Multi-Modal Fusion of Event and RGB for Monocular Depth Estimation Using a Unified Transformer-based Architecture

Anusha Devulapally
The Pennsylvania State University
akd5994@psu.edu

Md Fahim Faysal Khan
The Pennsylvania State University
mzk591@psu.edu

Siddharth Advani
Samsung Electronics America
siddharth.advani@ieee.org

Vijaykrishnan Narayanan
The Pennsylvania State University
vijaykrishnan.narayanan@psu.edu

Abstract

In the field of robotics and autonomous navigation, accurate pixel-level depth estimation has gained significant importance. Event cameras or dynamic vision sensors, capture asynchronous changes in brightness at the pixel level, offering benefits such as high temporal resolution, no motion blur, and a wide dynamic range. However, unlike traditional cameras that measure absolute intensity, event cameras lack the ability to provide scene context. Efficiently combining the advantages of both asynchronous events and synchronous RGB images to enhance depth estimation remains a challenge. In our study, we introduce a unified transformer that combines both event and RGB modalities to achieve precise depth prediction. In contrast to individual transformers for input modalities, a unified transformer model captures inter-modal dependencies and uses self-attention to enhance event-RGB contextual interactions. This approach exceeds the performance of recurrent neural network (RNN) methods used in state-of-the-art models. To encode the temporal information from events, convLSTMs are used before the transformer to improve depth estimation. Our proposed architecture outperforms the existing approaches in terms of absolute mean depth error, achieving state-of-the-art results in most cases. Additionally, the performance is also seen in other metrics like RMSE, absolute relative difference and depth thresholds compared to the existing approaches. The source code is available at:https://github.com/anusha-devulapally/ER-F2D.

1. Introduction

Scene depth estimation plays a significant role in computer vision, improving perception and understanding of three-dimensional environments. It has wide-ranging applications, including robotic navigation, autonomous driving, and virtual reality experiences [14, 21, 23, 28]. Accurately estimating scene depth enables these technologies to operate effectively in complex and dynamic settings, enhancing their spatial awareness and interaction capabilities. However, the limitations of conventional cameras, such as low dynamic range and sensitivity to motion blur, can adversely impact the quality of depth maps generated from their images. To address these challenges, the use of event-based cameras [2], which capture pixel-level temporal changes, has emerged as a promising solution. Event-based vision provides a higher dynamic range and can robustly estimate depth for complex scenarios irrespective of the changes in motion or lighting conditions [26, 30]. However, the event camera primarily detects scene edges, resulting in sparse and asynchronous event-based data.

While RGB data effectively retains spatial contextual information, event cameras excel at capturing salient edges. Hence, combining both data modalities is an approach to enhance the overall accuracy of depth estimation. The fusion of event-based and RGB data for depth estimation presents several challenges. RGB imaging operates synchronously with fixed frame capture rates, whereas event-based sensing operates asynchronously in response to brightness changes, leading to dissimilarities in data throughput and temporal representation. Therefore, we require a model that can effectively incorporate both of these distinct asynchronous modalities, each complementing the other, to enhance the precision of depth estimation.

Our research is motivated by the success of transformer architectures in various tasks, such as natural language processing [4, 24] and computer vision [1, 15]. Transformers are known for their ability to capture complex relationships, context, and sequences through self-attention mechanisms. Their suitability is evident due to their capacity for parallel processing, which contrasts the sequential nature of recurrent neural networks (RNNs). They are also widely used in multi-modal fusion [10, 16, 27] but so far, most of the works use individual encoders for each modality which is computationally expensive. We want to enquire whether a single transformer-encoder for multiple modalities benefit us. This prompted us to explore transformers in the context of multi-modal depth estimation, specifically in harmonizing the unique characteristics of event cameras and RGB images.

We present a novel unified transformer integrating both the event and RGB modalities. Additionally, we have incorporated convLSTM blocks [20] to encode the temporal information of events for precise depth prediction, outperforming the performance of traditional RNN methods. This unified approach captures cross-modal dependencies via self-attention.

We have tested our model on both real dataset, multi vehicle stereo event camera (MVSEC) [31] and synthetic dataset, EventScape [3]. Our model outperforms the existing fusion techniques in both accuracy and performance metrics. Our architecture achieves state-of-the-art results in absolute mean depth error, Root Mean Square Error, Absolute Relative Difference and depth thresholds demonstrating better performance over existing methods.

To summarize, we made the following contributions:

- We propose a novel unified transformer architecture as an encoder for dense depth estimation fusing events and RGB frames.
- We incorporate convLSTM blocks as pre-processing to leverage temporal information from events before feeding to the transformer.
- We apply our approach to both real and synthetic datasets, where we outperform the state-of-the-art-fusion method in absolute mean-depth error, Root Mean Square Error, Absolute Relative Difference and depth thresholds.

