Event-Guided Super-Resolving Blurry Image via Asymmetric Integral Driven Consistency

Chi Zhang[1], Xiang Zhang[2], Lei Yu[3], Gui-Song Xia[3], Yuming Fang[4], Wenhan Yang[1]*

[1]Peng Cheng Laboratory
[2]ETH Zürich
[3]School of Artificial Intelligence, Wuhan University
[4]School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics
zhangch21@pcl.ac.cn, xiangz.ethz@gmail.com, ly.wd@whu.edu.cn, guisong.xia@whu.edu.cn,
fa0001ng@e.ntu.edu.sg, yangwh@pcl.ac.cn

Abstract

*S*uper-*R*esolution from a *B*lurry low-resolution image (SRB) constitutes a severely ill-posed inverse problem. Current learning-based SRB approaches primarily rely on synthetic, well-labeled paired datasets to regularize solution spaces, yet they exhibit limited generalizability in practical applications due to significant domain discrepancies between simulated degradations and real-world imaging conditions. To bridge this synthetic-to-real gap, we propose a novel *S*elf-supervised *E*vent-based SRB (SE-SRB) framework that leverages neuromorphic event streams as physical priors and adopts a lightweight neural architecture tailored for effective domain adaptation. Specifically, the proposed SE-SRB introduces a self-supervised learning paradigm based on asymmetric integral driven consistency, which enforces temporal coherence between predictions derived from RGB and asynchronous event streams at different time points. Extensive experiments validate that SE-SRB consistently outperforms state-of-the-art methods on both synthetic and real-world datasets. Built upon a lightweight parallel two-stream architecture, SE-SRB achieves high computational efficiency, featuring reduced parameter count, lower FLOPs, and real-time inference capability (40 FPS).

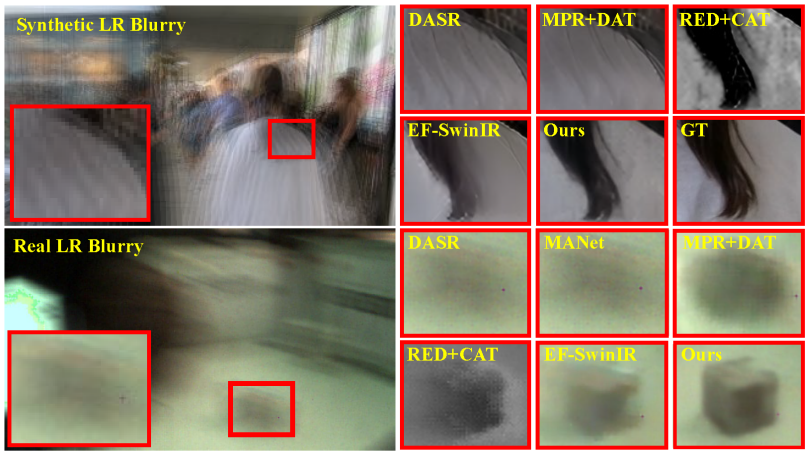
Code — https://github.com/bestrivenzc/SE-SRB

Introduction

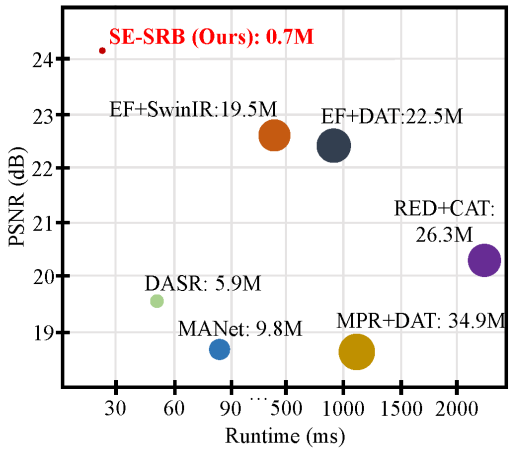
Image Super-Resolution (SR) aims to reconstruct High-Resolution (HR) images from Low-Resolution (LR) inputs and has been widely adopted in video restoration, photography, and efficient data transmission (Liang et al. 2024; Gong et al. 2024). However, motion blur in dynamic scenes often introduces ambiguities and texture loss, significantly impairing SR performance and downstream tasks like autonomous driving (Wang et al. 2022), object tracking (Xin et al. 2024; Wu et al. 2024b), and SLAM (Wu et al. 2023; Ge et al. 2024). While image SR (Chen et al. 2022, 2023; Zhang, Zhang, and Yu 2024) and motion deblurring (MD) (Han et al. 2023; Jung et al. 2021) have achieved significant advancements, naively cascading deblurring modules with SR architectures often amplifies artifacts due to error propagation (Zhang, Zuo, and Zhang 2018). Recent single-image approaches reveal that addressing motion ambiguities during SR enhances reconstruction quality (Fang and Zhan 2022; Niu et al. 2021), but existing approaches often struggle in real-world scenarios with complex motions. For instance, although kernel-based methods show promising performance by utilizing uniform motion assumptions (Yun et al. 2024), their effectiveness usually diminishes in real-world scenes involving non-uniform motions or non-rigid objects. Alternative strategies leverage video motion flow (Bai and Pan 2024) or end-to-end networks (Li, Zuo, and Loy 2023; Barman and Deka 2024), yet current solutions remain domain-specific (*e.g.*, faces (Li, Zuo, and Loy 2023; Zhou et al. 2025) or texts (Li et al. 2022; Noguchi, Fukuda, and Yamanaka 2024)) or overly dependent on the performance of deblurring submodules (Barman and Deka 2024), limiting their generalizability to natural scenes with complex motions.

Recent studies have underscored the benefits of Event-based Super-Resolving Blurry images (E-SRB) in scenarios characterized by complex motion dynamics (Han et al. 2021; Yu et al. 2023). These results demonstrate that event data with microsecond latency enables accurate reconstruction of fine details under non-linear motion, while maintaining temporal precision. Nonetheless, two primary challenges remain in practical applications: (i) **Generalization**: Most prior works rely on well-labeled synthetic datasets for supervised training (Han et al. 2021; Yu et al. 2023; Zhang, Liang, and Shao 2020), which often result in performance degradation in real-world scenarios due to the discrepancy between synthetic LR blurry images and their real-world counterparts, as illustrated by the synthetic (top) and real-world (bottom) examples in Fig. 1 (a); (ii) **Efficiency**: To improve reconstruction performance, existing MD and SR methods often rely on large model sizes and complex network architectures. However, the resulting computational overhead significantly slows inference, making real-time applications impractical

*Corresponding Author.



