Event-Guided Scene Text Image Super-Resolution

Zihan Qi[1], Zeyu Xiao[2*], Haoyi Zhao[1], Yang Zhao[1], Feng Xue[1], Wei Jia[1*]

[1]Hefei University of Technology
[2]National University of Singapore

Abstract

Scene text image super-resolution aims to enhance text legibility by recovering high-resolution text images from low-resolution inputs. However, maintaining fine details such as text strokes, edges, and textual accuracy remains challenging, particularly in low-light environments and high-speed motion scenarios, where degradation is more severe. Event cameras, with their high temporal resolution and ability to capture intensity changes, offer a promising solution for restoring lost fine details and mitigating degradation in these challenging conditions. In this paper, we propose EvTSR, the first framework that integrates Event data for scene Text image Super-Resolution. The core of EvTSR is the dual-stream frequency boost (DSFB) mechanism, which separates image features into high- and low-frequency components. High-frequency details like edges and strokes are enhanced using event data via the event-guided high-frequency (EGH) mechanism, while low-frequency components, responsible for global structure, are refined using the Text-Guided Low-frequency (TGL) mechanism with a pre-trained text recognizer, ensuring textual coherence. To further improve cross-modal integration, we introduce the cross-modal fusion (CMF) mechanism, which effectively aligns event and image features, enabling robust information fusion. Extensive experiments demonstrate that EvTSR achieves superior performance over existing methods.

Code — https://github.com/codes81/EVTSR

1 Introduction

Scene text image super-resolution (TSR) is crucial for enhancing text clarity and legibility in low-resolution (LR) images, thereby boosting the performance of recognition systems in applications like autonomous driving (Reddy et al. 2020) and intelligent transportation (Al-Shemarry, Li, and Abdulla 2022; Liem et al. 2018). While deep learning-based methods (Mao et al. 2025; Xiao, Li, and Jia 2025; Xiao and Wang 2025) using convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Transformers have significantly advanced the field, they still face a critical challenge: performance degrades severely in extreme conditions.

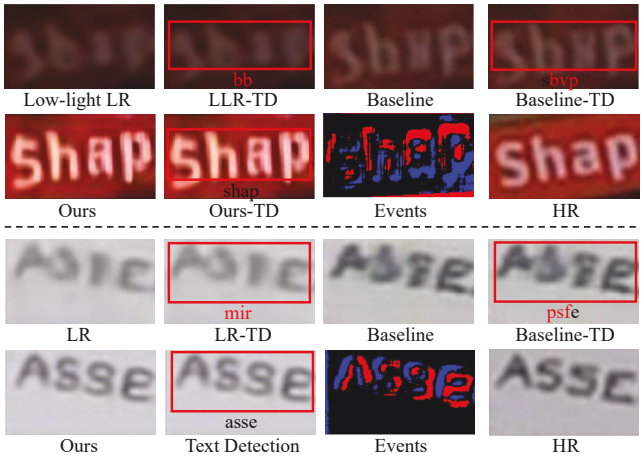

Figure 1: Examples of EvTSR. We illustrate two major challenges in TSR: poor lighting and motion blur. Both hinder accurate text localization in LR images and degrade recognition. Existing TSR methods struggle with these issues. By incorporating event cameras and our proposed EvTSR approach, text clarity and localization are significantly improved. (TD: text detection; HR: high-resolution)

In scenarios with low light or motion blur, fine-grained details such as text strokes and edges are often lost, limiting the effectiveness of existing methods (Rasheed and Shi 2022).

This limitation stems from the inherent lack of high-frequency details in degraded LR images captured by conventional cameras. Recently, event cameras have shown great promise in related tasks like low-light enhancement (Liang et al. 2024, 2023; Zhang et al. 2020), high-dynamic-range imaging (Yang et al. 2023; Han et al. 2023), and stereo depth estimation (Cheng, Knoll, and Cao 2025). Their advantages (high dynamic range (Brandli, Muller, and Delbruck 2014), high temporal resolution (Gallego et al. 2020), and rich "moving edge" data (Mitrokhin et al. 2020)) make them ideal for recovering sharp text edges. Motivated by this, we incorporate event signals as auxiliary information to enhance TSR, overcoming conventional limitations.

In this paper, we propose EvTSR, the *first* practical method that effectively integrates event data into the TSR process to restore fine-grained details and enhance text clar-

*Corresponding Author.

ity. The core of EvTSR lies in its dual stream frequency boost (DSFB) mechanism, which strategically separates image features into high-frequency (*e.g.*, sharp edges, fine details) and low-frequency (*e.g.*, text structure, layout) components, enabling targeted enhancement for each domain. High-frequency features, such as sharp text edges and fine details, are enhanced using event data through the event-guided high-frequency (EGH) mechanism, which leverages the event camera's strength in capturing details, particularly in dynamic or low-light conditions. The EGH mechanism employs the event-guided dynamic convolution to adaptively fuse the high-frequency texture and edge features from the event stream with the text features, significantly enhancing the high-frequency representation and ensuring accurate detail reconstruction. Meanwhile, low-frequency features, such as the overall text structure and layout, are refined using guidance from a pre-trained text recognizer through an effective attention mechanism. The DSFB mechanism ensures structural consistency and readability by dynamically focusing on the most relevant regions of the text, preserving the global layout and coherence of the reconstructed text. To enhance cross-modal fusion, we design the cross-modal fusion (CMF) mechanism, which seamlessly integrates high-frequency features, low-frequency features, and event data into a unified representation. The CMF mechanism dynamically balances the contributions of event and RGB features, ensuring robust performance in complex backgrounds and degraded environments. Thanks to these innovative designs, EvTSR achieves superior text detail recovery and clarity, particularly in highly challenging scenarios (see Figure 1).

