PRE-Mamba: A 4D State Space Model for Ultra-High-Frequent Event Camera Deraining
(Supplementary Material)

Ciyu Ruan[1,*], Ruishan Guo[1,*], Zihang Gong[2], Jingao Xu[3], Wenhan Yang[4], Xinlei Chen[1,†]
[1]Tsinghua University, [2]Harbin Institute of Technology,
[3]Carnegie Mellon University, [4]Pengcheng Laboratory,
{rcy23, grs24}@mails.tsinghua.edu.cn, gongzihang0201@gmail.com,
jingaox@andrew.cmu.edu, yangwh@pcl.ac.cn, chen.xinlei@sz.tsinghua.edu.cn

Abstract

This is the supplementary material for PRE-Mamba: A 4D State Space Model for Ultra-High-Frequent Event Camera Deraining.

We provide the following materials in this manuscript:

- *§1 Formulation of raindrop event generation model.*
- *§2 Implementation details of PRE-Mamba.*
- *§3 More analysis about frequency loss.*
- *§4 More visual examples of EventRain- 27K dataset*
- *§5 More visual comparisons.*
- *§6 Future works.*

1. Formulation

Event cameras, with their high temporal resolution and low latency, are well-suited for dynamic environments [1, 2]. However, rain poses a significant challenge, as raindrops generate dense noise events that overwhelm valid scene information. An event $e \triangleq (x, y, t, p)$ is triggered [3]when the logarithmic intensity change $\Delta L(x, y, t)$ exceeds a threshold C:

$$pC = \Delta L(x, y, t) = L(x, y, t) - L(x, y, t - \Delta t). \tag{1}$$

For small Δt, this is approximated as:

$$\frac{\partial L(x, y, t)}{\partial t} \approx \frac{pC}{\Delta t}. \tag{2}$$

In rainy conditions, the observed intensity $I(x, y, t)$ combines raindrop intensity $I_r(x, y, t)$ and background intensity $I_b(x, y, t)$:

$$I(x, y, t) = \alpha(x, y, t)I_r(x, y, t) + (1 - \alpha(x, y, t))I_b(x, y, t), \tag{3}$$

where $\alpha(x, y, t)$ is the rain coverage fraction, representing the proportion of the pixel area covered by rain. The temporal derivative of $L(x, y, t) = \log(I(x, y, t))$ is:

$$\frac{\partial L(x, y, t)}{\partial t} = \frac{\alpha \frac{\partial I_r}{\partial t} + (1 - \alpha)\frac{\partial I_b}{\partial t} + \frac{\partial \alpha}{\partial t}(I_r - I_b)}{I}. \tag{4}$$

*These authors contributed equally.
†Corresponding author.

Rain motion (v_{r_x}, v_{r_y}) and camera motion (v_{c_x}, v_{c_y}) drive intensity changes:

$$\frac{\partial I_r}{\partial t} = -\nabla I_r \cdot (v_{r_x}, v_{r_y}), \tag{5}$$

$$\frac{\partial I_b}{\partial t} = -\nabla I_b \cdot (v_{c_x}, v_{c_y}). \tag{6}$$

The change in rain coverage is:

$$\frac{\partial \alpha}{\partial t} = -\nabla \alpha \cdot (v_{r_x} + v_{c_x}, v_{r_y} + v_{c_y}). \tag{7}$$

Combining these effects, the event generation model for rain becomes:

$$pC = \frac{1}{I}\Bigg[- \alpha \nabla I_r \cdot (v_{r_x}, v_{r_y}) - (1-\alpha)\nabla I_b \cdot (v_{c_x}, v_{c_y}) - \nabla \alpha \cdot (v_{r_x} + v_{c_x}, v_{r_y} + v_{c_y})(I_r - I_b)\Bigg]\Delta t,$$

The formula can be expressed in matrix form by defining the gradient vectors

$$\nabla I_r = \begin{bmatrix} \frac{\partial I_r}{\partial x} \\ \frac{\partial I_r}{\partial y} \end{bmatrix}, \quad \nabla I_b = \begin{bmatrix} \frac{\partial I_b}{\partial x} \\ \frac{\partial I_b}{\partial y} \end{bmatrix}, \quad \nabla \alpha = \begin{bmatrix} \frac{\partial \alpha}{\partial x} \\ \frac{\partial \alpha}{\partial y} \end{bmatrix}, \tag{8}$$

the velocity vectors

$$\mathbf{v}_r = \begin{bmatrix} v_{r_x} \\ v_{r_y} \end{bmatrix}, \quad \mathbf{v}_c = \begin{bmatrix} v_{c_x} \\ v_{c_y} \end{bmatrix}, \tag{9}$$

the combined matrix

$$\mathbf{M} = \begin{bmatrix} \alpha \nabla I_r & (1-\alpha)\nabla I_b & \nabla \alpha (I_r - I_b) \end{bmatrix}, \tag{10}$$

and the combined velocity vector

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_r \\ \mathbf{v}_c \\ \mathbf{v}_r + \mathbf{v}_c \end{bmatrix}. \tag{11}$$

The event generation model can then be compactly written as

$$pC = \frac{1}{I}\Big[\mathbf{M} \cdot \mathbf{V}\Big]\Delta t. \tag{12}$$

Under rainy conditions, events are generated by the joint effects of raindrop motion $\mathbf{v}_r$, camera motion$\mathbf{v}_c$, and temporal variations in rain coverage $\nabla \alpha$. These factors collectively modulate the intensity changes of the scene I_r and background I_b, ultimately determining the triggering conditions of events.

