Timestep-Aware Diffusion Model for Extreme Image Rescaling

Supplementary Material

1. Training Details

TADM is built based on the SD 2.1-base model, and the training is divided into three stages. In the first stage, we train the DFRM together with the pre-trained VAE model for 200K iterations using Adam optimizer. The learning rate is initialized as 2×10^{-4} and reduced by half every 50k iterations. Then, in the second stage, we jointly train the LoRA layers of the denoising U-Net, the TPM and the time scheduler module for 50K iterations using AdamW optimizer. The LoRA rank for U-Net and VAE decoder is set as 48 and 16 respectively. The learning rate is initialized as 3×10^{-4} and reduced by half every 15k iterations. Finally, we fine-tuned the DFRM, LoRA layers, TPM module, and time scheduler module jointly for 10k iterations using a learning rate of 1×10^{-5}.

2. Discussions

2.1. Visual Comparison with JPEG Compression

In Fig. 1, we present the visual quality, the bpp and the LPIPS of the reconstructed images. Compared to JPEG compression, our method requires comparable or less storage space while achieving superior objective metrics. Additionally, JPEG compression tends to produce noticeable block artifacts and color distortions in the background regions of the images, whereas our model maintains superior reconstruction quality across all three scales.

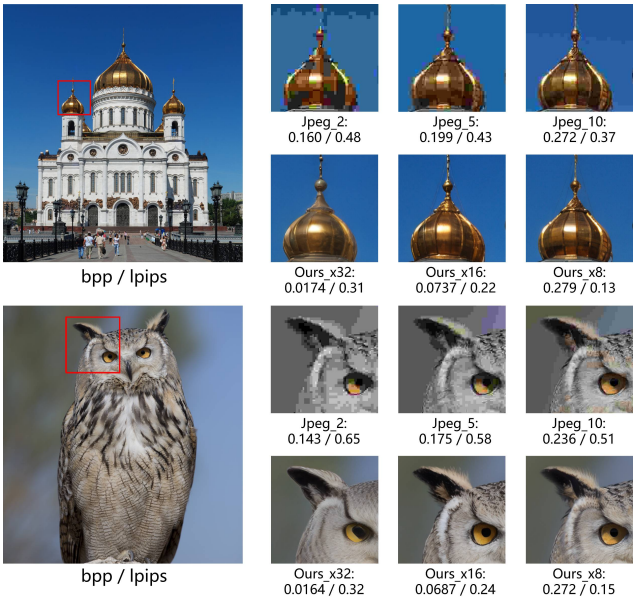

Figure 1. Visual comparisons between our model and JPEG compression. bpp/LPIPS of the reconstructed images are shown below the images.

2.2. Details about Tiled Inference

When TADM is employed for rescaling ultra-high-resolution images, it often necessitates dividing the input image into multiple patches for separate processing. In the main paper, this is referred to as the tiled inference strategy. In Algorithm 1, we elaborate on the process of tiled inference. Unlike works such as StableSR [4], our tiled inference algorithm can predict different time steps t for each image patch, thereby achieving dynamic allocation of generative capacity.

The aforementioned algorithm encompasses two hyperparameters: the patch size and the stride length, both defined in the latent space. If the patch size is set too large, although inference efficiency may improve, the prediction of time steps would become overly sparse, leading to performance degradation. Conversely, if the patch size is set too small, the computational load will be significantly higher, and there might be a misalignment between the inference size and the pre-training image size of Stable Diffusion.

Therefore, to achieve a trade-off between performance and inference efficiency, we conduct experiments on the patch size, as shown in Fig. 2. It can be observed that the model achieves optimal performance when the patch size is between 60 and 120. Consequently, we select 96 as the

Algorithm 1 Tiled Inference Process

Input: Input image x, latent encoder $\mathcal{E}$, decoupled feature rescaling module DFRM, denoising U-Net ϵ_θ, time prediction module TPM, time scheduler TS, latent decoder $\mathcal{D}$, patch size p, stride length s

