This ICCV Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MIPI 2025 Challenge on Deblurring for Hybrid EVS Camera: Methods and Results

Challenge Organizers
Yaqi Wu† Zhihao Fan† Hirotaka Shinozaki Frank Zhang Xander Li Alexis Baudron
Wenbin Feng Shuang Zhao
Challenge Participants
Jin Han* Cheng Li* Yihui Shi* Dehua Song Zheng Chen Wenbo Li Fenglong Song
Yihong Leng Siming Zheng Peng-Tao Jiang Linxiao Shi Jinwei Chen Bo Li Jiaojiao Li
Cheng Li Jinao Song Yan Chen Yajing Wei Yuqiang Yang Jian Tang Long Bao
Heng Sun Yang Qiu Kean Liu Senyan Xu Zhijing Sun Jiaying Zhu Chengjie Ge
Xingbo Wang Yidi Liu Xin Lu Xueyang Fu Zheng-Jun Zha Qinglin Liu Wei Yu
Yunfan Lu Xiangming Wang Shuqi Ren Haijin Zeng Jinchao Li Shiyang Zhou
Yongyong Chen Qichao Dong Junqi Lu

Abstract

The demand for computational photography on mobile devices is growing, pushing the development of better image sensors and algorithms. To support this progress and connect industry with academia, we are launching the forth MIPI Challenge in conjunction with ICCV 2025, following successful workshops at CVPR 2024 and CVPR 2023. This challenge includes three tracks on new image sensors and imaging methods. This paper reviews the Deblurring Challenge for Hybrid EVS Cameras at MIPI 2025. With 130 participants from both industry and academia, the challenge attracted a wide range of innovative solutions and pushed reconstruction quality to a new level. This paper presents a detailed overview and analysis of the solutions developed for this challenge. More information is available at: https://mipi-challenge.org/MIPI2025

1. Introduction

Motion blur remains a persistent challenge for conventional frame-based cameras when imaging fast-moving scenes. The usual remedy—shortening the exposure time to “freeze” motion—inevitably degrades the signal-to-noise ratio (SNR) and overall image quality, a limitation that becomes especially severe in low-light environments. Recent progress in hybrid visual sensing offers a compelling alternative. The hybrid vision sensor, also referred to as the Hybrid Vision Sensor (HVS) [1] or Hybrid Event-Vision Sensor (Hybrid-EVS) [9], stands at the forefront of this progress, as shown in Fig. 1. A key distinguishing feature of this sensor is its intrinsic capability to simultaneously acquire high dynamic range (HDR) RGB image information alongside high-temporal-resolution, low-latency event stream data, eliminating the need for additional inter-sensor calibration.

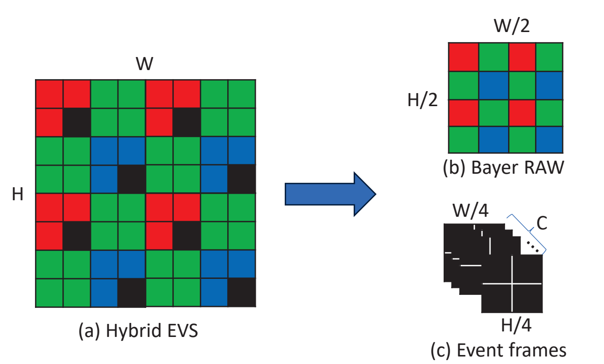

Figure 1. An introduction to Hybrid-EVS. Sub-figure(a) is Quad Bayer, where the black positions are event pixels. It is then processed into sub-figure(b) Bayer RAW and sub-figure(c) Event Frames during the exposure time.

Notably, the EVS operates asynchronously, registering micro-second brightness changes at each pixel. By decoupling image capture from exposure-time constraints, it largely suppresses motion blur and delivers sharper imagery even in fast, low-light scenes. Motivated by the unique ca-

† These organizers contributed equally.
* Both authors contributed equally.

pabilities of hybrid event-vision sensing and its promise for motion-robust imaging, we launch a new benchmark to accelerate algorithmic advances in this area.

We are organising this challenge in conjunction with the **4th Mobile Intelligent Photography and Imaging (MIPI) Challenge**, to be held at ICCV 2025. Building on the success of previous editions [6, 19, 23, 24, 30], the 2025 challenge seeks efficient, high-performance algorithms that recover high-quality RGB frames by jointly exploiting asynchronous event streams and their corresponding blurred images. To cover complementary yet practical imaging problems, MIPI 2025 is divided into three competition tracks:

- **Track 1 - Detailed Image Quality Assessment** Achieve detailed image quality assessment (IQA) with three subtasks: fine-grained quality grounding, detailed quality perception, and descriptive quality reasoning.
- **Track 2 - Deblurring for HVS/HybridEVS Camera** is to restore high-quality RGB frames by jointly leveraging captured event data and blurred RGB images.
- **Track 3 - Aberration Correction for Mobile Camera** is to restore high-quality RAW images by correcting spatially varying optical aberrations introduced during lens manufacturing.

2. MIPI 2025 Deblurring for HVS/HybridEVS Camera

To foster the development of accurate and efficient deblurring algorithms, we release a high-quality dataset for training and testing, along with standard evaluation metrics. This challenge focuses on deblurring for Hybrid-EVS cameras and aims to push the state of the art in event-guided restoration.

2.1. Problem Definition

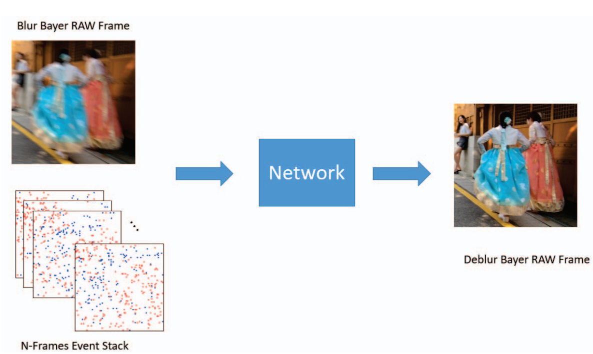

Figure 2. Illustrations of Hybrid-EVS deblurring track.

