Pre-Trained Image Processing Transformer (Supplementary Material)

1. Results on Deblurring

We further evaluate the performance of our model on image deblurring task. We use the GoPro dataset [4] to fine-tune and test our model. We modify the patch size as 256, patch dim as 8 and number of features as 9 to achieve a higher receptive field. Table 1 reported deblurring results, where $^+$ denotes applying self-ensemble technique. As a result, our IPT achieves the best results among all deblurring methods. Figure 2 shows the visualization of the resulted images. As shown in the figure, our pre-trained model can well achieve the best visual quality among all the previous models obviously.

2. Architecture of IPT

In the main paper, we propose the image processing transformer (IPT). Here we show the detailed architecture of IPT, which consists of heads, body and tails. Each head has one convolutional layer (with 3×3 kernel size, 3 input channels and 64 output channels) and two ResBlock. Each ResBlock consists of two convolutional layers (with 5×5 kernel size, 64 input channels and 64 output channels) which involved by a single shortcut. The body has 12 encoder layers and 12 decoder layers. The tail of denoising or deraining is a convolutional layer with 3×3 kernel size, 64 input channels and 3 output channels. For super-resolution, the tail consists of one pixelshuffle layer with upsampling scale 2 and 3 for $\times 2$ and $\times 3$ SR, two pixelshuffle layer with upsampling scale 2 for $\times 4$ SR.

The whole IPT has $114M$ parameters and $33G$ FLOPs, which have more parameters while fewer FLOPs compared with traditional CNN models (*e.g.*, EDSR has $43M$ parameters and $99G$ FLOPs).

3. Impact of Multi-task Training

We train IPT following a multi-task manner and then fine-tune it on 6 different tasks including $\times 2$, $\times 3$, $\times 4$ super-resolution, denoising with noise level 30,50 and deraining. We find that this training strategy would not harm the performance on these tasks which have been pre-trained on large scale dataset (ImageNet). In other words, the performance of multi-task training and single-task training remains almost the same. However, when transferring to other tasks (*e.g.*, Section 4.4 in the main paper), the pre-trained model using multi-task training is better than that of single-task training for about 0.3dB, which suggests the multi-task training would learn universal representation of image processing tasks.

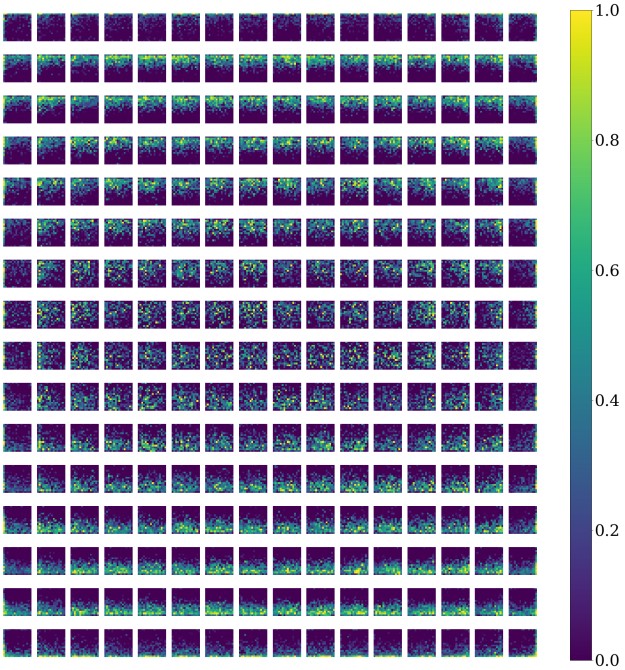


Figure 1. Visualization of cosine similarity of position embeddings.

4. Visualization of Embeddings

We visualize the learned embeddings of IPT. Figure 1 shows the visualization results of position embeddings. We find that patches with similar columns or rows have similar embeddings, which indicate that they learn useful information for discovering the position on image processing. We also test to use fixed embeddings or do not use embeddings, whose performance are lower than that of using learnable position embeddings (vary from 0.2dB to 0.3dB for different tasks).

