EveCamNet: A Fusion Framework of Event and RGB Camera Towards Detecting Drivable Area for Autonomous Vehicles

Jyoti M Satihal[a], Safa Imtihaz Sayyad[b] and Ujwala Patil[c]

KLE Technological University, Vidyanagar, Hubballi, Karnataka, 580031, India

Keywords: Event Camera, Drivable Area, Dynamic Conditions.

Abstract: We propose EveCamNet, a novel framework that fuses features from RGB and event cameras to detect drivable areas in dynamic scenarios. This framework enhances semantic segmentation for identifying drivable regions in autonomous driving systems. While RGB cameras excel at capturing spatial and textural details, they often struggle in dynamic conditions, such as motion blur and varying lighting. In contrast, event cameras offer high temporal resolution and low latency, but they lack detailed spatial context. Our proposed framework effectively combines the strengths of both modalities through attention-based feature fusion mechanisms and a robust loss function incorporating Cross-Entropy and Dice losses. Using an open-source dataset, we evaluate the performance of EveCamNet, achieving a mean Intersection over Union (mIoU) of 69.94% and Pixel Accuracy (PA) of 90.54%. These results highlight the potential of RGB-Event fusion as a promising approach for advancing autonomous driving systems.

1 INTRODUCTION

The rapid advancements in autonomous driving have made intelligent mobility an important aspect of modern transportation. It is crucial to understand the environment to ensure safe and efficient navigation. However, to achieve this, it is important to understand the challenges it poses, especially in diverse and complex scenarios. One of the fundamental tasks in this domain is the detection of drivable areas, which serves as the basis for path planning, obstacle avoidance, and overall navigation. This task becomes even more difficult in unstructured environments, such as settings with dynamic objects, varying terrains, or rural roads that lack clear lane markings.

Semantic segmentation has emerged as a key technique to address the challenges of drivable area detection (Qiao and Zulkernine, 2021) (Jain et al., 2023). Unlike traditional object detection (Zou et al., 2023), which provides bounding-box outputs, semantic segmentation offers pixel-level classification of road surfaces and surrounding elements. This detailed scene understanding is crucial for accurately distinguishing between drivable and non-drivable areas, even in the presence of environmental variations.

RGB cameras are widely used for semantic segmentation due to their ability to capture detailed spatial and color information. They are cost-effective, scalable, and capable of handling a variety of road scenarios. However, they face significant limitations in dynamic environments, such as those involving fast-moving objects, motion blur, or changes in illumination. These challenges necessitate the use of complementary sensors to fill in the gaps left by RGB data.

In recent applications of autonomous driving, event cameras have attracted attention due to their unique functionality. Unlike conventional cameras, event cameras operate asynchronously, capturing pixel-level changes in brightness with microsecond-level temporal resolution. Their advantages consist of extremely high temporal resolution and low latency, both of which are measured in microseconds. Additionally, they offer a dynamic range of 140dB, significantly higher than the 60dB typically found in standard cameras, along with lower power consumption (Shariff et al., 2024) (Gallego et al., 2022). This makes them highly effective in scenarios with rapid motion or abrupt lighting changes, where traditional cameras often struggle. Their sparse data output and resilience to motion blur make them particularly suitable for detecting dynamic objects. Despite their

[a] https://orcid.org/0009-0009-5800-0496
[b] https://orcid.org/0009-0002-9888-5865
[c] https://orcid.org/0000-0001-5776-785X

Satihal, J. M., Sayyad, S. I. and Patil, U.
EveCamNet: A Fusion Framework of Event and RGB Camera Towards Detecting Drivable Area for Autonomous Vehicles.
DOI: 10.5220/0013588200004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 2*, pages 135-143
ISBN: 978-989-758-763-4

strengths, event cameras lack the spatial richness required for tasks such as semantic segmentation. Additionally, aggregating event data into frame-based representations requires careful handling to align temporal and spatial information effectively.

To overcome the individual limitations of RGB and event cameras, RGB-Event fusion has emerged as a promising approach. By combining the complementary strengths of these modalities, fusion-based frameworks can generate richer and more reliable scene representations. For semantic segmentation, this fusion enables models to leverage the spatial detail of RGB data and the motion sensitivity of event data, resulting in improved performance across diverse driving conditions. While RGB-Event fusion has been explored for tasks like object detection and tracking (Zhou et al., 2023) (Tomy et al., 2022), its application to semantic segmentation for drivable area detection remains underexplored.

