GAN Prior Based Null-Space Learning for Consistent Super-resolution

Yinhuai Wang[1], **Yujie Hu**[1], **Jiwen Yu**[1], **Jian Zhang**[1,2]

[1]Peking University Shenzhen Graduate School, China
[2]Peng Cheng Laboratory, China
{yinhuai; hhuyujie; yujiwen}@stu.pku.edu.cn, zhangjian.sz@pku.edu.cn

Abstract

Consistency and *realness* have always been the two critical issues of image super-resolution. While the *realness* has been dramatically improved with the use of GAN prior, the state-of-the-art methods still suffer inconsistencies in local structures and colors (e.g., tooth and eyes). In this paper, we show that these inconsistencies can be analytically eliminated by learning only the null-space component while fixing the range-space part. Further, we design a pooling-based decomposition (PD), a universal range-null space decomposition for super-resolution tasks, which is concise, fast, and parameter-free. PD can be easily applied to state-of-the-art GAN Prior based SR methods to eliminate their inconsistencies, neither compromise the *realness* nor bring extra parameters or computational costs. Besides, our ablation studies reveal that PD can replace pixel-wise losses for training and achieve better generalization performance when facing unseen downsamplings or even real-world degradation. Experiments show that the use of PD refreshes state-of-the-art SR performance and speeds up the convergence of training up to **2**$\sim$**10** times.

Introduction

Image super-resolution (SR) is a process that generates a high-resolution (HR) image that is consistent with the input low-resolution (LR) image. It consists of two critical constraints: (1) Low-frequency consistency, i.e., the downsampling of the HR image should be identical to the LR image. (2) Realness, i.e., the HR image should be within the distribution of natural images. In this paper, we abbreviate these two constraints to *consistency* and *realness*.

Prevailing deep neural network (DNN) based SR methods usually relies on pixel-wise losses (e.g., ℓ_1, ℓ_2) to learn both *realness* and *consistency*. Specifically, they first downsample the ground-truth (GT) images to generate LR-GT image pairs, then feed the LR images into a DNN to generate an HR image and optimize the DNN by minimizing the pixel-wise losses between HR and GT. When the GT images are of high-quality, the pixel-wise losses measures both the *realness* and *consistency* constraints. Generative adversarial network (GAN) and the adversarial training objectives (Goodfellow et al. 2014) are proved effective as a guidance toward *realness* (Ledig et al. 2017). The Recent breakthrough in GANs enables generating high-resolution images (Karras et al. 2020), and encourages utilizing pretrained GAN models as the GAN prior for SR tasks (Menon et al. 2020; Wang et al. 2021; Chan et al. 2021; Wang, Hu, and Zhang 2022).

Though superior in hallucinating realistic details, GAN prior may bring undesirable bias in generating those details, causing inconsistencies in structures and colors, e.g., eye, tooth, and skin, as can be seen in Fig. 1. Further increasing the weight of ℓ_1 or ℓ_2 losses may ease these inconsistencies while over-smoothing the rich details that GAN Prior brings. Since *realness* and *consistency* are both critical for SR, we need to find a way to assure *consistency* for GAN prior based SR without compromise its superiority in *realness*.

We observe there exists a fundamental physical prior in SR tasks. That is, the LR image (assuming the LR image is clean) already contains low-frequency information that constitutes the GT image. However, recent GAN prior based SR methods do not fully leverage this physical prior. To utilize this physical prior, we can resort to the range-null space decomposition, which is well studied in linear algebra and inverse problems (Schwab, Antholzer, and Haltmeier 2019).

Theoretically, given a downsampler and its pseudo-inverse, any HR image can be decomposed into a range-space part and a null-space part. The range-space part is fully responsible for the LR image while the null-space part is irrelevant to the LR image. Actually, the range-space part of the GT can be calculated explicitly. Then we can explicitly assure the *consistency* of a predicted HR image by replacing its range-space part with the GT's range-space part.

However, the pseudo-inverse of the downsampler is usually hard to obtain, let alone so many types of downsamplers in SR tasks. We observe that most of the downsampling with antialiasing shares very similar results. We also notice a wildly used downsampling method, average-pooling, possesses concise forms of pseudo-inverse. Hence we use average-pooling as a universal measurement for the *consistency*. This provides an efficient solution to use range-null space decomposition for consistent SR.

Specifically, we use the pseudo-inverse to directly upsample the LR image as the low-frequency part of the result, then use the GAN prior based network to predict a raw HR im-

This paper was supported in part by Shenzhen General Research Project under Grant JCYJ20220531093215035. *Corresponding author: Jian Zhang*.

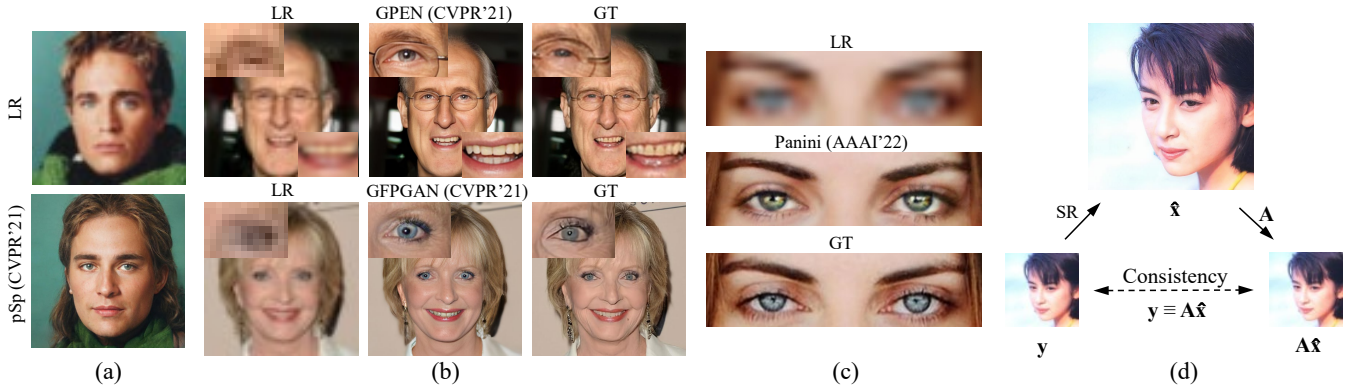


Figure 1: GAN prior based SR methods suffer inconsistencies. Here we take four state-of-the-art GAN prior based SR methods for example. (a) pSp fails to maintain the identity of the low-resolution (LR) image; (b) GFP-GAN and GPEN can not faithfully restore the eye structures and colors of the LR; (c) Panini sometimes appears evident bias of eye colors during training. In these cases, it is clear that the LR images already contain rough structures and colors. However, the mentioned methods neglect these physical prior and entirely rely on networks to generate the SR result, causing inconsistencies. Instead, our proposed method assures downsampling consistency (defined by $\mathbf{A}\hat{\mathbf{x}} \equiv \mathbf{y}$) analytically, as is illustrated in (d), where $\mathbf{y}$, $\hat{\mathbf{x}}$, and $\mathbf{A}$ represents the LR image, SR result, and downsampling operator, respectively. Codes: https://github.com/wyhuai/RND.