Contributions of this paper are summarized as follows: (1) We propose EvTSR, the *first* method that leverages event data to enhance the task of TSR, restoring fine details and improving text clarity, especially in low-light and dynamic scenes. (2) We design the DSFB mechanism, which separates image features into high-frequency and low-frequency components. High-frequency features are enhanced using event data through adaptive event-guided dynamic convolution, while low-frequency features are refined using a pre-trained text recognizer with an attention mechanism, ensuring structural consistency. To further improve cross-modal integration, we introduce the CMF mechanism, which effectively combines high-frequency features, low-frequency features, and event data, enabling robust performance. (3) Extensive experiments show that EvTSR significantly outperforms existing TSR methods in text detail recovery, clarity, and recognition accuracy.

2 Related Work

2.1 Scene Text Image Super-Resolution

TSR recovers fine details like strokes and edges from low-resolution text images. Unlike general single image super-resolution (Xiao et al. 2023), TSR is a much more complex and task-dependent field. Early CNN-based methods (Dong et al. 2015; Wang et al. 2019) often failed to preserve critical text features effectively. Recent work shifted towards multi-task learning and textual priors. For instance, TextSR (Wang et al. 2019) uses adversarial training, while MCGAN (Wang et al. 2020b) employs multiple loss functions for text recovery. To handle irregular text and complex backgrounds, methods like C3-STISR (Zhao et al. 2022a) and TATT (Ma, Liang, and Zhang 2022) utilize attention mechanisms. TSRN (Mou et al. 2020) and STN (Jaderberg et al. 2016) also use attention, but they still rely solely on traditional image data, failing in low-light or dynamic conditions. In contrast, our method, EvTSR, integrates event data to overcome these limitations, enhancing fine detail recovery in all of these challenging environments.

2.2 Event-Guided Super-Resolution

Incorporating neuromorphic cameras into super-resolution holds strong potential (Zhao et al. 2023, 2024b; Xiao and Xiong 2025; Xiao et al. 2020), as their asynchronous, high-temporal-resolution signals can restore fine details that RGB cameras lose under motion blur or poor illumination (Ding et al. 2022; Hu et al. 2022; Zhao et al. 2022b, 2024a; Xiao, Li, and Jia 2025; Xiao et al. 2024b,c, 2022). Foundational works like eSL-Net (Wang et al. 2020a) used sparse learning to fuse event streams with low-resolution (LR) images for single-frame super-resolution (SR). Building on this, EvIntSR (Han et al. 2021) introduced a dual-path architecture that first reconstructs intermediate frames from events, then fuses them with LR inputs. For video SR, Jing et al. (Jing et al. 2021) proposed a two-stage pipeline to interpolate motion details, while Kai et al. (Kai, Zhang, and Sun 2023) designed a bidirectional framework using nonlinear motion cues for more robust temporal alignment. More recent breakthroughs in this area include implicit neural representation methods for arbitrary-scale SR (Lu et al. 2023), and advanced VSR models like AsEVSRN (Xiao et al. 2024a), which uses a content hallucination mechanism and recurrent cells. Addressing these limitations, EvTSR leverages event data to enhance high-frequency text features such as edges and strokes, while using a text recognizer to guide the refinement of low-frequency structural details, achieving superior text readability and overall quality.

3 Method

3.1 Overview

EvTSR is a novel text image super-resolution framework that is designed to integrate event-based information and frequency-aware enhancement. Given a low-resolution text image $I^{LR} \in \mathbb{R}^{H\times W\times 3}$, our goal is to reconstruct a high-resolution text image $I^{SR} \in \mathbb{R}^{H\times W\times 3}$ while preserving text details. To address motion artifacts and frequency loss, we incorporate event-driven representations and dual-stream frequency boosting, which work together to preserve text details and enhance visual clarity.

To simulate real-world camera motion, we generate a 25 fps video from I^{LR} by applying random transformations (*i.e.*, translation, rotation, zoom) and convert it into an event stream E^{LR} using the V2E method (Gehrig et al. 2020), following the parameters in its official implementation. The event stream is then transformed into a voxel grid representation via temporal bilinear interpolation ($B = 5$ bins).

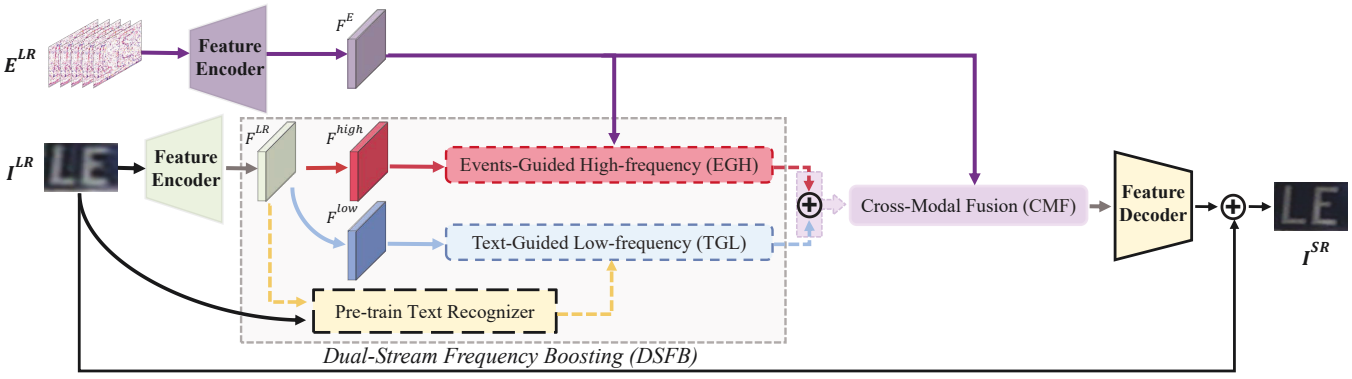

Figure 2: Overview of EvTSR, detailing the DSFB (with EGH and TGL mechanisms) and the CMF mechanism.