When there is no rain, the coverage fraction $\alpha = 0$, the rain intensity $I_r = 0$, and the observed intensity $I = I_b$. Substituting these conditions into the model, the formula degenerates to

$$pC = -\nabla I_b \cdot \mathbf{v}_c \cdot \Delta t. \tag{13}$$

This simplified form represents the influence of camera motion $\mathbf{v}_c$ on the background intensity I_b, reflecting the normal operation of the event camera in rain-free conditions.

2. Implementation Details

In the 4D event cloud representation, x, y denote 2D-spatial coordinates (pixel indices, integers), while z, T represent 2D-temporal information (z is the normalized timestamp in $[0, 1]$). We adopt a grid size of $(1, 1, 0.1)$ for (x, y, z), ensuring compatibility between discrete spatial and continuous temporal domains.

The architecture of PRE-Mamba remains consistent with the U-Net [4] framework. It consists of two stage encoders and one decoders, with respective block depths of [2, 4] and [2]. The encoder employs serialized pooling (scale factor = 2) beyond the first stage, reducing spatial resolution while increasing feature channels from 16 to 32. The decoder uses serialized unpooling to restore resolution, aided by skip connections for multi-scale feature fusion. Additionally, an efficient position encoding block is used at the beginning of each block to capture local attention, following the design of point transformer

works [5–7]. Each block integrates a Multi-Scale State Space Model and an MLP, enhanced by residual connections and DropPath. Event clouds are serialized by four scanning modes (z-order, Hilbert curve, trans-z-order, and trans-Hilbert curve). The entire training is conducted on six NVIDIA 416 RTX A6000 GPUs for 50 epochs with a batch size of 6. The learning rate is scheduled using OneCycleLR, with a 5% warm-up phase and cosine annealing for adjustment. The λ in loss function is 1.

3. Frequency Loss

Frequency loss $\mathcal{L}_{\text{fft}}$ complements cross-entropy loss $\mathcal{L}_{\text{ce}}$ by operating at a global scale, suppressing overfitting to local noise through dual constraints on amplitude and phase. The amplitude constraint preserves the energy distribution differences between rain and background, while the phase constraint enforces spatiotemporal continuity. The normalization term enhances robustness by making the loss invariant to absolute energy intensity, ensuring adaptability across diverse rain densities. This relationship ensures that the spectral features of ground truth labels encode meaningful spatiotemporal rain noise distributions, enabling FFT Loss to align predictions with physically consistent global patterns.

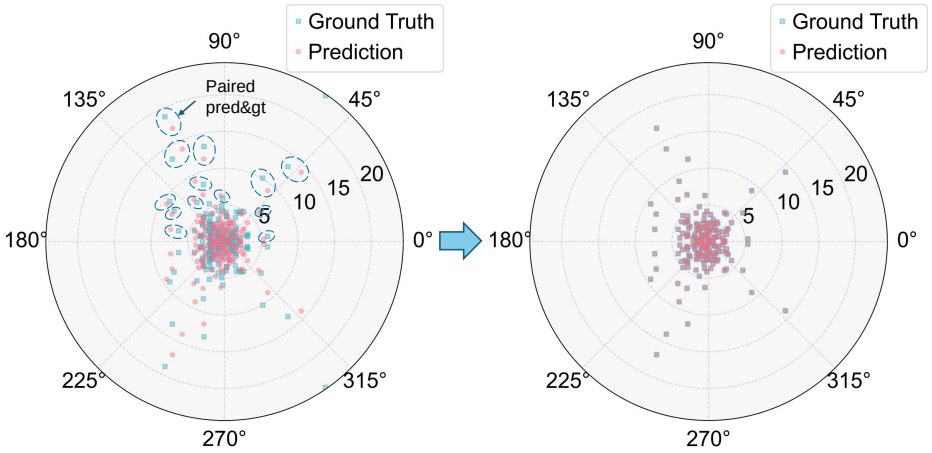

Figure 1. The frequency loss guides model through dual constraints on amplitude and phase.

4. EventRain-27K

In this section, we provide a comprehensive presentation of our dataset.

4.1. Synthetic Dataset

KITTI.

