Supplementary for ZFusion: Efficient Deep Compositional Zero-shot Learning for Blind Image Super-Resolution with Generative Diffusion Prior

Alireza Esmaeilzehi[1], Hossein Zaredar[1], Yapeng Tian[2], Laleh Seyyed-Kalantari[1,3]
[1]York University, Toronto, ON, Canada [2]The University of Texas at Dallas, Richardson, TX, US
[3]Vector Institute, Toronto, ON, Canada
{alirezae, hzaredar, lsk}@yorku.ca, {yapeng.tian}@utdallas.edu

1. Implementation Details

We train the proposed scheme with training images of *DIV2K* [1], *Flickr2K* [5], and *OutdoorSceneTraining* [7] datasets. We have extracted sub-images of size $512 \times 512 \times 3$ from the training dataset. We train the CZSD module of our method for 150K iterations. Afterwards, the training of the LR degradation suppression stage (SwinIR with our degradation embeddings) and image detail reconstruction (diffusion model with our degradation embeddings) are continued for 500K and 300K iterations, respectively. We employ the batch size of 32 in all our experiments. The training process of each module of the proposed ZFusion is performed with Adam optimizer [3] with the learning rate of 0.0001. The training of the proposed diffusion-based Blind SR scheme is carried out on a machine with four NVIDIA A40 GPUs.

All compared diffusion-based Blind SR methods were trained solely on synthetic data (no real-world data finetuning). As stated in the manuscript, for synthetic degradations, we followed the BSRGAN [11] pipeline, using randomly selected degradation operations, as done by all SOTA methods. For *RealSR* [2] and *DRealSR* [8], we used LR images across scaling factors 2, 3, and 4 to ensure thorough analysis. Due to the randomness in synthetic degradation, results vary across papers [4, 6, 9, 10]. For realistic degradations, some methods report results for only one scaling factor, limiting comparison. However, by using the official code of each SOTA method on the same evaluation set, we ensure a fair and conclusive analysis. The use of 3000 synthetic LR images and 560 realistic LR images further guarantees statistical reliability.

2. Ablation Studies

Tab. 1 reports PSNR/SSIM/LPIPS for the ablation studies done in the original manuscript. As shown, each component of the CZSD module improves not only perceptual quality but also fidelity metrics.

We have performed several ablation studies in the original manuscript in order to scrutinize the effectiveness of various components of ZFusion. We now show how coefficient α in the training of the CZSD module could influence the Blind SR performance. In order to show how we chose the values of this coefficient, Fig. 1 illustrates the SR performance in terms of CLIP-IQA, MANIQA, and CNN-IQA metrics as a function of α on the images of *RealSR* [2] dataset. It is seen from this figure that the best Blind SR performance is achieved when the value of coefficient α is set to 0.1.

Network	PSNR↑	SSIM↑	LPIPS↓
$Variant_1$	20.89	0.5708	0.2500
$Variant_2$	20.98	0.5734	0.2472
$Variant_3$	20.95	0.5729	0.2466
$Variant_4$	20.86	0.5719	0.2497
Proposed	21.05	0.5788	0.2456

Table 1. Impact of different components of the proposed ZFusion on SR performance in terms of PSNR/SSIM/LPIPS. Red indicates the best.

In order to assure that the realistic degradation attributes of the RDAL block are meaningful and of good quality, we demonstrate in Fig. 2 the t-SNE plot of 100 textual data containing sentences `image is degraded with unknown realistic degradation resulting in HyperIQA/BRISQUE/TOPIQ/NIMA image quality assessment` value, where the value of the IQA metrics are changed linearly from the minimum value of the IQA range to the maximum value. It is seen from Fig. 2 that the t-SNE plot of the real-world degradation embeddings also change in a linear regime, confirming that this block properly understands the unknown degradation processes.

3. Comparison with State-of-the-art Schemes

We now compare the performances of our ZFusion with those of the other SOTA schemes in the case of the images of *DRealSR* [8] dataset. Similar to *RealSR* [2] LR images,

Metric	Real-ESRGAN	ResShift	PASD	SeeSR	SinSR	StableSR	DiffBIR	ZFusion (Ours)
PSNR↑	22.81	22.50	23.03	21.84	21.15	19.87	21.87	20.73
SSIM↑	0.6910	0.6541	0.6668	0.5928	0.5950	0.5822	0.5772	0.5472
LPIPS↓	0.1891	0.1791	0.1774	0.2185	0.2109	0.2207	0.2209	0.2198
CLIP-IQA↑	0.5193	0.5895	0.5879	0.6384	0.6512	0.6294	0.6593	0.6652
NIQE↓	4.7980	6.5872	4.5025	4.6161	5.6998	5.1306	4.9159	4.6418
MUSIQ-AVA↑	4.6710	4.7150	4.8628	5.0406	4.8066	4.6768	5.1564	5.0579
MANIQA↑	0.2570	0.2481	0.2875	0.3462	0.3152	0.2485	0.3759	0.5707
CNN-IQA↑	0.6534	0.6254	0.6294	0.6786	0.6811	0.6247	0.6898	0.6900

Table 2. Comparison between performances of various Blind SR schemes on the images of DRealSR dataset with real-world unknown degradations. Red indicates the best, while Blue shows the second best.

