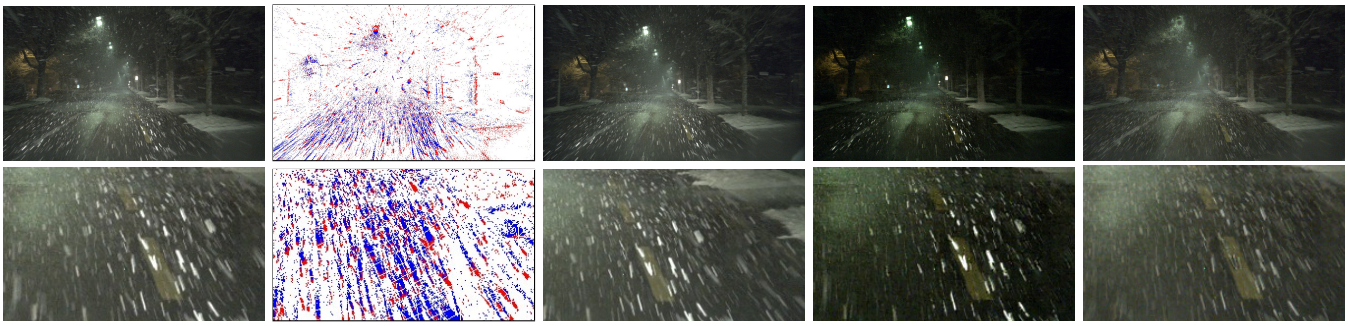
This paper has been accepted for publication at the
IEEE Transactions on Robotics, 2026. © IEEE

Event-Based De-Snowing for Autonomous Driving

Manasi Muglikar Nico Messikommer Marco Cannici Davide Scaramuzza



(a) Snowfall image (b) Events (c) Restormer (d) Snowformer (e) **Ours**

Fig. 1: **Event-based video snow removal in challenging nighttime scenes**: (a) Sample image from our dataset captured while driving in snowfall, and (b) corresponding event data highlighting the motion of snowflakes. The presence of dense, dynamic snow and low visibility presents significant challenges for conventional image restoration methods, as illustrated by the results of (c) Restormer [1] and (d) Snowformer [2]. Our proposed event-based approach (e) utilizes event information to address these challenges and restore clearer scene content under adverse weather conditions.

***Abstract*—Adverse weather conditions, particularly heavy snowfall, pose significant challenges to both human drivers and autonomous vehicles. Traditional image-based desnowing methods often introduce hallucination artifacts as they rely solely on spatial information, while video-based approaches require high frame rates and suffer from alignment artifacts at lower frame rates. Camera parameters, such as exposure time, also influence the appearance of snowflakes, making the problem difficult to solve and heavily dependent on network generalization. In this paper, we propose to address the challenge of desnowing by using event cameras, which offer compressed visual information with submillisecond latency, making them ideal for desnowing images, even in the presence of ego-motion. Our method leverages the fact that snowflake occlusions appear with a very distinctive streak signature in the spatiotemporal representation of event data. We design an attention-based module that focuses on events along these streaks to determine when a background point was occluded and use this information to recover its original intensity. We benchmark our method on DSEC-Snow, a new dataset created using a green-screen technique that overlays pre-recorded snowfall data onto the existing DSEC driving dataset, resulting in precise ground truth and synchronized image and event streams. Our approach outperforms state-of-the-art desnowing methods by 3 dB in PSNR for image reconstruction. Moreover, we show that off-the-shelf computer vision algorithms can be applied to our reconstructions for tasks such as depth estimation and optical flow, achieving a 20% performance improvement over other desnowing methods. Our work represents a crucial step towards enhancing the reliability and safety of vision systems in challenging winter conditions, paving the way for more robust, all-weather-capable applications.**

Code, dataset and video are available under: https://rpg.ifi.uzh.ch/evsnow.html

This work was supported by the European Research Council (ERC) under grant agreement No. 864042 (AGILEFLIGHT). The authors are with the Robotics and Perception Group, Department of Informatics, University of Zurich, Switzerland.

I. Introduction

Imagine driving through heavy snowfall. Bright, swirling flakes, windshield accumulation, and reduced visibility severely degrade scene perception, making driving not only unpleasant but also potentially dangerous. Autonomous vehicles and assistive driving systems, designed to enhance road safety, also struggle in these conditions as snowflakes obscure both camera and LiDAR sensors. To advance vehicle autonomy and automotive safety, addressing adverse weather challenges is crucial. Thus, effective desnowing techniques are necessary to ensure the reliability and safety of vision systems in snowy environments.

Existing solutions, such as training a network on specific desnowing datasets [3], [4], [5], [6], [7], [8] or using Gated Cameras [3], fall short—either failing to generalize across snow conditions or losing intensity information in low light. In comparison to a single image, a video provides richer temporal context about the dynamic features of the scene. Building on this, existing works [3], [8], [9], [10], [11], [12], [13], [14] have studied the effect of incorporating this temporal information for image desnowing, showing promising results. However, the performance of video desnowing relies on the framerate of the camera. A high framerate ensures reliable alignment across individual frames, which can be used to align the background and remove snow occlusion. For effective image desnowing, it is crucial to capture high-speed scene information. We, therefore, propose to solve this problem by using event cameras, which provide extremely high temporal resolution (on the order of 1 MHz), without a huge bandwidth demand.

Event cameras measure intensity changes with very low

latency (up to $1\,\mu s$) and asynchronously [15]. This produces a stream of events that encode the time, location, and polarity of the brightness change. The main advantages of event cameras include sub-millisecond latency, very high dynamic range ($> 120\,\mathrm{dB}$), and strong robustness to motion blur. This work leverages these unique properties to remove snow occlusions from images effectively.

Since event cameras have high temporal resolution and no exposure time, snowflakes always appear as streaks in the space-time representation of events (c.f. Fig. 2). This is unlike conventional cameras, where the exposure time plays a major role in the appearance of the snowflake. Thus, our method takes advantage of this unique signature in the space-time domain to track these streaks. In the absence of ego-motion, the task of recovering background intensity is quite simple, as one only has to look at the intensity changes (or events) per pixel. However, with ego-motion, the point corresponding to the pixel that is occluded at a certain time may not be the same at a future time. Therefore, keeping track only along the pixel will result in significant errors. Instead, we propose to look along the streak, as a point in 3D space will be occluded and de-occluded along this streak. To recover the background intensity where the occlusion covers the background, it is necessary to look along the streak in time. By focusing only on points along these streaks, we can accurately determine when a point in the scene was occluded. Using this information, we reconstruct the background intensity, enabling us to recover a clear image even under challenging weather conditions. Our approach improves the performance of image-based and video-based approaches by over $3\,\mathrm{dB}$ in terms of PSNR of image reconstruction. Not only that, we also outperform image-based approaches in further downstream tasks such as depth estimation, optical flow, and object detection by over 20% in terms of accuracy.

We list our contributions as follows:

- **A new event-based desnowing dataset generated by chroma composition method**: We develop a synthetic dataset (DSEC-Snow) by overlaying recorded snowfall onto the DSEC [16] driving dataset using green-screen technology, providing precise ground truth and synchronized image-event streams without the need for complex snow rendering.
- **A novel approach for desnowing images using event cameras**: Our approach utilizes the rich temporal information of events to detect and remove snowflake occlusions by tracking their spatiotemporal streaks, enabling accurate background recovery. We also propose a simple learning framework to fuse the event and image information for improved desnowing performance.
- **A real-world driving dataset collected in snowy conditions**: We present a real-world event-camera dataset captured while driving in snowfall, offering valuable data for testing the robustness of event-based desnowing methods in practical scenarios.

II. Related work

We summarize the related works for image and video desnowing in Section II-A. There has been some progress in the context of image deraining with event cameras, which is summarized in Section II-B. Lastly, we also summarize the existing datasets published so far for image desnowing in Section II-C.

Dataset	Simulation	Sensors	GT
Snow100K [4]	Yes	Image	Yes
SnowCityScapes [5]	Yes	Image	Yes
SnowKITTI2012 [5]	Yes	Image	Yes
SRRS [6]	Yes	Image	Yes
CSD [7]	Yes	Image	Yes
SnowVideo [8]	Yes	Video	Yes
ESnowBR [17]	No	Events	No
DSEC-Snow (ours)	Yes	Video + Events	Yes
SnowDriving (ours)	No	Video + Events	No

TABLE I: **Comparison of snow datasets.** Existing snow datasets primarily focus on synthetic image-based data with ground truth (GT), while our DSEC-Snow and SnowDriving datasets provide real and simulated video sequences with event data, addressing the need for benchmarks suitable for event-based snow removal under more realistic conditions

A. Image and video desnowing

Estimating the background image in the presence of rain or snow is challenging as it requires hallucination of the background content. Some earlier works focused on modeling the rain streaks and snowflakes and decomposed the image into background and foreground components [18], [19]. However, even if the perfect model of the rain streaks or snowflakes is known, the background content is still challenging to estimate [20]. To address this issue, recent works [4], [6] have proposed using deep learning methods to estimate the background content. [5] proposed learning a representation for snow using the geometric and semantic properties of snow. On the other hand, Snowformer [2] proposes using a vision transformer that fully combines local and global information and obtains state-of-the-art results. Other image-based approaches focus more on image restoration [1] or image deraining [21], both of which do not generalize well for image desnowing [8]. Recent work [21] tackles the unique challenges of nighttime deraining—caused by non-uniform local illumination and complex rain-light interactions—by introducing a Rain Location Prior (RLP) learned via a recurrent residual model, along with a Rain Prior Injection Module (RPIM) to enhance feature representation and boost deraining performance at night. Instead of handling each image restoration task separately, Restormer [1] proposes a unified framework for image restoration tasks, including image deraining, deblurring, and denoising. However, the performance of a generalized image restoration method is often limited for specialized tasks such as image desnowing. Therefore, Snowformer [2] proposes a dedicated desnowing vision transformer with a scale-aware snow query and local-patch embedding, resulting in state-of-the-art results for image desnowing. These methods, however, rely on the spatial information available in the image to estimate the background content. In complex weather conditions, such as heavy snowfall, the spatial information is often not sufficient to accurately estimate the background content. Recently, Sun et

al. [22] proposed a histogram transformer network that leverages histogram self-attention and a dynamic-range convolution to efficiently captures long-range dependencies across similar-intensity regions, achieving superior restoration performance compared to existing methods.

Increasing temporal resolution for this task gives rise to video-based desnowing methods. Video-based desnowing methods [9] have explored the use of temporal information to improve the quality of desnowed images. The seminal work of Garg et al. [9] proposed a model-based approach to remove occlusions such as rain by characterizing the photometric and temporal properties of rain streaks. Subsequent works [23], [24], [25], [26], [27] have extended this work by proposing various priors to model the rain streaks and snowflakes. Another line of approach used matrix factorization to encode the correlation of background video along the temporal dimension [14], [28], [29], [30].

More recently, deep learning methods have also shown significant improvements in video desnowing [8], [31], [32], [33], [34], [35], [36], [37]. Li et al. [31] proposed a multi-scale convolutional neural network (CNN) for sparse coding to encode and remove the repetitive local patterns of rain streaks at different scales. Liu et al. [33] proposed a recurrent neural network (RNN) to turn this problem into classification of rain pixels and then recover the background. While these methods have shown promising results on synthetic data, they often struggle with complex weather conditions and fail to generalize to real-world scenarios. To address the domain gap between synthetic and real rain data, recent work [37] proposes a semi-supervised video deraining method that leverages a deep-learning-based dynamical rain generator and Monte Carlo EM optimization, jointly exploiting both labeled synthetic and unlabeled real videos for improved performance in real-world scenarios. Recently, Chen et al. [8] addressed the challenging task of video snow removal by introducing a high-quality dataset that simulates realistic snow and haze through advanced rendering and augmentation techniques.

B. Event-based image deraining and desnowing

Event cameras, known for their high temporal resolution and robustness to motion blur, have been increasingly utilized to enhance traditional imaging systems [38]. This high temporal resolution of events was used for de-occluding frames [39]. This technique utilizes the asynchronous nature of event cameras to provide additional temporal information, enabling more effective de-occlusion of images in real-time scenarios. Similarly, [40] introduced an unsupervised video deraining method that combines event data with traditional video frames. By integrating these two data types, the method effectively removes rain streaks and other occlusions from video sequences, demonstrating significant improvements in visibility and clarity.

Compared to image deraining, there has been limited work on event-based image desnowing. To the best of our knowledge, the only existing work is [17], which proposes a model-based approach for identifying and removing snow occlusions from events. The method relies on statistical modeling of snowflake events to partition event streams into snowflake and background events. While this method shows promising results, it is limited by the assumptions made in the model (e.g. thin structures are often misclassified as snowflakes instead of background) and does not leverage the complementary information available from intensity images, thereby limiting its performance in complex scenarios. Additionally, for complex tasks such as image reconstruction, the lack of intensity information can lead to loss of fine details and color in the reconstructed images. In this paper, we propose a learning-based approach that leverages both event and intensity information to effectively remove snow occlusions and reconstruct high-quality background images in color.