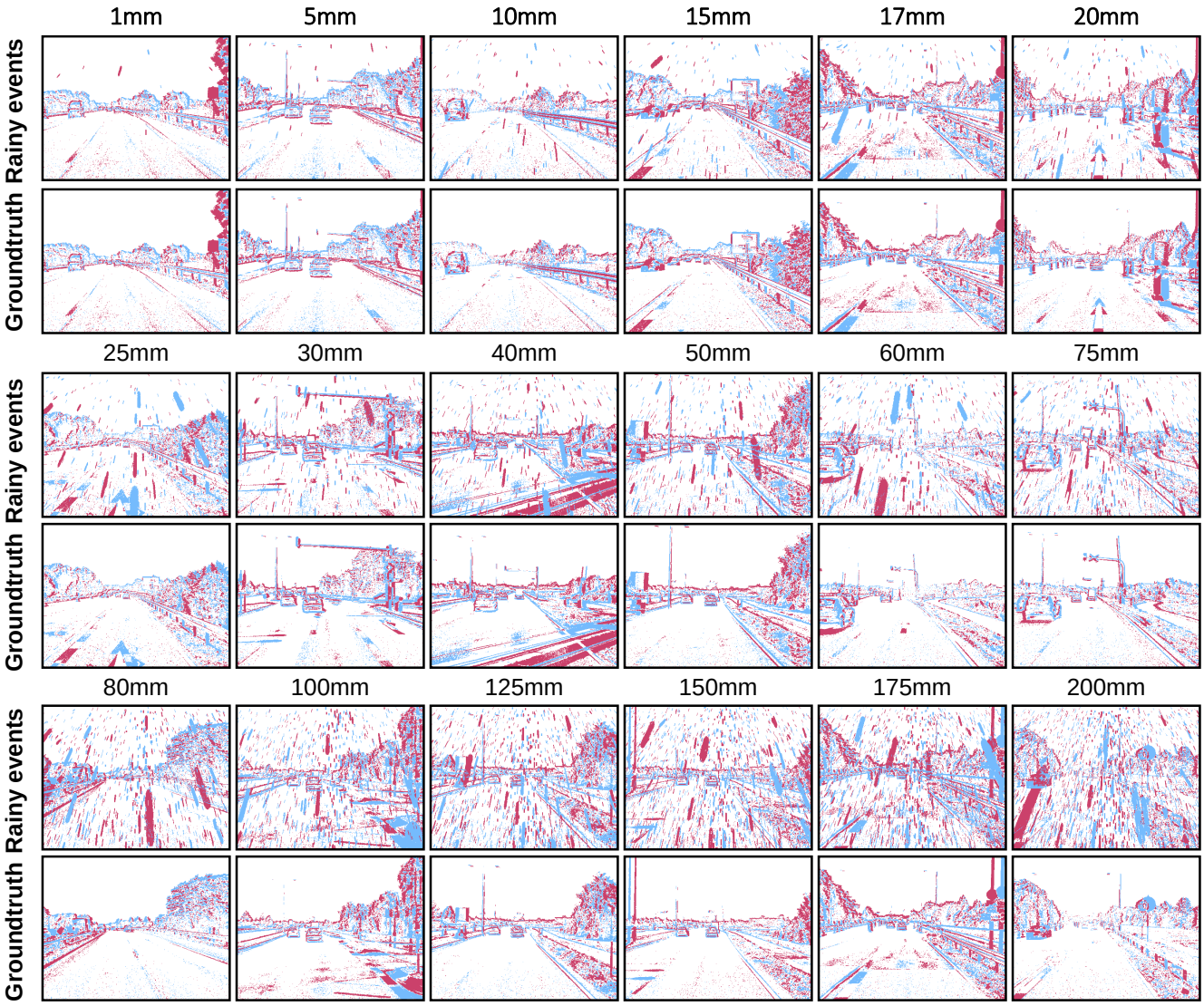

Figure 2. Representative samples of KITTI at different rainfall intensity levels.

The KITTI event-based synthetic rain dataset is a comprehensive benchmark designed for evaluating event-based deraining algorithms under various precipitation conditions. This dataset encompasses 18 distinct rainfall intensity levels, ranging from light rain (5mm/h) to heavy rainstorms (200mm/h), and incorporates varying camera motion speeds to accurately simulate real-world driving conditions. The scenarios are meticulously crafted to include multiple critical elements such as vehicles, roads, traffic signs, street lights, and pedestrians, ensuring realistic and challenging conditions for algorithm evaluation.

The dataset contains 7,002 rain event sequences, each paired with synchronized ground truth, providing a robust foundation for comprehensive performance evaluation. The event data is captured at a resolution of 352×460, with each rainfall intensity level comprising 389 event sequences. Representative samples of KITTI can be found at Figure 2.

SPAC.