Method	Real-ESRGAN	ResShift	PASD	SeeSR	SinSR	StableSR	DiffBIR	ZFusion (Ours)
Time (s)	0.05	0.19	6.81	6.45	0.07	10.44	9.67	9.92

Table 3. Complexity of various Blind SR schemes on an image of size 128 × 128.

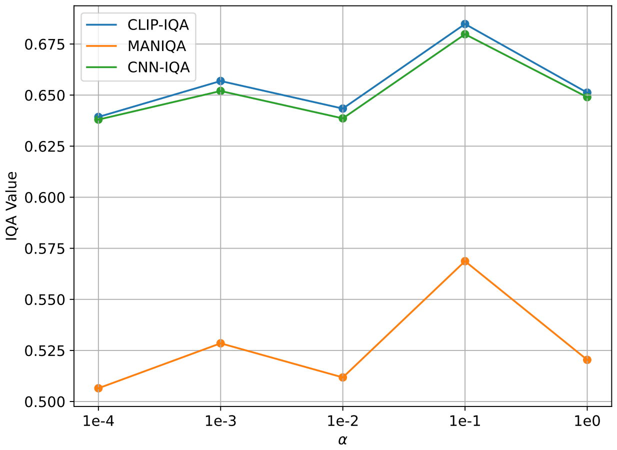

Figure 1. Impact of hyper-parameter α in the loss function on SR performance.

the *DRealSR* LR images are also acquired with camera realistic degradation pipeline. The results of various methods on the images of this dataset are given in Tab. 2. As seen from this table, our proposed ZFusion provides 4 best and 1 second best IQA values. By considering all the results of synthetic and real-world degradation processes (Tab. 2 and Tab. 3 of the manuscript and Tab. 2 of SM), our ZFusion provides 13 best and 1 second best IQA values out of a total number of 24 metrics. DiffBIR gives 1 best and 11 second best metrics, as the second best performing method. This shows the significant superiority of our ZFusion over the SOTA methods, no matter LR images are degraded with synthetic or real-world degradations.

We compare the complexity of the proposed ZFusion with those of the other SOTA Blind SR methods on an LR image of size 128 × 128 × 3. Specifically, Tab. 3 gives the average inference time of various methods on a machine with an NVIDIA A40 GPU. It is seen from the results of this table that the proposed ZFusion employs a comparable inference time as those of the high-performing diffusion-based SR methods. Specifically, the execution time of our method is very similar to that of DiffBIR [4], which stands out as the second best performing method considering the results of (Tab. 2 and Tab. 3 of the manuscript and Tab. 2 of SM) all together.

We now present a perception-distortion plot in Fig. 3, comparing all methods. The vertical axis shows normalized SSIM and 1 − LPIPS (averaged), and the horizontal axis reflects the average of MANIQA and CLIP-IQA metrics. ZFusion achieves the best overall balance, offering superior perceptual quality with only a slight drop in fidelity—highlighting its effectiveness in managing the perception-distortion trade-off.

Finally, we illustrate more qualitative results generated by our ZFusion and the other SOTA schemes, in Figs. 4 and 5. As seen from these figures, our ZFusion better enhances the quality of LR images comparing to the other Blind SR algorithms. Generally, the SR images of our ZFusion are sharper and less blurred than those of the other SR counterparts. This is achieved on the LR images that are obtained by synthetic and real-world degradations.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 1

[2] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for

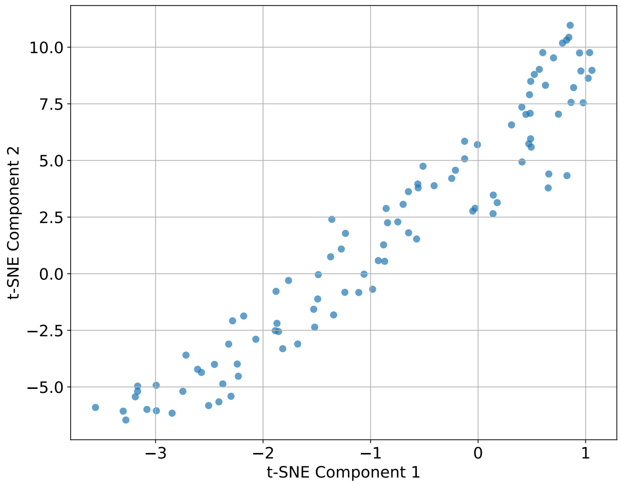

Figure 2. t-SNE visualization of the RDAL block's embeddings $\mathbf{f}_2$ in response to changing the severity of unknown degradations in a linear fashion.