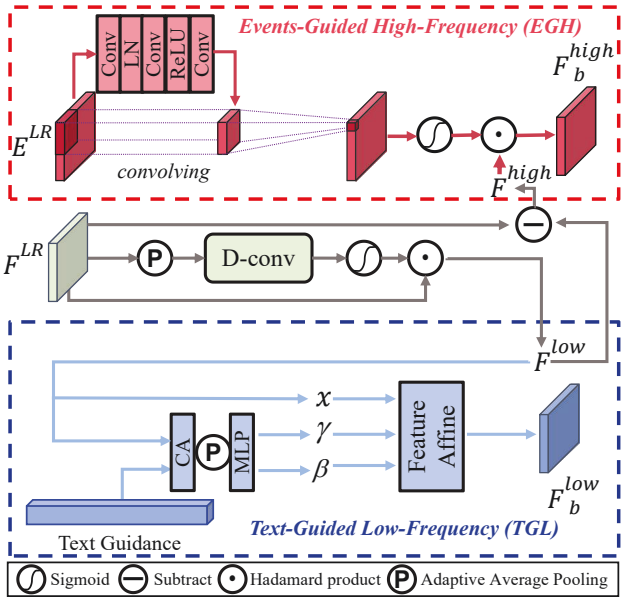

Figure 3: The structure of the DSFB mechanism.

While large-scale real datasets are unavailable, we demonstrate generalization on real-world data (Figure 9) to validate our simulation-based approach.

As illustrated in Figure 2, EvTSR consists of a feature encoder, our core DSFB mechanism, and a CMF mechanism. The DSFB mechanism decomposes features into high and low frequencies, which are then refined by the EGH and TGL branches, respectively. The CMF mechanism integrates the frequency-enhanced features with the event voxel grid to generate a refined representation. Finally, a feature decoder maps the output back to the RGB space using a 1×1 convolution, producing the super-resolved text image I^{SR}.

By leveraging event motion cues and frequency-aware enhancement, EvTSR effectively reconstructs high-quality text images, improving OCR accuracy and image quality metrics. The overall architecture is illustrated in Figure 2.

3.2 Dual-Stream Frequency boost Mechanism

Scene text super-resolution requires not only restoring fine-grained details such as text strokes but also preserving the overall text structure. However, traditional methods struggle to maintain this balance, as they rely on spatial features without explicitly modeling the frequency components. To tackle this challenge, we propose the DSFB mechanism (Figure 3), which decomposes the image into high-frequency and low-frequency components and enhances them separately to recover text details effectively.

Given an input feature F^{LR}, we first generate an adaptive filtering weight W to guide the frequency-aware transformation. This process begins with adaptive average pooling (AAP) to extract dominant structural information, followed by transformation through two 1×1 convolution layers and a gating mechanism. The complete formulation is given by:

$$W = \sigma(\Phi_D(P(F^{LR}))), \tag{1}$$

where $P(\cdot)$ represents the AAP operation, which extracts the primary structural features from F^{LR}. $\Phi_D(\cdot)$ denotes the transformation module consisting of two consecutive 1×1 convolution layers, responsible for further refining the extracted features and generating the filtering weight. The first convolution layer extracts channel-wise representations, while the second convolution layer refines them further through a gating mechanism. $\sigma(\cdot)$ is the sigmoid activation function, which normalizes the adaptive weight W.

Using this frequency decomposition, we obtain the low-frequency and high-frequency components:

$$F^{low} = W \odot F^{LR}, \tag{2}$$

$$F^{high} = F^{LR} - F^{low}, \tag{3}$$

where $\odot$ denotes element-wise multiplication.

F^{high} encodes text edges and strokes but often suffers from noise and artifacts. To refine these features, we leverage event-based information, which is well-known for its excellent ability to capture rapid scene changes and motion dynamics. The event representation E^{LR} is first processed through a convolutional transformation:

$$\hat{E}^{LR} = \Psi(E^{LR}) * E^{LR}, \tag{4}$$

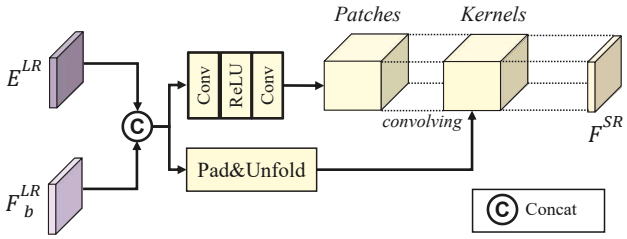

Figure 4: The structure of the CMF mechanism.

where $\Psi(\cdot)$ generates an adaptive convolutional kernel from E^{LR}. It consists of three consecutive 1×1 convolution layers. A LayerNorm operation is applied between the first and second convolution layers, while a ReLU activation function is inserted between the second and third convolution layers to enhance feature non-linearity. This enables $\Psi(\cdot)$ to extract more meaningful event-driven features.

Furthermore, $*$ denotes the standard convolution operation. By applying the generated kernel $\Psi(E^{LR})$ to E^{LR} via convolution, we obtain $\hat{E}^{LR}$, which serves as a refined event-enhanced feature.

To enhance the high-frequency features, we use an event-driven modulation strategy:

$$F_b^{high} = \sigma(\hat{E}^{LR}) \odot F^{high}, \tag{5}$$

This operation ensures that high-frequency enhancement is guided by event-based cues, improving text clarity. For low-frequency enhancement, we incorporate text guidance from a pre-trained recognizer, refining the overall structure. Specifically, the low-frequency feature F^{low} first undergoes channel attention (CA), adaptive average pooling (P), and a multilayer perceptron (MLP) to generate the modulation parameters γ and β:

$$\gamma, \beta = \text{MLP}(P(\text{CA}(F^{low}))), \tag{6}$$

where CA is channel attention, P is adaptive average pooling, and MLP consists of two 1×1 convolutions with a ReLU activation. With these learned modulation parameters, an affine transformation is then applied to F^{low}, ultimately producing the refined feature:

$$F_b^{low} = \Gamma(F^{low}, \gamma, \beta), \tag{7}$$

where $\Gamma(\cdot)$ represents the affine transformation that modulates F^{low} based on γ and β, ensuring the preservation of text structure with guidance from the pre-trained recognizer.

