Timestep-Aware Diffusion Model for Extreme Image Rescaling

Ce Wang, Zhenyu Hu, Wanjie Sun*, Zhenzhong Chen
Wuhan University
{cewang, zhenyuhu, sunwanjie, zzchen}@whu.edu.cn

Abstract

Image rescaling aims to learn the optimal low-resolution (LR) image that can be accurately reconstructed to its original high-resolution (HR) counterpart, providing an efficient image processing and storage method for ultra-high definition media. However, extreme downscaling factors pose significant challenges to the upscaling process due to its highly ill-posed nature, causing existing image rescaling methods to struggle in generating semantically correct structures and perceptual friendly textures. In this work, we propose a novel framework called Timestep-Aware Diffusion Model (TADM) for extreme image rescaling, which performs rescaling operations in the latent space of a pre-trained autoencoder and effectively leverages powerful natural image priors learned by a pre-trained text-to-image diffusion model. Specifically, TADM adopts a pseudo-invertible module to establish the bidirectional mapping between the latent features of the HR image and the target-sized LR image. Then, the rescaled latent features are enhanced by a pre-trained diffusion model to generate more faithful details. Considering the spatially non-uniform degradation caused by the rescaling operation, we propose a novel time-step alignment strategy, which can adaptively allocate the generative capacity of the diffusion model based on the quality of the reconstructed latent features. Extensive experiments demonstrate the superiority of TADM over previous methods in both quantitative and qualitative evaluations. The code will be available at: https://github.com/wwangcece/TADM.

1. Introduction

With the explosive growth of ultra high-resolution (HR) images, image rescaling has become essential for enhancing data storage and transmission efficiency [39], optimizing resource usage in real-time applications [22], and ensuring proper display on devices with varying resolutions [21]. During image rescaling, HR images are downscaled to low-resolution (LR) images for storage, transmission, or display. When HR images are needed, an upscaling process is employed to restore the downscaled images to their original resolution. Unlike image super-resolution (SR), image rescaling allows for joint optimization of both downscaling and upscaling processes, preserving detailed information crucial for effective upscaling.

Recently proposed image rescaling methods can be classified into two main categories: encoder-decoder architecture [12, 25] and invertible neural networks [17, 37]. The encoder-decoder architecture employs two neural networks to perform downscaling and upscaling. While, invertible neural networks view image rescaling as a reversible process, using the forward and backward passes of an invertible neural network to simulate the downscaling and upscaling. Despite the significant improvements in reconstruction quality achieved by these methods, they are limited to small rescaling factors, such as $2\times$ or $4\times$.

However, with the advancement of ultra HR imaging devices, there is an increasing need for extreme rescaling factors (*e.g.*, $16\times$ or $32\times$) to achieve more efficient storage and transmission of ultra HR data [7, 47]. Existing image rescaling methods often struggle to generate sufficient detail under extreme rescaling factors. Therefore, some approaches propose using pre-trained GANs as priors to constrain the solution space of the reconstructed results. GRAIN [47] and BDFlow [14] use pre-trained StyleGAN [9, 10] to generate high-quality upscaled face images. VQIR [34] leverages the high-quality visual embeddings encapsulated in pre-trained VQGAN [3] to achieve extreme rescaling on natural images. Although these methods have made significant progress in perceptual metrics, their reconstruction results still suffer from incorrect semantics and structure.

To address the aforementioned issues, we propose a novel framework named Timestep-Aware Diffusion Model (TADM) for extreme image rescaling, which performs image rescaling in the latent space of a pre-trained autoencoder and leverages the natural image priors stored in large text-to-image diffusion models to restore HR images. Specifically, first, the HR images are encoded by the via the VAE encoder to obtain the latent features. Then, we use an au-

*Corresponding author

toencoder to rescale the latent code of the HR image to the target size, and employ a set of invertible neural networks (INN) to perform the mapping from the feature space to the pixel space to obtain the LR image. Next, the rescaled features are fed into a pre-trained diffusion model to perform a single denoising step, thereby enhancing their perceptual quality. Finally, the enhanced features are decoded via the VAE decoder to obtain the rescaled image. Additionally, we find that the rescaled latent features exhibit non-uniform degradation, which can be interpreted as the addition of noise at varying levels. As a result, different images require different time-steps to perform the denoising process. Previous studies have shown that the time-step in diffusion models directly controls the generative capacity of the model [29]. Therefore, we propose a time-step alignment mechanism that predicts the appropriate time-step based on the reconstruction quality of the rescaled latent features, thereby adaptively tuning the generative capacity of the diffusion model.

Our contributions can be summarized as follows:

- We propose a Timestep-Aware Diffusion Model that performs well in global semantic and structural reconstruction for the extreme image rescaling task.
- We propose a novel decoupled strategy for feature rescaling. It decouples the generation of LR images from the rescaling operation, thereby improving the reconstruction accuracy of the rescaling process.
- We design a time-step alignment strategy to address the spatially non-uniform reconstruction quality of the rescaled latent features, allowing for more effective utilization of the prior information from the text-to-image diffusion model.

2. Related Work

2.1. Image Rescaling

Image rescaling aims to downscale HR images to visually coherent LR images and then plausibly reconstruct the original HR images. Unlike image SR, image rescaling involves joint modeling of both downscaling and upscaling processes, leading to higher reconstruction accuracy. Earlier researches [12, 16] utilize CNNs to build an encoder-decoder framework, treating image rescaling as a unified task by jointly training the downscaling and upscaling processes. Instead of directly generating LR images, CAR [25] proposes generating content-adaptive resampling kernels based on HR images to perform the downsampling operation, avoiding pixel-level constraints on the downscaled images. To fully leverage the inherent reversibility of image rescaling, there has been a surge in the use of INNs for this task [15, 38, 45, 48]. IRN [37] is the first invertible framework that models image rescaling as a bijective transformation, embedding residual high-frequency (HF) components into a case-agnostic latent distribution for efficient reconstruction. HCFlow [17] assumes that the HF information depends on the LR image, thereby achieving better performance by incorporating LR conditions in HF information estimation. In extreme image rescaling, generative priors are usually employed to constrain the outcomes. GRAIN [47] and VQIR [34] leverage pre-trained StyleGAN [9, 10] and VQGAN [3], respectively, to achieve notable enhancements in perceptual quality. Nonetheless, these approaches are often limited in image contents, either confined to handling images from specific domains or encountering difficulties when reconstructing faces and textual elements.

