Supplemental Material: Real-SRGD: Enhancing Real-World Image Super-Resolution with Classifier-Free Guided Diffusion

Kenji Doi, Shuntaro Okada, Ryota Yoshihashi, and Hirokatsu Kataoka

LY Corporation

A Model Architecture

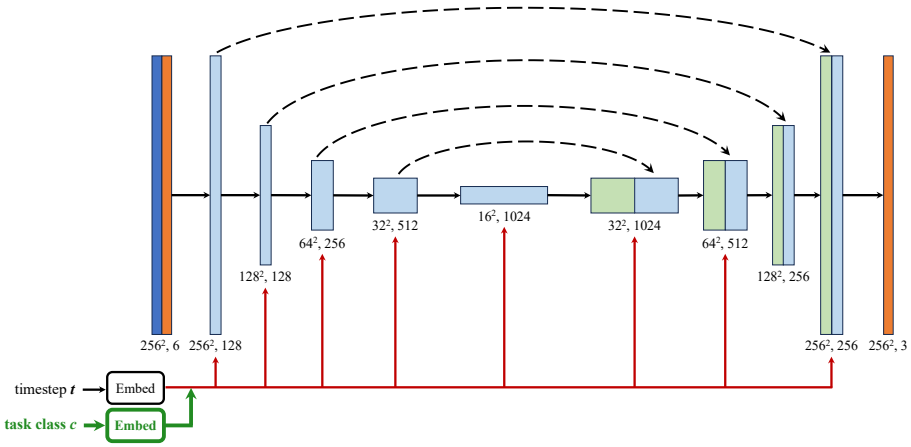

Fig. A: U-Net architecture utilized in our Real-SRGD. Channel Dim = 128, Depth Multipliers = {1,2,4,8}, ResNet Blocks = 2, input resolution = 256×256. Full attention is applied to the deepest ResNet block with timestep and task class embeddings, while linear attention is used for the other ResNet blocks.

Fig. A provides an overview of the U-Net [11] architecture used in our Real-SRGD, which follows the model referenced in DDPM [7]. We set the dimension of the first U-Net layer to 128, and the depth multipliers for U-Net are {1,2,4,8}. The number of ResNet Blocks is 2. In addition, an attention mechanism is applied to the output of each layer's ResNet block. This mechanism involves the use of timestep and task class embeddings. Full attention is directed toward the deepest layer, while linear attention is applied to the remaining layers. The model accepts an input resolution of 256×256 pixels. Low-resolution images are upscaled to this input resolution through Bicubic interpolation before being concatenated with noisy, high-resolution images, which then serve as the model's inputs.

B Exploring the Settings for Effective Diffusion Models in RISR

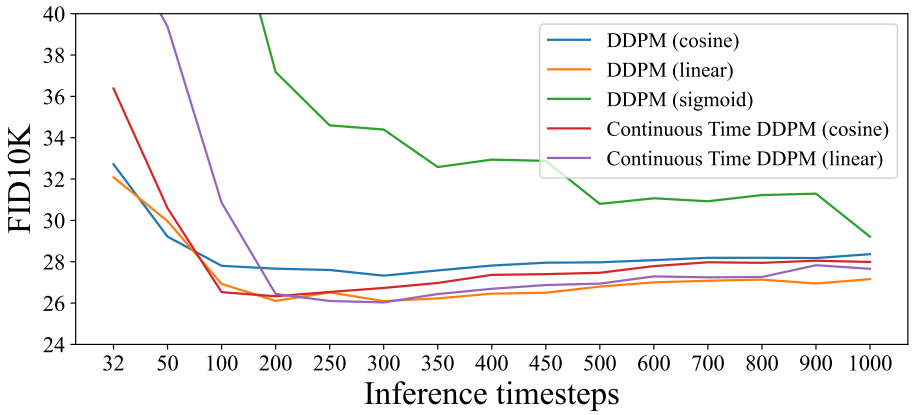

Fig. B: FID10K among our models at various inference timesteps.

In our quest for the optimal training methodology for diffusion models suitable for real-world image super-resolution (RISR), we evaluated the performance of models trained under a broad array of settings. More specifically, we assessed combinations of diffusion model training methods (DDPM [7] and EDM [8]), training timesteps (T=1000, or continuous timesteps), and noise scheduling methods (Linear, Cosine, Sigmoid). Like many existing methods, we implemented a super-resolution enlargement factor of four.

We trained the models using the DIV2K [9], Flickr2K [12], and OST [15] datasets We set the initial layer of U-Net to 64 dimensions, defined the input image size as 256×256 pixels, configured the training model's EMA decay to be 0.995, and limited the training iterations to $40K$. In addition, during this preliminary screening, we trained super-resolution diffusion models considering only LR images as conditions, instead of task-conditional diffusion models.