3.3 Cross-Modal Fusion Mechanism

To seamlessly integrate features from DSFB and avoid any potential modality mismatches, we therefore propose the CMF mechanism (Figure 4).

To enable effective interaction between the event stream representation and the low-frequency image features, we first concatenate the event feature E^{LR} with the enhanced low-frequency feature F_b^{low}:

$$F_c = \text{Concat}(E^{LR}, F_b^{low}), \tag{8}$$

This concatenation operation aggregates event and low-frequency information at the channel level while preserving the complementary nature of event data, which captures motion changes over time, and RGB features, which provide structural consistency and spatial context. This interaction ensures that low-frequency text components effectively receive the necessary event guidance, thereby enhancing overall text clarity and detail preservation.

To further refine the fused representation, we introduce a learnable dynamic convolution kernel that modulates the information exchange:

$$K = \Theta(F_c), \tag{9}$$

where K represents the dynamically generated convolutional filter, and $\Theta(\cdot)$ is a set of convolution layers that extract feature relationships and generate adaptive modulations. This adaptive mechanism tailors the fusion process to text structures and event distributions, ensuring robust and context-aware information integration.

Subsequently, we apply the generated convolution kernel to the transformed feature representation obtained by passing F_c through an MLP, allowing the network to modulate the contributions of event-based details adaptively:

$$F^{SR} = K * \text{MLP}(F_c), \tag{10}$$

where K represents the dynamically generated convolutional filter. This adaptive mechanism incorporates event data effectively, maintaining structural consistency without artifacts and preventing over-enhancement.

Finally, the fused representation F^{SR} is passed into the feature decoder to generate the high-resolution text image. The CMF mechanism is designed to ensure complementary interactions between event data and RGB structures, leveraging the strengths of both modalities to recover fine text details while preserving overall text consistency.

4 Experiments

4.1 Experimental Settings

Datasets. TextZoom (Wang et al. 2020c) is a widely used benchmark for TSR (STISR) tasks within the research community. This dataset is constructed from two single-image super-resolution datasets, RealSR (Cai et al. 2019) and SR-RAW (Zhang et al. 2019), where images are captured in real-world scenarios using digital cameras. The dataset consists of 17,367 low-resolution (LR) and high-resolution (HR) image pairs for training and 4,373 pairs for testing. The test set is divided into three subsets: simple (1,619 pairs), medium (1,411 pairs), and hard (1,343 pairs). Images with a resolution lower than 32×32 pixels are removed from the training set to ensure meaningful learning. To assess the robustness of our method across different text styles, we evaluate our model on four widely used scene text recognition (STR) benchmarks: ICDAR2015 (Karatzas et al. 2015), CUTE80 (Risnumawan et al. 2014), SVT (Wang, Babenko, and Belongie 2011), and SVTP (Phan et al. 2013), which encompass diverse real-world text scenarios. Since these specific benchmark datasets lack predefined LR-HR pairs and mainly consist of high-quality images, we must therefore

DSFB	CMF	Recognition Accuracy			
		Easy	Medium	Hard	avgAcc
-	-	59.5%	52.2%	38.8%	50.8%
✓	-	65.4%	58.5%	43.3%	56.4%
-	✓	65.5%	58.7%	43.3%	56.5%
✓	✓	**66.0%**	**60.1%**	**44.6%**	**57.5%**

Table 1: Combination of different components in EvTSR.

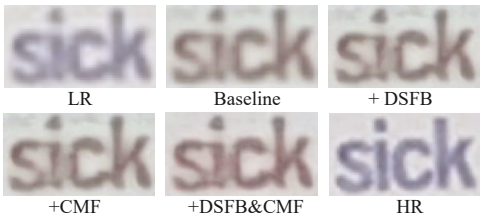

Figure 5: Comparison of SR results generated by different configurations of EvTSR.

Variant	Recognition Accuracy			
	Easy	Medium	Hard	avgAcc
Baseline	59.5%	52.2%	38.8%	50.8%
DSFB-Conv	61.2%	54.2%	39.3%	52.2%
DSFB-w/o EGH	62.4%	55.3%	40.6%	53.4%
DSFB-w/o TGL	63.1%	56.4%	41.6%	54.3%
DSFB	**66.0%**	**60.1%**	**44.6%**	**57.5%**

Table 2: The ablation results of the DSFB mechanism and its variants.

Figure 6: Qualitative comparison of SR results generated using different configurations of the DSFB mechanism.

follow the preprocessing strategy of Guo et al. (Guo et al. 2023a) to systematically generate all LR images for evaluation.

Evaluation Metrics. To evaluate the performance of TSR, we employ both text recognition accuracy and image fidelity metrics. For text recognition evaluation, we follow prior works (Wang et al. 2020c) and adopt three commonly used text recognizers: CRNN (Shi, Bai, and Yao 2017), MORAN (Luo, Jin, and Sun 2019), and ASTER (Shi et al. 2019), in order to ensure the robustness of our method across different recognition models.

Implementation Details. We use the public ABINet (Fang et al. 2021) as our text recognizer. The input LR images are set to a size of 32×32. The training process includes 30K iterations, utilizing the Adam optimizer (Kingma and Ba 2014) with an initial learning rate of 2×10^{-3} and a cosine annealing learning rate scheduler (Loshchilov and Hutter 2016). We use random horizontal and vertical flips for data augmentation. The Charbonnier loss function is used for supervision. The entire training process is conducted on an NVIDIA RTX 3090 GPU.

4.2 Ablation Studies

In this section, we conduct an ablation study to evaluate each component's effectiveness in EvTSR. CRNN (Shi, Bai, and Yao 2017) is chosen as the text recognizer for uniformity.

