6-2022

Faithful extreme rescaling via generative prior reciprocated invertible representations

Zhixuan ZHONG

Liangyu CHAI

Yang ZHOU

Bailin DENG

Jia PAN

See next page for additional authors

Citation

ZHONG, Zhixuan; CHAI, Liangyu; ZHOU, Yang; DENG, Bailin; PAN, Jia; and HE, Shengfeng. Faithful extreme rescaling via generative prior reciprocated invertible representations. (2022). *Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5698-5707.

Available at: https://ink.library.smu.edu.sg/sis_research/8444

Author

Zhixuan ZHONG, Liangyu CHAI, Yang ZHOU, Bailin DENG, Jia PAN, and Shengfeng HE

Faithful Extreme Rescaling via Generative Prior Reciprocated Invertible Representations

Zhixuan Zhong[1] Liangyu Chai[1] Yang Zhou[1] Bailin Deng[2] Jia Pan[3] Shengfeng He[1]*
[1] School of Computer Science and Engineering, South China University of Technology
[2] School of Computer Science and Informatics, Cardiff University
[3] Department of Computer Science, The University of Hong Kong

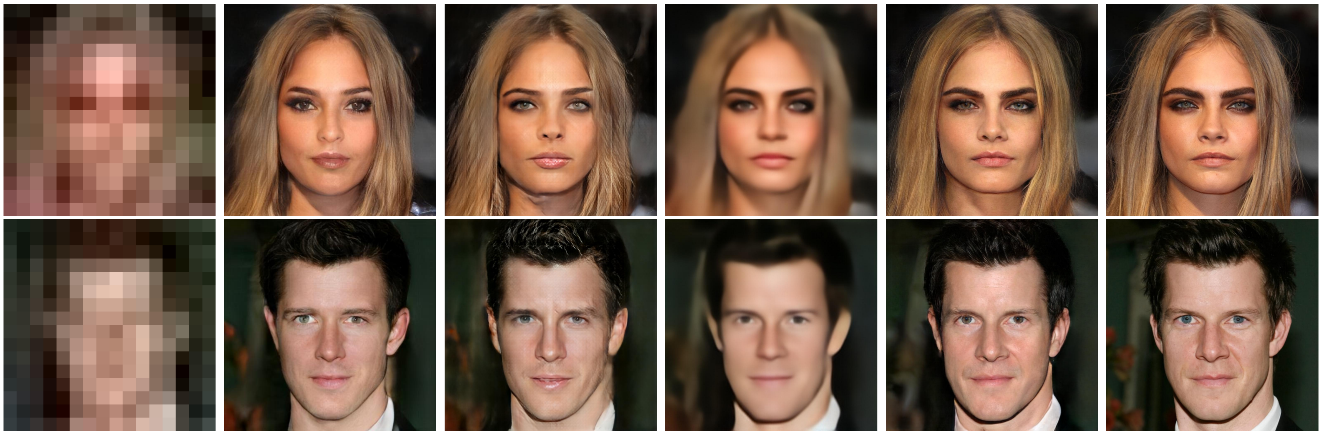

(a) Input LR (b) GPEN [30] (c) GLEAN [3] (d) IRN [28] (e) Ours (f) GT

Figure 1. We propose a faithful rescaling method that enables pixel-accurate upscaling with an extreme factor (64×). Previous generative prior approaches (b) & (c), or invertible approach (d) suffer from the mapping ambiguity between the extreme low-resolution input and GT. We embed high-resolution information both into the invertible low-resolution output as well as the generative prior, reciprocally enhancing upscaling results.

Abstract

*This paper presents a **G**enerative prior **R**eciprocAted **I**nvertible rescaling **N**etwork (GRAIN) for generating faithful high-resolution (HR) images from low-resolution (LR) invertible images with an extreme upscaling factor (64×). Previous researches have leveraged the prior knowledge of a pretrained GAN model to generate high-quality upscaling results. However, they fail to produce pixel-accurate results due to the highly ambiguous extreme mapping process. We remedy this problem by introducing a reciprocated invertible image rescaling process, in which high-resolution information can be delicately embedded into an invertible low-resolution image and generative prior for a faithful HR reconstruction. In particular, the invertible LR features not only carry significant HR semantics, but also are trained to predict scale-specific latent codes, yielding a preferable utilization of generative features. On the other hand, the enhanced generative prior is re-injected to the rescaling process, compensating the lost details of the invertible rescaling. Our reciprocal mechanism perfectly integrates the advantages of invertible encoding and generative prior, leading to the first feasible extreme rescaling solution. Extensive experiments demonstrate superior performance against state-of-the-art upscaling methods. Code is available at https://github.com/cszzx/GRAIN.*

1. Introduction

Due to the explosive growth of image data, image downscaling is a typical way for fast data processing and efficient storage. Therefore, the capability of rescaling a low-resolution image to a high-resolution one is of great importance for many multimedia applications. However, due to the highly underspecified mapping process from low-resolution (LR) to high-resolution (HR), different image priors have to be introduced to reduce the learning ambiguity, especially for an extreme upscaling setting (*e.g.*, 64×).

*Corresponding author (hesfe@scut.edu.cn).

Large-scale *data prior* is typically used in deep learning based super-resolution to mimic the transformation from LR to HR [6,8,13]. Due to the dominated pixel-wise constraints, these methods can smoothly recover the overall structure. For the same reason, small scale details cannot be "synthesized" from scratch and therefore these methods are typically limited to 8× upscaling.

Recent advances in generative adversarial networks (GANs) demonstrate that the *generative prior* [3,22,30,35] derived from a pretrained GAN model provides rich and diverse supplementary representations for the extreme upscaling process. The basic principle is that a small LR image can be mapped to the input latent code of a pretrained GAN model, such that a perceptually similar output can be produced by the generator. Although the inherent knowledge of a pretrained GAN enables a plausible extreme upscaling (64×), the vague LR input prevents them from locating a perfect latent code and therefore cannot preserve the original characteristics (see Fig. 1b and Fig. 1c).