2.2. Diffusion Model for Image SR

Here, we briefly introduce the application of diffusion models in a related task, *i.e.*, image SR. Diffusion models promote the development of natural image SR in two main ways: training diffusion models from scratch [4, 13, 41] and leveraging pre-trained SD model [19, 36, 40]. The former is represented by SR3 [24], which achieves perceptual quality comparable to GANs. The latter is represented by StableSR [28], which aims to build a model capable of handling image restoration under any complex degradation in real-world scenarios.

However, the above methods typically require dozens or even hundreds of iterations during inference, which significantly increases computational costs. Researchers have begun exploring how to achieve image SR using one-step diffusion models. SinSR [31], based on ResShift [41], derives a deterministic one-step sampling process and employs a consistency preserving loss to distill the prior knowledge from a multi-step teacher network into a one-step student network. A representative work, OSEDiff [35], directly feeds the LQ image into the SD model and employs variational score distillation for regularization. S3Diff [43] proposes a degradation-aware LoRA and utilizes a pre-trained DINO [44] model as a discriminator for adversarial training. InvSR [42] designs a partial noise prediction strategy to provide a dynamic starting sampling point and supports inference with an arbitrary number of sampling steps ranging from one to five.

3. Methodology

3.1. Overview of TADM

As illustrated in Fig. 1, TADM consists of four parts: **Latent Encoding**, **Feature Rescaling in the Latent Space**, **Denoising Guided Perception Enhancement** and **Latent Decoding**. First, the input HR images x are mapped to the latent features z via the pre-trained VAE encoder $\mathcal{E}$. Subsequently, the proposed Decoupled Feature Rescaling Module (DFRM) rescales z to the target size, outputting LR images y and the rescaled latent features $\hat{z}$. Our aim is to make $\hat{z}$

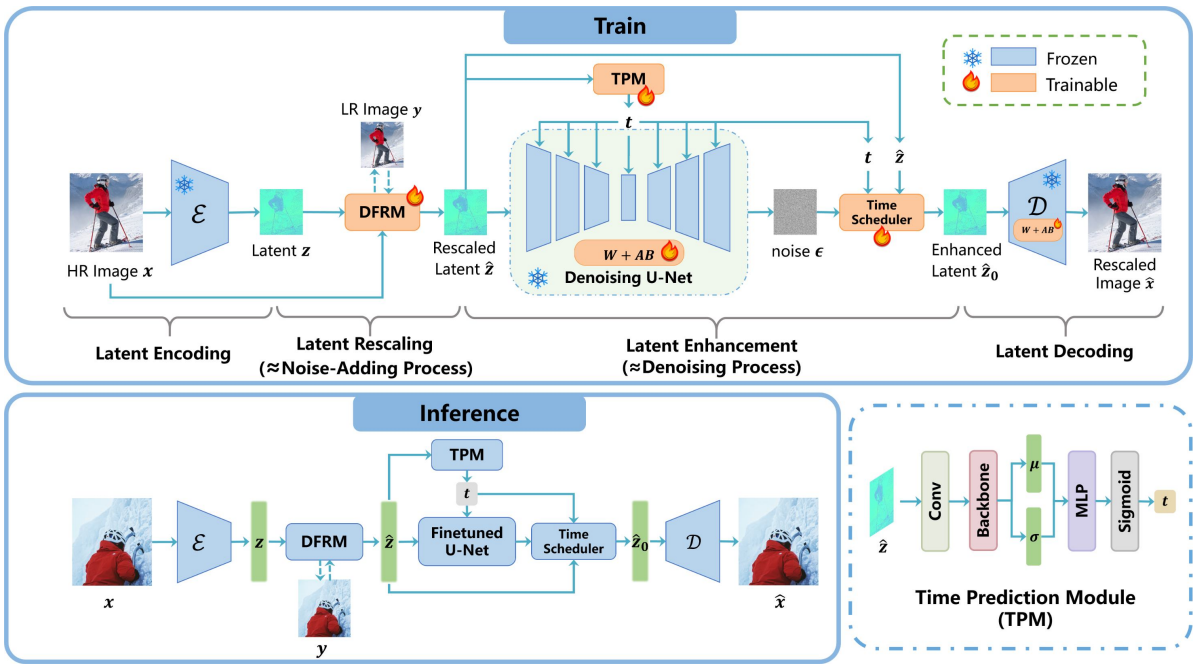

Figure 1. Overview of the proposed Timestep-Aware Diffusion Model (TADM). First, the HR images x are encoded to obtain the latent features z. The latent features are then rescaled to the target size using the decoupled feature rescaling module (DFRM), producing the LR images y and outputting the rescaled latent features $\hat{z}$. Next, the rescaled latent features $\hat{z}$ are passed into a pre-trained diffusion model to perform a single denoising step for perceptual enhancement to obtain $\hat{z}_0$. In this process, a time prediction module (TPM) is used to estimate the time step t based on $\hat{z}$ and the predicted t is then fed into both the U-Net and the time scheduler. Finally, the perceptually enhanced latent features $\hat{z}_0$ are decoded to obtain the rescaled image $\hat{x}$.

as close as possible to z so that the rescaled results align with the latent space of the pre-trained VAE, while ensuring that y and the original HR images x share the same content. Then, the rescaled features $\hat{z}$ are fed into a denoising U-Net network to perform a single denoising step to get the perceptually enhanced latent features $\hat{z}_0$. Finally, $\hat{z}_0$ is decoded via the pre-trained VAE decoder $\mathcal{D}$ to obtain the rescaled image $\hat{x}$. Overall, the latent rescaling process can be regarded as a noise-adding process, while the latent enhancement process can be viewed as a denoising process. To dynamically estimate the 'noise' density introduced by latent rescaling, we design a time alignment strategy, which consists of a time-step prediction module (TPM) and a time scheduler module. It can better address the non-uniform degradation caused by rescaling, thereby enabling the dynamic allocation of generative capabilities of pre-trained SD.

