network, an SSL-based solution enjoys a good generalization property.

In the second stage, $f_2 : X^{MR} \rightarrow X^{HR}$, we propose to cast FSR as an SSL-based semantic editing problem for face images. We advocate an *interpolation* instead of an exploration approach (e.g., in self-supervised PULSE, Menon et al., 2020) based on a recent study of semantics in latent space (Shen et al., 2020a). In previous work (Menon et al., 2020), the latent vector in the W^+ space was searched by exploring the local neighborhood that satisfies the data / likelihood constraint specified by the SLR image. Inspired by the latest work (Shen et al., 2020a), we propose to project face images of varying resolutions (e.g., MR vs. HR) onto their corresponding subspaces. By analogy with conditional manipulation of particular attributes (e.g., pose or age), we can generate the latent code of SR images by interpolating along the principal normal direction toward the HR-subspace. Improved image inversion with better artifact suppression can be achieved by interpolating between the estimated latent codes in the Gaussianized latent space (Wulff and Torralba, 2020).

In summary, an overview of the proposed two-step SSL-based FSR system is shown in Figure 3. It can be interpreted as combining the strengths of both self-supervised CycleGAN (Liu et al., 2021b) and StyleGAN (Karras et al., 2020). The introduction of an MR image sized by 128 × 128 kills two birds with one stone. On the one hand, it helps overcome the critical resolution barrier, as shown in Figure 1, which was overlooked by PULSE (Menon et al., 2020). On the other hand, it aims to reduce the simulated-to-real gap by a self-supervised CycleGAN (Liu et al., 2021b).

4. FSR *via* self-supervised CycleGAN

Given an image of extreme LR (e.g., 16 × 16), we propose to introduce the MR image as an intermediate result to facilitate degradation learning and image reconstruction tasks. Such an MR image also supports the extraction of latent code in the second step by overcoming the barrier of critical resolution.

4.1. Overview of network design

In the first step of our FSR approach, we focus on the design of self-supervised CycleGAN (refer to Figure 4), consisting of reconstruction networks (low-to-high generator and discriminator) and degradation networks (high-to-low generator and discriminator). Despite the structural similarity to self-supervised CycleGAN (Liu et al., 2021b), we note that our design contains the following important differences. (1) We have taken into account the simulated to real gap (Köhler et al., 2019). For simulated LR, our high-to-low generator becomes downsampling, and the high-to-low discriminator is bypassed, which degenerates into a vanilla GAN for supervised learning similar to SRGAN (Ledig et al., 2017). In other words, our FSR solution to simulated LR is *embedded* into that of real LR and serves as a pre-training network following the SSL principle. (2) Our design of high-to-low and low-to-high networks uses the latest advances, including a style-based

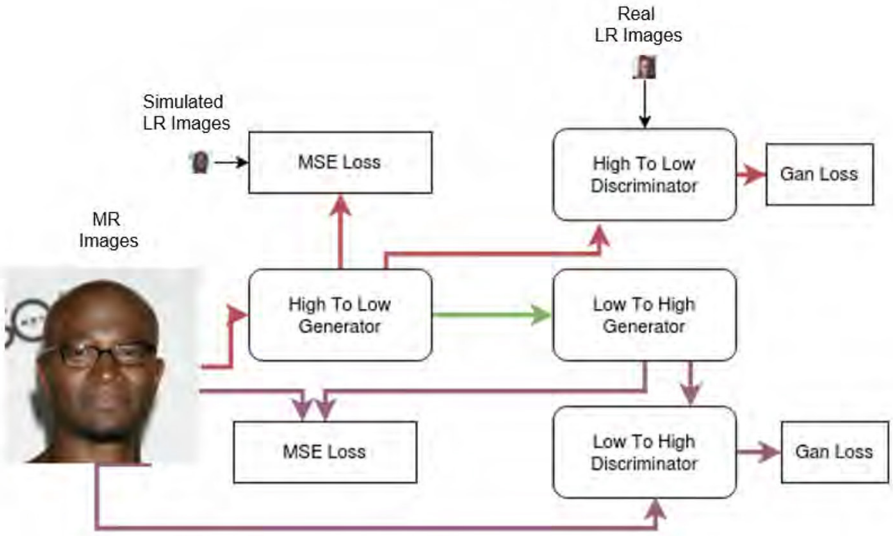

FIGURE 4
Overview of the CycleGan (Liu et al., 2021b) based architecture of step 1 of our method. Details of all components are presented in Section 4.

generator (Karras et al., 2019) and a self-attention mechanism (Zhang et al., 2019) for FSR. These designs allow us to handle various uncertainty factors (e.g., pose and expression variations) in face images without explicitly estimating facial landmarks such as FSRNet (Chen et al., 2018). Note that erroneous landmark estimation results often have a catastrophic impact on FSR reconstruction. (3) Our training targets joint end-to-end optimization of both High-to-Low and Low-to-High networks, which shows faster convergence and improved stability over separated training in Bulat's approach (Bulat et al., 2018). Both the selection of loss functions and the exclusion of normalization strategies [e.g., AdaIN (Huang and Belongie, 2017) and spectral normalization (Miyato et al., 2018)] jointly contribute to better preservation of facial identity and suppression of artifacts.

