[Supplementary Material] AdaDiffSR: Adaptive Region-aware Dynamic Acceleration Diffusion Model for Real-World Image Super-Resolution

Yuanting Fan[1,2*], Chengxu Liu[2,3*], Nengzhong Yin[2], Changlong Gao[2], and Xueming Qian[2,3✉]

[1] School of Software Engineering, Xi'an Jiaotong University, Xi'an, China
[2] Xi'an Jiaotong University, Xi'an, China
[3] Shaanxi Yulan Jiuzhou Intelligent Optoelectronic Tech. Co., Ltd, Xi'an, China
{retofan,chengxuliu,ynz0608,gaochanglong}@stu.xjtu.edu.cn,
qianxm@mail.xjtu.edu.cn

In this supplementary material, Sec. 1 illustrates the details about the selection of the IQA metrics and the structure of the regressor in MMLE. Sec. 2 describes more details of region integration strategy and corresponding visualization results. Finally, Sec. 3 shows more visualization results of AdaDiffSR.

1 Details of MMLE

1.1 The selection of IQA metrics

As mentioned in Sec. 3.2 in the main paper, we select four FR-IQA metrics and two NR-IQA metrics to measure the information gain from six perspectives, including PSNR, LPIPS [8], AHIQ [3], NLPD [7], BRISQUE [4], and MUSIQ [2]. In this section, we further analyze the mutual interaction of IQA metrics within a single category on the DIV2K validation set [6]. As shown in Tab. 1, we conduct ablation studies on the impact of different NR-IQA metrics during the denoising process. Only with a single NR metric to guide the denoising process, the image achieve sup-optimal realism. As the number of NR metrics increases, the reconstruction performance improves slightly while the training cost increases considerably. Moreover, the specific categories of IQA metrics have no significant impact on the restoration quality. Therefore, to achieve a better trade-off between the efficiency and capacity of the MMLE regressor, we select the above four FR-IQA metrics and two NR-metrics.

1.2 The structure of MMLE regressor

As mentioned in Sec. 3.2 in the main paper, we propose a lightweight regressor to estimate the multi-dimensional information gain. The network structure of the regressor is mainly based on the lightweight network MobileNetV3 [1], and we

* Equal contribution.

Table 1: Ablation studies of NR-IQA metrics in MMLE on DIV2K validation set [6]. The best and second-best performance are in red and blue color, respectively.

Exp.	Components						PSNR↑	SSIM↑	LPIPS↓
	MANIQA	MUSIQ	DBCNN	NRQM	NIQE	BRISQUE			
(a)	✓						23.46	0.6941	0.2458
(b)			✓				23.24	0.7013	0.2436
(c)	✓	✓					24.19	0.7356	0.2197
(d)		✓				✓	24.25	0.7355	0.2153
(e)	✓		✓		✓		24.29	0.7344	0.2146
(f)		✓	✓	✓			24.17	0.7378	0.2209
(g)		✓	✓	✓	✓		24.22	0.7376	0.2175
(h)	✓	✓	✓	✓			24.11	0.7359	0.2155

modified the average pooling layers to match our multi-dimensional information gain which can be interpreted as a six-dimensional classification problem, and the output will be normalized between zero and one. Then we calculate the multi-dimensional information gain via the multi-dimensional representation score of the input features. The input and output of the regressor are current timestep latent feature and information gain, respectively. We decode this latent feature using VAE decoder, as mentioned in Sec. 4.4 in the main paper, we use PY-IQA toolbox on the decoded feature to get GT information gain. The regressor achieves 97.3% and 96.1% (when the difference less than $\frac{\tau}{15}$) accuracy on RealSR and DRealSR dataset, making the DTSS strategy effective.