3.2. Feature Rescaling in the Latent Space

3.2.1. Network Architecture.

In this part, we perform rescaling of the HR latent features z, obtaining LR images y and rescaled features $\hat{z}$. A straightforward idea is to construct a transformation chain as $y = G_{\mathrm{e}}(z)$, $\hat{z} = G_{\mathrm{d}}(y)$. Here, G_{e} is the encoder used to downscale the input features to the target sized LR images, while G_{d} represents the decoder used to reconstruct the original HR features. Guidance loss $\mathcal{L}_{\text{gui.}}$ and reconstruction loss $\mathcal{L}_{\text{rec.}}$ are applied to y and $\hat{z}$ respectively, to force y to have the same content as x, and $\hat{z}$ to be as similar as possible to z:

$$\mathcal{L}_{\text{gui.}} = \| y - \text{Bicubic}(x) \|_1 \tag{1}$$

$$\mathcal{L}_{\text{rec.}} = \| \hat{z} - z \|_1 \tag{2}$$

However, these two losses are contradictory to each other. When the guidance loss approaches zero, the rescaling problem degenerates into a classical SR problem [33], embedding little information about the downscaling process into the LR image. In contrast, when the reconstruction loss reaches its global optimum, the rescaling problem degenerates into an image compression problem [20], causing the LR image to lack meaningful visual content. Therefore, directly applying the guidance loss $\mathcal{L}_{\text{gui.}}$ on y could adversely impact subsequent upscaling.

Considering the above issues, we devise the DFRM as shown in Fig. 2, which divides the rescaling operation into two sub-tasks: feature rescaling and LR image generation. Inspired by FGRN [15], we use a set of CNNs to perform downscaling and upscaling, and a separate INN to

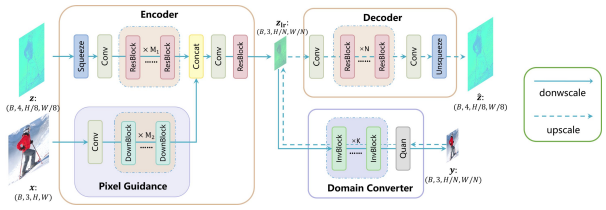

Figure 2. Detailed architecture of Decoupled Feature Rescaling Module (DFRM).

perform the bidirectional mapping between the feature domain and the pixel domain. To achieve a balance between feature rescaling accuracy and visual quality of LR images, we introduce another pixel guidance module in the encoder to incorporate pixel-level information, which can improve the quality of the LR image. Specifically, DFRM perform feature rescaling through two independent transformation chains $((x, z) \rightarrow z_{\text{lr}} \rightarrow \hat{z})$; $(z_{\text{lr}} \leftrightarrow y)$ as follows:

$$z_{\text{lr}} = G_{\text{e}}\left(x, z\right) \tag{3}$$

$$z_{\text{lr}} \approx F^{-1}\left(F\left(z_{\text{lr}}\right)\right) = F^{-1}\left(y\right) \tag{4}$$

$$\hat{z} = G_{\text{d}}\left(z_{\text{lr}}\right) \tag{5}$$

Here, F and F^{-1} represent the forward and backward processes of the INN, respectively. Given the latent variable $z \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 4}$ of the pre-trained SD VAE as input, the encoder G_{e} first transforms it into a compact representation $z_{\text{lr}} \in \mathbb{R}^{\frac{H}{N} \times \frac{W}{N} \times 3}$, whose size is the same as that of the target LR image and N represents the target downscaling factor. Subsequently, a set of INNs perform the forward process to transform z_{lr} from the feature domain to the pixel domain and obtain the LR image $y \in \mathbb{R}^{\frac{H}{N} \times \frac{W}{N} \times 3}$ by quantization. In the upscaling process, the inverse of the INN is used to process the LR image y and reconstruct z_{lr}. Finally, the decoder G_{d} is utilized to obtain the rescaled feature $\hat{z}$.

3.2.2. Loss Function.

Training of this part involves two losses: reconstruction loss and guidance loss. Considering the information lost in quantization operations, the bidirectional mapping between z_{lr} and y is not exactly reversible. To ensure the robustness of the network to quantization operations, we consider two types of transformation chains in designing the reconstruction loss: $(x, z) \rightarrow z_{\text{lr}} \rightarrow \hat{z}$ and $(x, z) \rightarrow z_{\text{lr}} \rightarrow y \rightarrow \hat{z}_{\text{lr}} \rightarrow \hat{z}$, which can be expressed as the following formula:

$$\begin{aligned} \mathcal{L}_{\text{rec.}} = &\parallel G_{\text{d}}\left(G_{\text{e}}\left(x, z\right)\right) - z \parallel_1 + \\ &\parallel G_{\text{d}}\left(F^{-1}\left(F\left(G_{\text{e}}\left(x, z\right)\right)\right)\right) - z \parallel_1 \end{aligned} \tag{6}$$

This combination can facilitate the network to reconstruct $\hat{z}$ with high fidelity in both the existence and absence of quantization operations. Additionally, the guidance loss is designed as follows:

$$\mathcal{L}_{\text{gui.}} = \parallel F\left(G_{\text{e}}\left(x, z\right)\right) - \text{Bicubic}(x) \parallel_1 \tag{7}$$