Fig. B presents the FID10K scores of the RealSRv3 [1] dataset at various inference timesteps. In the legend of Fig. B, parentheses indicate the type of noise scheduling. We found that the sigmoid noise scheduling generally results in poorer FID10K scores, with only moderate improvements as the number of inference timesteps increases. On the other hand, for the cosine and linear noise schedules, the best scores are achieved around 200 to 300 inference timesteps, after which there seems to be no improvement and even a slight worsening trend. Following preliminary experiments, we adopted two models: the continuous time DDPM with a linear noise schedule, hereafter called CDM, and the EDM model.

Furthermore, we conducted a thorough examination into the impacts of the U-Net dimensions (64 or 128 dim), the effects of EMA decay, and the benefits

of supplementing the training data with the DIV8K [6] dataset. Based on these investigations, we resolved to increase the dimension of the first U-Net layer to 128, adjust the EMA decay rate to 0.9999, and change our training dataset to be DIV2K, DIV8K, Flickr2K, and OST datasets. These revised EDM and CDM models are the ones employed in our study.

C Training and Inference Details

C.1 Training Details

We trained our Real-SRGD models using AdamW with L2 loss as the loss function. Training commences with a learning rate of 1×10^{-4}, incorporating a linear warmup for the initial 4K iterations, and afterwards employs a cosine annealing scheduler that reduces the learning rate from 1×10^{-4} to 1×10^{-7} over the course of total 40K iterations.. We adopt an exponential moving average (EMA) with decay = 0.9999 for more stable training and better performance. Our implementation builds upon the DDPM as implemented by Phil Wang [14].

Fig. 2 in our paper provides an overview and a data augmentation pipeline of our proposed method. In the case of task conditions pertaining to RISR, our Real-SRGD employs the same data augmentation pipeline as used in Real-ESRGAN. However, when the task conditions are related to BIR or SR, our implementation modifies the data augmentation pipeline to include only the elements relevant to the respective tasks.

The LR image's condition are concatenated with input noise along the channel axis before inputting them into the model. Simultaneously, the task-class condition is encoded into an embedding, similar to the timestep embedding, and is incorporated into the model's residual blocks. During inference, the LR condition remains unchanged, whereas the noise is iteratively updated.

Furthermore, to enable classifier-free guidance after the model's training, we conduct learning without inputting the task-class condition to the model 10% of the time.

C.2 Inference Details

Based on the results presented in Fig. B, we selected 250 generation timesteps for model inference because it provided an optimal balance between quality and efficiency for our CDM model. The EDM model operates with a setting of 32 steps for both training and generation. This model serves as a lightweight alternative, providing a more streamlined generation process compared to the continuous-time DDPM model. Furthermore, to generate samples, the EDM model leverages the DPM++ sampler [10]. This approach achieves generation speeds approximately twice as fast as those in its original configuration.

Table A: Impact of generation steps on DDPM cosine noise schedule

inference timesteps	RealSRv3			DRealSR		
	FID10K ↓	NIQE ↓	MANIQA ↑	FID10K ↓	NIQE ↓	MANIQA ↑
1	287.29	30.859	0.2224	297.96	32.780	0.2294
2	273.74	16.010	0.3407	285.33	16.810	0.3613
4	207.76	17.820	<u>0.4582</u>	229.41	19.288	<u>0.4639</u>
8	111.48	19.123	**0.4652**	135.41	20.167	**0.4728**
16	52.97	12.647	0.4258	67.02	13.755	0.4244
32	32.72	9.357	0.3803	38.52	9.911	0.3687
50	29.22	8.742	0.3674	32.30	9.186	0.3514
100	27.80	8.128	0.3584	29.22	8.426	0.3391
200	<u>27.66</u>	7.833	0.3538	27.82	7.904	0.3319
250	27.60	7.699	0.3503	27.38	7.909	0.3303
300	**27.33**	7.658	0.3495	27.28	7.670	0.3275
400	27.82	7.549	0.3470	27.24	7.695	0.3261
500	27.97	7.578	0.3475	27.41	7.577	0.3252
600	28.08	7.436	0.3461	**26.95**	7.653	0.3243
700	28.19	7.414	0.3454	27.23	<u>7.497</u>	0.3241
800	28.19	<u>7.411</u>	0.3461	<u>27.03</u>	7.577	0.3234
900	28.18	7.439	0.3459	27.08	**7.430**	0.3238
1000	28.36	**7.404**	0.3439	27.06	7.507	0.3231

D Additional Comparisons Between Real-SRGD and Other Models

We present additional side-by-side comparisons of RISR results between our Real-SRGD and other benchmark models. Figs. G, H, and I show comparisons on the RealSRv3 dataset while Figs. J, K, and L show comparisons on the DRealSR dataset. Figs. M, and N represent comparisons on the DIV2K-Wild [13] dataset.

E Comparison of RISR Results with Classifier Guidance (CG) and Classifier-Free Guidance (CFG)

We evaluated our method with classifier guidance (CG) [3] instead of CFG. The implementation of CG was based on the proposed paper [3], utilizing a U-Net-based classifier that was trained to classify images with noise added according to the timestep. Our experiments involved two different training scenarios for the classifier. The first scenario was a binary classification between real images and images degraded for RISR, and the second was a quaternary classification among real images and images with degradations applied according to the three tasks in our proposed method. The validation results showed that the classifier trained in the latter configuration performed better, so we compared the results of CG

using this classifier with our proposed method that employs CFG on RealSRv3 and DRealSR datasets.