C. Snow Datasets

While there has been significant progress in the field of image and video desnowing, the availability of datasets for training and evaluating these methods remains limited. Some existing datasets focus on image desnowing, such as Snow100K [4], which provides a large collection of synthetic images with snow occlusions generated using Photoshop rendering techniques [41], providing ground truth images. In SnowKITTI2012 and SnowCityScapes [5], the authors used a similar approach to generate different densities of snow occlusions on images. SRRS [6], [7] improved these models and included more realistic rendering of snowy scenes by introducing a veiling effect and then using Photoshop to render snowflakes. Since these datasets rely on Photoshop, the realism of the snow occlusions is lacking, and therefore causes limited generalization to real-world desnowing of images. Therefore, [8] proposed a new video desnowing dataset which is rendered using Unreal Engine, providing a more realistic simulation of snow occlusions. These datasets, however, do not provide events. Recently, [17] proposed a dataset for event-based image desnowing by recording snowfall using an event camera and manually annotating objects in the scene. This dataset, however, does not provide images and ground truth background images, limiting its use for training learning-based desnowing methods. A comparison of these datasets is shown in Table I.

To the best of our knowledge, there currently exists no dataset that provides events and images for the task of desnowing. We therefore propose two datasets for this task, namely the DSEC-Snow dataset and Slider-Snow Dataset. The DSEC-Snow dataset overlays foreground data of snowfall recorded with an event camera onto a background consisting of the driving dataset DSEC [16]. This process removes the complexity of rendering snow particles that are both photo-realistic and physically accurate. Another advantage this dataset provides is the availability of synchronized events, images, and ground truth images. In addition to this, we also record real data of driving in snowfall using a color DAVIS event camera, called Slider-Snow .

III. Understanding the effect of snow on images and events

The appearance of snow in images and events is quite different due to the nature of the sensors. Garg et al. [9]

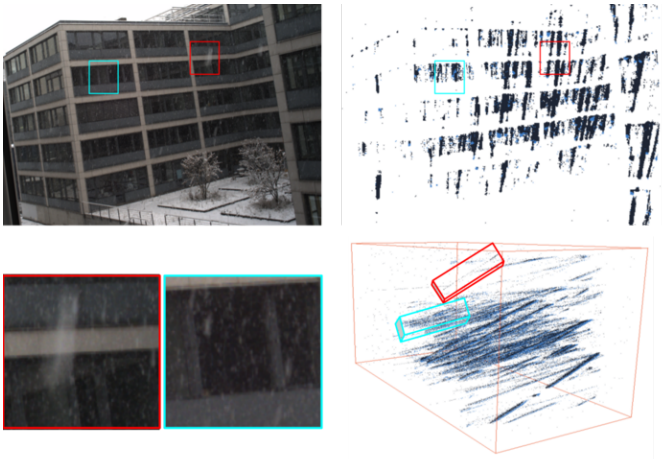


Fig. 2: **Effect of snow on images (left) and events (right)** (Left) The appearance of snowfall in an image depends on multiple factors, such as snowflake size, density, ambient illumination, and camera exposure settings. Two such examples are shown within the red and blue regions. (Right) The same snow occlusions in event data are visualized on the image plane and x-y-t volume (bottom-right). Irrespective of the snowflake size, snow occlusions have a unique spatio-temporal pattern in the form of streaks (bottom-right).

Algorithm 1 Background Intensity Estimation under Snow Occlusion Using Event Data

Input: Polarity event stream $p(t)$ at pixel X; reference intensity I_r; contrast threshold C; time window τ; [optional] warping function $W(\cdot)$

Output: Estimated background intensity I_b at pixel X

$I_b \leftarrow I_r$
if background is static **then**
    Accumulate events: $E \leftarrow \int_0^\tau p(t)\,dt$
    Update background intensity: $I_b \leftarrow I_r - C \times E$
else
    Identify warping $W(\cdot)$ along motion streaks (e.g., using velocity prior)
    Accumulate warped events: $E_w \leftarrow \int_0^\tau W(p(t) \times C)\,dt$
    Update background intensity: $I_b \leftarrow I_r - E_w$
end if
return I_b

analyzed snowflakes and showed that their appearance depends not only on the size, shape, and distance from the camera, but also on the ambient illumination and camera exposure settings. This results in a wide variety of snowflake appearances in images, as shown in Fig. 2 (top). For example, [9] provided a mathematical model of how snow can appear either as bright spots, streaks, or even haze depending on the distance of the snowflake from the camera, for the same exposure settings. This makes the challenge of identifying snowflakes in images quite significant.

On the other hand, snow occlusions in event data have a unique spatio-temporal pattern in the form of streaks, as shown in Fig. 2 (bottom). This is because of the high temporal resolution of the event camera, which captures the snow occlusions as streaks in the x-y-t volume. This makes the task of identifying snow occlusions in events much easier compared to images. The streaks in the x-y-t volume are independent of the distance and only depend on the relative difference between background intensity and snowflake intensity. For example, in Fig. 2 (bottom, left), the snow streak corresponding to the same snowflake is visible in front of a darker background like the window, but not visible in front of a brighter background like the sky. This makes the task of identifying snow occlusions in events much easier compared to images.

A. De-occluding with events

We now describe a geometric way to solve for background extraction using events. This algorithm is also summarized in Algo. 1. As a snowflake moves in front of the background, it causes an intensity change, resulting in events being triggered. The intensity at a pixel $\mathbf{X}$ at time t caused by a snowflake of intensity I_r occluding a background intensity of I_b can be written as:

$$\Delta I(t) = I_r - I_b = pC \tag{1}$$

In general, since the snowflake is usually brighter than the background, the intensity change is positive when the snowflake is occluding the background, triggering positive polarity events. When the snowflake moves away from the background, the intensity change is negative, triggering negative polarity events. Therefore, assuming we know the intensity of the snowflake, estimating the background intensity is simply a matter of integrating the events over time τ, scaling it by the contrast threshold C, and subtracting it from the snowflake intensity. Thus, the background intensity can be estimated as:

$$I_b = I_r - \sum_0^\tau pC \tag{2}$$

Therefore, once we have identified the snow occlusions (using events), we can recover the background intensity from the equation above and attend to events that occur at the same pixel at different times. In the presence of background motion, Equation. 2 no longer holds true, as the pixel before and after the occlusion can belong to a different point. This implies that the pixel at location X_0 being occluded at time t_0 could be at location X_i when it is de-occluded. Therefore, the search for the corresponding de-occluded pixel is no longer only along the time dimension but also along the spatial dimension. What does remain true, however, is that the equation still holds in 3D space, and the above equation is modified as follows:

$$I_b = I_r - \sum_0^\tau W(pC) \tag{3}$$

where W represents the warping of events along the motion streak. However, estimating this warp is quite challenging, especially when the snow occlusions are dense and overlapping. We follow the approach of [9], [42] to identify the streaks using the rain-velocity prior and use this to recover the background intensity. The main difference with respect to [9], [42] is that we use events instead of a set of images as

input to this method. We take events in a short time window (e.g., 5 ms) to ensure that the motion of the background is minimal, and therefore the warping can be approximated using a constant velocity model. Thus, the dominant motion is because of the snowflakes and can be estimated by approach proposed in [9], [42]. An event pixel (which we assume is a snow event), will move by $v \times t$ where v is velocity and t is time. In the spatial neighborhood of this event, we search for other events that fit this motion model. This can be done by iterating over a set of velocity hypotheses and for each hypothesis, and selecting the hypothesis which results in the maximum number of events being aligned along the motion streak. However, it can be computationally expensive to search over all possible velocities. Therefore, we fix the speed and only change the direction. Instead of a dense velocity grid, pick a small number of directions (8 uniformly spaced around a circle). For each direction, we project local events along that direction and measure alignment by calculating event density along the line.

This forms our model-based approach to de-occlude the snow occlusions in images using events. As this approach requires several assumptions about the snow occlusions, we also propose a data-driven approach to learn to de-occlude the snow occlusions using events, which we describe in the following sections.

IV. Data-driven Desnowing with Event Cameras

In this section, we present our approach for desnowing images using event cameras. Similar to the model-based approach, we model the observed image $I_{input}(x)$ as a combination of the clean background image $I_{clean}(x)$ and the occlusion caused by snowflakes $I_{occl}(x)$. However, instead of explicitly modeling the occlusion using geometric priors, we aim to learn a mapping from the input image and event data to the clean image using a neural network. The design of each component is inspired by classical geometric approaches but replaces explicit modeling with learnable modules. An overview of our proposed method is shown in Fig. 3.

Problem formulation: The event stream $E_{input} = \{e_i\}_{i=1}^{N}$ consists of asynchronous events $e_i = (x_i, y_i, t_i, p_i)$, where (x_i, y_i) represents the spatial coordinates, t_i is the timestamp, and $p_i \in \{-1, +1\}$ indicates the polarity of the brightness change. The events are accumulated over the exposure time of the camera to form a voxel grid representation $V \in \mathbb{R}^{H \times W \times B}$, where H and W are the height and width of the image, and B is the number of temporal bins. Given an input image $I_{input} \in \mathbb{R}^{H \times W \times 3}$ and the corresponding event voxel grid V, we aim to reconstruct the clean image $I_{pred} \in \mathbb{R}^{H \times W \times 3}$ where I_{pred} is the reconstructed clean image. Our approach consists of three main components: (1)**EventNet**, which processes the event stream to extract spatio-temporal patterns of snowflake streaks; (2) **Image Reconstruction**, which fuses the image and event features using a transformer; and reconstructs the clean image by combining the network prediction and the input image guided by a learned mask.

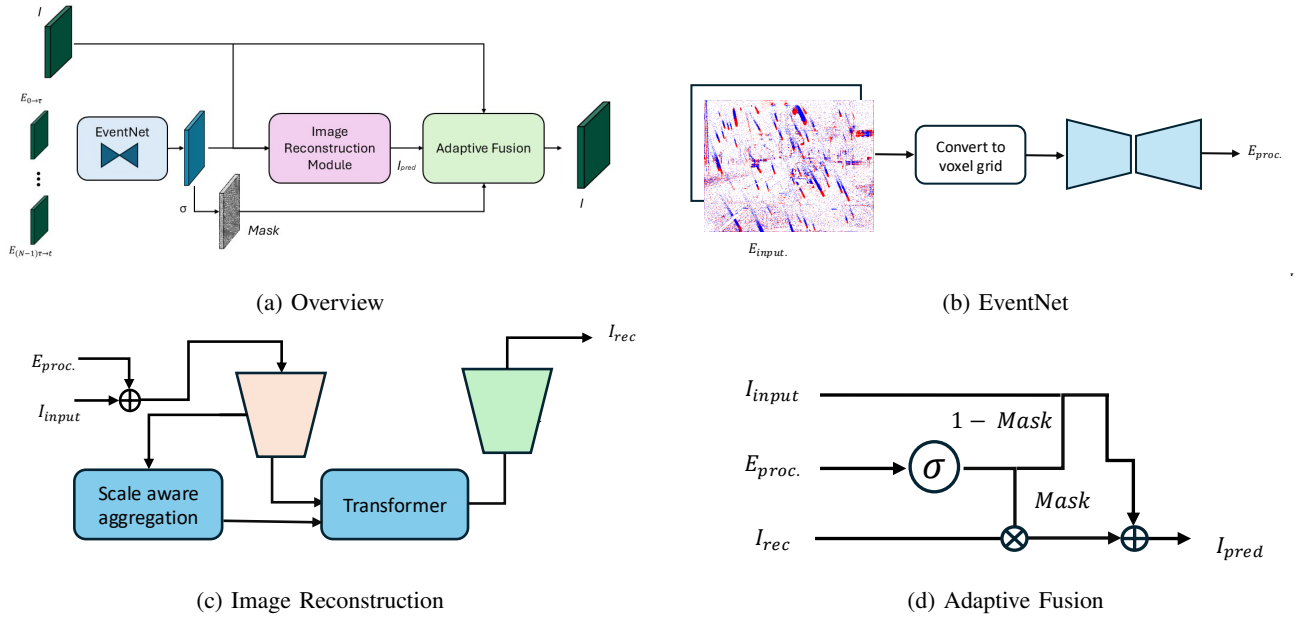
EventNet: The EventNet module is designed to extract spatio-temporal features from the raw event data, which encode the motion and geometry of snowflake occlusions. The events are first converted to a voxel grid representation. Instead of relying on explicit warping or velocity priors as in classical geometric models, EventNet employs a convolutional LSTM (ConvLSTM) to capture temporal dependencies, followed by a U-Net architecture for hierarchical spatial feature extraction. This structure enables the network to implicitly learn the geometric properties of snow streaks from data. The output of EventNet is a feature map that encodes the spatio-temporal patterns of snowflake occlusions E_{proc}, as shown in Fig. 3b.