age from the LR image, and use the calculated null-space operator to extract its high-frequency part. Finally, we add the low-frequency part and the high-frequency part to form an SR result that owns the *consistency*. The whole process is called Pooling-based Decomposition (PD) in this paper.

As a physical prior of *consistency*, the involvement of PD not only enables existing state-of-the-art GAN prior based SR methods to reach obviously higher performance in *consistency* but also significantly accelerates training convergence, as is revealed in our experiments. Besides, our ablation studies show that PD can replace pixel-wise losses for training and achieves better generalization performance when facing unseen downsamplings or real-world degradation, implying its potential for further explorations.

Our contributions include:

- We theoretically and experimentally reveal that combining GAN prior with physical prior yields significantly better *consistency*, outperforming state-of-the-art GAN prior based SR methods in several datasets.
- We propose a novel Pooling-based Decomposition (PD) for GAN prior based null-space learning. It is parameter-free, fast, and mathematically rigorous. PD can be used for GAN prior based SR methods to improve the *consistency* and accelerate training convergence significantly.

Related Work

GAN Prior based Super-Resolution. Typical DNN based super-resolution (SR) methods use pixel-wise constraints to encourage learning the inverse mapping of downsampling (Dong et al. 2015; Kim, Lee, and Lee 2016; Zhang et al. 2018a,b; Huang et al. 2017; Li et al. 2018; Lin, Zhou, and Chen 2018; Lin et al. 2020; Chong Mou 2021, 2022). These methods usually perform well in pixel-wise metrics, e.g., PSNR and SSIM, but tend to generate smooth structures with poor details. Owing to the excellent performance in generating realistic details, GANs and adversarial training strategies (Goodfellow et al. 2014) are applied for SR to produce higher visual quality (Wang et al. 2018; Yu and Porikli 2017; Yang et al. 2020; Wan et al. 2020; Zhang and Ling 2021; Mou et al. 2022). Recently, Karras et al. proposed a series of StyleGAN that can generate excellent realistic images of certain types, typically human faces (Karras, Laine, and Aila 2019; Karras et al. 2020, 2021). The success of StyleGAN inspired a lot of works in image editing (Abdal, Qin, and Wonka 2019, 2020; Nitzan et al. 2020; Tewari et al. 2020; Zhu et al. 2020; Patashnik et al. 2021; Gu, Shen, and Zhou 2020; Xia et al. 2022) and encouraged the utilizing of StyleGAN as GAN prior for SR tasks. Some methods attempt utilizing a pretrained GAN to solve SR by optimizing the latent code iteratively to maximize the data consistency (Yeh et al. 2018; Luo and Wu 2020; Menon et al. 2020). Richardson et al. further propose using an encoder to map the input image into the latent space of StyleGAN (Richardson et al. 2021). Since they directly utilize the latent space of StyleGAN, the result owns excellent *realness*. However, the latent space is not expressive enough to preserve the identity of the input image, thus suffering identity shifts. To improve the *consistency*, some methods build skip connections between the encoder and StyleGAN to better capture structural information (Yang et al. 2021; Wang et al. 2021; Chan et al. 2021). Wang et al. further propose interpolating StyleGAN features and encoder features to yield better *consistency* and *realness* (Wang, Hu, and Zhang 2022). These GAN prior based methods achieve unprecedented results in face SR. Though pixel-wise losses are already applied during their training, the use of GAN prior may still bring undesirable inconsistency in facial structures and colors, as is shown in Fig. 1. Besides, the success of these methods relies heavily on trial and error, lacking rigorous explanation and mathematical proof. Instead, our proposed PD follows rigorous derivation and can be applied to the mentioned methods

to completely resolve inconsistency while preserving the *realness* that GAN prior brings.

Consistent Super-Resolution. Traditional approaches for solving super-resolution (SR) are typically model-based, which usually regularize the results with prior knowledge about distributions of natural images, e.g., sparsity (Yang et al. 2010). Most of the recent state-of-the-art SR methods are DNN based and are model-free (Liang et al. 2021; Chan et al. 2021), thus lack interpretation and heavily relying on extensive experiments on network structures. To implement a consistent solution for inverse problems, Chen et al. (Chen and Davies 2020) analytically decompose the inverse result into range-space and null-space of a specific linear operator and learn them respectively. However, they take bicubic interpolation as the pseudo-inverse of bicubic interpolation for image SR, which is a lossy approximation. Bahat et al. (Bahat and Michaeli 2020) further resort to the Fourier domain to calculate the pseudo-inverse of the downsampler $\mathbf{A}$. However, their formulation contains inverse matrix $(\mathbf{A}\mathbf{A}^\top)^{-1}$, which may not exist for not-full rank $\mathbf{A}\mathbf{A}^\top$. Hence the pseudo-inverse calculated by FFT and IFFT may still be lossy in theory, let alone the complexity and inaccuracy in implementations. Besides, they did not combine null-space learning with GAN prior, thus can not yield high-quality details. Instead, we show that the GAN prior is an ideal tool to fill in the missing null-space in SR problems, and the implementations of our proposed PD perfectly fit the theory and possess concise forms.

Method

Preliminaries: Range-Null Space Decomposition

Given a non-zero linear operator $\mathbf{A} \in \mathbb{R}^{d\times D}$, it usually has at least one pseudo-inverse $\mathbf{A}^\dagger \in \mathbb{R}^{D\times d}$ that satisfies $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = \mathbf{A}$. In particular, an analytical solution by SVD:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \quad \mathbf{A}^\dagger = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^\top, \tag{1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrix and $\mathbf{\Sigma}$ is a diagonal matrix with eigenvalues as its diagonal elements. $\mathbf{A}^\dagger\mathbf{A}$ can be seen as the operator that projects samples to the range-space of $\mathbf{A}$, since $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} \equiv \mathbf{A}$. While $(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})$ can be seen as the operator that projects samples to the null-space of $\mathbf{A}$, since $\mathbf{A}(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}) \equiv \mathbf{0}$.

