

High-Quality and Efficient Turbulence Mitigation with Events

Supplementary Material

Summary Organization of the supplementary material.

- Sec. 1 provides characteristics analysis under turbulence, covering more statistics under high-speed observation, in-depth study of event-polarity behaviors, and detailed theoretical derivations of event tubes.
- Sec. 2 details the network architectures and the training scheme of the proposed method.
- Sec. 3 introduces the setup of hardware and features of our constructed datasets.
- Sec. 4 discusses more ablation studies and presents extended results on LATH, CTH and UDET datasets.

1. Characteristics Analysis under Turbulence

1.1. More Statistics under High-Speed Observation

We conduct statistical experiments to further investigate the advantages of high-speed observations for turbulence mitigation. A representative visual example is shown in Fig. 1 (a). We quantitatively compare temporal average results between high-speed and low-speed observations. The high-speed observation required only a substantially shorter duration (a 75% reduction in cumulative time) while achieving a lower error (a 199.33 reduction in MSE). Furthermore, we analyze the motion distribution of corners, as illustrated in Fig. 1 (b). Compared with low-speed observations, corners motion under high-speed conditions follows an approximately zero-mean distribution within a significantly shorter temporal window, providing fine-grained motion cues that reveal the original position of corners.

1.2. In-Depth Study of Event Polarity Alternation

To gain deeper insight into how event polarity alternation relates to spatial gradients under turbulence, we conduct an additional in-depth analysis. As shown in Fig. 2 (a), the spatiotemporal visualization of event streams from a static scene reveals a clear polarity-alternation pattern along the time axis. Fig. 2 (b) tracks the temporal trajectory of a single event pixel and shows that its GT position closely matches the maximum PAEP response, reflecting a consistent correspondence between structurally stable regions and PAEP statistics. In Fig. 2 (c), the PAEP map provides much sharper structural cues than the event density map, offering clear benefits for restoring turbulence-degraded edges. Finally, Fig. 2 (d) shows that PAEP counts remain positively correlated with spatial gradients across turbulence levels, indicating that stronger turbulence increases brightness-fluctuation frequency while preserving the intrinsic polarity-alternation pattern.

1.3. Justification of the Formation of Event Tubes

Let the video signal captured by the camera be the function $I : \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}$, mapping from continuous spatio-temporal coordinates to intensity. The event generation process is described by:

$$\log I(x, y, t) - \log I(x, y, t - \Delta t) = pC, \quad (1)$$

where $\log I(x, y, t)$ is the logarithmic brightness at pixel (x, y) and time t , Δt is the event time interval, $p \in \{-1, 1\}$ is the event polarity, and C is the contrast threshold.

We analyze $\frac{\partial \log I}{\partial t}$ as a continuous approximation of the event response. To capture the ensemble motion of scenes, we utilize the concept of *particle* from the Lagrangian particle tracking [2] method. In this way, a particle can be described by the 3-dimensional spatial-temporal coordinate (\mathbf{x}, t) , where $\mathbf{x} \in \mathbb{R}^2$ is the spatial coordinates.

Suppose that the trajectory of a particle (\mathbf{x}_0, t_0) in some time interval $\Delta > 0$ can be defined as a function $T_{\mathbf{x}_0}^{t_0} : [t_0 - \Delta, t_0 + \Delta] \rightarrow \mathbb{R}^2$ mapping from temporal coordinates to the corresponding spatial coordinates (e.g., $T_{\mathbf{x}_0}^{t_0}(t_0) = \mathbf{x}_0$), allowing the trajectory to be represented as the point set $\{(T_{\mathbf{x}_0}^{t_0}(t), t) : t \in [t_0 - \Delta, t_0 + \Delta]\}$. When Δ is small enough (e.g., 20 ms), the trajectory function can be approximated linearly according to Taylor's expansion at t_0 :

$$T_{\mathbf{x}_0}^{t_0}(t) \approx \mathbf{x}_0 + (t - t_0) \cdot T_{\mathbf{x}_0}^{t_0'}(t_0). \quad (2)$$

Therefore, the trajectory of all particles in the uniform grid at time t_0 :

$$T^{t_0} = \begin{bmatrix} T_{1,1}^{t_0} & \dots & T_{1,w}^{t_0} \\ \vdots & \ddots & \vdots \\ T_{h,1}^{t_0} & \dots & T_{h,w}^{t_0} \end{bmatrix}, \quad (3)$$

forms a set of event tube as mentioned in the main body of the paper.

Given these tube-shaped trajectories, we next analyze how events are generated along them. We assume that the intensity of the particle remains constant over time:

$$\forall t \in [t_0 - \Delta, t_0 + \Delta] : I(T_{\mathbf{x}_0}^{t_0}(t), t) = I(\mathbf{x}_0, t_0). \quad (4)$$

For any $t_1, t_2 \in [t_0 - \Delta, t_0 + \Delta]$, denote $\mathbf{x}_1 = T_{\mathbf{x}_0}^{t_0}(t_1)$, $\mathbf{x}_2 = T_{\mathbf{x}_0}^{t_0}(t_2)$. Using Eq. (2), the first-order approximation of $I(\mathbf{x}_1, t_1)$ yields:

$$\begin{aligned} I(\mathbf{x}_1, t_1) &\approx I(\mathbf{x}_2, t_2) \\ &+ (t_1 - t_2) \left\langle T_{\mathbf{x}_0}^{t_0'}(t_0), \frac{\partial I}{\partial \mathbf{x}}(\mathbf{x}_2, t_2) \right\rangle \\ &+ (t_1 - t_2) \cdot \frac{\partial I}{\partial t}(\mathbf{x}_2, t_2). \end{aligned} \quad (5)$$