Finally, the loss function of this part is a combination of these two losses:

$$\mathcal{L}_{\text{res.}} = \lambda_{\text{rec.}}\mathcal{L}_{\text{rec.}} + \lambda_{\text{gui.}}\mathcal{L}_{\text{gui.}} \tag{8}$$

3.3. Denoising Guided Perception Enhancement

The latent variable $\hat{z}$ obtained in Sec. 3.2 retains sufficient semantic information about the HR image, but lacks adequate texture. Therefore, we propose to leverage the strong priors stored in the pre-trained SD model to enhance the perceptual quality of the rescaling results in this part. Original diffusion model requires a multi-step sampling algorithm, significantly increases inference time and makes it difficult to scale to practical applications. Inspired by recent work on accelerating diffusion models [8], we adopt an one-step inference strategy. Specifically, we can directly feed the rescaled latent variable $\hat{z}$ into the denoising U-Net $\epsilon = \epsilon_\theta\left(\hat{z}, t\right)$ and perform a single denoising step as follows:

$$\hat{z}_0 = \frac{\hat{z} - \sqrt{1 - \bar{\alpha}_t}\epsilon}{\sqrt{\bar{\alpha}_t}} \tag{9}$$

where $\hat{z}_0$ represents the perceptually enhanced latent features, $\bar{\alpha}_t$ is the noise scheduler of the forward diffusion process of the SD model, and t represents the time-step associated with the noise intensity. Finally, $\hat{z}_0$ can be transformed to the pixel domain through the pre-trained VAE decoder $\mathcal{D}$ to obtain the upscaled image $\hat{x} = \mathcal{D}\left(\hat{z}_0\right)$. Then, we can adopt the loss function between $\hat{x}$ and x to optimize the U-Net using LoRA layers.

However, the key problem of the aforementioned approaches lies in the selection of time-step t. We can regard the rescaling process as the noise-adding process of diffusion model, while Eq. (9) represents the denoising process. For different rescaling factors and different images, the 'noise' introduced by the rescaling operation varies in intensity. In Fig. 3, we present the global rescaling error of an image under different rescaling factors, as well as the local rescaling error of different patches within the image. Additionally, we plot the error curve of the forward noise-adding process of the SD model. It can be observed that as the rescaling factor and image content change, rescaling MSE also varies and corresponds to the diffusion MSE at different time-steps. Thus, we propose to dynamically predict the time-step t based on the quality of the rescaled latent features $\hat{z}$. A simple yet effective network named time-step prediction module (TPM), whose structure is presented in Fig. 1, is used to predict t based on the rescaled latent features: $t = \text{TPM}\left(\hat{z}\right)$. During the training phase, we compute a single time-step for each input image patch. In the inference phase, we adopt a tiled inference approach (which will be elaborated in detail in the supplementary materials), allowing for predicting different time-steps for various regions of a high-resolution input image.

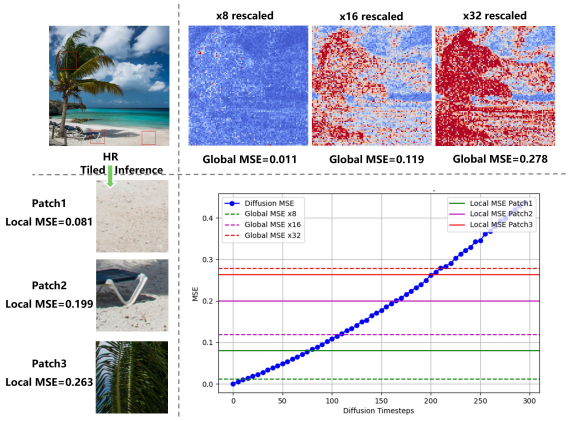

Figure 3. We compare the MSE introduced by the rescaling operation with that caused by the forward diffusion process. The rescaling degradation severity is correlated with both the image content and the rescaling scale. Therefore, the model requires adaptive time-step to align the rescaling MSE with the diffusion MSE.

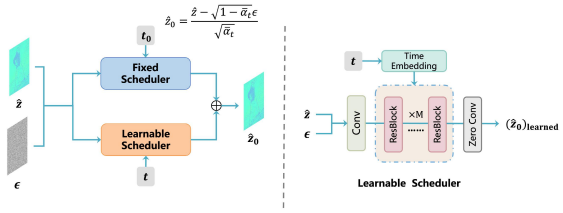

Figure 4. The structure of the time scheduler module. It combines both the fixed scheduler and learnable scheduler to perform the denoising process. The two schedulers are connected by a zero-convolution layer to stabilize the training.

Subsequently, the predicted time-step t is used to perform a single denoising step. However, the vanilla SD time scheduler as shown in Eq. (9) is not differentiable with respect to t, which prevents gradients from being back-propagated to the TPM. Considering that given $\hat{z}$, ϵ and t, Eq. (9) essentially represents a linear transformation, which can be learned by a neural network. Therefore, we propose a hybrid time scheduler module consisting of a fixed scheduler and a learnable scheduler, whose structure is presented in Fig. 4. The former uses Eq. (9) to perform the denoising step, while the latter employs a neural network to simulate the denoising scheduler. To ensure training stability, the final layer of the learnable scheduler is a zero-initialized convolution, allowing it to gradually correct the results of the fixed scheduler as training progresses. This process can be described as follows:

$$\hat{z}_0 = \mathcal{S}_{\text{fixed}}\left(\hat{z}, \epsilon, t_0\right) + \mathcal{S}_{\text{learned}}\left(\hat{z}, \epsilon, t\right) \tag{10}$$

where t_0 is the fixed pre-set time-step and $\mathcal{S}_{\text{fixed}}$ denotes the vanilla SD time scheduler as shown in Eq. (9).