Image Reconstruction: The image reconstruction module uses a hierarchical Transformer-based architecture proposed in Snowformer [2]. This Transformer fuses features from both the intensity image and the processed event stream. It comprises a U-Net-style encoder-decoder structure built upon Transformer blocks, which are capable of capturing long-range dependencies and aggregating contextual information across multiple scales, as shown in Fig. 3c. Channel attention mechanisms and context interaction layers further enhance feature fusion at each scale. The image and event features are concatenated and passed through the Transformer backbone to produce a de-snowed reconstruction I_{rec} It replaces hand-crafted priors and explicit motion modeling with multi-head self-attention and hierarchical feature aggregation, allowing the network to learn complex, non-linear dependencies across both spatial and temporal domains. The Transformer's ability to capture global context serves as a learnable analog to geometric reasoning over motion and occlusion structure, integrating information from both local neighborhoods and the entire image.

The next stage of image reconstruction combines the Transformer's prediction with the original input image using the spatial mask produced by EventNet, as detailed below. **Mask Prediction and Adaptive Fusion**: In the final stage, we perform an adaptive, pixel-wise fusion between the input image and the network's de-snowed prediction. The fusion is controlled by the learned spatial mask generated by applying a sigmoid activation to EventNet's output feature map, which indicates the likelihood of occlusion at each pixel. Since the mask is learned from data, it serves as a soft, adaptive counterpart to the explicit occlusion identification in geometry-based approaches.

$$I_{pred} = mask \odot I_{rec} + (1 - mask) \odot I_{input}, \qquad (4)$$

where I_{rec} is the output of the Transformer backbone, I_{input} is the original input image, and $mask$ is the spatial mask produced from EventNet. The operator $\odot$ denotes element-wise multiplication. This formulation can be interpreted as a data-driven generalization of the subtraction operation in geometry-based approaches, where the mask adaptively determines the contribution of the predicted clean image and the original input. Rather than explicitly subtracting a warped occlusion estimate(as done in model-based approach), the network learns the optimal blending strategy for each pixel, guided by the inferred occlusion geometry from the event data.



(a) Overview

(b) EventNet

(c) Image Reconstruction

(d) Adaptive Fusion

Fig. 3: **Overview of our data-driven method for reconstructing deoccluded images using event-camera data.** Events within the camera's exposure time are segmented into non-overlapping spatio-temporal windows, converted into voxel-grid representations, and processed alongside RGB images through modality-specific feature extraction. Event features are extracted via EventNet, which also produces a spatial mask. The reconstruction module fuses event and image features to reconstruct the image, and adaptive fusion uses a learned mask to blend this reconstruction with the original image.

A. Loss Function

For training supervision, we adopt the **L1 loss** as our primary reconstruction loss. The loss function is defined as:

$$L_{\text{L1}} = \|S(I(x)) - Y\|_1, \tag{5}$$

where $S(\cdot)$ denotes the proposed SnowFormer network, $I(x)$ is the input snowy image, and Y is the corresponding ground truth image.

To further enhance the perceptual quality of the restored images, we also incorporate a perceptual loss. This loss is computed on feature maps extracted from specified layers of a pretrained VGG-19 network and is formulated as:

$$L_{\text{perceptual}} = \sum_{j=1}^{2} \frac{1}{C_j H_j W_j} \|\phi_j(S(I(x))) - \phi_j(Y)\|_1, \tag{6}$$

where ϕ_j represents the activation of the j-th selected layer in VGG-19, and C_j, H_j, and W_j correspond to the number of channels, height, and width of the feature map at that layer, respectively.

The overall loss function is expressed as a weighted sum of the reconstruction and perceptual losses:

$$L = \lambda_1 L_{\text{L1}} + \lambda_2 L_{\text{perceptual}}, \tag{7}$$

where λ_1 and λ_2 are empirically set to 1 and 0.2, respectively.

V. Implementation Details

We implement our method using PyTorch [43] on a single NVIDIA RTX 4090 GPU. During training, we use a batch size of 4 and a learning rate of 10^{-4} with the Adam optimizer [44]. The images and events are cropped and resized to 256×256. The events are accumulated over 10 ms and converted to a voxel-grid representation with 10 channels.

VI. Evaluation

We now describe the evaluation of our proposed method on the DSEC-Snow dataset and the real snow dataset. We compare our method with existing state-of-the-art methods for the task of snow occlusion removal. We also perform ablation studies to understand the effect of events and images on the performance of our method. Our method is evaluated using the following metrics:

- Peak Signal-to-Noise Ratio (PSNR): Higher PSNR indicates better quality of the reconstructed image.
- Structural Similarity Index (SSIM): Higher SSIM indicates greater similarity between the reconstructed image and the ground-truth image.

In addition, we evaluate the performance of our method on downstream tasks such as object detection, depth estimation, and optical flow using their corresponding standard metrics.

Baselines We consider state-of-the-art single-image desnowing approaches proposed in [1], [2]. In addition, we consider the RLP model proposed in [21], which introduces a novel method for night-time deraining. This is specifically considered as the rain in the night-time sequences has a similar appearance to snow occlusions. We also consider the state-of-the-art publicly available video-based desnowing approach S2VD [37]. Additionally, we include the E2VID [45] method, for event-based image reconstruction method.

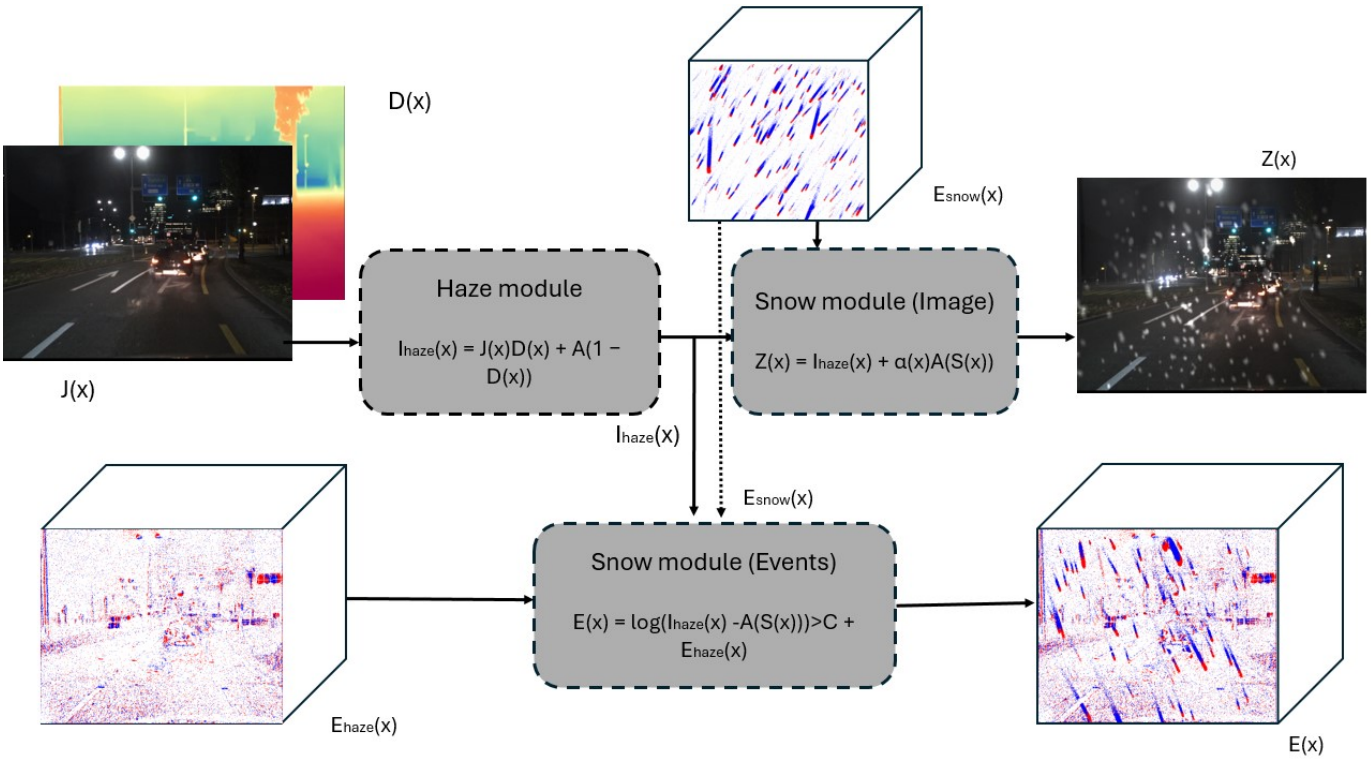


Fig. 4: **Overview of the synthetic snow dataset generation process.** Given a clean background image $J(x)$ and its corresponding event stream $E_{\text{haze}}(x)$ and foreground snow event stream $E_{\text{snow}}(x)$, we generate a synthetic snow-occluded image $Z(x)$ and a synthetic event stream $E(x)$. The haze module generates a hazy image $I_{\text{haze}}(x)$ using an atmospheric scattering model [46] based on the background image $J(x)$ and its depth map. The snow module overlays snow onto the hazy image to produce the snow-occluded image $Z(x)$. Simultaneously, background events $E_{\text{haze}}(x)$ and snow events $E_{\text{snow}}(x)$ are combined to generate the final event stream $E(x)$.

VII. Dataset

To the best of our knowledge, there exists no dataset consisting of events for desnowing. Moreover, obtaining accurate ground-truth for such a task is also quite challenging. Previous image-based and video-based approaches relied on simulated datasets such as SnowCity [5], SnowKITTI [5] to train and evaluate their models. However, due to the limited framerate of the cameras, simulating events from these videos is not possible. We therefore propose to use real events recorded by a physical event camera for both the snow and background movement. Our dataset uses a popular visual effect technique called "Chroma key compositing". We record two independent sequences consisting of background motion and foreground motion using a real event camera and RGB camera. The foreground motion is recorded in front of a black screen to have maximum contrast between the snow particles and the background. These events are then overlaid on the background events as described in Section VII. To show the performance of our approach with real snowfall, we also record driving sequences in snowfall using a real event camera. We describe this in detail in Section VII-B.

A. DSEC-Snow dataset

Accurate simulation of image degradation due to weather conditions has been a popular research topic as it provides a supervision signal in the form of ground-truth clean images, the ability to generate large-scale datasets, and a benchmark to evaluate methods. Generating synthetic snow-occluded images and video has been proposed in the past [5], [8], which overlay realistic snow occlusions on top of clean images, providing ground-truth for training and evaluation. We adopt a similar approach for generating the snow-occluded images in our dataset. The clean image ($J(x)$) is taken from the DSEC dataset [16], which consists of driving sequences recorded using a Prophesee event camera and RGB camera mounted on a car while driving in different cities in Switzerland. The snow occlusions are separately recorded using a real event camera during a snowfall. These snow occlusions are overlaid on the background image to produce a snow-occluded image ($I_{snow}(x)$), similar to the approach used in [8]. To add realistic snow image degradation, we also render haze on the background image based on the atmospheric scattering model [46]:

$$I_{haze}(x) = J(x) \cdot t(x) + A \cdot (1 - t(x)), \tag{8}$$

where $t(x)$ is the transmission map, A is the atmospheric light, and $J(x)$ is the background image. The transmission map is computed using the depth map of the image computed by [47]. The snow rendering is done by overlaying the snow occlusion on the background image:

$$Z(x) = I_{haze}(x) + \alpha \cdot Aug(E_{snow}(x)), \tag{9}$$

Algorithm 2 Synthetic Event Stream Generation Using Real Background and Snow Occlusion Events

Input: Clean background image $J(x)$, background events $E_{\text{haze}}(x)$, snow events $E_{\text{snow}}(x)$, haze parameters A.
Output: Snow-occluded image $Z(x)$, synthetic events $E(x)$.

Compute depth map $D(x)$ from $J(x)$ using a depth estimation model [47]
Generate haze image $I_{\text{haze}}(x)$ using $D(x)$,, haze parameters A, and atmospheric scattering model [46]
Overlay snow events on $I_{\text{haze}}(x)$ to obtain snow-occluded image $Z(x)$
Initialize synthetic event stream: $E(x) \leftarrow \emptyset$
for all events $e_{\text{snow}} \in E_{\text{snow}}(x)$ **do**
    Let (x, y) be the event location and t the timestamp
    if $|I_{\text{haze}}(x, y) - I_{\text{snow}}(x, y)| > C$ **then**
        Add e_{snow} to $E(x)$
    end if
end for
for all events $e_{\text{haze}} \in E_{\text{haze}}(x)$ **do**
    Let (x, y) be the event location and t the timestamp
    if there is no overlapping snow event e_{snow} at (x, y, t) in $E(x)$ **then**
        Add e_{haze} to $E(x)$
    end if
end for
return $Z(x)$, $E(x)$

where $Z(x)$ is the final image, α models ambient illumination of the scene [8], and $Aug(E_{snow}(x))$ is the augmentation function that simulates the occluded image from foreground snow events.

While such realistic rendering of snow occlusion is possible for images, for events, generating events that simulate the real world and the real sensor is quite challenging and still an open problem [48]. We therefore propose to instead use a unique method to generate a synthetic dataset by combining multiple event streams recorded with a real event camera, as shown in Fig. 4. It consists of recording two event streams, one corresponding to background motion resulting from camera movement (E_{haze}) and a second event stream corresponding to the motion of occlusion such as snow (E_{snow}). These events are combined using the process described below.