Moreover, we visualize the task embeddings in figure 3. We can find that for $\times 2$ super-resolution task, the simi-

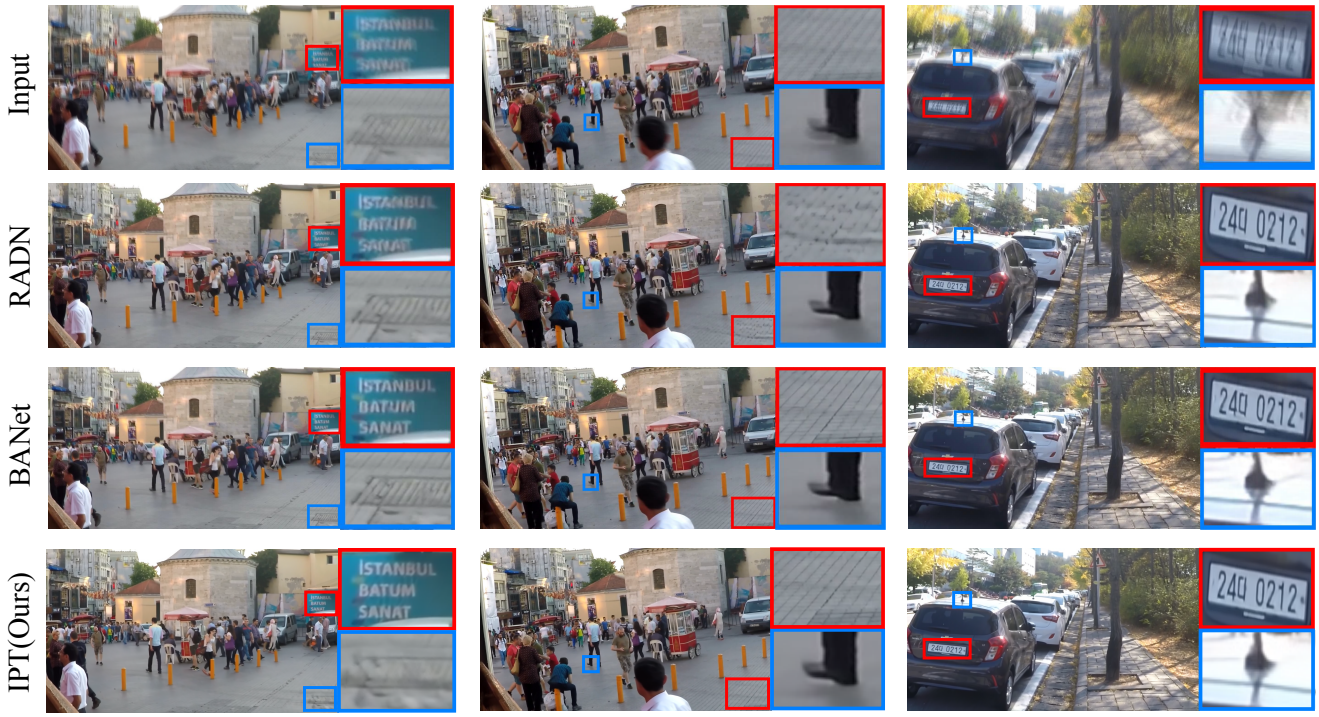


Figure 2. Image deblurring results on the GoPro dataset. Compared images are derived from [10].

Table 1. Quantitative results on image deblurring. Best and second best results are **highlighted** and underlined.

Method	MSCNN [4]	SRN [9]	DSD [1]	DeblurGANv2 [3]	DMPHN [12]	LEBMD [2]	EDSD [11]
PSNR	30.40	30.25	30.96	29.55	31.36	31.79	29.81
DBGAN [13]	MTRNN [6]	RADN [7]	SAPHN [8]	BANET [10]	MB2D [5]	IPT (Ours)	IPT$^+$ (Ours)
31.10	31.13	31.85	32.02	32.44	32.16	32.58	**32.91**

larity between the embeddings on each position and their neighbours are higher than $\times 3$ super-resolution, while that of $\times 4$ super-resolution is the smallest. This results indicates that each patches in $\times 2$ super-resolution can focus on other patches with farther distance than $\times 3$ and $\times 4$, since their downsampling scale are smaller and the relationship between different patches are closer. The similarity of task embedding for deraining in figure 3 (d) shows that the patches pay more attention on the vertical direction than horizontal direction, which is reasonable as the rain is dropped vertically. The similarity of task embedding for denoising is similar with Gaussian noise, and figure 3 (f) with higher (50) noise level shows higher similarity between neighbours than figure 3 (e) with 30 noise level. The visualization results suggests that our task embeddings can indeed learn some information for different tasks. We also test to not use task embeddings, which results in significant accuracy drop (vary from 0.1dB to 0.5dB for different tasks).

