

Combining Events and Frames using Recurrent Asynchronous Multimodal Networks for Monocular Depth Prediction

Daniel Gehrig*, Michelle Rüegg*, Mathias Gehrig, Javier Hidalgo-Carrió, Davide Scaramuzza

Abstract—Event cameras are novel vision sensors that report per-pixel brightness changes as a stream of asynchronous “events”. They offer significant advantages compared to standard cameras due to their high temporal resolution, high dynamic range and lack of motion blur. However, events only measure the varying component of the visual signal, which limits their ability to encode scene context. By contrast, standard cameras measure absolute intensity frames, which capture a much richer representation of the scene. Both sensors are thus complementary. However, due to the asynchronous nature of events, combining them with synchronous images remains challenging, especially for learning-based methods. This is because traditional recurrent neural networks (RNNs) are not designed for asynchronous and irregular data from additional sensors. To address this challenge, we introduce Recurrent Asynchronous Multimodal (RAM) networks, which generalize traditional RNNs to handle asynchronous and irregular data from multiple sensors. Inspired by traditional RNNs, RAM networks maintain a hidden state that is updated asynchronously and can be queried at any time to generate a prediction. We apply this novel architecture to monocular depth estimation with events and frames where we show an improvement over state-of-the-art methods by up to 30% in terms of mean absolute depth error. To enable further research on multimodal learning with events, we release EventScape, a new dataset with events, intensity frames, semantic labels, and depth maps recorded in the CARLA simulator.

Index Terms—Deep Learning for Visual Perception; Sensor Fusion; Data Sets for Robotic Vision

MULTIMEDIA MATERIAL

Visit the project home page for code, dataset and more at http://rpg.ifi.uzh.ch/RAM_Net.html.

I. INTRODUCTION

EVENT cameras are bio-inspired vision sensors, that work radically different from conventional cameras. Instead of capturing the scene at a fixed frame rate, they report per-pixel brightness changes as a stream of asynchronous *events* and do this with low latency and with microsecond resolution. The resulting stream of events, encodes the timestamp, pixel location and polarity (sign) of the brightness changes.

Manuscript received: October 15th, 2020; Revised December 8th, 2020; Accepted February 13th, 2021.

This paper was recommended for publication by Editor Cesar Cadena Lema upon evaluation of the Associate Editor and Reviewers’ comments.

This work was supported by Prophesee, the National Centre of Competence in Research (NCCR) Robotics through the Swiss National Science Foundation (SNSF), and the SNSF-ERC starting grant.

*These authors contributed equally. The authors are with the Robotics and Perception Group, University of Zurich, Switzerland (<http://rpg.ifi.uzh.ch>)

Digital Object Identifier 10.1109/LRA.2021.3060707

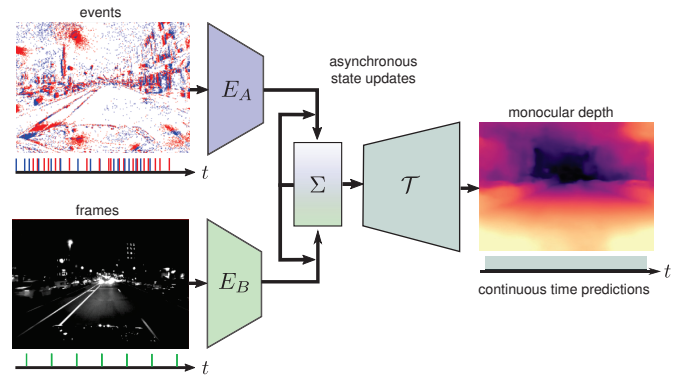


Fig. 1. In this work, we introduce Recurrent Asynchronous Multimodal (RAM) networks for monocular depth estimation. Our method generalizes traditional recurrent neural networks (RNN) for asynchronous and irregular data from multiple sensors. It does this by maintaining an internal state Σ that is updated asynchronously (through E_A or E_B) and can be decoded to a task prediction \mathcal{T} at any time instance. We apply this novel architecture to the task of monocular depth estimation from events and frames.