Incorporating the "chroma key compositing" technique, we record the snow particles in front of a black screen. These foreground events are merged with the background events using two main criteria:

- Background intensity correction: Events are only generated if there exists a contrast between the snowflake (I_{snow}) and the background. Since snowflakes are typically bright [9], brighter backgrounds will not generate events even if a snowflake moves across a bright pixel.
- Background events overlap: In the case where background activity generates events at the same time as the occlusion, priority is given to the occlusion event as the occlusion is typically in front of the scene.

The algorithm described above is also summarized in Algo. 2

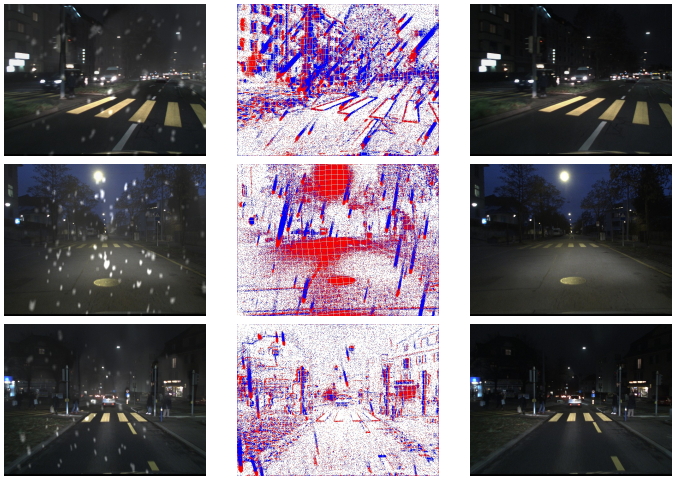


Fig. 5: **Example scenes from our DSEC-Snow dataset**. It consists of synchronized RGB frames(Left), Events (Middle) and Groundtruth (Right).

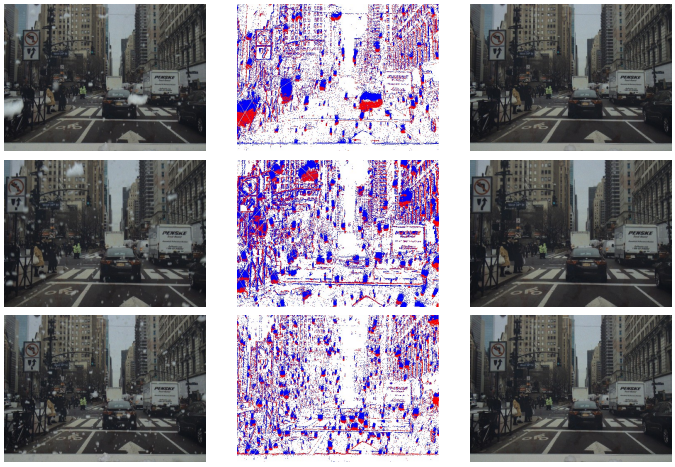


Fig. 6: **Examples scenes from our Slider-Snow dataset** It consists of synchronized RGB frames(Left), Events (Middle) and Groundtruth (Right).

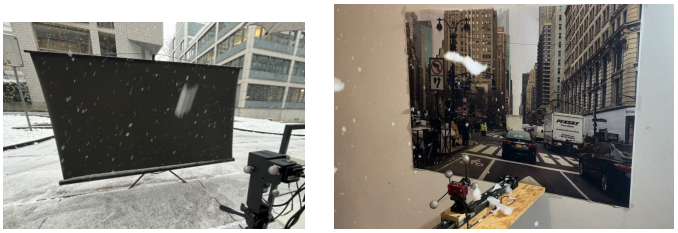
Examples sequences are shown in Fig. 6. More details about our dataset can be found in the supplementary material.

B. Slider-Snow dataset

To evaluate the model on real data, we propose to collect our own dataset. Evaluating with real snowfall, however, is quite challenging as there is no ground-truth available. We therefore propose to use a controlled setup to evaluate the performance in the real world, using a snow machine. We use a linear slider to move the event camera at a fixed velocity while recording the snowfall using a snow machine. This allows us to record sequences with snowfall and ground-truth data. The details about the setup and data generation process are shown in Fig. 7 and elaborated in the supplementary material.

C. Real Snowfall Driving Sequences

Finally, we also collect a new dataset consisting of real snowfall driving sequences. We use the BeamSplitter setup of the Prophesee event camera [49] and FLIR BlackFly S global shutter RGB camera mounted on the dashboard of the car



(a) (b)

Fig. 7: **Overview of the experimental setups** (a) Our setup for recording the foreground occlusion events used for generating DSEC-Snow sequence. (b) Experimental setup for our controlled real-world dataset. The scene was printed on a poster and the camera was placed on a linear slider with a fixed velocity. The snowfall was simulated using a snow machine.

		DSEC-Snow		Slider-Snow	
Method	**Input**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
Restormer [1]	I	23.12	0.8909	**25.16**	**0.8939**
SnowFormer [2]	I	25.12	0.9240	17.63	0.5593
RLP [21]	I	11.20	0.5383	10.10	0.6062
S2VD [37]	V	23.37	0.8758	**25.16**	0.8706
E2VID [45]	E	11.84	0.3369	15.77	0.3416
Model-based	E+I	21.24	0.5656	19.81	0.6254
Ours	E + I	**31.76**	**0.9686**	*24.18*	*0.8467*

TABLE II: **Quantitative comparison of image desnowing methods on our DSEC-Snow and Slider-Snow datasets** We report PSNR and SSIM for each method using different input modalities: intensity images (I), video (V), events (E), and both (E+I). The proposed approach, leveraging both event and image data, demonstrates higher image reconstruction quality compared existing image-based and event-based methods.

while driving in the snowfall. Of course, as there is no ground-truth available, we only use these sequences for qualitative evaluation. See examples in Fig. 6.

VIII. Results

We evaluate our method on three datasets: the DSEC-Snow dataset, the Slider-Snow dataset and real-world driving sequences. We benchmark against existing state-of-the-art methods on the DSEC-Snow dataset in Section VIII-A and on the Slider-Snow dataset in Section VIII-B. Furthermore, we provide qualitative results and comparisons on real-world driving datasets in Section VIII-C. The supplementary material includes more results, ablation studies on network architecture, and evaluations on downstream tasks such as optical flow.

A. DSEC-Snow Dataset

Table II presents a quantitative comparison of various image desnowing methods evaluated on the DSEC-Snow and Slider-Snow datasets, reporting PSNR and SSIM for each method. The image-based approaches (Restormer, SnowFormer, and RLP) which relying solely on intensity images (I) show moderate performance. Since RLP was trained for night-time deraining, it struggled to generalize to day-time occlusions which is evident from the low PSNR and SSIM scores. Video-based methods (S2VD) leverage temporal information from video sequences, achieving better results than single-image methods, but still fall short of the performance of our proposed method, as events provide high resolution temporal information that videos of 20 fps cannot capture. Notably, event-only methods (E2VID) perform significantly worse, particularly in terms of SSIM, indicating that events alone are insufficient for effective desnowing in highly occluded scenes. Model-based fusion of events and images (E+I) shows improved PSNR over single-modality approaches however, the overall image quality is low (indicated by low SSIM scores), suggesting that naive fusion of modalities does not capture the complex spatio-temporal patterns of snow occlusions effectively. In comparison to all, our method, which fuses both event and intensity information (E+I), achieves the highest PSNR and SSIM across both datasets, outperforming all baselines by a significant margin. Qualitative results are shown in Fig. 9, where our method effectively removes snow occlusions while preserving fine details of the background scene.

Ablation on occlusion density We study the robustness of desnowing methods under varying occlusion conditions in Fig. 10. Each row in the figure represents a distinct occlusion density, increasing from top to bottom. As the occlusion density increases, we observe a substantial degradation in scene visibility and reconstruction quality for the image-only and event-only baselines. SnowFormer and RLP exhibit noticeable artifacts and loss of detail under severe occlusions, leaving substantial snow artifacts in the reconstructed outputs. In contrast, our method demonstrates consistent robustness across all occlusion densities, effectively removing snow particles and preserving fine scene details. These results highlight the benefit of jointly leveraging event and image data, enabling our approach to maintain high-quality reconstructions even in highly adverse weather conditions with dense occlusions.

B. Results on Slider-Snow Dataset

We further evaluate our method on a real-world snowfall dataset to assess its generalization ability beyond the synthetic domain. Specifically, we deploy the model trained on our synthetic DSEC-Snow dataset and apply it directly to the Slider-Snow dataset, which is captured under real snowfall conditions. Fig. 11 compares our method with state-of-the-art image-based baselines such as Restormer [1] and SnowFormer [2]. Our approach (column d) consistently produces better reconstructions compared to the original occluded image (column a), Restormer (column b), and SnowFormer (column c). Notably, in the first and third rows, where dense snowflakes heavily obscure vehicle details, our method is able to restore the underlying vehicle contours and even fine-grained elements such as traffic signs, which are either blurred or completely lost in the baseline. Table II presents a quantitative comparison of all baselines on Slider-Snow dataset. Although certain image-based baselines, such as Restormer [1] achieve slightly higher PSNR, qualitative comparisons indicate that they fail to effectively remove occlusions, as illustrated in Fig. 11. These qualitative improvements highlight the strength of using event data for temporally consistent occlusion-aware image reconstruction. By attending to motion patterns in the event stream, our model infers occluded content more robustly than

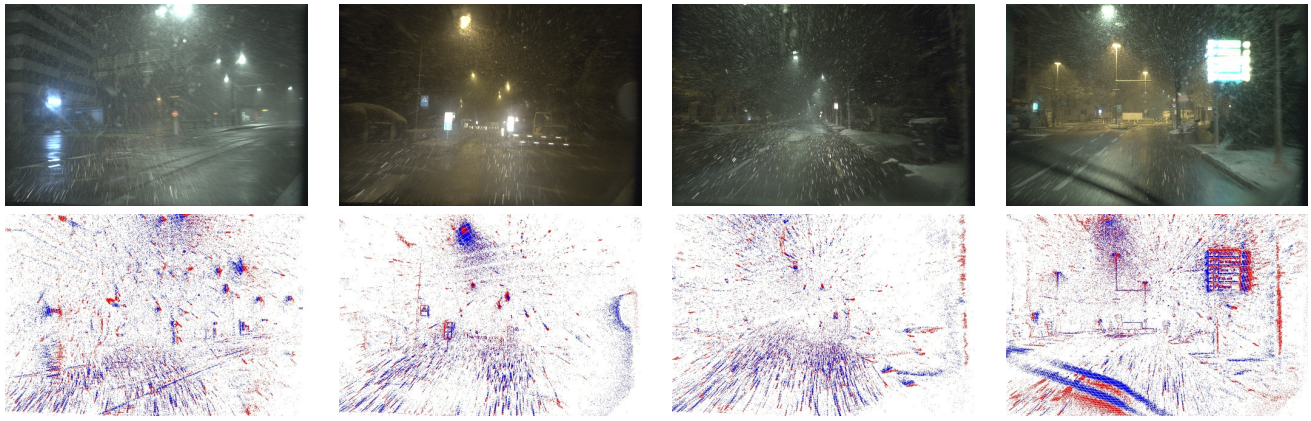


Fig. 8: **Samples from our real-world snowfall driving dataset** The images are recorded using the synchronized and aligned setup of RGB camera (top) and event camera (bottom) and mounted on the dashboard of the car while driving in the snowfall.

(a) Image (b) Restormer [1] (c) Snowformer [2] (d) **Ours** (e) GT

Fig. 9: **Qualitative comparison on DSEC-Snow dataset.** (a) Input images with synthetic snow,(b)-(c) image restoration baselines , (d) our result, and (e) ground truth (GT) images. On our synthetic DSEC-Snow dataset, our fusion of event and image data enables more effective snow removal and better recovery of occluded details. The proposed approach recovers clearer scene structures and more faithfully restores the underlying content compared to prior methods.

purely image-based approaches. Importantly, the model is not fine-tuned on real-world snow data, indicating strong cross-domain generalization. This result validates the effectiveness of our synthetic dataset generation pipeline and supports its applicability to training models deployable in real-world conditions.

