EvTurb: Event Camera Guided Turbulence Removal

Yixing Liu†, Minggui Teng†, Yifei Xia, Peiqi Duan,
Boxin Shi‡, *Senior Member, IEEE*

***Abstract*—Atmospheric turbulence degrades image quality by introducing spatially varying blur and geometric tilt distortions, posing significant challenges to downstream computer vision tasks including segmentation and recognition. Existing single-image and multi-frame methods struggle with the highly ill-posed nature of this problem due to the compositional complexity of turbulence-induced distortions. To address this, we propose EvTurb, an event guided turbulence removal framework that leverages high-speed event streams to decouple blur and tilt effects. EvTurb uses event-derived cues in a novel two-step event-guided network: event integrals are first employed as empirical guidance to reduce blur in the coarse outputs. This is followed by employing a variance map, derived from raw event streams, to eliminate the tilt distortion for the refined outputs. Additionally, we present TurbEvent, the first real-captured dataset featuring diverse turbulence scenarios. Experimental results demonstrate that EvTurb surpasses state-of-the-art methods while maintaining computational efficiency.**

***Index Terms*—Computational Photography, Event Camera, Turbulence Removal.**

I. Introduction

Atmospheric turbulence refers to small-scale, irregular air motions, typically induced by heat, that degrade image quality through spatially varying changes in the index of refraction along the optical path. These variations introduce complex blur and geometric tilt effects (the top of Figure 1), compounded over a fixed exposure duration. While recent methods [1], [3] have been proposed for single-image atmospheric turbulence removal, their effectiveness remains limited (Figure 1(d)) due to the extremely ill-posed nature of the problem.

To address this challenge, multi-frame methods [4], [5] leverage temporal information from image sequences, helping to mitigate the ill-posed nature of the problem (*e.g.*, utilizing the lucky frame phenomenon [6]). While these approaches generally outperform single-image methods [1], [3], frame-based methods require large amounts of data and are further complicated by object motion. Computational photography approaches [7], [8] require additional hardware to mitigate turbulence effects and further enhance the performance of turbulence removal. For example, narrow-band filters [7] have been employed to mitigate turbulence by integrating them with conventional RGB imaging. Nevertheless, owing to the limitations of traditional camera frame rates, tilt and blur distortions are only partially alleviated and remain compounded in the recorded images.

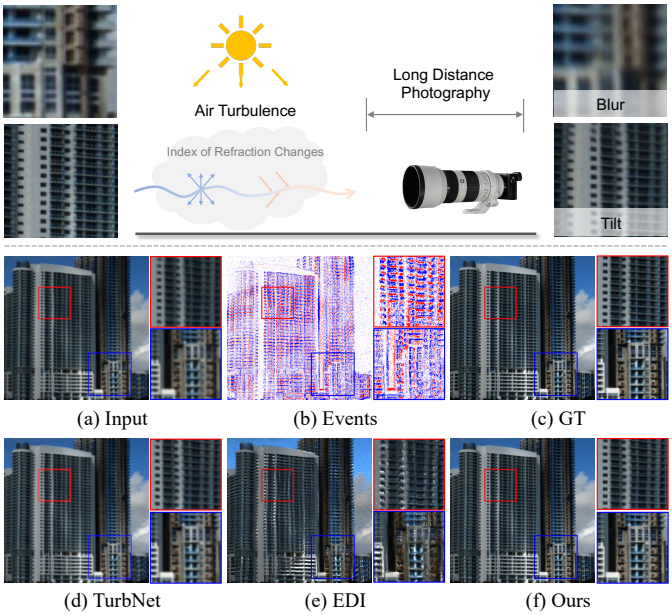


Fig. 1. Turbulent images are typically captured through atmospheric turbulence in long-distance photography, leading to distortions such as blur and geometric tilt effects (top). Given a turbulent image (a) and corresponding events (b) triggered during its exposure, we aim to restore a turbulence-free image (c). The single-image method TurbNet [1] struggles to fully remove turbulence (d), while the event-based deblurring method EDI [2] introduces significant artifacts (e). In contrast, our EvTurb method produces a sharper image with fewer distortions (f).

Event cameras [9], a class of neuromorphic sensors, produce asynchronous streams of events whenever changes in pixel intensity exceed predefined thresholds. Unlike conventional frame-based cameras that record redundant information at fixed intervals, event cameras operate with microsecond-level temporal resolution, thereby capturing scene dynamics with exceptional efficiency and low latency. This capability enables a wide range of applications, including high-speed video reconstruction from event streams [10]–[12] and motion deblurring [13]–[15], where the fine-grained temporal cues provide advantages unattainable by traditional imaging sensors. This feature further facilitates high-speed observation of dynamic scenes, which motivates us to explore: *Can event cameras capture turbulence field variations and separately characterize blur- and tilt-related cues using high-speed event signals?*

In this paper, we introduce the **EvTurb** framework, an

† Contributed equally to this work as first authors.

‡ Corresponding author.

Minggui Teng, Yifei Xia, Peiqi Duan, and Boxin Shi are with the State Key Laboratory for Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100871, China. Email: {minggui_teng, yfxia, duanqi0001, shiboxin}@pku.edu.cn

Yixing Liu is with the School of Engineering and Computing, Rice University, Houston 77005, USA. This work was completed when he was an undergraduate student at Peking University. Email: yl366@rice.edu

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. The material includes additional dataset details, robustness analysis, and qualitative results. This material is 8.6 MB in size.

Project page: https://github.com/ChipsAhoyM/EvTurb

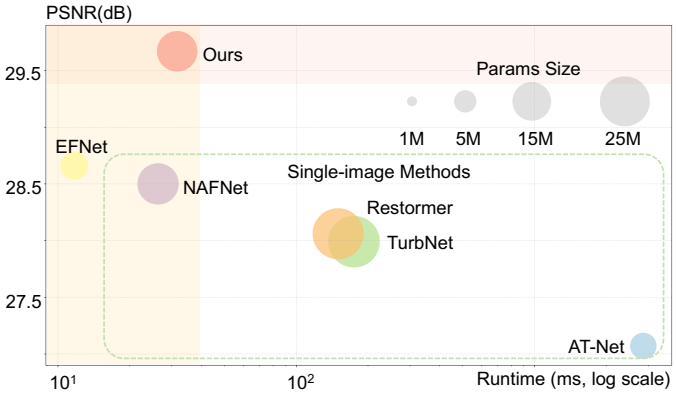


Fig. 2. A comparison of PSNR, runtime, and model size. PSNR is evaluated on the TurbEvent dataset, and runtime is measured on an RTX 3090 GPU. By using an event-guided formulation for image turbulence removal and incorporating high-speed information into the restoration process, our approach not only achieves the highest PSNR, delivering superior output quality but also maintains a relatively short runtime.

Event camera-guided **Turb**ulence removal approach designed to mitigate atmospheric turbulence effects in single-image restoration by exploiting high-speed event streams. Unlike conventional cameras, event sensors capture fine-grained variations in the turbulence field during the exposure period, recording both the intensity and frequency of these fluctuations. However, existing event-based deblurring methods primarily address motion blur and thus cannot be directly extended to turbulence scenarios, where tilt and blur effects are coupled, often leading to residual distortions and additional artifacts (Figure 1(e)).

To overcome this limitation, we formulate an event-driven turbulence removal model in which event-recorded intensity changes provide an empirical surrogate for temporally accumulated blur-related degradation, while event-triggering frequency characterizes tilt-related turbulence fluctuations. This formulation leverages the unique property of events to provide complementary high-speed cues for decoupling blur and tilt components and reducing the ill-posedness of the restoration problem. Building on this model, we design a two-stage event-guided turbulence removal network that explicitly incorporates temporal information from events. In the first stage, event integrals, representing the cumulative intensity of turbulence-induced variations, are fused with RGB inputs to produce coarse reconstructions that primarily address blur. In the second stage, a variance map computed from raw event streams captures the frequency of turbulence fluctuations and is integrated with the coarse output to suppress tilt distortions. This progressive design yields refined results with sharper structures and more stable geometry, as illustrated in Figure 1(f). Overall, our contributions include:

- establishing an event-guided formulation that links high-speed event data with turbulence-distorted images by decoupling blur and tilt effects;
- proposing a two-step framework that integrates turbulence field variations from event data and RGB images to separately address blur and tilt distortions; and
- curating the first real-captured TurbEvent benchmarking dataset, featuring diverse turbulence patterns for comprehensive evaluation.

Quantitative and qualitative results on our newly collected TurbEvent dataset demonstrate that our method surpasses state-of-the-art techniques in recovering turbulence-free images with reduced distortion. As shown in Figure 2, our method achieves an improvement of over 1 dB in PSNR while maintaining a relatively short runtime.

II. Related Work

In this section, we provide a brief overview of image-based turbulence removal methods. Additionally, we review event-based deblurring techniques, which are partially relevant to blur removal from event data.

Atmospheric turbulence mitigation. The study of turbulence mitigation has long relied on traditional methods such as deconvolution for images with a limited field of view, assuming spatially invariant turbulence [16], and selecting lucky frames when addressing spatially variant turbulence [6], [17]. To improve upon these classical techniques, especially for anisoplanatic conditions, unified methods have been developed to handle both static and dynamic scenes by integrating physics-constrained priors and advanced frame selection metrics [18]. More recently, deep learning approaches have emerged, categorized into single-frame and multi-frame methods. Single-frame strategies employ backbones including CNNs [19], Transformers [1], Generative Adversarial Networks (GANs) [20]–[23], and Denoising Diffusion Probabilistic Models for turbulence correction [24], alongside methods integrating RGB with narrow-band images [7]. An alternative to purely data-driven models are hybrid approaches that combine the modeling power of deep neural networks with the accuracy of physics-based models, using techniques like digital holography and model-based iterative reconstruction to correct phase errors from a single measurement [25]. Recent efforts have further strengthened this direction by introducing physics-driven turbulence image restoration with stochastic refinement, which explicitly incorporates physical imaging priors into the restoration pipeline while leveraging learned refinement for improved image quality [26]. However, these single-image techniques often overlook crucial temporal information.

In contrast, multi-frame methods effectively utilize temporal data, either using transformer based architectures [4], [27] or recurrent models [5]. Some methods use special techniques to facilitate this process, such as statistical prior [27], neural representation [28], [29], or adaptations to special domains (*e.g.*, planetary images [30]). Despite their efficacy, they require substantial datasets of 10 to 12 turbulent frames and face challenges related to varying turbulence and object motion between frames. The significant data requirement for training these deep learning models underscores the importance of efficient and realistic turbulence simulation, with recent advancements enabling real-time, high-resolution simulation that overcomes the computational bottlenecks of traditional methods [31].

Event-based deblurring. Event-based image deblurring exploits the intrinsic connection between motion events and blurry images to reconstruct sharp details. A basic model of this field is the Event-based Double Integral (EDI) model [2], which represents event streams and sharp/blurry images in the log domain. To improve robustness against noise and enhance fine details, CNN-based methods were later introduced [32], [33]. More recently, transformer architectures [15], [34] have been leveraged, allowing attention mechanisms to better bridge the domain gap between event data and images, leading to more precise reconstructions. Additionally, event representations [13] were explored to achieve a better fusion of image and event modalities. To address the challenge of image and event data domain gap, cross-modal calibration strategy was incorporated to address spatial and temporal alignment [35]. Meanwhile, Yu *et al.* [36] proposed the Scale-Aware Spatio-Temporal Network, which integrates spatial and temporal features for more effective image restoration.

Prior studies [13], [15], [32], [33] have demonstrated that event cameras offer a promising solution to the motion deblurring challenges faced by conventional imaging systems, owing to their ability to efficiently capture dynamic scene changes. Unlike traditional cameras that record intensity frames, event cameras asynchronously encode changes in scene brightness, enabling accurate representation of rapid motion and making them particularly suitable for modeling turbulence-induced variations. Boehrer *et al.* [37] utilized event streams to reconstruct high-speed videos and subsequently applied video-based turbulence removal methods; however, their approach did not explicitly model the relationship between events and turbulence-degraded images. More recently, Li *et al.* [38] proposed an event-guided turbulence mitigation method inspired by lucky-frame selection. Yet, this approach inherently relies on multiple frames and assumes a static or quasi-static scene, substantially limiting its applicability to real-world scenarios involving dynamic motion. To address these limitations, we integrate motion cues from event streams directly into turbulence removal frameworks at the pixel level, thereby leveraging the rich temporal information in continuous event data for more effective turbulence restoration.

III. Method

In this section, we first introduce the concepts of event cameras and event-based turbulence formation in Section III-A. Next, we present a two-step strategy for reconstructing turbulence-free images, guided by turbulence information encoded in event streams, in Section III-B. We then introduce our EvTurb framework in Section III-C. The curation process of the TurbEvent dataset is described in Section III-D, followed by implementation details in Section III-E.

A. Formulations

Clear image formation with events. Event cameras [9] are designed to detect changes in radiance in the logarithmic domain. When the intensity change surpasses a preset threshold c, an event signal is generated as a quadruple, denoted as (x, y, t, p), *i.e.*,

$$\log(\mathbf{I}^{t}_{x,y}) - \log(\mathbf{I}^{t_0}_{x,y}) = p \cdot c, \tag{1}$$

where $\mathbf{I}^{t}_{x,y}$ and $\mathbf{I}^{t_0}_{x,y}$ denote the intensity values of a pixel of (x, y) at times t and t_0, respectively. The polarity $p \in \{1, -1\}$ indicates whether the intensity is increasing or decreasing. Since Equation (1) is applied independently for each pixel, the pixel indices are omitted henceforth.