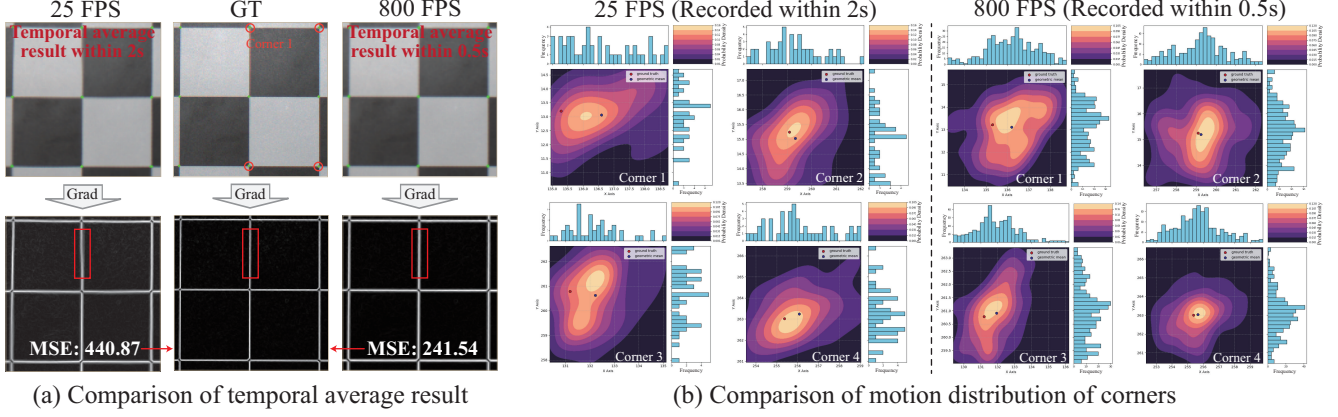


Figure 1. More statistics under high-speed observation. (a) High-speed observation requires a much shorter cumulative time to achieve lower errors compared with low-speed observation. (b) Compared with low-speed observation, corners motion under high-speed conditions conforms to an approximately zero-mean distribution over significantly shorter time windows.

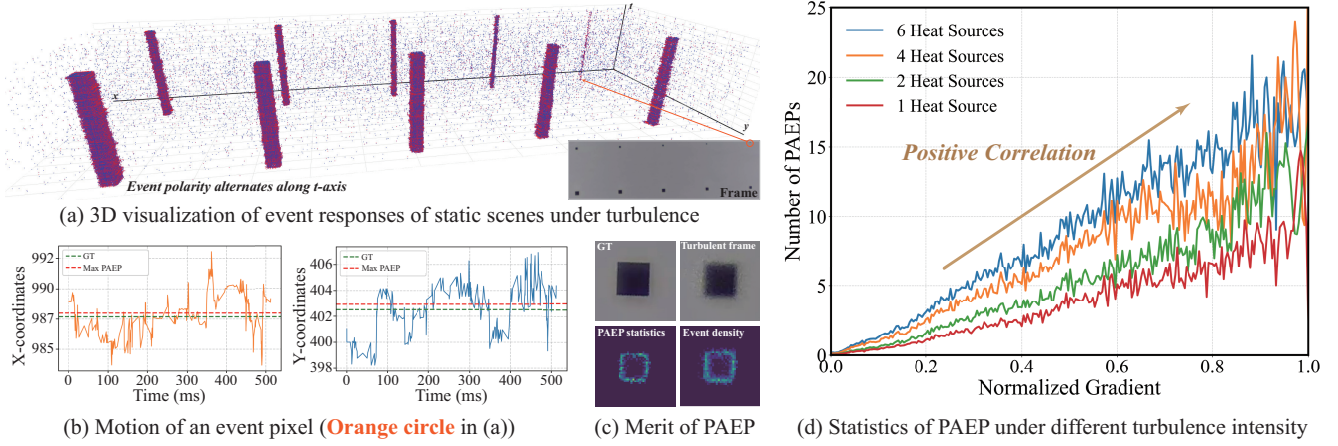


Figure 2. In-depth study of event polarity alternation. (a) Event responses in static scenes exhibit distinct polarity alternation along t-axis under turbulence. (b) The maximum PAEP value is closely approximated by the temporal average of a single-pixel event trajectory. (c) PAEP maps yield sharp structural cues for edge restoration. (d) The number of PAEPs maintain a strong positive correlation with spatial gradients across different turbulence intensities.

Under the consistent intensity assumption, replacing t_2 by t gives:

$$\frac{\partial \log I}{\partial t}(\mathbf{x}, t) \approx -\frac{1}{I(\mathbf{x}_0, t_0)} \left\langle T_{\mathbf{x}_0}^{t_0'}(t_0), \frac{\partial I}{\partial \mathbf{x}}(\mathbf{x}, t) \right\rangle, \quad (6)$$

where $\mathbf{x} = T_{\mathbf{x}_0}^{t_0}(t)$. If the background intensity around the particle is approximately flat, then

$$\frac{\partial \log I}{\partial t}(\mathbf{x}, t) \approx -\frac{1}{I(\mathbf{x}_0, t_0)} \left\langle T_{\mathbf{x}_0}^{t_0'}(t_0), \frac{\partial I}{\partial \mathbf{x}}(\mathbf{x}_0, t_0) \right\rangle, \quad (7)$$

which is a constant and independent of t .

In this case, the event response becomes temporally uniform along the linear particle trajectory, giving rise to the uniform event response along tube-shaped trajectories, i.e., the formation of *event tubes*.