Output: Rescaled image $\hat{x}$, LR image y

$z = \mathcal{E}(x)$ ▷ latent encoding
$\hat{z}, y = \mathrm{DFRM}(z)$ ▷ latent rescaling
$[\hat{z}_i] = \mathrm{ToPatch}(\hat{z}, p, s)$ ▷ split $\hat{z}$ to patches
$z_0\_list = []$ ▷ initialize list of z_0
for $\hat{z}_i$ in $[\hat{z}_i]$ **do**
  $t_i = \mathrm{TPM}(\hat{z}_i)$ ▷ time-step prediction
  $\epsilon_i = \epsilon_\theta(\mathrm{patch}, t_i)$ ▷ noise prediction
  $z_0^i = \mathrm{TS}(\hat{z}_i, \epsilon, t_i)$ ▷ denoising by time scheduler
  $z_0\_list.\mathrm{append}(z_0^i)$
end for
$z_0 = \mathrm{Merge}(z_0\_list)$ ▷ merge patches of z_0
$\hat{x} = \mathcal{D}(z_0)$ ▷ latent decoding
return $\hat{x}, y$ ▷ output

patch size in our work. For the stride length, we adopt the value from previous works [4] and directly set it to 64.

In Fig. 3, we present the time step predictions when employing different patch sizes during the tiled inference process. It can be observed that smaller patch sizes allow for finer-grained time step predictions. However, setting the size too small leads to misalignment with the pre-training image size of Stable Diffusion, resulting in increased computational load and performance degradation. In contrast, our selected patch size of 96 achieves a balance between performance and computational load, while also enabling relatively accurate prediction of the time step mask.

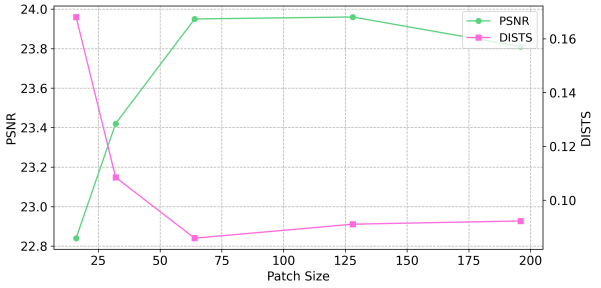

Figure 2. Ablation study about patch size.

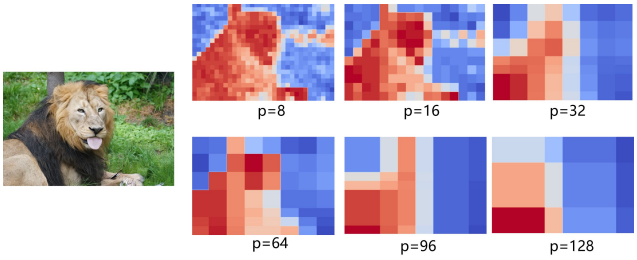

Figure 3. The time step predictions at different patch sizes..

3. Supplementary Results of Ablation Study

3.1. Image Rescaling in the Pixel Space

In the main paper, we compare the performance between rescaling operations performed in the pixel space and latent space. Here, we provide a more detailed comparison of the two approaches, as shown in Fig. 4. Specifically, the rescaling operator in pixel space takes HR image as input and simultaneously outputs both the LR image and the rescaled image. Then, to use SD for perceptual enhancement, we need to map the rescaled image to the latent space using a VAE encoder and perform the denoising process. Finally, the enhanced latent features are mapped back to the pixel space using the VAE decoder.

However, due to the nonlinear mapping nature of the

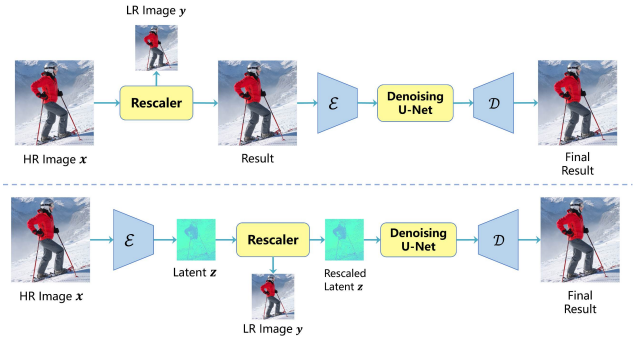

Figure 4. Image rescaling within pixel space and latent space with Stable Diffusion (SD) prior.

VAE encoder, the minimum distance in pixel space does not correspond to the minimum distance in latent space. This causes the rescaling operator trained in pixel space to misalign with the prior of the pre-trained SD model, resulting in a degradation of perceptual quality. As shown in Fig. 5, we conduct visual comparisons between rescaling in latent space and pixel space in two different scenarios. Our latent-space rescaling method is able to reconstruct more realistic textures and more accurate structural features. This indicates that latent-space rescaling operation can preserve sufficiently fine-grained contextual information about the HR image, such as structure, texture, and semantics.