As illustrated in Figure 2, deblurring for Hybrid-EVS seeks to reconstruct a sharp image from a blurred Bayer Raw frame and the asynchronous events recorded during the same exposure interval. Given a blurred Bayer Raw frame $\mathbf{I}_{\text{in}} \in \mathbb{R}^{H\times W}$ and its corresponding N-frame event stack $\mathbf{E}_{\text{in}} \in \mathbb{R}^{N\times H\times W}$, the deblurring network $\mathcal{F}$ produces a restored sharp Bayer Raw frame $\mathbf{I}_{\text{out}} \in \mathbb{R}^{H\times W}$. We define the reconstruction task using the following formula:

$$\mathbf{I}_{\text{out}} = \mathcal{F}\big(\mathbf{I}_{\text{in}}, \mathbf{E}_{\text{in}}\big). \tag{1}$$

Figure 3. Some sample images from training dataset.

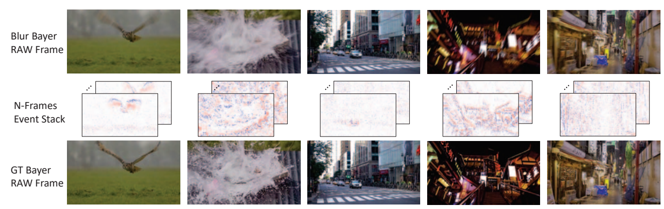

Figure 4. Some sample images from test dataset.

2.2. Datasets

For this competition, we introduce a new dataset comprising 1,100 data pairs, which are partitioned into training (1,000 pairs), validation (50 pairs), and testing (50 pairs) sets. Each pair consists of three components: (i) a 10-bit blurry Bayer RAW frame with a resolution of 720×1280; (ii) a corresponding N-frame, 2-bit event stack of size 360×640 captured during the exposure period; and (iii) the ground-truth 10-bit sharp Bayer RAW frame at the same 720×1280 resolution. To ensure a robust and comprehensive evaluation, the test set is specifically designed to feature a wide array of challenging scenarios. As illustrated in Fig. 4, these include high-speed wildlife, chaotic fluid dynamics, and complex urban landscapes under both bright daylight and low-light night conditions. This diversity guarantees that algorithms are benchmarked under varied and demanding situations.

2.3. Evaluation

The performance of submitted models is evaluated by quantitatively comparing the recovered images against their corresponding ground-truth counterparts. For this purpose, we employ two widely-adopted metrics: the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity (SSIM)

Table 1. Final test results of MIPI 2025 challenge on the Deblurring for HVS/HybridEVS Camera. PSNR and SSIM are computed between the test results and ground truth. The running time of input of 720×1280 was measured. The measurement was taken on an NVIDIA Geforce GTX 3090.

Rank	Team	PSNR	SSIM	Time (s)
1	NoahTCV	**33.076**$_{(1)}$	**0.9260**$_{(1)}$	5.275$_{(6)}$
2	vivo Camera Research	33.068$_{(2)}$	0.9253$_{(2)}$	7.823$_{(7)}$
3	MiVideoDeblur	33.017$_{(3)}$	0.9247$_{(3)}$	5.023$_{(5)}$
4	404NotFound	32.897$_{(4)}$	0.9229$_{(4)}$	1.832$_{(4)}$
5	IVISLAB	32.324$_{(5)}$	0.9171$_{(5)}$	1.640$_{(3)}$
6	Yunfan Lu	31.954$_{(6)}$	0.9093$_{(7)}$	0.816$_{(2)}$
7	suospring	31.941$_{(7)}$	0.9105$_{(6)}$	0.312$_{(1)}$
8	Attempt++	16.668$_{(8)}$	0.5613$_{(8)}$	/

index [21]. Participants can use the scores from these metrics on their submissions to guide the optimization of their models.

2.4. Challenge Phases

The challenge was structured into the following three phases:

1. **Development Phase:** Participants were provided with the training data and baseline code. This allowed them to develop and train their models and locally evaluate computational performance (e.g., runtime).
2. **Validation Phase:** Participants could submit their trained models to an online evaluation server. The server automatically evaluated submissions on the validation set and refreshed a public leaderboard, allowing teams to benchmark their methods and track their progress.
3. **Final Testing Phase:** For the final evaluation, participants submitted their source code, final model, and a methodology factsheet. The final ranking was then determined by the performance of these solutions on a sequestered test set.

3. Challenge Results

Among 130 registered participants, 8 teams successfully submitted their results, code, and factsheets in the final test phase. Table 1 reports the final test results and rankings of the teams. Only one team employed extra datasets to augment training procedures. The methods evaluated in Table 1 are briefly described in Section 4 and the team members are listed in Appendix. Finally, the NoahTCV team is the first place winner of this challenge, while vivo Camera Research team win the second place and MiVideoDeblur team is the third place, respectively.

The top-performing teams achieved significantly higher PSNR and SSIM values, demonstrating superior reconstruction quality. Furthermore, our empirical validation of participant results revealed that enhanced performance consistently corresponds to increased model runtime. Future competitions will evaluate both inference efficiency (e.g., latency, model size) and image quality (e.g., PSNR, SSIM) to enable smartphone deployment of hybrid EVS - deblur image enhancement functions on smartphones.

We found many exciting and innovative solutions from the participating teams. Notably, all teams designed their network architectures through the fusion of APS and EVS modalities, fully leveraging information from both modalities to address the deblurring task. Approximately two-thirds of participating teams adopted the Transformer Block or its variants as their foundational architecture, while the remaining teams utilized traditional CNN architectures, reflecting the coexistence of emerging Transformer-based architectures and established CNN-based approaches in the field. Furthermore, two teams employed an ensemble strategy, which involved either integrating outputs from multiple models or aggregating results generated by the same model with different patch sizes, both of which contributed to their favorable rankings. Additionally, one team first trained models from on traditional RGB image deblurring datasets and subsequently switched to fine-tuning on Hybrid EVS data. This approach not only enhanced deblurring performance for the EVS component but also improved sharpness in non-motion event regions, presenting a valuable technical insight. Interestingly, one team stood out with an innovative approach: it adopted a unified image degradation model optimized via the proximal gradient method, which learns task-specific degradation matrices for different tasks while demonstrating exceptionally fast inference performance. These diverse and creative approaches will effectively facilitate the development of Hybrid EVS applications.