2. Network Details and Training Scheme

2.1. Rigid Motion-Aware Block

As illustrated in Fig. 3, to provide a more intuitive understanding of the internal structure of the RMAB, we present the detailed designs of the 3D-ResConv and 3D-CAU. The diagrams specify layer configurations, kernel sizes, and activation functions. RMAB can effectively extract multi-scale spatiotemporal features in a lightweight manner by the integration of three 3D-ResConv and 3D-CAUs.

2.2. Stable Edge-Guided Bi-Mamba

Mamba shows high efficiency in modeling long-range dependencies via linear computational complexity. While effective for sequential data, its original 1D causal sequence design is less naturally aligned with 2D or 3D visual inputs. Inspired by [9], we adopt the bidirectional scanning strat-

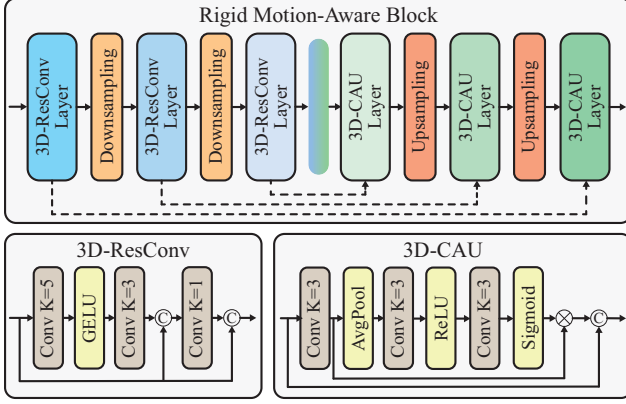


Figure 3. Architecture of the proposed RMAB.

egy to better accommodate the spatial-temporal characteristics of video data, preserving Mamba’s modeling strengths while effectively enhancing visual understanding.

Hilbert Curve Scanning. The Hilbert curve is a space-filling curve that preserves proximity among elements when mapping multi-dimensional data to one dimension. As presented in Fig. 4 (b), the Hilbert curve connects all elements in the spatial domain while maintaining locality, which enhances feature clustering [5]. Moreover, [4] shows that the Hilbert curve achieves superior clustering in multi-dimensional space. Thus, the Hilbert curve scanning can strengthen correlations between spatial-temporal neighboring features by promoting clustering of adjacent tokens.

Stable Edge-guided Bi-Mamba Block. To exploit stable motion cues for TM, we propose the Stable Edge-guided Bi-Mamba (SE-Mamba) Block, as shown in Fig. 4 (a). It integrates stable edge guidance into the Bi-Mamba architecture to enhance the learning and perception of sharp structural information in turbulent scenes. At its core, the SE-Mamba layer distinguishes itself from the conventional Bi-Mamba by incorporating stable edge information as guidance, as illustrated in Fig. 4 (c). This mechanism enables the model to effectively aggregate and leverage these stable cues, directing the restoration process toward preserving both structural consistency and object boundaries with few frames. Through this design, the SE-Mamba block achieves a coherent fusion of stable structural priors and visual features, resulting in superior restoration quality in dynamic scenes. Moreover, with linear-complexity design, the block remains lightweight and enables efficient TM.

2.3. Training Scheme of the Proposed Method

The training procedure is organized into three stages, progressively enhancing the quality of stable edge guidance while ensuring effective video restoration. Each stage follows a three-step learning strategy to promote feature learning, following the principle that gradually increasing the spatial resolution allows the model to learn from coarse to

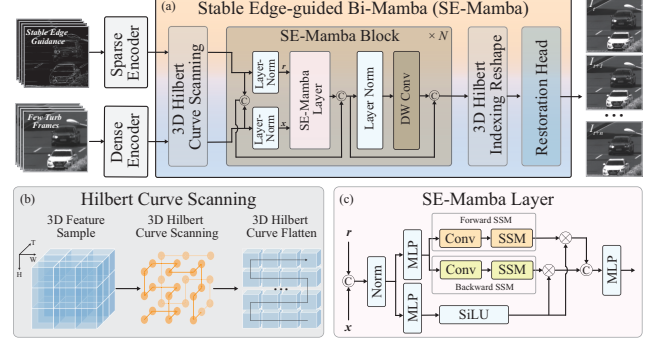


Figure 4. Details of the proposed Stable Edge-guided BiMamba for TM. (a) The overall architecture of the SE-Mamba. (b) Hilbert curve scanning order. (c) Details of the SE-Mamba layer.

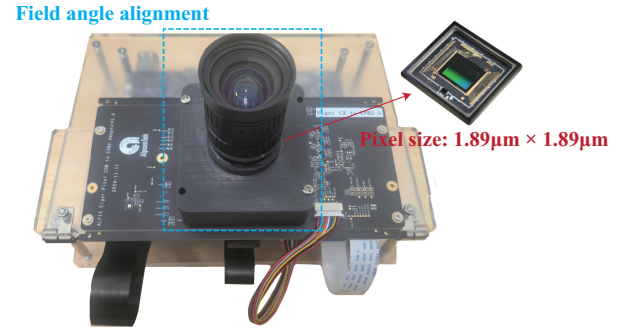


Figure 5. ALPIX-Pizol camera.

fine representations, which has been shown effective in image restoration tasks [6]. Specifically, during the first 100 epochs, a patch size of 128×128 , batch size of 8, and 20 input frames are used; in the next 100 epochs, the patch size is increased to 256×256 with a batch size of 4 and 10 input frames; and in the final 100 epochs, the full-resolution (512×512) images are used with a batch size of 2 and 5 input frames.

Stage 1: ET-Stable Training. ET-Stable module is trained on our CTTH dataset. Benefiting from the availability of GTs of dynamic objects, ET-Stable can obtain stable motion fields to constrain turbulence-degraded event tubes. It focuses on extracting temporally dense features from event voxels and generating stable edge guidance for dynamic objects, enabling accurate motion reconstruction and suppression of turbulence-induced distortions, while providing a solid foundation for subsequent scene stabilization.