We provide side-by-side comparisons of RISR results with CG and CFG. In addition, we will also include the results from the baseline model.

Fig. E shows comparisons on the RealSRv3 dataset while Fig. F shows comparisons on the DRealSR dataset.

For CG, increasing the scale tends to improve perceptual evaluation metrics, although a granular noise often appears in the image. The results when using both CFG and CG tend to be better than using CG alone, but the results of CFG alone often give a stronger impression of higher resolution.

F Impact of Different Task Composition on Performance

In the proposed method, the task classes were one-hot encoded from three classes: RISR/SR/BIR. To examine the impact of different task partitions on performance, we also evaluated a partitioning into five classes. These five classes are Blurring, Resolution change, Noise addition, Compression (JPEG), and RISR, which encompasses all of these. Apart from the task partitioning, the model was trained with the same settings as the CDM model described in the paper, and comparisons were made with the EDM and CDM models. The evaluation data used were the DIV2K-Wild and DPED-iPhone datasets. The model trained with three classes is referred to as EDM3 and CDM3, respectively, and the model trained with five classes is referred to as CDM5. During inference, all models specify the RISR class.

The results are shown in Table B. CDM5 achieves better results in perceptual evaluation metrics such as NIQE, CLIPIQA, and MUSIQ, indicating that the method of task partitioning can also lead to performance differences. A human subject study has not been conducted for the CDM5 model, so qualitative evaluations by humans have not been confirmed.

Additionally, for the CDM3 model, each of the three classes corresponds to existing tasks, providing the advantage of being usable for different tasks per class. However, in the CDM5 model, classes other than RISR correspond to specific degradations, which makes them difficult to use unless the degradation of the input image is known.

G Effect of Classifier-Free Guidance Scale

We compare the effect of the scale of classifier-free guidance (CFG) on image quality in our Real-SRGD (Both EMD and CDM models). Figs. O, P, Q, and R represent comparisons of RISR results on the RealSRv3 and DRealSR datasets, conducted while varying the scale of CFG.

In both models, as the scale increases, there appears to be an improvement in resolution. However, at higher scales, some results exhibit unnatural effects, such as color bleeding. Comparing between EDM and CDM, the resolution seems to be relatively higher using the CDM model.

Table B: Evaluation of Performance Differences Due to Task Decomposition Variations.

Methods	DIV2K-Wild						DPED-iPhone		
	PSNR ↑	SSIM ↑	LPIPS ↓	NIQE ↓	CLIP-IQA ↑	MUSIQ ↑	NIQE ↓	CLIP-IQA ↑	MUSIQ ↑
baseline	17.55	0.4554	0.4035	3.269	0.5881	55.40	3.918	0.3831	45.58
EDM3 ($s = 0$)	16.98	0.3454	0.5630	3.389	0.5080	46.68	3.993	0.2697	34.61
EDM3 ($s = 1$)	16.47	0.3381	0.5092	2.928	0.6373	55.55	3.495	0.3672	46.99
EDM3 ($s = 2$)	15.96	0.3265	0.4951	2.729	0.6844	59.19	**3.213**	0.4322	52.00
CDM3 ($s = 0$)	**17.73**	0.4600	0.4367	3.732	0.4707	47.51	4.148	0.3273	39.81
CDM3 ($s = 1$)	17.00	0.4275	**0.4000**	2.866	0.7125	62.24	3.522	0.4887	53.13
CDM3 ($s = 2$)	16.34	0.3984	0.4183	**2.722**	0.7711	65.85	3.365	0.5757	57.75
CDM5 ($s = 0$)	17.71	**0.4620**	0.4488	4.111	0.4567	45.35	4.690	0.2791	34.92
CDM5 ($s = 1$)	15.82	0.3907	0.4230	2.996	0.7754	66.46	3.315	0.5876	57.20
CDM5 ($s = 2$)	14.34	0.3420	0.4805	3.104	**0.7953**	**67.49**	3.423	**0.6769**	**61.44**

Table C: NIQE of models using CFG under LR Conditions. (Baseline model's NIQE: 3.269 for DIV2K-Wild and 3.918 for DPED-iPhone)

Methods	DIV2K-Wild					DPED-iPhone				
	CFG start step (total 250)					CFG start step (total 250)				
	0	50	100	150	200	0	50	100	150	200
Ours$_{LR}$ ($s = 1$)	3.581	3.534	3.409	3.613	**3.351**	3.461	3.448	**3.346**	3.482	3.639
Ours$_{LR}$ ($s = 2$)	6.220	6.128	5.420	5.052	3.936	4.629	4.631	4.398	4.272	4.069
Ours$_{LR}$ ($s = 3$)	8.598	8.493	7.859	6.206	4.688	5.813	5.806	5.892	5.347	4.760
Ours$_{LR}$ ($s = 4$)	10.744	10.757	10.225	7.286	5.355	6.930	6.914	7.034	6.369	5.446

H LR Condition-Based Classifier-Free Guidance

We also conducted experiments applying CFG using the LR image conditions, instead of the task conditions. We verified starting the CFG from the middle steps of generation as well. The LR image condition provides the foundational information for generating high-resolution images, and particularly in the initial stages of generation where the input is predominantly noise, the model would struggle to effectively denoise without the LR image condition. As a result of auditing multiple starting steps of CFG, as expected, we noted an improvement in results when the CFG was initiated from the mid-generation steps, compared to applying the CFG from the beginning.