Effectiveness of Two Mechanisms in EvTSR. The DSFB and CMF mechanisms serve as two essential components in EvTSR. To assess their contributions, we conduct ablation studies by removing each mechanism and analyzing the performance changes. Table 1 presents the results across different difficulty levels. Additionally, Figure 5 provides a visual comparison of the SR results obtained under different configurations, illustrating each mechanism's contribution to fine-grained details. Removing both DSFB and CMF drops accuracy to 50.8%, suggesting the model struggles to effectively reconstruct fine-grained text details without our frequency decomposition and event-aware fusion. Introducing DSFB alone improves the average accuracy to 56.4%, clearly showing frequency boosting's effectiveness in preserving text structures. Similarly, incorporating CMF without DSFB results in a comparable improvement, yielding an average accuracy of 56.5%, highlighting event-based enhancement's effectiveness. Finally, when both mechanisms are included, EvTSR achieves the best performance, with an average accuracy of 57.5%. This confirms DSFB and CMF are complementary in enhancing text restoration, with DSFB focusing on structural integrity, while CMF refines fine details via event information.

Investigation of the DSFB Mechanism. The DSFB mechanism is designed to separately enhance the high-frequency and low-frequency components of the image, improving the clarity and readability of super-resolved text images. To evaluate its effectiveness, we analyze different variants and present the results in Table 2. Additionally, Figure 6 provides a visual comparison of the SR results obtained using different DSFB configurations, illustrating how each component contributes to text restoration. First, we replace the dynamic convolution in DSFB with a standard 1×1 convolution, denoted as DSFB-Conv. The results indicate that although this variant improves super-resolution performance, its accuracy is lower than the full DSFB mechanism. This highlights the importance of dynamic convolution in enhancing high-frequency features. Next, we evaluate two ablated versions by removing the EGH branch (DSFB-w/o EGH) and the TGL branch (DSFB-w/o TGL), respectively. The results show that both variants lead to performance degradation. Specifically, removing EGH reduces the model's ability to recover high-frequency details, such as text edges and fine strokes, while removing TGL deteriorates the structural consistency of the reconstructed text. This observation suggests that EGH is essential for high-frequency enhancement by leveraging event data, whereas TGL plays a crucial role in preserving text structure and maintaining proper alignment in low-frequency components. The full DSFB mechanism achieves the best performance across all test sets, further

Method	ASTER (Shi et al. 2019)				MORAN (Luo, Jin, and Sun 2019)				CRNN (Shi, Bai, and Yao 2017)			
	Easy	Medium	Hard	Average	Easy	Medium	Hard	Average	Easy	Medium	Hard	Average
Bicubic	67.4%	42.4%	31.2%	48.2%	60.6%	37.9%	30.8%	44.1%	36.4%	21.1%	21.1%	26.8%
HR	94.2%	87.7%	76.2%	86.6%	91.2%	85.3%	74.2%	84.1%	76.4%	75.1%	64.6%	72.4%
SRCNN	70.6%	44.0%	31.5%	50.0%	63.9%	40.0%	29.4%	45.6%	41.1%	22.3%	22.0%	29.2%
SRResNet	69.4%	50.5%	35.7%	50.7%	66.0%	47.1%	33.4%	45.6%	45.2%	32.6%	25.5%	35.1%
RCAN	67.3%	46.4%	35.1%	50.7%	63.1%	42.9%	35.7%	47.5%	46.8%	27.9%	26.5%	34.5%
SAN	68.1%	46.7%	35.1%	50.7%	66.5%	44.4%	35.7%	49.4%	50.1%	31.2%	28.1%	37.2%
TSRN	75.1%	56.3%	41.6%	58.3%	70.1%	55.3%	37.9%	55.4%	52.5%	38.2%	31.7%	41.4%
TBSRN	75.7%	59.9%	41.6%	60.1%	74.1%	57.0%	40.0%	58.4%	59.6%	47.1%	35.3%	48.1%
PCAN	77.5%	60.2%	42.4%	61.3%	73.9%	57.6%	41.0%	58.5%	59.6%	45.4%	34.8%	47.4%
TG	77.9%	63.4%	45.4%	61.8%	75.2%	57.6%	41.0%	59.4%	61.2%	47.6%	35.5%	48.9%
TATT	78.9%	63.3%	46.8%	64.1%	72.5%	55.3%	41.4%	59.5%	62.6%	53.6%	39.8%	52.6%
C3-STISR	79.1%	63.6%	46.4%	64.7%	74.2%	61.0%	43.2%	59.5%	65.2%	53.6%	39.7%	53.7%
LEMMA	81.1%	66.3%	47.4%	66.0%	77.7%	64.4%	44.6%	63.2%	**67.1**%	58.8%	40.6%	56.3%
EvTSR	**83.1**%	**68.3**%	**49.2**%	**67.9**%	**78.0**%	**65.5**%	**47.6**%	**64.6**%	66.0%	**60.1%**	**44.6%**	**57.5%**

Table 3: Comparison of the downstream text recognition accuracy on the TextZoom dataset. The best result is in bold.

Variant	Recognition Accuracy			
	Easy	Medium	Hard	avgAcc
Baseline	59.5%	52.2%	38.8%	50.8%
CMF-Sum&Conv	60.7%	53.4%	39.1%	51.7%
CMF-Conv	61.3%	54.7%	39.6%	52.5%
CMF-Sum	64.2%	57.6%	42.3%	55.3%
CMF	**66.0**%	**60.1**%	**44.6**%	**57.5**%

Table 4: The ablation results of the CMF mechanism.

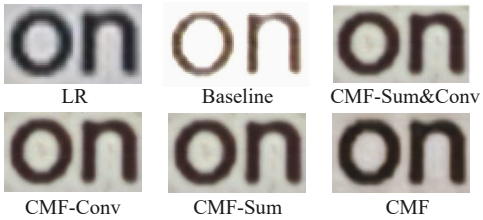

Figure 7: Qualitative comparison of SR results generated using different configurations of the CMF mechanism.

confirming its effectiveness in the text super-resolution task.