(a) Visual comparison results of different method



(b) Performance and efficiency comparisons

Figure 1: SE-SRB achieves superior reconstruction and efficiency over state-of-the-art methods. In (a), top and bottom examples are from synthetic and real-world datasets, respectively. In (b), the top-left model shows the best trade-off among quality, speed, and complexity (indicated by blob size).

(see Fig. 1 (b)).

In this paper, we propose a novel Self-supervised E-SRB (SE-SRB) framework to recover HR sharp frames from LR blurry images and concurrent event streams. Unlike existing methods that rely on well-labeled paired data, SE-SRB leverages the consistency of fused non-homogeneous data at any given time to construct a self-supervised learning algorithm, achieving superior performance with high computational efficiency (see Fig. 1) and making it well-suited for real-world applications without the need for well-labeled paired data. In detail, SE-SRB consists of several key modules: the Learnable Double Integral (LDI) for event integration, the Learnable Division Reconstruction (LDR) for refining coarse results and enabling re-blurring, and the Up-Scaling Reconstruction (USR) for spatial detail enhancement. Such a self-supervised learning paradigm is driven by asymmetric integral consistency, *i.e.*, temporal coherence between predictions from RGB and asynchronous event streams, which enables the model to learn modality fusion and reconstruct sharp HR images based on underlying physical patterns. The main contributions of this work are three-fold:

- We present a fully self-supervised SE-SRB framework that trains directly on real data without well-labeled paired data, effectively matching real data distributions and improving generalization by reducing domain gaps.
- We propose a simple yet effective architecture that leverages asymmetric integral consistency to simultaneously address motion deblurring and super-resolution efficiently, making the proposed SE-SRB suited for real-time applications.
- We validate our proposed SE-SRB on a new E-SRB dataset containing real LR blurry images and events with various scenarios. Extensive experiments demonstrate the superiority of SE-SRB in terms of reconstruction quality and computational efficiency (Fig. 1).

Related Work

Frame-based Super-Resolving Blurry Image. Frame-based Super-Resolving Blurry images (F-SRB) aims to reconstruct high-resolution, sharp images from low-resolution, blurry inputs, which is a highly ill-posed task due to the limited information in texture and motion. Previous straightforward cascading methods (Zamir et al. 2021; Liang et al. 2021a; Chen et al. 2022, 2023), which perform deblurring followed by super-resolution, often amplify artifacts and yield sub-optimal results due to error propagation. Decoupling motion ambiguities is essential to tackle SRB in an integrated manner. To address this, early attempts often assume uniform motion and represent blur with a unified degradation model parameterized by a blurring kernel (Yun et al. 2024; Zhang et al. 2021). Jointly optimizing the blurring kernel and image super-resolution can substantially improve performance in both non-blind (Zhang, Zuo, and Zhang 2019) and blind (Nah et al. 2019; Wang et al. 2021) approaches. Gu et al. (Gu et al. 2019) introduced an iterative kernel correction method to refine the blur kernel based on prior SR results. Recently, domain-specific priors (Li, Zuo, and Loy 2023; Noguchi, Fukuda, and Yamanaka 2024; Zhou et al. 2025) are employed to mitigate the ill-posedness, but these methods are mainly designed for face and text images. Despite significant progress being achieved, existing F-SRB methods mainly depend on kernel estimation or specific assumptions, and thus often suffer from performance degradation in real-world scenarios with complex and non-uniform motions.

Event-based Super-Resolving Blurry Image. Thanks to the low latency and high temporal resolution, event cameras (Gallego et al. 2020) can effectively capture high-contrast textures and fine motion details in dynamic scenes, making them well-suited for reconstructing sharp images under complex, non-uniform motion. While E-SRB can be achieved by applying event-based deblurring (Zhang and Yu 2022; Sun et al. 2022; Xu et al. 2025) followed by conventional super-

resolution methods (Chen et al. 2023; Zhang, Zhang, and Yu 2024; Wu et al. 2024a), such cascaded pipelines are prone to error accumulation (Zhang, Zuo, and Zhang 2018), often resulting in suboptimal performance. To address this, recent works (Wang et al. 2020; Han et al. 2021) have proposed unified E-SRB frameworks that jointly optimize deblurring and super-resolution in an end-to-end manner, benefiting from well-annotated paired datasets. However, these supervised approaches suffer from domain gaps between synthetic and real-world low-resolution blurry inputs, leading to performance degradation in practical scenarios. To this end, we propose SE-SRB, a unified self-supervised framework to fit real-world data distributions and enable efficient training without relying on well-labeled data.

Self-supervised Event-based SRB

Problem Formulation

Due to the imperfection of image sensors, the captured image B may suffer from motion blur and low spatial resolution, *i.e.*,

$$B = \frac{1}{T}\int_{t\in\mathcal{T}} I(t)dt, \tag{1}$$

$$I(t) = \mathcal{D}^{\downarrow}(L(t)), \tag{2}$$

where L indicates High-Resolution (HR) sharp images, $\mathcal{T} \triangleq [0, T]$ denotes the exposure interval of B, and I is the Low-Resolution (LR) version of L downscaled via a spatial degradation model $\mathcal{D}^{\downarrow}(\cdot)$. Taking into account event data, the goal of E-SRB is to restore the HR sharp image $\hat{L}(t)$ at arbitrary timestamps $t \in \mathcal{T}$, *i.e.*,

$$\hat{L}(t) = \mathcal{G}_\theta\left(t, B, \mathcal{E}_\mathcal{T}\right), \forall t \in \mathcal{T}, \tag{3}$$

where $\mathcal{G}_\theta$ denotes E-SRB models with parameters θ. $\mathcal{E}_\mathcal{T} \triangleq \{(\mathbf{x}_n, p_n, t_n)\}_{t_n\in\mathcal{T}}$ are the events triggered within $\mathcal{T}$, where t_n and $\mathbf{x}_n$ respectively represent the timestamps and the pixel locations of the n-th event, and $p_n \in \{+1, -1\}$ is the polarity.