4.1.1. Low-to-high generator

Low-to-high generator consists of four sections of six, three, two, and one successive residual attention block separated by a bilinear upsampling of two. Similarly to the GAN progressive growth strategy (Karras et al., 2017), we start with the input of the patch 16 × 16 and gradually increase the spatial resolution by a factor of two after each section. In this way, the LR patch is upsampled by a factor of eight after three sections reach the final dimension of 128 × 128.

4.1.2. Low-to-high discriminator

Low-to-high discriminator consists of four convolution layers followed by a leaky relu layer and a last convolution layer. We have also added two layers of self-attention (Zhang et al., 2019) at the end of the network, which progressively increases the complexity of the SR task (by matching the design of the style-based generator).

4.1.3. High-to-low generator

High-to-low generator has an encoder-decoder type architecture (Badrinarayanan et al., 2017): the encoder consists of five residual + attention blocks (Zhang et al., 2018b) each followed by an average pooling layer and the decoder consists of four residual + attention blocks, where the first two are followed by a bilinear upsampling layer (Karras et al., 2017). Therefore, the input is downsampled by a factor of 32 and upsampled by a factor of four, which produces a downsampled image by a factor of eight, but with more flexibility in modeling real-world degradation (e.g., unknown blur, Gu et al., 2019). We also concatenate a noise vector into the input image of the network using a fully connected layer, which contributes to the robustness of our degradation modeling.

4.1.4. High-to-low discriminator

High-to-low discriminator consists of three convolution layers followed by a leaky relu layer and a last convolution layer. Similarly, we have added two layers of self-attention at the end of the network. Compared to the high-to-low discriminator in Bulat's method (Bulat et al., 2018), ours shares a similar strategy of skipping batch normalization, but differs in the introduction of self-attention layers and LeakyRelu units. In other words, Bulat's is closer to Resnet (Ledig et al., 2017), while ours is more similar to non-local neural networks (Wang et al., 2018). Note that our High-to-Low and Low-to-High Discriminators share similar architectures: In the presence of simulated LR, we simply skip this module.

4.2. Loss functions and training procedures

4.2.1. High-to-low loss functions

The High-to-Low generator loss is the weighted sum of the content loss and the GAN loss, as shown in Equation (1) where $\alpha = 1$ and $\beta = 0.001$.

$$L_G = \alpha L_{pixel} + \beta L_{GAN}^{G} \tag{1}$$

The relativistic GAN loss and the pixel loss follow the formula in Equations (2) and (3) (Jolicoeur-Martineau, 2018).

$$\begin{aligned} L_r(u,v) = \frac{1}{2}\Big\{ & \mathop{\mathbb{E}}_{u\sim P_u}\Big[\max\big(0, 1-\big(D(u)-\mathop{\mathbb{E}}_{v\sim P_v}\big[D(v)\big]\big)\big)\Big]+ \\ & \mathop{\mathbb{E}}_{v\sim P_v}\Big[\max\big(0, 1+\big(D(v)-\mathop{\mathbb{E}}_{u\sim P_u}\big[D(u)\big]\big)\big)\Big]\Big\} \end{aligned} \tag{2}$$

$$L_2(u,v) = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}(u_{i,j}-v_{i,j})^2 \tag{3}$$

In a supervised setting, we have $L_{GAN}^{G} = L_r(I_{FHR}, I_{HR})$, $L_{GAN}^{D} = L_r(I_{HR}, I_{FHR})$ and $L_{pixel} = L_2(I_{HR}, I_{FHR})$ where FHR refers to fake HR images; In an unsupervised setting, we have $L_{GAN}^{G} = L_r(I_{RLR}, I_{FLR})$, $L_{GAN}^{D} = L_r(I_{FLR}, I_{RLR})$, and $L_{pixel} = L_2(I_{SLR}, I_{FLR})$ where $I_{SLR}, I_{RLR}, I_{FLR}$ denote the simulated/real/fake LR images, respectively.