Investigation of the CMF Mechanism. The CMF mechanism is designed to integrate event data with RGB image features to enhance super-resolution reconstruction. To evaluate its effectiveness, we conduct ablation studies by designing multiple variants, as shown in Table 4. Furthermore, Figure 7 provides a qualitative comparison of the SR results under different CMF configurations, illustrating the impact of each component on text clarity and structure. Compared to the complete CMF model achieving 57.5% accuracy, the CMF-Sum&Conv variant leads to 51.7% recognition accuracy, suggesting that concatenation is more suitable for fusing event data with RGB structural features, while the dynamic convolution plays a key role in cross-modal interactions. Replacing the dynamic kernel with a 1×1 convolution (CMF-Conv) drops accuracy by 5.0%, showing adaptive filtering is crucial. The CMF-Sum variant only replaces concatenation with element-wise summation, leading to a 2.2% reduction in recognition accuracy, indicating that explicit concatenation is crucial for cross-modal feature interactions. These results further validate the complementary nature of concatenation and dynamic convolution: while concatenation preserves complete information, the dynamic convolution adaptively adjusts based on varying event inputs, achieving optimal super-resolution performance.

4.3 Quantitative and Qualitative Results

We compare the proposed EvTSR against previous leading methods. These include general-purpose SR methods such as SRCNN (Dong et al. 2016), SRResNet (Xiao et al. 2025), RCAN (Zhang et al. 2018), and SAN (Dai et al. 2019), as well as leading TSR-specific methods including TSRN (Mou et al. 2020), TBSRN (Chen, Li, and Xue 2021), PCAN (Zhao et al. 2021), TG (Chen et al. 2021), TATT (Ma, Liang, and Zhang 2022), C3-STISR (Zhao et al. 2022a), and LEMMA (Guo et al. 2023b). This quantitative evaluation is conducted on the TextZoom dataset and STR benchmarks using three widely-used text recognizers: ASTER (Shi et al. 2019), MORAN (Luo, Jin, and Sun 2019), and CRNN (Shi, Bai, and Yao 2017), thereby ensuring a fair and comprehensive comparison.

Results on TextZoom. Table 3 presents the recognition accuracy across different difficulty levels (Easy, Medium, Hard) on the TextZoom dataset. Our EvTSR achieves the best results among all competing methods, consistently outperforming prior works across all recognizers. Specifically, when using ASTER, our method improves the average recognition accuracy to 67.9%, surpassing the previous best method LEMMA by 1.9%. Notably, our model achieves this with 37.1M parameters, a slightly smaller size compared to LEMMA (37.9M), demonstrating strong performance with minimal complexity. Similarly, for MORAN and CRNN, EvTSR attains 64.6% and 57.5% accuracy, exhibiting superior performance over prior approaches. Furthermore, our method also shows superior performance over recent diffusion-based models; for instance, TCDM

Figure 8: Qualitative comparison with other methods. Zoom in for better visualization.

Method	STR Datasets			
	IC15	CUTE80	SVT	SVTP
Bicubic	9.5%	35.8%	3.3%	10.2%
SRResnet	13.0%	48.3%	9.3%	12.1%
TBSRN	20.7%	75.0%	12.2%	17.4%
TATT	28.6%	74.0%	14.0%	25.9%
C3-STISR	22.7%	71.5%	10.2%	17.7%
EvTSR (Ours)	**33.4%**	**77.1%**	**23.3%**	**31.1%**

Table 5: Comparison on scene text recognition benchmarks.

Figure 9: Real-world test results of EvTSR.

(Noguchi, Fukuda, and Yamanaka 2024) achieves 65.5% with ASTER, whereas EvTSR reaches 67.9%. Furthermore, our method also shows strong image fidelity (PSNR/SSIM): Easy (23.05/0.7981), Medium (21.03/0.7401), and Hard (19.75/0.6747). These quantitative improvements, along with the qualitative comparisons in Figure 8, highlight the significantly enhanced clarity and improved legibility of text images reconstructed by EvTSR.

Evaluation on STR Benchmarks. To assess the generalization capability of our model on more challenging datasets, we evaluate EvTSR on four widely-used STR benchmarks: IC15 (Karatzas et al. 2015), CUTE80 (Risnumawan et al. 2014), SVT (Wang, Babenko, and Belongie 2011), and SVTP (Phan et al. 2013). The results, shown in Table 5, indicate that our model consistently outperforms previous approaches, achieving the highest recognition accuracy on all datasets. Notably, EvTSR obtains a significant improvement on IC15 (33.4%) and SVTP (31.1%), demonstrating its robustness in handling complex backgrounds and challenging real-world scenarios. These results validate that EvTSR not only excels in the constrained TextZoom dataset but also generalizes well to diverse text images in the wild.

Generalization to Real-World Testing. To further validate the effectiveness of EvTSR beyond synthetic datasets, we conduct real-world testing using images captured in real environments along with event data obtained from an event camera. Figure 9 presents the qualitative results. Compared to the LR input, EvTSR significantly enhances text clarity and sharpness, recovering fine-grained details such as text strokes and edges. These results further demonstrate the practicality of event-based super-resolution in real-world applications, showcasing the robustness of our method in handling real captured event streams. Furthermore, our language-agnostic EGH mechanism enables potential extension to low-resource languages.

5 Conclusion

In this paper, we propose EvTSR, a novel event-driven scene text super-resolution framework designed to address the challenge of recovering fine-grained text details under low-light conditions and fast motion scenarios. EvTSR effectively integrates event data into the super-resolution process, leveraging the high temporal resolution and motion sensitivity of event cameras to enhance high-frequency details such as text strokes and edges. To achieve this, we introduce two key mechanisms: the DSFB mechanism, which separates and enhances high- and low-frequency components, and the CMF mechanism, which facilitates cross-modal feature integration. Extensive experiments on both synthetic and real-world scene text datasets validate that EvTSR consistently outperforms advanced methods in both text clarity and recognition accuracy, providing a robust solution for scene text image super-resolution.