Event cameras excel at sensing motion and do this at high speed and with low latency. However, compared to standard cameras that measure absolute brightness, they measure *changes in brightness*. Event cameras are, thus, complementary to standard cameras, which motivates the development of novel computer vision algorithms which can combine the advantages of both sensors. For this reason, the Dynamic and Active-pixel Vision Sensor (DAVIS) [1] was introduced. It combines a standard image sensor with an asynchronous event sensor on the same pixel array, making the sensor modalities synchronized and aligned. More details about algorithms and applications of event cameras can be found in [2].

Despite these hardware alignments, designing algorithms for fusing both sensor modalities remains a challenging endeavor. An ideal algorithm must take into account that events and frames are measured asynchronously and at different rates.

While traditionally, filtering techniques, such as the Kalman Filter, have addressed this type of problem, they rely on accurate models designed by experts. By contrast, data-driven sequence models such as LSTM [3] directly learn a model from data and are thus capable of scaling to more complex problems. They are responsible for many of the recent breakthroughs in complex time sequence modeling but require that inputs be presented at fixed rates and synchronized with auxiliary inputs. For these reasons classical RNN-based methods are bound to fail when the input data violates these underlying assumptions [4], [5], as is the case for events and frames.

Methods for combining events and frames have taken two

approaches. Model-based methods fuse events and frames by using a generative model that explains how events are triggered by the underlying brightness signal. However, due to nonlinearities and uncertainties in the event generation model, model-based fusion approaches are brittle and are sensitive to hyper-parameter tuning. For this reason, they have yet to demonstrate state-of-the-art performance on complex pixel-level tasks such as depth prediction [6] or image reconstruction [7], [8]. By contrast, learning-based methods have leveraged large datasets to generate more accurate predictions. However, current learning-based methods for events are limited in that they group events and frames into synchronized stacks which are passed to feed-forward neural networks [9], [10]. Not only does this strategy sacrifice the asynchronicity and high temporal resolution of events, but it also limits the temporal context by using simple feed-forward networks instead of RNNs. Instead, RNNs maintain a memory of previous states which encode temporal context over longer time windows and adapt it based on new inputs. Thus RNNs combine information from different views to guide depth prediction.

In this work we introduce Recurrent Asynchronous Multimodal (RAM) networks that leverage the complementarity of events and frames by (i) explicitly taking into account the asynchronous and irregular nature of events and frames and (ii) integrating temporal contexts from several sensors through the use of recurrency. Inspired by traditional RNNs, RAM networks maintain an internal state that is updated asynchronously by either input modality and decoded to a prediction at any time (Fig. 1). We apply this novel network architecture to the task of event and frame-based monocular depth prediction which is an important building block in modern algorithms such as obstacle avoidance, path planning, and 3D mapping. By using event cameras these algorithms achieve increased robustness thanks to the advantages conferred by the sensor in high-speed and high-dynamic-range scenarios.

Despite this interest, high-quality event and frame-based datasets for learning depth are scarce. To fill this gap we release we EventScape which we use to benchmark several baseline algorithms as well as RAM Net. EventScape is a new synthetic dataset that will enable further research on multimodal learning using events. It is recorded in the CARLA [11] simulator and comprises events, intensity frames, semantic labels, depth maps and vehicle navigation parameters in diverse automotive scenes. Finally, we test our method on the multi-vehicle stereo event camera (MVSEC) dataset recorded with a real stereo DAVIS, where we show an improvement over state-of-the-art event-based methods of up to 30% in terms of mean absolute depth error. This is because depth predictions from RAM Net are more robust to changes in the frequency of input data. In this work, we make the following contributions:

- We propose a new framework that can process multimodal inputs asynchronously and generalizes better to unseen asynchronous data compared to classical RNNs.
- We apply our approach to the task of monocular depth estimation from events and frames, where we outperform state-of-the-art event-based methods by up to 30%.

- We release EventScape, a new multimodal dataset recorded in the CARLA simulator.