4.2.2. Low-to-high loss functions

Similarly, the Low-to-High generator loss is the weighted sum of content loss and GAN loss, as shown in Equation (1) where $\alpha = 1$ and $\beta = 0.001$. The GAN losses and the pixel loss follow the same formula as in Equations (2) and (3): $L_{GAN}^{G} = L_r(I_{HR}, I_{FHR})$, $L_{GAN}^{D} = L_r(I_{FHR}, I_{HR})$ and $L_{pixel} = L_2(I_{HR}, I_{FHR})$ where I_{FHR} is the fake HR image generated by the Low-to-High generator and I_{HR} is the real-world HR image. It should be noted that a major difference between this work and Bulat et al. (2018) lies in the selection of regularization parameters. In Bulat et al. (2018), the specification of (α, β) satisfies the constraint $\alpha L_{pixel} < \beta L_{GAN}$ in general. We argue that this is not desirable from the perspective of preserving facial identity. We advocate the setting of parameters of $\alpha = 1$ and $\beta = 0.001$, leading to comparable loss terms (i.e., $\alpha L_{pixel} \approx \beta L_{GAN}$). According to our own experience, such a balanced choice of regularization terms is beneficial for end-to-end optimization.

4.2.3. Training strategy

It is worth mentioning that we have *NOT* augmented the data during training using standard techniques such as image flipping, scaling, and rotation. Our experience suggests that, for unsupervised learning, data augmentation does not help improve the accuracy of face SR reconstruction, but increases the computational burden and the risk of introducing artifacts (due to unpaired LR-HR training data). We have used a batch of size 32 and the total training requires about 20 epochs or $\sim$ 143,000 generator and discriminator updates (by contrast, Bulat et al., 2018 requires 570,000 updates). The learning rate is maintained at 0.001 throughout the training process (in contrast to the decoupled training in Bulat et al., 2018). We also use Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and adopt a PyTorch-based implementation (Paszke et al., 2019).

5. FSR *via* self-supervised latent space interpolation

In the second step of our FSR approach, we extract the latent code from the intermediate MR result and reconstruct the HR image from the latent code using StyleGAN. A similar idea exists in PULSE (Menon et al., 2020); however, PULSE assumes a synthetic LR image and lacks generalization to realistic LR images.

5.1. FSR *via* semantic face manipulation

The success of StyleGAN-based face image synthesis (Karras et al., 2020) is largely due to the construction of a latent space (e.g., W or $W+$) in which semantic information is exploited for the reconstruction of high-quality face images. A pre-trained GAN model such as Karras et al. (2020) can be formulated as a deterministic function $g: W \rightarrow X$ where W denotes the low-dimensional latent space and X is the image space (here $X = X^{HR}$). Instead of constructing a semantic scoring function as in Shen et al. (2020a), we propose to consider an embedding function $e: X^{MR} \rightarrow W$ as in Image2StyleGAN (Abdal et al., 2019). In this way, we can bridge the MR image space and the HR image space with $f_2 = g(e(X^{MR}))$. Due to the two properties (hyperplane separation and large deviation) of semantics in the latent space (Shen et al., 2020a), we can formulate the construction of f_2 as a generalized problem of semantic face editing.

The key new insight brought by this work is that the spaces of X^{LR}, X^{MR}, X^{HR}—when projected onto the latent space W—are separated just like the semantic attributes of face images (e.g., pose, age, gender). It follows that there exists a hyperplane in the latent space such that all samples from the same side have shared characteristics (e.g., visual quality or semantic attributes). Instead of the subspace for attribute manipulation

(Viazovetskyi et al., 2020), we target the one associated with X^{HR} and define the distance from a sample latent code $\vec{z} \in W$ to the unit normal vector $\vec{n}$ (called "principal normal", Shen et al., 2020a) of this HR-face subspace by the following:

$$d(\vec{n}, \vec{z}) = \vec{n}^T \vec{z} \tag{4}$$

Note that $d(\cdot, \cdot)$ can be negative because the latent code can be on either side of the hyperplane. Conceptually similar to the semantic score used in Shen et al. (2020a), we can define the visual quality of a latent code z as linearly dependent on its projected distance from the target hyperplane (the closer, the better).

$$e(g(\vec{z})) = \lambda d(\vec{n}, \vec{z}), \tag{5}$$

where $\lambda > 0$ is a scalar that controls how fast the quality varies along with the change in projected distance. In summary, we assume that the target X^{HR} can be approximately modeled by a linear subspace defined by the principal normal vector $\vec{n}$ in the latent space.