2 Details of region integration strategy

2.1 The algorithm of region integration strategy

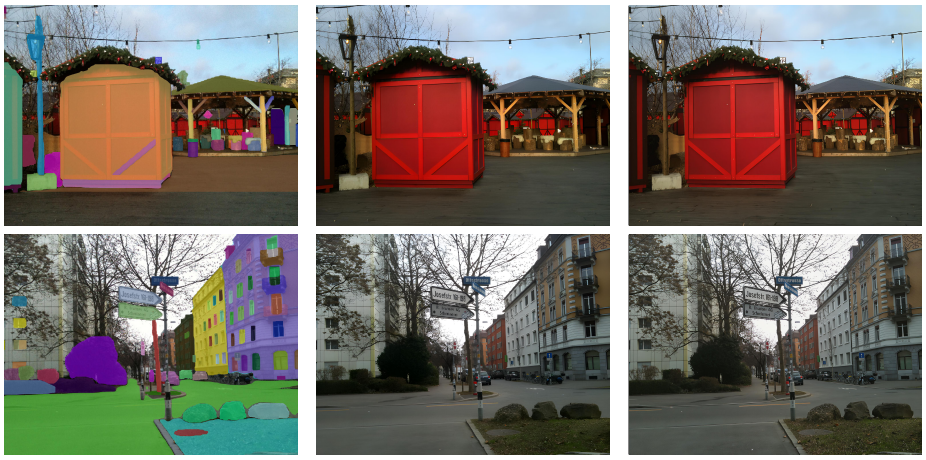
As mentioned in Sec. 3.4 in the main paper, we adopt the region integration strategy to eliminate the discontinuities in the boundaries of different image regions. Here, we provide more details about this strategy. First, we use VAE within the Stable Diffusion [5] to downsampling the original input image into the latent feature with less spatial size, and crop this latent feature into overlapping regions with 64×64 spatial resolution. Then, we generate the Gaussian weight map with 64×64 resolution using the Gaussian filter, and use this weight map to aggregate these regions into result, the specific calculation process is as follows:

$$\hat{O} = \sum_{i=1}^{N} \frac{\omega}{\bar{\omega}} \times o_i, \tag{1}$$

where o_i and $\hat{O}$ indicate the latent feature of the particular region and entire image. ω represents the Gaussian weight map, and $\bar{\omega} = \sum_i \omega$. Note that we only conduct this strategy for images with resolution larger than 512×512, and adopt this strategy only in the first and final denoising step to reduce the computational resource overhead.

2.2 The visual comparisons between different slicing strategy

As mentioned in Sec. 4.4 in the main paper, we conduct the ablation studies to demonstrate the effectiveness of the static slicing strategy, which slices the input LR images into multiple overlapping regions with the fixed resolution (*i.e.*, 512×512) as the pre-trained DMs. In this section, to compare the differences between different slicing methods more intuitively, we further show more visualization results. As shown in Fig. 1, we visualize the reconstruction results, it can be noticed that using the super-pixel segmentation methods or not have a significant influence on the restored results, since the fact that the feature variations in different image regions can adaptively adjust the timesteps during the reconstruction process, thus achieving more precise and efficient reconstruction. Besides, we also show the equivalent timesteps of different image regions during the reconstruction process, as shown in Fig. 2. The foreground and background regions are already well distinguished by the information gain. Furthermore, to validate the boundary continuity of static slicing and integration strategy, we show boundaries visualization case in Fig. 3. With the integration strategy, the discontinuity of the slicing boundaries disappeared significantly.



a) FastSAM b) Ours(W/ FastSAM) c) Ours(W/O FastSAM)

Fig. 1: The visualization comparisons between the super-pixel segmentation method and the static slicing strategy.