Stage 2: ET-Stable and EPAW-Stable Jointly Training. ET-Stable and EPAW-Stable modules are jointly trained on the CTTH dataset. In the first 100 epochs, the ET-Stable are frozen while only the EPAW-Stable module is trained. In the following 200 epochs, both modules are jointly optimized to achieve coordinated learning. EPAW-Stable extends stabilization to scene regions by leveraging event polarity alternation statistics and temporally averaged gradi-

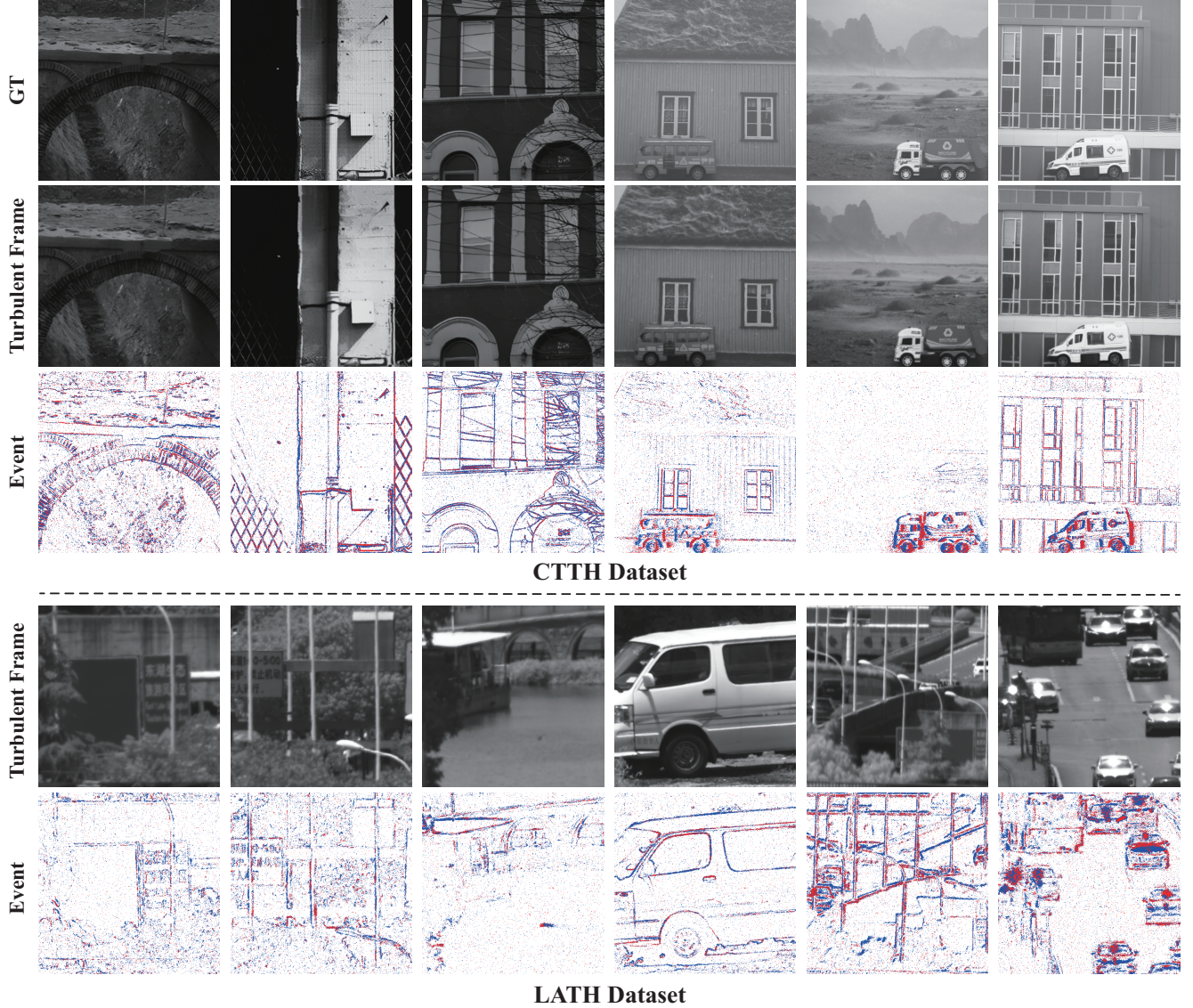


Figure 6. Examples in our CTTH and LATH datasets.

Table 1. Ablation Studies of Multi-Stage Training

Training Scheme	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
w/o Stage 1 + Stage 2	29.85	0.8759	0.2353
w/o Stage 1	30.47	0.9014	0.1972
w/o Stage 2	32.58	0.9207	0.1789
Multi-Stage Training	35.17	0.9425	0.1281

ents. During joint training, ET-Stable module preserves dynamic object consistency, while EPAW-Stable module refines scene edges, ensuring that the combined guidance effectively captures both object and scene structures.

Stage 3: Restoration Network Fine-tuning. SE-Mamba video restoration network is first pretrained on the ATSyn-dynamic [7] dataset, where spatial gradients of degraded

images are employed as guidance to facilitate the learning of turbulence-aware representations. Subsequently, the SE-Mamba video restoration network is fine-tuned jointly with the ET-Stable and EPAW-Stable modules using few-frame inputs on our CTTH dataset. This fine-tuning stage integrates high-quality motion and edge guidance, enabling accurate restoration of fine object structures and scene textures, efficiently mitigating turbulence distortions.