4. Methods

4.1. NoahTCV

This team proposes a hybrid ensemble framework named Noah-HEF, as shown in Figure 5. The proposed framework consists of several base models, where each model focuses on different levels and types of motion blur. Specifically, this team incorporates base models including: TEIDNet [18], XRestormer [5], and NAFNet [2] to process different levels of blur. The overall pipeline includes self-ensemble operation for each base model, and a cross ensemble operation that merges the output of each base model to achieve the final sharp results.

Framework This team applies a 2-stage training strategy to optimize the base models. The first stage is for pretraining. During the first stage, since the dataset of blurry-sharp image pairs are easier to obtain compared to event-based

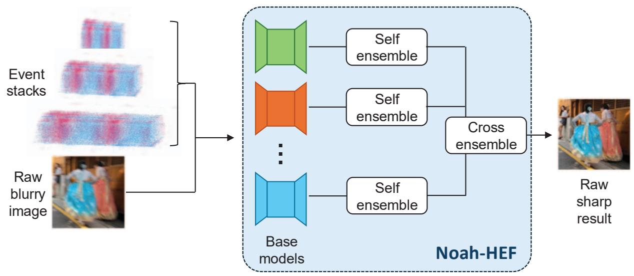

Figure 5. The pipeline of Hybrid Ensemble Framework: Noah-HEF. The final sharp results are achieved by a cross ensemble operation of different base models.

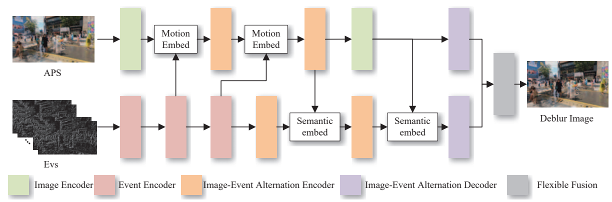

Figure 6. Framework of MotionDissolve proposed by vivo Camera Research.

deblurring dataset, this team uses blurry-sharp image pairs without event data as input. This team collects blurry-sharp image pairs from LSDIR dataset [10], GoPro dataset [14], Adobe240 dataset [16], and RBI dataset [28] for the first stage pretraining. The event stacks are set to zero-tensors during at this stage. For the second stage, this team finetunes the pretrained base models using the official training dataset from the HybridEVS Camera MIPI-challenge. During the second finetuning stage, the base models learn to merge the sharp temporal information from event stacks into the blurry raw images for effective deblurring. For event stacks, this team applies the multi-scale stacking strategy. Specifically, for each event stream, this team slices the long-term, mid-term, and short-term to form 3 event stacks with different numbers of channels, which focus on different levels of motion perception and high-frequency textures.

The proposed Noah-HEF framework capitalizes on both self-ensemble and cross-ensemble techniques to enhance robustness and accuracy. Within each base model, this team implements a self-ensemble strategy, which involves aggregating predictions from 4 inference-time augmentations (original, 180° rotation, 90° rotation+flip, and 270° rotation + flip), thereby stabilizing its output and reducing variance. Following this intra-model ensemble, a cross-ensemble mechanism is employed. The predictions from different base models potentially specialized or trained under distinct configurations, which are fused to form the final sharp results. Specifically, this team refer to the non-reference evaluation metrics like MUSIQ and DACLIP [13] as a filter to select the best reconstruction images for ensemble. This two-tiered ensemble approach allows the system to benefit not only from the diversity within each model's prediction space but also from the heterogeneity across models. Compared to relying on a single sub-network, the ensemble strategy significantly improves performance.

Implementation Details The overall framework is implemented using PyTorch and trained on HUAWEI Ascend 910B. During the first pretraining stage, a batch size of 16 is utilized. The base models are optimized for 250k iterations using AdamW optimizer with a learning of 1e-4. For the input images, this team applies a degradation pipeline to get raw images from the sRGB images. For the second finetuning stage, all of the base models are optimized for 200k iterations with a learning rate of 5e-4. This team crops the images and event stacks into patch size of 256×256. The Raw images are unshuffled into a 4-channel RGGB input. The event stacks are binned into 4, 8, and 16 channels to get the multi-scale event stacks. This team uses L1 loss in the raw domain during the first pretraining stage. For finetuning stage, they combines the Raw domain L1 loss and sRGB domain L1 loss to optimize the base models. Data augmentation strategy including flip, rotation, and reverse events is applied during the finetuning stage.

4.2. vivo Camera Research

This team proposes MotionDissolve, an Image-Event Alternation Deblurring Network specifically designed for deblurring in HVS/Hybrid-EVS cameras. The core objective is to fully exploit the complementary characteristics of APS images and EVS event streams for improved deblurring performance.

Network Architecture As illustrated in Figure 6, MotionDissolve introduces a cross-modality fusion strategy that leverages both early- and late-stage interactions, inspired by the asymmetric fusion mechanism proposed in [25]. The network employs two dedicated encoder branches to process APS and EVS inputs independently: 1) The APS branch begins with a hierarchical image encoder, followed by two image-event alternation encoders, and ends with a final image encoder, producing four levels of multi-scale features. 2) The EVS branch consists of three event-specific encoders, interleaved with two image-event alternation encoders, enabling hierarchical feature extraction tailored to the event domain.

Instead of using cross-attention, MotionDissolve designs two modality-specific self-attention mechanisms: motion-aware and image-aware self-attention, aiming to capture motion semantics while promoting mutual guidance between modalities. These mechanisms are designed to respect the distinct characteristics of APS and EVS data, enabling the extraction of complementary information and their progressive alignment in the decoder via hierarchical fusion. Moreover, a dual-path embedding strategy is adopted to enhance both low-level motion and high-level semantic representation learning: 1) Motion Embedding (ME): Enhances the APS stream by injecting motion priors derived from the EVS branch. 2) Semantic Embedding (SE): Enhances the EVS stream with semantic context from the APS branch. In detail, giving the middle representation from APS and EVS branches, $\mathbf{F}_{aps}$ and $\mathbf{F}_{evs}$, a spatial attention and a channel-wise attention are performed on $\mathbf{F}_{aps}$ to obtain the spatial-wise content $\mathbf{F}^{spa}_{aps}$ and the motion-wise correlation $\mathbf{F}^{mot}_{aps}$. Then, the semantic prior $\mathbf{F}^{sem}_{aps}$ can be obtained by $\mathbf{F}^{spa}_{aps} \cdot \mathbf{F}^{mot}_{aps}$. To achieve an injection to $\mathbf{F}_{evs}$, $\mathbf{F}^{sem}_{aps}$ is concatenated with $\mathbf{F}_{evs}$, followed by a spatial-motion joint learning as a spatial-spectral domain learning (SDL) in [8]. Subsequently, multi-scale features from both branches are decoded through separate image-event alternation decoders, which refine representations at each scale. Finally, the deblurred output is generated via an adaptive fusion of the decoded features from both modalities.