2. Literature Review

2.1. Event Based Depth Estimation

Monocular depth estimation from events has been of interest for quite some time due to the characteristics of event-based cameras. Earlier papers have explored the depth estimation from stereo images by leveraging the left-right consistency [5] or maximizing the temporal consistency between event streams [29, 30]. However, these approaches only give a semi-dense depth. Later, [17, 22]leveraged the stereo event streams to estimate dense depth. Hidalgo-Carrio et al. [7] is the first work to estimate depth from a single camera which uses a simple U-Net architecture to estimate the dense depth maps from events by using convolutions followed by convLSTM blocks at each level in the encoder. Later, the U-Net architecture is enhanced using transformers as a bottleneck layer for the generator model in a GAN setting [12] or transformer blocks at each level of the encoder-decoder, including the skip connection [13]. Even though using transformers enhances the accuracy, the architectures built are computationally expensive. Even though events give a dense map they do not utilize the rich spatial information from RGB images. Building upon the achievements demonstrated by transformers in this domain, we aim to leverage further their capabilities in the fusion of event and RGB data considering the resource constraints and emphasizing the applicability of our findings within such environments.

2.2. Event-RGB Fusion Based Depth Estimation

Recent advancements have focused on incorporating multiple data modalities to enhance the performance of specific tasks, see Tab. 1. Gehrig et al. [3] proposed the first work on combining events and RGB for monocular depth estimation using a recurrent asynchronous encoder-decoder network. It follows a U-Net architecture with an encoder each for events and RGB. Each level of encoder consists of convolutions followed by convLSTMs and these encoder outputs are combined with convGRU at each level, followed by a decoder with enabled skip connections. EVT+ [19] proposed a patch-based event representation and a backbone to process input modalities for classification and depth estimation tasks. It uses attention blocks for event encoder and is fused in the later stages with images. Most recent work, HMNet [6] proposed a generic low-latency multi-level memory hierarchy to process events, and the final level of the hierarchy is fused with RGB. This model encodes both modalities separately and fuses at the later stages, whereas, we perform both early and late fusion. **In this work, we leverage multi-modal fusion that uses a single vision transformer-based encoder to input two distinct modalities, events and RGB, to learn the dependencies among them. To leverage the temporal information of the events, we introduce convLSTM blocks before the patch embedding.** Unlike the conventional methods [18] which uses multiple transformer blocks in the encoder phase, we use single transformer and perform the patch embedding on the feature space generated from convLSTM outputs rather than the raw events. This novel approach demonstrated a significant improvement both in the accuracy and performance of the model.

3. Methodology

3.1. Input Representations

Event cameras capture the change in brightness and generate events asynchronously. Each event is of the shape (x, y, t, p). Where x, y are the event coordinates in the x and y direction, t is the time stamp and p is the polarity which ranges from $\{-1, 1\}$ depending on the direction of

Features	E2Depth [7]	RAMNet [3]	HMNet [6]	EVT+ [19]	Transformer-based (Ours)
Network Architecture	U-Net	U-Net	Hierarchial Memory Stack	Transformer-based	Vision Transformer-based
Input	Events (Voxels)	Events (Voxels) + frames	Events (Memory Cells) + frames	Events (patches) + frames	Events (Voxels) + frames
Output	Dense Depth Map	Dense Depth Map	Dense Depth Map	Dense Depth Map	Dense Depth Map
Fusion Type	No Fusion	Middle Fusion	Late Fusion	Late Fusion	Early + Late Fusion
No. of Encoders	1 (only events)	2	2	2	1
Contribution	Learns from only events	Combines events and frames to improve accuracy	Efficient event processing and low latency	Patch-based event representation and robust data preprocessing	A single transformer encoder to combine the inputs and to enhance the accuracy further

Table 1. In Comparison with Existing State-of-the-art Models

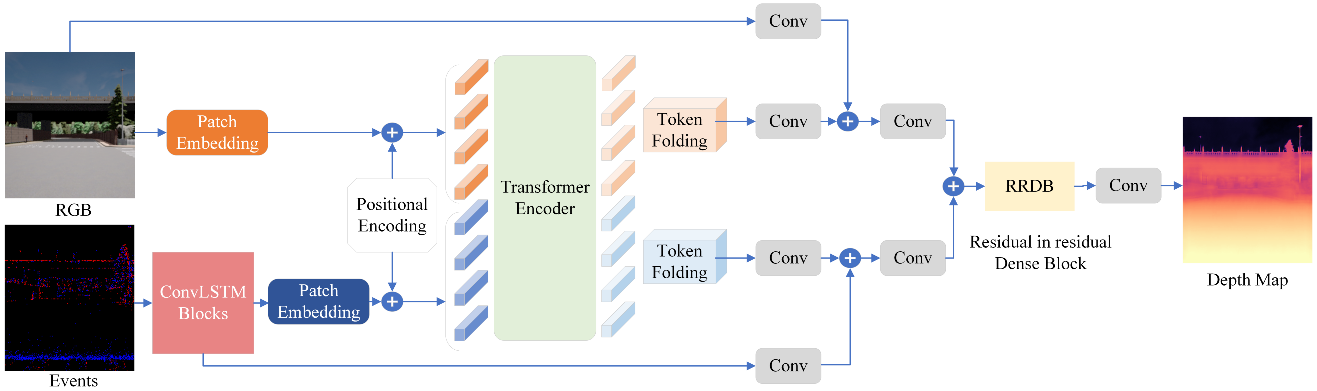

Figure 1. Transformer-based architecture for Monocular Dense Depth Estimation by fusing Events and RGB Images. A single transformer encoder block is used for both the inputs and the folded outputs are given to fusion module with skip connections and RRDB [25] to estimate depth map.