Supervised schema: Given a dataset $D_{\text{Syn}} \triangleq \{(B, \mathcal{E}_\mathcal{T}, L)_i\}_i$ containing synthetic LR blurry images, event streams, and ground-truth HR sharp images, the parameters θ of the model $\mathcal{G}_\theta$ can be optimized by minimizing the following objective:

$$\underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{D_{\text{Syn}}}\left\{\mathcal{L}\left(\mathcal{G}_\theta(t, B, \mathcal{E}_\mathcal{T}), L(t)\right)\right\}, \tag{4}$$

where $\mathcal{L}$ is a loss function, *e.g.*, $\mathcal{L}_1$ loss, that quantifies the discrepancy between the reconstructed output $\hat{L}(t)$ and the ground-truth image $L(t)$. Most existing F-SRB (excluding $\mathcal{E}_\mathcal{T}$ as input) and E-SRB methods (Zhang, Zuo, and Zhang 2018, 2019; Park and Mu Lee 2017; Han et al. 2021) rely on D_{Syn} for supervised learning. However, this often leads to a performance drop in real-world scenarios due to the domain gap between synthetic and real data. Additionally, as illustrated in Fig.1 (b), previous approaches usually employ complex architectures with a large number of parameters θ, which hinders their suitability for real-time applications.

Self-supervised schema: Our objective is to design a self-supervised framework that addresses the limitations inherent in existing supervised learning approaches. Given a dataset collected in real-world scenarios $D_{\text{Real}} \triangleq \{(B, \mathcal{E}_\mathcal{T})_i\}_i$ *without ground-truth images*, we aim to design an objective function to utilize only LR blurry images B and the corresponding events $\mathcal{E}_\mathcal{T}$ for supervision, *i.e.*,

$$\underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{D_{\text{Real}}}\left\{\mathcal{L}\left(\mathcal{G}_\theta(t, B, \mathcal{E}_\mathcal{T}), B, \mathcal{E}_\mathcal{T}\right)\right\}. \tag{5}$$

However, achieving superior real-world performance while ensuring computational efficiency presents significant challenges in both model architecture and self-supervision strategy:

- **Complex Models**. E-SRB relies on effectively leveraging complementary information from events and images. However, existing fusion methods are overly complex and impractical for real-time use, underscoring the need for architectures that ensure both accuracy and efficiency.
- **Reliance on Synthetic Paired Data**. Real-world blurry frames B suffer from texture loss and motion ambiguity due to long exposure and dynamic scenes (Eq. 1). While event streams $\mathcal{E}_\mathcal{T}$ provide fine-grained motion cues, both modalities lack high-resolution detail, making their fusion difficult without ground-truth supervision.

Framework

Unlike existing methods that rely on complex transformer-based models, our approach introduces a novel self-supervised dual-branch paradigm based on asymmetric integral-driven consistency, which enforces temporal coherence between RGB and event-based predictions across time. By effectively fusing both modalities under this constraint, the SE-SRB framework adopts a lightweight design (Fig. 2) built on streamlined convolutional modules and residual dense blocks, achieving high efficiency with low computational cost. Each branch comprises three key modules: Learnable Double Integral (LDI), Learnable Division Reconstruction (LDR), and Up-Scaling Reconstruction (USR).

Learnable Double Integral. We first revisit the physical model of event generation. In event-based vision, an event is triggered when the logarithmic change in brightness intensity surpasses a predefined threshold $c > 0$, *i.e.*,

$$\log(I(t_n, \mathbf{x}_n)) - \log(I(t_n + \Delta t, \mathbf{x}_n)) = p_n \cdot c, \tag{6}$$

where $I(t_n, \mathbf{x}_n)$ and $I(t_n + \Delta t, \mathbf{x}_n)$ denote the instantaneous intensity at time t_n and $t_n + \Delta t$ at the pixel position $\mathbf{x}_n$. Combining Eq. 1, Eq. 6, and the EDI model (Pan et al. 2019), one can obtain

$$I(m) = \frac{B}{E(m, \mathcal{T})}, \quad \text{with} \tag{7}$$

$$E(m, \mathcal{T}) = \frac{1}{T}\int_{t\in\mathcal{T}} \exp(c\int_m^t e(s)ds)dt \tag{8}$$

that converts blurry frames B to latent sharp images $I(m)$ at arbitrary time $m \in \mathcal{T}$ with events $\mathcal{E}_\mathcal{T}$. Moreover, we omit the pixel position in Eqs. 7 and 8 for better readability. In Eq. 8,

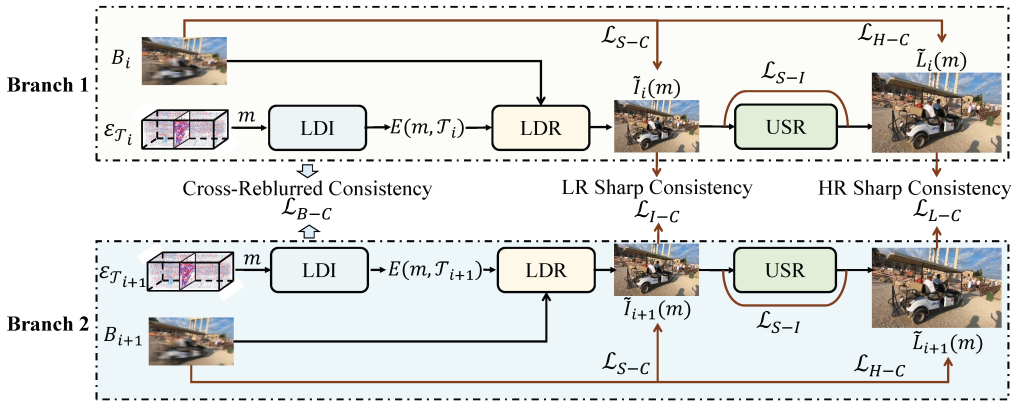


Figure 2: Data flow of the proposed SE-SRB: Given two consecutive LR blurry frames B_i, B_{i+1} and their events, the Learnable Double Integral (LDI) and Learnable Division Reconstruction (LDR) modules predict LR sharp images $\tilde{I}_i(m)$ and $\tilde{I}_{i+1}(m)$ at time m, then upsampled by the Up-Sampling Reconstruction (USR) module to produce HR outputs $\tilde{L}_i(m)$ and $\tilde{L}_{i+1}(m)$.