3 Additional Qualitative Results

In this section, to demonstrate the effectiveness of AdaDiffSR, we show more qualitative results on real-world images. We follow Sec. 4.3 of the main paper

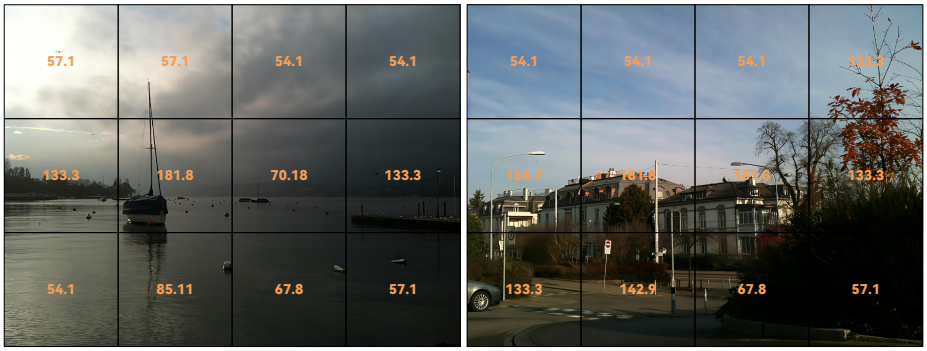


Fig. 2: The visualization results of Equivalent timesteps in different image regions.

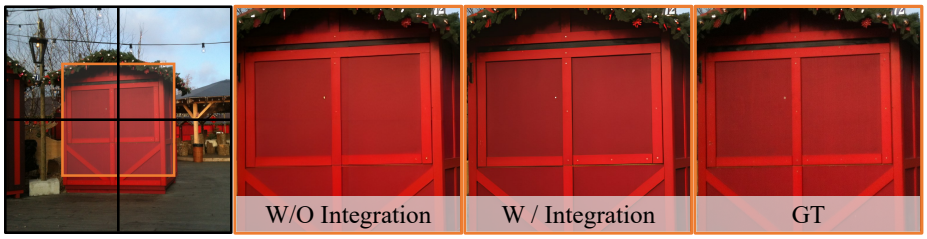


Fig. 3: The boundaries visualization comparisons with and without the region slicing and integration strategy.

to report the visual results on static input resolution (*i.e.*, 512×512) and arbitrary resolution, respectively. As shown in Fig. 4 and Fig. 5, AdaDiffSR achieves realistic details and sharp edges, significantly outperforms existing methods.

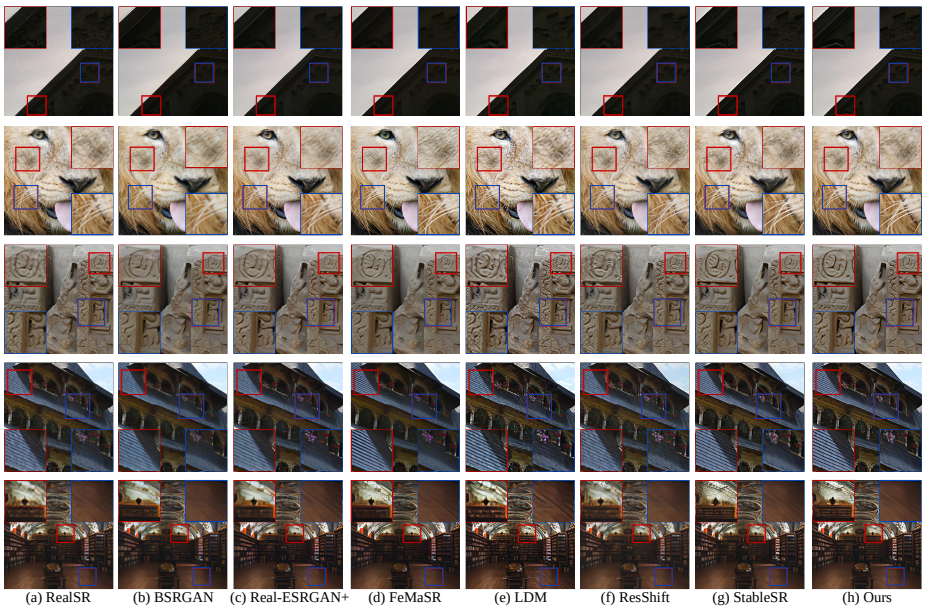


Fig. 4: More qualitative comparisons on real-world images (static input resolution as 512×512). Zoom in for the best view.