C. Results on real snowfall driving sequences

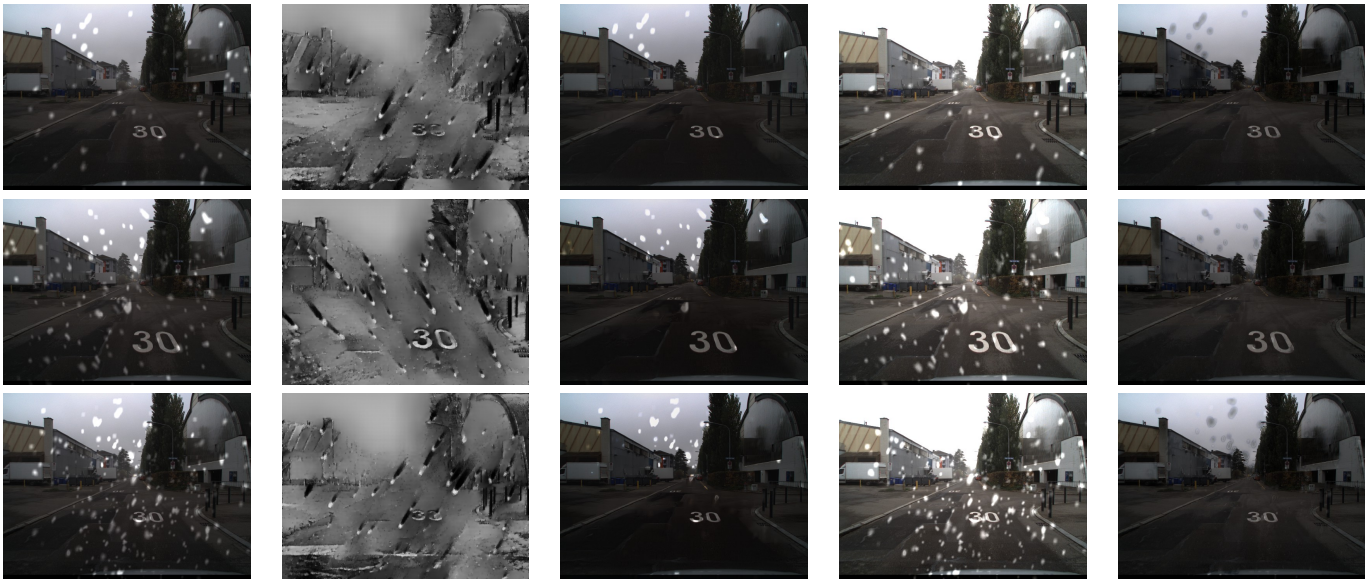
We also evaluate our method on real-world snowfall driving sequences acquired using a beamsplitter-based sensor rig comprising a Prophesee Gen4 event camera (1280×720) and a FLIR BlackFly S global shutter RGB camera (1440×1080). This setup enables temporally aligned acquisition of high-dynamic-range (HDR) event data and intensity frames under challenging low-light and high-motion conditions. Qualitative results are presented in Fig. 12.

Our method demonstrates a clear advantage in removing snow-induced occlusions and recovering underlying scene structure. The baseline RGB-only methods, Restormer [1] and Snowformer [2], struggle with motion blur and fail to distinguish snowflakes from meaningful scene content, often resulting in over-smoothed or distorted reconstructions. In contrast, our method preserves high-frequency details and recovers semantically relevant features such as traffic signs, barriers, and road markings, even under heavy snowfall.

In Fig. 12, where oncoming headlights and snowflakes dominate the visual field, the image-only baselines suffer from halo artifacts and flare-induced saturation. Our approach effectively suppresses such artifacts, allowing visibility of distant road features, including lane markers and background vehicles. This is largely attributable to the asynchronous, high-temporal-resolution nature of the event data, which captures scene dynamics without the integration-based blur inherent to conventional RGB sensors.

The qualitative comparisons in Fig. 12 validate the effectiveness of our approach in real-world adverse weather conditions. By leveraging the complementary sensing modalities of events and images, our method achieves robust occlusion removal and scene restoration beyond the capability of state-of-the-art RGB-only models. Please see the supplementary material for additional results on real snowfall driving sequences.

Model Complexity and Parameters Comparison We evaluated the computational efficiency of our method by measuring the average runtime on NVIDIA GeForce RTX 4090 GPU. Our approach processes a single image with resolution



(a) Image (b) E2VID [45] (c) Snowformer [2] (d) RLP [21] (e) Ours

Fig. 10: **Effect of occlusion density on image reconstruction quality** Each row corresponds to a different level of occlusion density, increasing from top to bottom. As occlusion density increases, visibility of scene details and robustness of the image-only algorithm degrade significantly. Our method performs consistently well across all occlusion densities, effectively removing snow occlusions and preserving scene details.

(a) Image (b) Restormer [1] (c) Snowformer [2] (d) **Ours** (e) GT

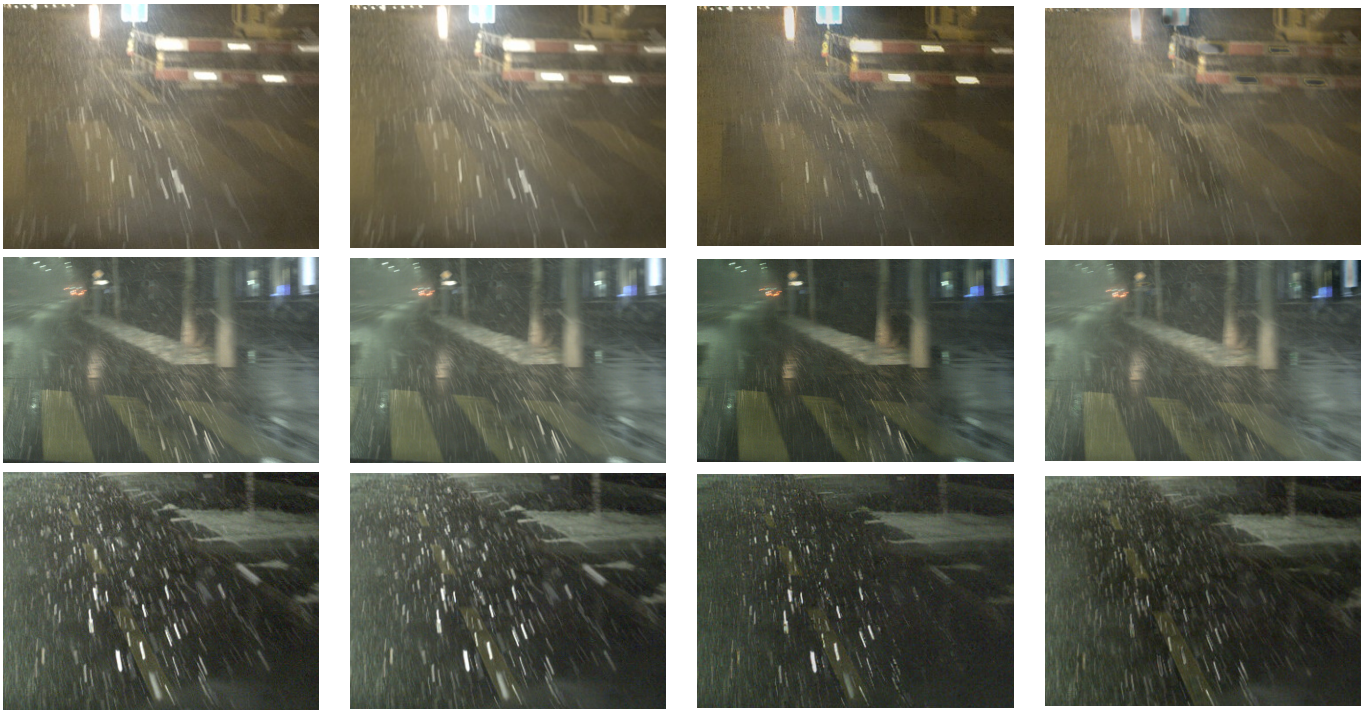
Fig. 11: **Qualitative comparison of image desnowing results on Slider-Snow data.** (a) Input images with synthetic snow,(b)-(c) image restoration baselines , (d) our result, and (e) ground truth (GT) images. On real-world Slider-Snow data, we show similar improvements in snow removal and detail recovery, demonstrating the strong generalization of our method to real snow conditions. The proposed approach recovers clearer scene structures and more faithfully restores the underlying content compared to prior methods.

512×512 in approximately 0.06 seconds, which is comparable to recent state-of-the-art methods while providing improved desnowing performance. The model contains 10.3 million parameters, which is on par with Snowformer [2] ($8.38M$) and significantly lower than Restormer [1] ($26M$).

IX. Discussion and Limitations

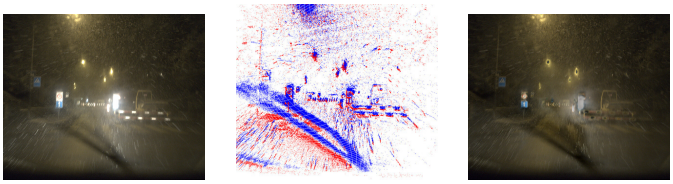
While our proposed synthetic dataset provides a practical and physically grounded approach to simulating event-based occlusions, it relies on the assumption that independently captured motions, such as background and occluder (e.g., snowflake) events, can be linearly combined without introducing artifacts. This simplification does not account for interdependencies such as lighting interactions, occlusion ordering, or nonlinear sensor responses, which may be present in real-world scenes. For example, in Fig. 13, we show a real snowfall scene where the street lamps create flickering light sources and the car wipers create dynamic occlusions that are not captured by our compositing approach and lead to artifacts in the reconstructed images.

Another inherent limitation arises from the characteristics of event cameras themselves. Events are triggered only when a local brightness change exceeds a sensor-defined contrast threshold (typically around 15%). As a result, regions with flat textures, uniform lighting, or minimal motion fail to produce meaningful event activity. This restricts the utility of our approach in low-texture environments or scenes where fine-



(a) Image (b) Restormer [1] (c) Snowformer [2] (d) **Ours**

Fig. 12: **Qualitative comparison on real driving scenes.** (a) Input images, (b)-(c) image restoration baselines, and (d) our results. This shows a magnified region with dense snow and reflective surfaces, highlighting the ability of our method to reduce snow occlusion and preserve scene details under adverse snowfall conditions.



(a) Image (b) Events (c) Ours

Fig. 13: **Limitations of the compositing approach**. The synthetic dataset generation method relies on the assumption that independently captured motions (e.g., background and occluder events) can be linearly combined without introducing artifacts. However, in real snowfall scenes, there are complex interactions such as flickering light sources from the street lamps, and car wipers which create dynamic occlusions that are not captured by our compositing approach.

grained intensity variations fall below the contrast threshold. Consequently, the reconstruction quality is strongly tied to the presence of high-frequency spatial and temporal information in the scene.

Despite these limitations, our approach offers a scalable and flexible alternative to traditional physics-based simulators, which often require complex modeling of event camera characteristics. Notably, the dataset generation methodology can be extended beyond weather simulation: any independently recordable motion (e.g., pedestrians, vehicles, or dynamic objects) can be composited over different backgrounds to augment data diversity in a controlled manner. This makes the method valuable for a wide range of applications such as robotics, autonomous driving, and dynamic scene understanding, where controlled yet realistic event-based data are scarce.

X. Conclusion

We introduce a novel event-based approach for background image reconstruction in the presence of dynamic occlusions. It leverages the complementary nature of event cameras and frames to reconstruct true scene information instead of hallucinating occluded areas as done by image inpainting approaches. Specifically, our proposed data-driven approach reconstructs the background image using only one occluded image and events. The high temporal resolution of the events provides our method with additional information on the relative intensity changes between the foreground and background, making it robust to dense occlusions. To evaluate our approach, we present the first large-scale dataset recorded in the real world containing challenging scenes with synchronized events, occluded images, and ground-truth images. Our method achieves an improvement of 3 dB in PSNR over state-of-the-art frame-based and event-based methods on both synthetic and real datasets. We believe that our proposed method and dataset lay the foundation for future research.

SUPPLEMENTARY MATERIAL

XI. DSEC-SNOW DATASET

We describe the setup used here to record our datasets. In this section, we provide a detailed overview of the data generation process that was used to create the DSEC-Snow dataset. We used a black background to maximize the contrast between the foreground and background as shown in the experimental setup. The underlying assumption in this setup is that snow particles tend to be brighter than most objects in the surroundings. Therefore, to maximize the contrast between every snowflake movement and the background, we use a black screen. These events can then be pruned to account for an arbitrary background.

The second assumption which we make is that snowflakes will always be in the foreground, i.e. they will never be occluded by the background, which for most common scenarios is a reasonable assumption. Of course, this also means we do not model any depth perception in this fusion and treat foreground events as far away from the background.

Lastly, as of now, we do not model the motion of the snow particles according to ego-motion of the camera. In typical driving scenarios, the ego-motion of the car makes the snowflake appear to come towards the camera rather than simply flying down. This of course is a function of the speed of the camera and makes this matter of accurately simulating the motion of snowflakes quite challenging. We therefore simplify this problem and only apply a homography transformation to the foreground events. Overall, our dataset consists of around 200 training and 50 test sequences, with each sequence consisting of a short duration of driving with snowfall during both day and night.