II. RELATED WORK

Due to the complementarity of events and frames, various algorithms have sought to leverage the advantages of both modalities by fusing them. This trend has accelerated since the introduction of DAVIS in 2014 [1]. Such algorithms include Simultaneous Localization and Mapping (SLAM) [12], feature tracking [13], [14], [15], high-dynamic-range intensity reconstruction [16] and image deblurring [17]. Although promising, these approaches depend strongly on the fidelity of the underlying models for combining events and frames. Many algorithms today rely heavily on ideal sensor models [18], [2], which do not model noise or other dynamic effects and thus lead to performance degradations in non-ideal conditions. This is why for more complex tasks such as image reconstruction or monocular depth, state-of-the-art methods use a data-driven approach [7], [6], [19], [20], [21]. Of these, many rely on recurrent architectures which can leverage long time windows of events for improved prediction [21], [7]. Although there exist many purely event-based learning methods, few address the fusion of images and events [9], [10], [22]. These approaches fuse both modalities by synchronizing and concatenating both inputs and passing them to a standard feed-forward network [9], [10], [22]. While this strategy improves over passing each input individually, it discards the asynchronous nature and high temporal resolution of the events through stacking and synchronization. Moreover, fusion methods do not involve recurrency, further limiting the leveraged temporal context.

It is well known in the field of multimodal learning¹ that traditional methods like RNNs struggle when sampling rates of different inputs is highly variable, and asynchronous [5], [4], [23]. Typical solutions involve padding, copying, or sampling rate conversions of the input, but at the cost of frequency reductions or temporal misalignments of input data [24]. Alternative approaches use continuous-time models such as neural ODE's [5] or phased LSTMs [4] to fuse such data but were so far only demonstrated on simple tasks.

RAM networks address all of these limitations by (i) leveraging mature data-driven models such as RNNs for robust predictions on complex tasks such as monocular depth estimation, (ii) maximizing the temporal context from events available through the use of recurrency and (iii) introducing individual and asynchronous state update rules for each input modality which retain the asynchronous nature and high temporal resolution of events.

A. Monocular Depth Estimation with Events

Monocular depth estimation is the task of estimating pixel-wise scene depth from a monocular input. Its applications range across different disciplines like medical imaging [25], [26] or autonomous driving applications [27], [28]. Several learning-based approaches based on frames have been

¹In this paper we use multimodality to mean multiple data inputs, and do not refer to multimodal distributions used in statistics.

proposed with promising results [29], [27], [30], [31], [32], [33]. In particular, the use of event cameras for monocular depth estimation has become increasingly popular due to their unique properties, especially for autonomous driving, where low-latency obstacle avoidance and high-speed path planning is of particular importance. There exist different model-based approaches for purely event-based monocular depth estimation which predict sparse [34], [35] or semi-dense depth maps [36], [37]. Recently, various new learning-based methods have been introduced relying on feed-forward [19], [6] and recurrent architectures [21]. Of these methods, all are purely event-based and thus cannot take advantage of complementary frames which can provide important queues in scenarios with little to no ego-motion, where few events are triggered. By contrast, our method, combines frames and events thus leveraging the advantages of both. It does this by generalizing the recurrent architecture in [21], [7] to handle asynchronous and irregular inputs from both sensors.

III. METHOD

A. Recurrent Asynchronous Multimodal Networks

The working principle of RAM Networks is shown in Fig. 2, where we consider the task of monocular depth prediction from events and frames. Let us first consider a scenario where we want to fuse the data from N sensors which provide measurements at steadily increasing time stamps t_j . We may assign a sensor $k_j \in \{1, 2, \dots, N\}$ and measurement $x_{k_j}(t_j)$ from this sensor to each time t_j . We are thus faced with fusing a stream of measurements $\{x_{k_j}(t_j)\}_{j=1}^T$. Since the measurements from different sensors may have vastly different formats, RAM Networks first maps the stream of measurements to a stream of *intermediate features*, s_{k_j} through the use of sensor specific learnable encoders \mathbf{E}_{k_j}

$$s_{k_j} = \mathbf{E}_{k_j}(x_{k_j}, \theta_{k_j}). \quad (1)$$

Where θ denote the network parameters of RAM Net. In this work we collect intermediate features at several scales of the encoder, as can be seen in Fig. 2. The resulting stream $\{s_{k_j}(t_j)\}_{j=1}^T$ has the following properties: (i) it is asynchronous, which means that features from different sensors may appear in any order in the sequence and (ii) the data rates are variable, which means that the time interval between features changes over time. We, therefore, need to make sure to fuse these features correctly. To retain both these properties we use sensor-specific state combination operators h_{k_j} which take a feature s_{k_j} from sensor k_j and use it to update the latent variable Σ_{j-1} at time t_{j-1} :