The above issue raises a question: can we enrich the informativeness of an extremely downscaled image to better incorporate with the generative prior? To answer this question, we turn to the alternative *invertible prior*. Previous invertible rescaling [17,28] embeds the high-resolution input image into inconspicuous and reconstructible patterns appended on the LR image, such that it can be easily rescaled to the original resolution. In the scenario of extreme upscaling, however, the LR image (16×16) is too small to embed sufficient hints for recovery (see Fig. 1d). To activate different upscaling priors under extreme scenarios, we propose in this paper a **G**enerative prior **R**eciproc**A**ted **I**nvertible rescaling **N**etwork (**GRAIN**), which maximizes the potential of invertible and generative priors. In particular, it consists of three reciprocal components, *i.e.*, an invertible extreme rescaling module, a scale-specific generative prior module, and an upscaling priors decoding module. The rescaling module is trained for two purposes. On one hand, it embeds the HR information to an extreme LR form that can be reverted back to the original. On the other hand, the invertible features are optimized to produce scale-specific latent codes of the pretrained GAN model. In this way, the HR information is maximally entangled with the rescaling process in image-level as well as the generative latent space. The enhanced generative features are then reciprocated back to the invertible representations for decoding the final upscaling result. Extensive experiments demonstrate faithful and superior upscaling performance against state-of-the-art upscaling methods using different types of image priors. Our proposed method is the first feasible extreme rescaling solution that can be beneficial for storing and transferring data due to our perceivable yet invertible extremely downscaled LR results (Fig. 1e). We show that our method can break through the limitation of the pretrained GAN data distribution and recover outlier inputs, while being also applicable to other domains by switching to different pretrained models.

Our contributions can be summarized as follows:

- We propose the first invertible extreme rescaling framework that allows perceivable downscaling and faithful upscaling with an extreme scaling factor (64×).
- We design a reciprocal strategy that elegantly connects the invertible prior with the generative prior. It maximizes the advantages of these two priors, largely mitigating the ill-posed nature of the image upscaling process.
- The proposed model sets new state-of-the-art in extreme image rescaling. We also explore and analyze the image invertibility with respect to different influencing factors like network structure, scaling factor, and data domain.

2. Related Work

2.1. Image Super-Resolution

CNN-based Methods. Super-Resolution (SR) aims to reconstruct a realistic HR image from its LR version. Early works [8, 11, 18, 20, 25, 29, 34] learn a direct mapping between LR and HR, which perform well for small upscaling factors (2× or 4×) but produce blurry results when the upscaling factor increases. WSRNet [12] proposes a wavelet-based CNN approach with additional wavelet coefficients prediction that can handle a scaling factor of up to 16×. RFB-ESRGAN [24], based on ESRGAN [26], employs receptive field blocks to achieve 16× SR. A main application of SR is face restoration, and some works utilize the extra facial priors to improve the quality of recovery. For example, [4] introduces a parsing map and facial landmarks as prior knowledge, and [19] designs an attribute-guided face transfer and enhancement network.

GAN Inversion Methods. The purpose of GAN inversion is to search for a latent code in the GAN latent space which can generate the closest result to a given input image. Thanks to the powerful generative priors inherent in large-scale GAN models, SR methods based on GAN inversion are able to achieve extreme upscaling (32× or 64×), while maintaining a good fidelity. For example, PULSE [22] and its extension Pro-PULSE [35] iteratively optimize the latent code of StyleGAN [15, 16] to narrow the gap between input and output. pSp [23] introduces a ResNet encoder to extract pyramidal features which are further mapped to the latent space. GPEN [30] builds a U-Net framework and finetunes the StyleGAN for blind face restoration. GLEAN [3] proposes an encoder-bank-decoder architecture with multi-resolution skip connections to upscale images from various categories. Although GAN inversion can recover a high-quality result, the characteristics of the original input cannot be well preserved due to the limited details of LR input.

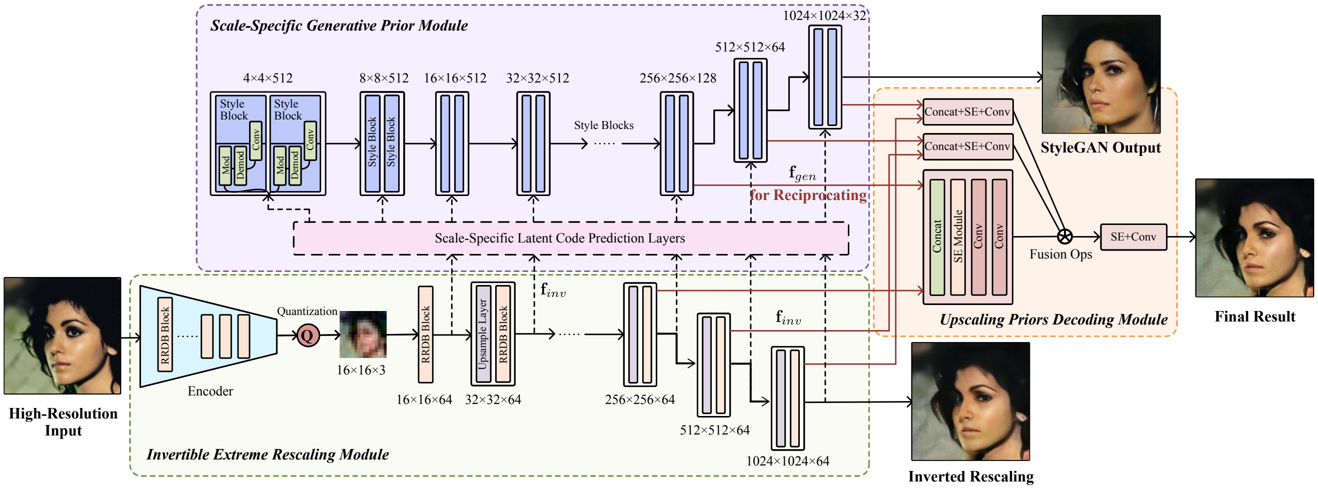


Figure 2. Overview of the GRAIN framework. GRAIN is composed of three modules, invertible extreme rescaling module, scale-specific generative prior module and upscaling priors decoding module. With a pre-trained StyleGAN capturing the image prior, the invertible encoder-decoder and reciprocal design learns both recovery and naturalness of images.