Table C shows the NIQE scores for the DIV2K-Wild and DPED-iPhone datasets. The NIQE scores achieved with CFG, using the LR condition, are worse than those of the baseline model, with the best scores being 3.351 for DIV2K-Wild and 3.346 for DPED-iPhone. Our experimental results confirmed that the application of CFG under task conditions resulted in superior quality of the generated images compared to the application of CFG under the LR image conditions.

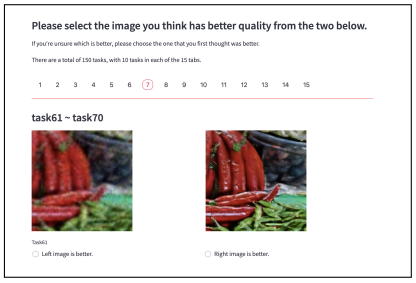

Fig. C: User interface for image comparison and selection.

Table D: Glicko-2 [4] rating results in human subject study.

Methods	Rating Score ↑	Deviation ↓	Volatility ↓
Ours$_{CDM}$ ($s = 2.0$)	**1806.58**	69.103	0.06000
Ground Truth	<u>1742.88</u>	66.928	0.06006
Ours$_{CDM}$ ($s = 1.0$)	1692.10	66.305	0.05996
Ours$_{EDM}$ ($s = 2.0$)	1673.95	65.498	0.06003
Real-ESRGAN+	1598.33	**62.757**	0.05998
Ours$_{EDM}$ ($s = 1.0$)	1574.03	64.869	0.06000
SwinIR-GAN	1541.21	64.869	0.05998
FeMaSR	1522.70	<u>64.387</u>	0.05997
StableSR	1495.92	64.935	0.05999
RealDAN	1424.05	64.968	0.05997
Ours$_{CDM}$ ($s = 0.0$)	1418.40	65.821	0.06003
DiffBIR	1412.07	65.054	0.06007
Ours$_{EDM}$ ($s = 0.0$)	1326.28	68.923	0.06000
Swin2SR	1271.22	68.543	**0.05993**
Bicubic	839.47	100.460	<u>0.05993</u>

I Supplementary Details of Human Subject Study

In our human subject study, we used a pairwise comparison strategy wherein participants were presented with pairs of images and asked to indicate their preference on the basis of the perceived image quality. To easily distinguish the differences when the subjects compared the images, we used the top 2,000 images with the largest variance after super-resolution from the 10,000 RealSRv3 and DRealSR images. For the DIV2K-Wild dataset, we used a 256×256 crop with the maximum variance between methods from each of the 100 samples after super-resolution.

We developed a web-based system that randomly presented pairs of image regions for users to choose from. Fig. C illustrates the interface used by the participants in the study. We asked 14 participants to select the image they thought had better quality from each of 150 pairs and thus acquired a total of 2,100 voting results. When randomly selecting any two methods from a total of 15, there were a total of 105 pairs of method comparisons. Each pair of methods was evaluated on average about 20 times.

In addition to Elo, we also tried the Glicko-2 rating system [4] for calculating ratings from the pairwise comparison results. Like Elo, Glicko-2 is often used for player evaluations in paired competitive games, but unlike Elo, it also considers the reliability of the ratings, and is also used for the quality evaluation of super-resolution models [2].

The Glicko-2 rating system estimates three parameters: Rating score (μ), Rating Deviation (ϕ), and Volatility (σ), and the final rating is in the range $(\mu - 2\sigma, \mu + 2\sigma)$ with 95% confidence. Consequently, higher-ranked methods are those with a high score and low uncertainty.

Final ratings are shown in Table D. The ranking according to the Glicko-2 score was identical to the top 5 ranking according to the Elo score. There was some shuffling of ranks below the 6th place, but overall, the trend was similar to the Elo results. Considering the reliability of the Glicko-2 rating, the CDM

model (with $s = 2$) significantly surpassed the top-rated existing method, which is Real-ESRGAN+, with a considerable margin in the confidence interval.

J Disparity Analysis between Perceptual Image Quality and MANIQA Scores

We also explored the use of the perceptual quality metric MANIQA [16], which won the NTIRE 2022 Perceptual Image Quality Assessment Challenge [5].

Fig. D presents comparisons between RISR results and MANIQA scores, with the MANIQA score for each image indicated in parentheses. Despite the perceived image quality, our method achieves a higher MANIQA score at step=8 (0.6473) compared to step 250 (0.4292).

