Perception-Distortion Balanced ADMM Optimization for Single-Image Super-Resolution Supplementary Material

Yuehan Zhang[1], Bo Ji[1], Jia Hao[2], and Angela Yao[1]

[1] National University of Singapore, {zyuehan,jibo,ayao}@comp.nus.edu.sg
[2] HiSilicon Technologies, Shanghai, hao.jia@huawei.com

The Supplementary features the following sections:

- Section A: detailed architecture of our model.
- Section B: full comparison between our method and other perception-distortion balanced methods, which traverse a trade-off curve.
- Section C: comparison of the inference complexity of our model and the post-processing based method, WDST [1].
- Section D: more ablation studies, including
  - comparison of training processes with regularizer-based variants.
  - comparison of the results from objective and perceptual stage.
  - weights of the loss terms of the no-constraint variant.
  - bandwidth of LF information involved in the low-frequency constraint.
- Section E: more qualitative examples and comparisons with competing methods.
- Section G: challenging cases for our method.

A Model Architecture

The architecture of the objective and perceptual-focused stage is borrowed from other state-of-the-art work. As shown in Fig. 1, we adopt the architecture of HAN [10] as the objective-focused stage. HAN is constructed based on RCAN [20] with a novel layer attention module and a channel spatial attention module, the composition of which is in Fig. 2. The objective-focused stage has ten residual groups with 20 RCAB blocks in each group, and the reduction ratio in the attention layer is 16. All convolution layers use 64 channels, 3×3 kernels, and a padding size of 1. Finally, we upsample feature maps with a PixelShuffle [11] module.

The perception-focused stage works in the wavelet domain. We first apply the discrete wavelet transform (DWT) at the beginning of the network and an inverse discrete wavelet transform (IWT) at the very end. We use 15 Res-Clique Blocks introduced in Zhong *et al.* [21] as the building blocks, as shown in Fig. 3. The perception-focused stage takes the output from the objective-focused module as input. The input has a size of $c \times H \times W$. We apply DWT to the input to split it into four half-resolution channels (LL, LH, HL, HH), which are then concatenated into a feature map of size $4c \times \frac{H}{2} \times \frac{W}{c}$. Note that DWT

is a lossless operation, even though it reduces the spatial size of feature maps. Processed features are converted by the last inverse wavelet transform (IWT) operation to an HR image of size $c \times H \times W$.

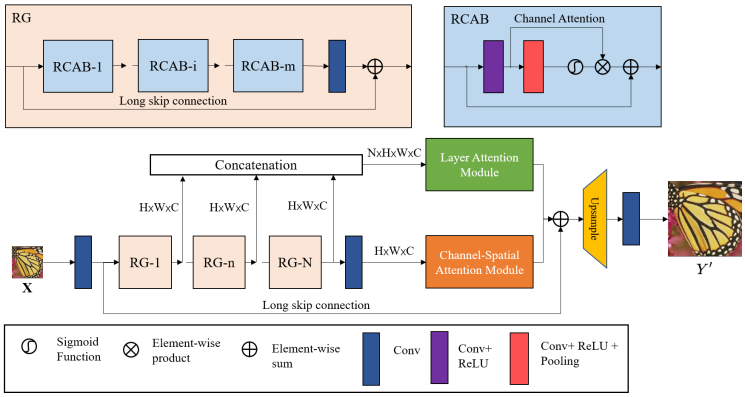
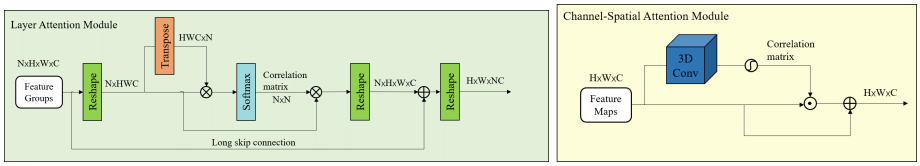


Fig. 1: The architecture of the objective-focused stage. We adopted HAN [10] here.