2.2. Invertible Image Restoration

Different from ordinary image restoration works that directly reconstructs an image from its degraded version, invertible image restoration aims to embed significant information into the degraded image to aid the process of restoration. Invertible grayscaling [9, 27] proposes a encoder-decoder framework to embed the original color information into a synthesized grayscale image which is further used to recover its corresponding color image. Zhu *et al*. [36] embeds the spatio-temporal information of a video into a "live" image, which can be converted back to a video preview. Recently, Cheng *et al*. [5] proposes an Invertible Image Conversion Net (IICNet) for various tasks including spatio-temporal video embedding, dual-view images embedding, and mononizing binocular images embedding. As for the SR category, TAR [17] proposes an auto-encoder-based framework that enables joint learning of downscaling and upscaling to maximize the restoration performance. IRN [28] captures the distribution of the lost information during downscaling using a latent variable and then upscales by inversely passing a randomly-drawn latent variable with the low-resolution image. In the extreme setting, however, the LR image is too small to embed sufficient details, leading to over-smoothed reconstructions. We effectively resolve this problem by introducing generative priors in the loop.

3. Approach

3.1. Overview of GRAIN

We describe the GRAIN framework in this section. Our main aim is to generate high-resolution images from extreme low-resolution ones. Given a high-resolution image $\mathbf{y}$, and compressed to a low-resolution (16×16) image as input, GRAIN estimates a high-resolution image $\hat{\mathbf{y}}$ that is as similar as possible to the ground-truth image $\mathbf{y}$ in terms of pixel-level accuracy and visual perception.

The overall framework of GRAIN is depicted in Fig. 2. GRAIN is composed of three complementary modules: an invertible extreme rescaling module (encoder-decoder), a scale-specific generative prior module (code prediction layers and a pretrained StyleGAN [15, 16]), and an upscaling priors decoding module. Specifically, the invertible extreme rescaling module is designed to encode an HR ground-truth image $\mathbf{y}$ as an extreme LR image, and decode it to restore an HR image $\hat{\mathbf{y}}_{inv}$ that is as close to $\mathbf{y}$ as possible. The features produced by the decoder provides an invertible representation $\mathbf{f}_{inv}$ of the ground-truth HR image. Then in the scale-specific generative prior module, $\mathbf{f}_{inv}$ is mapped to the corresponding latent codes for modulating the StyleGAN features $\mathbf{f}_{gen}$. After that, the upscaling priors decoding module can reciprocate the enhanced generative features $\mathbf{f}_{gen}$ back to the invertible features $\mathbf{f}_{inv}$ for decoding the final faithful and realistic HR results.

3.2. Invertible Extreme Rescaling Module

Most downscaling methods are intended to save the data transfer cost while maintaining moderate image qualities. However, the downscaling process makes super-resolution (SR) highly ill-posed and causes a blurry high-resolution restoration result. To reduce the loss of information when downscaling, we adopt a encoder-decoder scheme, invertible rescaling, to find a semantically reasonable LR image that also benefits the HR restoration performance. Our Invertible Extreme Rescaling Module consists of three parts:

(1) **An encoder** that embeds necessary information into the LR image (16×16) to help restore the HR image. We use an RRDB-Net [26] to extract features and embed them into the RGB image space.

(2) **A quantization operation** is unavoidable when saving the embedding PNG-format image for end-to-end training. We employ the approach in [2] to add uniform noise when training, and perform integer rounding to quantize the embedding image during inference. The quantized image is clamped between 0 and 255.

(3) **A decoder** which is an upscaling network, composed of a stack of RRDB blocks [26] and upsampling layers, to generate an HR image $\hat{\mathbf{y}}_{inv}$. Each block produces a feature vector $\mathbf{f}^{i}_{inv}$ (which is of dimension 16×16×64, 32×32×64, …, 1024×1024×64 respectively for different blocks) that represents the image feature for a particular scale. The vectors are combined into the invertible features $\mathbf{f}_{inv}$ as a multi-scale representation that can be decoded for follow-up generation.

3.3. Scale-specific Generative Prior Module

The scale-specific generative prior module maps the invertible features $\mathbf{f}_{inv}$ to the StyleGAN $\mathcal{W}+$ latent space, and utilizes the pretrained StyleGAN to generate faithful and realistic HR results. StyleGAN has a strong representation ability to generate lifelike images from a code in its latent space. Inspired by [23], we propose the scale-specific latent code prediction layers to embed the hierarchical invertible features $\mathbf{f}_{inv}$ into 18 different 512-dimensional latent vectors, which form a latent code in $\mathcal{W}+$. These latent vectors are fed into the generator according to their scales to produce generative features $\mathbf{f}_{gen}$ and an HR image $\hat{\mathbf{y}}_{style}$.