Finally, we fine-tune the denoising U-Net network and the VAE decoder using LoRA, while also jointly training the TPM and time scheduler module. To achieve a balance between fidelity and realism, we supervise the reconstruction results $\hat{x}$ using pixel loss and perceptual loss as follows:

$$\begin{aligned}\mathcal{L}_{\text{enh.}} = &\left\|\hat{x} - x\right\|_1 + \\ &\lambda_{\text{pec.}}\left(\mathcal{L}_{\text{lpips}}\left(\hat{x}, x\right) + \mathcal{L}_{\text{dists}}\left(\hat{x}, x\right)\right)\end{aligned} \tag{11}$$

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets and Evaluation Metrics.

We train the proposed TADM using the DF2K dataset [1, 18]. Considering that traditional benchmarks such as Set5, Set14, and BSD100 often have lower quality and resolution [5], they are not suitable for evaluating extreme rescaling methods. Therefore, we utilize four high-resolution datasets: DIV2K, CLIC2020 [26], Urban100 [6], and DIV8K [5] to evaluate models been compared (results on the latter two datasets are presented in the supplementary materials). We adopt PSNR and SSIM [32] for quantitative evaluation. For perceptual quality evaluation, reference-based metrics LPIPS [46] and DISTS [2] are used. For non-reference image quality assessment, we use metrics including MUSIQ [11] and CLIPIQA [27].

4.1.2. Training Details

TADM is built based on the SD 2.1-base model, and the training is divided into three stages. First, we train the DFRM using Eq. (8). Then we adopt Eq. (11) to jointly train the LoRA layers, the TPM and the time scheduler module. Finally, we use a smaller learning rate to jointly fine-tune DFRM, LoRA layers, TPM and scheduler module using a combination of Eq. (8) and (11). More training details are provided in the supplementary material.

4.2. Results and Comparison with SOTA

We compare the proposed TADM with three kinds of methods: (1) bicubic downscaling and SR, *i.e.*, ESRGAN [30], StableSR [28] (based on multi-step diffusion) and S3Diff (based on single-step diffusion)[43]; (2) flow-based rescaling models, *i.e.*, IRN [37] and HCFlow [17]; (3) prior-based extreme rescaling method VQIR [34]. Some of them are proposed for 4× rescaling factors and we extended them to 16× or 32×. To ensure a fair comparison, all methods are trained until convergence on the DF2K dataset following their original configurations.

4.2.1. Quantitative Results

Table 1 presents the 16× and 32× quantitative image rescaling results on four datasets. The performance of SR methods are generally poor. Both gan-based (ESRGAN) and diffusion-based (StableSR and S3Diff) methods can only achieve relative good performance on no-reference metrics.

Table 1. Quantitative comparisons with different methods at 16× and 32× (results on Urban100 and DIV8K are presented in the supplementary materials). The symbols ↑ and ↓ respectively represent that higher or lower values indicate better performance. Bold represents the best and underline represents the second best.

Dataset	Method	PSNR ↑		SSIM ↑		LPIPS ↓		DISTS ↓		MUSIQ ↑		CLIPIQA ↑	
		16×	32×	16×	32×	16×	32×	16×	32×	16×	32×	16×	32×
DIV2K	ESRGAN	23.15	19.94	0.5946	0.5689	0.4478	0.5892	0.2378	0.4230	59.80	43.35	0.6161	0.3734
	StableSR	22.02	19.92	0.5983	0.5588	0.4980	0.5736	0.2078	0.2603	51.28	48.94	0.3780	0.3203
	S3Diff	20.22	17.81	0.5491	0.4601	0.4033	0.4895	0.1309	0.1810	64.37	67.92	0.6228	0.6991
	IRN	26.04	22.84	0.7020	0.6212	0.5153	0.6095	0.3041	0.4036	43.72	28.70	0.2799	0.2336
	HCFlow	**26.66**	**23.89**	**0.7176**	**0.6467**	0.4885	0.5816	0.2866	0.3852	46.43	37.25	0.2735	0.2792
	VQIR	23.91	22.02	0.6498	0.5823	0.3174	0.4568	0.1024	0.2663	64.04	58.21	0.6350	0.6293
	Ours	23.98	22.18	0.6736	0.5869	**0.2979**	**0.4221**	**0.0886**	**0.1684**	**66.56**	**69.12**	**0.7189**	**0.7204**
CLIC2020	ESRGAN	25.43	22.47	0.7018	0.6804	0.4213	0.5429	0.2209	0.4011	57.80	41.28	0.6107	0.3558
	StableSR	24.11	22.24	0.7052	0.6726	0.4625	0.5288	0.1860	0.2282	52.69	49.80	0.3779	0.3325
	S3Diff	22.04	19.87	0.6615	0.5711	0.3968	0.4829	0.1336	0.1812	63.07	66.40	0.6001	0.6838
	IRN	28.55	24.93	0.7907	0.7228	0.4676	0.5528	0.2746	0.3776	47.66	27.81	0.2950	0.2495
	HCFlow	**28.82**	**26.21**	**0.7949**	**0.7451**	0.4542	0.5325	0.2635	0.3611	47.72	37.15	0.2733	0.2884
	VQIR	25.69	24.12	0.7363	0.6942	0.2854	0.4191	0.0901	0.2445	64.36	54.83	0.6105	0.6068
	Ours	26.04	24.26	0.7542	0.6903	**0.2724**	**0.3911**	**0.0792**	**0.1513**	**65.35**	**69.10**	**0.6815**	**0.7271**

This is because they rely solely on the LR image to reconstruct the original image and cannot jointly optimize the downscaling and upscaling processes. While methods based on INNs (HCFlow and IRN) achieve the highest PSNR and SSIM due to its consideration of the correlation between downscaling and upscaling. In all generative methods, VQIR achieves relatively good perceptual metrics by leveraging the priors of the pre-trained VQGAN. In contrast, our method achieves optimal performance across all perceptual metrics on all datasets. For instance, in the 32× rescaling on the DIV2K validation set, our method improves the DISTS metric by 36.76% compared to the second-best VQIR. Importantly, our approach is designed for applications like social media delivery, where semantic fidelity and perceptual realism are more critical. In such cases, metrics like LPIPS and DISTS better reflect image quality [23].