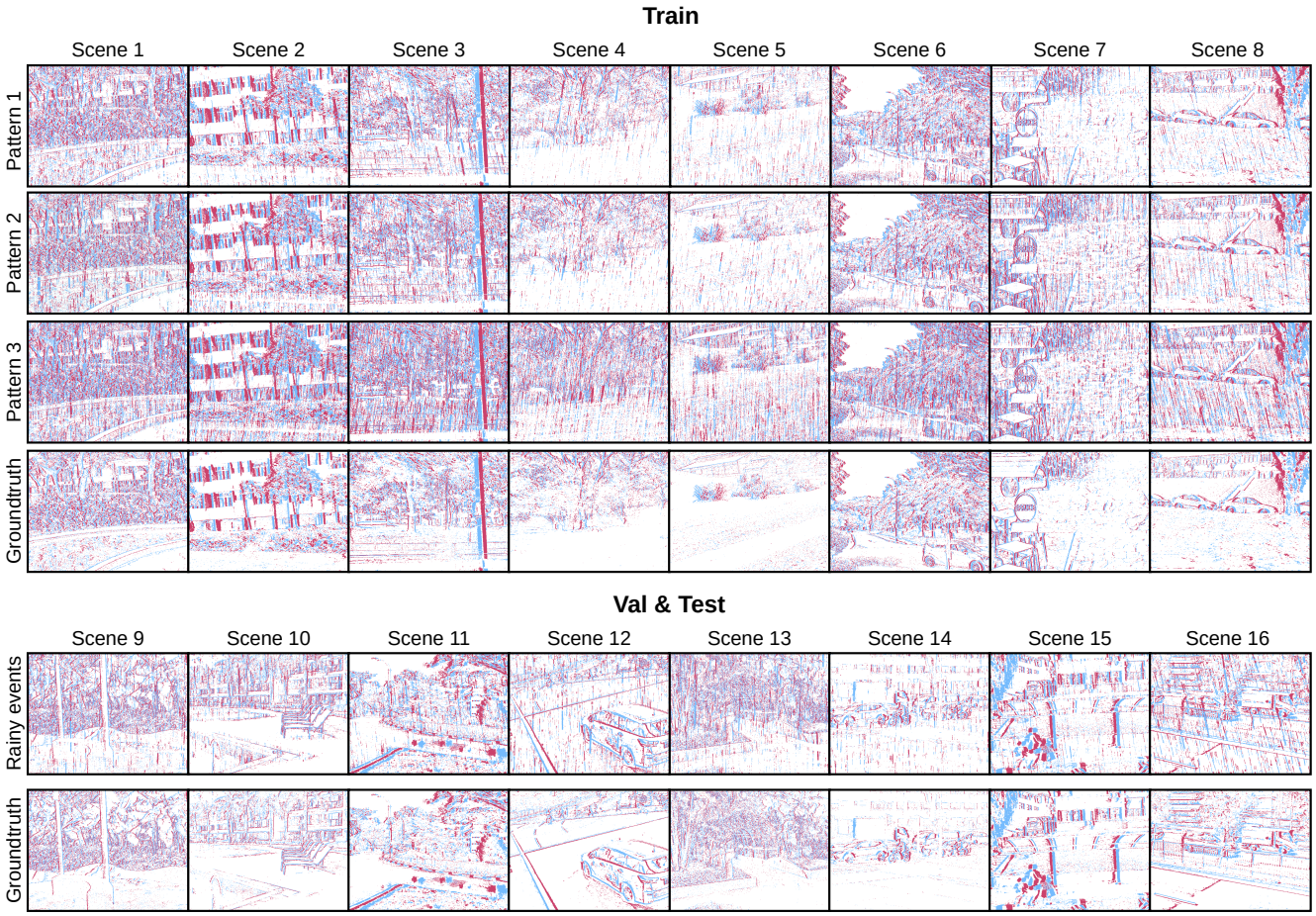

Figure 3. Representative sampling of SPAC across diverse scenarios and rainfall patterns.

The SPAC dataset comprises 16 distinct scenarios, each augmented with 3 to 4 distinct rain patterns. These scenarios includes cityscapes and natural environments, featuring diverse elements such as buildings, trees, grass, vehicles, roads, pedestrians, traffic signs and ponds.

The dataset contains a total of 4,635 event sequences, with the training set consisting of 8 scenarios and 2,961 sequences, and the validation and test set comprising 8 scenarios and 1,674 sequences. Each training sequence includes event data for 3 to 4 precipitation patterns paired with synchronized ground truth annotations. The event data is captured at a resolution of 480×640, providing a comprehensive resource for evaluating event-based deraining algorithms under varying rainfall conditions. Representative samples of SPAC can be found at Figure 3.