In this paper, we propose a novel RGB-Event fusion framework for semantic segmentation to improve drivable area detection in autonomous driving. The framework incorporates mid-level cues and utilizes an attention-based fusion mechanism to effectively integrate and refine RGB and event features across spatial and channel dimensions. This approach ensures accurate and consistent segmentation, making it suitable for challenging conditions. Our key contributions include:

- Integration of RGB and event data to enhance segmentation accuracy by combining the spatial richness of RGB images with the temporal sensitivity of event data, addressing the limitations of standalone modalities.
- Implementation of a robust combined loss function to balance pixel-level accuracy and boundary precision, improving the model's ability to handle segmentation tasks.
- Evaluation of the model's effectiveness using metrics such as mean Intersection over Union (mIoU) and Pixel Accuracy, demonstrating its applicability and reliability across diverse and real-world driving scenarios using DDD17.

The remainder of this paper is structured as follows: Section 2 reviews related work on various segmentation techniques that utilize different modalities. Section 3 describes the proposed methodology, focusing on input representation, feature extraction, fusion, and the segmentation process. Section 4 presents the experimental results and compares them with baseline methods. Section 5 discusses the findings from the ablation studies. Finally, Section 6 concludes the paper and explores potential future directions.

2 LITERATURE SURVEY

Semantic segmentation is an essential technique in autonomous driving that allows vehicles to categorize each pixel in an image into specific classes. By doing so, it significantly enhances the vehicle's ability to understand and perceive its environment at a detailed level. The introduction of deep learning, particularly through Convolutional Neural Networks (CNNs), has transformed semantic segmentation by providing more accurate and scalable solutions.

Early deep learning models, particularly Fully Convolutional Networks (FCNs) (Long et al., 2015), laid the groundwork for semantic segmentation by allowing for end-to-end learning, which demonstrated strong performance on RGB datasets like Cityscapes (Cordts et al., 2016) and KITTI (Geiger et al., 2013). Semantic segmentation for drivable area has progressed through two primary approaches: monocular vision-based methods and multimodal sensor-based methods (Rasib et al., 2021).

Monocular vision-based methods primarily utilize RGB cameras, which makes them more cost-effective and easier to implement. Notable examples of these methods include road detection using neural networks, as demonstrated by Li et al. (Li et al., 2022), and instance segmentation techniques employed by Chan et al (Chan et al., 2019). Models like ENet (Almeida et al., 2020) and EdgeNet (Han et al., 2021) have achieved significant improvements in pixel-level segmentation accuracy for both structured and unstructured roads. U-Net (Siddique et al., 2021) has gained popularity in autonomous driving due to its encoder-decoder architecture and skip connections, which effectively capture both spatial and semantic features. DeepLabv3+ (Liu et al., 2021) enhances segmentation capabilities even further by utilizing improved convolutions and spatial pyramid pooling, allowing for better handling of complex scenes.

Despite these advancements, methods relying solely on RGB cameras struggle in dynamic environments and varying weather conditions. While RGB cameras provide rich spatial and color information, they face limitations such as motion blur, changing lighting conditions, and difficulties in capturing unstructured roads without clear lane markings. These challenges underscore the need to explore complementary sensing modalities.

Multimodal methods that incorporate LiDAR and point clouds utilize spatial and depth information to enhance detection capabilities. For example, Demir et al. (Demir et al., 2017) developed a point-cloud-based adaptive method, while Yang et al. (Yang et al.,

2020) introduced a fusion network that combines LiDAR with image data. However, these approaches often face challenges such as high costs and limited datasets.

A recent advancement in the industry is the use of event cameras, which provide significant advantages over conventional cameras in various scenarios. Event cameras have emerged as a promising alternative due to their unique ability to capture changes in brightness at the pixel level with high temporal resolution. Unlike traditional cameras, event cameras operate asynchronously, making them resilient to motion blur and variations in lighting (Shariff et al., 2024) (Gallego et al., 2022).

Research has demonstrated that event cameras are particularly effective in dynamic scenes involving fast-moving objects or rapid transitions between light and shadow. Recent advancements have explored their utility in various fields, including robotics, augmented reality, and object tracking. The unique ability of event cameras to operate efficiently under challenging lighting conditions and their low power consumption further enhance their suitability for wearable devices and autonomous systems. However, the unconventional output of these sensors necessitates developing specialized algorithms to exploit their potential fully. Emerging methods, ranging from low-level vision tasks such as optical flow estimation to high-level applications like recognition and segmentation, underscore the transformative impact of this technology on computer vision and robotics (Chakravarthi et al., 2024).