In our proposed method, as the number of generation steps increases to 8, 100, and 250, the perceptual quality of the image improves. Nevertheless, in terms of MANIQA scores, a paradox arises where a fewer number of generation steps yield higher scores, despite the perceptual image quality. This trend is not limited to the samples in Fig. D; it is also apparent when evaluating the average of a large number of images. As observed in the MANIQA scores presented in Table A, a similar trend ensues.

Although MANIQA is recognized for its ability to produce scores that correlate with human subjective quality, as evidenced by its success in the NTIRE 2022 competition [5], the disparity between perceptual image quality and the scores shown here may be occurring due to a lack or insufficiency of training data containing images with unique quality characteristics, such as the spotty noise seen in the images in Fig. D. Images featuring noise with such unique quality characteristics might be scarce in the real world. Therefore, it's worth considering the inclusion of images generated by the diffusion model in the training data when designing learning-based evaluation metrics.

K Details of Comparative Models and Their Checkpoints

This section provides further details about the implementation and checkpoint files of the comparative models used in our experiments.

For RISR

- **Real-ESRGAN+**
  - `https://github.com/xinntao/Real-ESRGAN.git`
  - `RealESRGAN_x4plus.pth`
- **SwinIR-GAN**
  - `https://github.com/JingyunLiang/SwinIR.git`
  - `003_realSR_BSRGAN_DFO_s64w8_SwinIR-M_x4_GAN.pth`
- **FeMaSR**
  - `https://github.com/chaofengc/FeMaSR.git`

  - `FeMaSR_SRX4_model_g.pth`
- **StableSR**
  - `https://github.com/IceClear/StableSR.git`
  - `vqgan_cfw_00011.ckpt`
  - `stablesr_000117.ckpt`
- **RealDAN**
  - `https://github.com/greatlog/RealDAN.git`
  - `RealDAN_GAN.pth`
- **DiffBIR**
  - `https://github.com/XPixelGroup/DiffBIR.git`
  - `general_swinir_v1.ckpt`
  - `general_full_v1.ckpt`
- **Swin2SR**
  - `https://github.com/mv-lab/swin2sr.git`
  - `Swin2SR_RealworldSR_X4_64_BSRGAN_PSNR.pth`

For Conventional SR

- **SwinIR**
  - `https://github.com/JingyunLiang/SwinIR.git`
  - `001_classicalSR_DF2K_s64w8_SwinIR-M_x4.pth`
- **HAT**
  - `https://github.com/XPixelGroup/HAT.git`
  - `HAT_SRx4_ImageNet-pretrain.pth`

Fig. D: Comparisons between RISR results and MANIQA scores. Despite the perceived image quality, our method achieves a higher MANIQA score at step=8 (0.6473) compared to step=250 (0.4292)

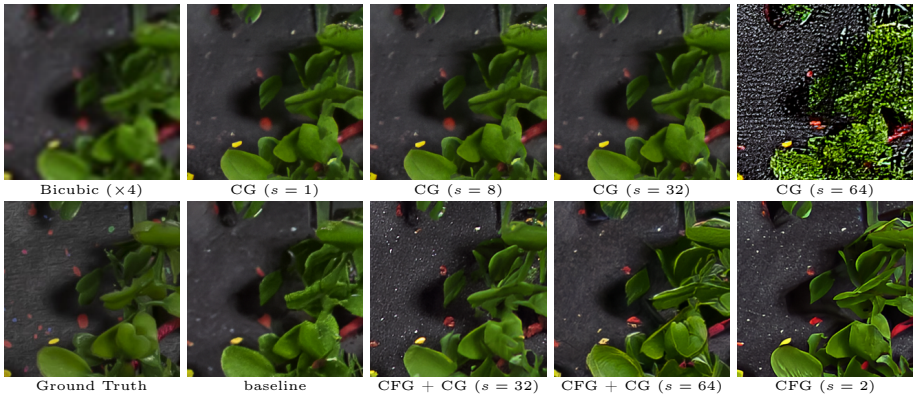

Fig. E: Comparison CG ($s = 1, 8, 32, 64$) and CFG ($s = 2$) on RealSRv3 results. (crop from `Nikon_172_LR4.png`)

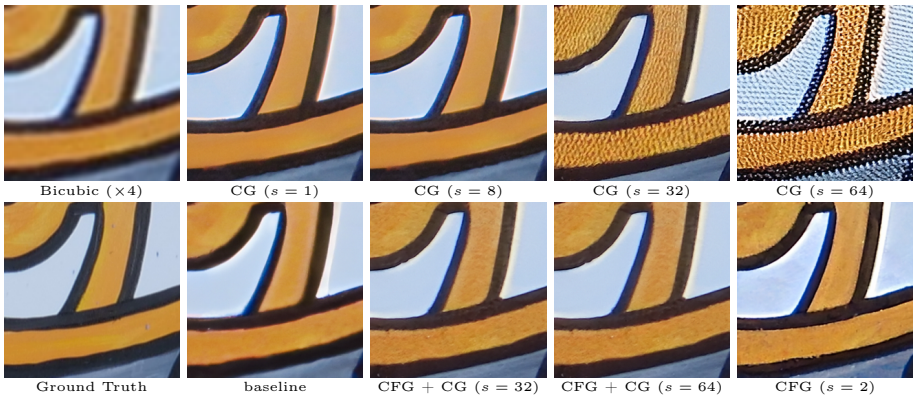

Fig. F: Comparison CG ($s = 1, 8, 32, 64$) and CFG ($s = 2$) on DRealSR results. (crop from `panasonic_131_x1_44.png`)