4.2. Artificial Rain Dataset

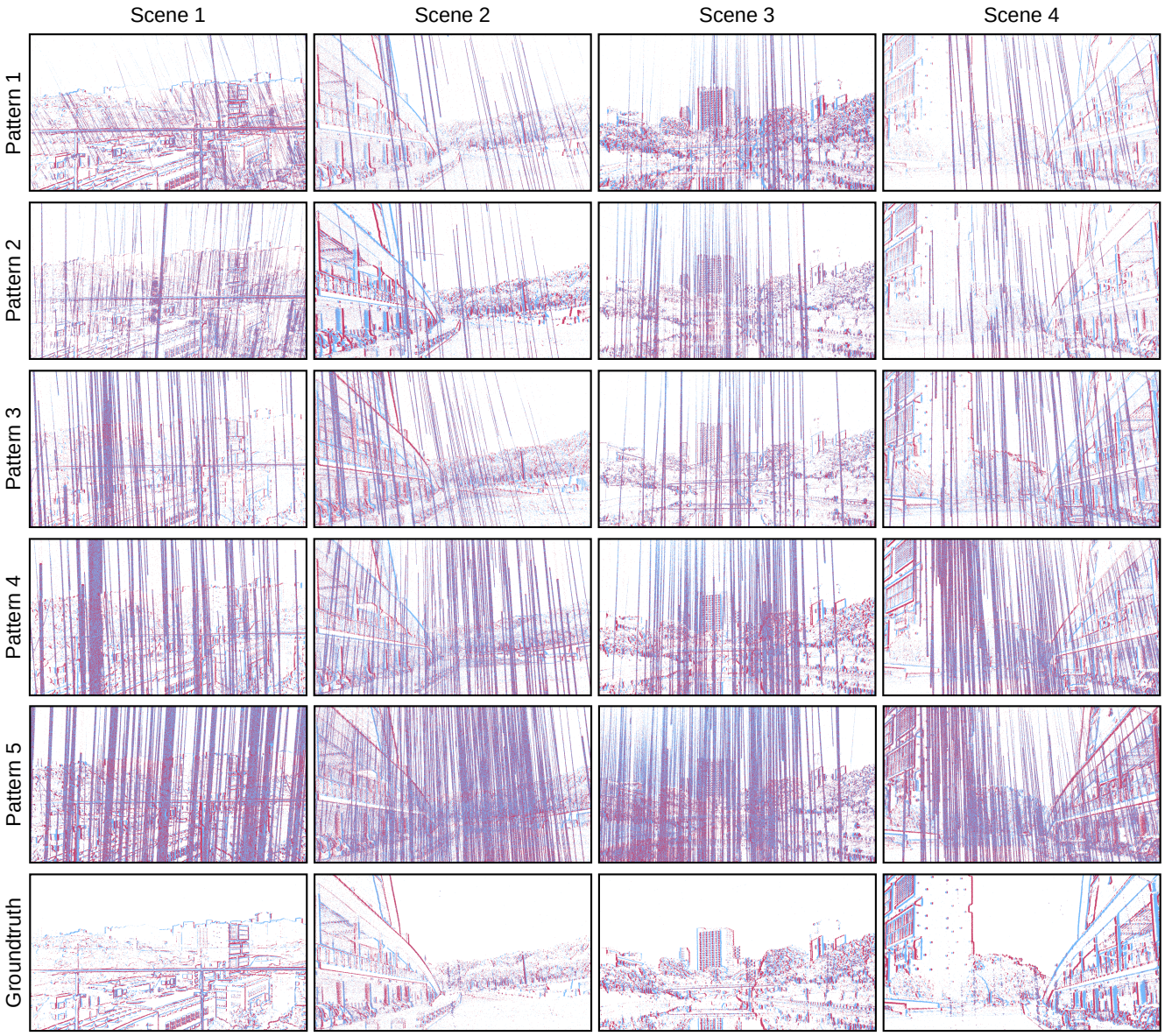

Figure 4. Representative sampling for self-collected artificial rainfall datasets across diverse scenarios and rainfall patterns.

The self-recorded artifical rainy dataset encompasses four distinct urban and natural scenarios, augmented with a complete spectrum of rainfall intensities, ranging from light drizzle to intense downpours. These scenarios includes buildings, trees, vehicles, and lakes. The dataset is captured by Prophesee EVK4 event cameras under controlled rainfall conditions, preserving the same acquisition characteristics as real-world scenarios while significantly reducing the domain gap compared to synthetic data.

The dataset contains a total of 6,438 event sequences, each containing diverse precipitation patterns with precisely aligned ground truth annotations. All sequences are captured at a resolution of 720×1280 under controlled rainfall conditions. Representative samples of the self-recorded artifical rainy dataset can be found at Figure 4.

4.3. Real-world Rain Dataset

Figure 5. Representative samples of real-world rainfall and snowfall scenarios.

The real-world dataset establishes the first large-scale point-based benchmark for event-based deraining research, comprising 9,471 sequences across 26 diverse scenes with a resolution of 720×1280. Notably, the dataset includes 4,000 samples of authentic snowfall data. Representative samples of the real-world dataset can be found at Figure 5.