Given two consecutive clear images $\mathbf{I}^0$ and $\mathbf{I}^1$ without turbulence, along with the corresponding N_e events $\{e_k\}^{N_e}$ triggered between them, we can connect these two frames using the event integral:

$$\mathbf{I}^1 = \mathbf{I}^0 \cdot \exp(\int_{t_0}^{t_1} c_k \cdot p_k \ \mathrm{d}k), \tag{2}$$

where c_k represents the spatial-temporal variant threshold, which is dependent on the scene conditions [39].

Turbulent image formation with events. A turbulent image, denoted as $\tilde{\mathbf{I}}$, can be generated using the classical split-step wave-propagation equation. For simplicity, this process can be implemented as a forward equation [1], approximating turbulence distortion as a combination of two degradation operators:

$$\tilde{\mathbf{I}} = \mathcal{B} \circ \mathcal{T}(\mathbf{I}) + \mathbf{N}, \tag{3}$$

where $\mathcal{B}$ represents the spatially varying blur, $\mathcal{T}$ denotes the geometric pixel displacement (commonly known as the tilt). An example is shown in Figure 1. $\mathbf{N}$ is the additive noise signal, and $\circ$ denotes functional composition. If only a single turbulent image is available, the blur and tilt operations are inseparable, as these distortions are coupled together during the exposure time.

As shown in Equation (2), the latent scene variations during the exposure are linked to the event stream. Under atmospheric turbulence, tilt distortion mainly originates from air refraction. Although each instantaneous scene state may still contain a certain amount of residual blur due to instantaneous wavefront distortion, this blur is substantially weaker than that in the long-exposure observation, where distortions are further accumulated over time [40]. Therefore, in the formation of the observed turbulence-degraded image, temporal averaging over these instantaneous states plays a more dominant role than the residual blur present in any individual state. By combining Equation (2) and Equation (3), we obtain the following formulation:

$$\begin{aligned}\tilde{\mathbf{I}} &\approx \mathcal{T}(\mathbf{I}) \cdot \frac{1}{T} \int_T \left(\exp(\int_t c_k \cdot p_k \ \mathrm{d}k) \right) \mathrm{d}t + \mathbf{N} \\ &= \mathcal{T}(\mathbf{I}) \cdot \mathbf{E} + \mathbf{N}.\end{aligned} \tag{4}$$

The event-integral term $\mathbf{E}$ represents a double integral over the events occurring within the exposure period T. In our formulation, $\mathbf{E}$ serves as an empirically motivated surrogate for the temporally accumulated degradation component associated with turbulence-induced blur, rather than a complete analytical model of all higher-order phase aberrations in

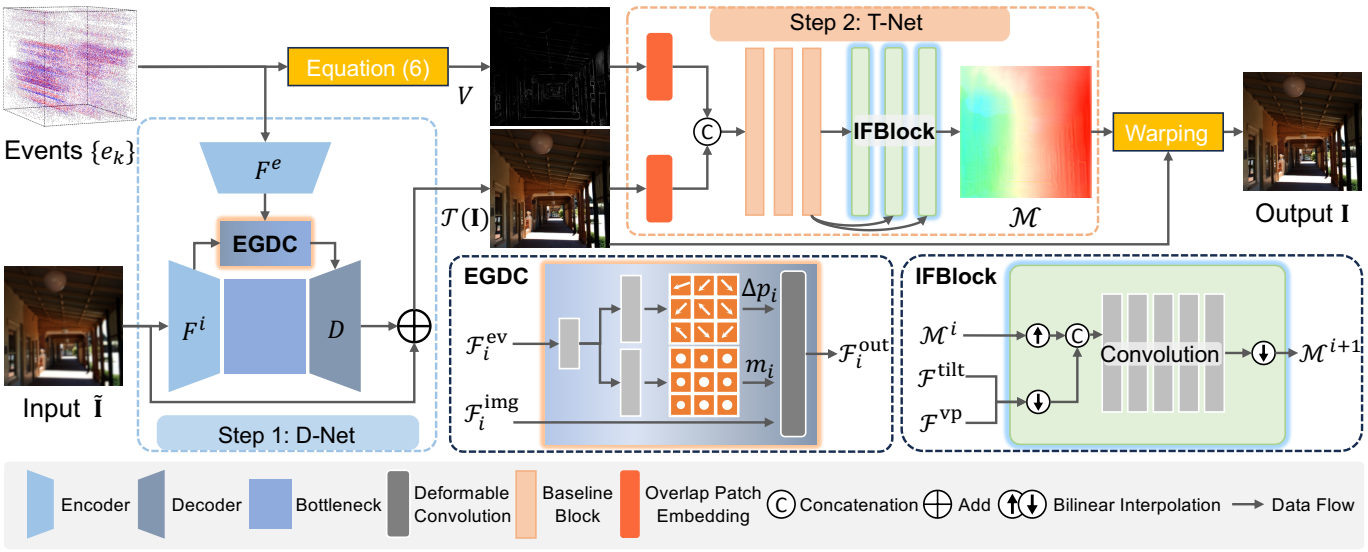


Fig. 3. The overall framework of the proposed EvTurb method. We first perform deblurring using a multi-scale architecture, D-Net, which estimates the event integral from two modalities: the RGB image $\tilde{\mathbf{I}}$ and the event stream $\{e_k\}$. The event-guided deformable convolution (EGDC) block is applied during the decoding process to enhance data fusion. After obtaining the blur-free image $\mathcal{T}(\mathbf{I})$, T-Net predicts the tilt flow $\mathcal{M}$ guided by the calculated variance map $\mathbf{V}$. The IFBlock is then used to iteratively refine the flow at lower resolutions. Finally, the estimated flow $\mathcal{M}$ is used to warp $\mathcal{T}(\mathbf{I})$, generating the output $\mathbf{I}$.

atmospheric wave propagation. In particular, this surrogate does not explicitly model scintillation or anisoplanatism, and any residual instantaneous wavefront blur is absorbed into the noise/residual term $\mathbf{N}$. By leveraging the capability of high-speed event cameras to capture defocus kernels [14], the proposed surrogate further incorporates local defocus effects, which can then be refined through event-based guidance, enabling a more accurate, though still approximate, representation of turbulence-induced degradation.

B. Event guided turbulence removal

Blur removal. The formation model in Equation (4) shows that turbulence distortion is compositional, comprising tilt and blur. By leveraging the high-speed turbulence field information captured in the event stream, the event-integral surrogate $\mathbf{E}$ provides an empirical cue for estimating the temporally accumulated degradation component associated with blur. We use $\mathbf{E}$ to guide an approximate least-squares inversion step, with the caveat that this inversion is approximate, since $\mathbf{E}$ is an empirically motivated surrogate for the temporally accumulated turbulence-induced variations rather than the exact blur operator. Recognizing the presence of noise, we further transform Equation (4) into the following least-squares optimization problem:

$$\mathcal{T}(\mathbf{I})^* = \underset{\mathcal{T}(\mathbf{I})}{\operatorname{argmin}} \|\tilde{\mathbf{I}} - \mathcal{T}(\mathbf{I}) \cdot \mathbf{E}\|^2. \tag{5}$$

Solving this objective facilitates the initial recovery of images $\mathcal{T}(\mathbf{I})^*$ that are less blurred yet exhibit tilt distortion.

Tilt removal. Since the event integral is also affected by tilt distortion, it is necessary to explore the temporal information within the event stream to further enhance tilt removal. Inspired by previous work [3], we recognize that variance maps can indicate pixel uncertainty related to tilt intensity. We construct a variance map from the raw event streams, capturing the frequency of variations in the turbulence field and indicating the tilt distortion uncertainty. As events record the intensity value changes of latent frames, we define the variance of the entire latent frame sequence as the variance map. Formally, we calculate the variance map as follows:

$$\mathbf{V} = \texttt{norm}\left(\frac{1}{N_e}\sum_{i=1}^{N_e}\mathcal{E}_i^2 - (\frac{1}{N_e}\sum_{i=1}^{N_e}\mathcal{E}_i)^2\right), \tag{6}$$

where $\texttt{norm}$ denotes the normalization operation and $\mathcal{E}_i = \sum_{k=1}^{i}\exp(c_k \cdot e_k)$ represents the accumulated events.

Utilizing the variance map $\mathbf{V}$ as guidance, tilt flow $\mathcal{M}$ can be formulated under Maximum-a-Posteriori:

$$\mathcal{M}^* = \underset{\mathcal{M}}{\operatorname{argmax}}\, \mathbb{P}(\mathcal{M} \circ \mathcal{T}(\mathbf{I})^* \mid \mathbf{V}), \tag{7}$$

where $\mathbb{P}(\cdot)$ represents the distribution of natural images without tilt effects, and $\circ$ denotes the image warping function. By estimating the optimal tilt flow $\mathcal{M}^*$ using natural image priors, we can effectively apply it to the initial image for the tilt correction.

It is worth noting that the reliability of the variance map may be affected by local image structures, especially high-contrast edges that can produce stronger event responses. Nevertheless, the proposed variance map is computed from temporally accumulated event stacks and mainly reflects turbulence-induced temporal fluctuations rather than instantaneous event density. Furthermore, it is introduced as an auxiliary prior and jointly processed with image features, which reduces the influence of structural contrast and improves the robustness of tilt estimation.

C. EvTurb Framework

Our pipeline is illustrated in Figure 3. Motivated by the turbulence formulation in Equation (4), we use the event-integral surrogate to guide the deblurring stage while keeping its approximate nature explicit, and then estimate the tilt component in a separate stage. Following the physical tilt-then-blur image formation process [41], we design a two-stage restoration framework. Specifically, the pipeline first performs deblurring with D-Net to recover sharper structural details. Then, T-Net estimates the tilt field, based on which a warping operation is applied to geometrically correct the turbulence-induced distortion.

D-Net. The key component for optimizing Equation (5) is to obtain an accurate event integral. Since the thresholds of event cameras exhibit spatial-temporal variations [39], directly integrating raw event data introduces artifacts into the restored images. Instead, we propose estimating this integral in a data-driven manner, converting Equation (5) into a regression problem, *i.e.*:

$$\mathcal{T}(\mathbf{I}) = D(F^i(\tilde{\mathbf{I}}), F^e(\{e_k\})), \tag{8}$$

where F^i and F^e are implicit functions for encoding images and events, respectively, and D is an implicit function for decoding and fusing information from both modalities.

To enhance feature extraction from both event data and RGB images, we propose a dual-branch network, termed D-Net, specifically designed to fuse these two modalities in a complementary manner. Previous studies [15], [33] have demonstrated that the U-Net architecture is highly effective for image restoration tasks, as its hierarchical design enables the extraction of features across multiple scales while preserving fine structural details. Motivated by these findings, we adopt a multi-scale U-Net structure that facilitates cross-modal fusion at different levels of resolution. Furthermore, dense blocks [42] are incorporated into both encoders to promote feature reuse and ensure more efficient propagation of refined representations, thereby improving the quality of the reconstructed outputs.

To further enhance local feature restoration, we introduce an Event-Guided Deformable Convolution (EGDC) module, which incorporates deformable convolution into the skip connections. In this design, event polarity provides critical cues for estimating sampling offsets, while accumulated event intensity highlights salient regions, effectively serving as a spatial mask. Specifically, event features at selected scales are employed to compute both the offset map and the mask map as follows:

$$\Delta p_i = C_p(B(\mathcal{F}_i^{\mathrm{ev}})), \quad m_i = C_m(B(\mathcal{F}_i^{\mathrm{ev}})), \tag{9}$$

where Δp_i, m_i, and $\mathcal{F}i^{\mathrm{ev}}$ denote the offset map, mask map, and the i-th event feature, respectively. Here, B represents a bottleneck convolution layer, while Cp and C_m correspond to two distinct convolutional layers. Based on the computed offsets and masks, the deformable convolution is defined as:

$$\mathcal{F}_i^{\mathrm{out}} = \texttt{Deform}(\mathcal{F}_i^{\mathrm{img}}, \Delta p_i, m_i). \tag{10}$$

where `Deform` denotes the deformable convolution operator [43]. By guiding the deformable convolution with event-derived offsets and masks, the EGDC module effectively enhances the alignment of structural information and restores salient local details within the feature space.

After estimating the event integral, we fuse it with the turbulent image features through a global connection to produce deblurred outputs. This global connection enables the model to capture long-range dependencies between event-derived information and turbulent image representations, ensuring that complementary cues are fully exploited. By leveraging these dependencies, the network more effectively suppresses blur and recovers sharper structures in the restored image.