Augmentation Parameters We describe in detail the augmentation parameters used to generate the dataset. Similar to the method proposed in [8], we render different effects of snowfall on the image such as haze and illumination-dependent snow appearance. To render haze, we follow the model proposed in [46] which uses the atmospheric scattering model for rendering hazy images. To blend the snow foreground with the background image, we use the strategy proposed in [8], by considering the ambient illumination and time of day to blend the snowflakes. For example, during the day, snowflakes blend with the sky and are therefore not easily visible with the brighter sky background. This is exactly the opposite during the night: snowflakes are more visible in brightly lit areas such as headlamps or streetlights [8]. These hazy images are combined with the foreground snow events to produce the final image. To simulate realistic snow occlusion, we apply different augmentations to the snow events. As described in Equation. . 9, we apply different augmentations to the snow:

- Snow Speed: The speed of the snowflakes is artificially controlled by scaling the timestamp of the foreground events.
- Snow Density: The density of snowflakes is increased by overlaying multiple snow events by staggering the timestamps and applying a homography transformation to the foreground events.
- Motion Direction: Motion direction is only controlled by flipping the foreground events along the horizontal axis.

Simulating Events Background events correspond to driving sequences recorded using DSEC [16]. It consists of an event camera and an RGB camera mounted next to each other on a car while driving in different cities in Switzerland. The background events ($E_{haze}(x)$) are recorded with a Prophesee Gen3.1 event camera with a resolution of 640×480 pixels. The background images ($J(x)$) are recorded at $20Hz$ with a resolution of 1440×1080 pixels and are aligned with the events. Since our approach relies on the temporal and spatial alignment of the events and images, we use the rectified and aligned events and images from the DSEC dataset.

The foreground events ($E_{snow}(x)$) are recorded using a Prophesee Gen4 event camera with a resolution of 1280×720 pixels. The foreground events are recorded in front of a black background. It was shown in [50], that the signal-to-noise ratio (SNR) of an event camera depends on the illumination and contrast of the scene. As illumination increases, the SNR of the events can drop significantly if the contrast is not high enough. As we were restricted to outdoor recording, we could not control the illumination of the scene, so we use a black background to maximize the contrast between the foreground (snow, typically bright) and the background, ensuring sufficient SNR when recording the foreground events. Dataset statistics are shown in Fig. 14. The dataset consists of 1000 training and 470 test pairs of images and events. In addition, we provide ground truth images and events for both the training and test sequences. A histogram illustrating the percentage of occluded pixels reveals the distribution of occlusion intensity across the dataset. Most images have occlusion levels concentrated between approximately 13% and 22%, indicating that moderate occlusion intensity is prevalent. Fewer instances exhibit extreme occlusion levels, highlighting the realistic variability in weather conditions captured by the dataset.

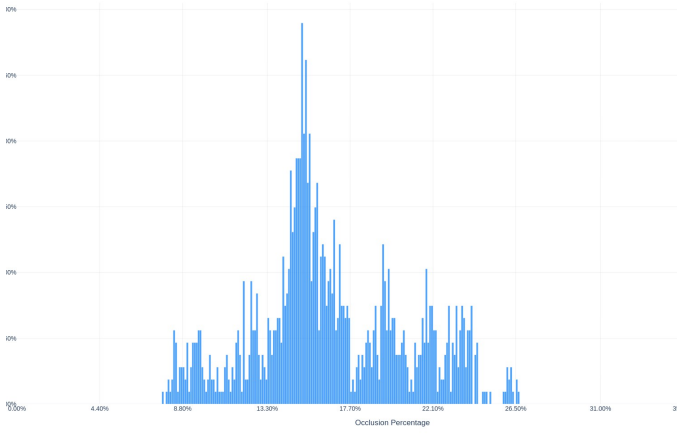


Fig. 14: Dataset statistics of DSEC-Snow dataset. The histogram shows the percentage of occluded pixels in the dataset, illustrating the distribution of occlusion intensity across the dataset.

A. Comparison with State-of-the-Art Methods

We show samples from our method and compare it with state-of-the-art desnowing methods on the DSEC-Snow dataset in Fig. 15.

We show samples from our method and compare it with state-of-the-art desnowing methods on the Slider-Snow dataset in Fig. 16 and Fig. 17.

In addition to the qualitative results, we also evaluate our method on a downstream task, depth estimation, using the real driving dataset in Fig. 18.

XII. Artifacts and limitation analysis

In some cases, the reconstructed images may exhibit a degree of blurriness or checkerboard artifacts. Checkerboard artifacts arise from the inherent checkerboard pattern in the event stream, which is a consequence of the alignment from the event camera to the RGB camera. As shown in Fig. 19, this checkerboard pattern is visible in the event stream.

In some cases, slight blurriness or artifacts can also appear in the reconstructed images. This effect is largely caused by depth-dependent alignment limitations in the DSEC dataset. In DSEC, events and frames are aligned using a global 2D homography. During camera motion in non-planar scenes, objects at different depths cannot be perfectly aligned simultaneously (see Fig. 20). As a result, certain scene elements may appear slightly misaligned between events and frames, which can lead to localized smoothing artifacts during reconstruction. For example, the degradation around the road marking "30" in is influenced by this depth-dependent misalignment.

Image reconstruction metrics such as PSNR often favor globally smooth reconstructions that reduce high-frequency differences with the reference image. Consequently, methods that slightly smooth textures can sometimes obtain higher PSNR values even when local sharpness is reduced. To better evaluate the impact of reconstruction quality on downstream perception tasks, we therefore also include motion-based metrics based on optical flow estimation. These metrics provide an indirect measure of how well the reconstructed images preserve motion-consistent scene structures.

As shown in Table III, our method achieves improved endpoint error (EPE) compared to competing approaches, indicating that the reconstructed images preserve motion information useful for downstream perception tasks. At the same time, the average pixel error (APE) remains within a very small margin ($< 1\%$), suggesting that the remaining blur artifacts mainly affect fine pixel-level alignment rather than the global structural consistency of the scene.

These observations highlight an important direction for future work. In particular, improving event-frame alignment through depth-aware calibration or incorporating more physically grounded snowfall simulation could further reduce these artifacts and improve local texture preservation while maintaining robust snow suppression.

XIII. Additional Results

A. Generalization to other occlusions

We also consider the all-weather driving dataset proposed in [51], which contains synchronized and calibrated events, images, LiDAR, and RADAR measurements. We only consider the sequences which have a mix of snow and rainfall. Unfortunately, the snowfall is not dense enough in these sequences to hinder the view, unlike our dataset which was collected in heavy snowfall. Nevertheless, we show the qualitative evaluation of our approach in Fig. 21. By fusing temporal information from events, we are able to recover the bus in the background of the raindrop and building, which was occluded by a raindrop causing lens flare.

Effect on downstream application We also evaluate the performance of our method on the downstream task of optical flow. We show that our method is able to reconstruct the background scene with high accuracy, resulting in significant improvement in downstream applications, see Fig. 22. We use the RAFT network [52] on the images reconstructed by all image restoration methods and compare the end-point-error (EPE) metric to evaluate the performance of optical flow.

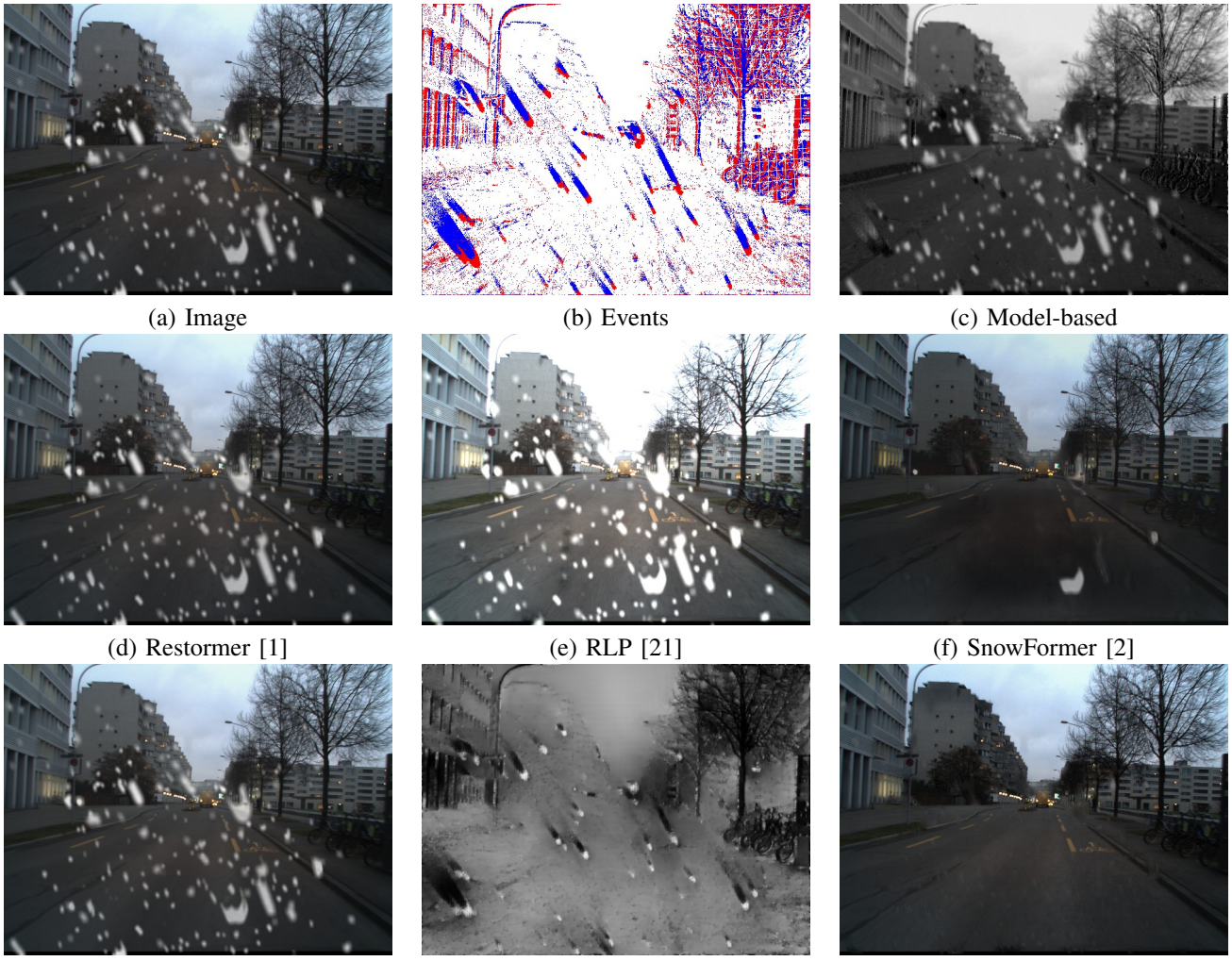
Table III reports the End-Point Error (EPE) and accuracy metrics ($AE < 1$, $AE < 3$, $AE < 5$) for optical flow computed on the Slider-Snow dataset. Event-only (E2VID) and video-based (S2VD) baselines exhibit high EPE and low accuracy, reflecting their limited ability to recover motion information in the presence of severe occlusions. Image-based methods (Restormer, SnowFormer, RLP) provide improved performance, but still suffer from significant error, particularly under challenging conditions. Our approach, which fuses event and intensity modalities, achieves the lowest EPE. The qualitative results in Fig. 22 further confirm these findings: optical flow maps generated using our method are visually closer to the ground truth, accurately capturing fine motion boundaries and overall scene structure, while baseline methods display artifacts and loss of detail. These results indicate that our event-image fusion approach not only improves desnowing quality, but also leads to substantial gains in the downstream tasks such as optical flow estimation.

Method	Input	Optical Flow			
		EPE ↓	$AE < 1$ ↑	$AE < 3$ ↑	$AE < 5$ ↑
Restormer [1]	I	19.36	0.06	0.36	0.48
SnowFormer [2]	I	30.80	0.02	0.20	0.30
RLP [21]	I	17.64	0.06	0.36	0.47
S2VD [2]	V	29.64	0.03	0.24	0.35
E2VID [45]	E	42.92	0.00	0.04	0.08
Ours (Model-based)	E+I	39.58	0.00	0.11	0.19
Ours	E+I	**10.42**	**0.11**	**0.56**	**0.67**

TABLE III: **Quantitative evaluation of downstream task (optical flow)**-on desnowed images using different restoration methods and input modalities. Our event-image fusion approach consistently achieves the best performance, as measured by EPE and accuracy for optical flow.

Model		**PSNR (dB)**	**SSIM**
Image-only	I	25.12	0.9240
Event-only	E	17.29	0.6864
Naive Fusion	E + I	29.10	0.9519
Ours	E + I	**31.76**	**0.9686**

TABLE IV: Ablation study on effect of sensor modalities.