5. More Visual Comparisons

5.1. Real-world Performance

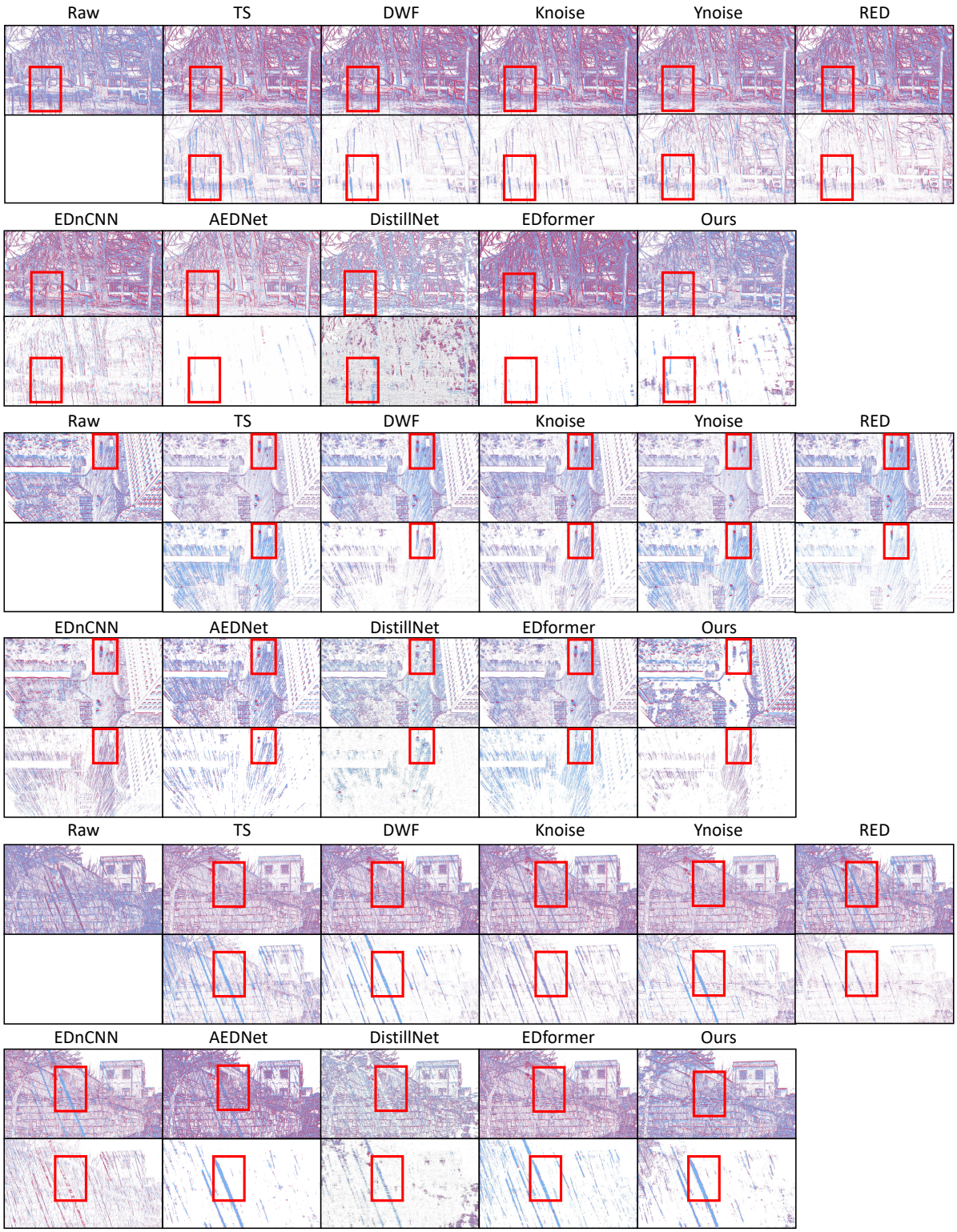

Figure 6. **Extended Qualitative Comparison on Real-World Datasets.** The first row illustrates the background after deraining, while the second row showcases the extracted rain layer. Our model exhibits superior deraining performance across diverse scenarios.

Figure 7. **Extended Qualitative Comparison on Real-World Datasets.** Rows 1-4: deraining results; Rows 5-10: desnowing results. Our model achieves state-of-the-art performance in both precipitation removal tasks across diverse scenarios.

5.2. Synthetic Data Performance

KITTI.