It has been noted that the input latent vectors to StyleGAN correspond to different levels of details in the output image [23]. To better utilize such relations, our code prediction layers generate each latent vector from an invertible feature vector $\mathbf{f}^{i}_{inv}$ with a corresponding resolution. Following the StyleGAN, the 18 latent vectors can be divided into nine pairs, with each pair corresponding to a resolution 4×4, 8×8, 16×16, …, 1024×1024, respectively. As shown in Fig. 3, for a latent vector pair corresponding to a resolution of 16×16 and upwards, we take the invertible feature $\mathbf{f}^{i}_{inv}$ of the matching resolution, aggregate it with a downsampled invertible feature from a higher resolution to incorporate more semantics information, and then downsample the feature to 16×16×64 through a series of convolution operations. The 16×16×64 feature is then fed into two branches of convolutional layers to obtain a pair of latent vectors. For the latent vector pairs corresponding to 4×4 and 8×8 resolutions, there is no invertible feature of matching resolutions, and we simply reuse the input features for the 16×16 latent vector branches to predict the 4×4 and 8×8 ones. Finally, the 18 latent vectors form a style latent code that is sent to StyleGAN to further generate $\mathbf{f}_{gen}$ and $\hat{\mathbf{y}}_{style}$.

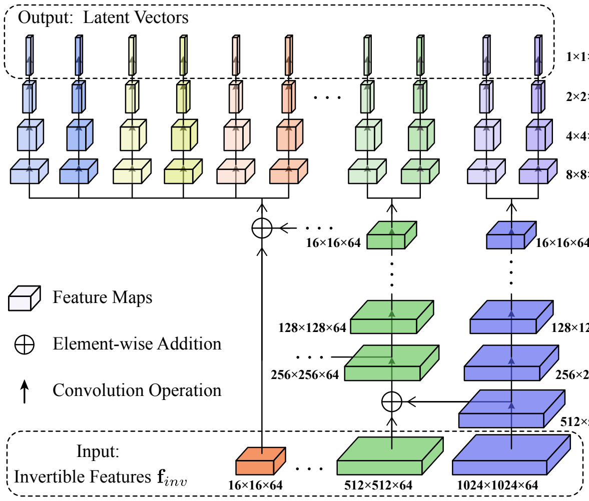


Figure 3. The details of scale-specific latent code prediction layers for mapping the invertible features to style vectors which form a style latent code for further generation in StyleGAN.

3.4. Upscaling Priors Decoding Module

The upscaling priors decoding module is introduced to reciprocate the StyleGAN prior features $\mathbf{f}_{gen}$ to combine with the invertible features $\mathbf{f}_{inv}$ to generate a faithful and realistic final result. As shown in Fig. 2, the module takes three pairs of corresponding features from $\mathbf{f}_{gen}$ and $\mathbf{f}_{inv}$, at the resolutions of 256×256, 512×512 and 1024×1024 respectively, for integrating and decoding. For each resolution, the pair of features are first concatenated and fed to a separate Squeeze-and-Excitation (SE) Module [10] and several convolutional layers with activation. Then we upsample all of them to 1024×1024 resolution. Finally, the resulting features for the three resolutions are concatenated again and fed to another SE module followed by convolutional layers, to obtain the final HR output image at 1024×1024 resolution.

3.5. Training Objectives and Strategy

We adopt a multi-stage strategy to stabilize the training of GRAIN. First in *Stage 1*, we train the invertible extreme rescaling module alone for several epoches to generate $\mathbf{f}_{inv}$ and $\hat{\mathbf{y}}_{inv}$. Next in *Stage 2*, we send $\mathbf{f}_{inv}$ to the scale-specific generative prior module and train it alone to generate $\mathbf{f}_{gen}$ and $\hat{\mathbf{y}}_{style}$. *Stage 3* takes both $\mathbf{f}_{inv}$ and $\mathbf{f}_{gen}$ as input, and the upscaling priors decoding module is trained alone to generate $\hat{\mathbf{y}}_{final}$. In the *Final Stage*, all three modules are trained together in an end-to-end manner. The pretrained StyleGAN model is fixed all the time. The loss functions used in different training stages share a common component $\mathcal{L}_{base}$, which is a weighted sum of the following terms:

(1) A pixel-wise reconstruction $\mathcal{L}_2$ loss for the ℓ_2 distance

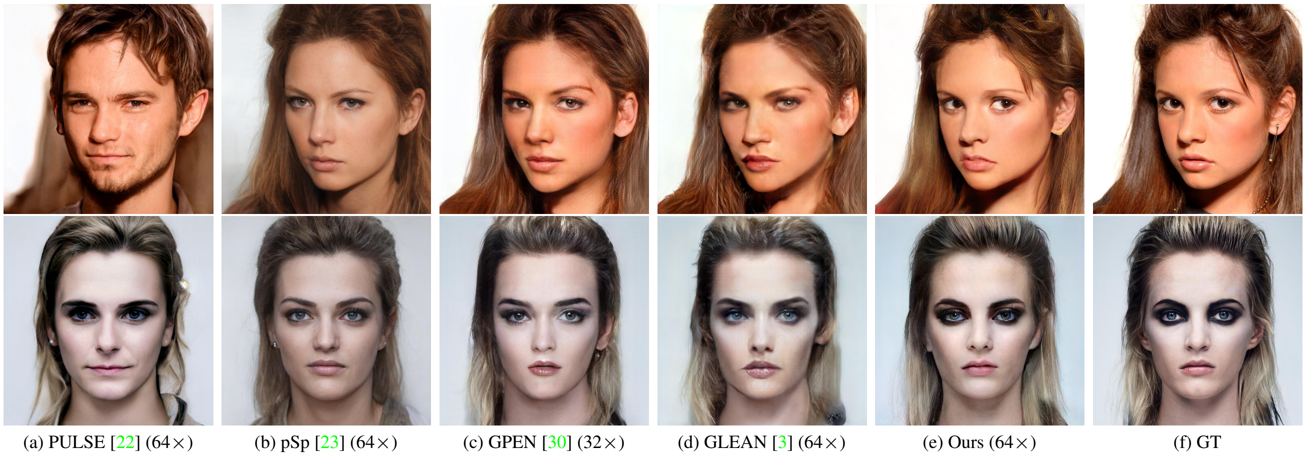


Figure 4. Qualitative comparisons with GAN inversion based methods. Our method outperforms the others in fidelity and details. (Zoom in for better view.)

between the generated image $\hat{\mathbf{y}}$ and the ground truth $\mathbf{y}$:

$$\mathcal{L}_2 = \|\mathbf{y} - \hat{\mathbf{y}}\|_2. \tag{1}$$

Here $\hat{\mathbf{y}}$ represents $\hat{\mathbf{y}}_{inv}$ for *Stage 1*, $\hat{\mathbf{y}}_{style}$ for *Stage 2*, and $\hat{\mathbf{y}}_{final}$ for both *Stage 3* and the *Final Stage*.