Any sample $\mathbf{x} \in \mathbb{R}^{D\times 1}$ can be decomposed into two parts: the part that locates at the range-space of $\mathbf{A}$ and the other part that locates at the null-space of $\mathbf{A}$, i.e.,

$$\mathbf{x} \equiv \mathbf{A}^\dagger\mathbf{A}\mathbf{x} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x} \tag{2}$$

GAN Prior Based Null-Space Learning

In image super-resolution, the linear operator $\mathbf{A}$ becomes a downsampler while its pseudo-inverse $\mathbf{A}^\dagger$ represents an upsampler. We consider naive downsampling of the form:

$$\mathbf{y} = \mathbf{A}\mathbf{x}. \tag{3}$$

Given a low-resolution (LR) image $\mathbf{y} \in \mathbb{R}^{d\times 1}$ that downsampled from a ground-truth (GT) image $\mathbf{x} \in \mathbb{R}^{D\times 1}$ using downsampling operator $\mathbf{A} \in \mathbb{R}^{d\times D}$, our goal is to get a high-resolution (HR) image $\hat{\mathbf{x}} \in \mathbb{R}^{D\times 1}$ that conforms to:

$$Consistency: \quad \mathbf{A}\hat{\mathbf{x}} \equiv \mathbf{y}, \tag{4}$$

$$Realness: \quad \hat{\mathbf{x}} \sim p(\mathbf{x}), \tag{5}$$

where $p(\mathbf{x})$ denotes the GT distribution. Though these two properties can both be optimized by training a DNN using pixel-wise and adversarial objectives, it does not thoroughly utilize the physical prior contained in the LR image.

We observe that the range-null space decomposition is an ideal tool to ensure *consistency*. Lets first decompose the GT $\mathbf{x}$ into two parts follow Eq. (2), then downsample $\mathbf{x}$ with $\mathbf{A}$, i.e., combining Eq. (2) with Eq. (3), it becomes:

$$\begin{aligned}\mathbf{A}\mathbf{x} &= \mathbf{A}\mathbf{A}^\dagger\mathbf{A}\mathbf{x} + \mathbf{A}(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x} \\ &= \mathbf{A}\mathbf{x} + \mathbf{0} = \mathbf{y}.\end{aligned} \tag{6}$$

We can see that the range-space part, $\mathbf{A}^\dagger\mathbf{A}\mathbf{x}$, after downsampling, becomes exactly the LR image $\mathbf{y}$, while the null-space part, $(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}$, is transparent to the downsampler $\mathbf{A}$. Following these observations, we can formulate our HR result $\hat{\mathbf{x}}$ into two parts: a range-space part that is set as the GT's range-space part $\mathbf{A}^\dagger\mathbf{A}\mathbf{x}$, i.e., $\mathbf{A}^\dagger\mathbf{y}$; a null-space part $(\mathbf{I}-\mathbf{A}^\dagger\mathbf{A})\hat{\mathbf{x}}_r$ extracted from the raw prediction $\hat{\mathbf{x}}_r$ of a GAN prior based network.

$$\hat{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\hat{\mathbf{x}}_r, \tag{7}$$

Now the result $\hat{\mathbf{x}}$ owns *consistency* since $\mathbf{A}\hat{\mathbf{x}} \equiv \mathbf{y}$ holds.

However, it is usually hard to get the pseudo-inverse of $\mathbf{A}$, let alone the situation that $\mathbf{A}$ is unknown.

Pooling-Based Decomposition

We observe that many downsampling methods with antialiasing share very similar results. Among them, average-pooling is the most simple one and owns desirable forms of pseudo-inverse. Since we focus on the structural and color inconsistencies in GAN prior based SR methods, the downsampling consistency measured by average-pooling is robust enough to assure that.

Specifically, for a certain downsampling rate s, average-pooling split the HR image $\mathbf{x}$ into patches of size $s \times s$, and calculate each patch's average value as the value of a corresponding pixel in the LR image $\mathbf{y}$. A simple pseudo-inverse of average-pooling replicates each pixel of the LR image into corresponding patches of size $s \times s$.

Considering operation on a single pixel, the average-pooling with scale s can be represented as a $s^2 \times 1$ matrix $\mathbf{A}$ with its pseudo-inverse as a $1 \times s^2$ matrix $\mathbf{A}^\dagger$:

$$\mathbf{A}^\dagger = (1, ..., 1)^\top, \quad \mathbf{A} = (1/s^2, ..., 1/s^2). \tag{8}$$

It is obvious that $\mathbf{A}\mathbf{A}^\dagger = \mathbf{I}$ holds. Since the operations between each pixel or each patch are independent, this formulation can be easily promoted to the whole image.

In practice, we use average-pooling in linear operator form, denoted as $\mathcal{P}_\downarrow(\cdot)$, and use broadcast mechanism to implement the pseudo-inverse of average-pooling, denoted as $\mathcal{P}_\uparrow(\cdot)$. Note that $\mathcal{P}_\downarrow(\mathcal{P}_\uparrow(\cdot)) = \mathcal{I}(\cdot)$, where $\mathcal{I}(\cdot)$ is the unit linear operator. Then we can rewrite our SR solution as:

$$\hat{\mathbf{x}} = \mathcal{P}_\uparrow(\mathbf{y}) + \hat{\mathbf{x}}_r - \mathcal{P}_\uparrow(\mathcal{P}_\downarrow(\hat{\mathbf{x}}_r)), \tag{9}$$