The first significant work in this domain introduced an event-based semantic segmentation dataset derived from DDD17, utilizing an Xception-type network to demonstrate robust performance, particularly in edge-case scenarios like overexposed images (Alonso and Murillo, 2019) (Binas et al., 2017) (Chollet, 2017). Semantic labels for the dataset were generated using a pre-trained network on grayscale frames from the DAVIS346B sensor, which aligns events with frames. However, the DAVIS346B's low resolution and image quality introduced significant artifacts and limited label granularity. Simulated datasets like EventScape, recorded in CARLA, offer higher-quality labels but suffer from sim-to-real gaps, reducing real-world applicability (Gehrig et al., 2021) (Hidalgo-Carrió et al., 2020). Follow-up studies leveraged synthetic events from video datasets to enhance performance, while others explored combining labeled image datasets like Cityscapes with unlabeled events to reduce dependence on video data (Cordts et al., 2016) (Gehrig et al., 2019) (Wang et al., 2021). Despite these advancements, methods relying solely on events face challenges due to their sparse data output and lack of spatial richness.

Standalone event-based segmentation struggles with providing detailed scene understanding due to the inherently sparse and asynchronous nature of event data. Techniques like Temporal Multi-Scale Aggregation (ETMA) (Zhou et al., 2023) have attempted to address this by aggregating temporal information into event frames, improving feature utility. However, even with such advancements, event-only approaches fall short in high-resolution semantic segmentation tasks, particularly in scenarios requiring fine-grained pixel-level classification.

To overcome these limitations, fusion-based techniques integrating RGB and event data have been proposed. Fusion frameworks combine the spatial richness of RGB images with the temporal sensitivity of event data, offering a more comprehensive understanding of the scene. While early works focused on object detection and tracking (Zhou et al., 2023) (Tomy et al., 2022), where fusion demonstrated improved accuracy in dynamic environments, applications to semantic segmentation remain limited. Existing methods often rely on simple feature concatenation or static weighting, failing to fully exploit the complementary strengths of the two modalities.

Our project addresses these gaps by proposing a novel RGB-Event fusion framework for semantic segmentation, specifically designed for drivable area detection in autonomous driving. By leveraging attention-based mechanisms for feature refinement and alignment, our framework dynamically integrates RGB and event features, addressing their individual limitations and enhancing segmentation performance. This approach aims to advance the field by optimizing RGB-Event fusion for pixel-level tasks, ensuring robust and scalable solutions for autonomous driving systems.

3 EVECAMNET'S PROPOSED FRAMEWORK

This section outlines the methodology for integrating RGB and event data to achieve semantic segmentation of drivable areas. The proposed framework, illustrated in Figure 1, describes the overall architecture, which consists of several essential components divided into five stages. These components include temporal multi-scale aggregation of event data, feature extraction specific to each modality, fusion of features from both types, and the final decoding process for segmentation. Each element is carefully designed to address the unique challenges posed by dynamic driving environments. The following subsections will provide a detailed explanation of each component and its role within the overall framework.

3.1 Input Representation

The proposed model of RGB and event fusion for semantic segmentation utilizes input derived from the DDD17 dataset (Binas et al., 2017), which contains synchronized RGB frames and event data. The RGB frames are captured at a standard rate, providing rich spatial and textural information vital for understanding scenes. Each RGB frame has a resolution of H × W, where H represents the height and W denotes the width.

In contrast, event data consists of asynchronous brightness changes recorded by event cameras. This data is presented as a stream of events, each defined by a timestamp, polarity (indicating the direction of brightness change), and spatial coordinates. To adapt the event data for deep learning architectures, it is aggregated over discrete temporal intervals and transformed into a multi-channel image format. For this study, the event data is categorized into three temporal scales: short-range, medium-range, and long-range, each highlighting different motion dynamics from fine details to broader, slower movements. The RGB frames and event data are preprocessed and resized to 288 × 288 to ensure consistency and compatibility with the model.

3.2 Feature Extraction

Feature extraction for RGB and event data is conducted separately using distinct encoders tailored to each modality.

RGB Data: The RGB data is processed through a ResNet-101 encoder, a deep convolutional neural network designed to capture high-resolution spatial and textural characteristics. The deeper layers of the ResNet-101 model capture abstract, global features, while the earlier layers retain detailed spatial information. This encoder analyzes the input RGB frame and produces a set of hierarchical feature maps, which are later utilized for fusion and skip connections.

Event Data (E-TMA): To handle event data, the Event-Based Temporal Multiscale Aggregation (E-TMA) module (Zhou et al., 2023) is used to harness the temporal richness inherent in asynchronous events. The event data from each temporal scale ξ_k (where $k \in \{1,2,3\}$) is transformed into a latent feature space as follows:

$$E_k = \phi(\xi_k) \tag{1}$$

where ϕ represents a shared convolutional projection layer that includes batch normalization and ReLU activation.