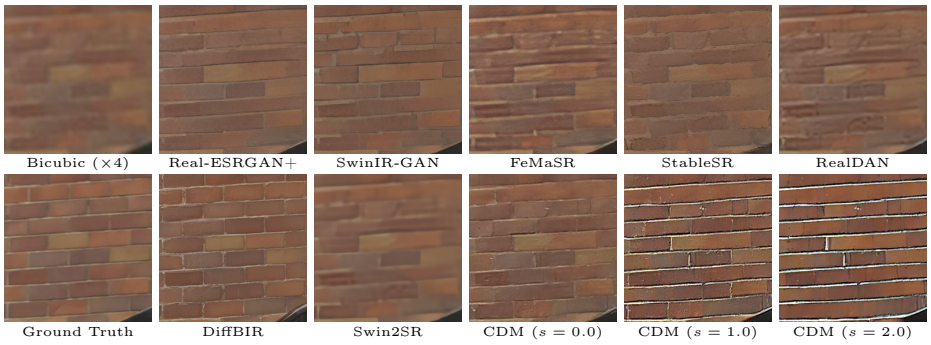

Fig. G: Comparison of RealSRv3 results. (crop from `Canon_180_LR4.png`)

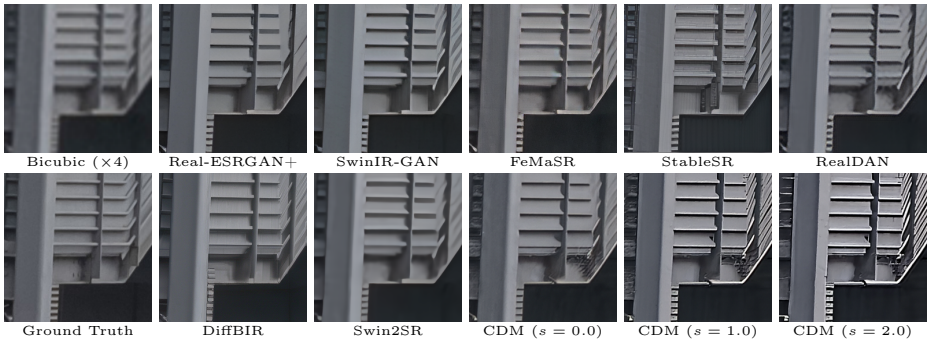

Fig. H: Comparison of RealSRv3 results. (crop from `Nikon_089_LR4.png`)

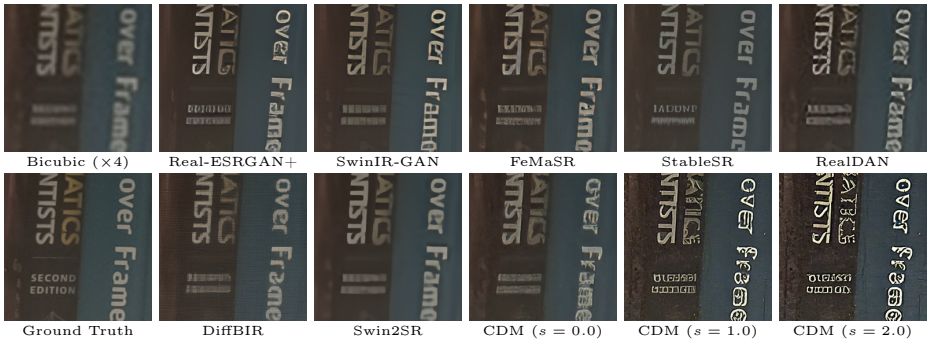

Fig. I: Comparison of RealSRv3 results. (crop from `Nikon_103_LR4.png`)

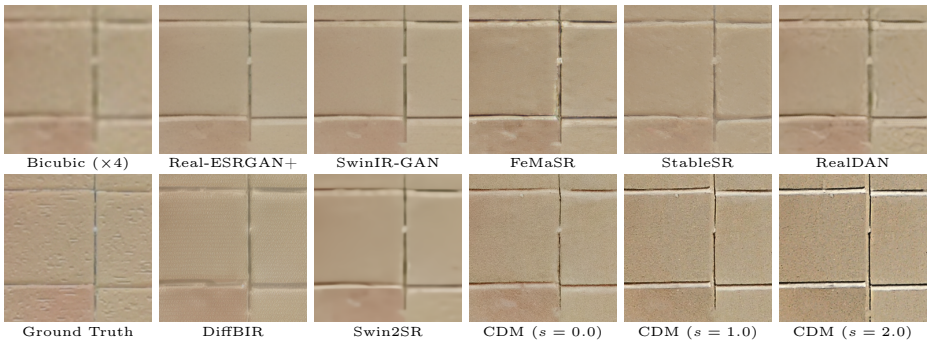

Fig. J: Comparison of DRealSR results. (crop from `DSC_1514_x1_33.png`)