Based on the above assumption, we can cast FSR as a manipulation problem in latent space. Similarly to facial attribute editing, we can manipulate the latent code as follows.

$$\vec{z}_{edit} = \vec{z} + \gamma \vec{n}, \tag{6}$$

where $\gamma > 0$ is a regularization parameter. To estimate the principal normal $\vec{n}$, we have collected two classes of MR/HR face images and trained a linear SVM to find the separation hyperplane. The unit vector orthogonal to the hyperplane that separates the two classes is taken as principal normal $\vec{n}$.

5.2. Improving FSR *via* latent space interpolation

Despite the impressive synthesis performance delivered by GAN and its variants, less has been studied about the inversion problem—i.e., given an input image X^{MR}, can we find a latent code $\vec{z}$ such that $g(\vec{z}) \approx I$? This is an ill-posed problem because the projected latent vectors are often unstable, and small perturbations in the latent space could result in significant quality distortions (e.g., noticeable artifacts) in the reconstructed images. Inspired by recent work on Gaussianized latent spaces (Wulff and Torralba, 2020), we propose regularizing the process of inversion or imposing a Laplacian prior on the data distribution in the latent space as follows.

Similarly to previous work (Abdal et al., 2019; Wulff and Torralba, 2020), we formulate image inversion as the following continuous optimization problem in the latent space.

$$\vec{w} = \underset{\vec{z} \in W}{\operatorname{argmin}}\, L(X^{MR}, g(\vec{z})), \tag{7}$$

where $L(\cdot)$ denotes reconstruction loss (e.g., LPIPS perceptual distance, Zhang et al., 2018a). Alternatively, extended latent space $W+$ has often been found to be preferred to better invert natural images (Menon et al., 2020). Therefore, we can also project an input image onto the extended latent space as follows.

$$\vec{w}^{+} = \underset{\vec{z} \in W+}{\operatorname{argmin}}\, L(X^{MR}, g(\vec{z})). \tag{8}$$

Note that the expanded $W+$ latent space having the dimension of 18 × 512 does not represent an 18-time repetition of the 512-dimensional vector W, which explains its tremendous expressive power. However, the price paid for greater flexibility is the tendency to produce artifacts, especially when regularization is absent. Since both W and $W+$ latent spaces have their own strengths and weaknesses, it is natural to combine them to simultaneously preserve facial identity and suppress artifacts.

Based on the analysis in Wulff and Torralba (2020), we have developed an interpolation-based approach to regularize the codes ($\vec{w}, \vec{w}+$ in the latent space. The interpolated latent codes are obtained by moving the original code toward the empirical mean $\bar{w}$ by some scaling factors (γ, γ^{+}), that is,

$$\vec{w}_I = \vec{w} + \gamma(\bar{w} - \vec{w}), \vec{w}_I^{+} = \vec{w}^{+} + \gamma^{+}(\bar{w} - \vec{w}^{+}) \tag{9}$$

Then we take the weighted code as the final output.

$$\vec{w}^{HR} = a\vec{w}_I + (1 - a)\vec{w}_I^{+}, \tag{10}$$

where $0 < a < 1$ is inversely proportional to the projection error in the latent space, that is,

$$a = \frac{||E_{P+}||^{-1}}{||E_{P+}||^{-1} + ||E_P||^{-1}}. \tag{11}$$

where E_P and E_{P+} denote the mean squared error (MSE) of the projection in the W and $W+$ space, respectively (note that both MSE values are strictly bounded away from zeros). For example, when projected into the W space, we have the following.

$$E_P = ||X - g(e(X))||_{L_2}. \tag{12}$$

6. Experimental results

6.1. Dataset collection

We have tested our two-step FSR method on simulated and real LR face images. The first is generated by downsampling HR images (following a procedure similar to PULSE, Menon et al., 2020); the latter consists of real LR face images taken from the Widerface dataset (Yang et al., 2016).

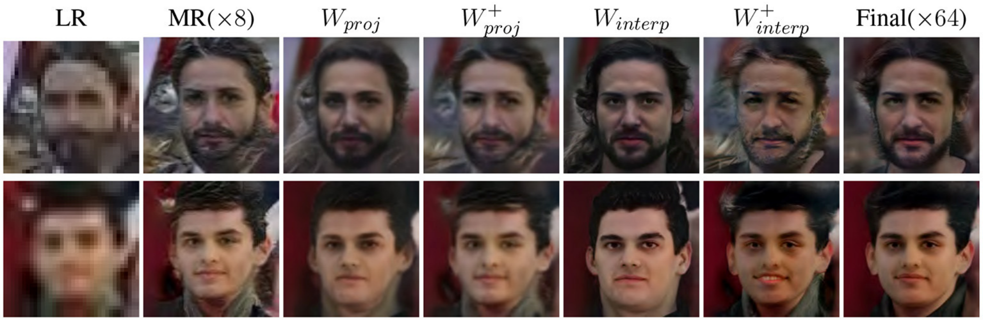

FIGURE 5
Comparison of visual quality among different stages of our two-step FSR method. From **left** to **right**: low resolution 16 × 16 input image, The mid-resolution 128 × 128 image. W_{proj} and W^+_{proj}: The projection of the Mid-resolution image into the W and $W+$ spaces of StyleGan (Karras et al., 2019). W_{interp} and W^+_{interp} indicate the interpolated vectors in the W and $W+$ spaces of StyleGan (Karras et al., 2019) using Equation (9). Final results is obtained by combining the interpolated vectors using Equation (10).

