A Comparison with general image restoration methods

We compare our TAPE-Net with two models, DnCNN [21] and UNet [17], which are widely used in image restoration tasks. As shown in Table S1, our model outperforms these two models on all datasets. Note that for a fair comparison, we set the compared models with a similar number of FLOPs. The parameters of DnCNN, UNet, and our TAPE are 0.5M, 9.8M, and 1.3M, respectively.

Table S1: Quantitative comparison for three models (in terms of PSNR (dB)).

Dataset	Rain200L	Rain200H	Raindrop800	SIDD	TIP2018	Snow100K	ISTD	FLOPs
DnCNN [21]	27.73	19.20	26.12	34.31	24.74	23.77	23.21	14.4 G
UNet [17]	31.50	22.50	26.41	33.72	25.58	23.48	25.64	18.2 G
TAPE-Net (Ours)	**33.17**	**23.84**	**27.69**	**37.90**	**27.52**	**26.33**	**26.57**	14.1 G

B Comparison with task-specific methods

As shown in Table S2, we firstly compare our methods with state-of-the-art denoising methods (DnCNN [21], FFDNet [22], RDN [23], and SADNet [1]). Note that the FLOPs number of our method (14.1G) is much smaller than that of RDN (46.6G) or SADNet (45.8G), when the input is a 64 × 64 RGB image. We replace our backbone model (original transformer) with a specially designed transformer, Swin transformer [14] (termed as 'TAPE-Net-swin-L' in Table S2), which outperforms all the SOTA methods with our pre-training strategy.

And we also compare with SOTA deraining methods in Table S3 and compare with SOTA demoireing methods in Table S4. Our method outperforms all these task-specific SOTA methods.

Table S2: Quantitative comparison with the state-of-the-art denoising methods on SIDD.

Method	DnCNN [21]	FFDNet [22]	RDN [23]	SADNet [1]	TAPE-Net (Ours)	TAPE-Net-swin-L (Ours)
PSNR/SSIM	34.31/0.892	33.26/0.890	38.70/0.901	38.41/0.900	37.90/0.896	38.76/0.901
FLOPS (G)	14.4	0.87	46.6	45.8	14.1	5.3

Table S3: Quantitative comparison with the state-of-the-art deraining methods on Rain200L and Rain200H. The best result are in **Bold**.

Dataset	Method	DDN [6]	SPANet [19]	RESCAN [12]	PreNet [16]	BRN [15]	PCNet [10]	TAPE-Net (Ours)
Rain200L	PSNR/SSIM	28.35/0.878	30.92/0.930	32.07/0.949	31.98/0.948	32.40/0.953	32.62/0.954	**33.17/0.959**
Rain200H	PSNR/SSIM	20.98/0.705	22.65/0.714	23.04/0.729	23.27/0.743	23.39/0.755	23.43/0.755	**23.84/0.759**

Table S4: Quantitative comparison with the state-of-the-art dermoireing methods on TIP2018. The best result are in **Bold**.

Method	DMCNN [18]	MopNet [8]	HRDN [20]	FHDe2Net [7]	WDNet [13]	MBCNN [24]	TAPE-Net (ours)
PSNR/SSIM	25.82/0.806	26.20/0.861	26.68/0.864	26.25/0.862	26.86/0.865	27.37/0.866	**27.52/0.866**

C Additional visualization results.

C.1 Visualization some results of the prior queries, Q.

We also visualize some other results of the prior queries (Q) of TAPE. As shown in Fig. S1, (a) and (b) are rain inputs and ground truth respectively; (c) are one of the predicted results of PLM. We can see that with the help of pre-training, the PLM module can correlate the information of similar textures or patches from a long distance. Thus, the transformer decoder of the backbone can utilize these long-distance similar areas/patches to restore the image.

C.2 Visualization some results of learned parameters (e).

We visualize the learned parameters (e) of TAPE. Fig. S2 shows some visualization results of learned parameters (e) and the position embeddings of IPT (these results are copied from [2]). We can find that our learned parameters (e) focus on other patches with farther distances than the position embeddings of IPT. Besides, our learned feature maps show richer patterns. For example, some of the patches focus on the four corners of the image at the same time (the patch on the 3rd row and 2nd column). Some of the patches focus on the characters of oblique directions (the patch on the 1st row and 5th column). These rich feature maps do not appear in the visualization results of IPT.

D Additional results on other transformer backbones.