$e(t) \triangleq \sum_n p_n \cdot \delta(t - t_n)$ denotes the continuous representation of events with $\delta(\cdot)$ indicating the Dirac function. Due to inherent non-idealities in physical event sensors (Gallego et al. 2020), the event threshold c in Eq. 8 is spatially and temporally unstable, leading to inaccuracies in $E(\cdot)$, especially under long-term integration where noise accumulates significantly(Zhang and Yu 2022). Following (Zhang and Yu 2022), we adopt a Learnable Double Integral (LDI) module composed of cascaded 3×3 convolutions with ReLU activations to statistically model real-world event data. For each target timestamp m, two separate double integrals are computed over $\mathcal{T}_i$ and $\mathcal{T}_{i+1}$ as follows:

$$\begin{aligned} E(m, \mathcal{T}_i) &= \text{LDI}(\mathcal{P}(\mathcal{E}_{\mathcal{T}_i})), \\ E(m, \mathcal{T}_{i+1}) &= \text{LDI}(\mathcal{P}(\mathcal{E}_{\mathcal{T}_{i+1}})), \end{aligned} \tag{9}$$

where $\mathcal{P}(\cdot)$ denotes the event pre-processing and spatiotemporal representation described in (Zhang and Yu 2022).

Learnable Division Reconstruction. The latent sharp LR image can be theoretically restored from LR blurry inputs B and corresponding event integrals $E(m, \mathcal{T})$ through the physical model in Eq. 7. However, directly applying this formulation faces practical challenges due to the instability of the threshold c. To this end, we further develop an Learnable Division Reconstruction (LDR) module that learns to approximate the ideal division mapping of Eq. 7, generating refined latent sharp LR images $\tilde{I}_i(m)$ and $\tilde{I}_{i+1}(m)$ through adaptive feature compensation, which can be formulated as:

$$\begin{aligned} \tilde{I}_i(m) &= \text{LDR}(B_i, E(m, \mathcal{T}_i)), \\ \tilde{I}_{i+1}(m) &= \text{LDR}(B_{i+1}, E(m, \mathcal{T}_{i+1})). \end{aligned} \tag{10}$$

The latent frame $\tilde{I}_{i+1}(m)$ is estimated by computing $E(m, \mathcal{T}_{i+1})$ and applying the LDR to approximate Eq. 7, despite $m \in \mathcal{T}_i$ falling outside the interval $\mathcal{T}_{i+1}$.

Up-Scaling Reconstruction. To maintain architectural simplicity, we design a lightweight Up-Scaling Reconstruction (USR) module composed of two convolutional layers followed by two upsampling layers. This compact structure progressively recovers spatial details with minimal computational cost. Given $\tilde{I}_i(m)$ and $\tilde{I}_{i+1}(m)$ as inputs, the USR modules reconstruct latent HR sharp images $\tilde{L}_i(m)$ and $\tilde{L}_{i+1}(m)$ with improved texture and structural fidelity.

Inference Procedure. At an arbitrary timestamp $m \in \mathcal{T}_i$, the proposed SE-SRB generates an HR sharp result $\tilde{L}_i(m)$ from a single LR image B_i and corresponding events. This design enables fair comparison with existing image deblurring and super-resolution methods.

Self-Supervised Learning

Leveraging the intermediate representations introduced in Section Framework, we establish temporal deblurring and spatial super-resolution constraints, which facilitate the training of our SE-SRB model under real-world scenarios without relying on ground-truth data.

Temporal Deblurring Constraints. Under the assumption that images captured at the same timestamp m should maintain consistent in both texture and brightness, we define a LR sharp image $\tilde{I}(m)$ consistency loss $\mathcal{L}_{I\text{-}C}$ as follows:

$$\mathcal{L}_{I\text{-}C} = \|\tilde{I}_i(m) - \tilde{I}_{i+1}(m)\|_1, \tag{11}$$

where $\tilde{I}_i(m)$ and $\tilde{I}_{i+1}(m)$ are generated by Eq. 10. Furthermore, since Eq. 7 can be rewritten as $B = I(m)/E^{-1}(m, \mathcal{T})$, LDR modules can also be applied to estimate the reblurred LR images $\tilde{B}_i$ and $\tilde{B}_{i+1}$, *i.e.*,

$$\begin{aligned} \tilde{B}_i &= \text{LDR}(\tilde{I}_{i+1}(m), E^{-1}(m, \mathcal{T}_i)), \\ \tilde{B}_{i+1} &= \text{LDR}(\tilde{I}_i(m), E^{-1}(m, \mathcal{T}_{i+1})). \end{aligned} \tag{12}$$

It is worth noting that $\tilde{I}_i(m)$ and $\tilde{I}_{i+1}(m)$ are used to reconstruct $\tilde{B}_{i+1}$ and $\tilde{B}_i$, respectively. This cross-supervision strategy exploits the similarity between $\tilde{I}_i(m)$ and $\tilde{I}_{i+1}(m)$ while avoiding identity mapping, effectively leveraging the complementary sharpness information in adjacent blurry frames. The reblurred $\tilde{B}_i$ and $\tilde{B}_{i+1}$ are then used to define the blurry consistency loss $\mathcal{L}_{B\text{-}C}$, enforcing alignment with real blur characteristics:

$$\mathcal{L}_{B\text{-}C} = \|\tilde{B}_i - B_i\|_1 + \|\tilde{B}_{i+1} - B_{i+1}\|_1. \tag{13}$$

Methods	Events	SSL	4× Upscaling				2× Upscaling				#Params.	FLOPs	Runtime
			GoPro		REDS		GoPro		REDS				
			PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑			
MANet	✗	✗	20.42	0.5868	18.76	0.4680	22.68	0.6694	19.55	0.4861	9.89	607.6	88.5
DASR	✗	✓	26.55	0.8108	19.56	0.5287	27.62	0.8554	20.18	0.5451	5.97	208.8	54.1
MPR+DAT	✗	✗	26.86	0.8236	19.48	0.5336	27.99	0.8784	20.04	0.5471	34.93	1983.2	1038.5
MPR+CAT	✗	✗	26.97	**0.8253**	19.50	0.5347	28.07	**0.8792**	20.06	0.5480	36.73	2068.1	2239.2
MPR+SwIR	✗	✗	26.87	0.8216	19.46	0.5319	28.03	0.8783	20.05	0.5464	27.83	1815.3	101.4
RED+DAT	✓	✗	22.89	0.7983	20.31	0.5885	23.30	0.8483	20.87	0.6636	24.59	436.7	977.09
RED+CAT	✓	✗	22.94	0.7985	20.26	0.5879	23.32	0.8484	20.90	0.6647	26.39	521.6	2177.8
RED+SwIR	✓	✗	22.92	0.7955	20.28	0.5882	23.34	0.8486	20.84	0.6601	21.64	969.6	487.4
EF+DAT	✓	✗	26.43	0.8182	22.21	0.6254	27.39	0.8701	23.05	0.6890	22.50	383.7	970.5
EF+CAT	✓	✗	26.50	0.8193	22.23	0.6259	27.44	0.8708	23.07	0.6900	24.3	468.6	2171.2
EF+SwIR	✓	✗	26.37	0.8158	22.18	0.6226	27.41	0.8697	23.03	0.6878	19.55	916.6	480.8
EVDI+DASR	✓	✓	23.70	0.8183	21.90	0.6511	24.19	0.8657	22.75	0.7293	6.36	222.3	72.2
SE-SRB (Ours)	✓	✓	**27.42**	0.8209	**24.04**	**0.6776**	**28.60**	0.8734	**25.21**	**0.7355**	**0.72**	**61.6**	**24.9**

Table 1: Quantitative comparisons of the proposed method to the state-of-the-art deblurring and super-resolving approaches on GoPro (Nah, Hyun Kim, and Mu Lee 2017) and REDS (Nah et al. 2019) datasets. The column SSL indicates the self-supervised lerning. The best and second-best results are **highlighted** and underlined.