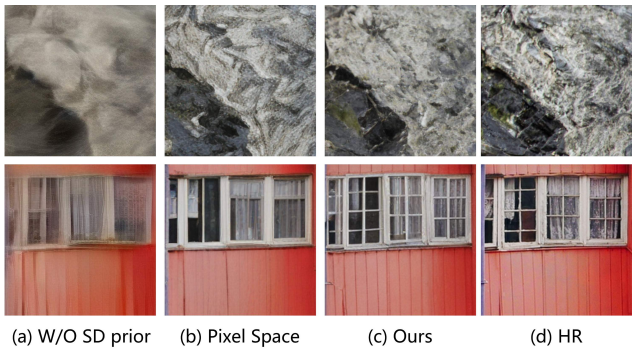

Figure 5. Visual comparisons of our method with rescaling in pixel space and without the SD prior.

3.2. Effectiveness of employing the SD Prior

In the main paper, we validate the effectiveness of introducing the SD prior. Here, we compare the latent features before and after perceptual enhancement by decoding them to the pixel space using the VAE decoder, as shown in Fig. 5. It can be seen that the latent features before perceptual optimization exhibit noticeable color shifts and blurriness when mapped to the pixel domain. In contrast, the perceptually enhanced latent features display richer textures and more accurate color information. Therefore, leveraging the

Table 1. Quantitative comparisons with different methods at 16× and 32×. 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×
Urban100	ESRGAN	19.36	17.68	0.4835	0.4344	0.4741	0.6142	0.2554	0.4288	65.82	57.02	0.5849	0.3860
	StableSR	19.93	18.08	0.5020	0.4447	0.5827	0.6839	0.3701	0.4734	41.97	24.80	0.3439	0.4173
	S3Diff	17.60	15.74	0.4368	0.3419	0.4204	0.5282	0.1917	0.2573	71.01	69.56	0.6658	0.6776
	IRN	22.15	18.93	0.6155	0.4772	0.5114	0.6567	0.3338	0.4457	58.73	32.03	0.3217	0.2210
	HCFlow	**22.59**	**19.86**	0.6335	**0.5185**	0.4841	0.6048	0.3125	0.4101	60.62	47.94	0.3182	0.2827
	VQIR	20.27	18.44	0.5782	0.4427	0.3038	0.5045	0.1418	0.3290	71.56	61.64	0.6619	0.6568
	Ours	21.46	18.84	**0.6495**	0.4685	**0.2577**	**0.4481**	**0.1154**	**0.2386**	**72.22**	**72.88**	**0.7155**	**0.6795**
DIV8K	ESRGAN	25.49	23.78	0.6247	0.6350	0.4628	0.5610	0.2518	0.4268	53.49	39.11	0.6401	0.4078
	StableSR	25.90	23.55	0.6602	0.6258	0.4844	0.5497	0.2134	0.2623	47.15	46.97	0.4017	0.3557
	S3Diff	22.20	20.19	0.5903	0.5063	0.4163	0.4954	0.1524	0.1997	**59.98**	62.58	0.6351	0.6901
	IRN	28.95	25.46	0.7402	0.6636	0.4932	0.5870	0.3044	0.4066	42.61	27.76	0.3216	0.2741
	HCFlow	**29.22**	**26.69**	**0.7470**	**0.6857**	0.4793	0.5616	0.2932	0.3906	42.71	35.53	0.2948	0.3128
	VQIR	25.79	24.26	0.6889	0.6350	0.3221	0.4457	0.1066	0.2628	56.76	51.44	0.5769	0.6203
	Ours	26.16	24.37	0.7163	0.6337	**0.3117**	**0.4358**	**0.0979**	**0.1888**	58.54	**63.34**	**0.6836**	**0.7229**

rich natural image priors in the pre-trained Stable Diffusion model can significantly enhance the visual quality of rescaled images.

3.3. Effectiveness of employing the Pixel Guidance

In the main paper, we demonstrate the necessity of employing the pixel guidance module for improving the quality of LR images. Ideally, the domain converter composed of a set of invertible neural networks (INN) should be able to accurately perform bidirectional mapping between the feature domain and the pixel domain, enabling the LR images to share the same content with HR images. However, due to the limited representation power of INN, the intermediate features z_{lr} obtained by simply downscaling the latent code z are not well-suited for conversion to the pixel domain. The introduction of pixel guidance allows for embedding pixel-domain information into the intermediate features z_{lr}, thereby achieving a balance between feature reconstruction performance and LR image quality. As shown in Fig. 6, the LR images exhibit significant noise and color distortions in the absence of pixel guidance. However, the introduction of pixel guidance significantly alleviates these issues.

