KNN Local Attention for Image Restoration - Supplementary material

Hunsang Lee[1] , Hyesong Choi[2] , Kwanghoon Sohn[1] , Dongbo Min[2†]
[1]Yonsei University, Korea, [2]Ewha W. University, Korea
hslee91@yonsei.ac.kr, hyesongchoi2010@gmail.ac.kr, khsohn@yonsei.ac.kr, dbmin@ewha.ac.kr

In this document, we provide more comprehensive results not provided in the original manuscript due to the page limit as below. The code to reproduce our results will be publicly available soon.

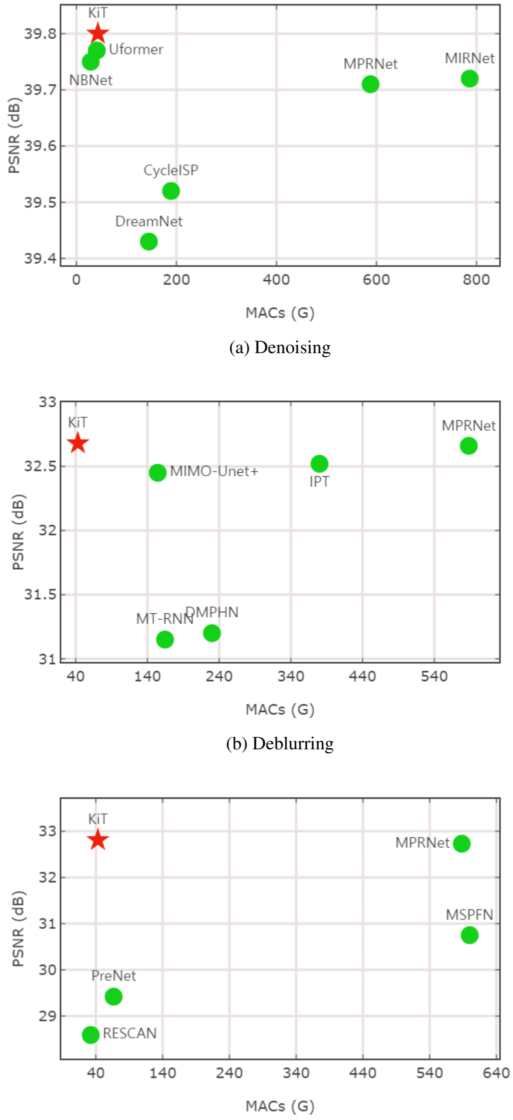
- Comparisons of computational cost with state-of-the-art image restoration methods (Section 1)
- Ablation study on pre-training the network (Section 2)
- More detailed architecture of the KiT (Section 3)
- More qualitative evaluation results for image restoration tasks with state-of-the-arts methods (Section 4)
- Performance evaluation for image deblurring with JPEG artifacts on REDS dataset (Section 5)

1. Computational cost

We first provide the performance comparison with state-of-the-art image restoration methods with respect to the accuracy and computational cost. Fig. 1 shows the graphs illustrating both the performance and computational cost of state-of-the-art methods. The proposed method is marked with a star symbol with red color, and other methods are marked with a circle symbol with green color. The x-axis and y-axis of the graphs represent the computational cost measured with Multiply-Accumulates (MACs) and the performance with the PSNR, respectively. The MACs of all graphs are measured when an input resolution is 256×256. In the image denoising on the SIDD dataset [1], the proposed method has comparable computational cost with Uformer [14] and NBNet [5], while achieving the best performance. In the image deraining and deblurring, the KiT shows a slightly better performance yet with much less computational cost. Compared to the MPRNet [16], the KIT has almost 92.7% fewer MACs.

2. Pre-training

Although the proposed method achieves state-of-the-art performance in various restoration tasks, it is well-known



(a) Denoising

(b) Deblurring

(c) Deraining

Figure 1. Performance *vs.* computational cost of state-of-the-art methods for the image restoration tasks.