T-Net. Although D-Net provides coarse restorations, residual tilt effects still remain. Consequently, an essential step in Equation (5) for image reconstruction is to fit a tilt-free natural image prior. Inspired by previous studies addressing spatially variant shifts [44], [45], we introduce a flow-based network, denoted as T-Net, which estimates the tilt flow $\mathcal{M}$ to warp distorted inputs into tilt-free representations. The corresponding optimization in Equation (7) can be expressed as:

$$\mathcal{M} = f_{\mathrm{VT}}(\mathcal{T}(\mathbf{I}), \mathbf{V}), \tag{11}$$

where f_{VT} denotes an implicit function parameterized by T-Net and specifically designed for tilt flow estimation. Here, $\mathbf{V}$ represents the variance map obtained from Equation (6), which encodes the uncertainty introduced by turbulence and provides critical guidance for accurate flow estimation.

In designing T-Net, we adopt overlapped patch embedding, which is effective in capturing both contextual and boundary information. This design choice ensures smooth transitions across neighboring patches and reduces artifacts in the reconstructed outputs. The embedded features are then processed through a sequence of convolutional layers and Baseline Blocks [46] to fuse information from the two modalities in a unified latent representation.

The subsequent step is to estimate a tilt flow from the fused features that can accurately warp the distorted image toward the ground truth. To achieve this, we incorporate IFBlocks [44] for flow estimation, which iteratively refines the predicted flow and gradually converges toward the correct tilt compensation. Formally, the refinement process is defined as:

$$\mathcal{M}^{i+1} = \texttt{IF}(\mathcal{F}^{\texttt{tilt}}, \mathcal{F}^{\texttt{vp}}, \mathcal{M}^i), \tag{12}$$

where $\mathcal{F}^{\texttt{tilt}}$ and $\mathcal{F}^{\texttt{vp}}$ denote the features extracted from the deblurred coarse image and the variance map, respectively, and `IF` refers to the IFBlock operator. As illustrated in Figure 3, the tilt flow is computed at a lower-resolution scale to enforce smoothness and stability in the estimated flow field. Finally, by applying the estimated tilt flow to the coarse results generated by D-Net, we obtain turbulence-free reconstructions with sharper details and more consistent geometry.

D. TurbEvent dataset

Simulating turbulence as a continuous process is inherently difficult due to the absence of a closed-form formulation,

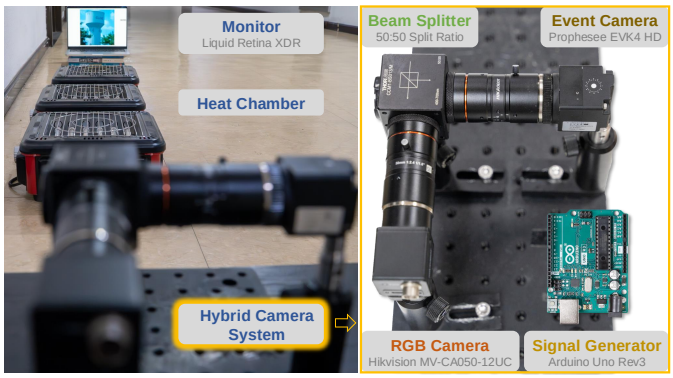


Fig. 4. We construct a hybrid camera system to investigate the relationship between event formation and image turbulence. Additionally, we implement a display-camera system with a heat chamber for the curation of the TurbEvent dataset.

and this challenge is further amplified when attempting to generate high-frame-rate videos for simulated event streams. To overcome these limitations, and inspired by the Event-NFS dataset [47], we construct a display–camera system that directly captures turbulent images, paired turbulence-free references, and the corresponding event streams. Specifically, we sample images from the ADE20K dataset [48], [49] and render them as video sequences at 10 FPS, a frame rate chosen to prevent frame drops and minimize artifacts from display refreshing. Through this process, we successfully collect 6318 triplets, each consisting of a turbulent image, its turbulence-free counterpart, and the associated event stream, all at a resolution of 592×592. We denote this curated resource as the **TurbEvent** dataset, which serves as the first event-based dataset specifically designed for turbulence removal.

The ADE20K dataset [48], [49] is a large-scale semantic segmentation benchmark containing more than $27,000$ images collected from the SUN [50] and Places [51] databases. Each image is densely annotated at the pixel level, covering a broad spectrum of objects, object parts, and, in certain cases, hierarchical part relationships. Key characteristics of ADE20K include:

- Scenes: ADE20K spans 623 scene categories, encompassing both indoor environments (*e.g.*, bedrooms, kitchens, and offices) and outdoor environments (*e.g.*, streets, parks, and beaches).
- Objects: The dataset provides annotations for more than 3000 object categories, ranging from well-defined entities (*e.g.*, cars and people) to amorphous background regions (*e.g.*, grass and sky).

Setup. The setup consists of a Liquid Retina XDR display (resolution: 3024×1964, refresh rate: 120 Hz) and a hybrid imaging system composed of two cameras: a machine vision camera (HIKVISION MV-CA050-12UC) and an event camera (PROPHESEE GEN4.0[1]). Both cameras are equipped with a 50mm F/1.8 lens and are precisely aligned using a beam splitter. The machine vision camera uses a SONY IMX264 sensor (crop factor $\approx$ 3.9), while the event camera uses a SONY IMX636 event sensor (crop factor $\approx$ 6.1). The machine vision images are further warped to align with the event camera view. Consequently, the aligned imaging system corresponds to an effective full-frame focal length of approximately 300mm, which falls in the telephoto range and is suitable for long-range atmospheric turbulence imaging.

[1]https://www.prophesee.ai/event-camera-evk4/

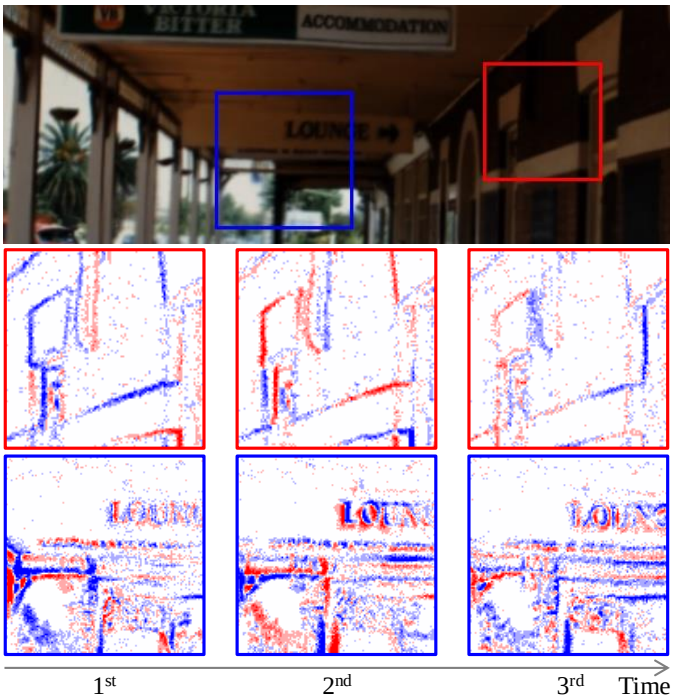


Fig. 5. Visualization of our TurbEvent turbulent video frame (upper) and three repeated event recordings (lower) of the turbulent video. The turbulence patterns in these three event clips are distinct and exhibit randomness.

For spatial calibration, we align the views of the two cameras using a checkerboard pattern to perform geometric calibration. Temporal synchronization is achieved via a signal generator, ensuring frame-level correspondence between the machine vision and event cameras. During data acquisition, we extract images from a predefined region on the display, marked for consistency across recordings.

For image data, raw frames are captured and subsequently processed with an ISP pipeline to ensure color consistency and maintain high image quality. Identical white balance and tone-mapping parameters are applied across all frames. Regarding dataset construction, we first select high-resolution images from the source dataset. Since these images are not evenly distributed across the original training and testing splits, we resample them into balanced train and test sets based on their labels to ensure a representative distribution.

Turbulence pattern. An illustrative example is shown in Figure 5, where a frame from the ADE20K dataset [48], [49] is paired with its corresponding event patches, captured at consecutive timestamps using our hybrid camera system. For each scenario, we consecutively record three event streams under identical imaging conditions. Despite the fixed setup, the resulting event patches exhibit distinct appearances due to the inherent randomness of turbulence. In some cases, edge signals

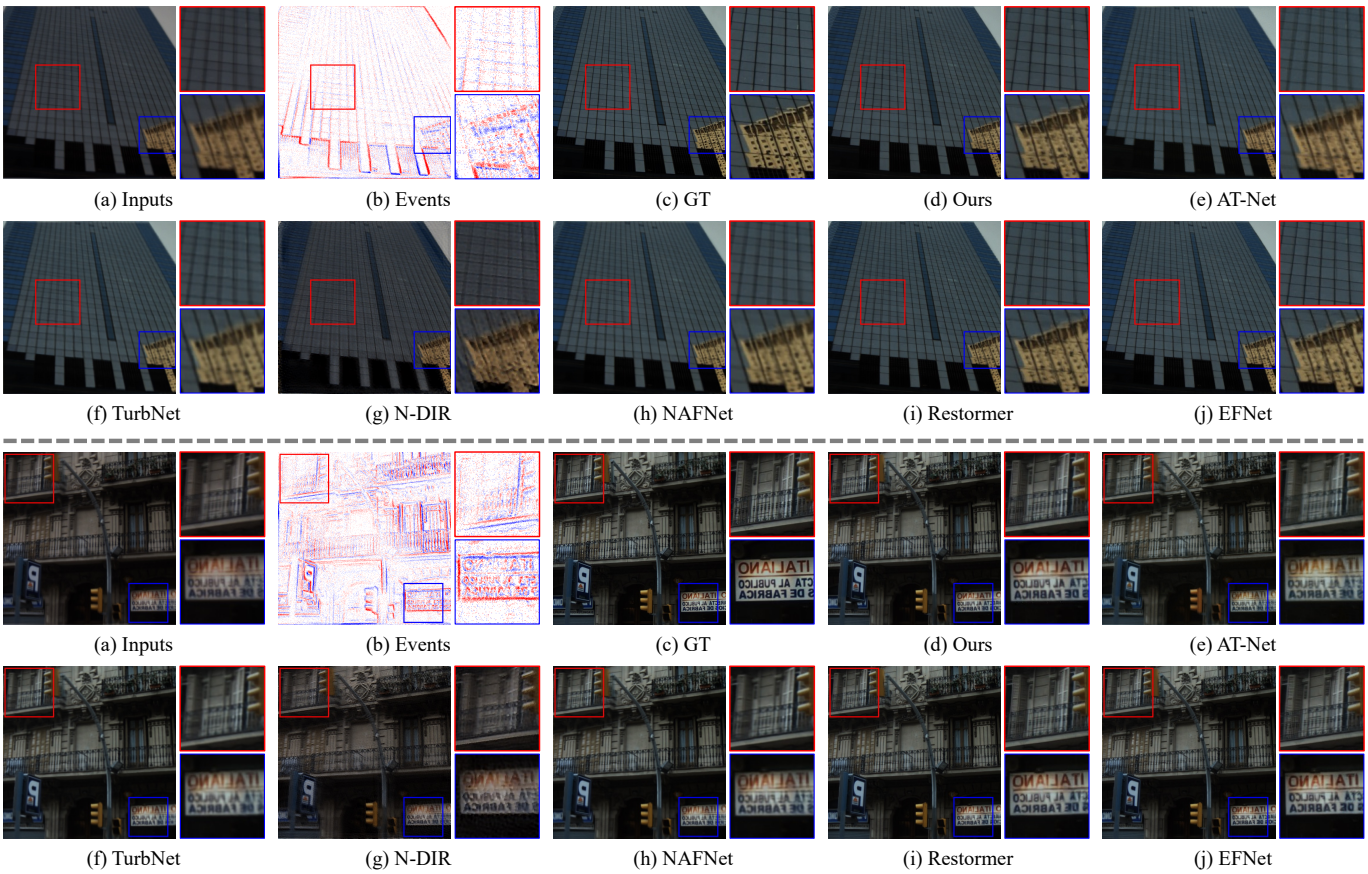


Fig. 6. Visual quality comparison with turbulence removal methods on the TurbEvent dataset. (a) Input images. (b) Events. (c) Ground truth. (d)∼(j) Results of ours, AT-Net [3], TurbNet [1], N-DIR [52], NAFNet [46], Restormer [53], and EFNet [15]. Close-up views are provided on the right side of each image. More results are in the supplementary material.

TABLE I
QUANTITATIVE COMPARISONS ON OUR NEWLY COLLECTED TURBEVENT DATASET. THE "EVENT" COLUMN SPECIFIES IF METHODS REQUIRE EVENTS AS INPUT (YES [✓] OR NO [✗]). ↑ (↓) INDICATES THE HIGHER (LOWER), THE BETTER THROUGHOUT THIS PAPER. THE BEST PERFORMANCES ARE HIGHLIGHTED IN **BOLD**.

	Event	PSNR ↑	SSIM ↑	LPIPS ↓
ATNet [3]	✗	27.07	0.7491	0.3416
TurbNet [1]	✗	27.99	0.7872	0.3418
N-DIR [52]	✗	23.19	0.6941	0.4560
NAFNet [46]	✗	28.50	0.8190	0.3218
Restormer [53]	✗	28.06	0.7995	0.2163
AT-DDPM [24]	✗	23.45	0.714	0.395
Restormer* [53]	✓	28.63	0.8114	0.2122
EFNet [15]	✓	28.66	0.8095	0.2140
Ours	✓	**29.67**	**0.8405**	**0.2010**

are missing, while in others, polarity reversals occur at the same spatial location across different captures. By collecting a large volume of such data, the TurbEvent dataset encompasses a diverse set of turbulence patterns. This randomness highlights the difficulty of synthesizing realistic events from simulated high-speed turbulence videos and further justifies the necessity of acquiring a real-world event-based turbulence dataset.