Implementation Details The MotionDissolve network is trained using a progressive two-stage strategy on NVIDIA L40 (48GB) and H20 (96GB) GPUs. The dataset is partitioned into 950 self-training samples and 50 self-validation samples. Each training sample consists of an APS image and its corresponding event representation, cropped into spatial patches of size $256 \times 256 \times 1$ (APS) and $128 \times 128 \times 17$ (EVS), respectively. In the first stage, only the self-training set is utilized. The initial learning rate is set to 2×10^{-4} and gradually decayed to 2×10^{-6} using the Cosine Annealing strategy. In the second stage, this team incorporate both the self-training and self-validation samples, along with additional training pairs derived from the GoPro dataset for enhanced generalization. The initial learning rate is set to 5×10^{-6} and decayed to 1×10^{-8} following the same cosine scheduling.

4.3. MiVideoDeblur

MiVideoDeblur This team proposes EFSwinIRAttn for event-based image deblurring. Unlike general Swin Transformer methods that only use blurred images for deblurring, this method efficiently incorporates event information to more accurately capture moving-object features in blur images. EFSwinIRAttn's architecture is shown in Fig 7. The RAW and the event image are concatenated and then input into the network.

Network Architecture Different from the original SwinIR[12], EFSwinIRAttn introduces an FFM (Feature Fusion Module) that uses learnable weights to effectively combine the features from six RSTB (Residual Swin Transformer Block) modules of SwinIR. For each STL (Swin Transformer Layer) within each RSTB, the MDTA (Multi-Dconv Transpose Attention)[3] module and the CAB (Channel Attention Block)[4] module are added in parallel to the original Window Attention, to

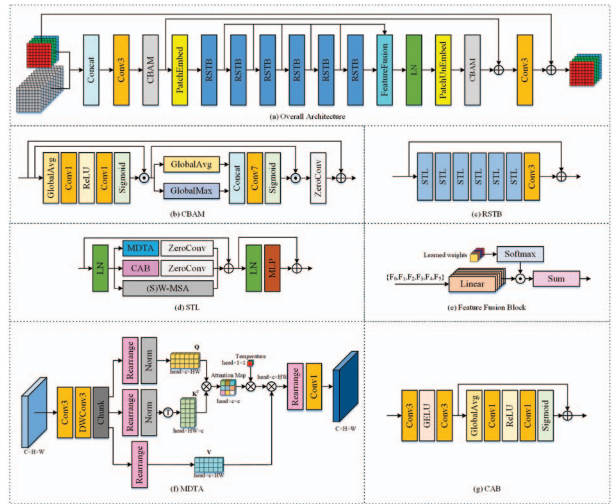

Figure 7. The architecture of EFSwinIRAttn proposed by MiVideoDeblur.

capture more spatial and channel features. Additionally, a CBAM (Convolutional Block Attention Module)[22] is introduced after the first convolutional layer and before the last convolutional layer of the SwinIR to enhance the robustness and accuracy of feature representation in both the channel and spatial dimensions. Finally, new attention modules are gradually added to the SwinIR model using zero convolution, ensuring that the network's performance is efficiently and steadily improved while maintaining its current level of performance.

Implementation Details This team's code is based on BasicSR[20]. Before training, the event data is transformed using the SCER method in EFNet[17] to improve the efficiency of model training. The model is trained in 4 stages. In each stage, EMA (Exponential Moving Average) technology is used to smooth the weight updates. The optimizer used for training is AdamW, with an input size of 128x128. The input is randomly flipped horizontally or vertically during training, and the model is gradually optimized in each stage. A detailed description of the staged training is as follows:

Firstly, this team introduces the FFM (Feature Fusion module) on top of SwinIR to aggregate features from each RSTB module. The model contains 6 RSTB modules, each with a window size of 8. The model is trained for a total of 400K steps with a batch size of 8. The loss function is L1 loss, and the feature representation channel number is 6. The initial learning rate is 2E-4, and the learning rate decay strategy is MultiStepLR, with a 50% reduction at 100K, 200K, 250K, and 275K steps.

Secondly, building on the model from stage one, this team introduces the CAB into the STL modules using zero convolution. The batch size for this stage is 4. We train the model for 120K steps using PSNR loss. The initial learning rate is 2E-4, and the learning rate decay strategy is MultiStepLR, with a 50% reduction at 50K and 100K steps. The other training configurations are the same as in stage one.

Thirdly, building on the model from stage two, this team introduces the CBAM module before the first and last convolutional layers and the MDTA module into the STL modules using zero convolution. The event representation channel number is changed to 8. The model is trained for 180K steps with an initial learning rate of 3E-5. The training loss is a combination of PSNR loss, Outlier-Aware[7] loss, and Cosine Similarity loss. The learning rate decay strategy is MultiStepLR, with a 50% reduction at 100K and 150K steps.

Lastly, Building on the model from stage three, this team changes the event representation channel number to 10 and set the STL window size to 16. The model is trained for a total of 285K steps with an initial learning rate of 2E-4. The learning rate decay strategy is MultiStepLR, with a 50% reduction at 100K, 150K, 200K and 250K steps.The loss function is the same as in stage three.

All models need to be trained on a single A100 80GB GPU. After four stages of training and model structure optimization, the final model training is completed. this team randomly selected 50 pairs of data as the local validation set. The model with the highest PSNR on the validation set is finally selected for testing.

