Supplementary Materials for "MB-TaylorFormer: Mutil-branch Efficient Transformer Expanded by Taylor Formula for Image Dehazing"

Yuwei Qiu[1] Kaihao Zhang[2] Chenxi Wang[1] Wenhan Luo[1] Hongdong Li[2] Zhi Jin[1]

[1] Sun Yat-sen University [2] Australian National University

Abstract

This is the supplementary material for the paper:"MB-TaylorFormer: Mutil-branch Efficient Transformer Expanded by Taylor Formula for Image Dehazing" submitted to the ICCV 2023. We first present an efficient implementation of Taylor expanded multi-head self-attention (T-MSA) via the pseudo-code. Next, we provide details on the configuration of our multi-scale path embedding, multi-scale attention refinement (MSAR) module, and two MB-TaylorFormer variants. Besides, we further explore the ablation experiments on the module and model and present visual comparison of the ablation studies for each module. In the end, we present more qualitative comparison on datasets of RESIDE [8], CSD [2], and RainCityscapes [6].

1. Efficient Implementation of T-MSA

In the main text, we give the vector-based calculation of T-MSA, where V_i' is the output of the T-MSA, $\tilde{Q}_i = \text{Norm}(Q_i)$ and $\tilde{K}_i = \text{Norm}(K_i)$, as follows:

$$\begin{aligned} V_i' &= \text{Taylor-Attention}\left(Q_i, K_i, V_i\right) \\ &= \frac{\sum_{j=1}^{N} V_j^T + \tilde{Q}_i^{\,T} \sum_{j=1}^{N} \tilde{K}_j V_j^T}{N + \tilde{Q}_i^T \sum_{j=1}^{N} \tilde{K}_j}. \end{aligned} \quad (1)$$

Algorithm 1 is the pseudo-code for the matrix implementation of T-MSA, which implements efficient self-attention operations.

Algorithm 1: Pseudocode of T-MSA in a PyTorch-like style.

input : A feature map I_f of shape $b \times c \times h \times w$
output: A feature map O_f of shape $b \times c \times h \times w$
$I_f{}' = \text{dwconv}(\text{project}(I_f))$ # $I_f' : b \times 3c \times h \times w$
$I_f{}'' : b \times head \times hw \times \frac{3c}{head}$
$I_f{}'' = rearrange(I_f')$
$Q, K, V : b \times head \times hw \times \frac{c}{head}$
$Q, K, V = \text{chunk}(I_f{}'', chunks = 3, dim = -1)$
$Q' = \text{normalize}(Q, dim = -1)$
$K' = \text{normalize}(K, dim = -1)$
mm: matrix multiplication
$numerator : b \times head \times hw \times \frac{c}{head}$
$K\_V = \text{mm}(K.\text{view}(b, head, \frac{c}{head}, hw), V))$
$Q\_K\_V = \text{mm}(Q, K\_V)$
$numerator = \text{sum}(V, dim = -2).\text{unsqueeze}(2)$
$+Q\_K\_V$
$denominator : b \times head \times hw \times \frac{c}{head}$
$K\_Ones = \text{sum}(K.\text{view}(b, head, \frac{c}{head}, h),$
$dim = -2).\text{unsqueeze}(2)$
$denominator = h \times w + K\_Ones + 1e^{-6}$
$att : b \times head \times hw \times \frac{c}{head}$
att=div($nominator$, $denominator$)
$att' : b \times c \times h \times w$
att' = rearrange(att)
out = project(att)

2. More Details on The Configuration

2.1. Details on The Configuration of Multi-scale Patch Emebdding

Fig. 1 is used to demonstrate the transition from the single-scale and single-branch patch embedding (Fig. 1a) to the proposed multi-scale patch embedding (Fig. 1e) to better illustrate our contributions. "-S" means two convolutional layers in series with the same kernel size of 3 to equal the convolution with the kernel size of 5, and the multi-scale patch embedding is paralleled; "-P" means two convolutional layers in parallel with the same kernel size of 3.