Method	SIDD	
	PSNR	SSIM
KiT	39.80	0.972
KiT†	39.85	0.974

Table 1. Effectiveness of the pre-training strategy

that the transformer architectures with pre-trained models using large-scale dataset have greater potential than those learned from scratch [7]. However, as relatively small datasets were available in low-level tasks, pre-training strategies mostly have been addressed in high-level tasks. IPT [3] first investigated that this strategy is also beneficial for low-level image processing by leveraging ImageNet dataset as the baseline dataset for pre-training their model. Following this approach, we trained the network using ImageNet dataset with synthetic Gaussian noise and fine-tuned on SIDD [1] dataset for image denoising. Tab. 1 shows the effectiveness of the pre-training approach. '†' means the model fine-tuned on small target dataset (SIDD) after trained with large-scale dataset (ImageNet). By pre-training the network, the results were further improved by 0.05 dB in terms of PSNR.

3. Detailed architecture

We provide the detailed architecture of the proposed KiT in Tab. 2. We omit an explanation of the decoder as it is the mirrored architecture of the encoder. As described in the original manuscript, each stage consists of the patch partition, k-NN transformer blocks (KTB) and an interpolation layer. The input feature maps are first splitted into non-overlapping N patches with the size of r^2, where $N = HW/r^2$. The chunk size k (equal to the number of NN patches) and patch size r were set to 4 in all stages of the encoder and decoder. In the bottleneck stage, k is set to 1 since there are only a few patches ($N/64$). As the stage progresses, the input feature resolution gradually decreases double in the encoder, and increases double in the decoder. The channel size of the input feature maps increases/decreases in contrast to resolution. The KTB consists of a sequence of layer normalization (LN), k-NN local attention (KLA), LN and feed-forward network (FFN). As the interpolation layer (downsample/upsample) is 2D spatial operation, the input feature map, size of (# of patches) × (patch size) × (channel size), is reshaped to (height) × (width) × (channel size) at the end of the KTB.

4. Qualitative results

In this section, we conducted more visual comparisons with the state-of-the-art methods not provided in the original manuscript due to the page limit. The qualitative results of the image denoising, image deblurring, and image deraining are shown in the Fig. 2, Fig. 3, and Fig. 4, respectively. Similar to the visual results of the original manuscripts, our method successfully restores degraded images with fine structures thanks to the capability of capturing locality with non-local connectivity.

Encoder			
Stage	Layer	Input Shape	Notes
-	Conv × 3	$H \times W \times 3$	-
1	Patch Partition	$H \times W \times C$	$r = 4$
	KTB=[LN KLA LN FFN]	$N \times r^2 \times C$	$k = 4, b = 2$
	Downsample	$H \times W \times C$	-
2	Patch Partition	$\frac{H}{2} \times \frac{W}{2} \times 2C$	$r = 4$
	KTB=[LN KLA LN FFN]	$\frac{N}{4} \times r^2 \times 2C$	$k = 4, b = 2$
	Downsample	$\frac{H}{2} \times \frac{W}{2} \times 2C$	-
3	Patch Partition	$\frac{H}{4} \times \frac{W}{4} \times 4C$	$r = 4$
	KTB=[LN KLA LN FFN]	$\frac{N}{16} \times r^2 \times 4C$	$k = 4, b = 2$
	Downsample	$\frac{H}{4} \times \frac{W}{4} \times 4C$	-
Bottleneck			
Stage	Layer	Input Shape	Notes
-	Patch Partition	$\frac{H}{8} \times \frac{W}{8} \times 8C$	$r = 4$
	KTB=[LN KLA LN FFN]	$\frac{N}{64} \times r^2 \times 8C$	$k = 1, b = 2$

Table 2. Detailed architecture of the KiT. k-NN transformer block (KTB) consists of layer noramlization (LN), k-NN local attention (KLA) and feed-forward network (FFN).