3. Hardware Setup and Datasets

3.1. ALPIX-Pizol Camera

The ALPIX-Pizol Camera [1] leverages a dual-focal-plane architecture that provides alignment of field angle between event and frame modalities at the sensor level, which helps

Table 2. Effect of Losses on Multi-Stage Training

$\mathcal{L}_{\text{motion}}$	$\mathcal{L}_{\text{guide}}$	\mathcal{L}_{pre}	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
		✓	31.74	0.9076	0.1884
	✓	✓	33.58	0.9282	0.1581
✓		✓	33.93	0.9279	0.1612
✓	✓	✓	35.17	0.9425	0.1281

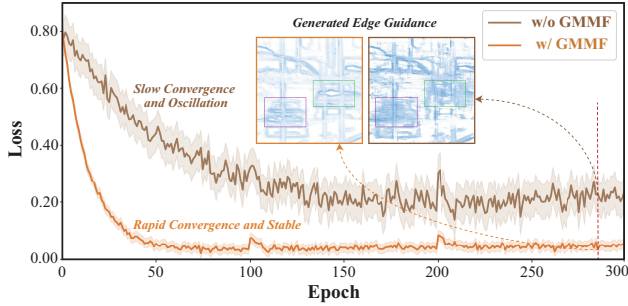


Figure 7. Gradient-masked motion fields enhance the stability of guidance while enabling faster and more reliable convergence.

maintain consistency when analyzing turbulence data. As shown in Fig. 5, its $1.89 \mu\text{m} \times 1.89 \mu\text{m}$ pixel pitch offers fine spatial resolution, making it suitable for observing fine-grained motion cues that appear in turbulence [3].

3.2. Datasets Features

The CTHH dataset contains thermal-turbulence sequences with dynamic moving objects and paired GTs, derived from the TMT dataset [8]. It includes both static and dynamic object configurations, and consists of about 30,000 event-frame pairs, with 90% used for training and 10% for evaluation. The LATH dataset provides long-range atmospheric turbulence event-frame data captured across diverse shooting distances and scenes, enabling comprehensive evaluation of model generalization under varying conditions. Representative samples are presented in Fig. 6.

4. Additional Ablation Studies and Results

4.1. Why Employ Multi-Stage Training?

Our multi-stage training progressively stabilizes event representations and enhances restoration quality. As presented in Table 1, removing both Stage 1 and Stage 2 yields the worst performance, as the restoration model becomes difficult to train and is prone to converging to poor local optima. Removing Stage 1 alone causes a significant performance drop, due to the loss of crucial object-motion priors needed to suppress turbulence. In contrast, removing Stage 2 weakens scene stabilization but has a less severe impact, as the model can still rely on the temporally consistent information from few frames. The multi-stage training pipeline provides coherent guidance, resulting in the most stable and

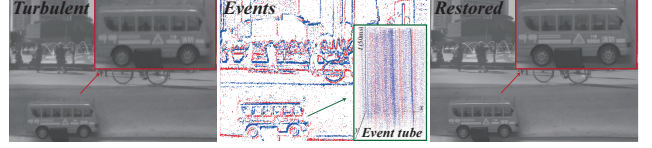
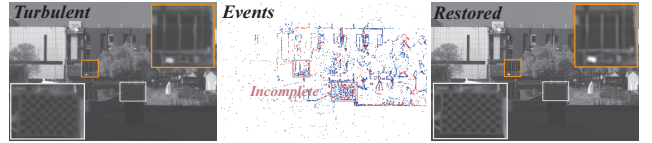


Figure 8. Robustness to camera shake with moving object.



(a) Complex lighting with low event camera threshold



(b) Complex lighting with high event camera threshold

Figure 9. Robustness to complex lighting with varying thresholds.

accurate restoration. In addition, Table 2 further provides the complementary roles of the loss functions used in our multi-stage training: the three losses work jointly to constrain both dynamic objects and scenes, providing balanced supervision that effectively enhances restoration quality.

4.2. Role of Gradient-Masked Motion Fields

As presented in Fig. 7, with gradient-masked motion fields (GMMF), the model focuses its attention on true boundaries of dynamic objects during training, allowing it to concentrate on structurally reliable motion cues. Consequently, the resulting edge guidance becomes sharper and clearer, particularly around dynamic objects. Furthermore, w/ GMMF leads to faster and more stable convergence, thereby enhancing the consistency of learned motion fields.

4.3. Robustness under Complex Conditions.

We conduct additional experiments under thermal turbulence to evaluate the robustness of our method.

Camera shake. As shown in Fig. 8, typical camera shake (e.g., wind-induced vibrations) resembles turbulence motion and preserves the coherence of event tube, under which our method remains effective for object restoration.

Complex lighting. The high dynamic range of event cameras enables reliable edge sensing under complex lighting conditions; hence, Fig. 9 shows that our method successfully restores the distorted railings and calibration board.

Sensitivity to event thresholds. As presented in Fig. 9, the high threshold produces incomplete event responses yet our method remains effective, while the low threshold provides richer edge cues and yields superior restoration quality.

Table 3. Computational Performance with Different Patch Sizes

Patch Size	FPS	Memory (MB)	Flops (G)	Data Size (%)
1024 * 1024	9.4	6959.1	710.9	100
512 * 512	29.5	4779.6	181.7	25
256 * 256	37.2	3983.5	44.5	6.25
128 * 128	43.3	3200.5	11.2	1.56

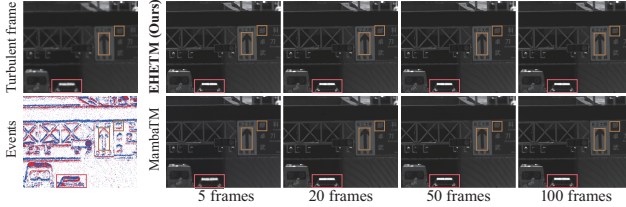


Figure 10. Visualization of input frame count on restoration.