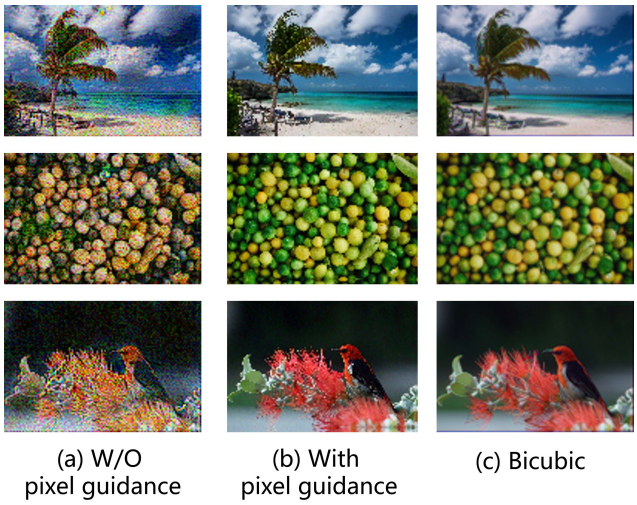

Figure 6. Visual comparisons of LR images with and without pixel guidance.

4. Supplementary Comparisons

4.1. Results on Urban100 and DIV8K

In Table 1, we present the quantitative results on the Urban100 [2] and DIV8K [1] datasets. It can be observed that our method demonstrates significant advantages in perceptual metrics. On the 16× rescaling task of the Urban100 dataset, our method not only leads in perceptual quality but also achieves fidelity comparable to regression-based methods (IRN [7], HCFlow [3]). On the DIV8K dataset, our approach attains the highest fidelity among all generative methods and achieves the best performance in almost all perceptual metrics.

Here, we provide a visual comparison of 16× rescaling results on the Urban100 dataset, as shown in Fig. 7. Traditional regression models, such as IRN [7] and HCFlow [3], tend to generate overly-smooth results. GAN-based methods, including ESRGAN [5] and VQIR [6], generate more details, but their edge information lacks accuracy and regularity. For diffusion-based super-resolution methods, such as S3Diff [8], they are capable of generating sharp edges. However, due to the lack of information about downscaling, their results often exhibit lower fidelity, which manifests in distorted structures. In contrast, our approach, by leveraging the prior knowledge encapsulated in the SD model, is