brightness change. In order to use ANNs, the continuous asynchronous stream of events requires to have a fixed tensor. So, they are divided into five temporal bins [7] of the same size. The ground truth depth maps are converted to normalized log depth maps to better distinguish small depth variations for the close ranges.

3.2. Network Architecture

The network architecture presented in the Fig. 1 incorporates convLSTM blocks [20] to effectively capture temporal information from events. ConvLSTMs [20] are recurrent neural network modules that learn spatial and temporal dependencies in sequential data. They take advantage of convolution and LSTM networks. Convolutions in the convLSTMs capture the spatial information in the data and LSTM cells capture the long-term dependencies by retaining the important information across time.

By employing this approach, events are processed through convLSTMs and the output, spatio-temporal feature maps along with RGB are fed to the patch embedding. We divide the input into non-overlapping patches of size 16 and positional embedding is applied. These RGB and event tokens are processed through a transformer encoder. Vision transformer (ViT) [1] architecture is the base for the encoder. It captures both local and global dependencies within the input using a self-attention mechanism.

The transformer encoder generates a set of tokens, which are then reshaped back to their original dimensions using token folding. These folded tokens are passed through a multi-modal fusion block, which employs convolution operations and skip connections. In this block, the reshaped images from token folding are fused with the initial input via skip connections and are further subjected to convolutional operations to facilitate the fusion of RGB and event modalities. This fusion block is crucial in information restoration, recovering any potentially lost information in the images.

The fused output is then directed into a residual in residual dense block (RRDB) [25], followed by a single channel convolutional layer. RRDB [25] architecture introduces residual connections within and between multiple dense blocks, increasing the network's depth and complexity, enhancing its performance. These layers capture the input's complex features, which aid in reconstructing depth in finer details, thus enhancing the overall depth maps. The network enables a comprehensive and accurate reconstruction of depth information by integrating RGB and event data and effectively fusing them through the proposed architecture.

3.3. Loss Functions

Our transformer-based model is trained in a supervised fashion using ground truth depth maps. We use a combination of L1 loss, normal loss [8] and multi-scale scale-invariant gradient matching loss [11] to compute the valid ground truth labels and the prediction outputs. For a se-

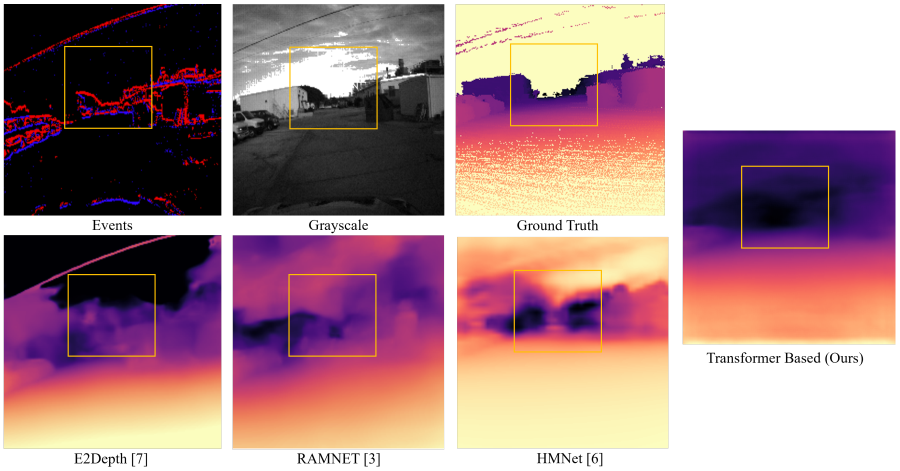

Figure 2. Qualitative comparison of the models on a test sample from MVSEC Dataset. Depth is depicted by the color gradient. Brighter the color, lower the depth. The yellow box in the images highlights the advantage of our proposed method over the baselines, E2Depth, RAMNet and HMNet. Our model can estimate the depth of the image to the farthest point similar to that in groundtruth.