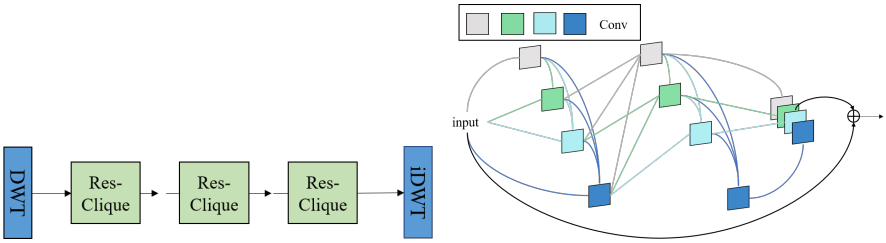


(a) Layer Attention Module (b) Channel-Spatial Attention Module

Fig. 2: The architecture of Layer Attention Module and Spatial-Channel Attention Module in the objective-focused stage.

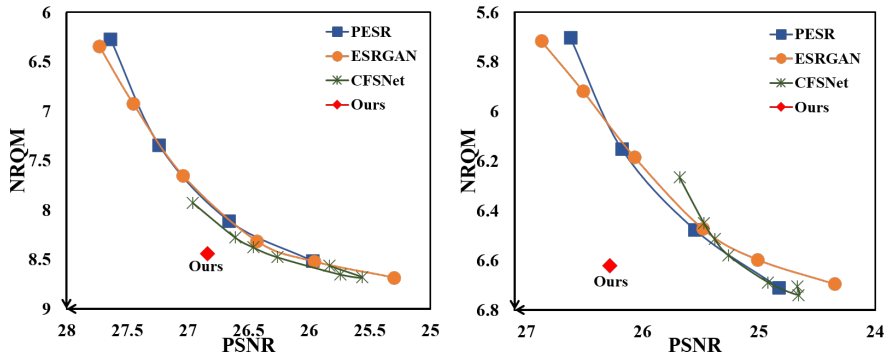
B Full Comparison with Transition Methods

We provide a complete comparison with CFSNet [16], PESR [15], and ESRGAN [17] with network interpolation. Section 4.2 of the main paper offers a table for convenient comparison between our method and other perception-distortion (PD) trade-off balanced methods. However, some related methods traverse a trade-off curve instead of having single-point performance. PESR uses the image interpolation method, ESRGAN interpolates network parameters, and CFSNet introduces a controlling factor, α, as model input. In the main paper, we chose the most balanced point of each traversing method for comparison; in this section, we compare our method with the full curve. As shown in Fig. 4, our method achieves a better balance of PD trade-off than those yielding a transition.



(a) Perceptual-focused stage. (b) Res-Clique Block

Fig. 3: The architecture of perceptual-focused stage.



(a) BSD100 [6] (b) Urban100 [3]

Fig. 4: Full comparison with CFSNet [16], PESR [15] and ESRAGN [17]. Our method can achieve better trade-off *(under the curves)* than those methods producing a transition.

C Inference Complexity

The closest competing method to ours is WDST [1], a post-processing method that fuses two images from separately trained objective- and perception-focused models with a style-transfer approach. To super-resolve a low-resolution image, WDST performs four steps. The second step is a style transfer procedure following [2] and requires inference-specific updates to the network. We elaborate on these computations below and then compare its complexity to our approach.

The fours steps of WDST are:

1. Super-resolution through two separately trained models, objective-focused EDSR [5] and perception-focused CX [8], yielding HR images $\hat{Y}^O$ and $\hat{Y}^P$ respectively.

	Model	Forward Pass	FLOPs	Run-time
Step 1	EDSR [5]; CX [8]	1	50.6 M	0.918s
Step 2	VGG19 [12]	$6 \times (i+2)$	$\geq 37445.7M$	$\geq 1152s$
Step 3	VDSR [4]	1	2.4 M	0.05s
Step 4	-	-	-	-

Table 1: Decomposition of the FLOPs calculation of WDST [1]. We calculate the FLOPs for the convolutional operations of each step. Step 2 applies 6 VGG19 models on 6 pairs of wavelet channels; each VGG19 needs to extract features from two source channels and update the merged channel. Assuming each model updates for i iterations, the total number of forward passes through all VGG19 models is $6 \times (i+2)$.