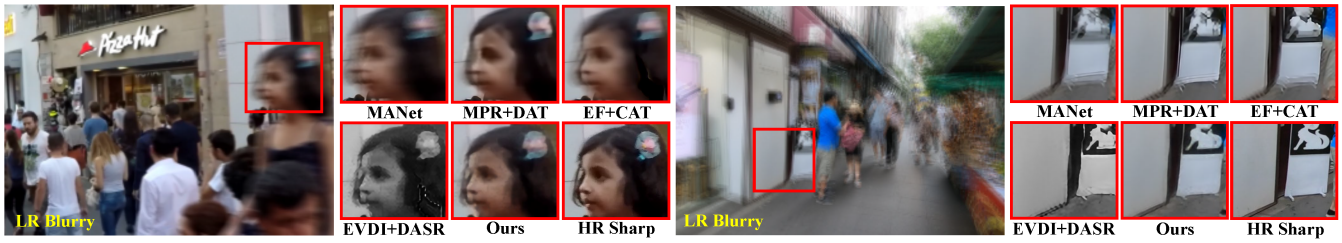


Figure 3: Qualitative comparison results on the GoPro (left) and REDS (right) datasets.

Additionally, the refined LR sharp images $\tilde{I}_i(m)$ and $\tilde{I}_{i+1}(m)$ preserve photometric consistency with the input blurry frames B_i and B_{i+1} in static regions, as dictated by the image formation model. Leveraging motion cues from event data to distinguish static from dynamic regions, we introduce a sharp-to-blurry consistency loss $\mathcal{L}_{S\text{-}C}$ to enforce brightness consistency between $\tilde{I}(m)$ and their blurry counterparts:

$$\begin{aligned}\mathcal{L}_{S\text{-}C} = &\|(\tilde{I}_i(m) - B_i) * \mathcal{M}(\mathcal{E}_{\mathcal{T}_i})\|_1 \\ &+ \|(\tilde{I}_{i+1}(m) - B_{i+1}) * \mathcal{M}(\mathcal{E}_{\mathcal{T}_{i+1}})\|_1,\end{aligned} \tag{14}$$

where $\mathcal{M}(\cdot)$ is a binary mask identifying static pixels (1, no events) versus dynamic pixels (0, with events).

Spatial Super-Resolution Constrains. Similar to $\mathcal{L}_{I\text{-}C}$, the reconstructed HR sharp images are expected to exhibit consistency in both structural details and photometric properties. Accordingly, we define an HR sharp image consistency loss $\mathcal{L}_{L\text{-}C}$ as follows:

$$\mathcal{L}_{L\text{-}C} = \|\tilde{L}_i(m) - \tilde{L}_{i+1}(m)\|_1. \tag{15}$$

Moreover, the reconstructed HR image $\tilde{L}(m)$ should maintain photometric consistency with the corresponding input blurry frame B, despite the difference in spatial resolution. To achieve this end, we first downsample $\tilde{L}(m)$ to match the resolution of B, and then formulate a HR sharp-to-blurry consistency loss $\mathcal{L}_{H\text{-}C}$ based on Eq. 14, thereby enforcing accurate brightness reconstruction in $\tilde{L}(m)$. The $\mathcal{L}_{H\text{-}C}$ is formulated as:

$$\begin{aligned}\mathcal{L}_{H\text{-}C} = &\|(\mathcal{D}^{\downarrow}(\tilde{L}_i(m)) - B_i) * \mathcal{M}(\mathcal{E}_{\mathcal{T}_i})\|_1 \\ &+ \|(\mathcal{D}^{\downarrow}(\tilde{L}_{i+1}(m)) - B_{i+1}) * \mathcal{M}(\mathcal{E}_{\mathcal{T}_{i+1}})\|_1.\end{aligned} \tag{16}$$

Additionally, inspired by the identity loss introduced in (Zhu et al. 2017), which aims to preserve color composition between input and output images, we design a Super-resolution Identity loss $\mathcal{L}_{\text{S-I}}$ to simultaneously constrain color fidelity and enhance the reconstruction of spatial details. The loss is defined as follows:

$$\mathcal{L}_{S\text{-}I} = \|\,\mathrm{USR}(\mathcal{D}^{\downarrow}(Y)) - Y\|_1, \tag{17}$$

where Y is a sharp HR image randomly drawn from publicly available natural image datasets.

Overall Loss. Finally, the total self-supervised framework can be summarized as

$$\begin{aligned}\mathcal{L}_{all} = &\mathcal{L}_{I-C} + \alpha_1\mathcal{L}_{B-C} + \alpha_2\mathcal{L}_{S-C} \\ &+ \alpha_3\mathcal{L}_{L-C} + \alpha_4\mathcal{L}_{H-C} + \alpha_5\mathcal{L}_{S-I},\end{aligned} \tag{18}$$

where α_1 to α_5 are weighting coefficients.