Dataset	Distance	E2Depth [7] (↓)	RAMNet [3] (↓)	HMNet [6] (↓)	Transformer-based (Ours) (↓)
Outdoor Night1	10m	3.38	2.50	**1.50**	1.58
	20m	3.82	3.19	2.48	**2.24**
	30m	4.46	3.82	3.19	**2.78**
Outdoor Night2	10m	1.67	**1.21**	1.36	1.54
	20m	2.63	2.31	2.25	**2.23**
	30m	3.58	3.28	2.96	**2.95**
Outdoor Night3	10m	1.42	**1.01**	1.27	1.24
	20m	2.33	2.34	2.17	**1.96**
	30m	3.18	3.43	2.86	**2.81**
Outdoor Day1	10m	1.67	1.39	**1.22**	1.34
	20m	2.64	**2.17**	2.21	2.25
	30m	3.13	2.76	2.68	**2.62**

Table 2. Absolute Mean Depth Error Results on Four Sequences of the MVSEC Dataset (in meters)

quence of inputs, ground truth labels and prediction outputs are denoted by D_k and $\hat{D}_k$ respectively and the difference $R_k = D_k - \hat{D}_k$. The L1 loss is defined as:

$$L_{\text{l1\_loss}} = \sum_{i=1}^{N} |R_k(i)| \tag{1}$$

where N is the number of valid ground truth pixels. The normal loss [8] is defined as:

$$L_{\text{normal}} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{\left\langle \nabla D_k(i), \nabla \hat{D_k}(i) \right\rangle}{\|\nabla D_k(i)\| \left\| \nabla \hat{D_k}(i) \right\|} \right) \tag{2}$$

Where ∇D_k, $\nabla \hat{D}_k$ are the corresponding gradient vectors of ground truth D_k and the prediction $\hat{D}_k$ respectively. The $\langle \cdot \rangle$ and $\| \cdot \|$ denote the corresponding vector's dot product and the norm respectively. It measures the alignment between the predicted and ground truth gradients and considers the relative similarity between gradients rather than their magnitudes, which is beneficial in the accurate orientation estimation. The multi-scale scale-invariant gradient matching loss [11] is defined as:

$$L_{\text{grad}} = \frac{1}{N} \sum_{s} \sum_{i} |\nabla_x R_k^s(i)| + |\nabla_y R_k^s(i)| \tag{3}$$

∇_x and ∇_y compute the edges in the x and y direction using the sobel operator and are calculated over four different scales (s). It tries to match the gradients in the ground truth depth and favours smooth gradient changes. The total loss is

$$L_{\text{total}} = \lambda \cdot L_{\text{l1\_loss}} + L_{\text{normal}} + \beta \cdot L_{\text{grad}} \tag{4}$$

The hyper-parameters used for λ and β are 0.5 and 0.25. We discuss the impact of the loss function in the ablation studies Sec. 6.3.

4. Experimental Setup

In this section, we provide an overview of the two datasets utilized, namely, MVSEC [31] and EventScape [3]. We also describe the evaluation metrics employed, and outline the implementation details.

MVSEC Dataset: Multi-Vehicle Stereo Event Camera dataset [31] (MVSEC) is a real-world dataset that includes driving sequences collected throughout the day and at night. We used Outdoor day2 for training and validation with 10,000 samples and tested on Outdoor night1, night2, night3 and day1, each containing 5000 samples. The MVSEC dataset consists of grayscale images, events and their respective ground truths calculated using LiDAR.

EventScape Dataset: EventScape [3] is a synthetic dataset generated from the CARLA event simulator. It consists of 743 sequences of driving data at different locations named as 'Town'. For training, we use sequences from Town 01, 02 and 03, validation and testing is performed on Town 05. The dataset consists of RGB images, events, segmentation labels, groundtruths and various vehicle controls.

4.1. Evaluation Metrics

We use absolute mean depth error [11] as the evaluation metrics. The error (Eq. (5)) is calculated between the groundtruth labels (D_k) and prediction output ($\hat{D_k}$) at three different depths, that is, 10m, 20m and 30m considering only the 'N' number of valid pixels at a particular depth measurements.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|D_k(i) - \hat{D_k}(i)| \tag{5}$$

We also evaluate our work on Absolute Relative Difference, RMSE and delta thresholds ($\delta < 1.25^t$, where t = 1, 2, 3) for the same distance cutoffs. For RMSE and Abs Rel, lower the value, better the performance and for the depth thresholds, higher the value, better the performance.

Additionally, we compute latency (in seconds), throughput (frames per second) and runtime memory usage (in MB) of our model for comparison with existing work.

4.2. Implementation Details

Our model is implemented in the PyTorch framework. We used pre-trained weights from vit-base [1] for transformer-based encoder block. During training, we normalize inputs with valid pixels to have mean 0 and variance 1. Furthermore, we applied a random crop of size 224 x 224 and random horizontal flip transformations. We used ADAM optimizer [9] with a batch size of 16, a learning rate of 0.0003, and trained it for 70 epochs.