References

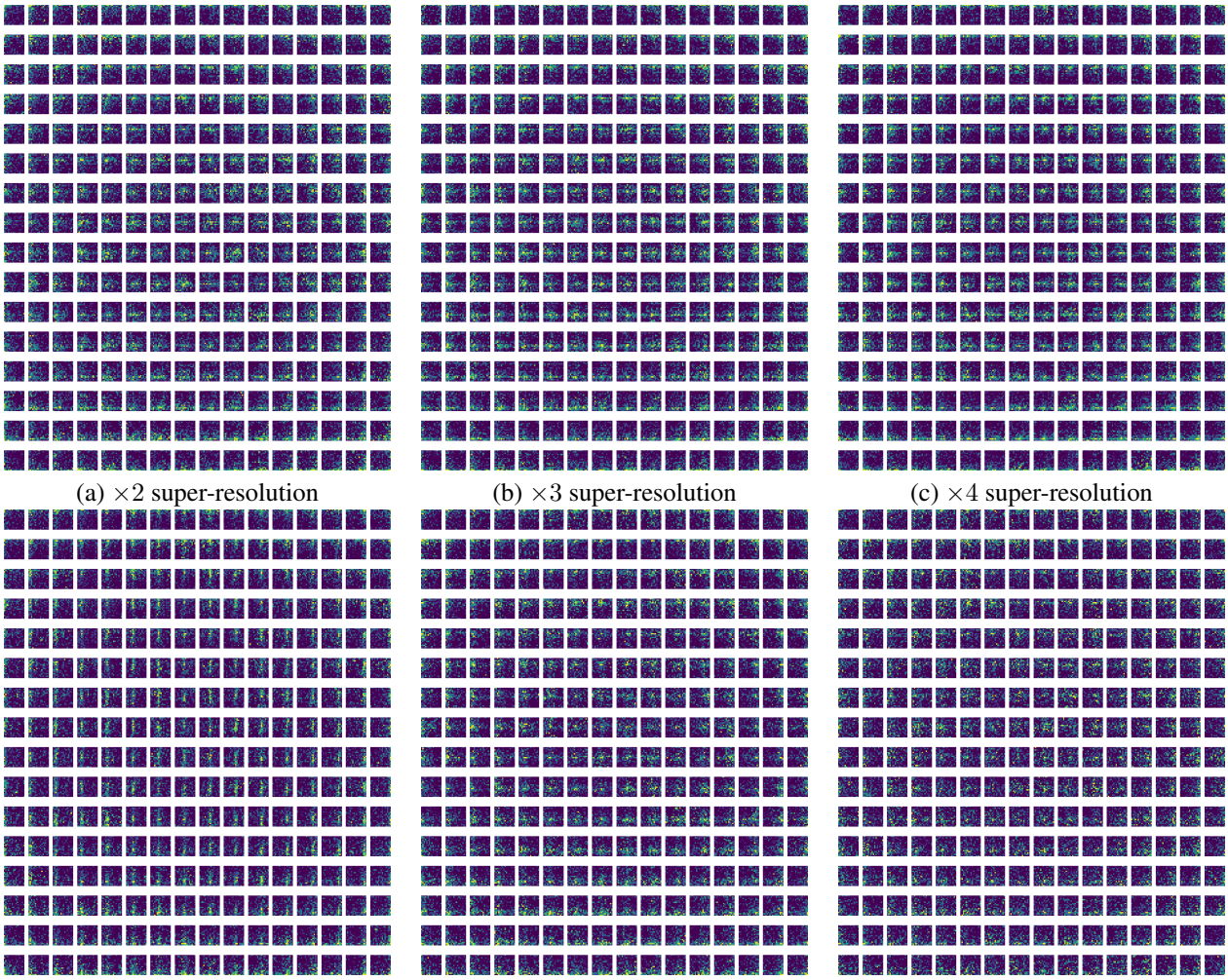
[1] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3848–3856, 2019. 2

[2] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3329, 2020. 2

[3] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8878–8887, 2019. 2

[4] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 1, 2

[5] Dongwon Park, Dong Un Kang, and Se Young Chun. Blur more to deblur better: Multi-blur2deblur for efficient video deblurring. *arXiv preprint arXiv:2012.12507*, 2020. 2



(a) $\times 2$ super-resolution (b) $\times 3$ super-resolution (c) $\times 4$ super-resolution

(d) deraining (e) denoising with 30 noise level (f) denoising with 50 noise level

Figure 3. Visualization of six different task embeddings.

[6] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *European Conference on Computer Vision*, pages 327–343. Springer, 2020. 2

[7] Kuldeep Purohit and AN Rajagopalan. Region-adaptive dense network for efficient motion deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11882–11889, 2020. 2

[8] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3606–3615, 2020. 2

[9] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018. 2

[10] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Banet: Blur-aware attention networks for dynamic scene deblurring. *arXiv preprint arXiv:2101.07518*, 2021. 2

[11] Yuan Yuan, Wei Su, and Dandan Ma. Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3555–3564, 2020. 2

[12] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5978–5986, 2019. 2

[13] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2737–2746, 2020. 2