E. Implementation details

Loss function. To effectively train both D-Net and T-Net, we adopt a composite loss function that balances pixel-level fidelity and perceptual quality. Specifically, we utilize a weighted combination of Mean Square Error (MSE) loss and perceptual loss [54]:

$$\mathcal{L} = \lambda_1 \mathcal{L}_2(\mathbf{I}_G, \mathbf{I}_O) + \lambda_2 \mathcal{L}_{\texttt{perc}}(\mathbf{I}_G, \mathbf{I}_O), \tag{13}$$

where $\mathbf{I}_G$ and $\mathbf{I}_O$ denote the ground-truth and generated images, respectively. The MSE term enforces low-level reconstruction accuracy by penalizing pixel-wise differences, while the perceptual term encourages structural and semantic consistency by comparing feature representations extracted from a pre-trained VGG16 network [55]. Through empirical validation, we set $\lambda_1 = 1.0$ and $\lambda_2 = 0.02$, which provides a balance between sharpness preservation and stability in training.

Training details. All experiments are implemented in the PyTorch framework and conducted on an NVIDIA RTX 3090 GPU. We adopt an end-to-end training strategy, jointly optimizing D-Net and T-Net for 150 epochs. The Adam optimizer is used with an initial learning rate of 1×10^{-4}. To stabilize convergence and prevent overfitting to a fixed learning rate,

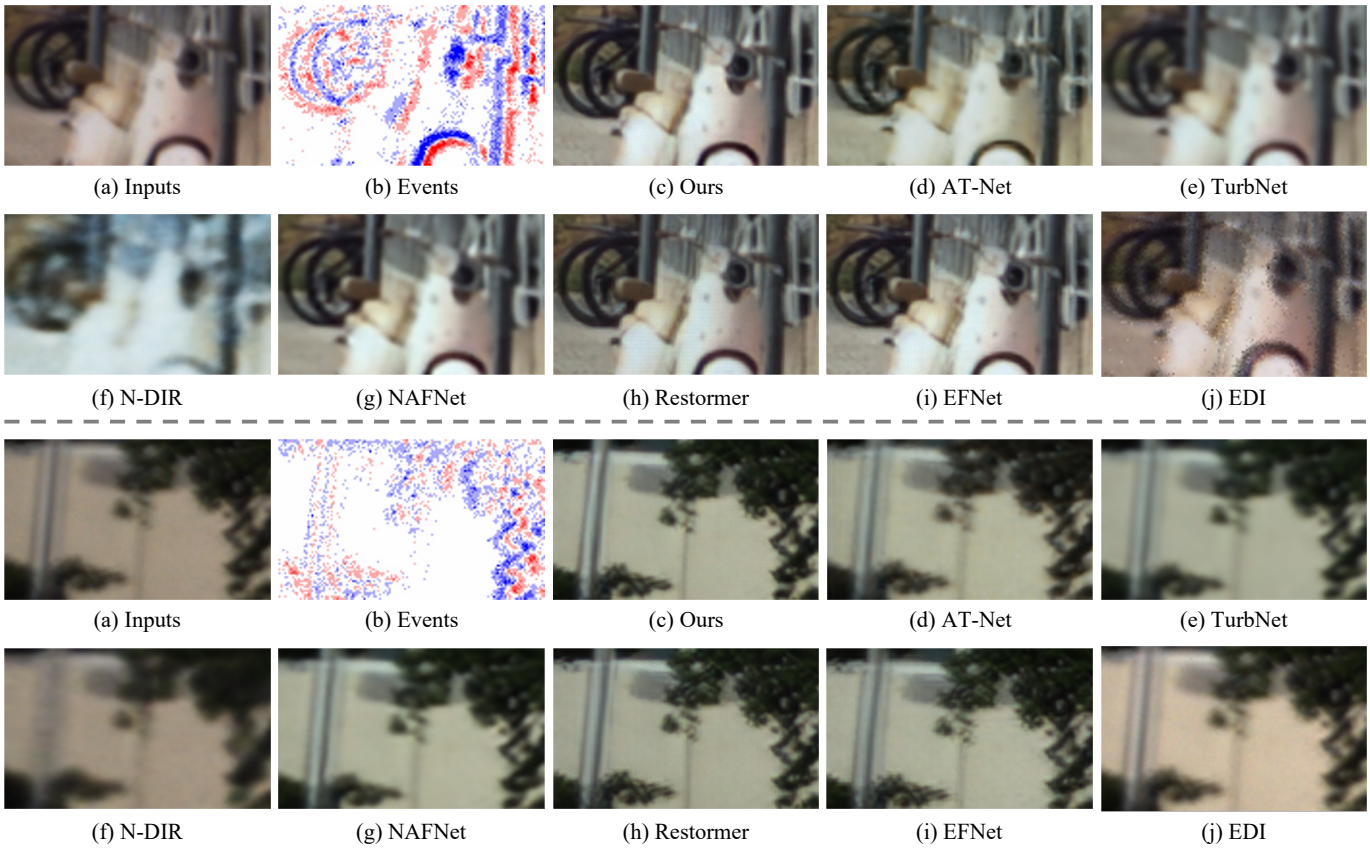


Fig. 7. Visual quality comparison with turbulence removal methods on the real-captured data. (a) Input images. (b) Events. (c)∼(j) Results of ours, AT-Net [3], TurbNet [1], N-DIR [52], NAFNet [46], Restormer [53], EFNet [15], and EDI [2].

we apply a cosine annealing scheduler that gradually decays the learning rate to zero throughout training.

To improve robustness and generalization to real-world scenarios, we employ standard data augmentation techniques. In particular, random cropping is applied to increase the diversity of training samples and to promote invariance to image scale and spatial variations. During training, the batch size is set to 4. In addition, events are accumulated within a fixed temporal window to construct an event stack [56] as input. This representation suppresses stochastic event noise while preserving turbulence-related structural cues, thereby improving robustness to noise in practice.

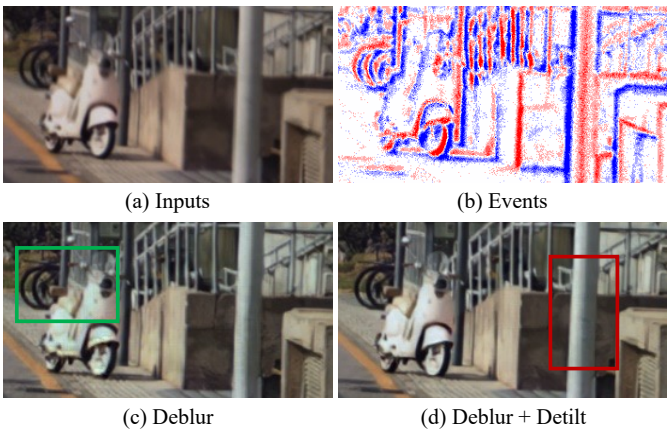


Fig. 8. An example illustrating the effectiveness of D-Net and T-Net. Starting from a turbulence-degraded image (a) and event signals (b), D-Net successfully removes the blur (c). When further combined with T-Net, the approach additionally corrects tilt distortions (d).

IV. Experiment

In this section, we present both qualitative and quantitative comparisons of our approach against state-of-the-art image-based turbulence removal methods on the proposed TurbEvent dataset (Section IV-A) as well as on real-captured data (Section IV-B). We further demonstrate the benefits of turbulence removal on a downstream vision task (Section IV-C), evaluate performance in dynamic scenes (Section IV-D), and assess computational efficiency (Section IV-E). Finally, ablation studies are conducted in Section IV-F to validate the contribution of each component in the proposed framework.

A. Evaluation on TurbEvent dataset

Dataset. For the training and testing data split in the TurbEvent dataset, we first leverage the provided image labels to categorize the data into multiple semantic groups. Within each group, the samples are then randomly partitioned into training and testing subsets using a 9:1 ratio. This stratified splitting strategy helps to preserve a similar distribution of content and

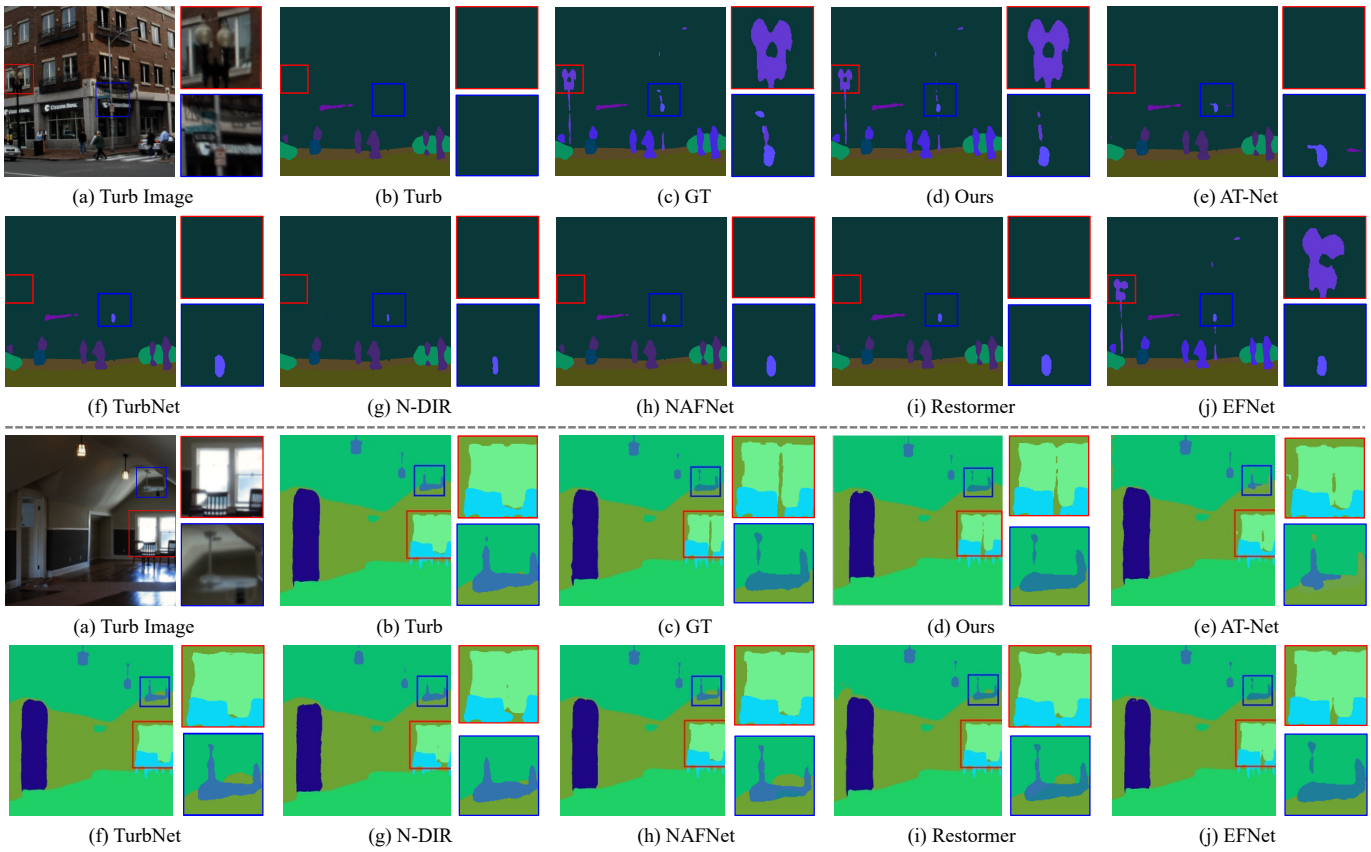


Fig. 9. Visual quality comparison for semantic segmentation on the TurbEvent dataset. (a) Turbulent images. (b) and (c) show segmentation results from the turbulent image and the ground truth image, respectively. (d)∼(j) present segmentation results from images restored using our method, AT-Net [3], TurbNet [1], N-DIR [52], NAFNet [46], Restormer [53], and EFNet [15]. Close-up views are provided on the right side of each image. Additional results are provided in the supplementary material.

turbulence conditions across both subsets, thereby reducing potential bias and ensuring fair evaluation. Following this procedure, we obtain a total of 5665 image–event pairs for training and 653 pairs for testing.

Comparison methods. We conduct extensive comparisons against seven representative baselines, including four turbulence removal approaches, two general-purpose image restoration methods, and one event-based deblurring model: AT-DDPM [24], AT-Net [3], TurbNet [1], N-DIR [52], NAFNet [46], Restormer [53], and EFNet [15]. In addition, we introduce an adapted variant of Restormer, denoted as Restormer*, which incorporates event data alongside image frames. To achieve this, we first encode the image and event streams separately using two independent convolutional layers, and subsequently concatenate the encoded features before forwarding them into the original Restormer architecture.