To capture hierarchical motion dynamics, pooling operations with varying kernel sizes are applied:

$$e_k = \mathrm{pool}_k(E_k) \tag{2}$$

where the kernel sizes increase with the temporal range. To standardize the resolutions of these features, smaller feature maps undergo upsampling and are then concatenated:

$$F_{\text{event}} = \mathrm{concat}(e_1, \mathrm{up}(e_2), \mathrm{up}(e_3)) \tag{3}$$

This combined event representation encapsulates motion patterns from fine to coarse and is subsequently fed into a lightweight ResNet-18 encoder for feature extraction. The event encoder is designed to be computationally efficient while preserving essential motion information.

3.3 Fusion Network

The fusion network integrates features obtained from both RGB and event encoders, effectively leveraging the unique advantages of each modality. Prior to fusion, the event features are transformed to match the dimensionality of the RGB features:

$$F_{\text{aligned}} = \psi(F_{\text{event}}) \tag{4}$$

where ψ denotes a 1×1 convolution.

Channel Attention: Channel attention mechanisms are utilized to highlight the most pertinent feature channels within each modality. The features, enhanced by attention, are computed as follows:

$$F_{\text{rgb}}^{\text{cal}} = \mathrm{CA}(F_{\text{aligned}}) \odot F_{\text{rgb}} + F_{\text{rgb}} \tag{5}$$

$$F_{\text{event}}^{\text{cal}} = \mathrm{CA}(F_{\text{rgb}}) \odot F_{\text{aligned}} + F_{\text{aligned}} \tag{6}$$

Here, CA represents the channel attention module, and $\odot$ signifies element-wise multiplication.

Spatial Attention: Spatial attention is applied to enhance the feature maps by concentrating on the most significant spatial areas:

$$F_{\text{rgb}}^{\text{sp}} = \mathrm{SA}(F_{\text{event}}^{\text{cal}}) \odot F_{\text{rgb}}^{\text{cal}} + F_{\text{rgb}}^{\text{cal}} \tag{7}$$

$$F_{\text{event}}^{\text{sp}} = \mathrm{SA}(F_{\text{rgb}}^{\text{cal}}) \odot F_{\text{event}}^{\text{cal}} + F_{\text{event}}^{\text{cal}} \tag{8}$$

In this context, SA denotes the spatial attention module.

Fusion : The refined features are then concatenated and processed through a convolutional layer to generate the fused representation:

$$F_{\text{fused}} = \mathrm{Conv3x3}(F_{\text{rgb}}^{\text{sp}} \oplus F_{\text{event}}^{\text{sp}}) \tag{9}$$

where $\oplus$ indicates concatenation.