6.1.1. High resolution data (HR)

We have used several publicly available HR face datasets with a resolution of 1024 × 1024: CelebA HQ (Karras et al., 2017) (30,000 images) and FFHQ (Karras et al., 2019) (70,000 images). Both datasets contain considerable variations in terms of the age, ethnicity, and background of the image. We have paid special attention to the issue of racial bias as suggested by Menon et al. (2020) and tried to achieve a good balance between different races in the training data.

6.1.2. Simulated LR data

Similar to previous work (Chen et al., 2018; Menon et al., 2020), we have downsampled our HR face images by a factor of eight using the bicubic method provided by Matlab. Note that the use of simulated LR is only for the study of supervised learning, which requires paired HR-LR training data.

6.1.3. Realistic LR data

To simulate the real world scenario (Bulat et al., 2018), we have followed a similar protocol to create our real LR dataset from Widerface (Yang et al., 2016). The face regions are cropped using Zhang et al. (2017), which ended up with a total of 156,557 real LR training images and 8,241 real LR testing images. All images have been resized to 16 × 16 for the study of unsupervised learning (there is no HR ground truth available).

6.2. Ablation study

StyleGAN-based face image editing in latent space is known to suffer from notorious artifacts (Karras et al., 2020). This is because latent codes are often unstable, so small perturbations in the latent space could lead to significant image distortions (Wulff and Torralba, 2020). To demonstrate how a Laplacian prior helps the suppression of artifacts, we have compared the reconstruction images without and with the prior (please refer to Figure 5). The suppression of artifacts by a Laplacian prior can be observed. More results of image comparison can be found in the study in different stages of our proposed method (see Figure 6). Table 1 shows the comparison of objective metrics between the results with and without the Laplacian prior (note that a lower FID or NIQE value corresponds to better image quality).

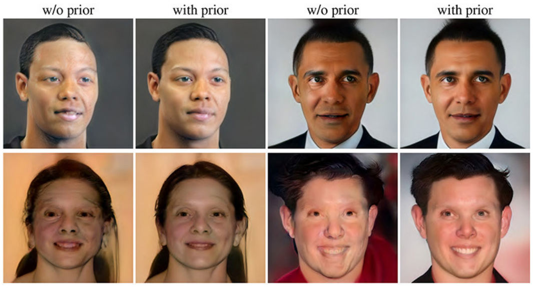

FIGURE 6
Artifact reduction by imposing the Laplacian prior.

TABLE 1 Quantitative comparison of visual quality improvement by imposing the Laplacian prior.

Configuration	FID	NIQE
Without prior	215.37	**12.7**
With prior	**207.5**	13.18

The bold values indicate the highest quality within each column.

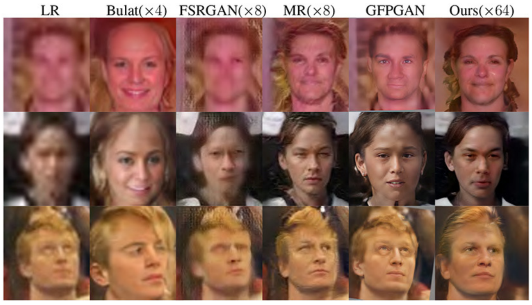

FIGURE 7
Comparison of FSR results on the Widerface dataset between ours, FSRGAN (Chen et al., 2018), Bulat's method (Bulat et al., 2018), and GFPGAN (Wang et al., 2021) (no GT is available).

TABLE 2 Quantitative comparison of visual quality between PULSE (Menon et al., 2020), GFPGAN (Wang et al., 2021) and ours.

Method	FID	NIQE
Pulse (Menon et al., 2020)	238.4	**9.62**
GFPGAN (Wang et al., 2021)	271.62	15.89
Ours	**181.7**	12.28

The bold values indicate the highest quality within each column.