For a fair and consistent evaluation, all supervised methods [1], [3], [15], [46], [53] are re-trained on our TurbEvent dataset. Each model is trained for 150 epochs with a batch size of 4 on an NVIDIA RTX 3090 GPU. For the diffusion-based method [24], we adopt the default setting and fine-tune the pretrained model for 50K iterations on our dataset. To ensure uniformity across experiments, we adopt the same optimizer, initial learning rate, and scheduler configurations as in our proposed method, while preserving the original loss functions defined in the respective works. This setup eliminates confounding factors and isolates the intrinsic capability of each model in handling turbulence degradations.

The quality of restored images is quantitatively evaluated using three widely adopted metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). Together, these metrics comprehensively capture both pixel-level fidelity and perceptual quality, providing a rigorous benchmark for comparing different approaches.

Results. Quantitative comparisons are summarized in Table I, while qualitative results are presented in Figure 6. Across all three evaluation metrics, our method consistently achieves the highest performance. In particular, event-guided approaches significantly outperform purely image-based methods in terms of PSNR, underscoring the value of leveraging high-speed event information. The performance of Restormer* further indicates that a naïve concatenation strategy for fusing event data with RGB frames provides only marginal improvements, highlighting the need for more principled integration mechanisms such as those employed in our framework.

From the perspective of visual quality, our method not only reduces blur more effectively than single-image-based techniques but also reconstructs images with fewer artifacts and sharper structural details compared to the event-based

TABLE II
QUANTITATIVE COMPARISON OF THE SEMANTIC SEGMENTATION TASK ON RESTORED IMAGES FROM OUR TURBEVENT DATASET. FOR REFERENCE, WE ALSO EVALUATE THE PERFORMANCE OF TURBULENCE-DISTORTED IMAGES AND GROUND TRUTH IMAGES. THE BEST PERFORMANCES FOR RESTORED IMAGES ARE HIGHLIGHTED IN **BOLD**.

	Event	mIoU ↑	Dice ↑	FWIoU ↑
Turb		0.923	0.959	0.931
Ground Truth		0.941	0.969	0.948
AT-Net	✗	0.922	0.958	0.932
TurbNet	✗	0.926	0.961	0.936
N-DIR	✗	0.910	0.951	0.921
NAFNet	✗	0.923	0.959	0.936
Restormer	✗	0.928	0.961	0.937
EFNet	✓	0.930	0.963	0.939
Ours	✓	**0.936**	**0.966**	**0.943**

deblurring method EFNet [15]. For Restormer*, the improvement remains limited, with residual distortions still visible in challenging regions. Notably, as shown Figure 6, only our approach is capable of accurately correcting pixel displacements, leading to geometrically consistent reconstructions that align closely with the underlying scene content.

B. Evaluation on real-captured data

We captured real-world turbulence-degraded images using our hybrid camera system, with the results presented in Figure 7. As shown, our method consistently outperforms other state-of-the-art approaches on real-world data, restoring finer details and exhibiting superior generalization capability. In real-world scenarios, we also compared our method with EDI-based approaches. As the results indicate, directly applying the original EDI model fails to remove turbulence effects effectively.

Notably, the real-world videos are collected from outdoor scenes with unseen semantic categories and uncontrolled illumination conditions. They also contain mild natural atmospheric turbulence, which is further amplified by the heat chamber during capture. As a result, the real-world test data differ from the display-based training data in scene content, lighting conditions, and turbulence statistics, providing a cross-domain generalization evaluation of the proposed framework. We note that the sensing pipeline (cameras, optics, alignment, and synchronization) and the heat-chamber-induced turbulence-generation regime largely overlap with the training setup; the present evaluation therefore probes robustness to shifts in scene content, illumination, and turbulence statistics, rather than to fully out-of-distribution sensing or turbulence-formation conditions. Stronger out-of-distribution evaluation, e.g., across different cameras, optics, or fully natural turbulence without the heat chamber, is left as future work.

Despite the clear performance advantage on this cross-domain evaluation, a partial domain gap still persists between the display-based training data and real-world outdoor scenes. Specifically, display-based captures may exhibit spectral characteristics different from those of natural scenes, whereas real outdoor data involve additional complexities, including atmospheric scattering, illumination variation, and camera-dependent white-balance changes. Even under these challenging conditions, the proposed method still effectively restores the dominant structural distortions, although slight color temperature shifts can occasionally be observed in the restored results. We attribute this issue primarily to the mismatch in color distribution and white-balance statistics between the training data and the real-captured scenes.

Component analysis of D-Net and T-Net. To further validate the design of our framework, we separately examine the contributions of D-Net and T-Net, which are tailored for deblurring and detilting, respectively. In Figure 8, we first present the outputs obtained using D-Net alone, where it effectively removes turbulence-induced blur (highlighted in the green box). When D-Net is combined with T-Net, the full pipeline also addresses tilt distortions, leading to sharper and more geometrically consistent reconstructions (highlighted in the red box).

From an architectural perspective, D-Net employs residual learning with global connections, a strategy widely recognized in image restoration tasks for its ability to preserve structural consistency while enhancing fine textures. In contrast, T-Net is designed to estimate a tilt flow field, enabling it to explicitly model geometric pixel displacements introduced by turbulence. Together, these complementary modules provide a robust solution for simultaneously handling both blur and tilt degradations in real-world scenarios.

C. Evaluation on downstream task

To comprehensively evaluate the effect of turbulence removal on downstream applications, we further examine how the proposed method improves image-based semantic segmentation and optical character recognition (OCR). Semantic segmentation is adopted as a downstream task because it reflects the recovery of global scene structure and object boundaries, both of which are often degraded by atmospheric turbulence. OCR is additionally included to assess the restoration of fine-grained textual details.

Semantic segmentation. We employ the ADE20K dataset [48], [49], which provides high-quality ground-truth semantic annotations. As the segmentation backbone, we select InternImage-XL [57], a state-of-the-art model that leverages deformable convolutions to enable adaptive spatial aggregation and capture effective receptive fields.

We adopt three widely used metrics for quantitative evaluation: mean Intersection over Union (mIoU), Dice coefficient, and frequency-weighted Intersection over Union (FWIoU). The results, summarized in Table II, show that our method consistently achieves the largest improvements across all metrics when compared to other turbulence removal approaches, and in many cases approaches the performance obtained on clean ground-truth images. A visual comparison in Figure 9 further confirms these findings, where our method produces segmentation maps that are more faithful to the underlying scene structure.

It is worth noting that our training and testing splits differ from those of the original ADE20K benchmark. Therefore, the

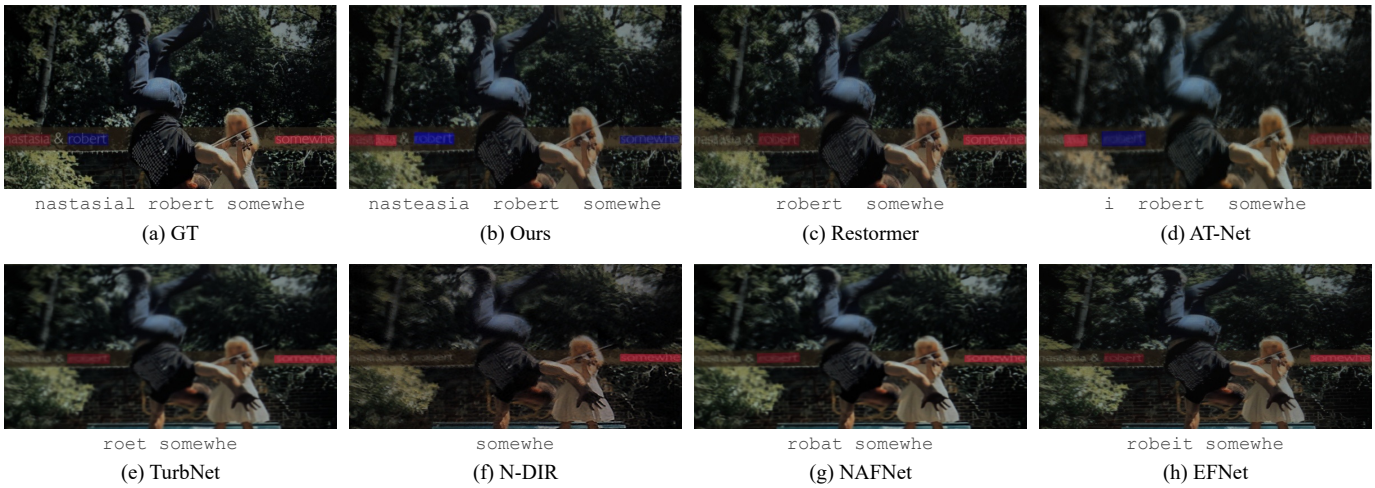


Fig. 10. Example images of OCR results. The colored blocks overlaid on the images indicate the words recognized, where different colors correspond to different recognized words. The recognized text strings are listed below the images. Notably, only our method successfully recognizes the first word in the image.

TABLE III
QUANTITATIVE COMPARISON OF OCR PERFORMANCE ON RESTORED IMAGES FROM THE TURBEVENT DATASET. WE USE THE OCR RESULTS OF CLEAN IMAGES AS THE REFERENCE.

Method	CER ↓	WER ↓
Turbulence	0.9797	0.975
AT-Net	0.9448	0.980
TurbNet	1.2047	0.970
N-DIR	0.9710	0.985
NAFNet	0.9866	0.975
Restormer	0.9260	0.960
EFNet	0.9440	0.950
Ours	**0.8921**	**0.945**

reported absolute values are not directly comparable to those from standard ADE20K evaluations. Nonetheless, the relative differences between restored and turbulence-degraded images clearly demonstrate that our method significantly enhances the effectiveness of downstream semantic segmentation, underscoring its practical utility in real-world vision applications.

OCR. We use the MMOCR toolkit [58], with DBNet [59] for text detection and CRNN [60] for recognition. We report Character Error Rate (CER) and Word Error Rate (WER), using the OCR results of the ground-truth clean images as references. Quantitative results are shown in Table III, and qualitative examples are provided in Figure 10. As shown by the results, our method achieves the best OCR performance among the compared methods.

D. Dynamic Scene

Event cameras capture scene textures with exceptionally high temporal resolution, enabling our method to process dynamic scenes in a frame-by-frame manner. Since Boehrer *et al.* [37] did not release their implementation, we instead compare our approach against the representative video-based turbulence removal method DATUM [5]. As illustrated in Figure 11, our framework achieves performance comparable to DATUM while operating on only a single-shot input, in contrast to DATUM, which requires 10 consecutive frames. This highlights the efficiency of our design, as our method reconstructs high-quality outputs by exploiting event information captured within a single image exposure, thereby avoiding the overhead and latency associated with multi-frame methods. Furthermore, this property makes our method particularly well-suited for real-world scenarios involving dynamic or rapidly changing scenes, where acquiring multiple stable frames is impractical.

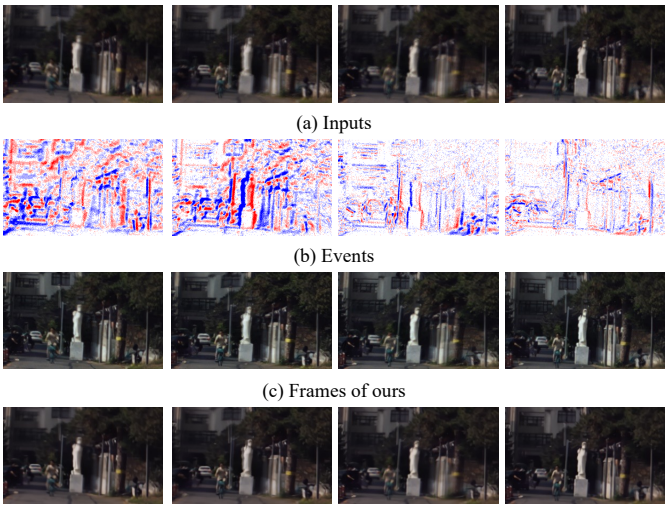


Fig. 11. Illustration of a dynamic scene. (a) Input frames. (b) Events. (c) Reconstructed frames using our method, which achieves high-quality results from a single-shot input. (d) Reconstructed frames using the video-based turbulence removal approach DATUM [5], which requires 10 input frames.

E. Network Efficiency

We report the multiply–accumulate operations (MACs), number of network parameters, and runtime of all comparative

TABLE IV
THE QUANTITY OF MULTIPLY-ACCUMULATE OPERATIONS (MACS) AND THE COUNT OF NETWORK PARAMETERS (#PARAM) ACROSS IMAGE-BASED AND EVENT-BASED METHODS

	MACs (G)	#Param (M)	Runtime (ms)
AT-Net	886.764	6.962	2861.2
TurbNet	687.872	26.57	174.80
N-DIR	232371	0.332	34445
NAFNet	85.3984	17.06	26.327
Restormer	753.969	26.10	149.93
EFNet	435.637	7.742	11.721
Ours	466.134	16.44	31.679

TABLE V
QUANTITATIVE RESULTS OF ABLATION STUDY

	PSNR ↑	SSIM ↑	LPIPS ↓
W/o EGDC	28.08	0.7957	0.2925
W/o Variance	28.42	0.8129	0.2384
W/o T-Net	27.70	0.7524	0.3352
W/o Flow	28.16	0.8088	0.2553
Our Full Model	29.67	0.8405	0.2010

methods in Table IV. For the self-supervised approach N-DIR [52], although it contains the fewest learnable parameters, it requires iterative optimization over each image for 1000 iterations, which leads to substantially higher MACs and significantly longer runtime. In contrast, our method achieves a more favorable trade-off between efficiency and model complexity. These results highlight that our framework not only attains state-of-the-art recovery performance but also maintains practical efficiency, making it more suitable for real-world applications where both accuracy and speed are critical.