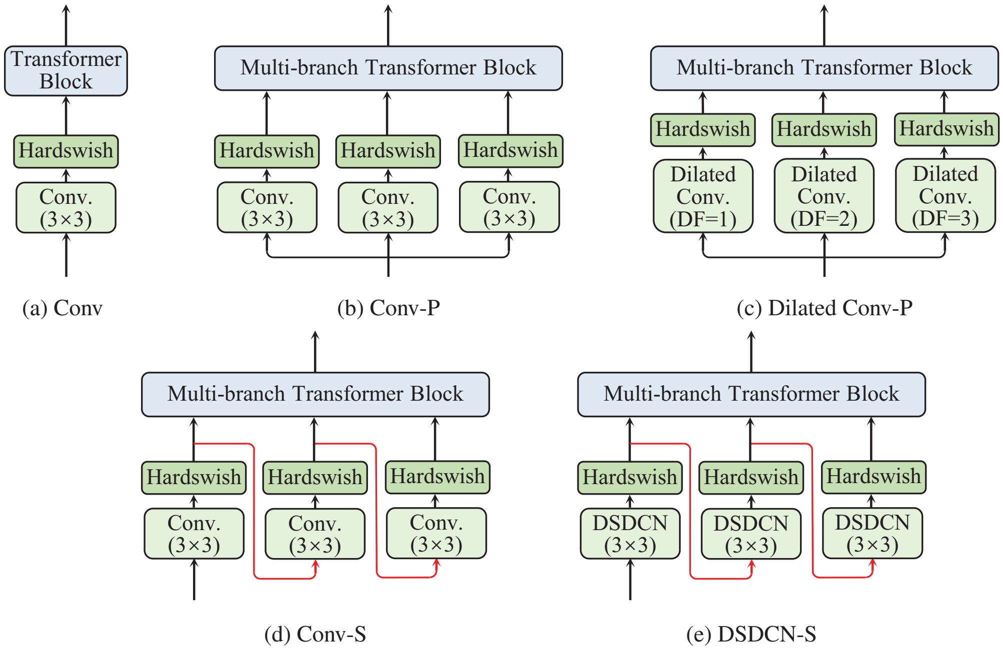


Figure 1: **The structure of patch embedding.**

Table 1: **Detailed structural specification of two variants of our MB-TaylorFormer.**

Model	Num. of Branches	Num. of Blocks	Num. of Channels	Num. of heads	#Params	MACs
MB-TaylorFormer-B	[2,2,2,2,2,2,2,2]	[2,3,3,4,3,3,2,2]	[24,48,72,96,72,48,24,24]	[1,2,4,8,4,2,1,1]	2.677M	38.51G
MB-TaylorFormer-L	[2,3,3,3,3,3,2,2]	[4,6,6,8,6,6,4,4]	[48,96,144,192,144,96,48,48]	[1,2,4,8,4,2,1,1]	2.652M	37.89G

Table 2: **Details on The Configuration of Our MSAR Module**

Num. of Heads	Num. of 3×3	Num. of 5×5	Num. of 7×7
1	1	0	0
2	2	0	0
4	2	2	0
8	2	3	3

2.2. Details on The Configuration of Our MB-TaylorFormer

We provide two variants of MB-TaylorFormer (-B and -L for basic and large, respectively) in the main paper. Table 1 lists the detailed configurations of these variants. MB-TaylorFormer-B is a lightweight Transformer network that enables efficient inference while guaranteeing good performance. MB-TaylorFormer-L is more concerned with improving performance while maintaining an appropriate number of parameters and MACs.

2.3. Details on The Configuration of MSAR Module.

In our MB-TaylorFormer, each level of encoder-decoder consists of transformers, where the number of channels increases progressively from top to bottom, we choose the number of heads depending on the number of channels. To strengthen the ability of multi-head to integrate multiple information, we use convolution with multi-scale kernels in our MSAR module. Table 2 presents the detailed configuration of our MSAR module. For example, in the fourth level, we pass the feature maps of the 8 heads through two 3×3 convolutions, three 5×5 convolutions, and three 7×7 convolutions to generate the scaling factor matrix.