Acknowledgments

This work is partly supported by the grants of the National Natural Science Foundation of China under Nos. 62476077, U24A20332, 62272142, and 62076086.

References

Al-Shemarry, M. S.; Li, Y.; and Abdulla, S. 2022. Identifying license plates in distorted vehicle images: detecting distorted vehicle licence plates using a novel preprocessing methods with hybrid feature descriptors. *IEEE Intelligent Transportation Systems Magazine*, 15(2): 6–25.

Brandli, C.; Muller, L.; and Delbruck, T. 2014. Real-time, high-speed video decompression using a frame-and event-based DAVIS sensor. In *ISCAS*.

Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; and Zhang, L. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*.

Chen, J.; Li, B.; and Xue, X. 2021. Scene Text Telescope: Text-Focused Scene Image Super-Resolution. In *CVPR*.

Chen, J.; Yu, H.; Ma, J.; Li, B.; and Xue, X. 2021. Text Gestalt: Stroke-Aware Scene Text Image Super-Resolution. arXiv:2112.08171.

Cheng, Y.; Knoll, A.; and Cao, H. 2025. UR-Net: uncertainty-aware refinement network for event-based stereo depth estimation. *Visual Intelligence*, 3(1): 18.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-Order Attention Network for Single Image Super-Resolution. In *CVPR*.

Ding, Z.; Zhao, R.; Zhang, J.; Gao, T.; Xiong, R.; Yu, Z.; and Huang, T. 2022. Spatio-temporal recurrent networks for event-based optical flow estimation. In *AAAI*.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2016. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 295–307.

Dong, C.; Zhu, X.; Deng, Y.; Loy, C. C.; and Qiao, Y. 2015. Boosting Optical Character Recognition: A Super-Resolution Approach. *CoRR*, abs/1506.02211.

Fang, S.; Xie, H.; Wang, Y.; Mao, Z.; and Zhang, Y. 2021. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *CVPR*.

Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; et al. 2020. Event-based vision: A survey. *IEEE TPAMI*.

Gehrig, D.; Gehrig, M.; Hidalgo-Carrió, J.; and Scaramuzza, D. 2020. Video to Events: Recycling Video Datasets for Event Cameras. In *CVPR*.

Guo, H.; Dai, T.; Meng, G.; and Xia, S.-T. 2023a. Towards Robust Scene Text Image Super-resolution via Explicit Location Enhancement. arXiv:2307.09749.

Guo, H.; Dai, T.; Meng, G.; and Xia, S.-T. 2023b. Towards robust scene text image super-resolution via explicit location enhancement. *arXiv preprint arXiv:2307.09749*.

Han, J.; Yang, Y.; Duan, P.; Zhou, C.; Ma, L.; Xu, C.; Huang, T.; Sato, I.; and Shi, B. 2023. Hybrid high dynamic range imaging fusing neuromorphic and conventional images. *IEEE TPAMI*.

Han, J.; Yang, Y.; Zhou, C.; Xu, C.; and Shi, B. 2021. EvIntSR-Net: Event Guided Multiple Latent Frames Reconstruction and Super-resolution. In *ICCV*.

Hu, L.; Zhao, R.; Ding, Z.; Ma, L.; Shi, B.; Xiong, R.; and Huang, T. 2022. Optical flow estimation for spiking camera. In *CVPR*.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2016. Spatial Transformer Networks. arXiv:1506.02025.

Jing, Y.; Yang, Y.; Wang, X.; Song, M.; and Tao, D. 2021. Turning Frequency to Resolution: Video Super-resolution via Event Cameras. In *CVPR*.

Kai, D.; Zhang, Y.; and Sun, X. 2023. Video Super-Resolution Via Event-Driven Temporal Alignment. In *ICIP*.

Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V. R.; Lu, S.; et al. 2015. ICDAR 2015 competition on robust reading. In *ICDAR*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liang, G.; Chen, K.; Li, H.; Lu, Y.; and Wang, L. 2024. Towards Robust Event-guided Low-Light Image Enhancement: A Large-Scale Real-World Event-Image Dataset and Novel Approach. In *CVPR*.

Liang, J.; Yang, Y.; Li, B.; Duan, P.; Xu, Y.; and Shi, B. 2023. Coherent event guided low-light video enhancement. In *ICCV*.

Liem, H. D.; Minh, N. D.; Trung, N. B.; Duc, H. T.; Hiep, P. H.; Dung, D. V.; and Vu, D. H. 2018. Fvi: An end-to-end vietnamese identification card detection and recognition in images. In *NICS*.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lu, Y.; Wang, Z.; Liu, M.; Wang, H.; and Wang, L. 2023. Learning Spatial-Temporal Implicit Neural Representations for Event-Guided Video Super-Resolution. In *CVPR*.

Luo, C.; Jin, L.; and Sun, Z. 2019. A Multi-Object Rectified Attention Network for Scene Text Recognition. arXiv:1901.03003.

Ma, J.; Liang, Z.; and Zhang, L. 2022. A Text Attention Network for Spatial Deformation Robust Scene Text Image Super-resolution. arXiv:2203.09388.

Mao, Y.; Xiao, Z.; An, P.; Liu, D.; and Shan, C. 2025. Deep Sparse-to-Dense Inbetweening for Multi-View Light Fields. *IEEE Transactions on Image Processing*.

Mitrokhin, A.; Hua, Z.; Fermuller, C.; and Aloimonos, Y. 2020. Learning visual motion segmentation using event surfaces. In *CVPR*.