5. Deblurring with JPEG artifacts

To verify the effectiveness of the proposed method, we further evaluated it with the image deblurring task on the REDS [11] dataset with JPEG artifacts. Namely, we restore an input image by removing both blur and compression artifacts. The REDS dataset contains 300 video sequences, where each sequence consists of 100 images with JPEG and blurry artifacts. For training, 24,000 images obtained with random cropping of 128 × 128 in the REDS dataset were used. Since a quantitative comparison was available only

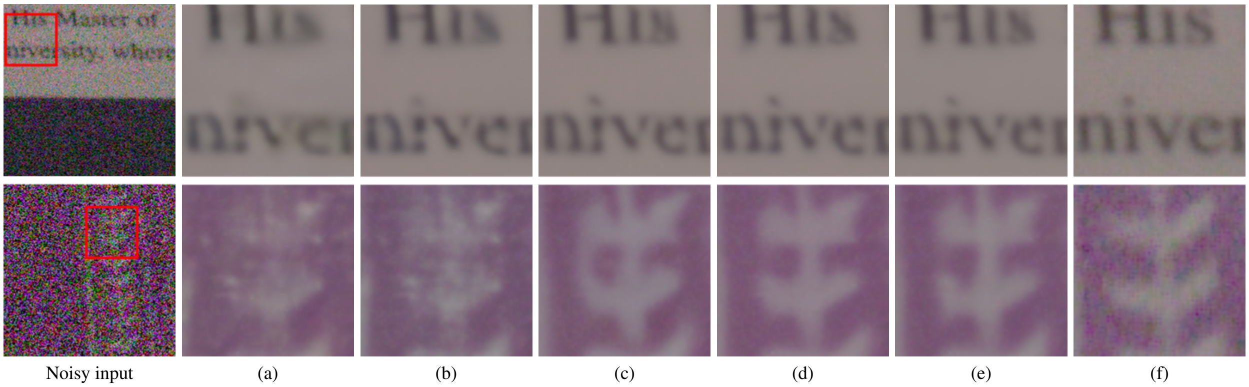


Figure 2. Visual results of image denoising: (a) RIDNet [2], (b) CycleISP [15], (c) MPRNet [16], (d) Uformer [14], (e) KiT, and (f) ground truth.