3. Additional Ablation Studies

3.1. The truncation range of offsets

We truncate the offsets of the DSDCN, and Table 3 shows the effect of different truncation ranges on the model. We find DSDCN with truncated offsets achieves better performance than DSDCN without truncated offsets. We attribute the improvement to the fact that the generated tokens in our approach focus more on local areas of the

Table 3: **Ablation study for the truncation range of offset and the normalization of q and k.** Appropriate normalization of q and k and inclusion of local correlations for tokens can help improve model performance.

Network	Component		PSNR	SSIM
	Truncation range	Norm of q & k		
Overall	[-3, +3]	0.5	**40.71**	**0.992**
Patch embedding	[-1, +1]	0.5	40.36	**0.992**
	[-2, +2]	0.5	40.51	**0.992**
	[-4, +4]	0.5	40.33	**0.992**
	[-5, +5]	0.5	40.24	0.991
	w/o truncation	0.5	39.16	**0.992**
TaylorFormer	[-3, +3]	1	40.51	**0.992**
	[-3, +3]	0.25	38.89	0.991

feature map. We further investigate the effect of different truncation ranges and finally choose $[-3, 3]$ as the truncation range for MB-TaylorFormer.

3.2. The normalization of q and k

Normalizing q, k to a smaller norm can bring qk^T close to 0, thus making the Taylor expansion more accurate, but it also restricts the value domain of each element in the attention map. We need to find a balance between them. In Table 3, we find that the value of PSNR is highest when we normalize the norm of q and k to 0.5. When normalizing the norm of q and k to 0.25, the value of PSNR decreases significantly. This could probably be attributed to that the value domain of the too-small attention map limits the Transformer expression capability.

3.3. Visual Comparisons for Ablated Models

We further investigate the qualitative comparison of ablation studies on the image dehazing task. The results in Fig. 2 demonstrate all methods we proposed could improve the dehazing performance of our model.

- "SSPE" means single-scale and single-branch patch embedding.
- "W/o truncation" means the offsets of DSDCN loss truncation.
- "W/o MSAR" means the MSAR module is removed.

As shown, SSPE and w/o truncation produce artifacts in the high-frequency region, and w/o MSAR generates coarse details in the result. In contrast, our full model achieves a more visually pleasing result, which produces clearer images and recovers better details. The visual comparisons show the effectiveness of our methods again.

Table 4: **Deeper, wider or more-branch model.** Only the structure of the encoder is given in the table, the decoder and encoder are symmetrical designs.

Component			PSNR	SSIM	#Params	MACs
Branch	Block	Channel				
[2,2,2,2]	[5,8,8,12]	[24,48,72,96]	42.36	**0.994**	7.131M	95.59G
[2,3,3,3]	[4,6,6,8]	[24,48,72,96]	**42.64**	**0.994**	7.432M	88.07G
[2,3,3,3]	[2,3,3,4]	[36,72,100,136]	41.53	**0.994**	7.500M	94.60G
[3,4,4,4]	[4,6,6,8]	[20,40,60,88]	42.01	0.993	7.572M	82.48G

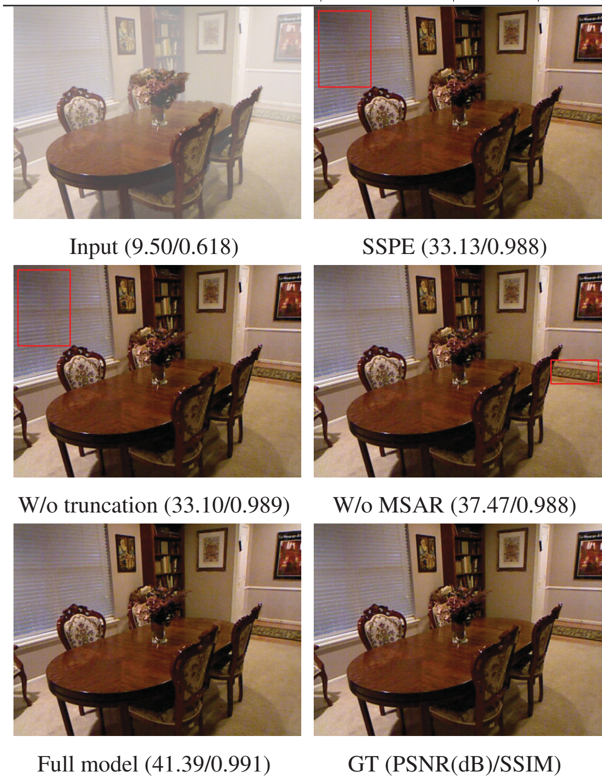


Figure 2: **The qualitative comparison of ablation studies.** The result of our full setting has the best visual quality and details.