6.3. SR performance comparison with others

6.3.1. Visual quality comparison on simulated LR

We first report our experimental results for SLR data (artificially created low-resolution face images) and compare them with the state-of-the-art PULSE method (Menon et al., 2020). Although synthetic, SLR images are still useful because they have ground truth (HR) available and appropriate to gauge the performance of supervised learning (with paired HR-LR training data). Figure 7 shows the qualitative comparisons between our supervised/unsupervised approach and the existing PULSE method (Menon et al., 2020). It can be easily verified that ours can produce visually more convincing and pleasant HR results than PULSE (e.g., sharper contrast and fewer artifacts around the earrings), as well as better ability to preserve the original ethnic information and head position of the ground truth image; first and second rows of Figure 7. Furthermore, we present an objective metric comparison between our method and the pulse (Menon et al., 2020) present in Table 2.

6.3.2. Visual quality comparison on realistic LR

We have tested our two-step method on the popular widerface LR dataset in the real world (Yang et al., 2016). This dataset is particularly challenging for face detection and SR because its 393,703 faces contain a high degree of variability in scale, pose, and occlusion. We report that our proposed method outperforms FSRGAN (Chen et al., 2018) and the unsupervised FSR method (Bulat et al., 2018) in visual quality (Figure 8). Furthermore, we calculate the Fréchet inception distance (FID) (Heusel et al., 2017) and the Naturalness Image Quality Evaluator (NIQE) (Mittal et al., 2012) scores to verify that our method can outperform existing methods on both metrics. FID has been widely used as a metric used to assess the quality of images created by GAN-based generative models; NIQE is a popular blind image quality metric that measures the distance based on natural scene statistics (NSS) between a given image and the training dataset. The metrics are calculated by averaging the results of different images. Table 3 shows the results.

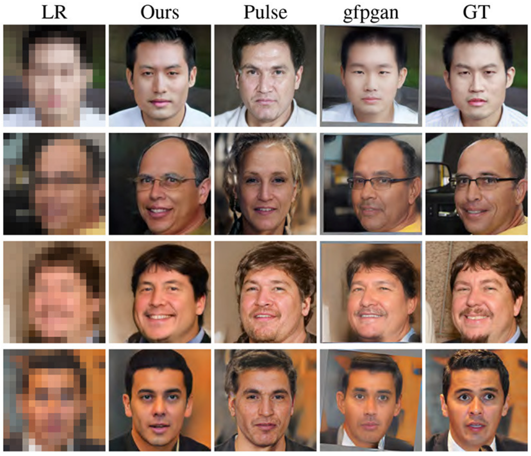

FIGURE 8
Comparison of the image quality of the extreme FSR between ours, PULSE (Menon et al., 2020) and GFPGAN (Wang et al., 2021).

TABLE 3 Quantitative comparison of visual quality between other competing methods and ours.

Method	FID	NIQE
FSRGAN (Chen et al., 2018)	420.8	18.72
Bulat et al. (2018)	481.4	19.05
Ours (mid-resolution)	447.4	18.73
GFPGAN (Wang et al., 2021)	361.63	15.59
Ours (Final)	**343.7**	**12.49**

The bold values indicate the highest quality within each column.

6.4. Discussions and limitations

6.4.1. Critical resolution for preserving facial identity

As mentioned above, the smallest spatial resolution for human operators to tell the identity of a person has been

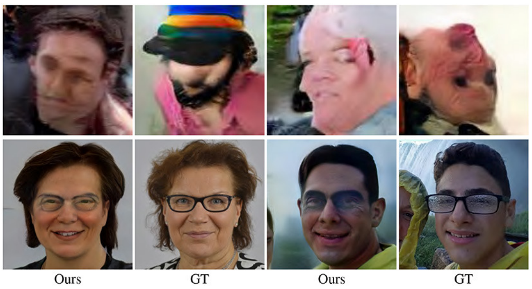

FIGURE 9
Failure cases in our study. **Top**: step-1 (due to extreme pose conditions and expressions); **bottom**: step-2 (due to poor handling of occlusion by eyeglasses).

found to be 18 × 24 (Bachmann, 1991). However, the relationship between spatial resolution and face recognition has not been well-studied in the literature; among the few existing works, seven usable pixels between eyes were shown to be sufficient to identify the Boston Bomber (Abiantun et al., 2019). To investigate in a more systematic way, we conducted an experiment with the Openface face matcher (Amos et al., 2016) at different scaling factors (×64, ×32, ×16 etc.). It can be observed from Figure 1 that (1) this work significantly outperforms PULSE (Menon et al., 2020) in terms of matching performance across all resolutions; (2) there is a critical resolution of 64 × 64 below which matching performance degrades rapidly. This observation indirectly justifies the plausibility of our two-step approach, where the magnetic resonance image has to serve as a step-stone to facilitate the extraction of identity information.