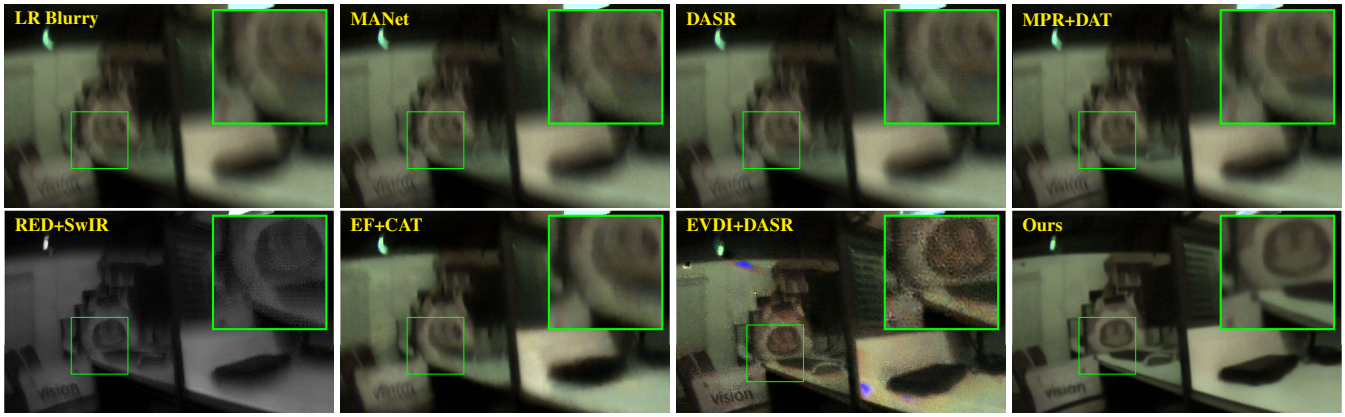


Figure 4: Qualitative comparison results on the real-world RLBE dataset.

Experiments and Analysis

Experimental Settings

Our SE-SRB model is implemented in PyTorch and trained on NVIDIA RTX 3090 GPUs with a batch size of 4. We use the Adam (Kingma and Ba 2015) and SGDR (Loshchilov and Hutter 2017) with T_{max}=200. During training, images are randomly cropped to 128×128. The loss weights α_1–α_5 in Eq. 18 are set to 1, 1, 0.1, 0.1, and 4. For evaluation, we report PSNR and SSIM(Wang, Simoncelli, and Bovik 2003).

We evaluate the proposed method on three datasets covering both synthetic and real-world scenarios. Two synthetic datasets are constructed from REDS (Nah et al. 2019) and GoPro (Nah, Hyun Kim, and Mu Lee 2017), widely used for SRB and recently adapted for E-SRB (Wang et al. 2020; Han et al. 2021). Original frames are downsampled to 180×320, and 7 intermediate frames are interpolated between each frame pair using RIFE (Huang et al. 2022). LR blurry frames are synthesized by averaging 11 (GoPro) or 7 (REDS) sharp frames, while event streams are generated using vid2e (Gehrig et al. 2020). Unlike synthetic datasets, we present the Real-world LR Blurry image and Events (RLBE) dataset, captured by a DAVIS346 camera, featuring diverse real-world scenes with complex real-world motions.

Experimental Results

SRB is a compound task that simultaneously addresses image super-resolution and motion deblurring. To comprehensively evaluate our SE-SRB framework, we compare it against state-of-the-art (SOTA) SRB approaches, including end-to-end solutions such as DASR (Wang et al. 2021) and MANet (Liang et al. 2021b), as well as two-stage pipelines that sequentially apply motion deblurring followed by super-resolution. For the deblurring stage, we consider both frame-based methods (*e.g.*, MPR (Zamir et al. 2021)) and event-based approaches (*e.g.*, RED (Xu et al. 2021), EF (Sun et al. 2022), and self-supervised EVDI (Zhang and Yu 2022)). The super-resolution component is implemented using representative methods including SwinIR (Liang et al. 2021a), DAT (Chen et al. 2023), and CAT (Chen et al. 2022). For a fair comparison, we retrain the self-supervised methods EVDI (Zhang and Yu 2022) and DASR (Wang et al. 2021) on each synthetic and real-world dataset using their official code and default parameters.

Evaluation on Synthetic Dataset. Quantitative and qualitative comparisons on the GoPro and REDS datasets are shown in Tab. 1 and Fig. 3, respectively. Tab. 1 reports PSNR and SSIM results, where our SE-SRB consistently outperforms prior methods across nearly all settings. In particular, SE-SRB achieves superior reconstruction quality for both $2\times$ and $4\times$ HR sharp images on synthetic REDS datasets, validating its ability to jointly address motion deblurring and super-resolution in a self-supervised manner. Fig. 3 further highlights visual improvements, where the $4\times$ HR results produced by SE-SRB exhibit clearer textures and sharper structures compared to SOTA methods. The right case particularly showcase its advantages: the proposed SE-SRB recovers fine structural details (e.g., the top-right corners) lost in other methods and reconstructs continuous sharp lines (e.g., the bottom-left corners) with reduced aliasing and blur.

Evaluation on Real Dataset. To assess the real-world performance of SE-SRB, we test it on the RLBE dataset, which contains real-world LR blurry frames paired with event data. As shown in Fig. 4, the retrained self-supervised EVDI+DASR exceeds prior supervised frame- or event-based methods, though its results are still affected by artifacts from the decoupled MD and SR pipeline. In contrast, SE-SRB produces sharper edges and more coherent textures, demonstrating stronger generalization in real-world scenes.

Complexity to Performance Analysis

SE-SRB achieves the fastest inference ($\approx$24.9ms) with the lowest complexity (0.72M parameters, 61.6G FLOPs) on 256×256 inputs (NVIDIA RTX 3090), as summarized in Tab. 1, significantly outperforming prior SRB methods.

Ablation Study

We study the importance of each loss functions and the network module in our self-supervised framework in Tab. 2, Figs. 5 and 6, and the following conclusion are drawn:

Temporal Deblurring Losses. When trained solely with the inter-frame consistency loss $\mathcal{L}_{I\text{-}C}$, the model tends to minimize the objective by predicting trivial solutions with near-

Figure 5: Qualitative ablations of each self-supervised loss on the REDS dataset.

Models	$\mathcal{L}_{I\text{-}C}$	$\mathcal{L}_{S\text{-}C}$	$\mathcal{L}_{B\text{-}C}$	$\mathcal{L}_{S\text{-}I}$	$\mathcal{L}_{L\text{-}C}$	$\mathcal{L}_{H\text{-}C}$	PSNR↑/SSIM↑	
							4× Upscaling	2× Upscaling
#0	✓						6.97/0.1013	6.98/0.1014
#1	✓	✓					21.22/0.3556	22.89/0.6119
#2	✓	✓	✓				21.78/0.3776	23.66/0.6489
#3	✓	✓	✓	✓			23.81/0.6640	24.96/0.7232
#4	✓	✓	✓	✓	✓		7.96/0.3854	8.01/0.3462
#5	✓	✓	✓	✓	✓	✓	**24.04/0.6776**	**25.21/0.7355**

Table 2: Ablation study on the REDS dataset.