4. More Qualitative Comparisons

4.1. The Qualitative Comparison on SOTS

We present more visual results in Fig. 3, Fig. 4, Fig. 5, and Fig. 6. As can be seen, the results of our method have better visual quality.

4.2. The Qualitative Comparison on CSD

In Fig. 7, we provide the visual comparison with other advance models on CSD [2]. We can clearly observed that our MB-TaylorFormer can reconstruct high-quality snow-free image very close to GT. Specifically, the MB-TaylorFormer restores better detail to the image than other methods (as shown in the red box).

4.3. The Qualitative Comparison on RainCityscapes

We illustrate the predictions from rain removal dataset like RainCityscapes [6] in Fig. 8. It can be seen that MB-TaylorFormer achieves visually pleasing results compared to the previous methods. It works very well in removing both fog and rain streaks. It can be seen from Fig. 8 that our method can even restore areas with very severe degradation, while other methods produce severe artifacts.

References

[1] Haoran Bai, Jinshan Pan, Xinguang Xiang, and Jinhui Tang. Self-guided image dehazing using progressive feature fusion. *TIP*, 31:1217 – 1229, 2022.

[2] Wei-Ting Chen, Hao-Yu Fang, Cheng-Lin Hsieh, Cheng-Che Tsai, I Chen, Jian-Jiun Ding, Sy-Yen Kuo, et al. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4196–4205, 2021.

[3] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *CVPR*, pages 2157–2167, 2020.

[4] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *CVPR*, pages 5812–5820, 2022.

[5] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *PAMI*, 33(12):2341–2353, 2010.

[6] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 8022–8031, 2019.

[7] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Zheng Wang, Xiao Wang, Junjun Jiang, and Chia-Wen Lin. Rain-free and residue hand-in-hand: A progressive coupled network for real-time image deraining. *IEEE Transactions on Image Processing*, 30:7404–7418, 2021.

[8] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *TIP*, 28(1):492–505, 2019.

[9] 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.

[10] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *ICCV*, pages 7314–7323, 2019.

[11] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI*, volume 34, pages 11908–11915, 2020.

[12] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *CVPR*, pages 5769–5780, 2022.

[13] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17683–17693, 2022.

[14] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022.

[15] Kaihao Zhang, Wenhan Luo, Yanjiang Yu, Wenqi Ren, Fang Zhao, Changsheng Li, Lin Ma, Wei Liu, and Hongdong Li. Beyond monocular deraining: Parallel stereo deraining network via semantic prior. *International Journal of Computer Vision*, 130(7):1754–1769, 2022.

Figure 3: **The qualitative comparison on SOTS-Indoor [8].** Our result has the best visual quality and details.

Figure 4: **The qualitative comparison on SOTS-Indoor [8].** Our result has the best visual quality and details.