6.4.2. Failure cases and computational issues

To suppress artifacts, our approach intentionally skips the step of data augmentation. It turns out that it is still sensitive to extreme variations in face pose, such as those shown in Figure 9. Meanwhile, artifacts may still appear due to severe occlusions (e.g., eyeglasses); those real LR examples are rare even among the Widerface dataset (refer to Figure 9). We have trained step-1 of the overall architecture on four Titan GTX GPUs for around 14 h. The low-to-high network has a 34 MB size for a test time of less than 1 s. For the second step, the projection time of an MR image in the latent space is ≈ 90 s; and it takes around 4 min to obtain the final output. However, we note that other semantic face image manipulation in the latent space (e.g., InterFaceGAN, Shen et al., 2020b) also have comparable complexity when StyleGAN was used for extracting the latent code. How to improve the computational efficiency of StyleGAN-based face image manipulation is left for future studies.

7. Conclusions

In this paper, we have studied the problem of extreme FSR and proposed a novel two-step FSR method that combines self-supervised CycleGAN with StyleGAN. In the first step of embedded face hallucination, we aim at learning an unknown degradation model in the real world using a self-supervised CylceGAN approach that combines a style-based generator with a relativistic discriminator. We demonstrate that the intermediate MR image is capable of preserving facial identity. In the second step of face reconstruction, the MR face image serves as a stepping stone to generate super-resolved HR images at ×64. A Laplacian prior is imposed to regularize the inversion process in the latent space for artifact suppression. Unlike previous works, our two-step approach takes advantage of the identity code embedded in the latent space of MR images as the connection bridging LR and HR. Extensive experimental results are reported to demonstrate the superiority of this work over other competing FSR approaches. Our two-step approach has achieved highly competitive and often better performance than others in terms of both subjective and objective qualities, especially in the extreme case of magnification ratios.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://shuoyang1213.me/WIDERFACE/. FFHQ: https://www.kaggle.com/datasets/arnaud58/flickrfaceshq-dataset-ffhq, CelebAHQ: http://mmlab.ie.cuhk.edu.hk/projects/CelebA/CelebAMask_HQ.html.

Author contributions

AC conducted all experiments and drafted the paper. XL discussed the technical ideas, provided financial support, and revised the paper. Both authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdal, R., Qin, Y., and Wonka, P. (2019). "Image2stylegan: how to embed images into the stylegan latent space?" in *Proceedings of the IEEE International Conference on Computer Vision*, 4432–4441.

Abiantun, R., Juefei-Xu, F., Prabhu, U., and Savvides, M. (2019). SSR2: sparse signal recovery for single-image super-resolution on faces with extreme low resolutions. *Pattern Recogn.* 90, 308–324. doi: 10.1016/j.patcog.2019.01.032

Amos, B., Ludwiczuk, B., Satyanarayanan, M., et al. (2016). Openface: a general-purpose face recognition library with mobile applications. *CMU Schl. Comput. Sci.* 6, 2. doi: 10.1080/09541449108406221

Bachmann, T. (1991). Identification of spatially quantised tachistoscopic images of faces: how many pixels does it take to carry identity? *Eur. J. Cogn. Psychol.* 3, 87–103. doi: 10.1109/TPAMI.2016.2644615

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). SEGNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495.

Bulat, A., Yang, J., and Tzimiropoulos, G. (2018). "To learn image super-resolution, use a GAN to learn how to do image degradation first," In *Proceedings of the European Conference on Computer Vision (ECCV)*, 185–200.

Cai, J., Zeng, H., Yong, H., Cao, Z., and Zhang, L. (2019). "Toward real-world single image super-resolution: a new benchmark and a new model," In *Proceedings of the IEEE International Conference on Computer Vision*, 3086–3095.

Chen, Y., Tai, Y., Liu, X., Shen, C., and Yang, J. (2018). "FSRNet: end-to-end learning face super-resolution with facial priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2492–2501.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al.. (2014). "Generative adversarial nets," in *Communications of the ACM* (ACM), 139–144.

Gu, J., Lu, H., Zuo, W., and Dong, C. (2019). "Blind super-resolution with iterative kernel correction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1604–1613.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, p. 30.

Huang, X., and Belongie, S. (2017). "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

Jia, K., and Gong, S. (2008). Generalized face super-resolution. *IEEE Trans. Image Process.* 17, 873–886. doi: 10.1109/TIP.2008.922421

Jolicoeur-Martineau, A. (2018). The relativistic discriminator: a key element missing from standard GAN. *arXiv [Preprint].* arXiv:1807.00734.