5. Results

In this section, we present and discuss the results on the two datasets, MVSEC [31] and EventScape dataset [3]. We also evaluated performance metrics such as absolute mean depth error, RMSE, Absolute Difference, delta thresholds, latency, throughput and runtime memory usage against other state-of-the-art models, E2Depth [7], RAMNet [3] and HMNet [6].

5.1. Results on MVSEC

Table 2 shows a quantitative comparison of our model with the state-of-the-art models. The numbers depict the absolute mean depth error calculated on four datasets, Outdoor night1, night2, night3 and day1. Our model's performance surpasses that of the baselines E2Depth, RAMNet and HMNet across four datasets, showcasing an significant decrease in mean depth error respectively.

Table 3 shows quantitative comparison of our model with current state-of-the-art model, HMNet [6] on 5 different metrics, absolute relative difference (Abs Rel), RMSE and three depth thresholds. All these 5 metrics for 10m, 20m and 30m distance cutoffs for both outdoor night1 and day1 sequences are tabulated. Due to space constraint, the results on night2 and night3 driving scenarios are tabulated in the supplementary material. From these tables, we observe that our transformer model has overall better performance compared to the existing state-of-the-art model. Figure 2 illustrates a visual comparison of our transformer-based model with the baselines, E2Depth, RAMNet and HMNet on a sample from the MVSEC dataset. We observe that our transformer-based model can predict the depth and the objects on the ground better compared to others. We observe that the sky in our depth estimation as well as other existing works is smeared. This is because during the training these regions are masked and do not have valid pixels because we do not have depth for the sky from lidar ground truth. We believe the difference could also be because of the difference in training and test samples where test samples have overcast sky as opposed to the clear sky in train samples. E2Depth only considers events and hence do not have artifacts in sky. However, as observed in Tab. 3 adding additional RGB information improves the overall depth estimation to only events as in E2Depth.

5.2. Results on EventScape

Table 4 shows a quantitative comparison of the model with the baselines on EventScape. The number depicts the absolute mean depth error calculated on the test dataset. We achieved a better accuracy in all cases, that is, 10m, 20m and 30m with an overall improvement of $\sim$ 66%,$\sim$ 21% and $\sim$ 8% over E2Depth, RAMNet and HMNet respectively.

Figure 3 shows a visual comparison of our transformer-based model with the three baselines, E2Depth, RAMNet

Metrics	Outdoor Night1						Outdoor Day1					
	HMNet [6]			Our transformer			HMNet [6]			Our transformer		
	10m	20m	30m	10m	20m	30m	10m	20m	30m	10m	20m	30m
Abs Rel (↓)	**0.21**	0.25	0.26	0.22	**0.24**	**0.25**	**0.28**	0.31	0.32	0.30	**0.30**	**0.31**
RMSE (↓)	2.61	4.44	5.14	**2.42**	**3.90**	**4.41**	2.78	4.11	5.15	**2.73**	**3.49**	**4.30**
$\delta < 1.25^1$ (↑)	0.82	0.75	0.72	**0.83**	**0.76**	**0.73**	**0.71**	0.61	0.56	0.70	**0.63**	**0.61**
$\delta < 1.25^2$ (↑)	0.91	0.86	0.86	**0.92**	**0.88**	**0.88**	**0.85**	0.79	0.76	**0.85**	**0.85**	**0.83**
$\delta < 1.25^3$ (↑)	**0.95**	0.93	0.93	0.94	**0.94**	**0.94**	**0.92**	0.89	0.88	**0.92**	**0.93**	**0.93**

Table 3. Different Metric Results on MVSEC Dataset for Outdoor Night1 and Outdoor Day1 Sequences

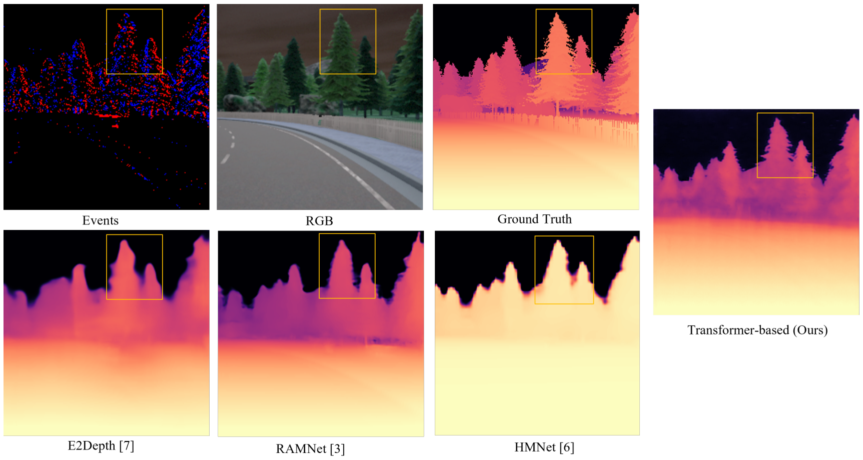

Figure 3. Qualitative comparison of the models on a test image from EventScape Dataset. The depth is depicted by color gradient, brighter the color, nearer the depth. We see that the tree structures highlighted in the yellow box are predicted precisely compared to the baselines.