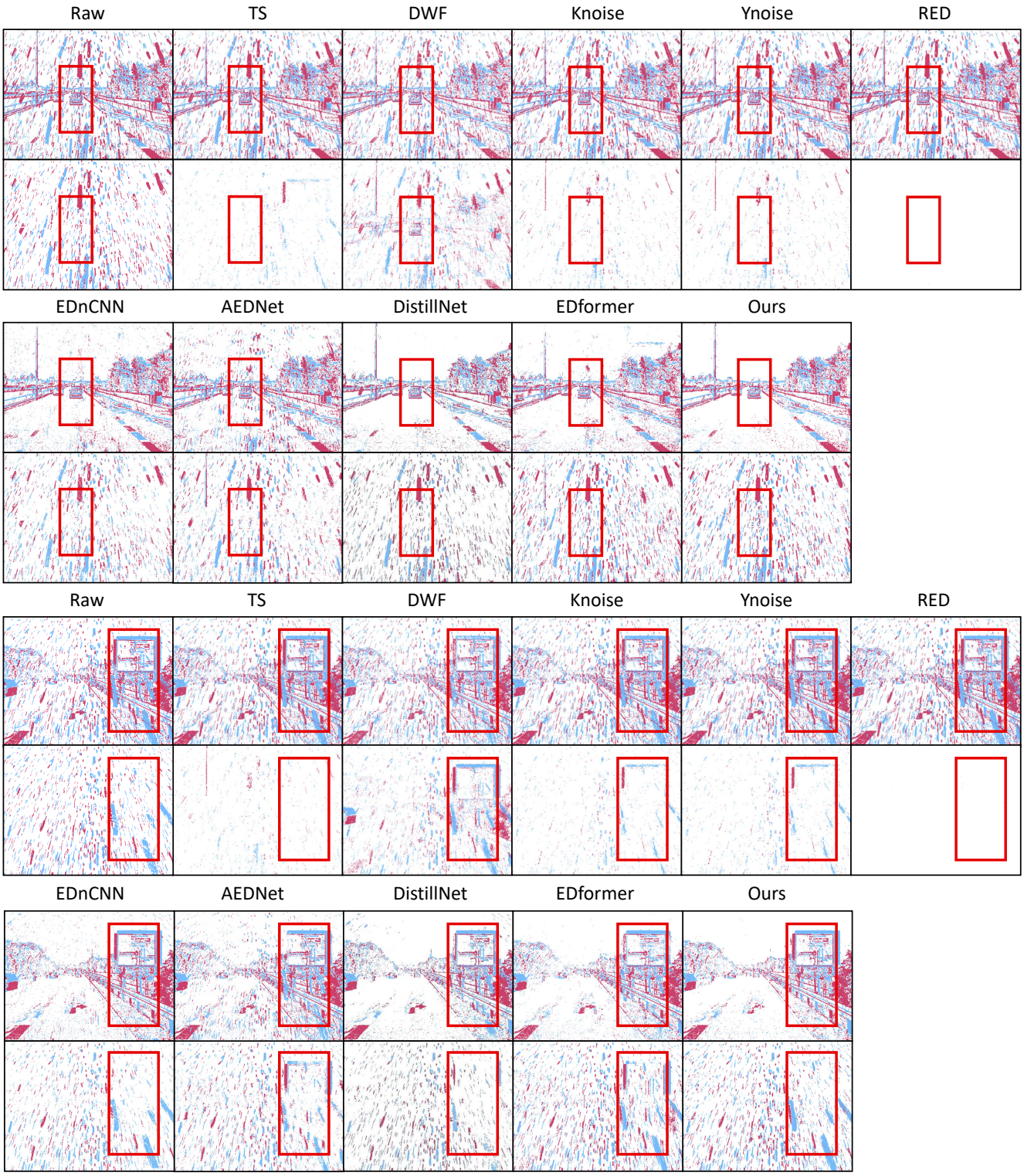

Figure 8. **Extended Qualitative Comparison on KITTI Datasets.** Please zoom in for details.

SPAC.

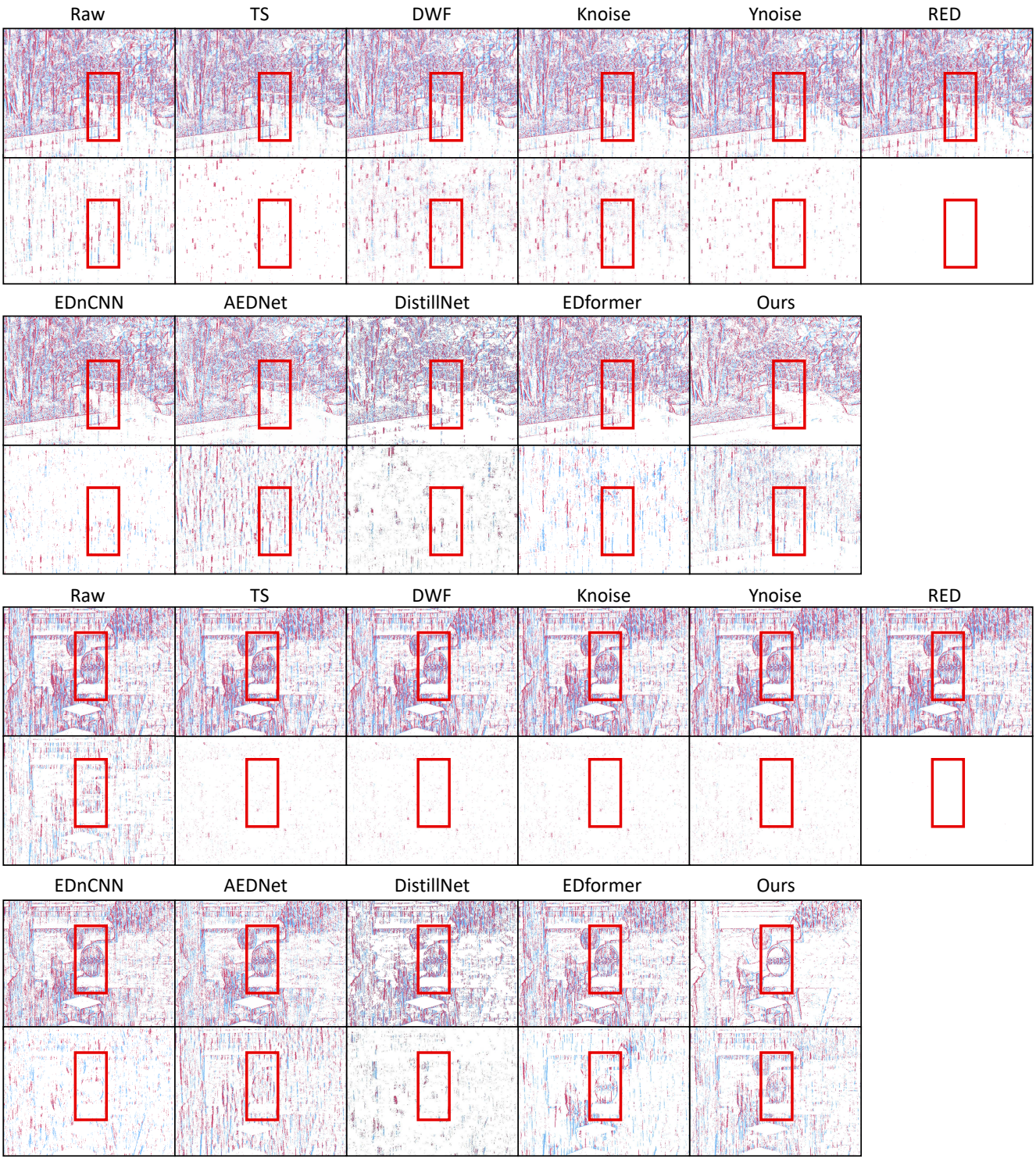

Figure 9. **Extended Qualitative Comparison on SPAC Datasets.** Please zoom in for details.