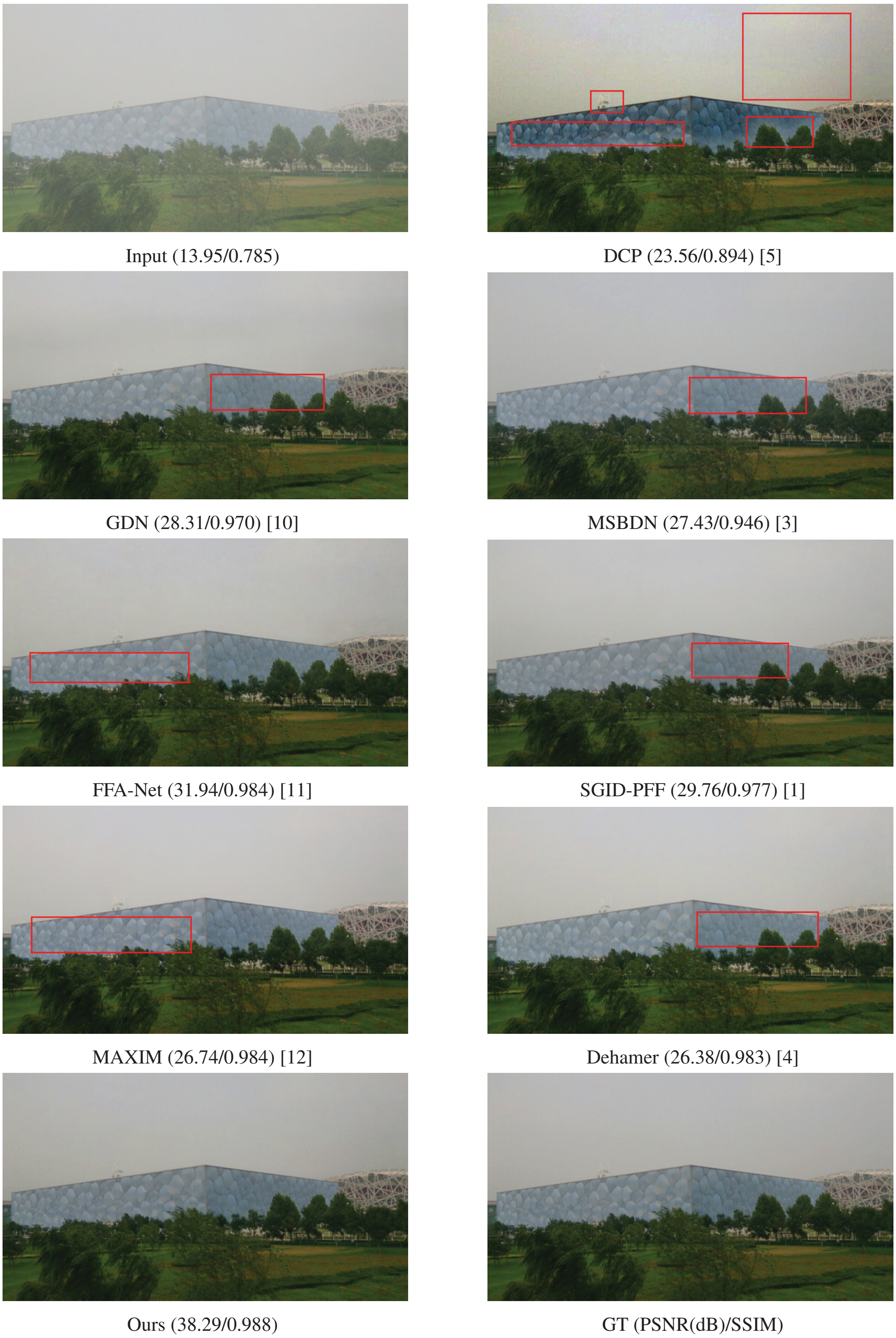
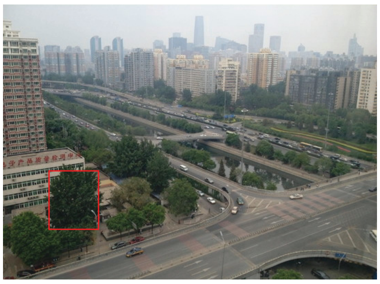


Figure 5: **The qualitative comparison on SOTS-Outdoor [8].** Our result has the best visual quality and details.

Input (22.72/0.961)

DCP (13.71/0.787) [5]

GDN (25.10/0.980) [10]



MSBDN (22.23/0.954) [3]

FFA-Net(30.35/0.990) [11]

SGID-PFF (18.53/0.930) [1]

MAXIM (25.95/0.979) [12]

Dehamer (30.22.38/0.988) [4]

Ours (40.94/0.994)

GT (PSNR(dB)/SSIM)

Figure 6: **The qualitative comparison on SOTS-Outdoor [8].** Our result has the best visual quality and details.

Input (13.89/0.755) [10]

TKL (28.81/0.939) [3]

HDCW-Net (29.60/0.922) [11]

SMGARN (29.81/0.950) [1]

Uformer (30.15/0.949) [12]

Restormer (33.26/0.957) [4]

Ours 38.34/0.980)

GT (PSNR(dB)/SSIM)

Figure 7: **The qualitative comparison on CSD [2].** Our result has the best visual quality and details.