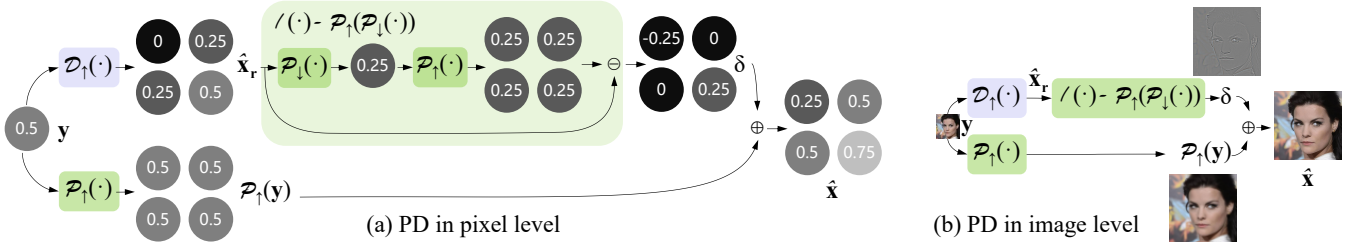


Figure 2: Illustration of pooling-based decomposition (PD). $\mathcal{D}_\uparrow(\cdot)$ denotes the network. $\mathcal{P}_\downarrow(\cdot)$ and $\mathcal{P}_\uparrow(\cdot)$ denote the average pooling and its pseudo-inverse. (a) shows a pixel level example of PD in $2\times$ SR. For each pixel in the LR image $\mathbf{y}$, we use $\mathcal{P}_\uparrow(\cdot)$ to replicate it to the size of 2×2. For the corresponding patch that predicted by the network $\mathcal{D}_\uparrow$, we subtract its own average value to get the high-frequencies $\hat{\mathbf{x}}_r - \mathcal{P}_\uparrow(\mathcal{P}_\downarrow(\hat{\mathbf{x}}_r))$, denoted as δ. Then we aggregate δ and $\mathcal{P}_\uparrow(\mathbf{y})$ as the result patch $\hat{\mathbf{x}}$. Note $\hat{\mathbf{x}}$ is consistent with $\mathbf{y}$, i.e., $\mathcal{P}_\downarrow(\hat{\mathbf{x}}) \equiv \mathbf{y}$. We can easily promote this paradigm to the image level, as is shown in (b).

where we use a GAN prior based DNN $\mathcal{D}_\uparrow$ to predict $\hat{\mathbf{x}}_r$:

$$\hat{\mathbf{x}}_r = \mathcal{D}_\uparrow(\mathbf{y}), \tag{10}$$

Note Eq. (9) satisfies the downsampling consistency:

$$\begin{aligned}\mathcal{P}_\downarrow(\hat{\mathbf{x}}) &= \mathcal{P}_\downarrow(\mathcal{P}_\uparrow(\mathbf{y})) + \mathcal{P}_\downarrow(\hat{\mathbf{x}}_r) - \mathcal{P}_\downarrow(\mathcal{P}_\uparrow(\mathcal{P}_\downarrow(\hat{\mathbf{x}}_r))), \\ &= \mathbf{y} + \mathcal{P}_\downarrow(\hat{\mathbf{x}}_r) - \mathcal{P}_\downarrow(\hat{\mathbf{x}}_r) \equiv \mathbf{y}.\end{aligned} \tag{11}$$

We name the whole operations of Eq. (9) as Pooling-based Decomposition, abbreviated as PD in this paper. Fig. 2 provides a detailed illustration of PD at pixel and image levels. $\mathcal{D}_\uparrow(\cdot)$ denotes the GAN prior based SR network. $\mathcal{P}_\downarrow(\cdot)$ and $\mathcal{P}_\uparrow(\cdot)$ denote the average pooling and its pseudo-inverse. Fig. 2(a) shows a pixel level example of PD in $2\times$ SR. For each pixel in the LR image $\mathbf{y}$, we use $\mathcal{P}_\uparrow(\cdot)$ to replicate it to the size of 2×2. For the corresponding patch $\hat{\mathbf{x}}_r$ that predicted by the network $\mathcal{D}_\uparrow(\cdot)$, we subtract $\hat{\mathbf{x}}_r$ by its low-frequency contents $\mathcal{P}_\uparrow(\mathcal{P}_\downarrow(\hat{\mathbf{x}}_r))$ to get the high-frequencies δ, i.e., $\delta = \hat{\mathbf{x}}_r - \mathcal{P}_\uparrow(\mathcal{P}_\downarrow(\hat{\mathbf{x}}_r))$. Then we aggregate δ and $\mathcal{P}_\uparrow(\mathbf{y})$ as the result patch $\hat{\mathbf{x}}$. It is worth noting that $\hat{\mathbf{x}}$ is consistent with $\mathbf{y}$, i.e., $\mathcal{P}_\downarrow(\hat{\mathbf{x}}) \equiv \mathbf{y}$. We can easily promote this paradigm to the image level, as is shown in Fig. 2(b).

Given a GAN prior based SR network, we can significantly elevate its *consistency* by simply imposing PD to its forward pipeline following Eq. (9) and Eq. (10) and applying the new forward pipeline to the training and inference. It is worth mentioning that PD is parameter-free and incurs negligible extra computations, making it an efficient tool for GAN Prior based SR networks to eliminate inconsistencies.

PD can be understood in many ways. Intuitively, $\mathcal{P}_\uparrow(\mathbf{y})$ can be seen as the low-frequency part (upsampled to match the size of $\mathbf{x}_r$) that is directly inhereted from the LR $\mathbf{y}$ and $\hat{\mathbf{x}}_r - \mathcal{P}_\uparrow(\mathcal{P}_\downarrow(\hat{\mathbf{x}}_r))$ be the high-frequency part that is extracted from the network prediction $\hat{\mathbf{x}}_r$). The operation $\mathcal{P}_\downarrow(\cdot)$ can be seen as a low-frequency filter, and the high-frequency part $\hat{\mathbf{x}}_r - \mathcal{P}_\uparrow(\mathcal{P}_\downarrow(\hat{\mathbf{x}}_r))$ yields no information after filtered by $\mathcal{P}_\downarrow(\cdot)$, since $\mathcal{P}_\downarrow(\hat{\mathbf{x}}_r - \mathcal{P}_\uparrow(\mathcal{P}_\downarrow(\hat{\mathbf{x}}_r))) \equiv \mathbf{0}$. Alternatively, we can see $\mathcal{P}_\uparrow(\mathbf{y}) - \mathcal{P}_\uparrow(\mathcal{P}_\downarrow(\hat{\mathbf{x}}_r))$ as the correction for the low-frequency contents of the initial prediction $\hat{\mathbf{x}}_r$. Or mathematically, we can see $\mathcal{P}_\uparrow(\mathbf{y})$ as the range-space part, which losses no information through observation $\mathcal{P}_\downarrow(\cdot)$, hence is set as fixed to assure consistency. While $\hat{\mathbf{x}}_r - \mathcal{P}_\uparrow(\mathcal{P}_\downarrow(\hat{\mathbf{x}}_r))$ is the null-space part, which leaves no information through observation $\mathcal{P}_\downarrow(\cdot)$, hence has fully flexibility for learning.