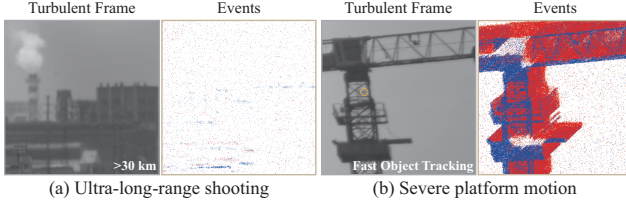


Figure 11. Limitation of our method. Ultra-long-range imaging limits fine texture capture by events. Severe platform motion entangles event information, complicating motion decoupling.

4.4. Impact of Input Frame Count on Restoration.

We provide visualizations comparing our method with MambaTM. Fig. 10 demonstrates that our method improves as frame count increases, and it achieves superior restoration with few frames compared to MambaTM.

4.5. Influence of Different Patch Sizes.

Table 3 reports the computational performance of our model under different input patch sizes. All results are obtained during inference on an NVIDIA RTX 3090 GPU.

4.6. Limitation and Future Works.

As shown in Fig. 11 (a), imaging at ultra-long-range conditions limits the ability of the event camera to capture fine texture details, thereby hindering the reliable restoration. Additionally, severe platform motion leads to the entanglement of event information from turbulence, the scene, and objects, making accurate motion decoupling highly challenging, as illustrated in Fig. 11 (b). In future, we plan to introduce inertial measurement unit and ultra-telephoto optics to overcome these problems.

4.7. More Quantitative and Qualitative Results.

User Study. We conduct a user study on LATH dataset. We randomly select 50 samples along with the corresponding restored results produced by each method, and invite 20

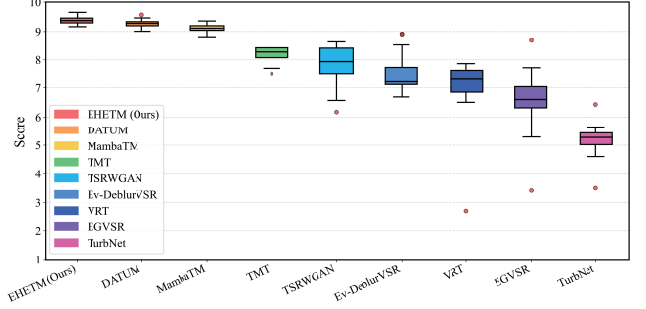


Figure 12. User study of turbulence mitigation on LATH dataset.

participants to score the restorations on a scale from 1 to 10, where higher scores indicate better perceptual quality. The score distributions for all methods are shown in Fig. 12. It is observed that our EHETM achieves the highest scores, suggesting that users prefer our results.

Visual Comparisons. We provide more visual comparisons against the eight typical methods on LATH, CTTH, and UDET datasets, as shown in Fig. 13, Fig. 14, and Fig. 15, respectively. On LATH dataset, EHETM shows strong generalization ability, producing clearer textures and more stable structures than other approaches. On CTTH dataset, EHETM achieves superior restoration performance, preserving motion details while suppressing turbulence-induced distortions. On UDET dataset, our EHETM further demonstrates its advantage by delivering the highest visual fidelity and recovering fine-grained details with excellent structural consistency. These results collectively verify that our method consistently outperforms existing approaches across diverse and challenging turbulence conditions.