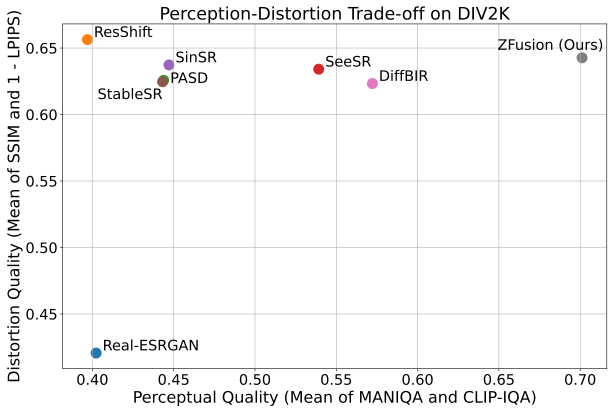

Figure 3. Perception-distortion trade-off of ZFusion and other SOTA schemes on the *DIV2K* dataset.

stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] 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. Springer, 2024. 1, 2

[5] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 1

[6] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024. 1

[7] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 1

[8] Pengxu Wei, Ziwei Xie, Hannan Lu, Zhengyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 101–117. Springer, 2020. 1

[9] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25456–25467, 2024. 1

[10] 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. Springer, 2024. 1

[11] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4791–4800, 2021. 1

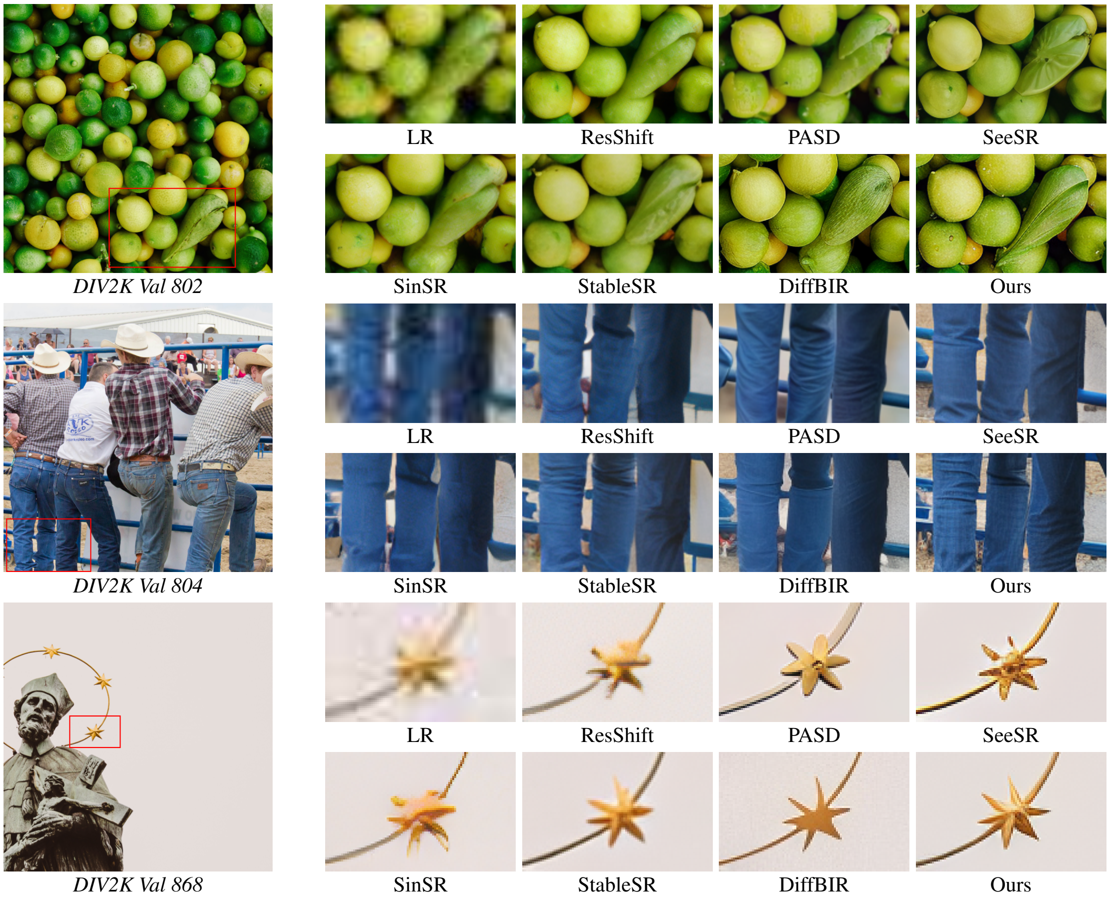

Figure 4. Visual qualities of images from the DIV2K Validation dataset with synthetic degradations that are super resolved by various state-of-the-art image super resolution schemes.