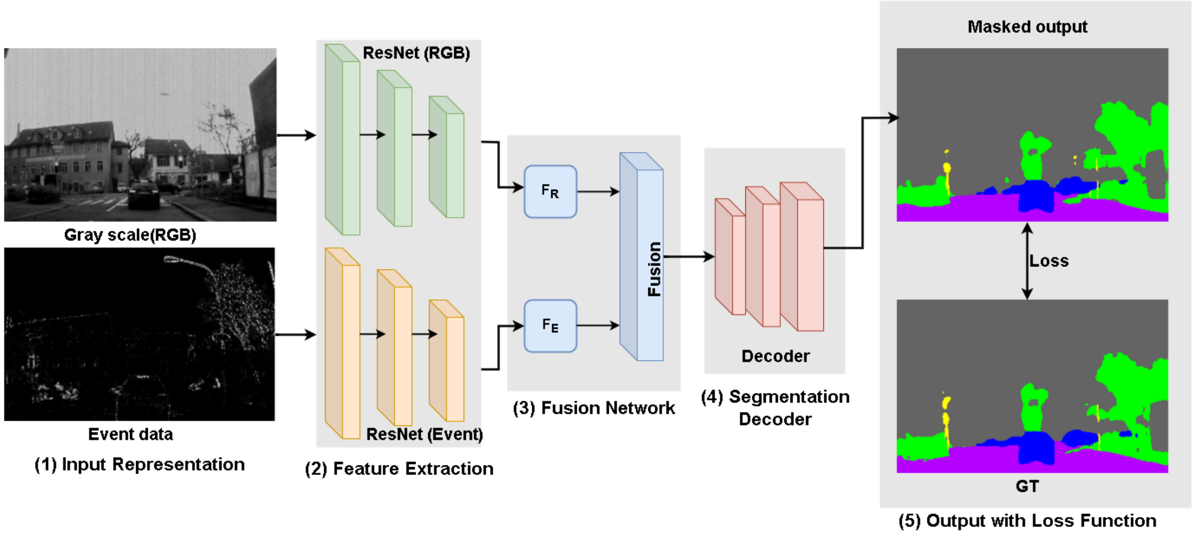

Figure 1: **Architecture for our proposed EveCamNet RGB+Event fusion model** (1) The framework initiates with the Input Representation, which effectively encompasses both grayscale RGB images and event data. (2) The inputs are independently processed through separate Feature Extraction networks, where RGB features are extracted using ResNet-101 and event features are extracted using ResNet-18. (3) The extracted feature maps, F_R (RGB features) and F_E (event features) are passed to a Fusion Network, which combines spatial and temporal information from both modalities. (4) The fused representation is fed into the Segmentation Decoder to generate pixel-wise segmentation masks. (5) Finally, the Output with Loss Function compares the predicted segmentation masks with the ground truth (GT) to compute the loss, optimizing the model for precise segmentation of drivable areas and objects. This pipeline enhances performance in dynamic conditions and by utilizing the complementary strengths of RGB and event data.

3.4 Segmentation Decoder

The combined feature map F_{fused} is transformed into a pixel-level segmentation map using an architecture inspired by UNet (Siddique et al., 2021). This decoder includes multiple upsampling layers along with convolutional operations that work together to gradually restore the original spatial resolution of the input.

Skip connections are used between matching layers in the encoder and decoder to retain important spatial information that might otherwise be lost during downsampling. For example, the output from an earlier encoder layer is added to the output of the corresponding decoder layer, which helps maintain fine details in the spatial data:

$$F_{\text{decoder}}^{(l)} = \text{concat}(F_{\text{encoder}}^{(l)}, \text{up}(F_{\text{decoder}}^{(l+1)})) \qquad (10)$$

where l represents the layer index.

The final segmentation map S is generated using a softmax activation function, which converts the output into class probabilities for each pixel:

$$S = \sigma(\text{Conv3x3}(F_{\text{decoder}}^{\text{final}})) \qquad (11)$$

.

3.5 Loss Function

The model produces a pixel-wise probability map, denoted as S, which is then compared to the actual segmentation map, P, to calculate the loss. During training, we aim to minimize a combined loss function represented as:

$$L = L_{\text{ce}} + \beta \cdot L_{\text{dice}}, \quad \beta = 0.5 \qquad (12)$$

In this equation:

- L_{ce} (cross-entropy loss) focuses on ensuring accurate classification at the pixel level. L_{dice} (Dice loss) helps tackle class imbalance and improves the segmentation of boundaries, calculated by the formula:

$$L_{\text{dice}} = 1 - \frac{2 \cdot \sum P \cdot \hat{P}}{\sum P + \sum \hat{P} + \varepsilon} \qquad (13)$$

Here, $\hat{P}$ is the predicted segmentation map, and ε is a small constant to avoid division by zero.

The cross-entropy loss penalizes wrong class predictions, while the Dice loss emphasizes enhancing the overlap between the predicted areas and the actual ground truth, especially at the edges. By combining these loss functions, the model effectively learns not only to classify accurately but also to delineate boundaries precisely, resulting in strong segmentation performance even in difficult situations.

4 RESULTS AND DISCUSSION

This section presents a concise evaluation of the RGB-Event fusion framework, covering the dataset, training process, and performance analysis of results. We compare our model's performance with established baselines, offering a clear understanding of the pipeline's capabilities.

4.1 Dataset

The DDD17 dataset (Binas et al., 2017) plays a crucial role in examining how RGB and event data can work together for segmenting drivable areas. It captures urban driving situations using an event-based camera paired with standard RGB frames, allowing for asynchronous brightness changes (events) that are aligned with detailed RGB images. This dataset is excellent for providing high temporal resolution and a wide dynamic range, which are essential for navigating complex driving environments.

For our project, we focused on a subset of sequences that met our criteria for urban environments with good visibility, using 15,950 frames for training and 3,890 frames for testing across the dataset's temporal range.

In the training phase, we integrated events over a 50ms period to capture the necessary temporal details. For testing, we evaluated the model at 10ms and 250ms intervals to check its performance across different speeds. This approach ensures that the model can adapt to both fast-paced and slower situations encountered in autonomous driving.