$$\Sigma_j = h_{k_j}(\Sigma_{j-1}, s_{k_j}, \theta). \quad (2)$$

The combination of features with latent variable can be performed via simple summation, concatenation or with more elaborate neural network layers. In this work, we choose to combine the states with a convolution gated recurrent unit (convGRU) [38] for each sensor and at each scale, which is depicted as a black square in Fig. 2. This is important for the

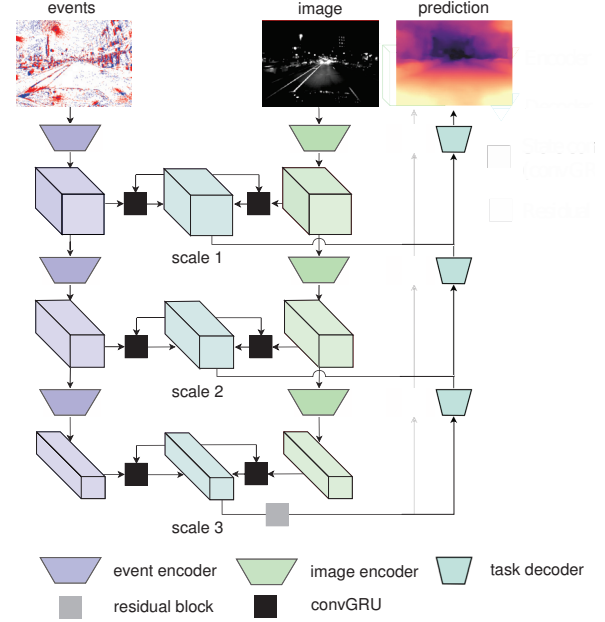


Fig. 2. Network architecture of RAM Net for monocular depth estimation with events and frames. The architecture is a fully convolutional encoder-decoder architecture based on U-Net [39]. Inputs can be processed asynchronously. Predictions are dependent on all previous inputs due to the recursive calculation of the state.

use of encoder-decoder architectures with skip connections. The update equation thus reads:

$$\tilde{\Sigma}_j = \Phi(r \odot \Sigma_{j-1}, s_{k_j}) \quad (3)$$

$$\Sigma_j = (1 - z) \odot \Sigma_{j-1} + z \odot \tilde{\Sigma}_j \quad (4)$$

$$\text{with } z = f(\Sigma_{j-1}, s_{k_j}) \text{ and } r = g(\Sigma_{j-1}, s_{k_j})$$

Where f and g denote single convolutions followed by a sigmoid and Φ denotes a convolution followed by tanh. By using dedicated state combination operators we ensure that features from different sensors are first mapped to a suitable space such that they can be easily fused with Σ_j . The state combination operators are thus applied sequentially, alternating between different sensors, as their measurements become available. Due to the recursive update of the state Σ_j , it always contains information of both modalities, independent of which modality was seen last. This produces a stream of latent variables $\{\Sigma_j\}_{j=1}^T$. These can be decoded to the task variable y_j at each timestep t_j with the task decoder \mathcal{T}

$$y_j = \mathcal{T}(\Sigma_j, \theta) \quad (5)$$

Since the y_j only depends on Σ_j , RAM Networks are Markovian, i.e. they are independent of the history of sensor measurements, given Σ_j . Thus Σ_j encodes the history of measurements. There is a strong connection with hidden Markov models (HMMs), which similarly model a latent variable given sensor measurements and impose the same conditional independence as RAM Networks. In the following section, we will explore how this framework is applied to the task of event and frame-based depth estimation.