Experiments

We validate PD on two typical GAN prior based SR networks: Panini (Wang, Hu, and Zhang 2022) and GLEAN (Chan et al. 2021). We experiment SR on three typical categories: human face, cat, and church.

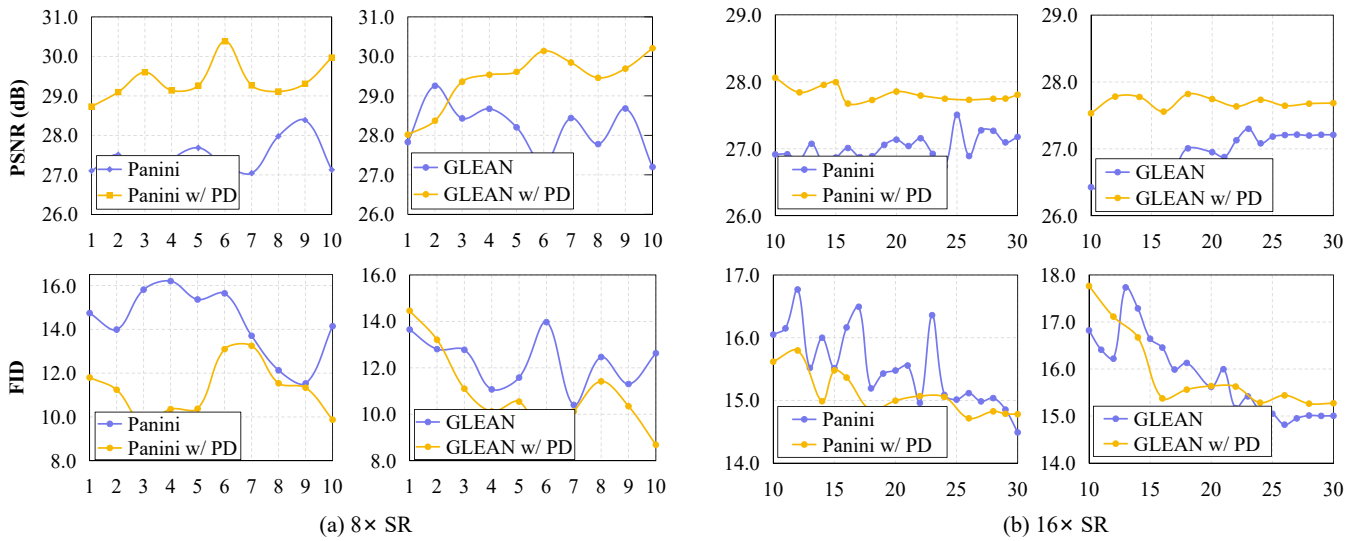
Validation of Theory

We theoretically proved that PD can inherently assure the low-frequency consistency: $\mathcal{P}_\downarrow(\hat{\mathbf{x}}) \equiv \mathbf{y}$. To verify that, we take an untrained Panini as the backbone (i.e., Panini with randomly initialized parameters) and use PD to generate the HR results. We take 100 images from the CelebA-HQ dataset and use $8\times$ average-pooling to generate the LR images. We take the LR image as the input $\mathbf{y}$ and use the "Panini w/ PD" to generate HR $\hat{\mathbf{x}}$, then calculate the PSNR($\mathbf{y}$, $\mathcal{P}_\downarrow(\hat{\mathbf{x}})$). The average PSNR of 100 image pairs reaches **145.7**, indicating that $\mathbf{y} \equiv \mathcal{P}_\downarrow(\hat{\mathbf{x}})$ holds with the use of PD. We also calculate the consistency of CEM, which use the Fourier transforms to calculate the pseudo-inverse $\mathbf{A}^\dagger$ of the $8\times$ bicubic(antialias) downsampler $\mathbf{A}$. However, the average PSNR($\mathbf{y}$, $\mathbf{A}^\dagger\hat{\mathbf{x}}$) is only **42.2**, indicating that the implementations of CEM suffer distortion.

We also compare the inference speed of PD and CEM. Specifically, we use random noise as $\hat{\mathbf{x}}_\mathbf{r}$ to exclude the use of networks, then use PD and CEM, respectively. We calculate the average running time of 100 inferences on a single Nvidia Terga4 GPU. We get 0.68ms for PD and 31.8ms for CEM, implying that PD is more efficient in execution.

$8\times$ Super-Resolution

To experiment $8\times$ SR on the human face, we train Panini, GLEAN, and their PD-based version on the FFHQ dataset (Karras, Laine, and Aila 2019). We use bicubic interpolation to synthesize LR images and follow the same training configuration as GLEAN and Panini. In detail, we use the Adam optimizer and Cosine Annealing Scheme with three training objectives: ℓ_1 loss, perceptual loss (Johnson, Alahi, and Fei-Fei 2016), and GAN loss (Goodfellow et al. 2014). For $8\times$ SR, the loss weights are set as 1, 1×10^{-2}, 1×10^{-2} respectively. The learning rate is set as 1×10^{-3}. See Appendix A for more details of training objectives.



(a) 8× SR (b) 16× SR

Figure 3: Convergence curves. Part (a) for the 8× face SR and (b) for the 16× face SR. With PD, both GLEAN and Panini yield significantly higher PSNR and comparable FID. The horizontal axis denotes the training iterations (10^4).

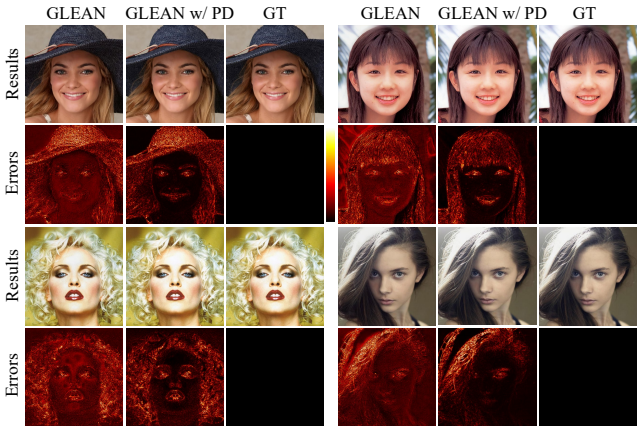


Figure 4: Qualitative results on 8× face SR. GLEAN yields high-quality results. However, it still suffers low-frequency inconsistencies, e.g., inconsistent structures and colors in lips, eyes, and large areas of skin color deviations. The use of PD helps eliminate all the low-frequency inconsistencies, as can be clearly observed from the error maps.