(a) Image (b) Events (c) Model-based

(d) Restormer [1] (e) RLP [21] (f) SnowFormer [2]

(g) S2VD [37] (h) E2VID [45] (i) Ours

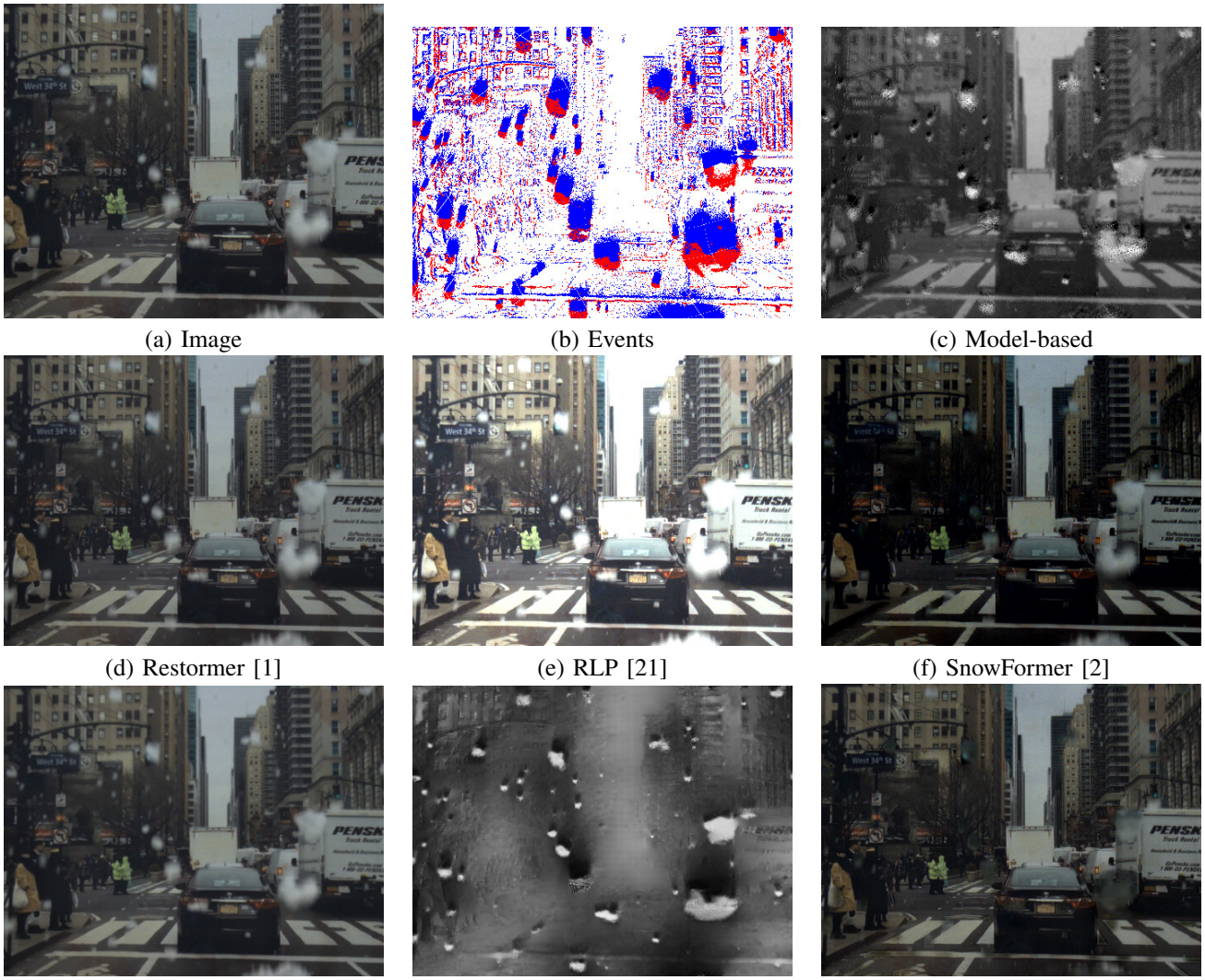
Fig. 15: **Comparison of our method with state-of-the-art desnowing methods on DSEC-Snow dataset.**

Network Components		PSNR (dB)	SSIM
EventNet	Mask Prediction		
×	×	29.10	0.9519
×	✓	29.45	0.9612
✓	×	31.09	0.9586
✓	✓	**31.76**	**0.9686**

TABLE V: Ablation study on network architecture.

Ablation on network architecture We also compare the contribution of each sensor modality in Table IV in first two rows. The image-only method is same as Snowformer [2], while the event-only method uses the same architecture as proposed method but discards the image input. While events provide valuable motion cues, they lack the low-frequency structural information present in image frames, explaining the lower performance of the event-only pipeline compared to the image-only baseline. Fig. 24 shows a qualitative comparison of the three approaches, where the event-only method fails to reconstruct the overall structure of the scene, while the image-only method struggles to remove occlusions, resulting in artifacts and loss of detail. Fusion of both modalities leverages the strengths of each sensor, resulting in better reconstruction quality as evidenced by the improvement in PSNR and SSIM. It can be seen that even with a naive fusion approach, we achieve higher PSNR and SSIM than image-only methods. However, our proposed method outperforms the naive fusion approach by a significant margin, demonstrating the effectiveness of our architecture.

In addition, we show how different components of our network architecture contribute to the overall performance in Table V. The naive fusion strategy performs a simple concatenation of the image and event inputs, which is then directly passed to the image reconstruction module. This approach lacks any form of temporal modeling, as it does not incorporate recurrence or sequential reasoning over the event stream. Incorporating mask prediction from events allows the network to adaptively combine the input image and the network's prediction, leading to improved reconstruction quality. However, the most significant performance gain comes from EventNet, which extracts salient spatio-temporal features from the event stream before fusion. Qualitatively, as shown in the Fig. 23 the naive fusion approach results in several artifacts and loss of detail, while our method effectively leverages



(a) Image (b) Events (c) Model-based

(d) Restormer [1] (e) RLP [21] (f) SnowFormer [2]

(g) S2VD [37] (h) E2VID [45] (i) Ours

Fig. 16: **Comparison of our method with state-of-the-art desnowing methods on Slider-Snow dataset.**

the spatio-temporal features from events to reconstruct high-quality images.

REFERENCES

[1] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022.

[2] S. Chen, T. Ye, Y. Liu, E. Chen, J. Shi, and J. Zhou, "Snowformer: Scale-aware transformer via context interaction for single image desnowing," *arXiv preprint arXiv:2208.09703*, 2022.

[3] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2020.

[4] Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang, "Desnownet: Context-aware deep network for snow removal," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3064–3073, 2018.

[5] K. Zhang, R. Li, Y. Yu, W. Luo, and C. Li, "Deep dense multi-scale network for snow removal using semantic and geometric priors," *IEEE Transactions on Image Processing*, 2021.

[6] W.-T. Chen, H.-Y. Fang, J.-J. Ding, C.-C. Tsai, and S.-Y. Kuo, "Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal," *Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[7] W.-T. Chen, H.-Y. Fang, C.-L. Hsieh, C.-C. Tsai, I. Chen, J.-J. Ding, S.-Y. Kuo *et al.*, "All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss," *Int. Conf. Comput. Vis. (ICCV)*, pp. 4196–4205, 2021.

[8] H. Chen, J. Ren, J. Gu, H. Wu, X. Lu, H. Cai, and L. Zhu, "Snow removal in video: A new dataset and a novel method," *Int. Conf. Comput. Vis. (ICCV)*, pp. 13 211–13 222, October 2023.

[9] K. Garg and S. Nayar, "When does a camera see rain?" *Int. Conf. Comput. Vis. (ICCV)*, vol. 2, pp. 1067–1074 Vol. 2, 2005.

[10] B. Yang, Z. Jia, J. Yang, and N. K. Kasabov, "Video snow removal based on self-adaptation snow detection and patch-based gaussian mixture model," *IEEE Access*, vol. 8, pp. 160 188–160 201, 2020.

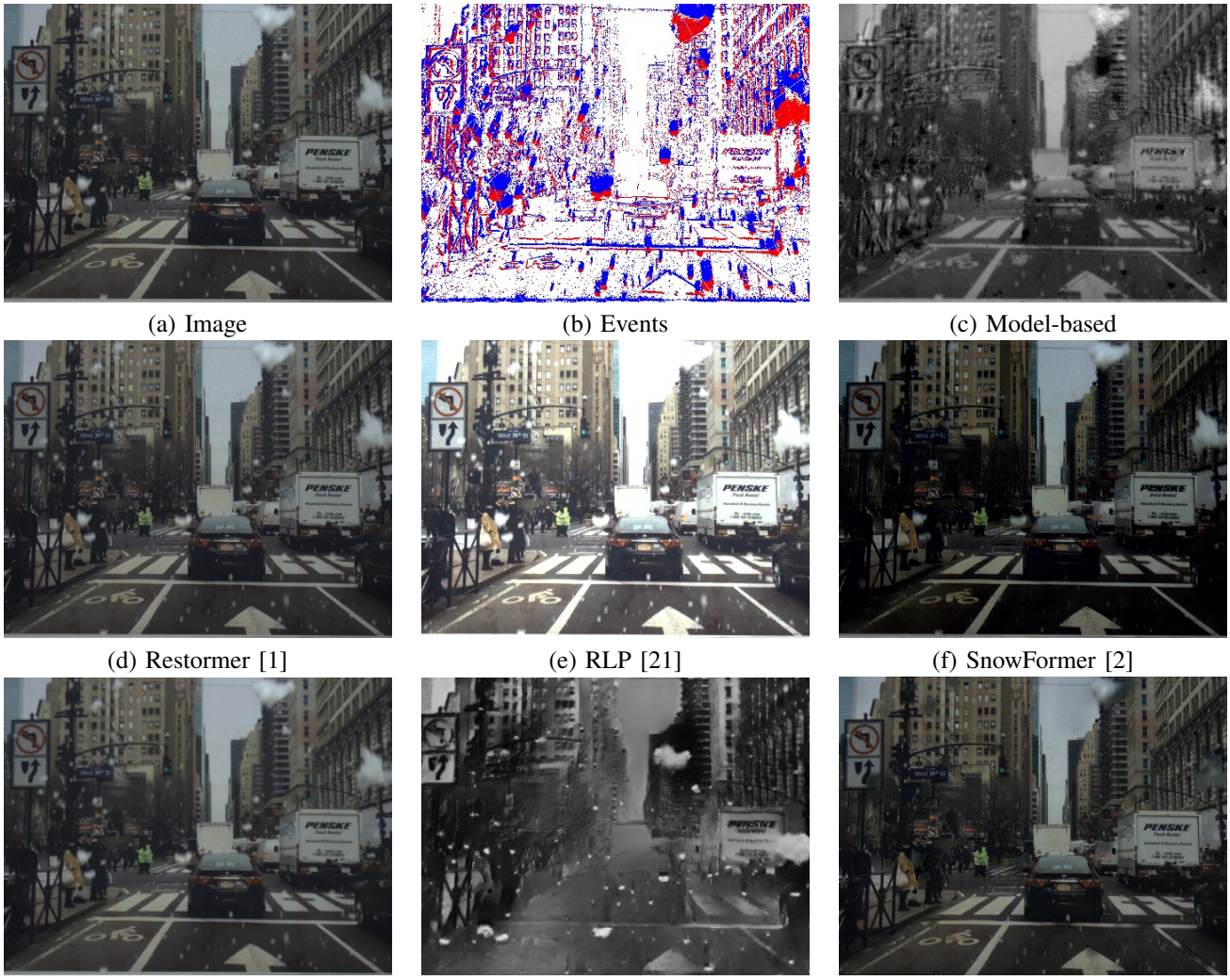
[11] W. Ren, J. Tian, Z. Han, A. Chan, and Y. Tang, "Video desnowing and deraining based on matrix decomposition," *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 2838–2847, 2017.

[12] M. Li, X. Cao, Q. Zhao, L. Zhang, and D. Meng, "Online rain/snow removal from surveillance videos," *IEEE Transactions on Image Processing*, vol. 30, pp. 2029–2044, 2021.

[13] M. Li, X. Cao, Q. Zhao, L. Zhang, C. Gao, and D. Meng, "Video rain/snow removal by transformed online multiscale convolutional sparse coding," *ArXiv*, vol. abs/1909.06148, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202572625

[14] J.-H. Kim, J.-Y. Sim, and C.-S. Kim, "Video deraining and desnowing using temporal correlation and low-rank matrix completion," *IEEE Transactions on Image Processing*, vol. 24, no. 9, pp. 2658–2670, 2015.

[15] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi,



(a) Image (b) Events (c) Model-based

(d) Restormer [1] (e) RLP [21] (f) SnowFormer [2]

(g) S2VD [37] (h) E2VID [45] (i) Ours

Fig. 17: **Comparison of our method with state-of-the-art desnowing methods on Slider-Snow dataset.**

S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[16] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "Dsec: A stereo event camera dataset for driving scenarios," *IEEE Robot. Autom. Lett.*, 2021.

[17] A. Wolf, O. Alsattam, S. Brooks-Lehnert, and K. Hirakawa, " EBSnoR: Event-Based Snow Removal by Optimal Dwell Time Thresholding ," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 01, pp. 1–13, Aug. 2025. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3603854

[18] K. Garg and S. Nayar, "Detection and removal of rain from videos," *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, vol. 1, pp. I–I, 2004.

[19] L.-W. Kang, C.-W. Lin, and Y.-H. Fu, "Automatic single-image-based rain streaks removal via image decomposition," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1742–1755, 2012.

[20] W. Yang, R. T. Tan, S. Wang, Y. Fang, and J. Liu, "Single image deraining: From model-based to data-driven and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4059–4077, 2021.