4.2 Training

The training process focused on fine-tuning a fusion-based segmentation model that integrates both RGB and event data. The RGB frames were resized to 288x288 pixels, while the event data was consolidated into multi-channel formats over various temporal intervals: short (10ms), medium (50ms), and long (250ms). This setup was intended to capture different motion characteristics, ranging from intricate details to broader, slower movements. The optimization was carried out using the Adam optimizer, starting with a learning rate of 5×10^{-4}. and progressively decaying it using a polynomial schedule. A batch size of 8 was maintained over 30 epochs, and we included data augmentations like random rotations, flips, and shifts to enhance generalization. The loss function combined cross-entropy loss for pixel-level accuracy with Dice loss to tackle class imbalance:

$$L = L_{\text{ce}} + \beta \cdot L_{\text{dice}}, \quad \beta = 0.5 \tag{14}$$

where the Dice loss L_{dice} is computed as:

$$L_{\text{dice}} = 1 - \frac{2\sum P \cdot \hat{P}}{\sum P + \sum \hat{P} + \varepsilon} \tag{15}$$

Here, P is the mask of ground truth and $\hat{P}$ is the mask predicted.

4.3 Experimental Results

The experimental results are divided into qualitative and quantitative analyses, providing a comprehensive evaluation of our RGB-event fusion model's performance across visual and metric-based perspectives.

4.3.1 Performance Analysis

Here, we analyzied our model through qualitative and quantitative comparisons with existing baseline models, deriving key observations and results that highlight its effectiveness.

Qualitative Results : In our qualitative analysis, we looked at how well our model segments by comparing its output to some baseline methods. These included RGB-only segmentation, event-only segmentation, and a combination of both RGB and event data. The models we compared were UNet for RGB U-Net (Siddique et al., 2021), EV-SegNet for events (Alonso and Murillo, 2019), and our own model that fuses RGB and event information.

The results are presented visually in Figure 2, where we arranged each input scenario in rows and the different models in columns. This setup allows for an easy comparison of the segmentation masks across methods, highlighting the differences in object boundaries and how well each method performs under tricky conditions. For instance, in high-motion scenarios, the event data significantly boosts segmentation performance by keeping the boundaries sharp, something the RGB-only models often struggle with. Our model successfully integrates the advantages of both RGB and event data, resulting in precise segmentation of drivable areas even in dynamic situations.

Quantitative Results : The performance of each model was quantitatively evaluated using mean Intersection over Union (IoU) and pixel accuracy metrics.

- mean Intersection over Union (mIoU)

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c} \tag{16}$$

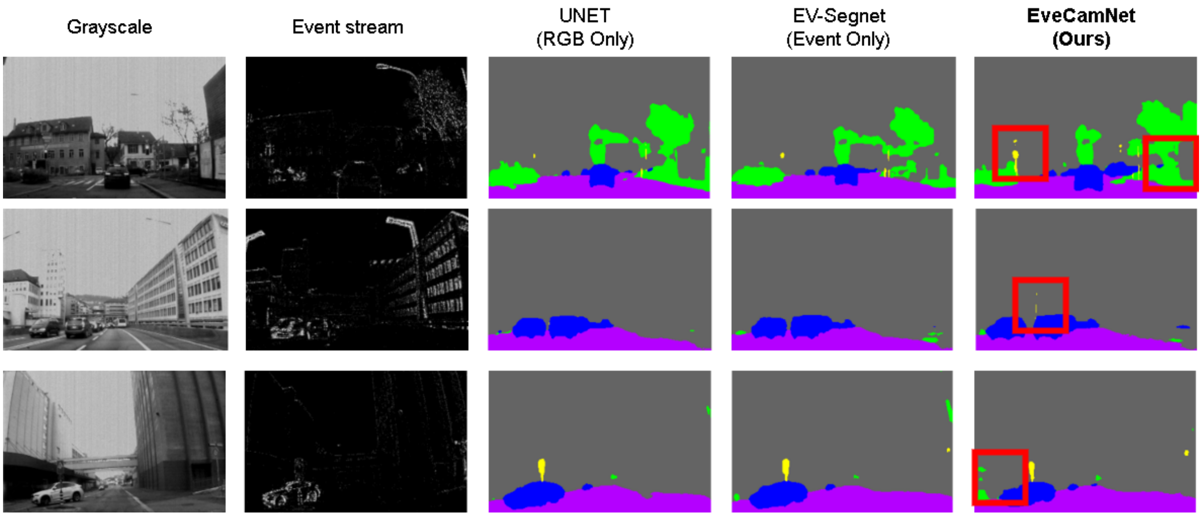

Figure 2: **Qualitative comparison of segmentation results across different input modalities and models.** The columns show grayscale RGB images, event streams, and outputs from three models: UNet (RGB-only), EV-SegNet (event-only), and our proposed EveCamNet (RGB+Event Fusion) model. The rows depict urban road scenes with vehicles and drivable areas. UNet struggles with object boundaries, while EV-SegNet captures motion details but lacks spatial context. Our RGB-Event Fusion model effectively integrates spatial and temporal information, resulting in accurate segmentation masks with clear object boundaries, even in dynamic scenarios. This analysis demonstrates our model's effectiveness in identifying drivable areas and objects. The red box highlights that EveCamNet achieves segmentation more efficiently than the baseline models.