Dataset	Method	PSNR↑	SSIM↑	FID↓
Face	Panini	27.13	0.729	14.15
	Panini w/ PD	**29.97**	**0.801**	**9.87**
	GLEAN	27.20	0.74	12.63
	GLEAN w/ PD	**30.21**	**0.81**	**8.69**
Cat	Panini	22.36	0.596	129.2
	Panini w/ PD	**23.52**	**0.623**	**118.9**
	GLEAN	22.74	0.588	62.92
	GLEAN w/ PD	**22.94**	**0.597**	**58.95**
Church	Panini	19.27	0.483	**67.98**
	Panini w/ PD	**19.80**	**0.491**	69.20
	GLEAN	19.59	0.485	24.49
	GLEAN w/ PD	**19.99**	**0.500**	**24.03**

Table 1: 8× SR on different categories. The use of PD significantly improves the PSNR, SSIM, and FID in most cases. It is worth noting that PD is parameter-free with negligible computational cost.

For 8× SR, we train Panini, "Panini w/ PD", GLEAN, and "GLEAN w/ PD" under the same training configuration for 100K iterations, with the batch size of 4 on a single Nvidia V100 GPU. For evaluation, we take 1K images from CelebA-HQ (Karras, Laine, and Aila 2019) dataset as the ground truth (GT), then use bicubic interpolation to generate the LR and yield the LR-GT testing pairs.

Likewise, we experiment 8× SR on LSUN cat and church datasets (Yu et al. 2015) for GLEAN and Panini and their PD based version. Tab. 1 shows the quantitative results of 8× SR on three categories. As can be seen, the use of PD significantly improves all metrics on both Panini and GLEAN, implying the overall enhancement of *consistency* and *realness*. Fig. 3(a) shows the convergence curve on 8× face SR. We can observe a steady improvement of PSNR and FID during training. Notably, "Panini w/ PD" in 10K iterations achieves comparable PSNR and FID to Panini in 90K iterations, showing nine times of convergence acceleration.

To intuitively understand the eliminated inconsistencies by PD, we take the GT as references and visualize the error maps in Fig. 4. Specifically, we subtract the result with GT to get the errors, then times five the absolute values and convert them into a color map. The results generated by GLEAN suffer color inconsistencies in most regions of the image. However, these inconsistencies are not a simple overall deviation

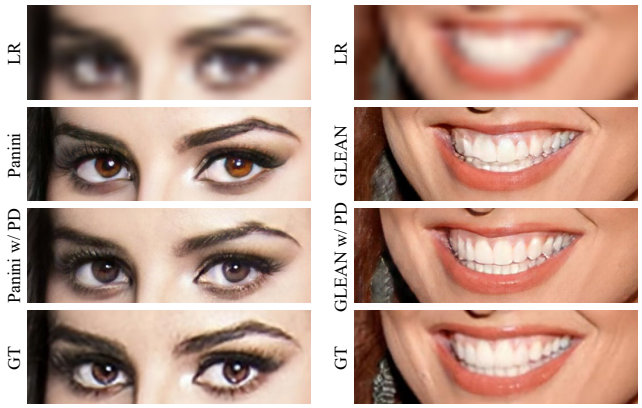


Figure 5: Qualitative results on 16× face SR. The use of PD can eliminate color deviation and reduce structural inconsistencies.

Method	PSNR↑	SSIM↑	MS-SSIM↑	FID↓
PULSE	21.68	0.676	0.596	42.71
pSp	18.91	0.680	0.526	39.88
GFPGAN	25.17	0.761	0.804	24.34
GPEN	26.07	**0.784**	0.820	31.89
Panini	27.18	0.758	0.843	**14.49**
Panini w/ PD	**27.81**	0.771	**0.851**	14.78
GLEAN	27.21	0.743	0.843	15.01
GLEAN w/ PD	27.69	0.754	0.848	15.27

Table 2: Comprehensive comparison on 16× face SR. We compare Panini and GLEAN and their PD-based versions with state-of-the-art face SR methods. The involvement of PD significantly elevates all consistency metrics, i.e., PSNR, SSIM, and MS-SSIM. We attribute the slight rise of FID to the training stochasticity. Actually, the FID is comparable during training, as can be seen in Fig. 3.

of hue but are semantically related (e.g., it changes along the facial structures) and thus can not be eliminated by simple post-processing. However, with the use of PD, these low-frequency inconsistencies are gone, leaving only the high-frequency inconsistencies, which are acceptable considering SR as an ill-posed inverse problem.

16× Super-Resolution

The experiment on 16× SR is similar to the 8× SR experiment, except for the following changes: (1) We take the pre-trained PULSE, pSp, GFPGAN, and GPEN, the state-of-the-art GAN Prior based face SR methods, for a comprehensive comparison. (2) The GAN loss weight are set as 1×10^{-3} for "Panini w/ PD" and "GLEAN w/ PD". (3) The training iteration is set as 300K.

Tab. 2 shows the quantitative comparisons on 16×SR. We can see a significant improvement in consistency metrics with the use of PD. "Panini w/ PD" gets the best score in PSNR and MS-SSIM and the second best score in SSIM and FID, achieving state-of-the-art in 16× face SR. Fig. 3(b) shows the convergence curve on 16×SR. The use of PD elevates PSNR for all the evaluated iterations and yields comparable FID. Fig. 5 shows the qualitative comparison of Panini and GLEAN and their PD based versions. Panini can generate highly realistic details that are even superior to the GT. However, like most GAN prior based methods, it suffers inconsistencies, typically the eye colors. The use of PD helps eliminate such inconsistencies while maintaining superior realistic details. GLEAN may generate unreasonable tooth shapes. However, the use of PD can help yield more reasonable structures.