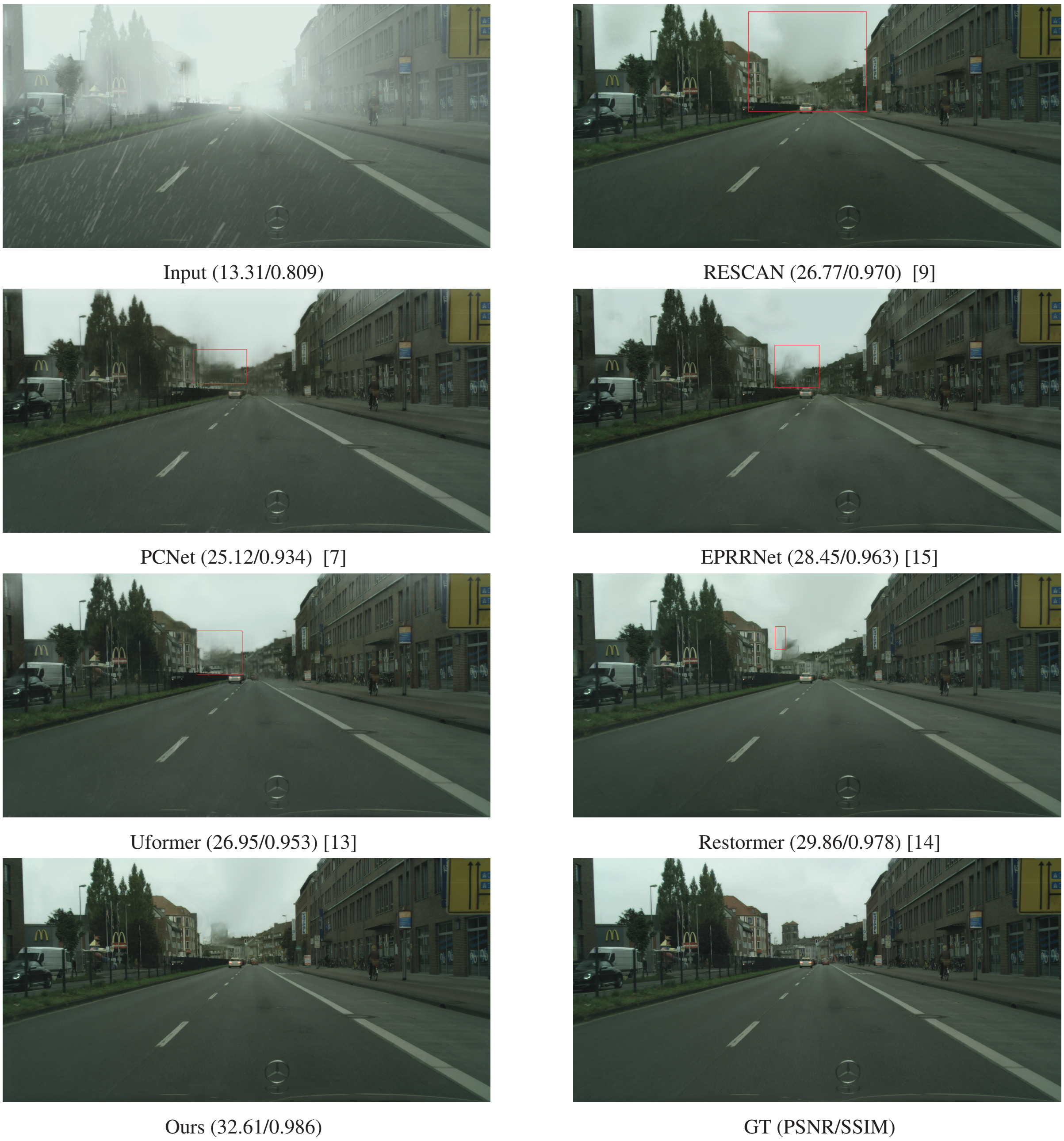


Figure 8: **The qualitative comparison on RainCiryscapes [6].** Our result has the best visual quality and details.