Mou, Y.; Tan, L.; Yang, H.; Chen, J.; Liu, L.; Yan, R.; and Huang, Y. 2020. PlugNet: Degradation Aware Scene Text

Recognition Supervised by a Pluggable Super-Resolution Unit. In *ECCV*.

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

Phan, T. Q.; Shivakumara, P.; Tian, S.; and Tan, C. L. 2013. Recognizing text with perspective distortion in natural scenes. In *ICCV*.

Rasheed, M. T.; and Shi, D. 2022. LSR: Lightening super-resolution deep network for low-light image enhancement. *Neurocomputing*, 505: 263–275.

Reddy, S.; Mathew, M.; Gomez, L.; Rusinol, M.; Karatzas, D.; and Jawahar, C. 2020. Roadtext-1k: Text detection & recognition dataset for driving videos. In *ICRA*.

Risnumawan, A.; Shivakumara, P.; Chan, C. S.; and Tan, C. L. 2014. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18): 8027–8048.

Shi, B.; Bai, X.; and Yao, C. 2017. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11): 2298–2304.

Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; and Bai, X. 2019. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9): 2035–2048.

Wang, B.; He, J.; Yu, L.; Xia, G.-S.; and Yang, W. 2020a. Event enhanced high-quality image recovery. In *ECCV*.

Wang, K.; Babenko, B.; and Belongie, S. 2011. End-to-end scene text recognition. In *ICCV*.

Wang, W.; Xie, E.; Liu, X.; Wang, W.; Liang, D.; Shen, C.; and Bai, X. 2020b. Scene Text Image Super-Resolution in the Wild. *CoRR*, abs/2005.03341.

Wang, W.; Xie, E.; Liu, X.; Wang, W.; Liang, D.; Shen, C.; and Bai, X. 2020c. Scene text image super-resolution in the wild. In *ECCV*.

Wang, W.; Xie, E.; Sun, P.; Wang, W.; Tian, L.; Shen, C.; and Luo, P. 2019. TextSR: Content-Aware Text Super-Resolution Guided by Recognition. *CoRR*, abs/1909.07113.

Xiao, Z.; Bai, J.; Lu, Z.; and Xiong, Z. 2023. A dive into sam prior in image restoration. *arXiv preprint arXiv:2305.13620*.

Xiao, Z.; Kai, D.; Zhang, Y.; Sun, X.; and Xiong, Z. 2024a. Asymmetric Event-Guided Video Super-Resolution. In *ACM MM*.

Xiao, Z.; Kai, D.; Zhang, Y.; Zha, Z.-J.; Sun, X.; and Xiong, Z. 2024b. Event-Adapted Video Super-Resolution. In *ECCV*.

Xiao, Z.; Kai, D.; Zhang, Y.; Zha, Z.-J.; Sun, X.; and Xiong, Z. 2025. Event-Adapted Video Super-Resolution. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision – ECCV 2024*, 217–235. Cham: Springer Nature Switzerland. ISBN 978-3-031-72946-1.

Xiao, Z.; Li, Z.; and Jia, W. 2025. Occlusion-Embedded Hybrid Transformer for Light Field Super-Resolution. In *AAAI*.

Xiao, Z.; Lu, Z.; Bi Mi, M.; Xiong, Z.; and Wang, X. 2024c. Unraveling Motion Uncertainty for Local Motion Deblurring. In *ACM MM*.

Xiao, Z.; and Wang, X. 2025. Event-based Video Super-Resolution via State Space Models. In *CVPR*.

Xiao, Z.; Weng, W.; Zhang, Y.; and Xiong, Z. 2022. EVA2: Event-Assisted Video Frame Interpolation via Cross-Modal Alignment and Aggregation. *IEEE Transactions on Computational Imaging*, 8: 1145–1158.

Xiao, Z.; and Xiong, Z. 2025. Incorporating degradation estimation in light field spatial super-resolution. *Computer Vision and Image Understanding*, 252: 104295.

Xiao, Z.; Xiong, Z.; Fu, X.; Liu, D.; and Zha, Z.-J. 2020. Space-time video super-resolution using temporal profiles. In *ACMMM*.

Yang, Y.; Han, J.; Liang, J.; Sato, I.; and Shi, B. 2023. Learning event guided high dynamic range video reconstruction. In *CVPR*.

Zhang, S.; Zhang, Y.; Jiang, Z.; Zou, D.; Ren, J.; and Zhou, B. 2020. Learning to see in the dark with events. In *ECCV*.

Zhang, X.; Chen, Q.; Ng, R.; and Koltun, V. 2019. Zoom to learn, learn to zoom. In *CVPR*.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. arXiv:1807.02758.

Zhao, C.; Feng, S.; Zhao, B. N.; Ding, Z.; Wu, J.; Shen, F.; and Shen, H. T. 2021. Scene text image super-resolution via parallelly contextual attention network. In *ACM MM*.

Zhao, M.; Wang, M.; Bai, F.; Li, B.; Wang, J.; and Zhou, S. 2022a. C3-STISR: Scene Text Image Super-resolution with Triple Clues. arXiv:2204.14044.

Zhao, R.; Xiong, R.; Zhang, J.; Yu, Z.; Zhu, S.; Ma, L.; and Huang, T. 2023. Spike camera image reconstruction using deep spiking neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6): 5207–5212.

Zhao, R.; Xiong, R.; Zhang, J.; Zhang, X.; Yu, Z.; and Huang, T. 2024a. Optical Flow for Spike Camera with Hierarchical Spatial-Temporal Spike Fusion. In *AAAI*.

Zhao, R.; Xiong, R.; Zhao, J.; Yu, Z.; Fan, X.; and Huang, T. 2022b. Learning optical flow from continuous spike streams. In *NeurIPS*.

Zhao, R.; Xiong, R.; Zhao, J.; Zhang, J.; Fan, X.; Yu, Z.; and Huang, T. 2024b. Boosting spike camera image reconstruction from a perspective of dealing with spike fluctuations. In *CVPR*.