(2) The LPIPS loss $\mathcal{L}_{LPIPS}$ from [32] to learn perceptual similarities.

$$\mathcal{L}_{LPIPS} = \|F(\mathbf{y}) - F(\hat{\mathbf{y}})\|_2. \tag{2}$$

where F denotes a fixed perceptual feature extractor.

(3) For human face images, an identity loss that requires the generated image to have the same identity as the ground-truth image:

$$\mathcal{L}_{id} = 1 - Cos(R(\mathbf{y}), R(\hat{\mathbf{y}})), \tag{3}$$

where R denotes the face recognition feature produced by the pre-trained ArcFace network [7], and $Cos(\cdot, \cdot)$ denotes the cosine similarity.

In summary, the base loss function is defined as

$$\mathcal{L}_{base} = \lambda_1 \mathcal{L}_2 + \lambda_2 \mathcal{L}_{LPIPS} + \lambda_3 \mathcal{L}_{id}, \tag{4}$$

where λ_1, λ_2, λ_3 are the weights. In addition, in *Stage* 1, the loss function also includes the Relativistic GAN loss [14] for injecting fine details. For other stages, the loss function also includes the adversarial WGAN loss [1], which adopts the pre-trained StyleGAN discriminator to guide the generation of $\hat{\mathbf{y}}_{style}$ and $\hat{\mathbf{y}}_{final}$ for better image quality and faster convergence. In order to produce semantically reasonable LR images, a $\mathcal{L}_2$ loss between the generated LR image and ground-truth LR image is used in *Stage 1* and *Final Stage*.

4. Experiments

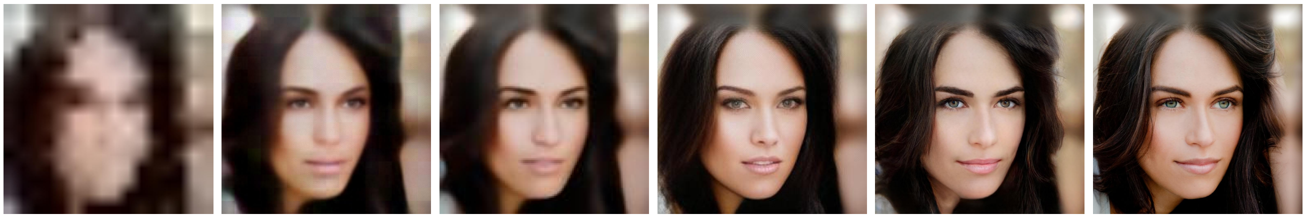
We adopt the pre-trained StyleGANv2 model [16] to produce generative prior, and train existing state-of-the-art methods based on the publicly available codes in the same training settings for fair comparison. All experiments are run on a PC with an Nvidia GeForce RTX 3090 GPU.

Datasets. For face super-resolution tasks, we use the CelebA-HQ dataset [16] which contains 30,000 human face images of resolution 1024×1024. We follow the original splitting, with 24,183 images used for training and 5,817 images for testing. To evaluate the generalization capacity in various domains of GRAIN, we also train our model on the Cat dataset [33] and the LSUN-Church dataset [31], both of which are of resolution 256×256. As for the input images, we resize these datasets to the resolution of 16×16 with bilinear downsampling, while invertible methods utilize primary images to get the 16×16 resolution images.

Evaluation Metrics. For quantitative evaluation, we adopt the widely-used pixel-wise metrics, PSNR and SSIM. We also employ the LPIPS metric [32] to measure the perceptual distance. It should be noted that LPIPS is resolution-dependent. Thus we upscale all results to the ground-truth resolution of 1024×1024 with bilinear interpolation especially for the methods that are unable to generate 1024×1024 images when evaluating the LPIPS metirc.

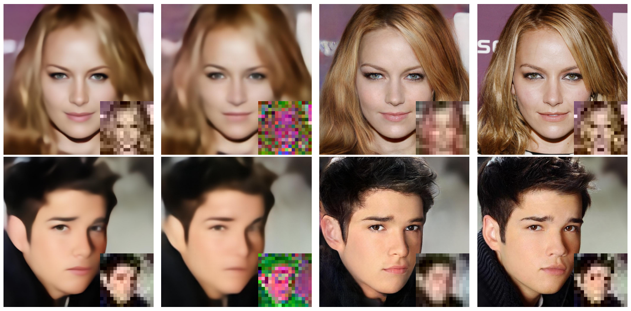
4.1. Comparisons with State-of-the-art Methods

We compare our GRAIN framework with several state-of-art methods, including GAN inversion based face upscaling methods (PULSE [22], pSp [23], GPEN [30], and GLEAN [3]), CNN-based face super-resolution methods (WSRNet [12], DIC [21], and ESRGAN [26]), and invertible image restoration methods (TAR [17] and IRN [28]). We train all these methods with their maximal upscaling factor except for GPEN [30] (for which we fine-tune its public 512×512 pretrained model due to the lack of released training code) and IRN [28] (which is out of memory when upscaling to the resolution of 1024×1024).