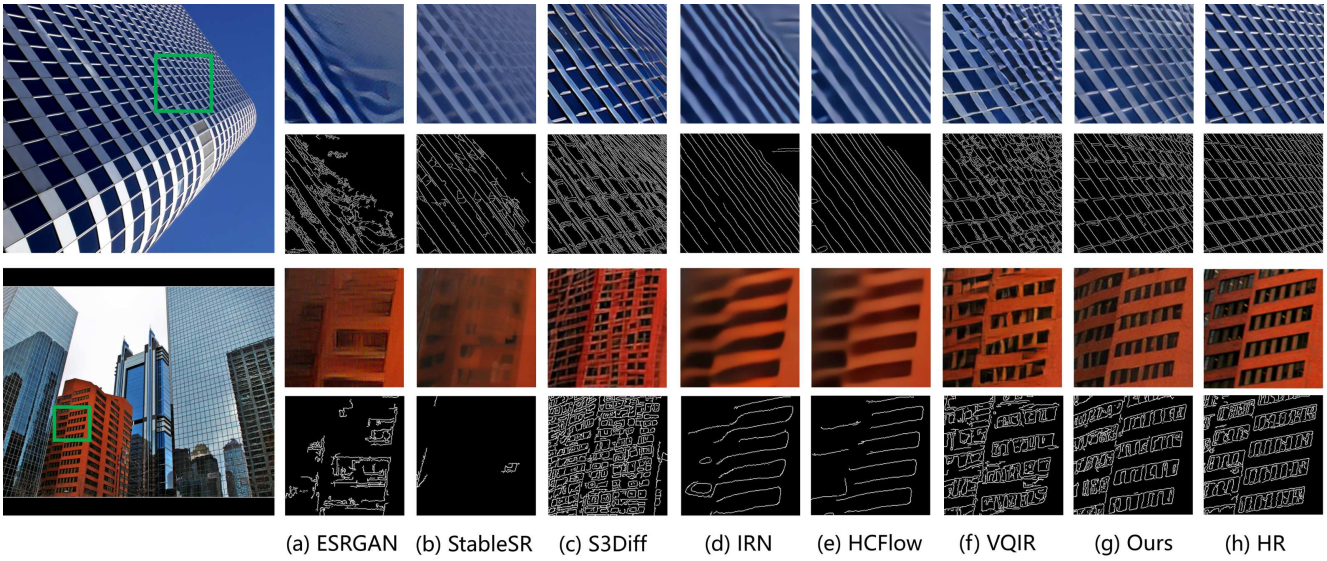

Figure 7. Visual comparisons of 16× rescaling methods on the Urban100 dataset, including both the rescaled images and corresponding edge maps. Our model is capable of restoring more regular structures and producing sharper, more accurate edges.

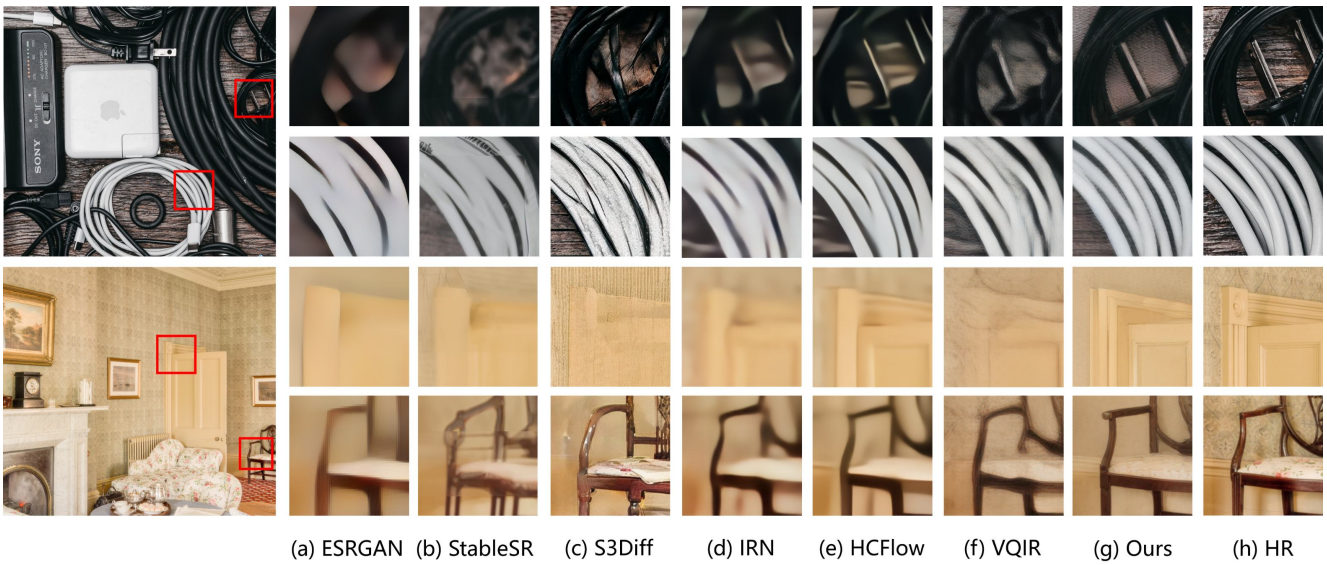

Figure 8. Visual comparisons of 32× rescaling methods on the DIV8K dataset. Our approach can still reconstruct correct semantic information even in such extreme scenarios.

able to reconstruct more accurate and rich edge details. In Fig. 8, we present a visual comparison of 32× rescaling results on the ultra-high-resolution dataset, DIV8K. It is evident that our method is capable of recovering images with accurate semantics, even at extreme rescaling factors. For instance, our method is capable of accurately reconstructing electronic devices, doors, and chairs, while retaining rich details.