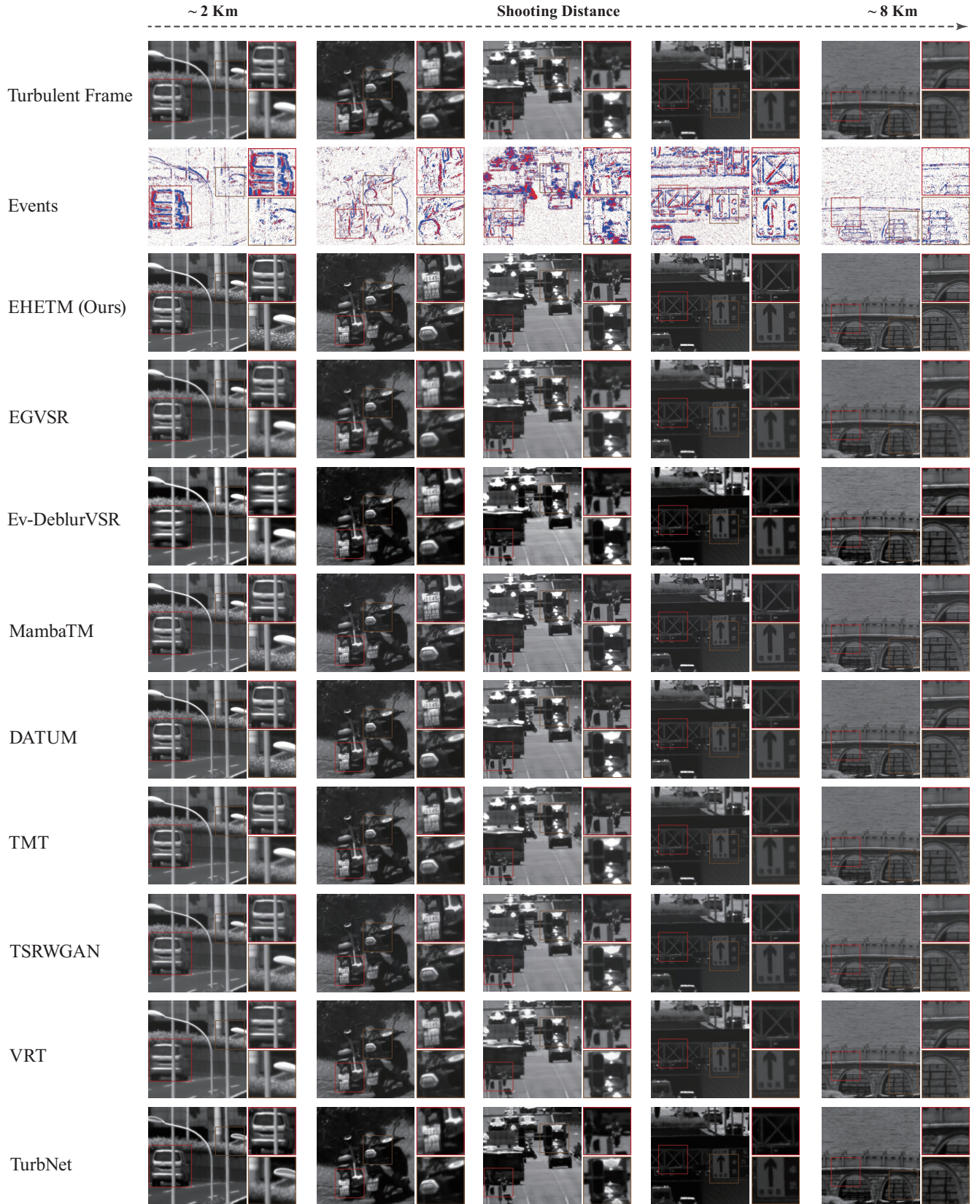


Figure 13. Additional visual comparison on LATH Dataset. Our EHETM produces clearer textures and more stable structures, demonstrating strong generalization ability to real-world atmospheric turbulence scenes.

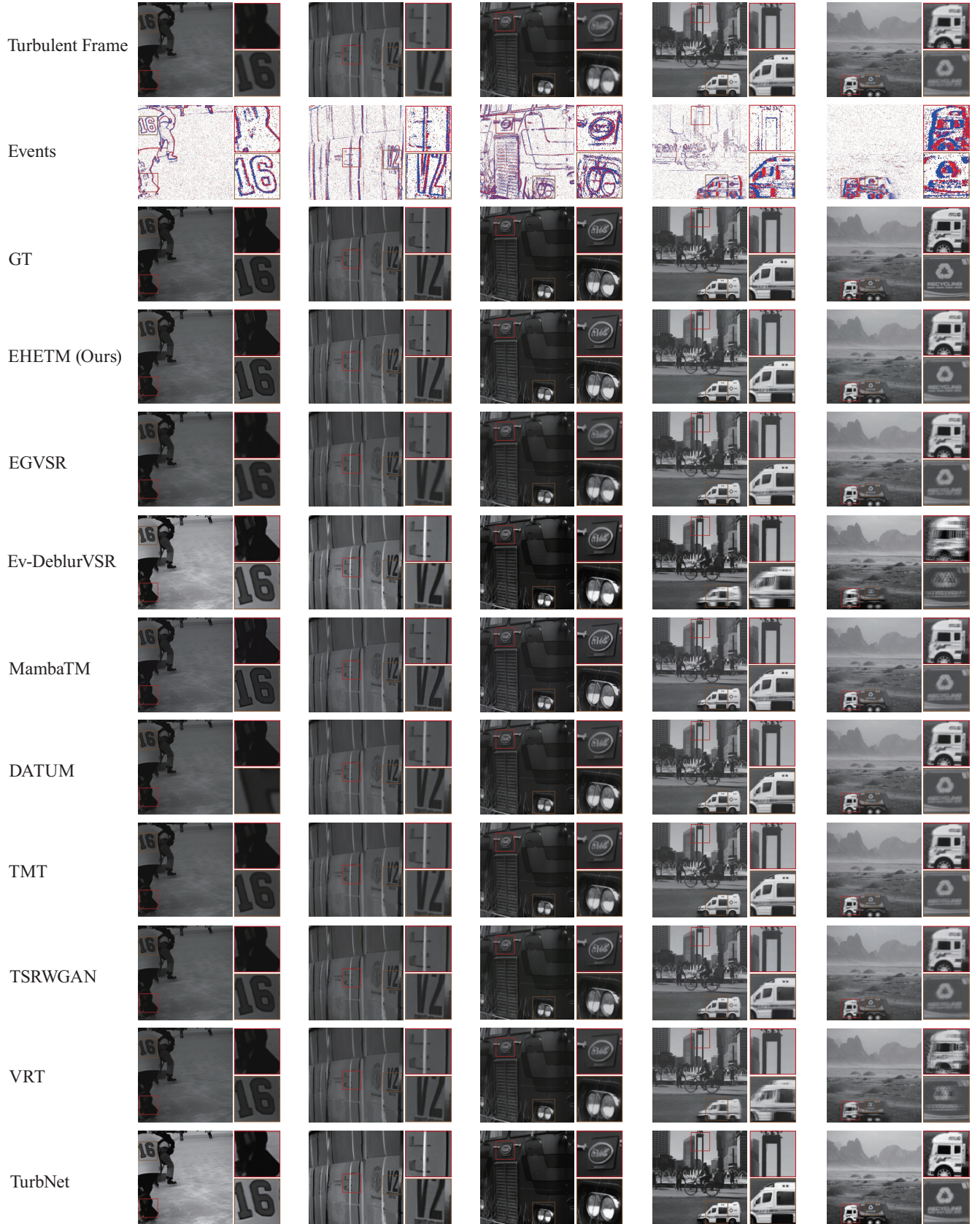


Figure 14. Additional visual comparison on CTTH Dataset. Our EHETM achieves the best restoration quality under dynamic object scenes, effectively preserving motion details and suppressing turbulence-induced distortions.