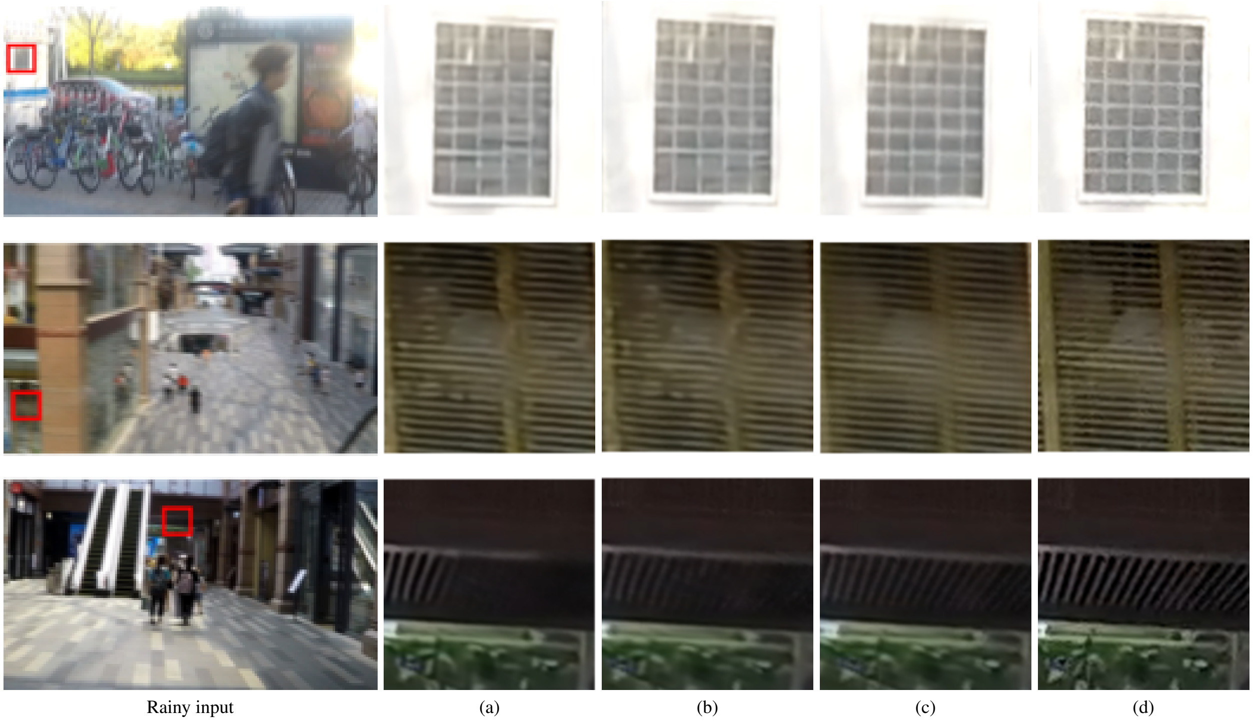


Figure 3. Visual results of image deblurring: (a) MPRNet [16], (b) MIMO-UNet+ [6], (c) KiT, and (d) ground truth.

with the REDS test data (which contains no ground truth) during the NTIRE 2021 challenge [12], we provided the qualitative evaluation results of the REDS validation data in Fig. 5. In the images on the upper row, the numbers of license plate in our result are more visible than others. In the case of the bottom row, the texture of rock is restored more vividly and accurately. Note that, the number of parameters of the KiT (20.6M), is remarkably smaller than those of WRCAN [9] (156.97M) and HINet [4] (88.91M). In addition, our network needs much lower MACs (43.08G) than WRCAN (704.01G) and HINet (170.73G).