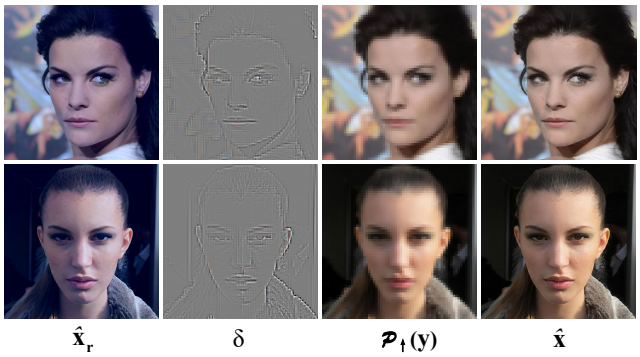


Figure 6: Visualization of PD. $\hat{\mathbf{x}}_r$ represent the raw prediction of GAN prior network, δ is the high-frequency part of $\hat{\mathbf{x}}_r$. $\mathcal{P}_\uparrow(\mathbf{y})$ denotes the low-frequency contents inherited from LR image. The final result $\hat{\mathbf{x}}$ is yielded by adding $\mathcal{P}_\uparrow(\mathbf{y})$ with δ. Zoom-in for the best view.

Ablation Studies

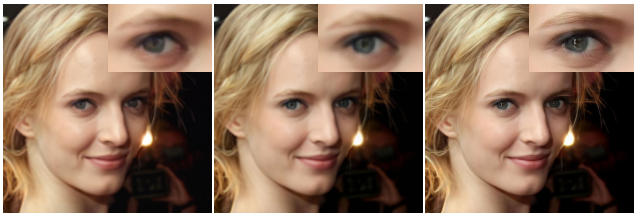
What Has the Network Learned?

With the use of PD, we take the upsampled LR, i.e., $\mathcal{P}_\uparrow(\mathbf{y})$ as the low-frequency part, then only the high-frequency part of $\hat{\mathbf{x}}_r$ that GAN prior generates is needed to constitute the final HR result $\mathbf{y}$. Since we only extract high-frequencies from $\hat{\mathbf{x}}_r$, it does not necessarily own reasonable low-frequency contents. Interestingly, $\hat{\mathbf{x}}_r$ turns out to be reasonable in most low-frequency contents. Fig. 6 visualizes $\hat{\mathbf{x}}_r$, high-frequency part δ, and low-frequency part $\mathcal{P}_\uparrow(\mathbf{y})$.

What If We Do Not Use Pixel-Wise Losses?

The use of pixel-wise losses provides a clear learning target. However, it does not generalize well when the GT or LR suffers degradation. That is because the pixel-wise losses encourage learning the inverse process of the downsampling of GT. Thus when GT already contains degradation, the SR network trained on pixel-wise losses tends to replicate such degradation.

Since PD can inherently assure *consistency*, we may train the PD-based network without pixel-wise losses (e.g., ℓ_1 and perceptual loss) and solely use adversarial losses for training. To stabilize the training, we take GLEAN as our backbone, with several critical changes: (1) we remove the "decoder" since it is proved to be redundant (Chan et al. 2022).



Panini GLEAN PDN(Ours)

Figure 7: Results on unseen downsamplings. PDN yields clearer results when facing unseen downsamplings. Here the networks are all trained on 8× bicubic(alias) and tested on 8× bicubic(antialias).

Method	Bicubic(antialias)		Bilinear(antialias)	
	PSNR↑	FID↓	PSNR↑	FID↓
Panini	29.40	30.47	29.17	37.09
GLEAN	30.13	24.26	29.76	29.50
PDN(ours)	**30.24**	**14.11**	**30.39**	**15.96**

Table 3: Quantitative comparisons on unseen downsamplings. We train PDN, Panini, and GLEAN on datasets that downsampled by 8× bicubic(alias) but test them on datasets that downsampled by 8× bicubic(antialias) and 8× bilinear(antialias) respectively. We can see that PDN achieves significantly better FID, indicating its robustness in restoring images from unseen downsamplings. Note that PDN is trained without using ℓ_1, ℓ_2, or perceptual losses.

(2) we do not predict the latent code from the "encoder" and instead set the latent code z as random to encourage diversity, as is done in StyleGAN. (3) we apply PD to the network. This revised network is denoted as PDN.

We train PDN with the same training configurations mentioned in the experiment chapter, except for the absence of ℓ_1 and perceptual loss. We find PDN works well on both 8× and 16× face SR tasks and shows superior robustness to GLEAN and Panini when facing unseen degradation. We evaluate PDN on two typical degradations.

Unseen Downsamplings. To evaluate the robustness of PDN on unseen downsampling methods, we generate training data by bicubic downsampling without antialiasing, noted as Bicubic(alias). In like manner, we generate Bicubic(antialias) and Bilinear(antialias) as testing data. We train PDN, GLEAN, and Panini on bicubic(alias) and evaluate them on Bicubic(antialias) and Bilinear(antialias). Note GLEAN and Panini are trained with ℓ_1, perceptual and adversarial loss, while PDN is only with adversarial loss. Fig. 7 shows the qualitative comparison. We can observe that PDN generates a clearer image than GLEAN and Panini. Tab. 3 shows the PSNR and FID under Bicubic(antialias) and Bilinear(antialias). PDN yields much better FID than GLEAN and Panini, which is in accord with our visual observation. Besides, evaluations under other types of downsamplers like Box and Lanczos come to a similar conclusion.

Real-World Degradation. To evaluate the robustness of PDN on unknown real-world degradation, we cherry-pick

GT Panini PDN(Ours)

Figure 8: Results on real-world degradation. We can see that GLEAN tends to replicate the degradation that GT suffers, while PDN is not affected and tends to generate clear results. Note PDN only uses 16× bicubic(antialias) to synthesize LRs for training, without any simulated degradation.

low-quality images from CelebA-HQ dataset and use bicubic interpolation to downsample them as LR images. We use PDN, GLEAN, and Panini which are trained on 16×SR. Not all three networks do not use any data augmentations to improve robustness during training. Fig. 8 shows the results. We can see that the GT suffers unknown blur and noise, and Panini's SR result reproduces these degradations. However, the result generated by PDN seems not disturbed by the degradation, showing clear and realistic details. We also tried to remove PD from PDN and train it with pure adversarial loss but got much worse results, indicating that PD can stand as a powerful *consistency* constraint to help the network converges to reasonable solutions.