4.4. 404NotFound

To address the deblurring problem, this team employed Restormer[26] ,which adopts an efficient Transformer-based encoder–decoder architecture combined with a U-Net-style design for high-resolution image restoration. The overall design balances global modeling capability and computational efficiency, making it suitable for various image restoration tasks.Their method leverages both blurry RAW images and stacked voxel-grid event data as inputs. The two are concatenated to form a 28-channel input (4 channels for RAW image and 24 channels for events).During training, this team use a combination of L1 loss and Focal Frequency Loss. The learning rate scheduler is CosineAnnealingRestartCyclicLR, and the optimizer is AdamW. The model is trained from scratch for approximately 500,000 iterations. In the early epochs, the network is trained on small image patches (crop size of 160) with a large batch size (batch size:4), and gradually shifts to larger patches with smaller batch sizes in later epochs. This training strategy helps Restormer learn contextual information from larger images, leading to improved performance during testing.During testing, Inference is performed using crop sizes of 184 and 360 respectively, and the results are then ensembled (see Figure 8).

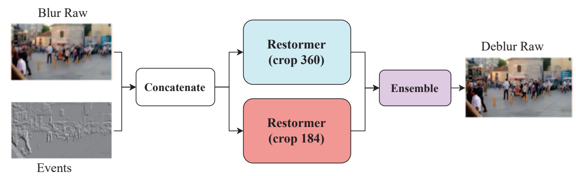

Figure 8. Overview of the method: This method feeds the blurry RAW image and corresponding event map into a Restormer network, employing dual-scale inference with crop sizes of 184 and 360 respectively. The results are then ensembled to generate the final sharp image.

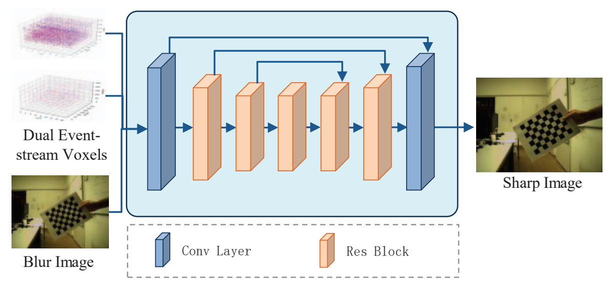

Figure 9. The architecture of DEDNet.

4.5. IVISLAB

To address image deblurring, team IVISLAB introduces the Dual Event-stream Image Deblurring Network (DEDNet). As illustrated in Figure 9, DEDNet converts consecutive events into event voxels at two temporal scales, enabling it to extract motion information from blurred images and capture fine edges for reconstructing clear images.

Network Architecture DEDNet utilizes an encoder-decoder architecture designed to process both images and dual-stream event voxels, aiming to estimate a deblurred raw image. Specifically, when deblurring the raw image at frame t, DEDNet considers the long-term event stream around frame t to aid in motion perception. As such, it voxelizes event data from frame $t - T_l$ to frame $t + T_l$ into a b-bin event voxel, $V_{l,t}$. Simultaneously, the short-term event stream around frame t is used to help reconstruct high-frequency textures, and event data from frame $t - T_s$ to frame $t + T_s$ is voxelized into a b-bin event voxel, $V_{s,t}$. The event voxels $V_{l,t}$, $V_{s,t}$, and $V_{m,t}$, along with the blurred raw image I_b, are concatenated and passed through the network. To fuse features from the raw image and event voxels, DEDNet employs convolutional layers and residual blocks to generate the fused feature representations. Finally, convolutional layers are used to predict the deblurred raw image I_t.

Loss Function To train DEDNet, a reconstruction loss function $\mathcal{L}_r$ is defined for the estimated deblurred raw im-

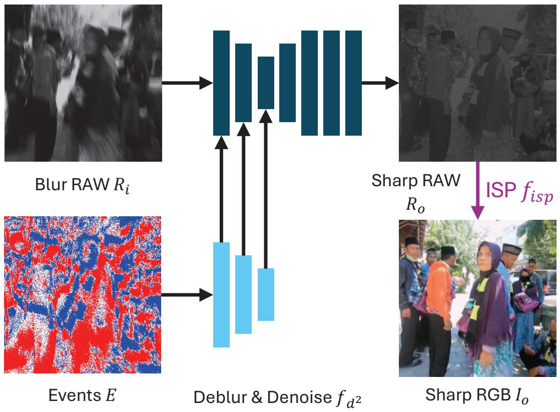

Figure 10. A blurry RAW image and the corresponding event map are input to a joint deblurring and denoising network. The network outputs a sharp RAW image, which is then processed by the ISP to generate the final sharp RGB result.

age I_t as follows:

$$\mathcal{L}_r = \lambda_1 \, \mathrm{L}_1(I_t, I_t^{gt}) + \lambda_2 \, \mathrm{L}_2(I_t, I_t^{gt}) \tag{2}$$

Here, $\lambda_1 = 1$ and $\lambda_2 = 4$ are the coefficients that balance the loss terms. $\mathrm{L}_1(\cdot,\cdot)$ represents the mean absolute error, and $\mathrm{L}_2(\cdot,\cdot)$ denotes the mean squared error. I_t^{gt} refers to the ground truth raw image at frame t.

Implementation Details DEDNet is implemented using PyTorch on four Nvidia A100 GPUs. During training, a batch size of 16 is used with input data dimensions of 320×320 pixels. The network weights are optimized over 2000 epochs using the AdamW optimizer, with an initial learning rate of 2×10^{-5}. A cosine annealing scheduler is applied to gradually decay the learning rate. To prevent overfitting, data augmentation techniques, such as random flipping and rotation, are applied.

4.6. Yunfan Lu

As illustrated in Fig. 10, their method is designed to jointly remove both blur and noise in the RAW domain. **Motivation**: The input consists of a blurry RAW image $R_i \in \mathbb{R}^{H\times W}$ and an event map $E \in \mathbb{R}^{B\times H\times W}$, and the output is a clean RAW image R_o. It is important to note that the ISP stage does not perform denoising; therefore, their network try to address both denoising and deblurring simultaneously. This joint approach is necessary, since blur and noise often interact, and optimizing for only one will often hurt the other [11, 29]. Therefore, this team design a unified training method that learns both tasks together.