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018. 1, 2

[2] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3155–3164, 2019. 3

[3] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and

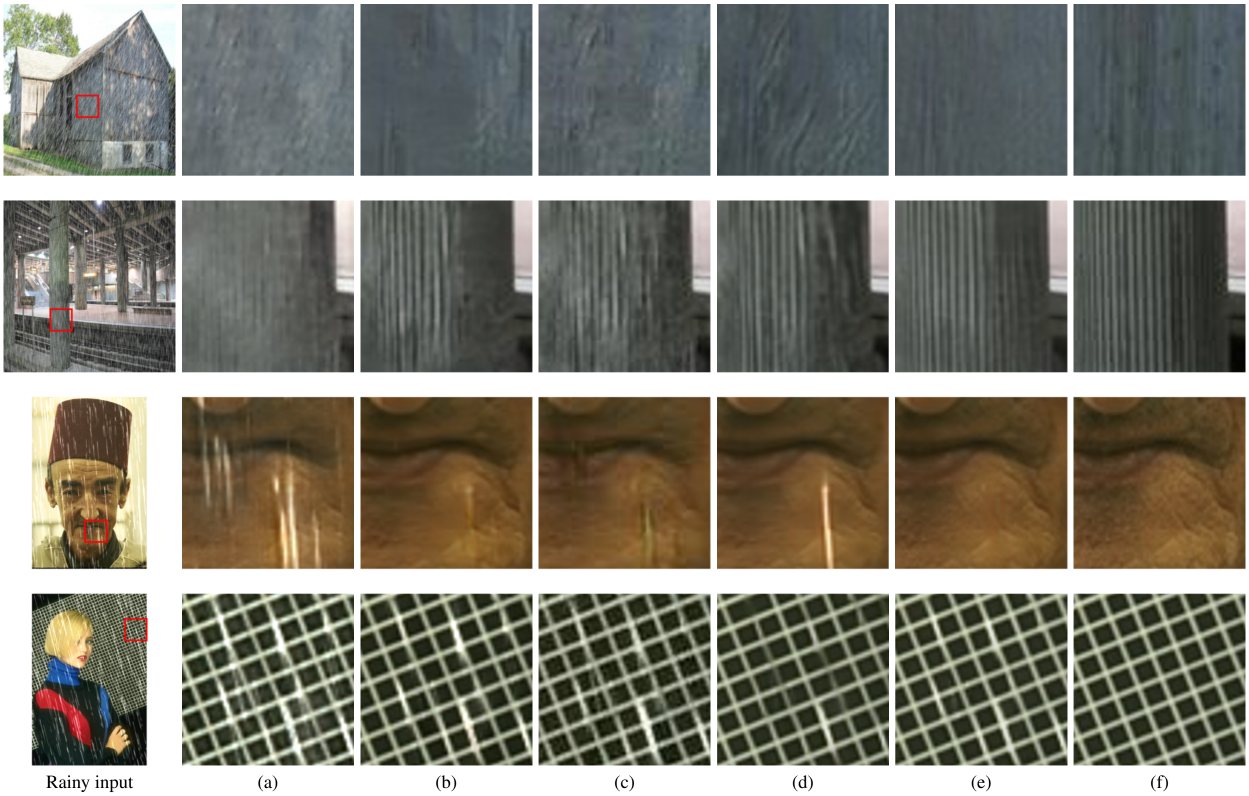


Figure 4. Visual results of image deraining: (a) DerainNet [8], (b) PreNet [13], (c) RESCAN [10], (d) MPRNet [16], (e) KiT, and (f) ground truth.

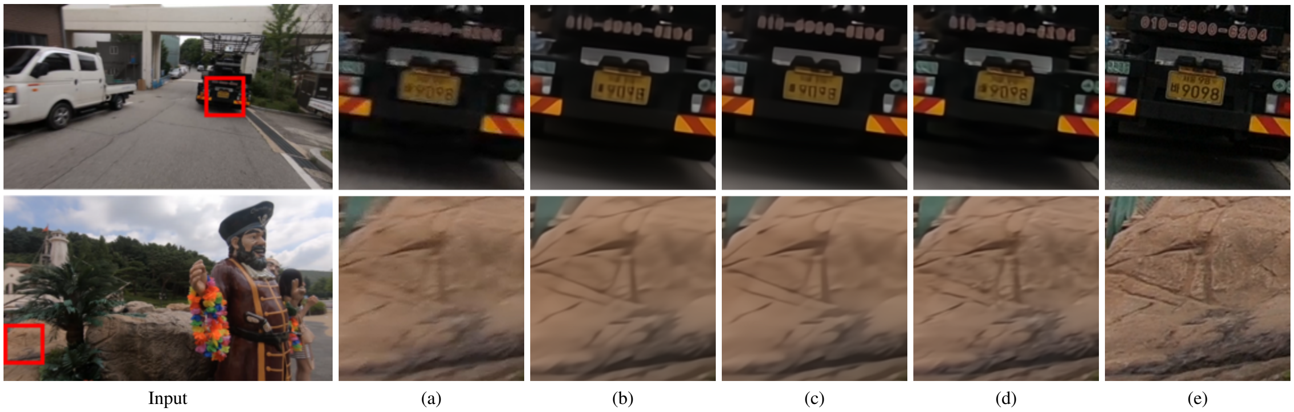


Figure 5. Visual comparisons on the REDS [11] dataset: (a) cropped image, (b) WRCAN [9], (c) HINet, (d) KiT, and (e) ground truth.

Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 2

[4] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021. 3

[5] Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, and Shuaicheng Liu. Nbnet: Noise basis learning for image denoising with subspace projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4896–4906, 2021. 1

[6] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4641–4650, 2021. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[8] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 4

[9] Donghyeon Lee, Chulhee Lee, and Taesung Kim. Wide receptive field and channel attention network for jpeg compressed image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2021. 3, 4

[10] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 254–269, 2018. 4

[11] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 4

[12] Seungjun Nah, Sanghyun Son, Suyoung Lee, Radu Timofte, and Kyoung Mu Lee. Ntire 2021 challenge on image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 149–165, 2021. 3

[13] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3937–3946, 2019. 4

[14] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv preprint arXiv:2106.03106*, 2021. 1, 3

[15] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2696–2705, 2020. 3

[16] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021. 1, 3, 4