- Pixel Accuracy (PA)

$$\text{PA} = \frac{\text{Number of correctly classified pixels}}{\text{Total number of pixels}} \tag{17}$$

The results in Table 1 clearly demonstrate that our RGB-event fusion model outperforms others in all evaluated metrics. The mean Intersection over Union (mIoU) for drivable areas shows a significant increase, indicating the model's strength in accurately identifying navigable regions. Moreover, the improved mIoU for vehicle detection highlights better boundary delineation, which can be attributed to the complementary capabilities of both RGB and event data. There are also notable gains in pixel accuracy (PA) for our proposed model.

By incorporating results from various temporal intervals (including 10ms and 250ms), we showcase the model's ability to adapt to both high-speed and slower scenarios. The performance at 50ms, which was used during training, acts as a reference point. The enhancements observed at 10ms illustrate its sensitivity to fine temporal details, while the 250ms performance indicates robustness to aggregated motion patterns. Overall, these findings underscore the model's versatility across different temporal dynamics, making it a strong candidate for a range of operational conditions.

5 ABLATION STUDY

This section explores the contribution of the individual component in our framework via the ablation study, allowing us to measure the impact on segmentation tasks and model efficiency.

Loss Importance Analysis : To assess the importance of the proposed loss functions in our RGB+Event fusion framework, we conducted ablation experiments by selectively omitting certain components of the loss function during training. This method enabled us to pinpoint the role of each component and evaluate how it affected segmentation performance.

The results presented in Table 2 illustrate how various loss functions impact critical metrics such as mean Intersection over Union (mIoU), accuracy, and final training loss. Notably, the removal of the Dice loss resulted in a significant drop in mIoU, highlighting its crucial role in addressing class imbalance and enhancing segmentation boundaries. Additionally, eliminating channel or spatial attention mechanisms led to performance declines, emphasizing their importance in effective feature fusion and alignment between modalities. The complete model consistently achieved the highest scores, highlighting the importance of the proposed loss functions for effective segmentation.

Table 1: **Performance comparison of different segmentation models (UNET, EV-SegNet, and our RGB+Event fusion approach) across various temporal intervals.** The proposed RGB+Event fusion model consistently outperforms others in both mean Intersection over Union (mIoU) and accuracy, highlighting its effectiveness in leveraging multi-modal data for improved segmentation performance.

Model	**mIoU (50ms)**	**Accuracy (50ms)**	**mIoU (10ms)**	**Accuracy (10ms)**	**mIoU (250ms)**	**Accuracy (250ms)**
UNET (Siddique et al., 2021)	65.77	87.92	65.34	84.5	65.47	83.89
EV-Segnet (Alonso and Murillo, 2019)	67.15	88.12	66.87	87.91	66.98	87.34
EveCamNet (Ours)	**69.94**	**90.54**	**69.83**	**89.97**	**68.73**	**90.12**

Table 2: **Adapted table for loss-based ablation**. The table compares the performance of the model under different configurations. Metrics include Accuracy (%), mIoU (%), and Final Loss.

Method	**Accuracy [%]↑**	**mIoU [%]↑**	**Final Loss↓**
w/o L_{dice}	88.76	60.12	0.25
w/o Channel Attention	89.65	62.40	0.23
w/o Spatial Attention	89.78	63.05	0.22
w/o Both Attention Mechanisms	89.40	61.50	0.24
EveCamNet (Ours)	93.50	71.50	0.18

6 CONCLUSION

In conclusion, the proposed RGB+Event fusion framework effectively combines the unique advantages of RGB and event data for robust segmentation. By incorporating both channel and spatial attention mechanisms, the model enhances feature fusion, allowing it to capitalize on the dense spatial details offered by RGB frames while also capturing dynamic motion cues from event data. This method has shown impressive segmentation performance across various temporal scales, particularly in the realm of drivable area segmentation for autonomous driving applications.

To tackle challenges like class imbalance and precise boundary segmentation, which are crucial for distinguishing between drivable and non-drivable areas, the framework employs specialized loss functions such as Dice loss. The results demonstrate the model's capability to adapt to a range of scenarios, including varying lighting conditions and high-motion environments, highlighting its practical applicability for real-world situations.

Looking ahead, future research could aim to expand this framework to tackle low-light and extreme lighting environments, where event cameras excel. Additionally, enhancing the framework for greater computational efficiency may facilitate real-time segmentation, opening up possibilities for its deployment in edge devices and autonomous systems with limited resources.