Figure 6: Qualitative results of intermediate results including B_i and B_{i+1}, $\tilde{I}_i(m)$ and $\tilde{I}_{i+1}(m)$, and $\tilde{B}_i$ and $\tilde{B}_{i+1}$.

zero intensity values, resulting in artificially high similarity between $\tilde{I}_i(m)$ and $\tilde{I}_{i+1}(m)$. This leads to suboptimal performance, as evidenced by the results in #0 of Tab. 2 and Fig. 5 (b). Introducing the additional intensity constraint $\mathcal{L}_{S\text{-}C}$ (#1) significantly improves both quantitative and qualitative outcomes. Moreover, incorporating the reblurring consistency loss $\mathcal{L}_{S\text{-}C}$ (#2) further enhances performance by enforcing constraints on both structural textures and luminance, as shown in Tab. 2 and Fig. 5 (d).

Spatial Super-Resolution Losses. While temporal deblurring losses improve LR reconstruction, they often introduce color shifts and jagged artifacts due to limited supervision in the HR domain. To address these issues, we introduce the SR identity loss $\mathcal{L}_{S\text{-}I}$, which significantly enhances visual quality, as demonstrated in #3 of Tab. 2 and Fig. 5 (e). Building on this, incorporating the HR-domain consistency loss $\mathcal{L}_{L\text{-}C}$ alone (#4) results in undesirable performance. As shown in Fig. 5 (f), $\mathcal{L}_{L\text{-}C}$ dominates when combined with $\mathcal{L}_{S\text{-}I}$, leading the network to produce overly uniform (white) outputs where $\tilde{T}_i$ and $\tilde{T}_{i+1}$ become nearly identical. This minimizes the consistency loss $\mathcal{L}_{L\text{-}C}$ but fails to preserve informative structures. Combined with the HR-domain intensity constraint $\mathcal{L}_{H\text{-}C}$ (#5)—which aligns outputs with input intensities—the network prevents overly uniform results and achieves better temporal coherence and visual fidelity.

Effectiveness of the proposed LDR module. As shown in Fig. 6, the latent LR sharp images $\tilde{I}_i(m)$ and $\tilde{I}_{i+1}(m)$ at arbitrary timestamp $m \in \mathcal{T}_i$, as well as the reblurred LR images $\tilde{B}_i$ and $\tilde{B}_{i+1}$, are computed using the proposed Learnable Division Reconstruction (LDR) module via Eqs. 10 and 12, respectively. Compared with the original blurry images B_i and B_{i+1}, the recovered $\tilde{I}_i(m)$ and $\tilde{I}_{i+1}(m)$ exhibit sharper textures and finer details, while the reblurred outputs $\tilde{B}_i$ and $\tilde{B}_{i+1}$ closely resemble the input blurs. These results validate the effectiveness of the proposed LDR module in accurately approximating the division process described in Eq. 7.

Conclusion

We propose SE-SRB, a self-supervised framework that reconstructs HR sharp images from LR blurry inputs. It leverages the intrinsic consistency among blurry frames, latent sharp images, and event data to jointly enforce temporal deblurring and spatial super-resolution constraints, enabling effective training without labels. Extensive experiments on synthetic and real-world datasets show that SE-SRB matches state-of-the-art performance while maintaining a lightweight design.

Acknowledgements

This work was supported in part by the Interdisciplinary Frontier Research Project of PCL (2025QYB013), the Major Key Project of PCL (PCL2025A03), the Fundamental Research Funds for the Central Universities (204205kf0063), and the National Natural Science Foundation of China (62271354, 62132006, U24A20220).

References

Bai, H.; and Pan, J. 2024. Self-Supervised Deep Blind Video Super-Resolution. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(7): 4641–4653.

Barman, T.; and Deka, B. 2024. A Deep Learning-Based Joint Image Super-Resolution and Deblurring Framework. *IEEE Trans. on Artificial Intell.*, 5(6): 3160–3173.

Chen, Z.; Zhang, Y.; Gu, J.; Kong, L.; Yang, X.; and Yu, F. 2023. Dual aggregation transformer for image super-resolution. In *Int. Conf. Comput. Vis.*, 12312–12321.

Chen, Z.; Zhang, Y.; Gu, J.; Kong, L.; Yuan, X.; et al. 2022. Cross aggregation transformer for image restoration. *Adv. Neural Inform. Process. Syst.*, 35: 25478–25490.

Fang, N.; and Zhan, Z. 2022. High-resolution optical flow and frame-recurrent network for video super-resolution and deblurring. *Neurocomputing*, 489: 128–138.

Gallego, G.; Delbruck, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; and Scaramuzza, D. 2020. Event-based Vision: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1): 154–180.

Ge, Y.; Zhang, L.; Wu, Y.; and Hu, D. 2024. PIPO-SLAM: Lightweight Visual-Inertial SLAM With Preintegration Merging Theory and Pose-Only Descriptions of Multiple View Geometry. *IEEE Trans. on Robotics*, 40: 2046–2059.

Gehrig, D.; Gehrig, M.; Hidalgo-Carrió, J.; and Scaramuzza, D. 2020. Video to Events: Recycling Video Datasets for Event Cameras. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Gong, Z.; Bao, G.; Zhang, Q.; Wan, Z.; Miao, D.; Wang, S.; Zhu, L.; Wang, C.; Xu, R.; Hu, L.; et al. 2024. NeuroClips: Towards high-fidelity and smooth fMRI-to-video reconstruction. *Adv. Neural Inform. Process. Syst.*, 37: 51655–51683.

Gu, J.; Lu, H.; Zuo, W.; and Dong, C. 2019. Blind super-resolution with iterative kernel correction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 1604–1613.

Han, G.; Wang, M.; Zhu, H.; and Lin, C. 2023. MPDNet: An underwater image deblurring framework with stepwise feature refinement module. *Engineering Appli. of Artificial Intell.*, 126: 106822.

Han, J.; Yang, Y.; Zhou, C.; Xu, C.; and Shi, B. 2021. Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. In *Int. Conf. Comput. Vis.*, 4882–4891.

Huang, Z.; Zhang, T.; Heng, W.; Shi, B.; and Zhou, S. 2022. Real-time intermediate flow estimation for video frame interpolation. In *Eur. Conf. Comput. Vis.*, 624–642. Springer.