Kalarot, R., Li, T., and Porikli, F. (2020). "Component attention guided face super-resolution network: CAGFace," in *The IEEE Winter Conference on Applications of Computer Vision*, 370–380.

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

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

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

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint].* arXiv:1412.6980.

Köhler, T., Bätz, M., Naderi, F., Kaup, A., Maier, A., and Riess, C. (2019). Toward bridging the simulated-to-real gap: benchmarking super-resolution on real data. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2944–2959. doi: 10.1109/TPAMI.2019.2917037

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., et al. (2017). "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4681–4690.

Liu, C., Shum, H.-Y., and Freeman, W. T. (2007). Face hallucination: theory and practice. *Int. J. Comput. Vis.* 75, 115–134. doi: 10.1007/s11263-006-0029-5

Liu, H., Liu, J., Hou, S., Tao, T., and Han, J. (2021a). Perception consistency ultrasound image super-resolution via self-supervised CycleGAN. *Neural Comput. Appl.* 1–11. doi: 10.1007/s00521-020-05687-9

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., et al. (2021b). Self-supervised learning: generative or contrastive. *IEEE Trans. Knowledge Data Eng.* doi: 10.1109/TKDE.2021.3090866

Lugmayr, A., Danelljan, M., and Timofte, R. (2019). "Unsupervised learning for real-world super-resolution," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (IEEE), 3408–3416.

Menon, S., Damian, A., Hu, S., Ravi, N., and Rudin, C. (2020). "Pulse: self-supervised photo upsampling via latent space exploration of generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2437–2445.

Mittal, A., Soundararajan, R., and Bovik, A. C. (2012). Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.* 20, 209–212. doi: 10.1109/LSP.2012.2227726

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv [Preprint].* arXiv:1802.05957.

Nguyen, N. L., Anger, J., Davy, A., Arias, P., and Facciolo, G. (2021). "Self-supervised multi-image super-resolution for push-frame satellite images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1121–1131.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "PyTorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems, Vol. 32*, 8026–8037.

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

Shen, Y., Yang, C., Tang, X., and Zhou, B. (2020b). InterfaceGAN: interpreting the disentangled face representation learned by GANs. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 2004–2018. doi: 10.1109/TPAMI.2020.3034267

Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H.-P., Pérez, P., et al. (2020). "StyleRig: rigging styleGAN for 3D control over portrait images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6142–6151.

Viazovetskyi, Y., Ivashkin, V., and Kashin, E. (2020). StyleGAN2 distillation for feed-forward image manipulation. *arXiv [Preprint].* arXiv:2003.03581. doi: 10.1007/978-3-030-58542-6_11

Wang, N., Tao, D., Gao, X., Li, X., and Li, J. (2014). A comprehensive survey to face hallucination. *Int. J. Comput. Vis.* 106, 9–30. doi: 10.1007/s11263-013-0645-9

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803.

Wang, X., Li, Y., Zhang, H., and Shan, Y. (2021). "Towards real-world blind face restoration with generative facial prior," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Wei, Y., Gu, S., Li, Y., and Jin, L. (2020). Unsupervised real-world image super resolution via domain-distance aware training. *arXiv [Preprint].* arXiv:2004.01178. doi: 10.1109/CVPR46437.2021.01318

Wulff, J., and Torralba, A. (2020). Improving inversion and generation diversity in styleGAN using a Gaussianized latent space. *arXiv [Preprint].* arXiv:2009.06529.

Yang, S., Luo, P., Loy, C.-C., and Tang, X. (2016). "Wider face: a face detection benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5525–5533.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2019). "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, 7354–7363.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018a). "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., and Li, S. Z. (2017). S^3FD: single shot scale-invariant face detector. *arXiv [Preprint].* arXiv: 1708.05237. doi: 10.1109/ICCV.2017.30

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. (2018b). "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 286–301.

Zhang, Z., Wang, R., Zhang, H., Chen, Y., and Zuo, W. (2022). Self-supervised learning for real-world super-resolution from dual zoomed observations. *arXiv [Preprint]*. arXiv:2203.01325. doi: 10.1007/978-3-031-19797-0_35

Zhao, C., Dewey, B. E., Pham, D. L., Calabresi, P. A., Reich, D. S., and Prince, J. L. (2020). Smore: a self-supervised anti-aliasing and super-resolution algorithm for mri using deep learning. *IEEE Trans. Med. Imaging* 40, 805–817. doi: 10.1109/TMI.2020.3037187

Zhu, J., Shen, Y., Zhao, D., and Zhou, B. (2020). In-domain gan inversion for real image editing. *arXiv [Preprint]*. arXiv:2004.00049. doi: 10.1007/978-3-030-58520-4_35