PDN learns a more generalized pattern for SR, other than the inverse process of the downsampling. However, PDN is no better than state-of-the-art SR methods when dealing with LR that is downsampled from high-quality GT. We believe that is because PDN lacks detailed guidance. We hope this guide can be solved in the future without resorting to pixel-wise loss.

Conclusions and Discussions

This paper presents a novel method to eliminate inconsistencies for GAN prior based super-resolution networks. As is shown in experiments, our proposed method can be easily applied to different backbones, accelerating their training convergence and yielding better *consistency*. Our method also shows potential in dealing with unseen downsamplings or real-world degradation.

There are a few limitations of our method. Firstly, our method is under an ideal assumption that the LR is clean. However, if the degradation that LR suffers damage the low-frequency contents, PD will pass the damaged low-frequencies directly to the SR result. A simple solution is to pre-process the degraded LR into a clean one. Secondly, PD may result in evident block artifacts at the early stages of training, but as the training goes on, those artifacts become imperceptible (see if you can observe block artifacts in Fig. 5). Thirdly, PD is sensitive to hyperparameter selection. See Appendix B for analysis.

References

Abdal, R.; Qin, Y.; and Wonka, P. 2019. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4432–4441.

Abdal, R.; Qin, Y.; and Wonka, P. 2020. Image2StyleGAN++: How to Edit the Embedded Images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8296–8305.

Bahat, Y.; and Michaeli, T. 2020. Explorable super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2716–2725.

Chan, K. C.; Wang, X.; Xu, X.; Gu, J.; and Loy, C. C. 2021. GLEAN: Generative Latent Bank for Large-Factor Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14245–14254.

Chan, K. C.; Xu, X.; Wang, X.; Gu, J.; and Loy, C. C. 2022. GLEAN: Generative Latent Bank for Image Super-Resolution and Beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–14.

Chen, D.; and Davies, M. E. 2020. Deep decomposition learning for inverse imaging problems. In *European Conference on Computer Vision*, 510–526. Springer.

Chong Mou, J. Z., Qian Wang. 2022. Deep Generalized Unfolding Networks for Image Restoration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chong Mou, Z. W., Jian Zhang. 2021. Dynamic Attentive Graph Learning for Image Restoration. In *International Conference on Computer Vision (ICCV)*.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2015. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 295–307.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. *Advances in Neural Information Processing Systems (NeurIPS)*.

Gu, J.; Shen, Y.; and Zhou, B. 2020. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3012–3021.

Huang, H.; He, R.; Sun, Z.; and Tan, T. 2017. Wavelet-SRNet: A Wavelet-Based CNN for Multi-Scale Face Super Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1689–1697.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 694–711.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34: 852–863.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 (CVPR)*, 8110–8119.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1646–1654.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; 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.

Li, X.; Liu, M.; Ye, Y.; Zuo, W.; Lin, L.; and Yang, R. 2018. Learning Warped Guidance for Blind Face Restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 272–289.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1833–1844.

Lin, J.; Zhou, T.; and Chen, Z. 2018. Multi-Scale Face Restoration With Sequential Gating Ensemble Network. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 32(1).

Lin, S.; Zhang, J.; Pan, J.; Liu, Y.; Wang, Y.; Chen, J.; and Ren, J. 2020. Learning to Deblur Face Images via Sketch Synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 34(07): 11523–11530.

Luo, F.; and Wu, X. 2020. Maximum a Posteriori on a Submanifold: a General Image Restoration Method with GAN. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7.

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 (CVPR)*, 2437–2445.

Mou, C.; Wu, Y.; Wang, X.; Dong, C.; Zhang, J.; and Shan, Y. 2022. Metric Learning based Interactive Modulation for Real-World Super-Resolution. In *European Conference on Computer Vision (ECCV)*.

Nitzan, Y.; Bermano, A.; Li, Y.; and Cohen-Or, D. 2020. Face identity disentanglement via latent space mapping. *ACM Transactions on Graphics*, 39: 1–14.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2085–2094.

Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in Style: A StyleGAN Encoder for Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2287–2296.

Schwab, J.; Antholzer, S.; and Haltmeier, M. 2019. Deep null space learning for inverse problems: convergence analysis and rates. *Inverse Problems*, 35(2): 025008.

Tewari, A.; Elgharib, M.; Bharaj, G.; Bernard, F.; Seidel, H.-P.; Pérez, P.; Zollhofer, M.; and Theobalt, C. 2020. StyleRig: Rigging StyleGAN for 3D Control Over Portrait Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6142–6151.

Wan, Z.; Zhang, B.; Chen, D.; Zhang, P.; Chen, D.; Liao, J.; and Wen, F. 2020. Bringing Old Photos Back to Life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2747–2757.

Wang, X.; Li, Y.; Zhang, H.; and Shan, Y. 2021. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9168–9178.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Loy, C. C. 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 63–79.

Wang, Y.; Hu, Y.; and Zhang, J. 2022. Panini-Net: GAN Prior Based Degradation-Aware Feature Interpolation for Face Restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Xia, W.; Zhang, Y.; Yang, Y.; Xue, J.-H.; Zhou, B.; and Yang, M.-H. 2022. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, J.; Wright, J.; Huang, T. S.; and Ma, Y. 2010. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11): 2861–2873.

Yang, L.; Wang, S.; Ma, S.; Gao, W.; Liu, C.; Wang, P.; and Ren, P. 2020. HiFaceGAN: Face Renovation via Collaborative Suppression and Replenishment. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 1551–1560.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2021. GAN Prior Embedded Network for Blind Face Restoration in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 672–681.

Yeh, R. A.; Lim, T. Y.; Chen, C.; Schwing, A. G.; Hasegawa-Johnson, M.; and Do, M. N. 2018. Image Restoration with Deep Generative Models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6772–6776.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Yu, X.; and Porikli, F. 2017. Hallucinating Very Low-Resolution Unaligned and Noisy Face Images by Transformative Discriminative Autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3760–3768.

Zhang, M.; and Ling, Q. 2021. Supervised Pixel-Wise GAN for Face Super-Resolution. *IEEE Transactions on Multimedia*, 23: 1938–1950.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018a. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 286–301.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2018b. Residual Dense Network for Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2472–2481.

Zhu, P.; Abdal, R.; Qin, Y.; and Wonka, P. 2020. SEAN: Image Synthesis With Semantic Region-Adaptive Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5104–5113.