(a) Bilinear(64×) (b) WSRNet [12] (16×) (c) DIC [21] (8×) (d) ESRGAN [26] (16×) (e) Ours (64×) (f) GT

Figure 5. Qualitative comparisons with CNN-based SR methods. They cannot produce high-resolution upscaling results. (Zoom in for better view.)



(a) TAR (64×) (b) IRN (32×) (c) Ours (64×) (d) GT

Figure 6. Qualitative comparisons with invertible image restoration methods. The bottom-right image is the invertible LR version.

Table 1. Quantitative comparisons with state-of-the-art methods.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
PULSE [22] (64×)	19.20	0.5515	0.4867
pSp [23] (64×)	17.70	0.5900	0.4456
GLEAN [3] (64×)	20.24	0.6354	0.3891
GPEN [30] (32×)	20.40	0.5919	0.3714
Bilinear (64×)	19.92	0.6840	0.6027
WSRNet [12] (16×)	22.91	0.6201	0.5432
DIC [21] (8×)	**25.55**	**0.7574**	0.5526
ESRGAN [26] (16×)	21.01	0.5959	0.4464
TAR [17] (64×)	25.15	0.7397	0.4733
IRN [28] (32×)	24.41	0.6943	0.5238
Ours (64×)	22.30	0.6467	**0.2686**

GAN Inversion Methods. Fig. 4 compares our method with the GAN inversion based methods. Guided by the ordinary LR input image, these methods are unable to generate plausible details and maintain a good fidelity. In particular, PULSE and pSp restore a face image with a different identity, and the results of GPEN and GLEAN have slight improvements in the quality but still with notable differences in the hair, eyebrows, eyeballs and facial expressions. This is because PULSE and pSp only search for the corresponding latent code that generates a result similar to the ground-truth. Although GPEN and GLEAN try to modify the GAN prior by fine-tuning StyleGAN and training a feature-connected decoder respectively, they are still restricted by the input with deficient information. With the integration of invertible prior and generative prior, our method succeeds in both verisimilitude and lifelikeness, including lively facial expressions, flying hair and realistic eyes in Fig. 4e.

CNN-based Methods. Fig. 5 presents a visual comparison with CNN-based SR methods. As discussed in Sec. 2.1, these methods have limited ability in conducting 64× SR and they produce results that are over-smoothed and lack details. In comparison, our method can generate faithful and realistic images with plausible details, such as correct pose, clear eyebrow and closed lips in Fig. 5e.

Invertible Image Restoration Methods. The invertible methods have a similar setting to our task, *i.e.*, they produce an embedded LR image in addition. Results are illustrated in Fig. 6. We can observe that these methods fail to generate legible HR images with faithful facial details. On the other hand, besides producing realistic HR results, our method is able to output a perceivable LR image, compared to IRN (bottom-right corner in Fig. 6c and Fig. 6b). As our invertible network embeds both information for rescaling and producing generative priors, the LR images are not as faithful as the ones produced by TAR; nevertheless, they still capture the main perceptual features in such an extreme low resolution.

Quantitative Scores. Table 1 presents a quantitative evaluation measuring different methods. It is not surprising that CNN-based methods report the best PSNR and SSIM, as they tend to produce blurry results that follow the overall structure. Our method achieves the lowest LPIPS value, which indicates that its result is perceptually similar with the ground-truth. It is understandable that our method is unable to achieve the best PSNR and SSIM, since they focus on image pixel-wise distance instead of a good fidelity.

4.2. Ablation Studies

Direct Invertible Prior Output. As shown in Fig. 7b, when we only utilize the invertible extreme rescaling module without generative prior, the restored images maintain the face structure well but lack details, and it is easy to spot the artifacts when zooming in. This is the drawback of CNN-based super-resolution methods without GAN prior

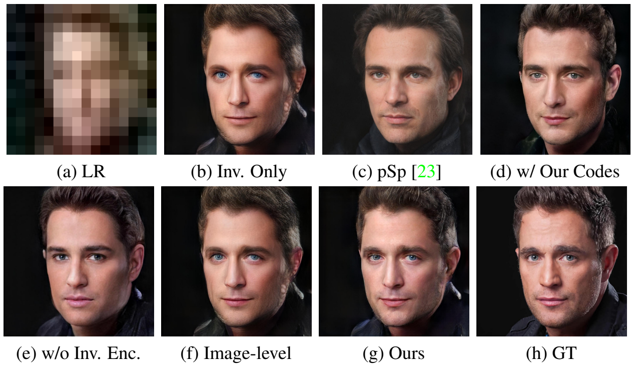


Figure 7. Qualitative comparisons of different variants of our method with a 64× upscaling factor.

Table 2. Quantitative comparisons of different variants of our method with a 64× upscaling factor.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
Invertible Only	21.93	0.5807	0.3379
pSp [23]	17.70	0.5900	0.4456
w/ Our Codes	18.33	0.5962	0.3591
w/o Invertible Encoder	19.99	0.6214	0.3299
Image-level Fusion	22.17	0.6389	0.2843
Ours	**22.30**	**0.6467**	**0.2686**

especially for large upscaling factors as discussed above.

Scale-specific Generative Prior Output. We compare our scale-specific latent code prediction layers with the pSp encoder module [23], collaborating with a pretrained StyleGAN which provides rich and diverse prior. We can observer in Fig. 7c that pSp can generate an HR image with a similar face profile, but the identity is significantly different from the ground-truth. This is because that the deficient input information is insufficient for guiding the encoder to produce a result with desired details. On the other hand, with one-to-one corresponding feature mapping, our code prediction layers can generate a faithful and superior result, *e.g.*, Fig. 7d can better capture facial expressions and restore realistic details.

Effects of Invertible Encoder. We discard the invertible encoder and apply ordinary LR images as input to investigate the contribution of information embedding mechanism. A performance drop is observed in Fig. 7e compared with Fig. 7g, especially for the eyes, mustache and hair. This is because the restoration capacity is limited after removing the encoder with embedding mechanism.