Jung, H.; Kim, Y.; Jang, H.; Ha, N.; and Sohn, K. 2021. Multi-Task Learning Framework for Motion Estimation and Dynamic Scene Deblurring. *IEEE Trans. Image Process.*, 30: 8170–8183.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*

Li, X.; Chen, C.; Lin, X.; Zuo, W.; and Zhang, L. 2022. From face to natural image: Learning real degradation for blind image super-resolution. In *Eur. Conf. Comput. Vis.*, 376–392. Springer.

Li, X.; Zuo, W.; and Loy, C. C. 2023. Learning generative structure prior for blind text image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 10103–10113.

Liang, J.; Cao, J.; Fan, Y.; Zhang, K.; Ranjan, R.; Li, Y.; Timofte, R.; and Van Gool, L. 2024. Vrt: A video restoration transformer. *IEEE Trans. Image Process.*

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021a. Swinir: Image restoration using swin transformer. In *Int. Conf. Comput. Vis.*, 1833–1844.

Liang, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021b. Mutual affine network for spatially variant kernel estimation in blind image super-resolution. In *Int. Conf. Comput. Vis.*, 4096–4105.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *Int. Conf. Learn. Represent.*

Nah, S.; Baik, S.; Hong, S.; Moon, G.; Son, S.; Timofte, R.; and Mu Lee, K. 2019. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 1974–1984.

Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 3883–3891.

Niu, W.; Zhang, K.; Luo, W.; and Zhong, Y. 2021. Blind motion deblurring super-resolution: When dynamic spatio-temporal learning meets static image understanding. *IEEE Trans. Image Process.*, 30: 7101–7111.

Noguchi, C.; Fukuda, S.; and Yamanaka, M. 2024. Scene text image super-resolution based on text-conditional diffusion models. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 1485–1495.

Pan, L.; Scheerlinck, C.; Yu, X.; Hartley, R.; Liu, M.; and Dai, Y. 2019. Bringing a blurry frame alive at high frame-rate with an event camera. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 6820–6829.

Park, H.; and Mu Lee, K. 2017. Joint estimation of camera pose, depth, deblurring, and super-resolution from a blurred image sequence. In *Int. Conf. Comput. Vis.*, 4613–4621.

Sun, L.; Sakaridis, C.; Liang, J.; Jiang, Q.; Yang, K.; Sun, P.; Ye, Y.; Wang, K.; and Gool, L. V. 2022. Event-Based Fusion for Motion Deblurring with Cross-modal Attention. In *Eur. Conf. Comput. Vis.*, 412–428.

Wang, B.; He, J.; Yu, L.; Xia, G.-S.; and Yang, W. 2020. Event enhanced high-quality image recovery. In *Eur. Conf. Comput. Vis.*, 155–171.

Wang, L.; Wang, Y.; Dong, X.; Xu, Q.; Yang, J.; An, W.; and Guo, Y. 2021. Unsupervised degradation representation

learning for blind super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 10581–10590.

Wang, Q.; Han, T.; Qin, Z.; Gao, J.; and Li, X. 2022. Multi-task Attention Network for Lane Detection and Fitting. *IEEE Trans. Neural Networks and Learning Syst.*, 33(3): 1066–1078.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multi-scale structural similarity for image quality assessment. In *IEEE Asilomar Conf. Sign. Syst. Comput.*, volume 2, 1398–1402.

Wu, R.; Yang, T.; Sun, L.; Zhang, Z.; Li, S.; and Zhang, L. 2024a. Seesr: Towards semantics-aware real-world image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 25456–25467.

Wu, Y.; Wang, L.; Zhang, L.; Bai, Y.; Cai, Y.; Wang, S.; and Li, Y. 2023. Improving autonomous detection in dynamic environments with robust monocular thermal SLAM system. *ISPRS Journal of Photogrammetry and Remote Sensing*, 203: 265–284.

Wu, Z.; Wen, J.; Xu, Y.; Yang, J.; Li, X.; and Zhang, D. 2024b. Enhanced Spatial Feature Learning for Weakly Supervised Object Detection. *IEEE Trans. Neural Networks and Learning Syst.*, 35(1): 961–972.

Xin, Z.; Chen, S.; Wu, T.; Shao, Y.; Ding, W.; and You, X. 2024. Few-shot object detection: Research advances and challenges. *Information Fusion*, 107: 102307.

Xu, F.; Yu, L.; Wang, B.; Yang, W.; Xia, G.-S.; Jia, X.; Qiao, Z.; and Liu, J. 2021. Motion deblurring with real events. In *Int. Conf. Comput. Vis.*, 2583–2592.

Xu, S.; Sun, Z.; Zhong, M.; Cao, C.; Liu, Y.; Fu, X.; and Chen, Y. 2025. Motion-adaptive Transformer for Event-based Image Deblurring. In *AAAI*, volume 39, 8942–8950.

Yu, L.; Wang, B.; Zhang, X.; Zhang, H.; Yang, W.; Liu, J.; and Xia, G.-S. 2023. Learning to super-resolve blurry images with events. *IEEE Trans. Pattern Anal. Mach. Intell.*

Yun, J.-S.; Kim, M. H.; Kim, H.-I.; and Yoo, S. B. 2024. Kernel adaptive memory network for blind video super-resolution. *Expert Systems with Applications*, 238: 122252.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-Stage Progressive Image Restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhang, D.; Liang, Z.; and Shao, J. 2020. Joint image deblurring and super-resolution with attention dual supervised network. *Neurocomputing*, 412: 187–196.

Zhang, K.; Li, Y.; Zuo, W.; Zhang, L.; Van Gool, L.; and Timofte, R. 2021. Plug-and-play image restoration with deep denoiser prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10): 6360–6376.

Zhang, K.; Zuo, W.; and Zhang, L. 2018. Learning a single convolutional super-resolution network for multiple degradations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 3262–3271.

Zhang, K.; Zuo, W.; and Zhang, L. 2019. Deep plug-and-play super-resolution for arbitrary blur kernels. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 1671–1681.

Zhang, X.; and Yu, L. 2022. Unifying motion deblurring and frame interpolation with events. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 17765–17774.

Zhang, X.; Zhang, Y.; and Yu, F. 2024. HiT-SR: Hierarchical transformer for efficient image super-resolution. In *Eur. Conf. Comput. Vis.*, 483–500.

Zhou, L.; Wang, M.; Huang, X.; Zheng, W.; Mao, Q.; and Zhao, G. 2025. An Empirical Study of Super-resolution on Low-resolution Micro-expression Recognition. *IEEE Transactions on Affective Computing*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis.*, 2223–2232.