2. The two-level stationary DWT decomposes $\hat{Y}^O$ into 7 channels $\{HL, LH, HH, LL', HL', LH', HH'\}$ [3]. It decomposes $\hat{Y}^P$ similarly. Aside from LL', the other six channels from each decomposition are merged through six independent style transfer procedures using a VGG19 model for deep feature extraction. During merging, the style transfer extracts features from the channels of $\hat{Y}^O$ and $\hat{Y}^P$ as content ground truth and style ground truth. The initial input then passes through the transfer to gather multiple-level features and calculate errors compared with content ground truth and style ground truth at different levels. The initial signal is updated for i iterations to get the merged wavelet channel. Fusing first- and second-level features usually needs 5000 and 1000 iterations, respectively, thus we assume $i \geq 1000$ in the FLOPs calculation.
3. The lowest-frequency channel LL of $\hat{Y}^O$ is refined with VDSR [4].
4. The final HR imaged is obtained by an inverse stationary discrete wavelet transform of the six merged channels and the refined LL.

The approximate FLOPs of the above four steps for a 128×128 input patch is tallied in Table 1.

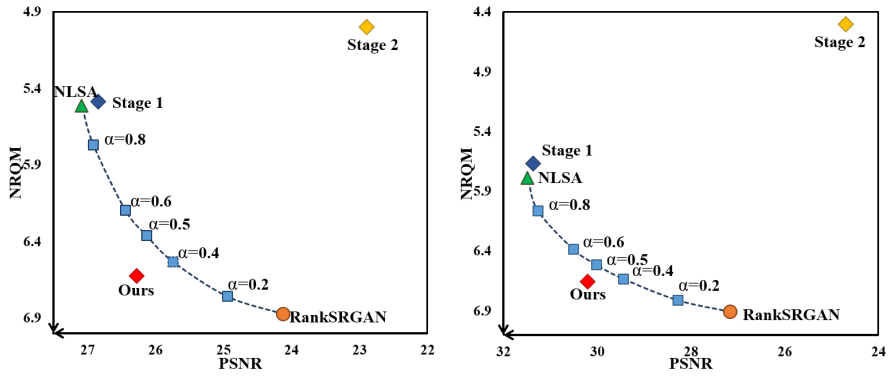
In comparison, our one-shot inference procedure applies only convolutional operations and does not require any back-propagation and updates like Step 2 of WDST. Despite the conservative estimate of WDST (we tabulate only 1000 iterations for all channels in Step 2 for FLOPs calculation), the complexity of our method, which uses only 26.8M FLOPs and 0.566s run-time, is three orders of magnitude smaller than WDST.

[3] Each stationary DWT decomposition results in 4 channels with the same resolution as the original image; the LL channel of the first level is decomposed further into $\{LL', HL', LH', HH'\}$, resulting in 7 channels in total.

D Ablations

a Stage-wise Performance

This experiment validates the necessity of each stage of LFc-SR. LFc-SR has two stages with different goals and allows a separate output Y' from the objective-focused stage. We compared Y' and the final output Y with $\tilde{Y}$ from the perceptual-focused stage only. The LR counterpart of $\tilde{Y}$ was upscaled by bicubic interpolation first and directly put into the perceptual-focused stage of LFc-SR. As shown in Fig. 5, we compare all results on a PD trade-off plane with the estimated trade-off boundary presented in Section 4.2 of the main paper. With only the objective-focused stage, Y' has a high PSNR score but fails to balance the objective and perceptual quality. On the other hand, using the perceptual-focused stage alone does not complete the reconstruction, as reflected by the low PSNR and NRQM performance of the outputs from stage 2. A good balance between objective and perceptual quality can only be achieved by using both stages in succession.



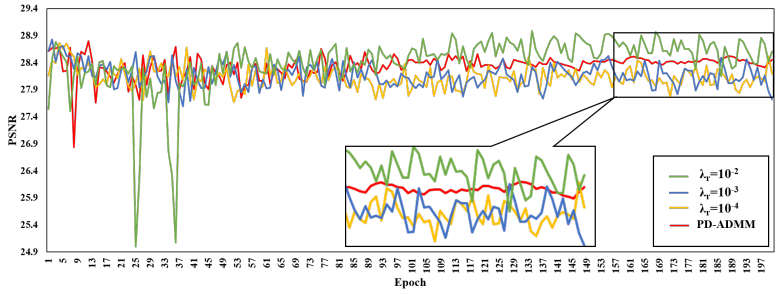
(a) Urban100 [3] (b) Manga109 [7]

Fig. 5: Comparison between the final output from LFc-SR and the HR images going through only one stage of the same LFc-SR model. The objective stage *(stage 1)* yields high PSNR HR images with a low perceptual score; the direct use of the perceptual stage *(stage 2)* fails to reconstruct a reasonable HR image. Only by using two stages together can we achieve a good balance of PD trade-off.