Our TAPE is a widely applicable method, where the backbone can be replaced by other transformer backbones. We replace our backbone network with the swin transformer backbone [14] for experiments. As shown in Table S5, 'Baseline-swintrans' is the Swin transformer backbone without pre-training. 'TAPE-Net-swintrans-S', 'TAPE-Net-swintrans-M', and 'TAPE-Net-swintrans-L' are our TAPE-Nets with the Swin transformer backbone and contain 1 Swin block, 3 Swin

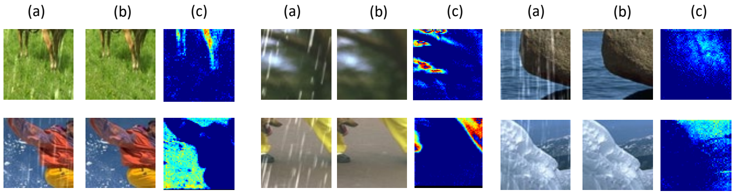
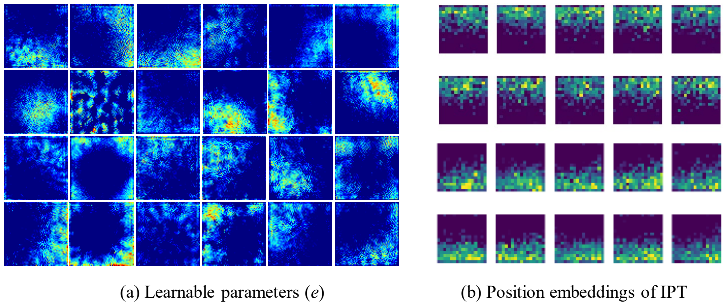


Fig. S1: Visualization of some results of the prior queries, Q on the Rain200L dataset. (a) Rain inputs. (b) Ground truth. (c) one of the predicted results of PLM.


(a) Learnable parameters (e)

(b) Position embeddings of IPT

Fig. S2: Visual comparison between our learnable parameters and the position embeddings of IPT. Our learnable parameters show more richer patterns.

blocks and 5 Swin blocks, respectively. The results illustrate that our 3-stage pre-training and adding blocks can significantly boost the performance.

E Ablation study about optimization in task-specific fine-tuning.

In Sec. 3.2.1, there are two ways to optimize the networks in the task-specific fine-tuning stage, namely: 1) The backbone ϕ is fine-tuned by loss between pseudo GT and GT firstly, and then fixed when fine-tuning other networks (denoted as Step by Step finetuning in Table S6); 2) All components are fine-tuned simultaneously (denoted as Joint finetuning in Table S6). As shown in Table S6, the performance of the two methods is equivalent. How to choose different optimization methods for different tasks is future work.

F Ablation Study of pixel-wise contrastive loss.

We remove the pixel-wise contrastive loss in the task-agnostic pre-training. And the PSNR/SSIM decrease by 0.12dB/0.001 on Rain200L without the proposed pixel-wise contrastive loss.

Table S5: Quantitative comparison for Baseline-swintrans and TAPE-Net-trans (in terms of PSNR (dB)). The numbers in () of the 2nd line are the PSNR gain compared with 'Baseline-swintrans'.

	Blocks numbers	Network parameters	Rain200L (dB)	SIDD (dB)	Raindrop800 (dB)
Baseline-swintrans	1	0.19M	33.52	38.01	27.79
TAPE-Net-swintrans-S	1	0.19M	34.07 (+0.55)	38.76 (+0.75)	28.31 (+0.52)
TAPE-Net-swintrans-M	3	0.61M	34.20	38.87	28.97
TAPE-Net-swintrans-L	5	0.97M	34.46	38.98	29.15

Table S6: Comparison between Step by step fine-tuning and Joint fine-tuning.

TAPE-Swin	Rain200L	Rain200H	Raindrop800-TestB
Step by step fine-tuning	35.10	26.18	26.41
Joint fine-tuning	35.06	26.05	26.49

G Ablation Study about Transformer or CNN.

We replace the transformer encoder and decoder with the ResNet encoder and decoder [9] respectively with the same model size. The feature map outputed from the encoder and the feature map outputed from the PLM are concatenated and served as the input of the ResNet decoder. We do the ablation study on the Rain200L dataset with the same setting as Sec. 4.5 of the main paper. The PSNR/SSIM drops 1.03dB/0.006 compared with the baseline with the transformer encoder and decoder. The result shows that using the transformer is better when fusing the information of the output of PLM and the encoder. We also made a comparison between pre-trained CNN and no-pre-trained CNN. Compared with transformer, CNN's performance improvement is much smaller.