4.2.2. Qualitative Results

In Fig. 5, we present visual comparisons at 16× and 32× rescaling factors. Due to the ill-posed nature of extreme rescaling, pixel-oriented optimization methods such as IRN, and HCFlow produce overly-smoothed results. Although ESRGAN incorporates perceptual optimization it still cannot generate adequate details. VQIR leverages the pre-trained VQGAN as a prior, allowing it to generate rich details and sharp edges. However, it struggles with faces and more complex structures. Although S3Diff can generate sharp edges and rich details, its reconstructed results have low fidelity and exhibit significant artifacts and color distortions. Thanks to the powerful prior provided by the pre-trained SD and the carefully designed architecture,, our method effectively addresses these issues. It not only generates sufficiently rich textures but also reconstructs accuracy semantics and regular structures, such as text, faces, vehicles, and leaves.

4.3. Ablation Study

We conduct ablation experiments to validate the effectiveness of each component. All ablations are trained on the DF2K dataset and tested on the validation set of the DIV2K for 16× rescaling task.

4.3.1. Effectiveness of Feature Rescaling

To verify the effectiveness of the feature rescaling operation, we move the rescaling operation from the latent space to the pixel space. Specifically, we use HCFlow [17], the comparative method with the best PSNR performance, to perform the rescaling operation in the pixel domain, followed by perceptual enhancement using the same approach described in Sec. 3.3. As shown the 1st row in Table 2, the rescaling operation in the latent space can improve reconstruction performance, as evidenced by much lower LPIPS and DISTS. Additional details and results are presented in the supplementary materials.

Table 2. Ablation study of rescaling space and SD prior.

Latent space	Pixel space	SD prior	LPIPS ↓	DISTS ↓
×	✓	✓	0.3630	0.1154
✓	×	×	0.4675	0.3109
✓	×	✓	**0.2979**	**0.0886**

4.3.2. Necessity of SD Prior

To validate that SD model provides rich priors about real-world HR images, we directly omit the perceptual enhancement part. As shown in Table 2 (2nd row), the introduction

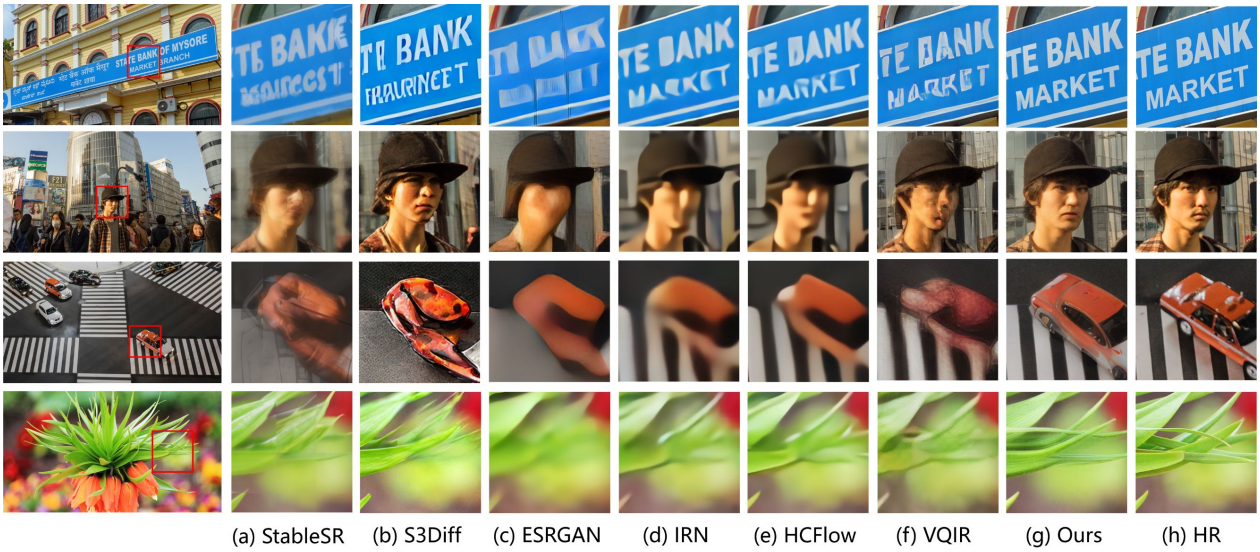

Figure 5. Qualitative comparisons of 16× (the 1st and 2nd rows) and 32× (the 3rd and 4th rows) rescaling results. Our method achieves higher semantic accuracy, such as more clearer text, more recognizable faces, and more realistic structures.

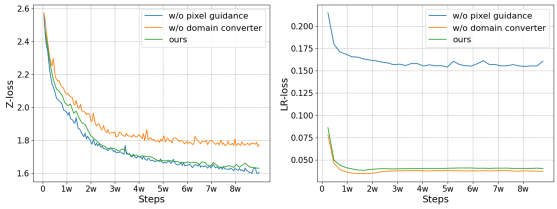

Figure 6. The validation loss curves when removing a specific component from the DFRM.

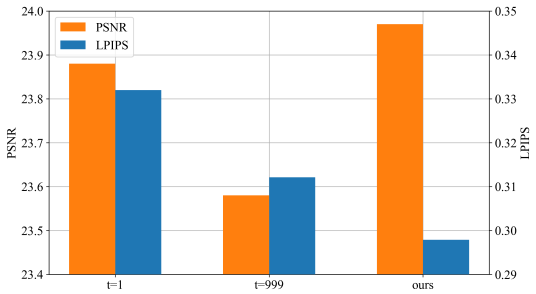

Figure 7. Ablation study about time alignment.

of the SD priors can significantly improve performance of LPIPS and DISTS and greatly enhance the perceptual quality of the reconstruction results.