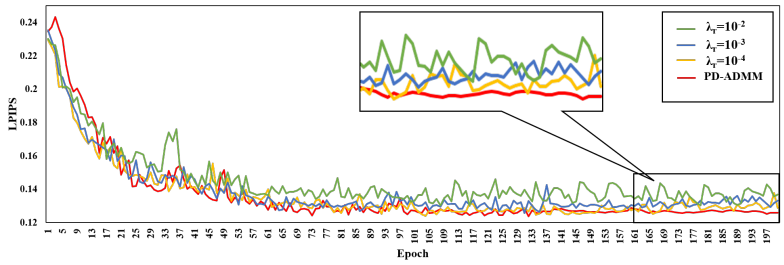
b Stability of Training

In Section 4.3 of the main paper, we compared the performance of our PD-ADMM method with regularizer-based models. In this section, we further prove that PD-ADMM can achieve a more stable performance than other models by providing the models' performance on the validation set during training. We used

the validation set of the DIV2K dataset [13]. For all the models, we used the same pretrained model and trained them with the same setting as introduced in Section 4.1 of the main paper, except for the gradients of the regularizer-based models that were clipped to 10^{-4} to prevent gradient explosion. We tested the model on the validation set after each epoch and recorded the PSNR and LPIPS scores. We trained each model for 200 epochs and visualize their performance in Fig. 6.



(a) PSNR



(b) LPIPS

Fig. 6: Visualization of the training performance on the validation set of our model and regularizer-based models. Higher PSNR and lower LPIPS indicate better performance. The model trained with PD-ADMM has a more flattened curve during the last tens of epochs with high performance on both PSNR and LPIPS.

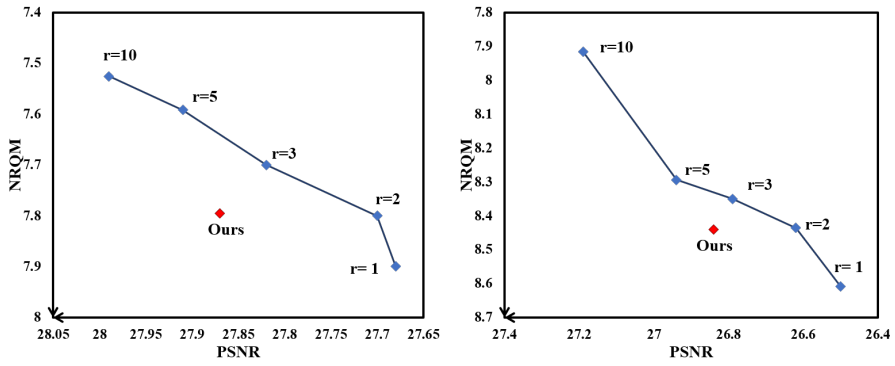
c Loss Terms of No-Constraint Variant

Section 4.3 of the main paper presents a comparison of our method with the regularizer-based methods, including the no-constraint case when the regularizer's weight equals 0. Based on the observation in Vasu *et al.* [14], the different relative weights of L_O and L_P (the loss terms of the objective-focused and perceptual-focused stage) may result in a transition between perceptual and objective quality even without a low-frequency constraint. It is because L_O and L_P both influence the gradients in the first stage of LFc-SR, even though they are designed to supervise their own stage only. We compare our method

with no-constraint models trained with different relative weights of L_O and L_P. Specifically, we weighted the losses of objective- and perceptual-focused stages as follows:

$$L = \lambda_O \cdot L_O + \lambda_P \cdot L_P, \tag{1}$$

where L_O and L_P have the same definitions as in Section 3.2 of the main paper. We tried different ratio r ($r = \lambda_O/\lambda_P$), and the results are shown in Fig. 7. As the ratio r increases to a significant value, *e.g.* $r = 10$, the results show a clear transition from better perceptual quality to better objective quality; however, the transition fails to achieve a good balance of PD trade-off compared with our method.