H Visual comparison with other SOTA methods on desnowing and shadow removal

We compare our method with several state-of-the-art desnowing and shadow removal methods. Fig. S3 and Fig. S4 show our method can remove the shadow or snow. Please note that these compared methods use the snow/shadow masks for training, while our method only uses snow/snow-free or shadow/shadow-free image pairs.

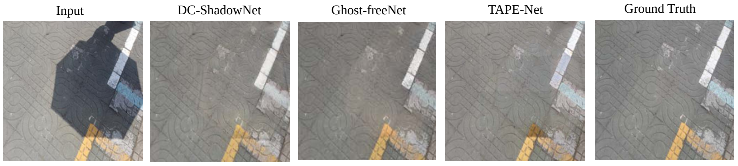


Fig. S3: Visual shadow removal comparison among ours and two other methods (DC-ShadowNet [11] and Ghost-freeNet [5]).

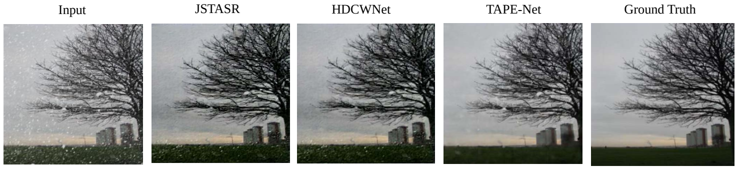


Fig. S4: Visual desnowing comparison among ours and two other methods (JS-TASR [3] and HDCWNet [4]).

References

1. Chang, M., Li, Q., Feng, H., Xu, Z.: Spatial-adaptive network for single image denoising. ECCV, 2020
2. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. CVPR, 2021
3. Chen, W.T., Fang, H.Y., Ding, J.J., Tsai, C.C., Kuo, S.Y.: Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In: ECCV, 2020
4. Chen, W.T., Fang, H.Y., Hsieh, C.L., Tsai, C.C., Chen, I., Ding, J.J., Kuo, S.Y., et al.: All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In: CVPR, 2021
5. Cun, X., Pun, C.M., Shi, C.: Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In: AAAI, 2020

6. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: CVPR, 2017
7. He, B., Wang, C., Shi, B., Duan, L.Y.: Fhde2net: Full high definition demoireing network. ECCV, 2020
8. He, B., Wang, C., Shi, B., Duan, L.Y.: Mop moire patterns using mopnet. In: ICCV, 2019
9. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: Resunet++: An advanced architecture for medical image segmentation. In: ISM, 2019
10. Jiang, K., Wang, Z., Yi, P., Chen, C., Lin, C.W.: Pcnet: Progressive coupled network for real-time image deraining. In: TIP, 2021
11. Jin, Y., Sharma, A., Tan, R.T.: Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In: ICCV, 2021
12. Li, X., Wu, J., Lin, Z., Liu, H., Zha, H.: Recurrent squeeze-and-excitation context aggregation net for single image deraining. In: ECCV, 2018
13. Liu, L., Liu, J., Yuan, S., Slabaugh, G., Leonardis, A., Zhou, W., Tian, Q.: Wavelet-based dual-branch network for image demoiréing. ECCV, 2020
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. CVPR, 2021
15. Ren, D., Shang, W., Zhu, P., Hu, Q., Meng, D., Zuo, W.: Single image deraining using bilateral recurrent network. TIP, 2020
16. Ren, D., Zuo, W., Hu, Q., Zhu, P., Meng, D.: Progressive image deraining networks: A better and simpler baseline. In: CVPR, 2019
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI, 2015
18. Sun, Y., Yu, Y., Wang, W.: Moiré photo restoration using multiresolution convolutional neural networks. TIP, 2018
19. Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q., Lau, R.W.: Spatial attentive single-image deraining with a high quality real rain dataset. In: CVPR, 2019
20. Yang, S., Lei, Y., Xiong, S., Wang, W.: High resolution demoire network. In: ICIP, 2020
21. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. TIP, 2017
22. Zhang, K., Zuo, W., Zhang, L.: Ffdnet: Toward a fast and flexible solution for CNN based image denoising. TIP, 2018
23. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image restoration. TPAMI, 2020
24. Zheng, B., Yuan, S., Yan, C., Tian, X., Zhang, J., Sun, Y., Liu, L., Leonardis, A., Slabaugh, G.: Learning frequency domain priors for image demoireing. TPAMI, 2021