Method: To implement this, they use a well-established deblurring backbone (EFNet) [17]. The network takes both the blurry RAW image and the event map as input. The event stream is processed by a feature extraction sub-network and fused with the RAW branch at multiple stages, which improves spatio-temporal information integration. The network is trained to map (R_i, E) to R_o as $R_o = f_{d^2}(R_i, E; \theta)$, where f_{d^2} denotes the joint deblurring and denoising network and θ are its trainable parameters. To improve robustness to real sensor noise, this team synthesize a variety of physical noise types during training. The overall RAW image degradation is modeled as:

$$R_i = \mathcal{Q}\big(\mathcal{N}_{\text{row}}(\mathcal{N}_{\text{read}}(\mathcal{P}(R_{gt}) + \mathcal{N}_{\text{dc}}))\big),$$

where $\mathcal{P}(R_{gt})$ is photon shot noise modeled by a Poisson process; $\mathcal{N}_{\text{read}}$ is read noise, $\mathcal{N}_{\text{dc}}$ is dark current bias, $\mathcal{N}_{\text{row}}$ is row noise, and $\mathcal{Q}$ is quantization noise. This augmentation strategy allows the network to learn not only deblurring but also the removal of real sensor noise, which is especially important for RAW data. They optimize the network using the loss $\mathcal{L} = |R_o - R_{gt}|_1$. The final output R_o is passed to the ISP function f_{isp} to produce the RGB image $I_o = f_{\text{isp}}(R_o)$. Since the ISP does not remove noise, our network must generate a clean RAW output.

Implementation Details: All experiments are conducted on an NVIDIA A40 GPU. During training, this team divide each image into 256×256 patches with a batch size of 1. They use the Adam optimizer with an initial learning rate of 0.0005 and train the model for 200 epochs. To improve generalization, they apply data augmentation methods such as random cropping, flipping, and rotation. As a result, the total training time is about 18 hours.

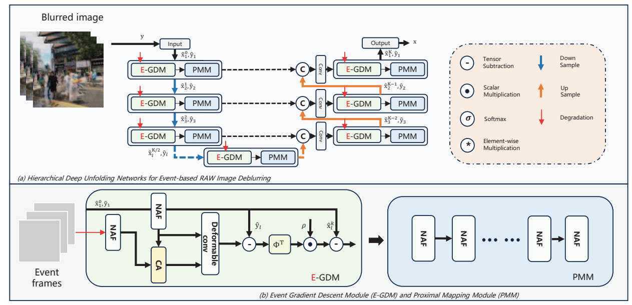

Figure 11. Overview of the method, an hierarchical DUN for Event-based deblurring. The IR process involves a $\mathbf{K}$-stage DUN with l levels, projection and back projection. Each stage utilizes a Event-based GDM (E-GDM) and PMM to adaptively handle degraded inputs and maintain multi-level information across stages.

4.7. suospring

The architecture of the team suospring is based on the VLUNet [27] and HCDNet [15]. Figure 11 shows the main framework of the method.

Network Architecture: This method adopts optimization in feature space and preserves multi-level degraded information by downsampling and upsampling $\mathbf{y}$ across levels, with residual connections to transfer original degraded features. For the k-th stage within the l-th level, its hierarchical degradation unfolding procedure is:

$$\hat{\mathbf{z}}_{(l)}^{(k)} = \hat{\mathbf{x}}_{(l)}^{(k-1)} - \rho \Phi^{\mathbf{T}} (\tilde{\Phi}(\hat{\mathbf{x}}_{(l)}^{(k-1)}, E) - \hat{\mathbf{y}}_{(l)}), \tag{3}$$

$$\hat{\mathbf{x}}_{(l)}^{(k)} = \arg\min_{\hat{\mathbf{x}}_{(l)}} \frac{1}{2} \|\hat{\mathbf{x}}_{(l)} - \hat{\mathbf{z}}_{(l)}^{(k)}\|_2^2 + \rho \lambda \Psi(\hat{\mathbf{x}}_{(l)}). \tag{4}$$

For the specific modules in the gradient descent module (GDM) in Eq. 3 and the proximal mapping module (PMM) in Eq. 4, this method adopts the Kernel Deformable Block (KDB) as the Φ operator, which receives the event information (E) and the blurry input to estimate the blurry kernel of the target image, which is called as E-GDM (Event-based GDM). There it should be noted that the E is inited as the shape of $25 \times H/2 \times W/2$ using interpolation algorithms. And the PMM contains several basic NAFNet blocks.

Implementation Details: The training was conducted on four NVIDIA RTX 4090 GPUs. During training, images were initially cropped into 256×256 patches, and the crop resolution was gradually increased as training progressed. The batch size was set to 4 per GPU. The Adam optimizer was used with an initial learning rate of 0.0002, and the model was trained for a total of 400 epochs. The loss function is $\mathcal{L} = 0.5 \times |X - X_{gt}|_1 + 0.5 \times |X - X_{gt}|_2$. The entire training process took approximately one day.

4.8. Attempt++

This team presents a motion deblurring approach leveraging event cameras, using a fine-tuning deblurring Network pipeline (FTDNet) based on the EFNet [17] architecture. This method focuses on data processing for training. It parses the training split of the Hybrid-EVS dataset and initializes the EFNet model using a checkpoint pretrained on the GoPro dataset. The model is then fine-tuned to adapt to real-world blurry data, enabling it to generalize effectively to real-world motion blur scenarios.

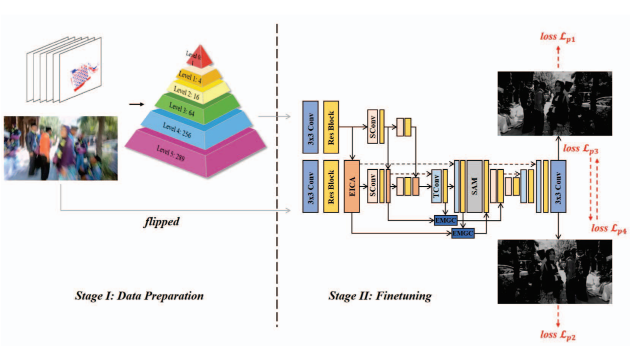

Figure 12. Architecture of FTDNet proposed by Attempt++.