Model	10m (↓)	20m (↓)	30m (↓)
E2Depth [7]	1.79	5.35	8.31
RAMNet [3]	0.81	2.26	3.58
HMNet [6]	**0.55**	1.80	3.27
Transformer-based (Ours)	0.67	**1.69**	**2.81**

Table 4. Absolute Mean Depth Error Results on EventScape Dataset (in meters)

Model	Latency(↓) (Seconds)	Throughput(↑) (FPS)	Runtime (↓) Memory (MB)
E2Depth [7]	0.015	30.74	**272.83**
RAMNet [3]	0.019	33.79	652.35
HMNet [6]	**0.004**	**83.72**	377.32
Transformer-based (ours)	0.015	49.91	**340**

Table 5. Quantitative Results on Performance Metrics. All Experiments are conducted on a single NVIDIA A100 GPU.

and HMNet on a sample from the EventScape dataset. The depth depicted by the color-gradient is correctly predicted and also the objects on ground. The edges of the tree are notably evident in our transformer-based approach in comparison to the other baselines, as depicted in the figure. Employing the single encoder for both modalities helped in learning the dependencies from each, resulting in enhanced depth map quality.

5.3. Performance Metrics

Table 5 provides a comprehensive comparison of latency, throughput, and runtime memory usage between our transformer-based model and the baseline models. The findings underscore the favorable performance of our transformer-based model across different metrics. Notably, although E2Depth showcases better latency and runtime memory, our model demonstrates a significant $\sim 63\%$ improvement in throughput. Moreover, our model outperforms RAMNet displaying an encouraging $\sim 48\%$ enhancement in both throughput and runtime memory usage, along with a $\sim 21\%$ improvement in latency. Although HMNet displays better latency and throughput, our model demonstrates a $\sim 10\%$ improvement in runtime memory usage. We thus propose an efficient model which achieve superior depth estimation accuracy while maintaining comparable or better latency, runtime memory, and throughput. All experiments were conducted using a single NVIDIA A100 80GB PCIe GPU. The performance gain we observe can be attributed to the parallel processing capability of transformers and the elimination of sequential dependencies

Experiments	Outdoor Night 1			Outdoor Day1		
	10m (↓)	20m (↓)	30m(↓)	10m (↓)	20m (↓)	30m (↓)
Transformer-based (best)	**1.58**	**2.24**	**2.78**	**1.34**	**2.25**	**2.62**
Without the convLSTM	3.63	3.83	4.28	3.11	3.23	3.47
Without the skip connections	3.01	3.27	3.90	2.22	2.47	2.78

Table 6. Absolute Mean Depth Error on Outdoor Night1 dataset and Outdoor Day1 for components impact analysis

Experiments	Outdoor Night 1			Outdoor Day 1			Model Size(MB)(↓)	Model Parameters(M)(↓)
	10m (↓)	20m (↓)	30m(↓)	10m (↓)	20m (↓)	30m (↓)		
Transformer-based (ours)	**1.58**	**2.24**	**2.78**	**1.34**	**2.25**	**2.62**	**336.37 MB**	**88 M**
Individual Encoders	2.03	3.23	3.65	1.78	3.11	3.61	660.31 MB	173 M
Cross-attention encoders	8.42	7.02	7.24	5.71	4.99	5.50	606.24 MB	158 M

Table 7. Absolute Mean Depth Error on Outdoor Night1 and Day1 dataset along with the Model Size and Number of Model Parameters for transformer encoder analysis

Row	Experiments	Outdoor Night 1			Outdoor Day 1		
		10m (↓)	20m (↓)	30m(↓)	10m (↓)	20m (↓)	30m (↓)
1	$0.5 \cdot L_{\mathbf{l1\_loss}} + L_{\mathbf{normal}} + 0.25 \cdot L_{\mathbf{grad}}$	1.58	2.24	**2.78**	**1.34**	**2.25**	**2.62**
2	$0.5 \cdot L_{\text{l1\_loss}} + 0.25 \cdot L_{\text{grad}}$	1.63	2.20	2.82	1.47	2.34	2.63
3	$0.5 \cdot L_{\text{l1\_loss}} + L_{\text{normal}}$	**1.56**	2.27	2.92	1.66	2.81	3.19
4	$L_{\text{si\_loss}} + 0.25 \cdot L_{\text{grad}}$	2.13	2.58	3.51	2.23	2.49	3.11
5	$0.5 \cdot L_{\text{si\_loss}} + L_{\text{normal}} + 0.25 \cdot L_{\text{grad}}$	3.72	4.01	4.28	3.36	3.61	3.74
6	$L_{\text{si\_loss}} + 0.5 \cdot L_{\text{normal}} + 0.25 \cdot L_{\text{grad}}$	1.70	**2.11**	3.04	1.72	2.60	2.82

Table 8. Absolute Mean Depth Error on Outdoor Night1 and Outdoor Day 1 dataset for Different Loss functions.