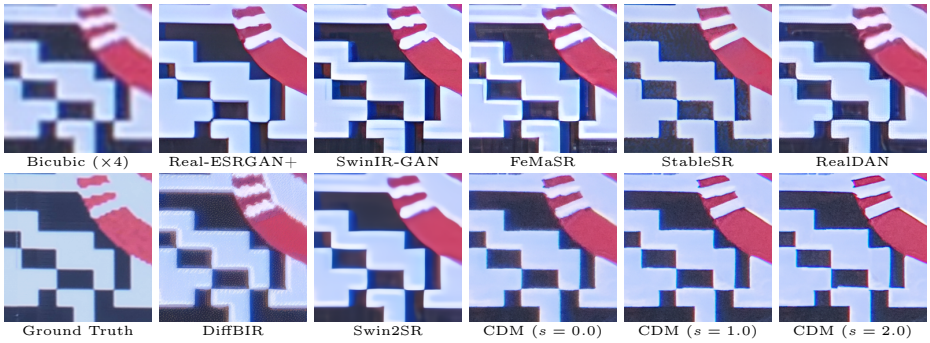

Fig. K: Comparison of DRealSR results. (crop from `panasonic_109_x1_22.png`)

Fig. L: Comparison of DRealSR results. (crop from `panasonic_253_x1_13.png`)

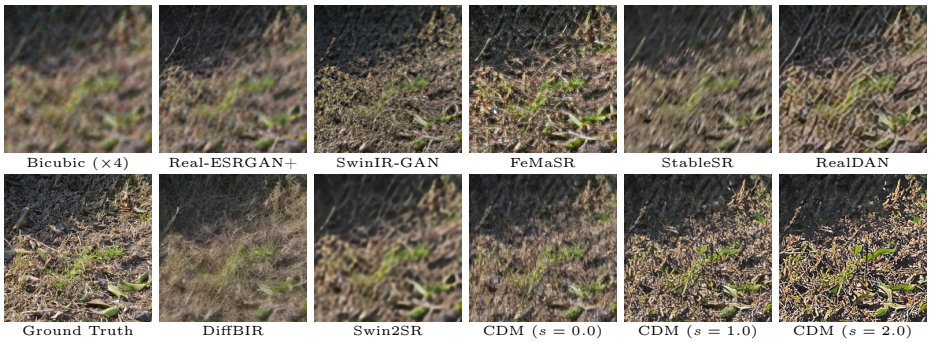

Fig. M: Comparison of DIV2K-Wild results. (crop from `0895x4w.png`)

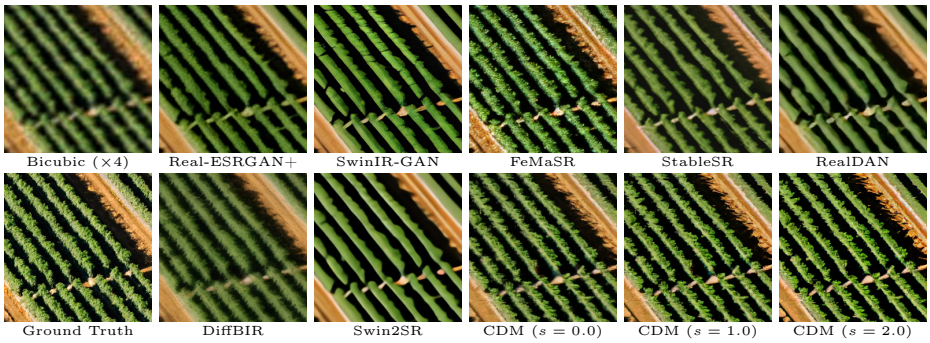

Fig. N: Comparison of DIV2K-Wild results. (crop from `0897x4w.png`)

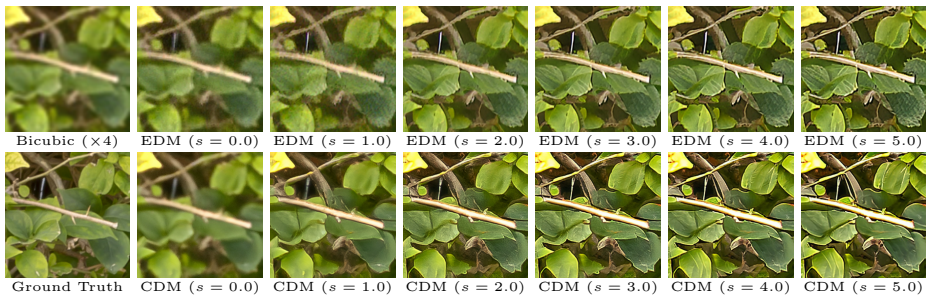

Fig. O: Comparison of CFG scale on RealSRv3 results. (crop from `Canon_169_LR4.png`)