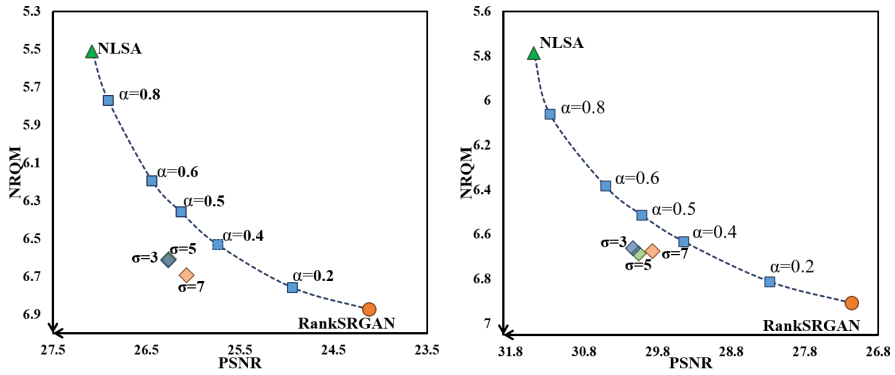
(a) Set14 [18] (b) B100 [6]

Fig. 7: Comparison between our model and the no-constraint model with different weights of objective stage and perceptual stage losses *($r = \lambda_O/\lambda_P$)*. Weighting stage losses differently can produce a transition between better perceptual quality and objective quality. However, the transition has an inferior PD balance compared with our method.

d Low Frequency Bandwidth of Constraint in Gaussian Blur Variant

In this experiment, we explored how the bandwidth of low-frequency information involved in the constraint of LFc-SR influences the overall performance. Although our method in the main paper used DWT to extract low-frequency bands, we used Gaussian blur here because of its convenient control of bandwidth by adjusting σ. Specifically, we implemented a convolution layer with 21×21 Gaussian Blur kernel with $\sigma = 1, 3, 5, 7$. Fig. 8 shows the results on a plane similar to what is used in Sec. a, except for the $\sigma = 1$ case, which suffered from constant training collapse. The $\sigma = 3$ and $\sigma = 5$ case have very close performance, while the $\sigma = 7$ case deviates more due to a lower PSNR. This

shows that our method is not sensitive to bandwidth change within an appropriate range, but an extremely high or low bandwidth will cause problems, *e.g.*, training collapse or objective quality drop.



(a) Urban100 [3] (b) Manga109 [7]

Fig. 8: Comparison of models trained with different low-frequency bandwiths involved in the constraint in LFc-SR. Within an appropriate range *($\sigma = 3, 5$)*, the performances are highly close; when offered too high or low bandwidth, the model will suffer from objective quality drop *($\sigma = 7$)* or training collapse (not shown in the figure).

E Visual Comparisons

This section shows more visual comparisons with state-of-the-art single-focused methods in Fig. 11. We also compare our method with the closest competing PD-balanced method, WDST [1]. Fig. 9 shows that WDST hallucinate non-existent textures in low-frequency regions, which does not happen in our results.