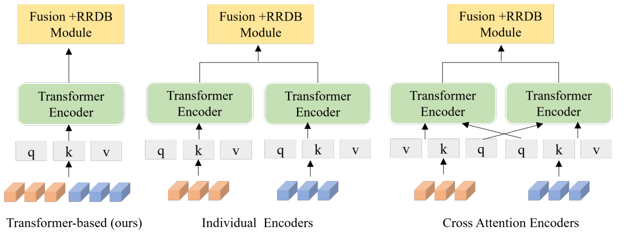

Figure 4. Different encoder combinations for the model architecture. We use these three different ways to fuse two modalities. Here, q is the query, k is the key and v is the values generated from the input tokens.

leads to faster training times, which is especially advantageous in large-scale applications.

6. Ablation Study

In this section, we perform an analysis of three key aspects. First, we explore the impact of different components on the model's performance. Second, we show a comparative study involving different types of transformer-encoders used for two input modalities. Finally, we show a comparative study of different loss functions on the model's performance. To ensure generalizability, we evaluate the performance our proposed model with above variations on two datasets with different lighting conditions, Outdoor Night1 and Outdoor Day1 driving datasets from MVSEC. The other two driving datasets, outdoor night2 and night3 are shown in supplementary.

6.1. Impact of convLSTM and Late Sensor Fusion

In this subsection, we want to highlight the importance of two key aspects of the network architecture, namely the convLSTM blocks to process the events data and the fusion of sensor data at the later stage of the pipeline. To investigate their role, we systematically remove each of these components individually and examine the quantitative results as depicted in Tab. 6.

ConvLSTM Block: ConvLSTM plays a major role in understanding the temporal information of the events. It elevates the information across the temporal domain and gives a better contextual output than a single channel event frame. Removing the convLSTM from the architecture did impact the entire model's performance as shown in Tab. 6.

Late Fusion through Skip Connections: From Fig. 1, we see skip connections each from RGB input and convLSTM followed by a convolution, which is later added to the intermediate output. These skip connections when removed impacts the model performance. They help in preserving the fine-grained details from the original input. The numbers from Tab. 6 also empirically confirm this.

6.2. Different Transformer Encoders for Each Modality

To underscore the advantages of a unified transformer, we conducted a series of experiments that involved distinct transformer encoders dedicated to events and RGB modalities for comparison as shown in Fig. 4 [27]. The first setting employs a single transformer which takes concatenation of tokens from both modalities. This is our proposed model in the Fig. 1. The second setting employs separate transformer encoders for each modality, followed by the fusion of their respective outputs. In the third setting, we introduced cross-attention between the transformer encoders where queries are interchanged between them. Table 7 shows the comparison of absolute mean depth error numbers at different distance cut-off for these three settings along with the insight on the model size and number of parameters. We observe that our proposed model, a unified transformer encoder is better both in terms of model size and also results in lower absolute mean depth error values.

6.3. Using Different Loss Functions

Finding and tuning loss functions is one of the fundamental aspects of training deep learning networks. While previous works [3, 7] have used scale-invariant loss ($L_{\text{si\_loss}}$) and multi-scale scale-invariant gradient loss (L_{grad}) primarily for training, we try with normal loss (L_{normal}) and the basic pixel to pixel distance, L1 loss ($L_{\text{l1\_loss}}$). In this study, we want to compare the performances of these different loss components on the depth prediction task and see which one works better.

Table 8 shows a comparison among different combinations of these loss functions. From row-1 to row-2, we observe that just by removing the normal loss, the errors increased for all three distances, the same is observed in row-6 to row-4. Removing the multi-scale scale-invariant gradient loss from row-1 to row-3 hurts the far field depth prediction performance. Furthermore, we compared the loss function taken from [3, 7] and observe that it doesn't surpass our best reported results. The same is observed when replacing L1 loss with scale invariant loss from row 1 to row 5. At the end, we tuned row 5's loss components set and reported the best results with corresponding loss co-efficients. We observe a similar performance as row-1 but doesn't surpass it. In order to tune for the different loss component co-efficients, we tried different values for a given loss functions and reported their best results with the corresponding loss co-efficients.

7. Conclusion

In this work, we introduce the transformer-based architecture for multi-modal fusion of events and RGB to estimate monocular depths. We also employ convLSTM block to leverage the temporal information obtained from the events. Our method uses the vision transformers to extract global context from both sensors to improve the depth maps. We reported the quantitative and qualitative results on both MVSEC and EventScape datasets and the performance metrics. Finally, we showed that the unified transformer-based model outperforms the existing baselines in most cases with better accuracy and performance metrics. This work demonstrates the significant potential of transformer-based multi-modal fusion for tasks involving event and RGB data. Future research could further explore different modalities and interpretability of the learned attention mechanisms. While our current approach demonstrates the effectiveness of event-based data with RGB for depth estimation and we acknowledge that our event pre-processing could be further refined. Exploring more advanced event-based preprocessing techniques has the potential to further enhance the extraction of temporal information and improve depth map accuracy.