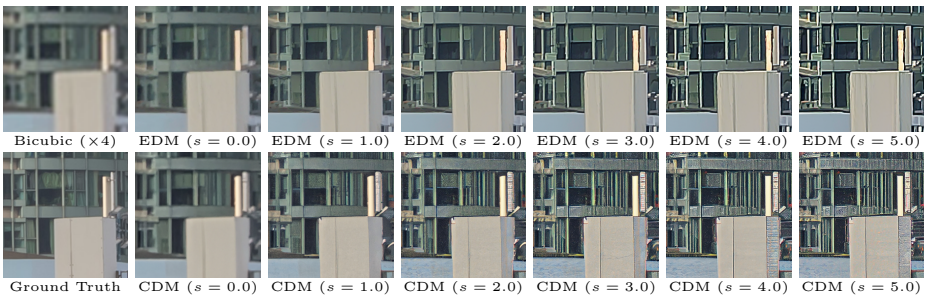

Fig. P: Comparison of CFG scale on RealSRv3 results. (crop from `Canon_170_LR4.png`)

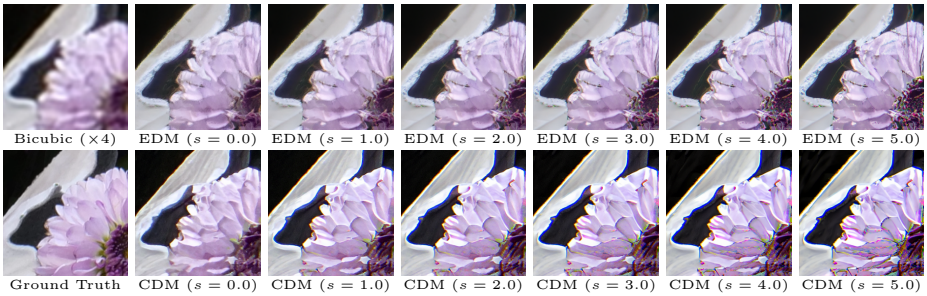

Fig. Q: Comparison of CFG scale on DRealSR results. (crop from `panasonic_195_x1_28.png`)

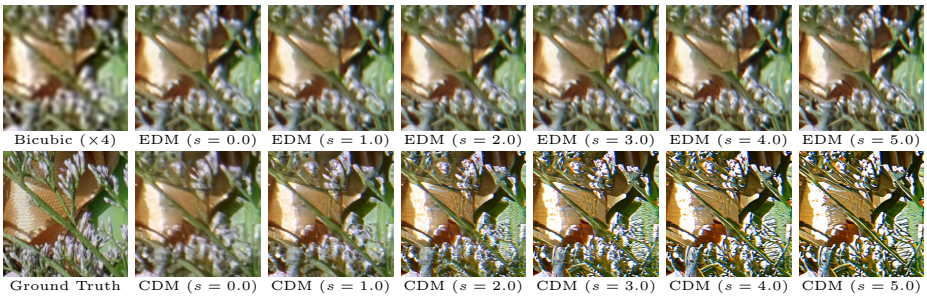

Fig. R: Comparison of CFG scale on DRealSR results. (crop from `panasonic_199_x1_27.png`)

References

1. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: ICCV (2019)
2. Cheng, S.: A new super-resolution measurement of perceptual quality and fidelity (2023)
3. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS. vol. 34, pp. 8780–8794 (2021)
4. Glickman, M.E.: Dynamic paired comparison models with stochastic variances. Journal of Applied Statistics **28**(6), 673–689 (2001)
5. Gu, J., Cai, H., Dong, C., Ren, J.S., Timofte, R.: Ntire 2022 challenge on perceptual image quality assessment. In: CVPR. pp. 950–966 (2022)
6. Gu, S., Lugmayr, A., Danelljan, M., Fritsche, M., Lamour, J., Timofte, R.: Div8k: Diverse 8k resolution image dataset. In: ICCV. pp. 3512–3516 (2019)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33, pp. 6840–6851 (2020)
8. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: NeurIPS (2022)
9. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: CVPR (2017)
10. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models (2023)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)
12. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L., Lim, B., et al.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: CVPR (2017)
13. Timofte, R., Gu, S., Wu, J., Van Gool, L., Zhang, L., Yang, M.H., Haris, M., et al.: Ntire 2018 challenge on single image super-resolution: Methods and results. In: CVPR (2018)
14. Wang, P.: denoising-diffusion-pytorch (2020), `https://github.com/lucidrains/denoising-diffusion-pytorch`
15. Xintao Wang, Ke Yu, C.D., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: CVPR (2018)
16. 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: CVPR. pp. 1191–1200 (2022)