Supplementary of Mutual Affine Network for Spatially Variant Kernel Estimation in Blind Image Super-Resolution

Jingyun Liang[1] Guolei Sun[1] Kai Zhang[1] Luc Van Gool[1,2] Radu Timofte[1]
[1]Computer Vision Lab, ETH Zurich, Switzerland [2] KU Leuven, Belgium
{jinliang, guosun, kaizhang, vangool, timofter}@vision.ee.ethz.ch
https://github.com/JingyunLiang/MANet

We first give more details on the training of MANet and the non-blind SR model RRDB-SFT. Then, we show the visualization of kernel distribution. We also report interesting results on consecutive degradation and image quantization. Last, we provide more visual comparisons of different methods on synthetic and real-world images.

1. Training Details of MANet

For MANet, Adam optimizer [5] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is used to train the model for 300,000 iterations. Learning rate is initialized as $2e-4$ and reduced by half every $50,000$ iterations. The training time is about 15 hours on a Tesla V100 GPU.

2. Non-Blind SR Model RRDB-SFT

We design a non-blind super-resolver based on RRDB block [11] and SFT layer [10]. The model is dubbed as RRDB-SFT and the architecture is shown in Fig. 1. As one can see, RRDB-SFT reconstructs the HR image by taking the LR image and the corresponding kernel as input. Specifically, it first reshapes the kernel from size of $h \times w$ to hw and reduces the dimensionality from hw to l by principal component analysis (PCA). After that, the kernel PCA vector is stretched to a kernel PCA map of size $l \times \frac{H}{s} \times \frac{W}{s}$, where H, W and s are HR image height, width and scale factor, respectively. Then, the kernel PCA map is concatenated with different levels of image features via the SFT layers. More details of the RRDB block and SFT layer can be found in [11] and [10], respectively.

Interestingly, although we train the model with spatially invariant kernels, RRDB-SFT can naturally deal with spatially variant kernels as the stretched kernel map includes kernel PCA vectors for every position on the LR image input. Thus, the network is able to learn the correspondences between local LR image patches and kernels. When combined with MANet for spatially variant blind SR, we first downscale the kernel prediction of MANet to match the spatial size of LR image and reduce the channel dimensionality by PCA. Then, the kernel PCA map is input to the SFT layers (kernel stretching is skipped) to help the reconstruction of HR image.

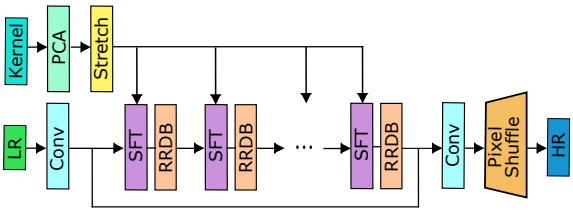


Figure 1: Architecture of the non-blind super-resolver RRDB-SFT. It takes blur kernels and the LR image as input, and outputs the HR image. Kernel feature and image feature are fused by the SFT layer.

In experiments, we use 10 RRDB blocks and 10 SFT layers. The kernel PCA vector dimension is set to 15. Similar to MANet, we randomly crop 192×192 image patches from DIV2K [1] and Flickr2K [8], and augment them by random flip and rotation in training. The image patches are blurred by random kernels for HR-LR image pair generation. We use mean absolute error (MAE) between SR image and HR image as the loss function. Adam optimizer [5] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ is used to train the model for $480,000$ iterations, with a batch size of 16. The learning rate is initialized as $2e-4$ and halved every $120,000$ iterations. It takes about two days to train RRDB-SFT on a Tesla V100 GPU. When combined with MANet, we freeze the parameters of MANet and fine-tune RRDB-SFT. The learning rate and total number of iterations are $5e-5$ and $200,000$, respectively. Note that the blind SR performance could be further improved with a better non-blind SR model.

3. Visualization of Kernel Distribution

We visualize the distribution of estimated kernels on "*img017*" in Urban100 [4] by t-SNE [6]. As Fig. 2 shows, the estimated kernels are diversified, ranging from kernels

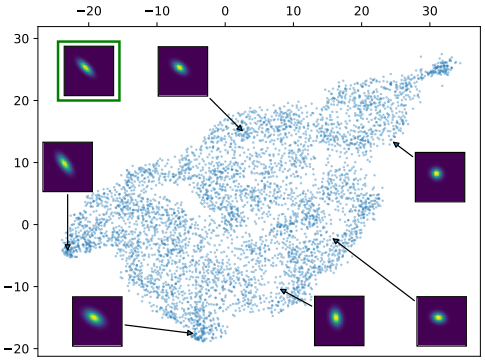


Figure 2: The t-SNE visualization of the distribution of estimated kernels on "*img017*" in Urban100 [4] when scale factor is 4. The corresponding HR image is blurred by a spatially invariant kernel in the top left green rectangle.