F. Ablation studies

We conducted four ablation studies to evaluate the contributions of each module in our method, as summarized in Table V. Specifically, we first removed the EGDC blocks (denoted as "W/o EGDC") to investigate their role in data fusion between two modalities. Next, we removed the T-Net to assess its effectiveness in mitigating tilt distortion (denoted as "W/o T-Net"). Then, we excluded the variance map to directly evaluate its specific impact on tilt correction (denoted as "W/o Variance"). Finally, we remove the flow-based architecture for directly restoring images to determine its role in refinement (denoted as "W/o Flow"). The results demonstrate that our complete model achieves the better overall performance. The model struggles with multi-scale image event data integration without EGDC blocks. Without T-Net, the results exhibit significant tilt degradation.

Event statistic for tilt intensity estimation. The proposed framework uses the variance map of accumulated event stacks as a proxy for tilt intensity. To examine whether this design choice is essential, we compare variance with two representative alternatives: local event rate and temporal gradient magnitude computed from the event stacks. All statistics are extracted from the same accumulated event representation and integrated into the same tilt estimation pipeline for a fair comparison. The quantitative results are reported in Table VI. Among the compared statistics, variance consistently provides more stable and discriminative cues for tilt intensity estimation, resulting in superior restoration performance.

TABLE VI
ABLATION STUDY OF DIFFERENT STATISTICS USED AS PROXIES FOR TILT INTENSITY. WE COMPARE THE PROPOSED VARIANCE-BASED METRIC WITH TWO REPRESENTATIVE ALTERNATIVES: LOCAL EVENT RATE AND TEMPORAL GRADIENT MAGNITUDE.

Statistic	PSNR ↑	SSIM ↑	LPIPS ↓
Local Event Rate	28.61	0.8021	0.2255
Temporal Gradient	28.36	0.8036	0.2311
Variance (Ours)	**29.67**	**0.8405**	**0.2010**

V. CONCLUSION

In this paper, we propose EvTurb, a novel framework for atmospheric turbulence removal that leverages high-speed event streams. By using event-recorded intensity changes as an empirical surrogate for blur-related temporal accumulation and event-triggering frequency as a cue for tilt-related turbulence fluctuations, we effectively decouple blur and tilt distortions. We further perform turbulence removal through a structured two-step framework. Evaluations on our real-captured TurbEvent dataset demonstrate that our method significantly outperforms other methods while maintaining competitive runtime efficiency.

Limitations. A limitation of our approach stems from potential misalignment between the event camera and the frame camera. Although we employ precise calibration and a beam-splitting optical setup to alleviate this issue, minor misalignment may still persist, potentially affecting the accuracy of data fusion. Additionally, our implementation assumes identical resolutions for both cameras, simplifying processing but overlooking cross-resolution discrepancies commonly encountered in real-world applications. Furthermore, because event data inherently lack color information, their monochromatic nature leads to a domain discrepancy when fused with RGB images.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (Grant No. 62136001), Beijing Natural Science Foundation (Grant No. L233024), and Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012).

REFERENCES

[1] Z. Mao, A. Jaiswal, Z. Wang, and S. H. Chan, "Single frame atmospheric turbulence mitigation: A benchmark study and a new physics-inspired transformer model," in *Proc. of European Conference on Computer Vision*, 2022.

[2] L. Pan, C. Scheerlinck, X. Yu, R. Hartley, M. Liu, and Y. Dai, "Bringing a blurry frame alive at high frame-rate with an event camera," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[3] R. Yasarla and V. M. Patel, "Learning to restore images degraded by atmospheric turbulence using uncertainty," in *Proc. of International Conference on Image Processing*, 2021.

[4] X. Zhang, Z. Mao, N. Chimitt, and S. H. Chan, "Imaging through the atmosphere using turbulence mitigation transformer," *IEEE Transactions on Computational Imaging*, vol. 10, pp. 115–128, 2024.

[5] X. Zhang, N. Chimitt, Y. Chi, Z. Mao, and S. H. Chan, "Spatio-temporal turbulence mitigation: A translational perspective," in *Proc. of Computer Vision and Pattern Recognition*, 2024.

[6] D. L. Fried, "Probability of getting a lucky short-exposure image through turbulence," *Journal of the Optical Society of America*, vol. 68, no. 12, pp. 1651–1658, 1978.

[7] Y. Xia, C. Zhou, C. Zhu, M. Teng, C. Xu, and B. Shi, "NB-GTR: Narrow-band guided turbulence removal," in *Proc. of Computer Vision and Pattern Recognition*, 2024.

[8] M. J. Steinbock, "Implementation of branch-point-tolerant wavefront reconstructor for strong turbulence compensation," *Master's thesis (Air Force Institute of Technology)*, 2012.

[9] T. Serrano-Gotarredona and B. Linares-Barranco, "A 128 × 128 1.5% contrast sensitivity 0.9% fpn 3 µs latency 4 mw asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 3, pp. 827–838, 2013.

[10] S. Tulyakov, D. Gehrig, S. Georgoulis, J. Erbach, M. Gehrig, Y. Li, and D. Scaramuzza, "Time Lens: Event-based video frame interpolation," in *Proc. of Computer Vision and Pattern Recognition*, 2021.

[11] S. Tulyakov, A. Bochicchio, D. Gehrig, S. Georgoulis, Y. Li, and D. Scaramuzza, "Time Lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion," in *Proc. of Computer Vision and Pattern Recognition*, 2022.

[12] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 1964–1980, 2019.

[13] M. Teng, C. Zhou, H. Lou, and B. Shi, "NEST: Neural event stack for event-based image enhancement," in *Proc. of European Conference on Computer Vision*, 2022.

[14] H. Lou, M. Teng, Y. Yang, and B. Shi, "All-in-focus imaging from event focal stack," in *Proc. of Computer Vision and Pattern Recognition*, 2023.

[15] L. Sun, C. Sakaridis, J. Liang, Q. Jiang, K. Yang, P. Sun, Y. Ye, K. Wang, and L. V. Gool, "Event-based fusion for motion deblurring with cross-modal attention," in *Proc. of European Conference on Computer Vision*, 2022.

[16] J. Primot, G. Rousset, and J. C. Fontanella, "Deconvolution from wave-front sensing: a new technique for compensating turbulence-degraded images," *Journal of the Optical Society of America*, vol. 7, pp. 1598–1608, 1990.

[17] M. A. Vorontsov and G. W. Carhart, "Anisoplanatic imaging through turbulent media: image recovery by local information fusion from a set of short-exposure images," *Journal of the Optical Society of America*, vol. 18, no. 6, pp. 1312–1324, 2001.

[18] Z. Mao, N. Chimitt, and S. H. Chan, "Image reconstruction of static and dynamic scenes through anisoplanatic turbulence," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1415–1428, 2020.

[19] R. Yasarla and V. M. Patel, "CNN-based restoration of a single face image degraded by atmospheric turbulence," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 2, pp. 222–233, 2022.

[20] K. Mei and V. M. Patel, "LTT-GAN: Looking through turbulence by inverting GANs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 3, pp. 587–598, 2023.

[21] C. P. Lau, C. D. Castillo, and R. Chellappa, "AtfaceGAN: Single face semantic aware image restoration and recognition from atmospheric turbulence," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 240–251, 2021.

[22] S. N. Rai and C. V. Jawahar, "Removing atmospheric turbulence via deep adversarial learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 2633–2646, 2022.

[23] D. Jin, Y. Chen, Y. Lu, J. Chen, P. Wang, Z. Liu, S. Guo, and X. Bai, "Neutralizing the impact of atmospheric turbulence on complex scene imaging via deep learning," *Nature Machine Intelligence*, vol. 3, no. 10, pp. 876–884, Oct 2021.

[24] N. G. Nair, K. Mei, and V. M. Patel, "AT-DDPM: Restoring faces degraded by atmospheric turbulence using denoising diffusion probabilistic models," in *Proc. of Winter Conference on Applications of Computer Vision*, 2023.

[25] C. J. Pellizzari, M. F. Spencer, and C. A. Bouman, "Coherent plug-and-play: Digital holographic imaging through atmospheric turbulence using model-based iterative reconstruction and convolutional neural networks," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1607–1621, 2020.

[26] A. Jaiswal, X. Zhang, S. H. Chan, and Z. Wang, "Physics-driven turbulence image restoration with stochastic refinement," in *Proc. of International Conference on Computer Vision*, 2023.

[27] S. Xu, R. Sun, Y. Chang, S. Cao, X. Xiao, and L. Yan, "Long-range turbulence mitigation: A large-scale dataset and a coarse-to-fine framework," in *Proc. of European Conference on Computer Vision*, 2024.

[28] W. Jiang, V. Boominathan, and A. Veeraraghavan, "Nert: Implicit neural representations for unsupervised atmospheric turbulence mitigation," in *Proc. of Computer Vision and Pattern Recognition Workshops*, 2023.

[29] H. Cai, J. Chen, B. Feng, W. Jiang, M. Xie, K. Zhang, C. Fermuller, Y. Aloimonos, A. Veeraraghavan, and C. Metzler, "Temporally consistent atmospheric turbulence mitigation with neural representations," in *Adv. of Neural Information Processing Systems*, 2024.

[30] Y. Xia, C. Zhou, C. Zhu, C. Xu, and B. Shi, "PlaNet: Learning to mitigate atmospheric turbulence in planetary images," in *Proc. of AAAI Conference on Artificial Intelligence*, 2025.

[31] N. Chimitt, X. Zhang, Z. Mao, and S. H. Chan, "Real-time dense field phase-to-space simulation of imaging through atmospheric turbulence," *IEEE Transactions on Computational Imaging*, vol. 8, pp. 1159–1169, 2022.

[32] H. Chen, M. Teng, B. Shi, Y. Wang, and T. Huang, "A residual learning approach to deblur and generate high frame rate video with an event camera," *IEEE Transactions on Multimedia*, vol. 25, pp. 5826–5839, 2023.

[33] C. Zhou, M. Teng, J. Han, J. Liang, C. Xu, G. Cao, and B. Shi, "Deblurring low-light images with events," *International Journal of Computer Vision*, vol. 131, no. 5, pp. 1284–1298, 2023.

[34] L. Sun, C. Sakaridis, J. Liang, P. Sun, J. Cao, K. Zhang, Q. Jiang, K. Wang, and L. Van Gool, "Event-based frame interpolation with ad-hoc deblurring," in *Proc. of Computer Vision and Pattern Recognition*, 2023.

[35] W. Yang, J. Wu, J. Ma, L. Li, and G. Shi, "Motion deblurring via spatial-temporal collaboration of frames and events," in *Proc. of AAAI Conference on Artificial Intelligence*, 2024.

[36] W. Yu, J. Li, S. Zhang, and X. Ji, "Learning scale-aware spatio-temporal implicit representation for event-based motion deblurring," in *Proc. of International Conference on Machine Learning*, 2024.

[37] N. Boehrer, R. P. Nieuwenhuizen, and J. Dijk, "Turbulence mitigation in imagery including moving objects from a static event camera," *Optical Engineering*, vol. 60, no. 5, pp. 053 101–053 101, 2021.

[38] H. Li, R. Fan, J. Guan, W. Hao, L. Rui, T. Wu, Y. Wang, and L. Gu, "EGTM: Event-guided efficient turbulence mitigation," 2025.

[39] Y. Hu, S.-C. Liu, and T. Delbruck, "v2e: From video frames to realistic dvs events," in *Proc. of Computer Vision and Pattern Recognition*, 2021.

[40] T. R. Rimmele and J. Marino, "Solar adaptive optics," *Living Reviews in Solar Physics*, vol. 8, no. 1, p. 2, 2011.

[41] S. H. Chan, "Tilt-then-blur or blur-then-tilt? clarifying the atmospheric turbulence model," *IEEE Signal Processing Letters*, vol. 29, p. 1833–1837, 2022.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. of Computer Vision and Pattern Recognition*, 2017.

[43] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. of International Conference on Computer Vision*, 2017.

[44] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-time intermediate flow estimation for video frame interpolation," in *Proc. of European Conference on Computer Vision*, 2022.

[45] X. Zhou, P. Duan, Y. Ma, and B. Shi, "EvUnroll: Neuromorphic events based rolling shutter image correction," in *Proc. of Computer Vision and Pattern Recognition*, 2022.

[46] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," *Proc. of European Conference on Computer Vision*, 2022.

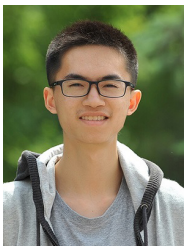
[47] P. Duan, Z. W. Wang, X. Zhou, Y. Ma, and B. Shi, "Eventzoom: Learning to denoise and super resolve neuromorphic events," in *Proc. of Computer Vision and Pattern Recognition*, 2021.