4.3.3. Effectiveness of DFRM Design

Our DFRM includes two specific designs: the domain converter module and the pixel guidance module. To analyze their impact, we remove each component individually and plot the validation loss curves, as shown in Fig. 6. It can be observed that the introduction of the INN can reduce reconstruction error. Although the pixel guidance module does not improve reconstruction performance, our experiments show that it can enhance the fidelity and visual quality of the LR images, as demonstrated by the lower guidance loss in Fig. 6. Therefore, the design of our DFRM strikes a good balance between feature reconstruction and LR image quality. More experiments and the visual quality comparisons of LR images are presented in the supplementary materials.

4.3.4. Effectiveness of Time Alignment

To validate the effectiveness of our proposed time-step alignment strategy, we retrain our model using fixed time-steps. This involves removing the TPM and time scheduler modules and employing a manually selected time-step to perform a single denoising step based on Eq. (9). Following the setting of previous work [35], we select fixed time-steps 1 and 999 for comparison. As shown in Fig. 7, when the time-step is set to a smaller value, the model achieves higher fidelity but lower perceptual quality, and the situation is reversed when a larger time-step is chosen. In contrast, our time-step alignment mechanism can dynamically determine the level of generative capability based on the image content, thereby achieving optimal performance in both fidelity and perceptual quality.

In Fig. 8, we present a schematic illustration of the tiled inference process along with the predicted time-step map.

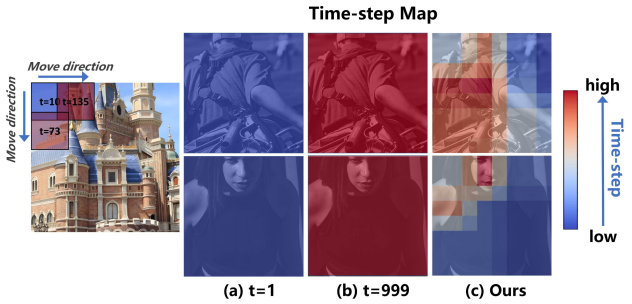

Figure 8. Illustration of the tiled inference process and the predicted time-step map.

It can be observed that the time alignment mechanism effectively captures more complex regions in high-resolution images and assigns larger time steps to these areas. Details regarding the selection of patch size and stride length are provided in the supplementary materials.

4.4. Discussions

4.4.1. Application to Image Compression

Our model can be combined with existing lossless image compression methods to achieve extreme compression ratios. To validate this, we use PNG to compress LR images and compare it with JPEG at various bit-rates. As shown in Fig. 9, at the same bits per pixel (bpp), the image reconstruction quality of our model significantly surpasses that of the JPEG compression. These comparisons demonstrate the potential of our model for image compression.

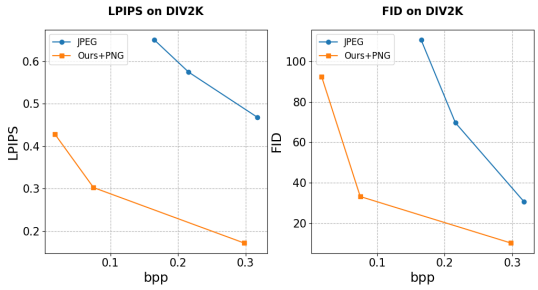

Figure 9. R-D curves of JPEG compression and our proposed model on the DIV2K validation set.

4.4.2. Inference Efficiency Comparison

In Table 3, we provide comparisons of inference time and memory consumption for representative methods. All tests are performed on an RTX 3090 using images with a resolution of 1024×1024. It can be observed that flow-based models (IRN and HCFlow) require less GPU memory consumption and enable fast inference. In contrast, StableSR, which is based on multi-step diffusion models, demands significantly more inference time and larger GPU memory consumption. Although VQIR requires less memory, the nearest neighbor search process in VQGAN results in longer inference time. For one-step diffusion-based method S3Diff, due to its degradation-guided LoRA which needs to dynamically predict LoRA parameters during inference, it requires longer ingerence time and higher memory consumption. In contrast, our proposed TADM demonstrates moderate efficiency and scalability compared to other methods. In the future, we can utilize distillation techniques to further reduce the model size.

Table 3. Comparison of inference time and memory consumption.

	IRN	HCFlow	VQIR	StableSR	S3Diff	Ours
Inference Time (s)	0.359	0.363	0.897	135.2	2.186	0.711
Memory Consumption (G)	0.498	0.098	0.344	6.380	5.025	2.598

4.4.3. Upper Bound of TADM

Since our method performs the rescaling operation in the latent space of a pre-trained VAE, its performance upper boud is inherently constrained by the VAE. In Table 4, we present a performance comparison between VAE and two $4\times$ rescaling models. For VAEs, we directly feed the high-resolution images to the encoder and decoder without performing any rescaling operations. This performance represents the upper bound of our method's performance across all rescaling factors. It can be observed that VAE significantly lags behind the latter. In the future, we could explore integrating stronger VAEs, such as those from SDXL, SD3, thereby enabling the TADM to be more applicable to small-scale rescaling tasks.

Table 4. Comparison of VAE and $4\times$ rescaling methods.

	IRN	HCFlow	VAE
PSNR	35.07	35.23	29.01
SSIM	0.9318	0.9346	0.8046

5. Conclusion

We propose a novel framework called TADM to address issues such as insufficient texture and semantic inaccuracies in extreme image rescaling. Specifically, we design a DFRM to achieve bidirectional mapping between HR features and LR images. Then, we utilize a pre-trained SD model to perform a single denoising step on the rescaled features, enhancing its perceptual quality. Considering the non-uniform quality of the rescaled latent features, we propose a novel time-step alignment strategy to achieve a balance between fidelity and perceptual quality. Both qualitative and quantitative experiments demonstrate the superiority of our approach.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62201403, in part by Hubei Provincial Natural Science Foundation of China under Grant 2024AFB549. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 5

[2] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020. 5

[3] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 1, 2

[4] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10021–10030, 2023. 2

[5] Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. In *IEEE International Conference on Computer Vision Workshop*, pages 3512–3516, 2019. 5