Figure 3: Kernel estimation results of the proposed MANet under consecutive degradation when scale factor is 4. The testing image is "*img077*" in Urban100 [4]. We blur and downsample the HR image twice with two different kernels (in total, $16\times$), which are shown in the blue and green rectangles. The shown image is the $4\times$ nearest neighbour interpolation of the final LR image.

close to the ground-truth, to nearly isotropic kernels. It is worth pointing our that most kernels are not obviously wrong, *e.g.*, being vertical to the ground-truth. Considering the high LR image PSNR achieved by MANet, we believe most of these kernels are "*correct*" kernels, leading to a high PSNR of the LR image. Besides, as can be seen from the image-kernel correspondences in Fig. 1 of the paper, kernels close to the ground-truth are mainly estimated from image patches with corners or rich textures, while flat patches are less discriminative, producing a fixed isotropic kernel.

4. Impact of Consecutive Degradation

As discussed in the paper, MANet estimates kernels based on image patch characteristics. Therefore, it is interesting to explore the effects of consecutive degradation. We first generate a LR image following the ordinary degradation process. Then, we blur the LR image by another kernel and downsample it again. As shown in Fig. 3, MANet tends to estimate kernels that are close to the second kernel, omitting the first kernel. This indicates that the image patch distribution is mainly determined by the latest degradation.

5. Impact of Image Quantization

Blurring the image with a kernel can lead to distinctive image patch characteristics. In image SR, the blurred image is further downsampled and quantized. In the paper, experiements on different scale factors have shown the impact of downsampling. Here, we try to explore the impact of quantization. Fig. 4(a) shows the kernel estimations at different positions when the LR image is unquantized. Surprisingly, MANet is able to roughly estimate the kernel from a blurred 1×1 cross, whose HR counterpart only has a single black point as shown in Fig. 4(b). In contrast, on quantized images, it can only deal with image patches whose sizes are at least 9×9. This can be attributed to information loss in image quantization after blurring and downsampling.

6. More Visual Comparisons

We provide more visual comparisons on both synthetic and real-world images in Fig. 5 to show the effectiveness of our model. Note that we train all these model with only L_1 pixel loss for simple and fair comparison, though it is known that GAN loss can further improve the visual quality.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 1

[2] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In *Advances in Neural Information Processing Systems*, pages 284–293, 2019. 4

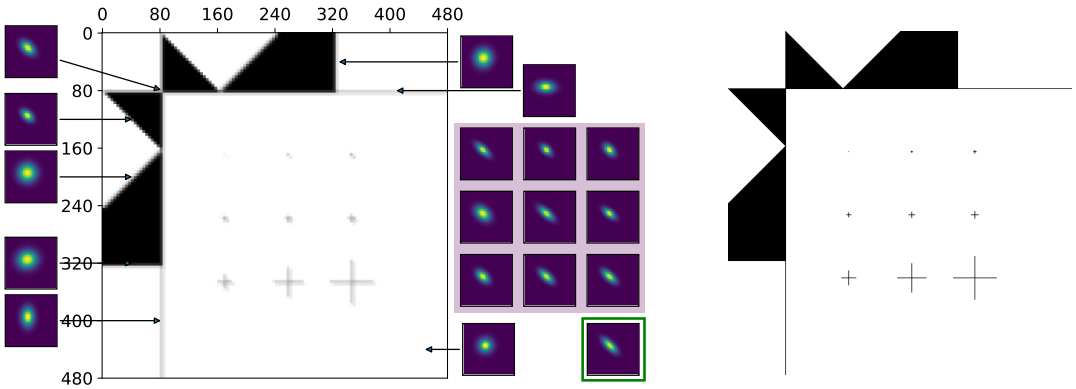
[3] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019. 4

[4] 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. 1, 2

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 1

[7] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic



(a) Kernel estimation results without image quantization

(b) HR image

Figure 4: (a) shows kernel estimations of MANet at different positions on a synthetic image for scale factor 4, **when the LR image is unquantized**. The shown image is the nearest neighbour interpolation of the LR image, which was generated by a Gaussian kernel with parameters $\sigma_1 = 6$, $\sigma_2 = 1$ and $\theta = \frac{\pi}{4}$, as shown in the down right green rectangle. The corresponding HR image has 9 black crosses (1×1, 3×3, 5×5, 7×7, 9×9, 11×11, 21×21, 41×41 and 61×61), whose kernel predictions are shown in the right purple rectangles. (b) shows the corresponding HR image.