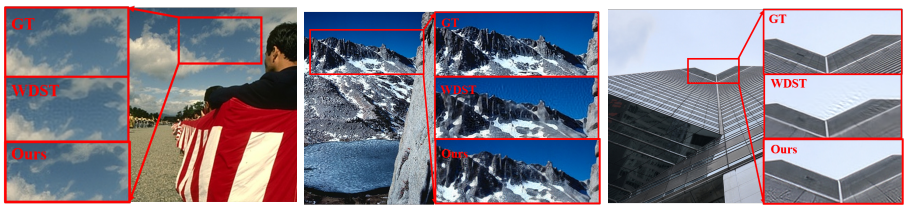
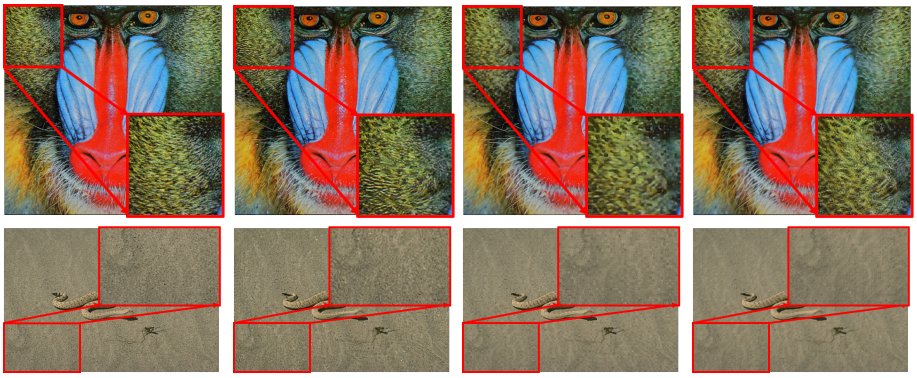


Fig. 9: Comparison between our method and WDST [1]. WDST generates non-existent textures in low-frequency regions, which do not occur in our results.



(a) Ground Truth (b) NLSA [9] (c) RankSRGAN [19] (d) Ours

Fig. 10: The challenging cases of our method. Compared to the perceptual-focused model, our method achieves inferior quality in very high-frequency regions, *e.g.* furry and sands. This phenomenon also happens with other trade-off balanced methods, such as WDST [1].

F User Study

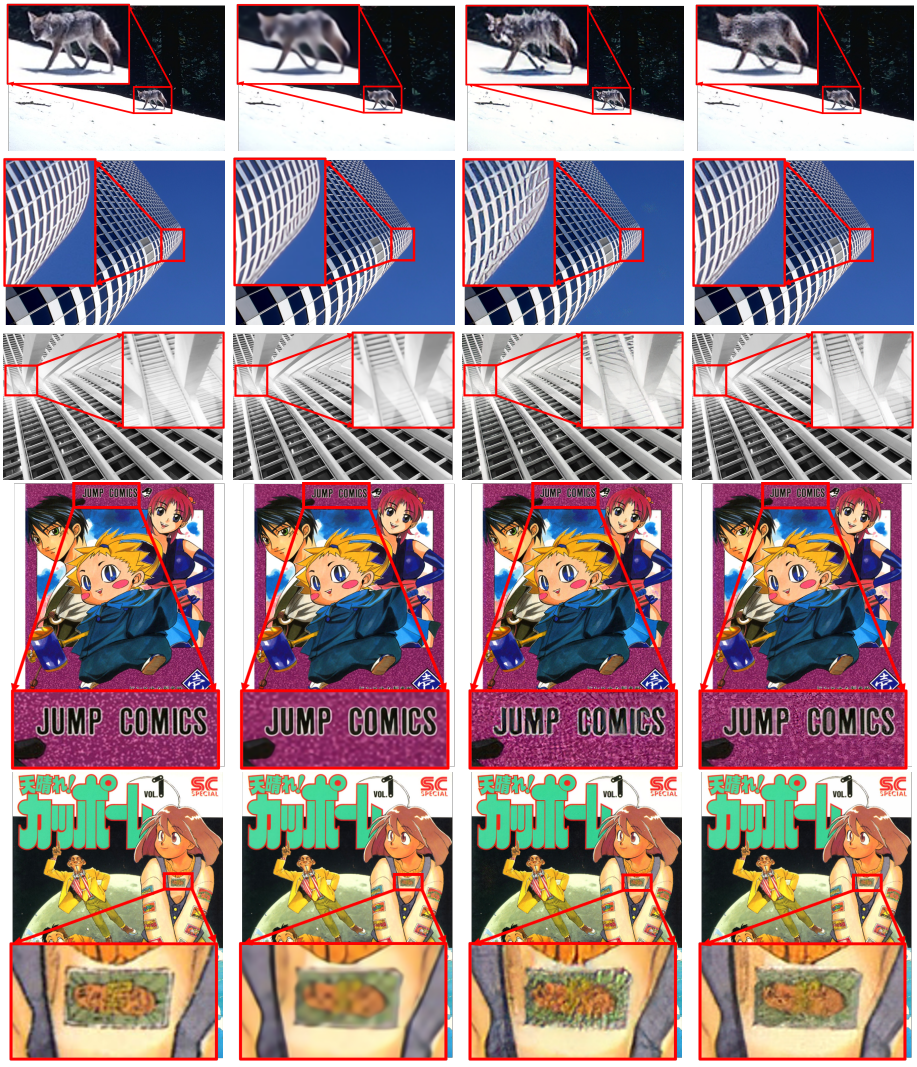
We ask 30 workers on *Amazon MTurk* to evaluate 40 images from datasets Urban100 and BSD100. Each worker saw five SR results of the same image and rated them based on their realness, from 1 (the worst) to 5 (the best). As shown in Table 3, our method surpasses others (even the perceptual-focused method, RankSRGAN) in terms of mean score with the lowest standard deviation.