[6] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015. 5

[7] Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Generative latent coding for ultra-low bitrate image compression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 26088–26098, 2024. 1

[8] Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling diffusion models into conditional gans. *arXiv preprint arXiv:2405.05967*, 2024. 4

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 2

[10] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1, 2

[11] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *IEEE International Conference on Computer Vision*, pages 5148–5157, 2021. 5

[12] Heewon Kim, Myungsub Choi, Bee Lim, and Kyoung Mu Lee. Task-aware image downscaling. In *European Conference on Computer Vision*, pages 399–414, 2018. 1, 2

[13] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 2

[14] Jinmin Li, Tao Dai, Jingyun Zhang, Kang Liu, Jun Wang, Shaoming Wang, Shu-Tao Xia, et al. Boundary-aware decoupled flow networks for realistic extreme rescaling. *arXiv preprint arXiv:2405.02941*, 2024. 1

[15] Shang Li, Guixuan Zhang, Zhengxiong Luo, Jie Liu, Zhi Zeng, and Shuwu Zhang. Approaching the limit of image rescaling via flow guidance. *arXiv preprint arXiv:2111.05133*, 2021. 2, 3

[16] Yue Li, Dong Liu, Houqiang Li, Li Li, Zhu Li, and Feng Wu. Learning a convolutional neural network for image compact-resolution. *IEEE Transactions on Image Processing*, 28(3): 1092–1107, 2018. 2

[17] Jingyun Liang, Andreas Lugmayr, Kai Zhang, Martin Danelljan, Luc Van Gool, and Radu Timofte. Hierarchical conditional flow: A unified framework for image super-resolution and image rescaling. In *IEEE International Conference on Computer Vision*, pages 4076–4085, 2021. 1, 2, 5, 6

[18] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 5

[19] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European Conference on Computer Vision*, pages 430–448, 2024. 2

[20] Siwei Ma, Xinfeng Zhang, Chuanmin Jia, Zhenghui Zhao, Shiqi Wang, and Shanshe Wang. Image and video compression with neural networks: A review. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6):1683–1698, 2019. 3

[21] Zhihong Pan, Baopu Li, Dongliang He, Mingde Yao, Wenhao Wu, Tianwei Lin, Xin Li, and Errui Ding. Towards bidirectional arbitrary image rescaling: Joint optimization and cycle idempotence. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 17389–17398, 2022. 1

[22] Chenyang Qi, Xin Yang, Ka Leong Cheng, Ying-Cong Chen, and Qifeng Chen. Real-time 6k image rescaling with rate-distortion optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14092–14101, 2023. 1

[23] Lucas Relic, Roberto Azevedo, Markus Gross, and Christopher Schroers. Lossy image compression with foundation diffusion models. In *European Conference on Computer Vision*, pages 303–319, 2024. 6

[24] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, 45(4):4713–4726, 2022. 2

[25] Wanjie Sun and Zhenzhong Chen. Learned image downscaling for upscaling using content adaptive resampler. *IEEE Transactions on Image Processing*, 29:4027–4040, 2020. 1, 2

[26] George Toderici, Lucas Theis, Nick Johnston, Eirikur Agustsson, Fabian Mentzer, Johannes Ballé, Wenzhe Shi, and Radu Timofte. Clic 2020: Challenge on learned image compression, 2020. 5

[27] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 5

[28] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 2, 5

[29] Luozhou Wang, Shuai Yang, Shu Liu, and Ying-cong Chen. Not all steps are created equal: Selective diffusion distillation for image manipulation. In *IEEE International Conference on Computer Vision*, pages 7472–7481, 2023. 2

[30] 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 *European Conference on Computer Vision Workshops*, pages 0–0, 2018. 5

[31] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 25796–25805, 2024. 2

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[33] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3365–3387, 2020. 3

[34] Hao Wei, Chenyang Ge, Zhiyuan Li, Xin Qiao, and Pengchao Deng. Towards extreme image rescaling with generative prior and invertible prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6181–6193, 2024. 1, 2, 5

[35] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems*, pages 92529–92553, 2024. 2, 7

[36] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 25456–25467, 2024. 2

[37] Mingqing Xiao, Shuxin Zheng, Chang Liu, Zhouchen Lin, and Tie-Yan Liu. Invertible rescaling network and its extensions. *International Journal of Computer Vision*, 131(1):134–159, 2023. 1, 2, 5

[38] Bingna Xu, Yong Guo, Luoqian Jiang, Mianjie Yu, and Jian Chen. Downscaled representation matters: Improving image rescaling with collaborative downscaled images. In *IEEE International Conference on Computer Vision*, pages 12237–12247, 2023. 2

[39] Jinhai Yang, Mengxi Guo, Shijie Zhao, Junlin Li, and Li Zhang. Self-asymmetric invertible network for compression-aware image rescaling. In *AAAI Conference on Artificial Intelligence*, pages 3155–3163, 2023. 1

[40] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European Conference on Computer Vision*, pages 74–91, 2024. 2

[41] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *Advances in Neural Information Processing Systems*, pages 13294–13307, 2023. 2

[42] Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. *arXiv preprint arXiv:2412.09013*, 2024. 2

[43] Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*, 2024. 2, 5

[44] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2

[45] Min Zhang, Zhihong Pan, Xin Zhou, and C-C Jay Kuo. Enhancing image rescaling using dual latent variables in invertible neural network. In *ACM International Conference on Multimedia*, pages 5602–5610, 2022. 2

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5

[47] Zhixuan Zhong, Liangyu Chai, Yang Zhou, Bailin Deng, Jia Pan, and Shengfeng He. Faithful extreme rescaling via generative prior reciprocated invertible representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5708–5717, 2022. 1, 2

[48] Yiming Zhu, Cairong Wang, Chenyu Dong, Ke Zhang, Hongyang Gao, and Chun Yuan. High-frequency normalizing flow for image rescaling. *IEEE Transactions on Image Processing*, 32:6223–6233, 2022. 2