[48] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. of Computer Vision and Pattern Recognition*, 2017.

[49] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019.

[50] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. of Computer Vision and Pattern Recognition*, 2010.
[51] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.
[52] N. Li, S. Thapa, C. Whyte, A. W. Reed, S. Jayasuriya, and J. Ye, "Unsupervised non-rigid image distortion removal via grid deformation," in *Proc. of International Conference on Computer Vision*, 2021.
[53] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. of Computer Vision and Pattern Recognition*, 2022.
[54] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. of Computer Vision and Pattern Recognition*, 2018.
[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of International Conference on Learning Representations*, 2015.
[56] L. Wang, Y.-S. Ho, K.-J. Yoon *et al.*, "Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks," in *Proc. of Computer Vision and Pattern Recognition*, 2019.
[57] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li *et al.*, "InternImage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. of Computer Vision and Pattern Recognition*, 2023.
[58] M. D. Team, "Mmocr: A comprehensive toolbox for text detection, recognition and understanding," 2022.
[59] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization," 2020.
[60] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

Yixing Liu is currently pursuing the Master degree in Electrical and Computer Engineering at Rice University, Houston, Texas, USA. He received the B.S. degree from Peking University, Beijing, China, in 2025. He was an undergraduate researcher at the National Engineering Research Center of Video Technology, School of Computer Science, Peking University. His research interests include computational imaging, polarimetric imaging, and event-based vision.



Minggui Teng received the B.S. degree from Peking University, Beijing, China, in 2021. He is currently working toward the Ph.D. degree with the National Engineering Research Center of Video Technology, School of Computer Science, Peking University. His research interests lie in computational photography and generative AI, with a primary focus on neuromorphic cameras and video generation. He has served as a reviewer for CVPR, ICCV, ECCV, TPAMI, TIP, *etc*.

Yifei Xia is an M.S. student in the School of Computer Science at Peking University. He received the B.S. from Peking University in 2023. His research interests include computational photography and video generation.

Peiqi Duan is currently a Boya Postdoctoral Fellow at the School of Computer Science, Peking University. He received his PhD degree from Peking University in 2023. His research interests span event-based imaging and vision, single image super-resolution, and HDR image recon struction. He has served as a reviewer/program committee member for TPAMI, IJCV, TIP, TCSVT, CVPR, ICCV, ECCV, NeurIPS, ICLR, etc.

Boxin Shi (Senior Member, IEEE) received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Associate Professor (with tenure) and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper, Runners-Up at CVPR 2024, ICCP 2015 and selected as Best Paper candidate at ICCV 2015. He is an associate editor of TPAMI/IJCV and an area chair of CVPR/ICCV/ECCV. He is a senior member of IEEE.

EvTurb: Event Camera Guided Turbulence Removal

Yixing Liu†, Minggui Teng†, Yifei Xia, Peiqi Duan,
Boxin Shi‡, *Senior Member, IEEE*

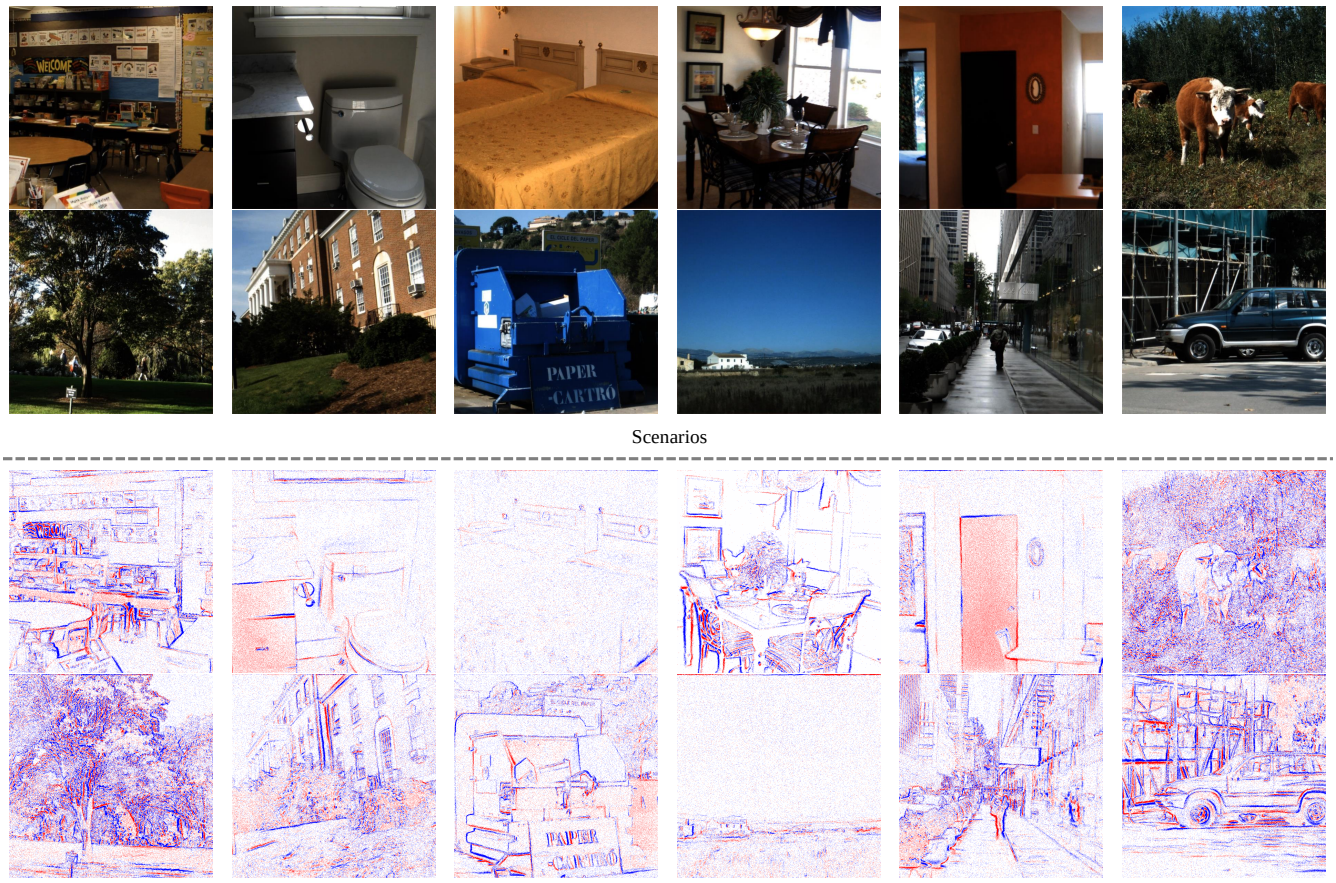


Fig. 12: Examples of scenes in our TurbEvent dataset.

VI. Details of TurbEvent dataset

In Figure 12, we present example samples from the TurbEvent dataset, including representative images and their paired event signals across diverse scenarios.

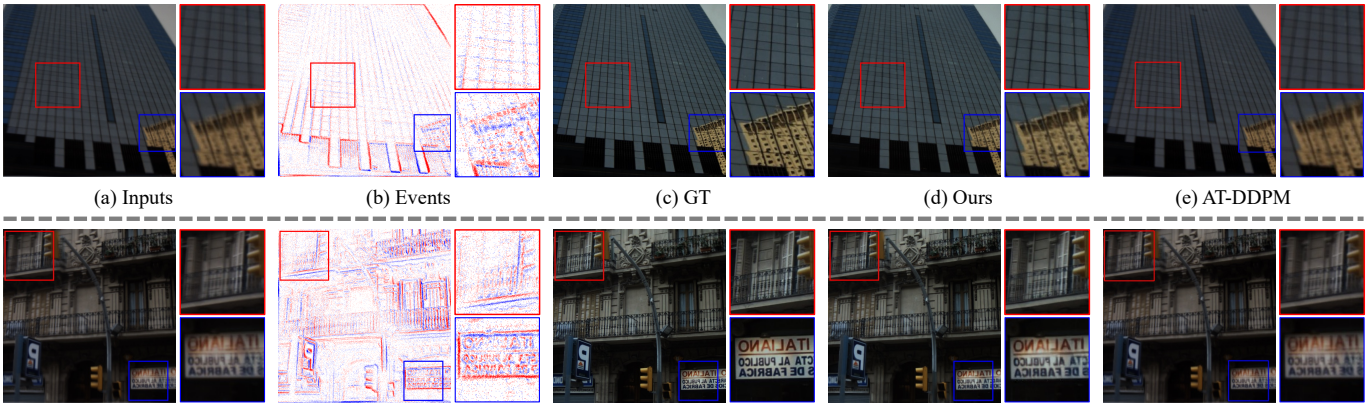
Because these samples are not evenly distributed in the original training and testing splits, we resample the data and construct new training and test sets at the image level within each semantic category. This protocol ensures that no identical scenes or image instances appear in both splits while preserving a balanced category distribution. Although some semantic categories are shared between the training and test sets, the corresponding scenes differ in spatial layout, viewpoint, object arrangement, and illumination conditions. As illustrated in Figure 13, even images from the same semantic category can exhibit substantial visual diversity.

VII. Analysis on Sensor Misalignment

Although the proposed framework performs pixel-level fusion between RGB images and event data, small residual misalignment may still arise in practical deployments due to calibration errors or imperfect synchronization. To evaluate the robustness of the method, we introduce synthetic spatial and temporal offsets between the RGB frames and the event stream. Specifically, spatial shifts ranging from 0 to 3 pixels are applied to the RGB frames relative to the event data, and temporal offsets ranging from 0 to 3 ms are introduced into the event stream. We then evaluate the restoration performance under these perturbations. As summarized in Table VII, the proposed method demonstrates strong robustness to moderate sensor misalignment. Even under a spatial shift of 3 pixels or a temporal offset of 3 ms, the performance drop remains limited, with less than 0.21 dB degradation in PSNR and less than 0.014 degradation in SSIM. These results indicate that the proposed fusion strategy is stable under realistic calibration and synchronization errors.

Test set Train set

Fig. 13: Example images from the same semantic category, **urban**, in the dataset. Although they belong to the same category, the scenes exhibit substantial diversity in layout, viewpoint, and lighting conditions.



(a) Inputs (b) Events (c) GT (d) Ours (e) AT-DDPM

(a) Inputs (b) Events (c) GT (d) Ours (e) AT-DDPM

Fig. 14: Qualitative comparison between AT-DDPM [1] and our method on TurbEvent Dataset.

TABLE VII: Robustness evaluation under synthetic sensor mis-calibration. Spatial shifts are introduced between RGB frames and event data, and temporal offsets are applied to the event stream. The reported ΔSSIM, ΔPSNR, and ΔLPIPS measure the performance degradation relative to the precisely aligned baseline. The proposed method maintains stable performance under moderate spatial and temporal misalignment.

Spatial	Temporal	ΔSSIM ↓	ΔPSNR ↓	Δ LPIPS ↑
$(0,0)$	1	−0.00631	−0.00498	0.00021
	2	−0.00843	−0.05674	0.00076
	3	−0.01314	−0.20246	0.00176
$(1,1)$	0	−0.00764	−0.05661	0.00009
	1	−0.00630	−0.00498	0.00021
	2	−0.00842	−0.05674	0.00076
	3	−0.01314	−0.20246	0.00176
$(2,2)$	0	−0.00764	−0.05662	0.00009
	1	−0.00630	−0.00500	0.00021
	2	−0.00842	−0.05674	0.00076
	3	−0.01314	−0.20246	0.00176
$(3,3)$	0	−0.00764	−0.05661	0.00009
	1	−0.00631	−0.00498	0.00021
	2	−0.00843	−0.05674	0.00076
	3	−0.01314	−0.20247	0.00176

VIII. More Results on TurbEvent Dataset

In this section, we first provide a qualitative comparison between AT-DDPM [1] and the proposed method in Figure 14. Besides, we provide additional qualitative comparisons between our method and AT-Net [2], TurbNet [3], N-DIR [4], NAFNet [5], Restormer [6], and EFNet [7] on the TurbEvent dataset, as shown in Figure 15 and Figure 16. The results demonstrate that our approach reconstructs images with fewer artifacts and sharper details.

IX. More Results on the Downstream Task

In this section, we provide additional qualitative comparisons of semantic segmentation performance among our method, AT-Net [2], TurbNet [3], N-DIR [4], NAFNet [5], Restormer [6], and EFNet [7] on the TurbEvent dataset, as shown in Figure 17. The results demonstrate that our method achieves more accurate semantic segmentation than other state-of-the-art methods.

REFERENCES

[1] N. G. Nair, K. Mei, and V. M. Patel, "AT-DDPM: Restoring faces degraded by atmospheric turbulence using denoising diffusion probabilistic models," in *Proc. of Winter Conference on Applications of Computer Vision*, 2023.

[2] R. Yasarla and V. M. Patel, "Learning to restore images degraded by atmospheric turbulence using uncertainty," in *Proc. of International Conference on Image Processing*, 2021.

[3] Z. Mao, A. Jaiswal, Z. Wang, and S. H. Chan, "Single frame atmospheric turbulence mitigation: A benchmark study and a new physics-inspired transformer model," in *Proc. of European Conference on Computer Vision*, 2022.