B. Network Architecture for monocular depth estimation

Fig. 2 shows in detail how the general network architecture is adapted for this task. The basic architecture is inspired by U-Net [39], which was previously used for monocular depth estimation with events [21]. At each scale, we use the skip connection as intermediate features s_{k_j} . The intermediate features s_{k_j} are combined with Σ_{j-1} via a ConvGRU layer to produce Σ_j . The latent variable at the lowest scale is directly fed into a residual block, followed by three decoder levels. At the skip connections, the decoder input is combined with the state of the corresponding stage by summation. Each encoder consists of a down-sampling convolution with kernel size 5 and stride 2. The residual block is composed of two convolutions with kernel size 3 and the decoders use bilinear upsampling followed by a convolution with kernel size 5. All layers use a ReLU activation function.

C. Event Generation Model

Before we see the application of RAM Networks to events and frames, we review the working principle of the event camera. Event cameras contain independent pixels \mathbf{u} that react to changes in the log brightness signal $L(\mathbf{u}, t)$. If the magnitude of the log brightness changes by more than a threshold C since the last event, a new event $e_k = (x_k, y_k, t_k, p_k)$ is triggered at that pixel location $\mathbf{u} = (x_k, y_k)^T$. The polarity p_k takes the values $\{-1, +1\}$, depending on the direction of brightness change. The generative event model for an ideal sensor [18], [2] states that an event with polarity p_k is triggered at pixel \mathbf{u}_k and timestamp t_k when:

$$\Delta L(\mathbf{u}_k, t_k) = p_k (L(\mathbf{u}_k, t_k) - L(\mathbf{u}_k, t_k - \Delta t_k)) \geq C. \quad (6)$$

Here Δt_k is the time since the last event at the same pixel.

D. Event Representation

To leverage existing convolutional neural networks architectures, events are typically converted to a fixed-size tensor. A set of input events $\epsilon = \{e_i\}_{i=0}^{N-1}$ in the time window $\Delta T = t_{N-1} - t_0$ is and drawn into a voxel grid with spatial dimensions $H \times W$ and B temporal bins [7], [6], [20].

$$\mathbf{V}(x, y, t) = \sum_i p_i \delta(x - x_i, y - y_i) \max\{0, 1 - |t - t_i^*|\}, \quad (7)$$

Here $t_i^* = \frac{B-1}{\Delta T}(t_i - t_0)$. As in [7], the number of bins B is set to 5 for all experiments.

E. Depth Representation

The metric depth $\hat{D}_{m,k}$ is first converted into a normalized log depth map $\hat{D}_k \in [0, 1]$ which facilitates learning of large depth variations [21], [31], [40]. The maps are constructed pixel-wise. The logarithmic depth is calculated as

$$\hat{D}_k = \frac{1}{\alpha} \log \frac{\hat{D}_{m,k}}{D_{\max}} + 1, \quad (8)$$

where α is chosen such that the closest observed depth is mapped to 0. For EventScape we choose $\alpha = 5.7$ and

$D_{\max} = 1000$ m whereas for MVSEC we choose $\alpha = 3.7$ and $D_{\max} = 80$ m. For monocular depth estimation, the depth can only be estimated up to scale. Some methods circumvent this problem by using additional information from a stereo setup, IMU data, or camera poses [19], [12], [35], [41]. However, using a recurrent deep learning architecture and manually setting the maximum depth observed in the dataset leads to satisfying results without additional data.

F. Training

RAM Net is trained in a supervised fashion by using ground truth depth map labels provided at sparse time points. In the case of EventScape, these labels are provided by CARLA and for MVSEC these are measured using a LiDAR. We use a combination of a scale-invariant loss and a multi-scale scale-invariant gradient matching loss which we compute for valid ground truth labels and sum over a sequence of events and images. The scale invariant loss [40] is defined as:

$$\mathcal{L}_{k,si} = \frac{1}{n} \sum_{\mathbf{u}} (R_k(\mathbf{u}))^2 - \frac{1}{n^2} \left(\sum_{\mathbf{u}} R_k(\mathbf{u}) \right)^2, \quad (9)$$

with n counting the valid ground truth pixels. The multi-scale scale-invariant gradient matching loss [31] is defined as:

$$\mathcal{L}_{k,grad} = \frac{1}{n} \sum_s \sum_{\mathbf{u}} |\nabla_x R_k^s(\mathbf{u})| + |\nabla_y R_k^s(\mathbf{u})|. \quad (10)$$