Effects of Upscaling Priors Decoding Module. To investigate the effects of the upscaling priors decoding module, we replace it with a simple image-level fusion module which concatenates the final invertible output and the GAN prior output and further sends them to a stack of convolutional layers. The result tends to copy the output of invertible prior directly while the image-level generative prior constrained in StyleGAN space has less similarity with ground-truth as shown in Fig. 7f. With the multi-scale intermediate features from both invertible and generative priors, the upscaling decoding module is able to extract and integrate rich information in a coarse-to-fine manner, further generating plausible details to enhance the output quality.

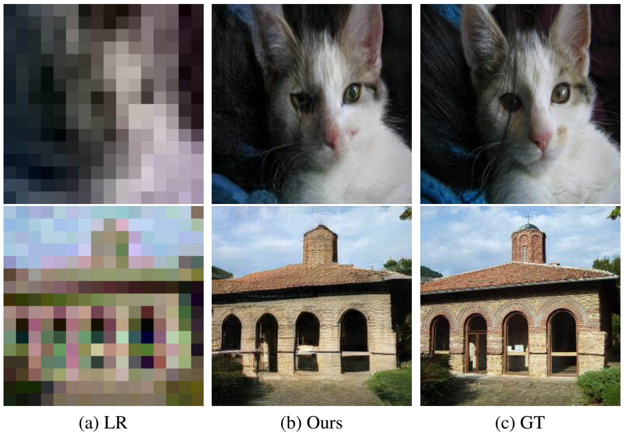


Figure 8. Results of our method on cat [33] domain and church [31] domain with a 16× upscaling factor.

Table 3. Quantitative comparisons on different domains.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
Cat [33] (16×)	21.97	0.5250	0.1945
Church [31] (16×)	18.48	0.4421	0.1860

Quantitative Scores. Quantitative ablation results are presented in Table 2. Our final setting receives the best performance in all three metrics than its variants.

4.3. Discussions and Analysis

Images in Different Domains. We demonstrate the generalization capacity of our network to reconstruct images from different domains by switching StyleGANs trained on various categories as shown in Fig. 8 and Table 3. This is due to the carefully designed invertible network which produces faithful prior of diverse data domains and further guides the StyleGAN to generate lifelike prior accurately, helping to achieve the final realistic results.

Invertibility of Different Scaling Factors. We extend our method using different LR resolutions to investigate the invertibility of diverse scaling factors, *i.e.*, our method produces a LR image of a resolution different from 16×16 and then upscales it to 1024×1024. The results are presented in Fig. 9, where we conduct experiments with LR resolutions of 16×16, 32×32, 64×64 and 128×128, respectively. It shows that an LR image of higher resolution can embed more information and lead to more a faithful reconstructed

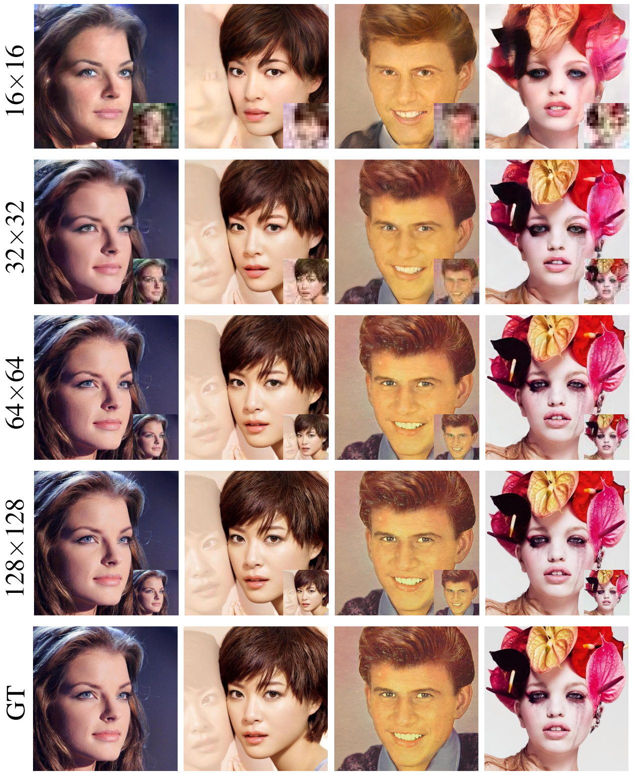


Figure 9. Invertibility of different scaling factors. The left annotation is the resolution of LR image and the LR image is depicted in bottom-right.

Table 4. Quantitative comparisons on different LR resolution sizes.

LR	PSNR ↑	SSIM ↑	LPIPS ↓
16 × 16	22.30	0.6467	0.2686
32 × 32	25.60	0.6932	0.1828
64 × 64	28.13	0.7451	0.1031
128 × 128	32.60	0.8757	0.0719

image. For example, the 64×64 variant is able to restore an authentic background with clear textures which is out of face domain, which indicates the powerful invertibility of our method. Quantitative results are shown in Table 4.

Comparison with JPEG Compression. JPEG is a widely used image compression standard with a controllable image quality range from 1 to 100 (where a higher value indicates better quality). Fig. 10 and Table 5 show a comparison in compression performance (in terms of the average PSNR and compressed storage size over the CelebA-HQ test set) between our method (with different LR resolutions) and JPEG (with different image quality values). We can see that for comparable PSNR values, our method requires significantly lower storage than JPEG. It shows the good potential of our method for image storage compression. Qualitative results can be found the supplementary material.