REFERENCES

Almeida, T., Lourenço, B., and Santos, V. (2020). Road detection based on simultaneous deep learning approaches. *Robotics and Autonomous Systems*, 133:103605.

Alonso, I. and Murillo, A. C. (2019). Ev-segnet: Semantic segmentation for event-based cameras. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1624–1633.

Binas, J., Neil, D., Liu, S.-C., and Delbrück, T. (2017). Ddd17: End-to-end davis driving dataset. *ArXiv*, abs/1711.01458.

Chakravarthi, B., Verma, A. A., Daniilidis, K., Fermuller, C., and Yang, Y. (2024). Recent event camera innovations: A survey. *arXiv preprint arXiv:2408.13627.*

Chan, Y.-C., Lin, Y.-C., and Chen, P.-C. (2019). Lane mark and drivable area detection using a novel instance segmentation scheme. In *2019 IEEE/SICE International Symposium on System Integration (SII)*, pages 502–506.

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, Los Alamitos, CA, USA. IEEE Computer Society.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Demir, S., Ertop, T. E., Koku, A. B., and Konukseven, E. İ. (2017). An adaptive approach for road boundary detection using 2d lidar sensor. In *2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 206–211.

Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A. J., Conradt, J., Daniilidis, K., and Scaramuzza, D. (2022). Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180.

Gehrig, D., Gehrig, M., Hidalgo-Carri'o, J., and Scaramuzza, D. (2019). Video to events: Recycling video datasets for event cameras. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3583–3592.

Gehrig, M., Aarents, W., Gehrig, D., and Scaramuzza, D. (2021). Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*.

Han, H.-Y., Chen, Y.-C., Hsiao, P.-Y., and Fu, L.-C. (2021). Using channel-wise attention for deep cnn based real-time semantic segmentation with class-aware edge information. *IEEE Transactions on Intelligent Transportation Systems*, 22(2):1041–1051.

Hidalgo-Carrió, J., Gehrig, D., and Scaramuzza, D. (2020). Learning monocular dense depth from events. In *2020 International Conference on 3D Vision (3DV)*, pages 534–542.

Jain, M., Kamat, G., Bachari, R., Belludi, V. A., Hegde, D., and Patil, U. (2023). Affordrive: Detection of drivable area for autonomous vehicles. In Maji, P., Huang, T., Pal, N. R., Chaudhury, S., and De, R. K., editors, *PReMI*, volume 14301 of *Lecture Notes in Computer Science*, pages 532–539. Springer.

Li, K., Xiong, H., Yu, D., Liu, J., Guo, Y., and Wang, J. (2022). An end-to-end multi-task learning model for drivable road detection via edge refinement and geometric deformation. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):8641–8651.

Liu, M., Fu, B., Xie, S., He, H., Lan, F., Li, Y., Lou, P., and Fan, D. (2021). Comparison of multi-source satellite images for classifying marsh vegetation using deeplabv3 plus deep learning algorithm. *Ecological Indicators*, 125:107562.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation . In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, Los Alamitos, CA, USA. IEEE Computer Society.

Qiao, D. and Zulkernine, F. (2021). Drivable area detection using deep learning models for autonomous driving. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5233–5238.

Rasib, M., Butt, M. A., Riaz, F., Sulaiman, A., and Akram, M. (2021). Pixel level segmentation based drivable road region detection and steering angle estimation method for autonomous driving on unstructured roads. *IEEE Access*, 9:167855–167867.

Shariff, W., Dilmaghani, M. S., Kielty, P., Moustafa, M., Lemley, J., and Corcoran, P. (2024). Event cameras in automotive sensing: A review. *IEEE Access*, 12:51275–51306.

Siddique, N., Paheding, S., Elkin, C. P., and Devabhaktuni, V. (2021). U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9:82031–82057.

Tomy, A., Paigwar, A., Mann, K. S., Renzaglia, A., and Laugier, C. (2022). Fusing event-based and rgb camera for robust object detection in adverse conditions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 933–939.

Wang, L., Chae, Y., Yoon, S.-H., Kim, T.-K., and Yoon, K.-J. (2021). Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 608–619.

Yang, F., Wang, H., and Jin, Z. (2020). A fusion network for road detection via spatial propagation and spatial transformation. *Pattern Recognition*, 100:107141.

Zhou, Z., Wu, Z., Boutteau, R., Yang, F., Demonceaux, C., and Ginhac, D. (2023). Rgb-event fusion for moving object detection in autonomous driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7808–7815.

Zou, Z., Chen, K., Shi, Z., Guo, Y., and Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276.