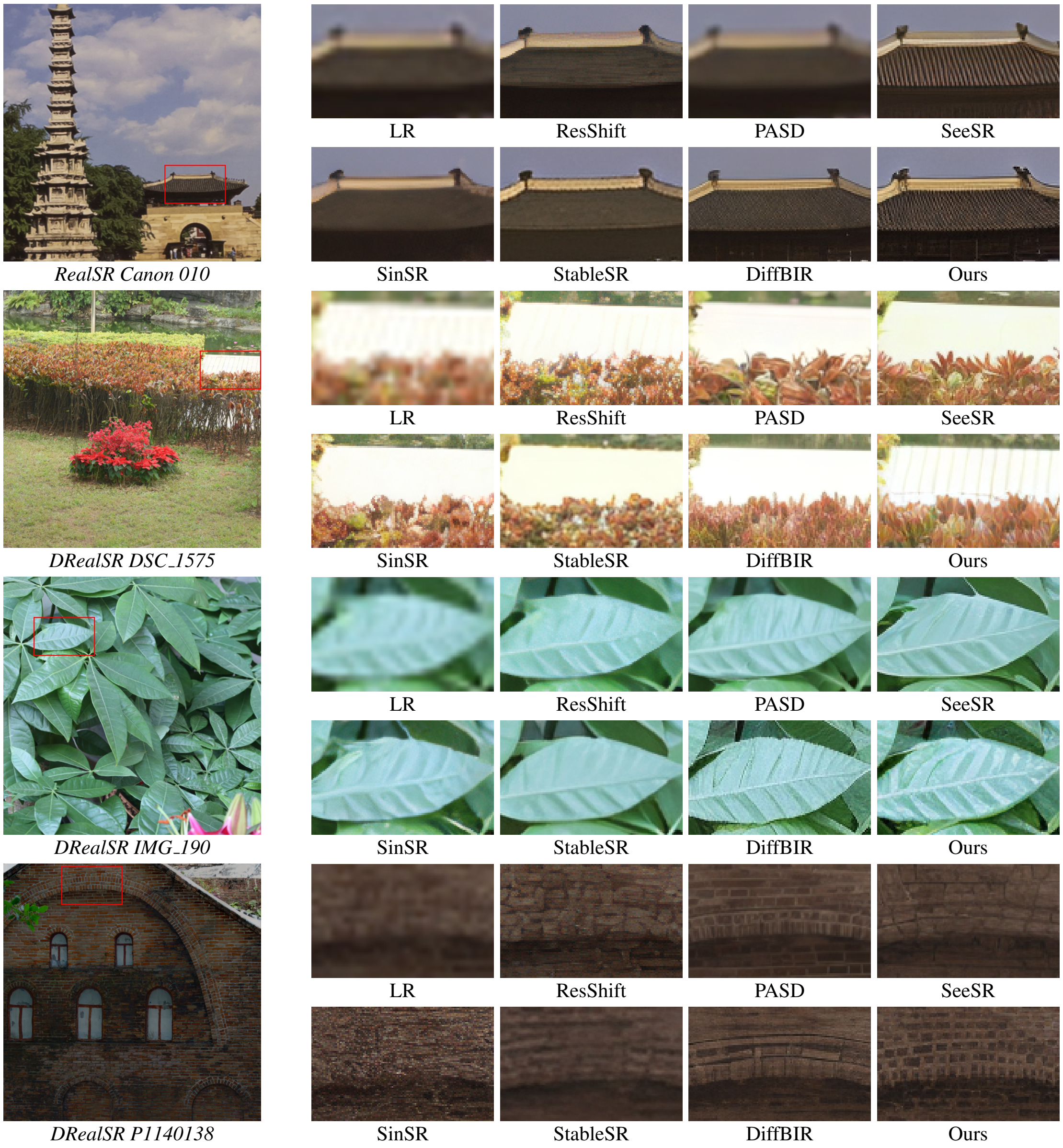

Figure 5. Visual qualities of images from the RealSR and DRealSR dataset with realistic unknown degradations that are super-resolved by various state-of-the-art image super resolution schemes.