Method: The approach builds upon the EFNet framework, and consists of two main stages: data preprocessing and model fine-tuning.

Stage I: Preprocessing. The training split of the Hybrid-EVS dataset is parsed, where raw `.bin` files contain Bayer APS images and event streams. The APS images are demosaiced and replicated across three channels to align with the RGB input requirements of the model. Event data are voxelized into six temporal bins to form a $(6, H, W)$ grid, effectively capturing spatio-temporal motion cues. The processed data are then stored in a structured `.h5` format for training.

Stage II: Fine-tuning. The EFNet model, pretrained on the GoPro dataset (`EFNet_GoPro.pth`), is fine-tuned using the processed Hybrid-EVS dataset. During fine-tuning, only the affine parameters of the normalization layers are updated, while other weights remain frozen. The objective is formulated as:

$$\theta^* = \arg\min_{\theta \in \Theta_{\text{affine}}} \mathcal{L}_{\text{total}}(\theta; \mathcal{D}_{\text{Hybrid-EVS}}) \tag{5}$$

where Θ_{affine} denotes the set of trainable affine parameters and $\mathcal{D}_{\text{Hybrid-EVS}}$ represents the input dataset.

The underlying network architecture employs a dual-encoder U-Net with symmetric branches, consisting of an APS image encoder and an event voxel encoder. A mid-level fusion module integrates both modalities, followed by an event mask-guided attention mechanism that refines key spatio-temporal features. The training process is supervised by a composite loss:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{mir}} + \beta(\mathcal{L}_{p3} + \mathcal{L}_{p4}) \tag{6}$$

where $\mathcal{L}_{\text{mir}}$ denotes the primary mirror loss, and $\mathcal{L}_{p3}, \mathcal{L}_{p4}$ are auxiliary flipped pixel losses. This loss design, derived from EFNet, encourages symmetric consistency and improves generalization to real-world motion blur.

Implementation Details The model was trained for 200 epochs on a single NVIDIA RTX 4090 GPU with a batch size of 16, using the AdamW optimizer (weight decay=0.05). A warm-up learning rate schedule was applied, starting at 1×10^{-6} and increasing to 2×10^{-4} over two epochs, before finally decaying to 1×10^{-5}.

5. Conclusions

This paper reviews the Deblurring Challenge for Hybrid EVS Cameras, part of the MIPI workshop at ICCV 2025. The challenge attracts 130 participants from industry and academia, yielding innovative solutions for restoring high-quality images from blurred frames and event streams. The competition advances image reconstruction quality and sets a new benchmark for mobile computational photography research.

References

[1] AlpsenTek. Alpsentek. hybrid vision® technology. [EB/OL] https://alpsentek.com, 2023. 1

[2] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 3

[3] Xiangyu Chen, Zheyuan Li, Yuandong Pu, Yihao Liu, Jiantao Zhou, Yu Qiao, and Chao Dong. A comparative study of image restoration networks for general backbone network design. *arXiv preprint arXiv:2310.11881*, 2023. 5

[4] Xiangyu Chen, Xintao Wang, Wenlong Zhang, Xiangtao Kong, Yu Qiao, Jiantao Zhou, and Chao Dong. Hat: Hybrid attention transformer for image restoration. *arXiv preprint arXiv:2309.05239*, 2023. 5

[5] Xiangyu Chen, Zheyuan Li, Yuandong Pu, Yihao Liu, Jiantao Zhou, Yu Qiao, and Chao Dong. A comparative study of image restoration networks for general backbone network design. In *ECCV*, 2024. 3

[6] Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, Qingpeng Zhu, Qianhui Sun, Wenxiu Sun, Chen Change Loy, Jinwei Gu, Shuai Liu, et al. Mipi 2023 challenge on nighttime flare removal: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2852–2862, 2023. 2

[7] Weiran Gou, Ziyao Yi, Yan Xiang, Shaoqing Li, Zibin Liu, Dehui Kong, and Ke Xu. Syenet: A simple yet effective network for multiple low-level vision tasks with real-time performance on mobile device. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12182–12195, 2023. 6

[8] Xiaowan Hu, Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17542–17551, 2022. 5

[9] Kazutoshi Kodama, Yusuke Sato, Yuhi Yorikado, Raphael Berner, Kyoji Mizoguchi, Takahiro Miyazaki, Masahiro Tsukamoto, Yoshihisa Matoba, Hirotaka Shinozaki, Atsumi Niwa, Tetsuji Yamaguchi, Christian Brandli, Hayato Wakabayashi, and Yusuke Oike. 1.22m 35.6mpixel rgb hybrid event-based vision sensor with 4.88m-pitch event pixels and up to 10k event frame rate by adaptive control on event sparsity. In *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 92–94, 2023. 1

[10] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *CVPR*, 2023. 4

[11] Chih-Hung Liang, Yu-An Chen, Yueh-Cheng Liu, and Winston H Hsu. Raw image deblurring. *IEEE Transactions on Multimedia*, 24:61–72, 2020. 7

[12] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 5

[13] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for universal image restoration. In *ICLR*, 2024. 4

[14] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 4

[15] Jaesung Rim, Junyong Lee, Heemin Yang, and Sunghyun Cho. Deep hybrid camera deblurring for smartphone cameras. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 7

[16] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *CVPR*, 2017. 4

[17] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European Conference on Computer Vision (ECCV)*, 2022. 5, 7, 8

[18] Lei Sun, Andrea Alfarano, Peiqi Duan, Shaolin Su, Kaiwei Wang, Boxin Shi, Radu Timofte, Danda Pani Paudel, Luc Van Gool, Qinglin Liu, et al. Ntire 2025 challenge on event-based image deblurring: Methods and results. In *CVPR*, 2025. 3

[19] Qianhui Sun, Qingyu Yang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, Yuekun Dai, Wenxiu Sun, Qingpeng Zhu, Chen Change Loy, Jinwei Gu, et al. Mipi 2023 challenge on rgbw remosaic: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2877–2884, 2023. 2

[20] Xintao Wang, Liangbin Xie, Ke Yu, Kelvin C.K. Chan, Chen Change Loy, and Chao Dong. BasicSR: Open source image and video restoration toolbox. https://github.com/XPixelGroup/BasicSR, 2022. 5