[4] N. Li, S. Thapa, C. Whyte, A. W. Reed, S. Jayasuriya, and J. Ye, "Unsupervised non-rigid image distortion removal via grid deformation," in *Proc. of International Conference on Computer Vision*, 2021.

[5] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," *Proc. of European Conference on Computer Vision*, 2022.

[6] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. of Computer Vision and Pattern Recognition*, 2022.

[7] L. Sun, C. Sakaridis, J. Liang, Q. Jiang, K. Yang, P. Sun, Y. Ye, K. Wang, and L. V. Gool, "Event-based fusion for motion deblurring with cross-modal attention," in *Proc. of European Conference on Computer Vision*, 2022.

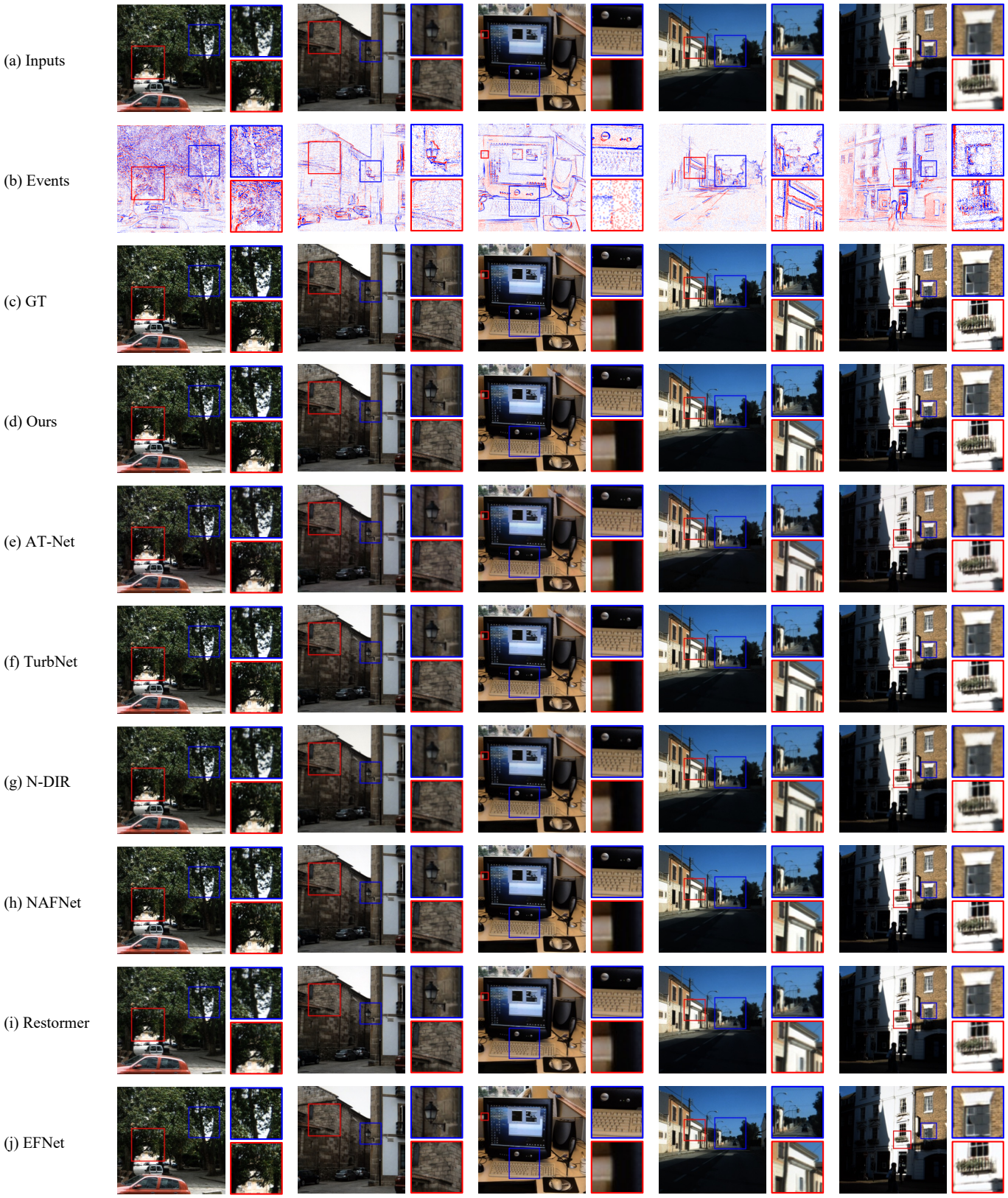


Fig. 15: Visual quality comparison with turbulence removal methods on the TurbEvent dataset (Part I). (a) Input images. (b) Events. (c) Ground truth. (d)∼(j) Results of ours, AT-Net [2], TurbNet [3], N-DIR [4], NAFNet [5], Restormer [6], and EFNet [7]. Close-up views are provided on the right side of each image.

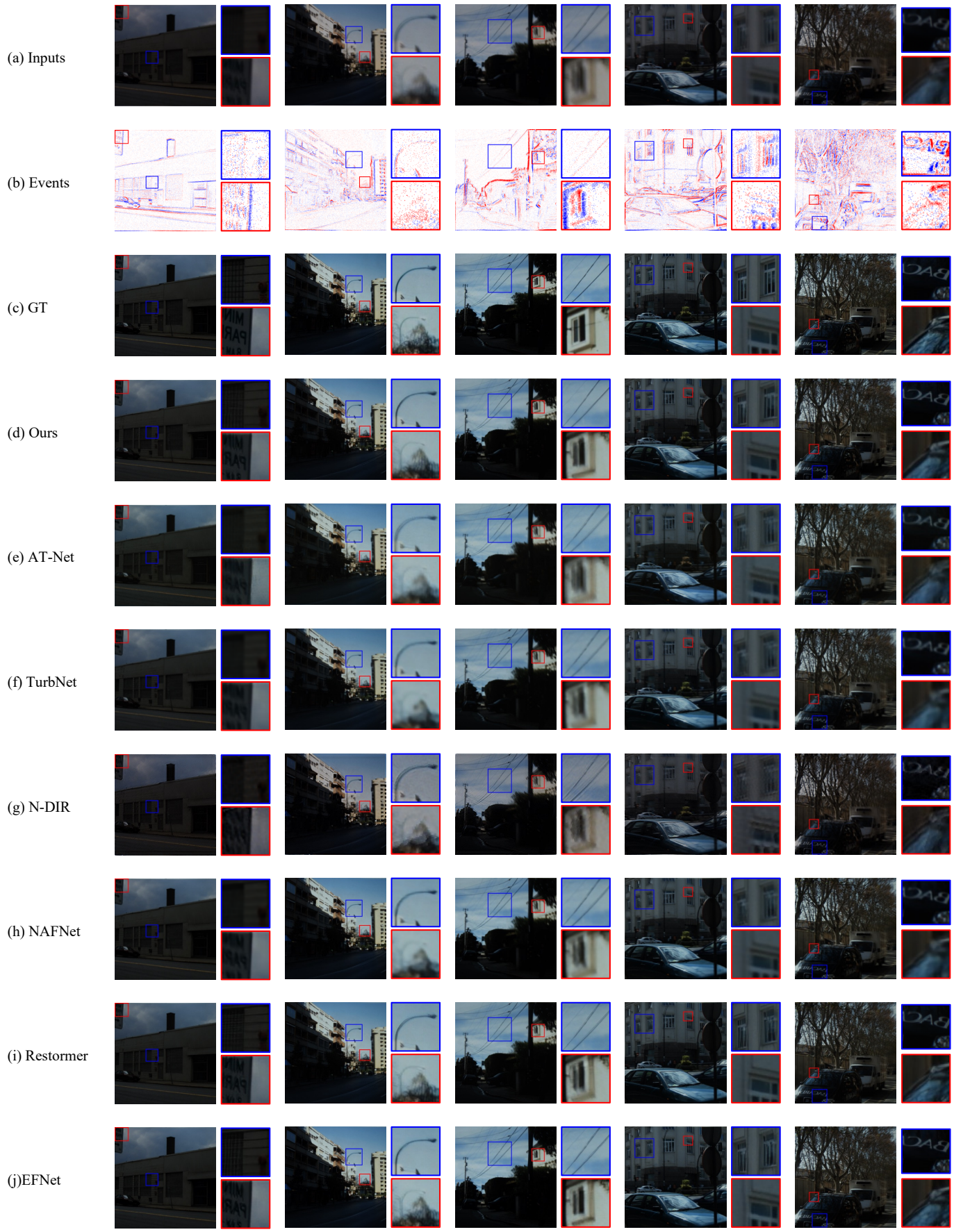


Fig. 16: Visual quality comparison with turbulence removal methods on the TurbEvent dataset (Part II). (a) Input images. (b) Events. (c) Ground truth. (d)∼(j) Results of ours, AT-Net [2], TurbNet [3], N-DIR [4], NAFNet [5], Restormer [6], and EFNet [7]. Close-up views are provided on the right side of each image.

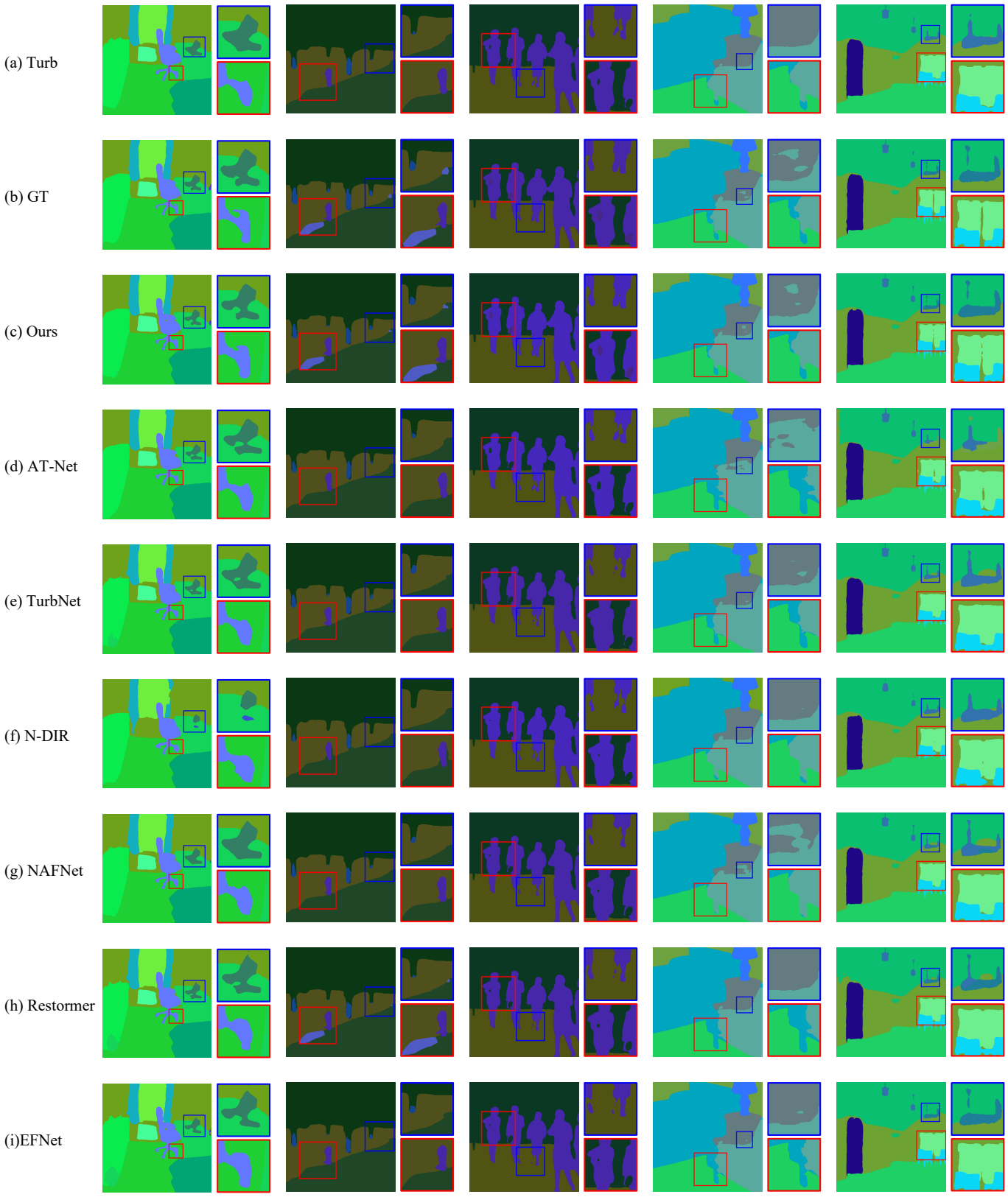


Fig. 17: Visual quality comparison for semantic segmentation on the TurbEvent dataset. (a) and (b) show segmentation results from the turbulent image and the ground truth image, respectively. (c)∼(i) present segmentation results from images restored using our method, AT-Net [2], TurbNet [3], N-DIR [4], NAFNet [5], Restormer [6], and EFNet [7]. Close-up views are provided on the right side of each image.