Artifical.

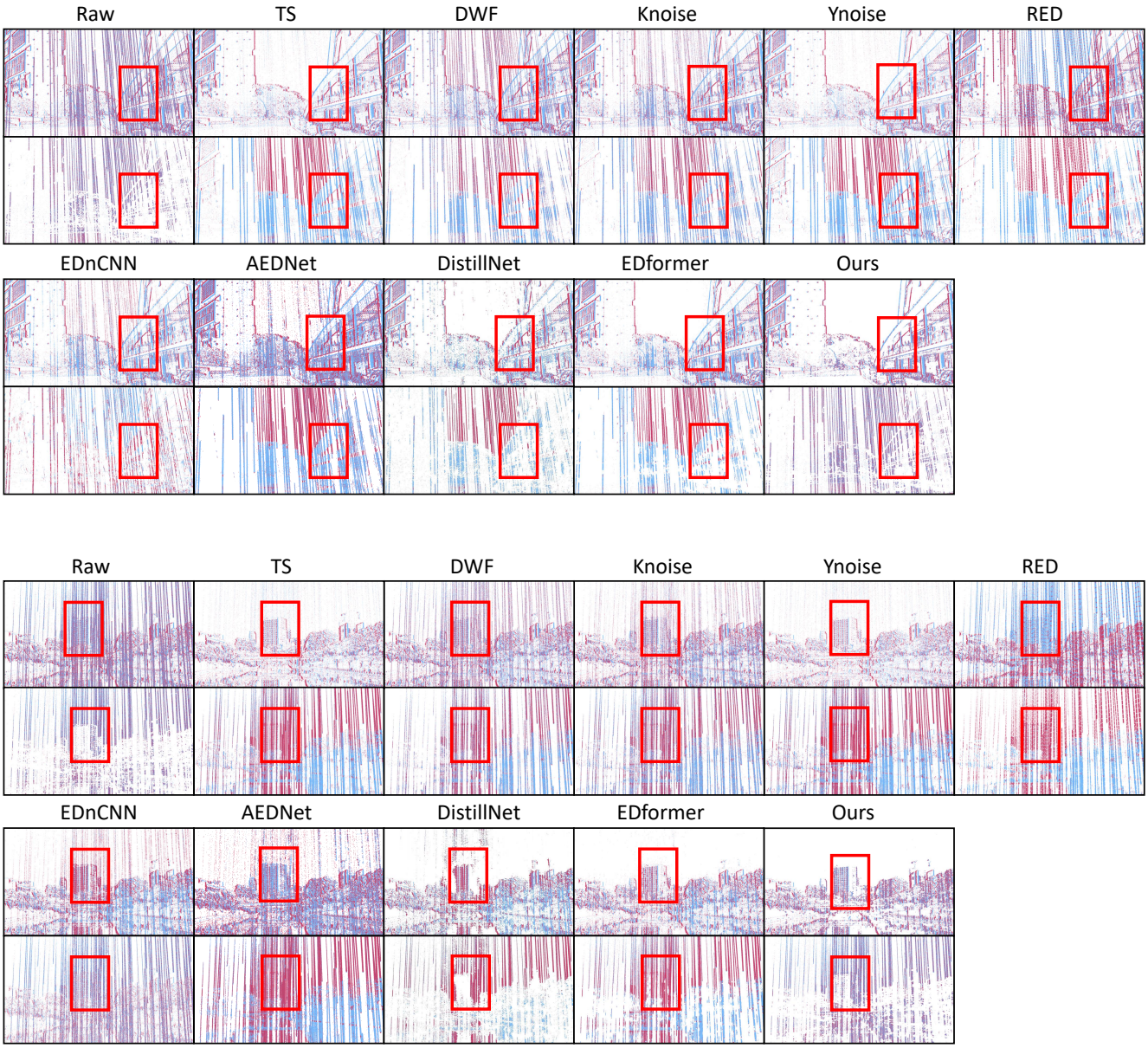

Figure 10. **Extended Qualitative Comparison on self-recorded artifical rainfall Datasets.** Please zoom in for details.

6. Future Work

In future work, we plan to further investigate efficient spatio-temporal feature representations for event clouds, leveraging their inherent sparsity and asynchronous nature to achieve effective and efficient deraining. Additionally, we aim to explore semi-supervised or unsupervised learning frameworks to reduce the reliance on labeled data and improve generalization capabilities. Furthermore, we will focus on model lightweighting techniques to optimize computational and memory efficiency, enabling seamless deployment on resource-constrained robotic platforms. These efforts will advance the practical application of event-based vision systems in dynamic and adverse environments, enhancing their robustness, scalability, and real-world usability.

References

[1] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1

[2] Tiziana D'Orazio and Cataldo Guaragnella. A survey of automatic event detection in multi-camera third generation surveillance systems. *International Journal of Pattern Recognition and Artificial Intelligence*, 29(01):1555001, 2015. 1

[3] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 884–892, 2016. 1

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022. 2

[5] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 3

[6] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022.

[7] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4840–4851, 2024. 3