[21] F. Zhang, S. You, Y. Li, and Y. Fu, "Learning rain location prior for nighttime deraining," *Int. Conf. Comput. Vis. (ICCV)*, pp. 13 148–13 157, October 2023.

[22] S. Sun, W. Ren, X. Gao, R. Wang, and X. Cao, "Restoring images in adverse weather conditions via histogram transformer," *Eur. Conf. Comput. Vis. (ECCV)*, 2025.

[23] P. Liu, J. Xu, J. Liu, and X. Tang, "Pixel based temporal analysis using chromatic property for removing rain from videos." *Computer and Information Science*, vol. 2, no. 1, pp. 53–60, 2009.

[24] X. Zhang, H. Li, Y. Qi, W. K. Leow, and T. K. Ng, "Rain removal in video by combining temporal and chromatic properties," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2006, pp. 461–464.

[25] P. C. Barnum, S. Narasimhan, and T. Kanade, "Analysis of rain and snow in frequency space," *International Journal of Computer Vision*, vol. 86, no. 2-3, p. 256, 2010.

[26] J. Bossu, N. Hautière, and J.-P. Tarel, "Rain or snow detection in image sequences through use of a histogram of orientation of streaks," *International Journal of Computer Vision*, vol. 93, no. 3, pp. 348–367, 2011.

[27] V. Santhaseelan and V. K. Asari, "Utilizing local phase information to remove rain from video," *International Journal of Computer Vision*, vol. 112, no. 1, pp. 71–89, 2015.

[28] Y.-L. Chen and C.-T. Hsu, "A generalized low-rank appearance model for spatio-temporally correlated rain streaks," *Int. Conf. Comput. Vis. (ICCV)*, pp. 1968–1975, 2013.

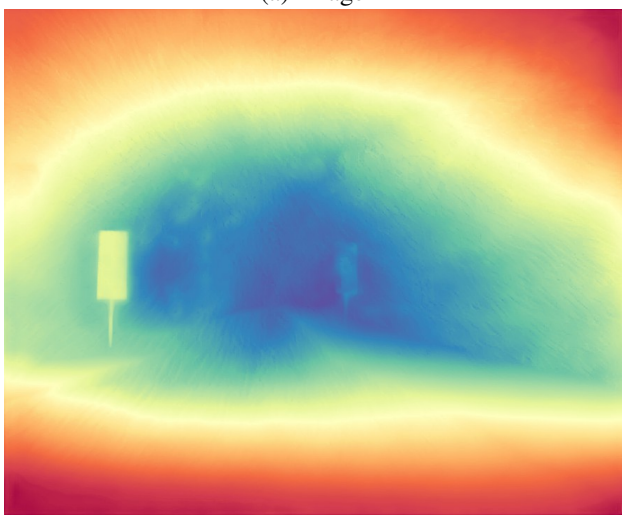
[29] T.-X. Jiang, T.-Z. Huang, X.-L. Zhao, L.-J. Deng, and Y. Wang, "A novel tensor-based video rain streaks removal approach via utilizing discriminatively intrinsic priors," *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 2818–2827, 2017.

[30] W. Ren, J. Tian, Z. Han, A. Chan, and Y. Tang, "Video desnowing and deraining based on matrix decomposition," *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 4210–4219, 2017.

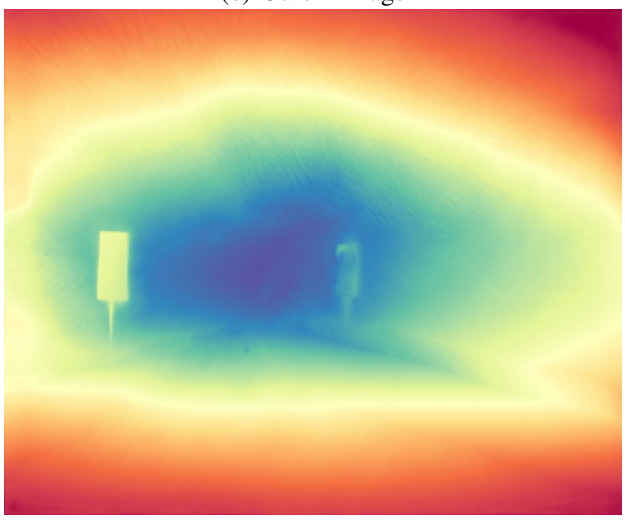
[31] M. Li, Q. Xie, Q. Zhao, W. Wei, S. Gu, J. Tao, and D. Meng, "Video rain streak removal by multiscale convolutional sparse coding," *IEEE*

(a) Image

(b) Ours - Image

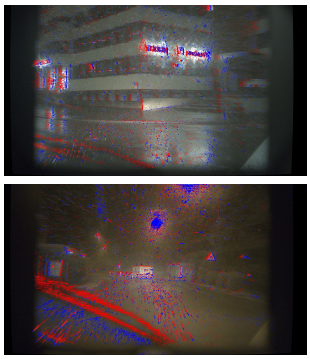


(c) Image-Depth

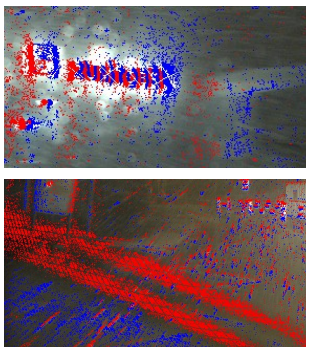


(d) Ours - Depth

Fig. 18: **Qualitative results of our method on downstream task - depth estimation.** We show the input image and depth map from the DSEC-Snow dataset, along with the depth map generated by our method. The depth map is estimated using the input image and events, demonstrating the effectiveness of our approach in leveraging both modalities for improved depth estimation in snowy conditions.



(a) Overlay events on image



(b) Zoomed in

Fig. 19: Checkerboard artifacts due to alignment: Overlaying events on groundtruth images from real-world dataset. The alignment of the events to the RGB image results in a checkerboard pattern in the event stream, which is reflected in the reconstructed image.

Conf. Comput. Vis. Pattern Recog. (CVPR), pp. 6644–6653, 2018.

[32] J. Chen, C.-H. Tan, J. Hou, L.-P. Chau, and H. Li, "Robust video content alignment and compensation for rain removal in a cnn framework," *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 6286–6295, 2018.

[33] J. Liu, W. Yang, S. Yang, and Z. Guo, "Erase or fill? deep joint recurrent rain removal and reconstruction in videos," *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 3233–3242, 2018.

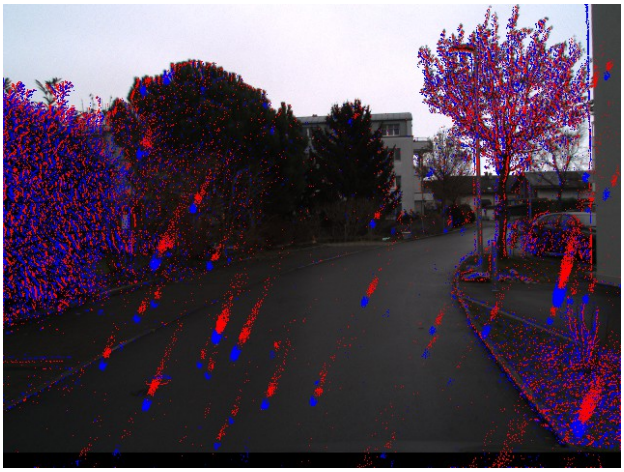
[34] J. Liu, W. Yang, S. Yang, and Z. Guo, "D3r-net: Dynamic routing residue recurrent network for video rain removal," *IEEE Transactions on Image Processing*, 2018.

[35] W. Yang, J. Liu, and J. Feng, "Frame-consistent recurrent video deraining with dual-level flow," *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 1661–1670, 2019.

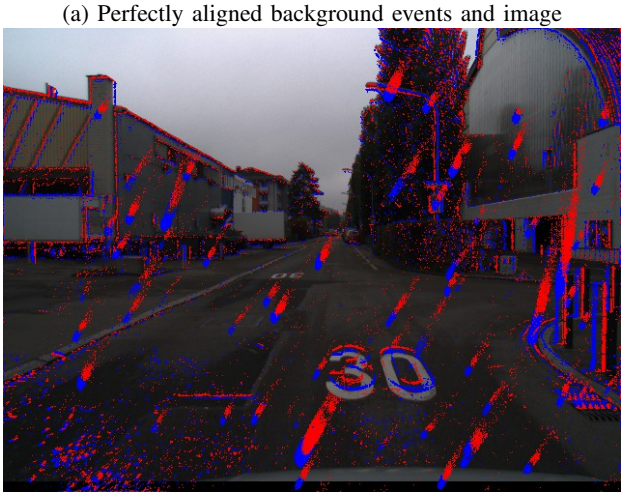
[36] H. Jin, P. Favaro, and S. Soatto, "A semi-direct approach to structure from motion," *The Visual Computer*, vol. 19, no. 6, pp. 377–394, 2003.

[37] Z. Yue, J. Xie, Q. Zhao, and D. Meng, "Semi-supervised video deraining with dynamical rain generator," *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021.

[38] 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022.



(a) Perfectly aligned background events and image



(b) Misaligned background events and image

Fig. 20: Depth-dependent alignment limitations in DSEC dataset: Overlaying events on groundtruth images from DSEC-Snow dataset. The left image shows a scenario where the background events and image is perfectly aligned, while the right image shows a scenario where the road marking ”30” is misaligned between events and images due to depth-dependent alignment limitations in the DSEC dataset.

[39] R. Zou, M. Muglikar, N. Messikommer, and D. Scaramuzza, “Seeing behind dynamic occlusions with event cameras,” 2023.

[40] J. Wang, W. Weng, Y. Zhang, and Z. Xiong, “Unsupervised video deraining with an event camera,” *Int. Conf. Comput. Vis. (ICCV)*, pp. 10 831–10 840, October 2023.

[41] A. Photoshop, https://www.adobe.com/products/photoshop.html.

[42] K. Garg and S. K. Nayar, “Vision and rain,” *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 3–27, 2007.

[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[44] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6628106

[45] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, “High speed and high dynamic range video with an event camera,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[46] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, “Benchmarking single-image dehazing and beyond,” *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2018.

[47] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision transformers for dense prediction,” *ArXiv preprint*, 2021.

[48] T. Delbruck, Y. Hu, and Z. He, “V2E: From video frames to realistic DVS event camera streams,” *arXiv e-prints*, 2020.

[49] T. Finateu, A. Niwa, D. Matolin, K. Tsuchimoto, A. Mascheroni, E. Reynaud, P. Mostafalu, F. Brady, L. Chotard, F. LeGoff, H. Takahashi, H. Wakabayashi, Y. Oike, and C. Posch, “A 1280x720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86μm pixels, 1.066geps readout, programmable event-rate controller and compressive data-formatting pipeline,” in *IEEE Intl. Solid-State Circuits Conf. (ISSCC)*, 2020.

[50] M. Muglikar, S. Somasundaram, A. Dave, E. Charbon, R. Raskar, and D. Scaramuzza, “Event cameras meet spads for high-speed, low-bandwidth imaging,” *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–12, 2025.

[51] T. Brödermann, D. Bruggemann, C. Sakaridis, K. Ta, O. Liagouris, J. Corkill, and L. Van Gool, “Muses: The multi-sensor semantic perception dataset for driving under uncertainty,” *arXiv preprint arXiv:2401.12761*, 2024.

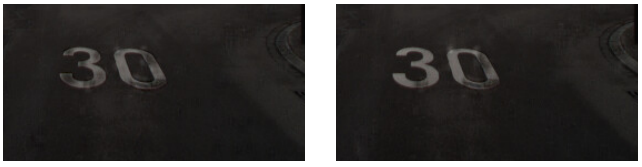
[52] Z. Teed and J. Deng, “RAft: Recurrent All-Pairs Field Transforms for Optical Flow,” *Eur. Conf. Comput. Vis. (ECCV)*, 2020.

(a) Image (b) Ours

Fig. 21: **Qualitative comparison on the MUSES Dataset [51] with rain and snow occlusions.** (a) Input images affected by rain or snow occlusions, and (b) results from our method. The examples show the effectiveness of our approach in handling adverse weather occlusions across different datasets.

(a) E2VID [45] (b) S2VD [37] (c) Restormer [1] (d) Ours (e) Groundtruth

Fig. 22: **Comparing the image reconstruction, optical flow estimation** for event-only baseline E2VID, video baseline S2VD, and image baseline Restormer with our method on the Slider-Snow dataset.



(a) Naive Fusion (b) Ours

Fig. 23: **Qualitative comparison of naive fusion and our proposed method** Our method effectively leverages the spatio-temporal features from events to reconstruct high-quality images, while the naive fusion approach results in several artifacts and loss of detail.

(a) Event Only (b) Image Only (c) Ours

Fig. 24: **Qualitative comparison of event only, image only, and our proposed method** The event-only method lacks the low-frequency structural information present in image frames, while the image-only method struggles to recover motion cues in the presence of severe occlusions. Our method effectively fuses both modalities, resulting in better reconstruction quality.