Table 3: Results of user study. PD-ADMM has highest mean score and lowest standard deviation (Std.).

Methods	Bicubic	NLSA	WDST	RankSRGAN	PD-ADMM
Mean	2.54	3.31	3.52	<u>3.65</u>	**3.66**
Std.	1.53	1.47	<u>1.36</u>	1.37	**1.32**

G Challenging Cases

Although our method shows competitive results on different benchmarks overall, it fails to generate sharp and realistic images in some challenging cases. As shown in Fig. 10, our method does not restore the information very well for the regions with very high-frequency information, like furry and sand. This is also an issue for WDST [1], indicating that it is a common challenge for perception-distortion balanced methods. A perception-focused method, such as RankSRGAN [19], can give better visual qualities for these cases.



(a) Ground Truth (b) NLSA [9] (c) RankSRGAN [19] (d) Our results

Fig. 11: Visual comparisons with objective-focused model, NLSA [9], and perception-focused model, RankSRGAN [19], on BSD100 [6] *(first row)*, Urban100 [3] *(second and third row)* and Manga109 [7] *(forth and fifth row)* datasets. Our method produces sharper HR images than the objective-focused method and less unnatural artifacts than the perceptual-focused method.

References

1. Deng, X., Yang, R., Xu, M., Dragotti, P.L.: Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3076–3085 (2019) 1, 3, 4, 8, 9
2. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016) 3
3. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5197–5206 (2015) 3, 5, 8, 10
4. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016) 4
5. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017) 3, 4
6. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. 8th Int'l Conf. Computer Vision. vol. 2, pp. 416–423 (July 2001) 3, 7, 10
7. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications **76**(20), 21811–21838 (2017). https://doi.org/10.1007/s11042-016-4020-z 5, 8, 10
8. Mechrez, R., Talmi, I., Shama, F., Zelnik-Manor, L.: Maintaining natural image statistics with the contextual loss. In: Asian Conference on Computer Vision. pp. 427–443. Springer (2018) 3, 4
9. Mei, Y., Fan, Y., Zhou, Y.: Image super-resolution with non-local sparse attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3517–3526 (2021) 9, 10
10. Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., Zhang, K., Cao, X., Shen, H.: Single image super-resolution via a holistic attention network. In: European conference on computer vision. pp. 191–207. Springer (2020) 1, 2
11. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016) 1
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 4
13. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 114–125 (2017) 6
14. Vasu, S., Thekke Madam, N., Rajagopalan, A.: Analyzing perception-distortion tradeoff using enhanced perceptual super-resolution network. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018) 6

15. Vu, T., Luu, T.M., Yoo, C.D.: Perception-enhanced image super-resolution via relativistic generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018) 2, 3
16. Wang, W., Guo, R., Tian, Y., Yang, W.: Cfsnet: Toward a controllable feature space for image restoration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4140–4149 (2019) 2, 3
17. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018) 2, 3
18. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: International conference on curves and surfaces. pp. 711–730. Springer (2010) 7
19. Zhang, W., Liu, Y., Dong, C., Qiao, Y.: Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3096–3105 (2019) 9, 10
20. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018) 1
21. Zhong, Z., Shen, T., Yang, Y., Lin, Z., Zhang, C.: Joint sub-bands learning with clique structures for wavelet domain super-resolution. Advances in neural information processing systems **31**, 165–175 (2018) 1