4.2. Qualitative Results of LR Image

To validate the effectiveness of our downscaling scheme, we conduct a visual comparison of the downscaled LR images with state-of-the-art (SOTA) methods, as shown in Fig. 9. Overall, the LR images generated by our method achieve comparable or even superior visual quality to SOTA image rescaling methods. The LR images produced by VQIR and HCFlow exhibit noticeable color shifts. Furthermore, VQIR generated LR images contain significant noise, whereas our method exhibits a certain degree of ringing artifacts. In fu-

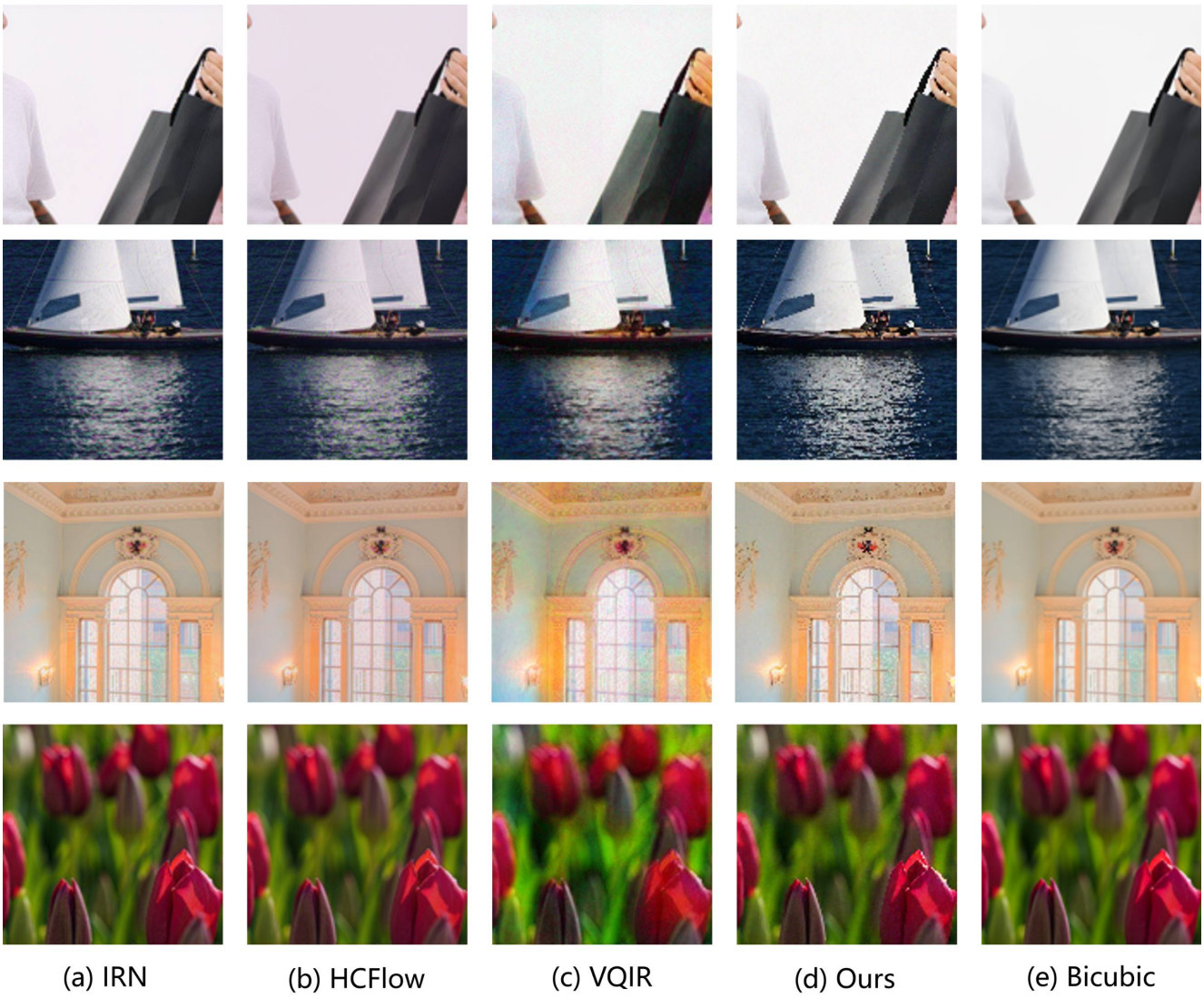

Figure 9. Visual comparisons of downscaled LR images with SOTA methods.

ture work, we plan to further optimize the quality of the LR images.

References

[1] 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. IEEE, 2019. 3

[2] 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. 3

[3] 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. 3

[4] 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. 1, 2

[5] 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. 3

[6] 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*, 2024. 3

[7] 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. 3

[8] 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. 3