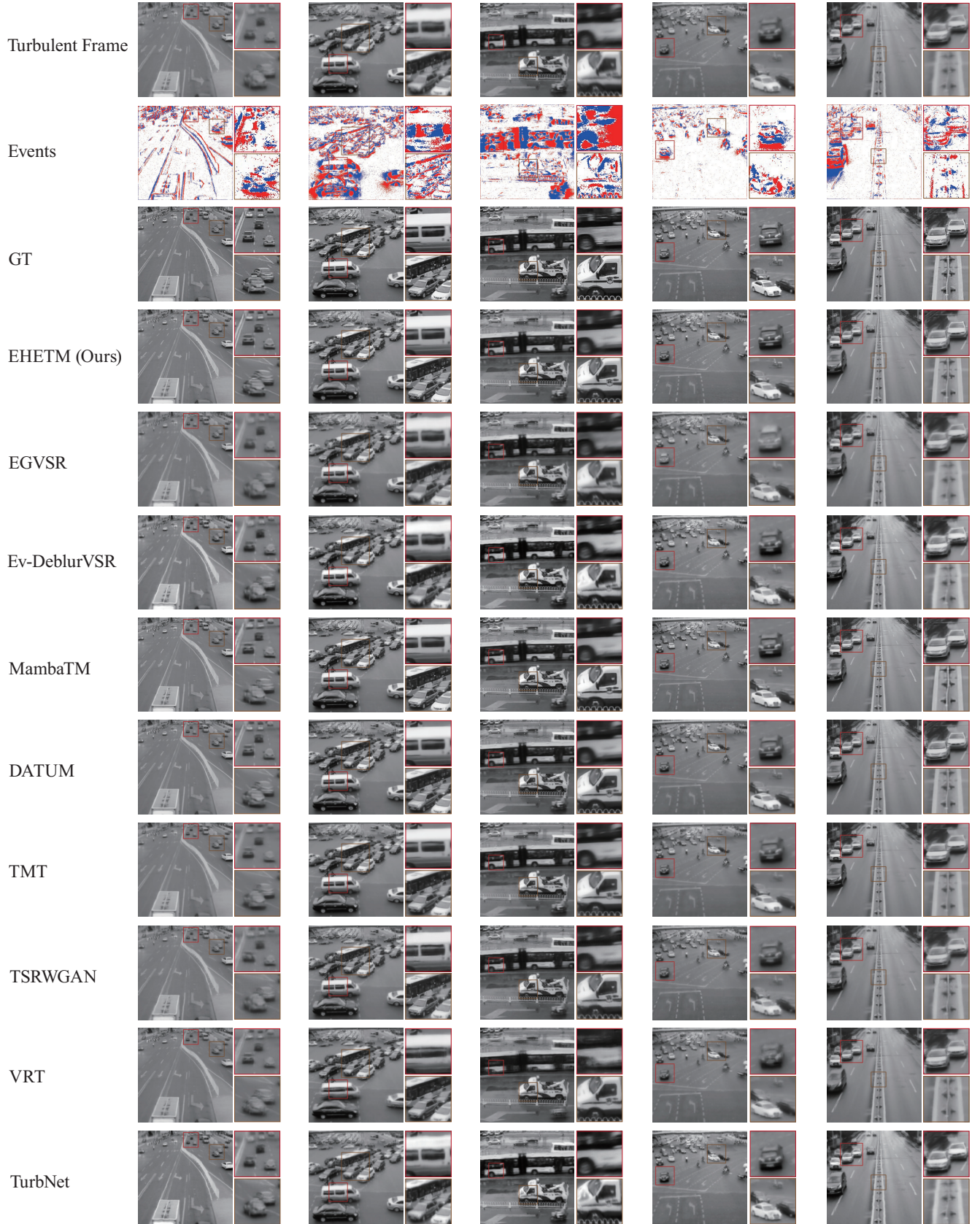


Figure 15. Additional visual comparison on UDET Dataset. Our EHETM achieves the highest visual quality, effectively recovering fine details and maintaining structural consistency.

References

- [1] Alpsentek official website. <https://www.alpsentek.com/>. 4
- [2] Saad Ali and Mubarak Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–6. IEEE, 2007. 1
- [3] Fanpeng Kong, Andrew Lambert, Damien Joubert, and Gregory Cohen. Shack-hartmann wavefront sensing using spatial-temporal data from an event-based image sensor. *Optics Express*, 28(24):36159–36175, 2020. 5
- [4] Bongki Moon, Hosagrahar V Jagadish, Christos Faloutsos, and Joel H. Saltz. Analysis of the clustering properties of the hilbert space-filling curve. *IEEE Transactions on knowledge and data engineering*, 13(1):124–141, 2001. 3
- [5] Hongtao Wu, Yijun Yang, Huihui Xu, Weiming Wang, Jinni Zhou, and Lei Zhu. Rainmamba: Enhanced locality learning with state space models for video deraining. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7881–7890, 2024. 3
- [6] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 3
- [7] Xingguang Zhang, Nicholas Chimitt, Yiheng Chi, Zhiyuan Mao, and Stanley H Chan. Spatio-temporal turbulence mitigation: A translational perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2889–2899, 2024. 4
- [8] Xingguang Zhang, Zhiyuan Mao, Nicholas Chimitt, and Stanley H Chan. Imaging through the atmosphere using turbulence mitigation transformer. *IEEE Transactions on Computational Imaging*, 10:115–128, 2024. 5
- [9] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 2