[21] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

[22] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 5

[23] Yaqi Wu, Zhihao Fan, Xiaofeng Chu, Jimmy S. Ren, Xiaoming Li, Zongsheng Yue, Chongyi Li, Shangcheng Zhou, Ruicheng Feng, Yuekun Dai, Peiqing Yang, Chen Change Loy, Senyan Xu, Zhijing Sun, Jiaying Zhu, Yurui Zhu, Xueyang Fu, Zheng-Jun Zha, Jun Cao, Cheng Li, Shu Chen, Liang Ma, Shiyang Zhou, Haijin Zeng, Kai Feng, Yongyong Chen, Jingyong Su, Xianyu Guan, Hongyuan Yu, Cheng Wan, Jiamin Lin, Binnan Han, Yajun Zou, Zhuoyuan Wu, Yuan Huang, Yongsheng Yu, Daoan Zhang, Jizhe Li, Xuanwu Yin, Kunlong Zuo, Yunfan Lu, Yijie Xu, Wenzong Ma, Weiyu Guo, Hui Xiong, Wei Yu, Bingchun Luo, Sabari Nathan, and Priya Kansal. Mipi 2024 challenge on demosaic for hybridevs camera: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1136–1143, 2024. 2

[24] Qingyu Yang, Guang Yang, Jun Jiang, Chongyi Li, Ruicheng Feng, Shangchen Zhou, Wenxiu Sun, Qingpeng Zhu, Chen Change Loy, Jinwei Gu, et al. Mipi 2022 challenge on quad-bayer re-mosaic: Dataset and report. In *European Conference on Computer Vision*, pages 21–35. Springer, 2022. 2

[25] Wen Yang, Jinjian Wu, Leida Li, Weisheng Dong, and Guangming Shi. Asymmetric hierarchical difference-aware interaction network for event-guided motion deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9265–9273, 2025. 4

[26] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 6

[27] Haijin Zeng, Xiangming Wang, Yongyong Chen, Jingyong Su, and Jie Liu. Vision-language gradient descent-driven all-in-one deep unfolding networks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7524–7533, 2025. 7

[28] Zhihang Zhong, Mingdeng Cao, Xiang Ji, Yinqiang Zheng, and Imari Sato. Blur interpolation transformer for real-world motion from blur. In *CVPR*, 2023. 4

[29] Chu Zhou, Minggui Teng, Jin Han, Jinxiu Liang, Chao Xu, Gang Cao, and Boxin Shi. Deblurring low-light images with events. *International Journal of Computer Vision*, 131(5): 1284–1298, 2023. 7

[30] Qingpeng Zhu, Wenxiu Sun, Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, Qianhui Sun, Chen Change Loy, Jinwei Gu, Yi Yu, et al. Mipi 2023 challenge on rgb+ tof depth completion: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2869, 2023. 2

A. Teams and Affiliations

MIPI 2025 HybridEVS Challenge Team
Title: MIPI 2025 Challenge on Deblurring for Hybrid EVS Camera Methods and Results
Members:
Yaqi Wu[1] (wuyaqi16@gmail.com)
Zhihao Fan[2], Hirotaka Shinozaki[1], Frank Zhang[1], Xander Li[3], Alexis Baudron[3], Shuang Zhao[2], Wenbin Feng[2]
Affiliations:
[1] Pentas Vision Pte Ltd, Singapore
[2] AlpsenTek, China
[3] Sony Electronics Inc, America

NoahTCV
Title: Hybrid Ensemble Framework for HybridEVS Image Deblurring
Members:
Jin Han[1](jin.han@huawei.com)
Cheng Li[1], Yihui Shi[1], Dehua Song[1], Zheng Chen[2], Wenbo Li[1], Fenglong Song[1]
Affiliations:
[1] HUAWEI Noah's Ark Lab
[2] Shanghai Jiaotong University

vivo Camera Research
Title: MotionDissolve: Image-Event Alternation Network for Event-based Image Deblurring
Members:
Yihong Leng[1,2](yhleng@stu.xidian.edu.cn)
Siming Zheng[2], Peng-Tao Jiang[2], Linxiao Shi[2], Jinwei Chen[2], Bo Li[2], Jiaojiao Li[1]
Affiliations:
[1] Xidian University
[2] vivo Mobile Communication Co., Ltd

MiVideoDeblur
Title: Event Fusion Deblurring via SwinIR-based Hybrid Attention Mechanism
Members:
Cheng Li (licheng8@xiaomi.com), Jinao Song, Yan Chen, Yajing Wei, Yuqiang Yang, Jian Tang, Long Bao, Heng Sun
Affiliations:
Xiaomi Inc., China

404NotFound
Title: Event-Aware Restormer for RAW Image Deblurring
Members:
Yang Qiu (qy1227@mail.ustc.edu.cn)
Kean Liu, Senyan Xu, Zhijing Sun, Jiaying Zhu, Chengjie Ge, Xingbo Wang, Yidi Liu, Xin Lu, Xueyang Fu, Zheng-Jun Zha
Affiliations:
University of Science and Technology of China, Hefei, China.

IVISLAB ***Title:*** Dual Event-stream Image Deblurring Network
Members:
Qinglin Liu[1] (qlliu@hit.edu.cn), Wei Yu[2]
Affiliations:
[1] Harbin Institute of Technology, Weihai
[2] Tsinghua University

Yunfan Lu
Title: Joint denoising and deblurring based on event in RAW domain
Members: Yunfan LU (ylu066@connect.hkust-gz.edu.cn)

Affiliations: HKUST(GZ)

suospring
Title: Hierarchical Deep Unfolding Networks for Event-based RAW Image Deblurring
Members: Xiangming Wang[1] (xmwang28@gmail.com), Shuqi Ren[1], Haijin Zeng[2], Jinchao Li[1], Shiyang Zhou[1], Yongyong Chen[1]
Affiliations:
[1] Harbin Institute of Technology, Shenzhen
[2] Harvard University

Attempt++
Title: Finetuning Based Event Motion Deblurring
Members: Qichao Dong, Junqi Lu
Affiliations:
University of Macau