8. Acknowledgement

This work was supported in part by Semiconductor Research Corporation C-BRIC, JUMP 2.0 PRISM, and the NSF SOPHIA (CCF-1822923). We thank Dr Mahmut Kandemir and Dr Jack Sampson for their insightful discussions and constructive suggestions which enhanced quality of the paper.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 3, 5

[2] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022. 1

[3] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrio, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotic and Automation Letters. (RA-L)*, 2021. 2, 3, 4, 5, 6, 8

[4] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting. In *2020 25th international conference on pattern recognition (ICPR)*, pages 10335–10342. IEEE, 2021. 1

[5] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pages 270–279, 2017. 2

[6] Ryuhei Hamaguchi, Yasutaka Furukawaa, Masaki Onishi, and Ken Sakurada. Hierarchical neural memory network for low latency event processing. *CVPR*, 2023. 2, 3, 4, 5, 6

[7] Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. In *2020 International Conference on 3D Vision (3DV)*, pages 534–542. IEEE, 2020. 2, 3, 4, 5, 6, 8

[8] Md Fahim Faysal Khan, Anusha Devulapally, Siddharth Advani, and Vijaykrishnan Narayanan. Robust multimodal depth estimation using transformer based generative adversarial networks. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 3559–3568, New York, NY, USA, 2022. Association for Computing Machinery. 3, 4

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[10] Wei-Yu Lee, Ljubomir Jovanov, and Wilfried Philips. Cross-modality attention and multimodal fusion transformer for pedestrian detection. In *European Conference on Computer Vision*, pages 608–623. Springer, 2022. 2

[11] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 3, 4, 5

[12] Xiuhong Lin, Chenhui Yang, Xuesheng Bian, Weiquan Liu, and Cheng Wang. Eagan: Event-based attention generative adversarial networks for optical flow and depth estimation. *IET Computer Vision*, 16(7):581–595, 2022. 2

[13] Xu Liu, Jianing Li, Xiaopeng Fan, and Yonghong Tian. Event-based monocular dense depth estimation with recurrent transformers. *arXiv preprint arXiv:2212.02791*, 2022. 2

[14] Hongcheng Luo, Yang Gao, Yuhao Wu, Chunyuan Liao, Xin Yang, and Kwang-Ting Cheng. Real-time dense monocular slam with online adapted depth prediction network. *IEEE Transactions on Multimedia*, 21(2):470–483, 2018. 1

[15] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3163–3172, 2021. 1

[16] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7077–7087, 2021. 2

[17] Ulysse Rançon, Javier Cuadrado-Anibarro, Benoit R Cottereau, and Timothée Masquelier. Stereospike: Depth learning with a spiking neural network. *IEEE Access*, 10:127428–127439, 2022. 2

[18] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 2

[19] Alberto Sabater, Luis Montesano, and Ana C Murillo. Event transformer+. a multi-purpose solution for efficient event data processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 3

[20] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 2, 3

[21] Jianyuan Sun, Zidong Wang, Hui Yu, Shu Zhang, Junyu Dong, and Pengxiang Gao. Two-stage deep regression enhanced depth estimation from a single rgb image. *IEEE Transactions on Emerging Topics in Computing*, 10(2):719–727, 2020. 1

[22] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[23] Julien Valentin, Adarsh Kowdle, Jonathan T Barron, Neal Wadhwa, Max Dzitsiuk, Michael Schoenberg, Vivek Verma, Ambrus Csaszar, Eric Turner, Ivan Dryanovski, et al. Depth from motion for smartphone ar. *ACM Transactions on Graphics (ToG)*, 37(6):1–19, 2018. 1

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[25] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 3

[26] Zhen Xie, Shengyong Chen, and Garrick Orchard. Event-based stereo depth estimation using belief propagation. *Frontiers in Neuroscience*, 11, 2017. 1

[27] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 8

[28] Xin Yang, Hongcheng Luo, Yuhao Wu, Yang Gao, Chunyuan Liao, and Kwang-Ting Cheng. Reactive obstacle avoidance of monocular quadrotors with online adapted depth prediction network. *Neurocomputing*, 325:142–158, 2019. 1

[29] Yi Zhou, Guillermo Gallego, Henri Rebecq, Laurent Kneip, Hongdong Li, and Davide Scaramuzza. Semi-dense 3d reconstruction with a stereo event camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[30] Alex Zihao Zhu, Yibo Chen, and Kostas Daniilidis. Realtime time synchronized event-based stereo. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VI*, page 438–452, Berlin, Heidelberg, 2018. Springer-Verlag. 1, 2

[31] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3 (3):2032–2039, 2018. 2, 5