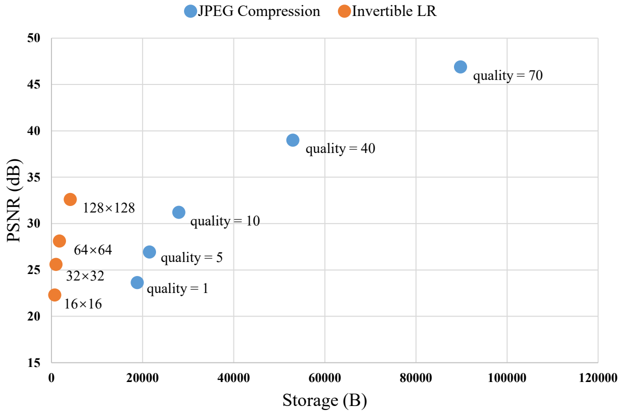


Figure 10. Scatter diagram of PSNR and compressed storage results using our method and JPEG respectively under diverse settings.

Table 5. Quantitative comparison with JPEG compression technology in storage size.

LR Resolution	16×16	32×32	64×64	128×128
PSNR ↑	22.30	25.60	28.13	32.60
Storage(B) ↓	724	978	1731	4103

JPEG Quality	1	5	10	40	70
PSNR ↑	23.64	26.94	31.21	38.99	46.89
Storage(B) ↓	18803	21511	27933	52979	89799

5. Conclusion and Limitations

We proposed the GRAIN framework for generating faithful high-resolution images from low-resolution invertible images with a challenging upscaling factor (64×). Our reciprocal mechanism utilizes both invertible prior and generative prior, allowing us to achieve a fine balance between pixel accuracy and fidelity. Extensive experiments demonstrate superior performance of GRAIN against state-of-the-art rescaling methods, even showing a better compression ratio than JPEG in high-resolution face image.

Limitations. Due to the StyleGAN generative prior, our framework may not perform well for images located outside the pretrained StyleGAN space. The issue can be addressed by fine-tuning the StyleGAN model. In addition, due to the setting of invertible rescaling, our method requires the HR version of image to encode information. It will be an interesting future work to extend our framework to directly perform super-resolution on ordinary LR images.

Acknowledgement. This project is supported by the National Natural Science Foundation of China (No. 61972162); Guangdong International Science and Technology Cooperation Project (No. 2021A0505030009); Guangdong Natural Science Foundation (No. 2021A1515012625); Guangzhou Basic and Applied Research Project (No. 202102021074); and CCF-Tencent Open Research fund (RAGR20210114).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223. PMLR, 2017. 5

[2] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016. 4

[3] Kelvin C.K. Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, pages 14245–14254, 2021. 1, 2, 5, 6

[4] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*, pages 2492–2501, 2018. 2

[5] Ka Leong Cheng, Yueqi Xie, and Qifeng Chen. Iicnet: A generic framework for reversible image conversion. In *ICCV*, pages 1991–2000, 2021. 3

[6] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, pages 11065–11074, 2019. 2

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 5

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 38(2):295–307, 2015. 2

[9] Yong Du, Yangyang Xu, Taizhong Ye, Qiang Wen, Chufeng Xiao, Junyu Dong, Guoqiang Han, and Shengfeng He. Invertible grayscale with sparsity enforcing priors. *ACM TOMM*, 17(3), 2021. 3

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 4

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 2

[12] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *ICCV*, pages 1689–1697, 2017. 2, 5, 6

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 2

[14] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018. 5

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2, 3

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 2, 3, 5

[17] Heewon Kim, Myungsub Choi, Bee Lim, and Kyoung Mu Lee. Task-aware image downscaling. In *ECCV*, pages 399–414, 2018. 2, 3, 5, 6

[18] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, pages 1637–1645, 2016. 2

[19] Mengyan Li, Yuechuan Sun, Zhaoyu Zhang, Haonian Xie, and Jun Yu. Deep learning face hallucination via attributes transfer and enhancement. In *ICME*, pages 604–609. IEEE, 2019. 2

[20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, pages 136–144, 2017. 2

[21] Cheng Ma, Zhenyu Jiang, Yongming Rao, Jiwen Lu, and Jie Zhou. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *CVPR*, pages 5569–5578, 2020. 5, 6

[22] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, pages 2437–2445, 2020. 2, 5, 6

[23] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, pages 2287–2296, 2021. 2, 4, 5, 6, 7

[24] Taizhang Shang, Qiuju Dai, Shengchen Zhu, Tong Yang, and Yandong Guo. Perceptual extreme super-resolution network with receptive field block. In *CVPRW*, pages 440–441, 2020. 2

[25] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *ICCV*, pages 4539–4547, 2017. 2

[26] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*, pages 0–0, 2018. 2, 4, 5, 6

[27] Menghan Xia, Xueting Liu, and Tien-Tsin Wong. Invertible grayscale. *ACM TOG*, 37(6):1–10, 2018. 3

[28] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *ECCV*, pages 126–144, 2020. 1, 2, 3, 5, 6

[29] Ke Xu, Xin Wang, Xin Yang, Shengfeng He, Qiang Zhang, Baocai Yin, Xiaopeng Wei, and Rynson WH Lau. Efficient image super-resolution integration. *The Visual Computer*, 34(6):1065–1076, 2018. 2

[30] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, pages 672–681, 2021. 1, 2, 5, 6

[31] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 5, 7

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 5

[33] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection-how to effectively exploit shape and texture features. In *ECCV*, pages 802–816. Springer, 2008. 5, 7

[34] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, pages 2472–2481, 2018. 2

[35] Yang Zhou, Yangyang Xu, Yong Du, Qiang Wen, and Shengfeng He. Pro-pulse: Learning progressive encoders of latent semantics in gans for photo upsampling. *IEEE TIP*, 31:1230–1242, 2022. 2

[36] Qianshu Zhu, Chu Han, Guoqiang Han, Tien-Tsin Wong, and Shengfeng He. Video snapshot: Single image motion expansion via invertible motion embedding. *IEEE TPAMI*, 43(12):4491–4504, 2020. 3