attention network. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020. 4

[8] 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 *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–125, 2017. 1

[9] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018. 4

[10] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018. 1

[11] 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 701–710, 2018. 1

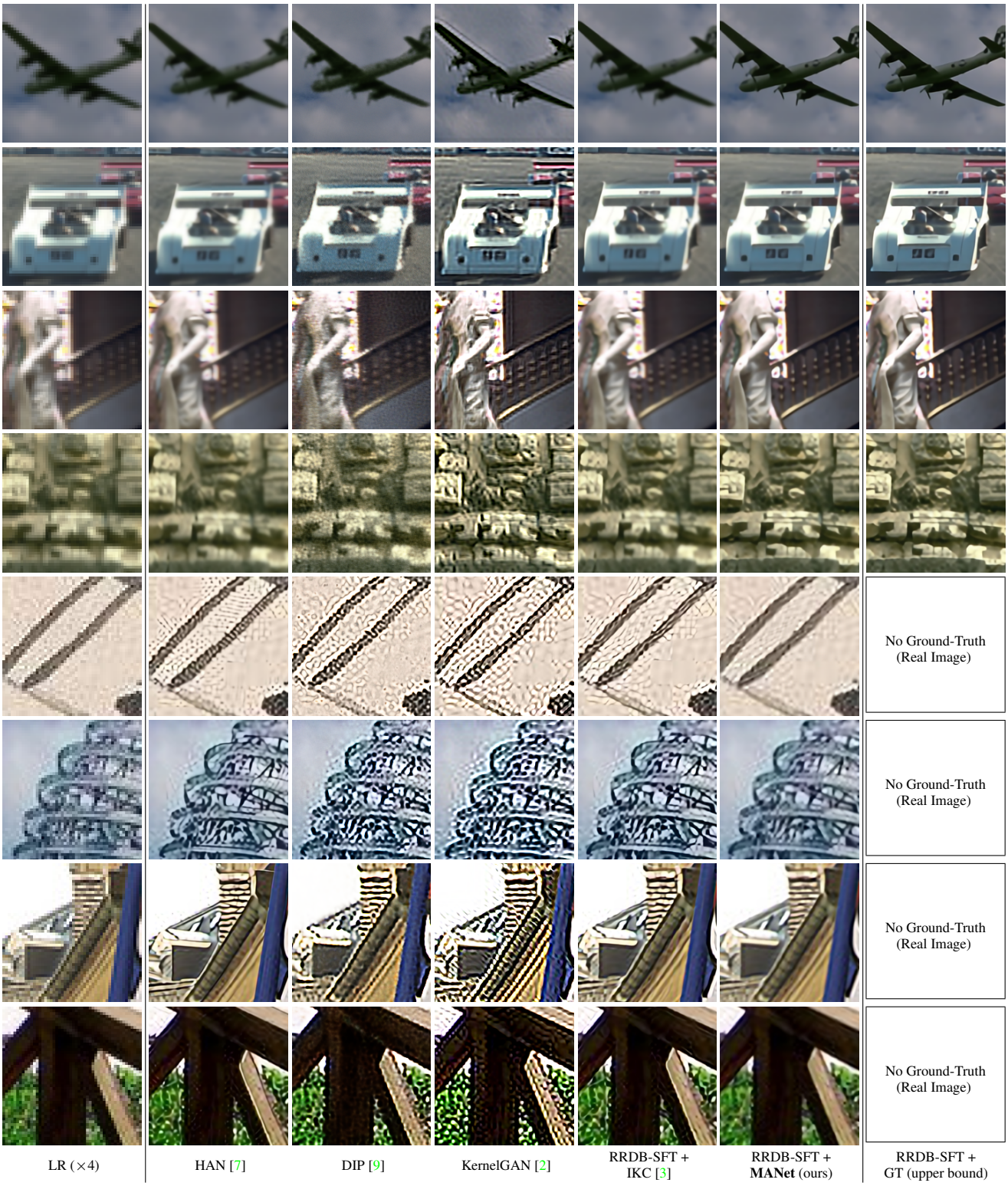


Figure 5: More visual results of different methods on synthetic and real-world images for scale factor 4.