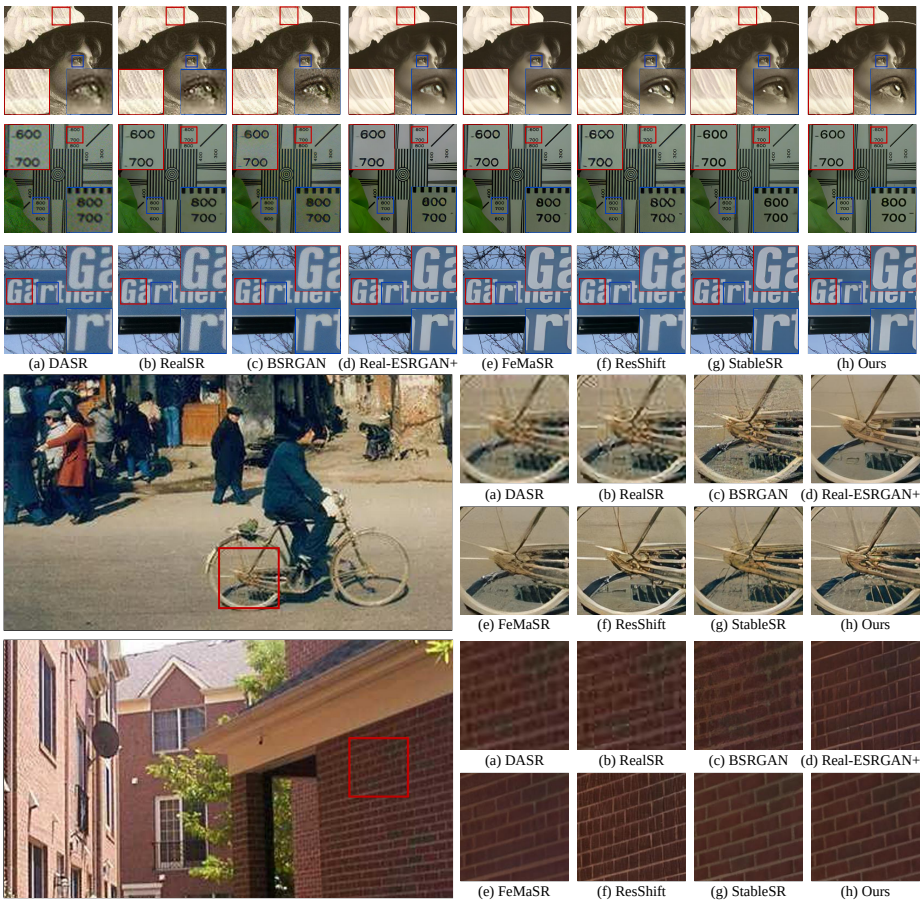


Fig. 5: More qualitative comparisons on real-world images (arbitrary input resolution). Zoom in for the best view.

References

1. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: ICCV. pp. 1314–1324 (2019)
2. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: ICCV (2021)
3. Lao, S., Gong, Y., Shi, S., Yang, S., Wu, T., Wang, J., Xia, W., Yang, Y.: Attentions help cnns see better: Attention-based hybrid image quality assessment network. In: CVPRW (2022)
4. Mittal, A., Moorthy, A.K., Bovik, A.C.: Blind/referenceless image spatial quality evaluator. In: ASILOMAR. pp. 723–727 (2011)
5. Rombach, R., Blattmann, A., etal.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
6. Timofte, R., Agustsson: Ntire 2017 challenge on single image super-resolution: Methods and results. In: CVPRW (2017)
7. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: CVPRW (2022)
8. Zhang, R., Isola, P., Efros, etal.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)