Where ∇_x and ∇_y compute the edges in the x and y direction respectively by using the well-known Sobel operator. The loss is calculated over four different scales, indicated with s . It favors smooth gradient changes and enforces sharp depth discontinuities in the prediction [31] since it tries to match the gradients in the ground truth depth. The resulting total loss for a sequence of length L is

$$\mathcal{L}_{tot} = \sum_{k=0}^{L-1} \mathcal{L}_{k,si} + \lambda \mathcal{L}_{k,grad}. \quad (11)$$

The weight of the gradient loss λ is chosen as 0.25. To facilitate learning, the input data is normalized, cropped randomly to 224×224 and we apply random horizontal flipping. Non-zero entries in the voxel grid normalized such that their mean and variance are 0 and 1 respectively.

The network is implemented in PyTorch [43] and optimized using the ADAM optimizer [44] with a batch size of 8. On EventScape, we choose a learning rate of 0.0003 and train for 27'000 iterations. For MVSEC on the other hand, we use a learning rate of 0.001 and train for 7'600 iterations.

IV. EXPERIMENTAL SETUP

In a first step, we validate RAM Net on the synthetic EventScape dataset where we compare against event-only, frame-only, and event and frame-based baselines. We show that our method not only achieves better performance compared to unimodal baselines but also outperforms multi-modal baselines. We then show results on the real world MVSEC dataset, where we compare against state-of-the-art frame-based [31] and event-based [21], [6] methods as well as our

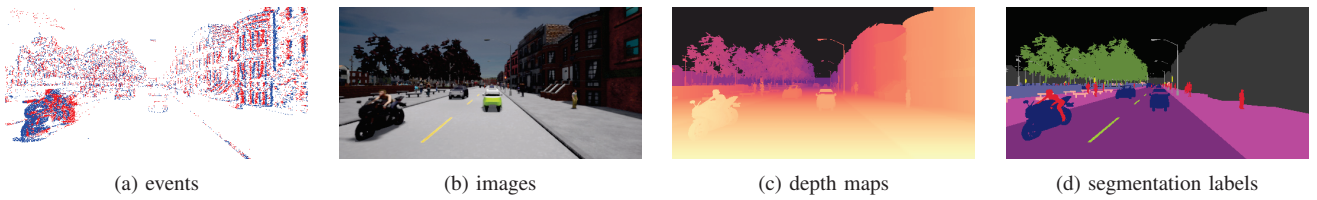


Fig. 3. Sample data from the EventScape dataset. It contains events (a), images (b), depth maps (c), segmentation labels (d), and vehicle navigation parameters. Images, depth maps, and segmentation labels are provided at 25Hz while vehicle navigation parameters are provided at 1000Hz. Events are generated by converting 500Hz images to events using the event camera plugin in CARLA [21] which is based on the event camera simulator [42]. We release this dataset to enable further research in multimodal learning with events.

image-only baseline. We show that RAM Net outperforms state-of-the-art methods by up to 30% in terms of absolute relative depth error.

A. Datasets

EventScape: EventScape is a large scale, synthetic dataset recorded in the CARLA simulator [11]. It contains 743 sequences of driving data with a total of 171'000 labels, equivalent to around 2 hours of driving. It is recorded in different cities with different dynamic actors like pedestrians and cars. In addition to events and image data, we recorded depth, semantic segmentation, and various vehicle control data. An example from the dataset can be viewed in Fig. 3. While frames, depth, and semantic labels are provided synchronously at 25Hz, vehicle control data is recorded at 1000 Hz. Events are generated by using the event camera plugin in CARLA [21] which renders images at 500Hz and converts them to asynchronous events using the event camera simulator ESIM [42]. ESIM generates events by randomly sampling a contrast threshold from the range $[0.15, 0.25]$ for each sequence and applying Eq.(6) to the 500Hz frames. To improve the realism of the events we include a refractory period $100\mu s$ which is similar to that observed in real sensors. Furthermore, this procedure was employed in previous works [7], where it was shown to boost performance. The events and frames are synchronized and aligned, resembling the output of a DAVIS sensor [1]. For training, CARLA towns 1, 2, and 3 are used, validation and test data are drawn from town 5 from geographically separated areas.

MVSEC: The multi-vehicle stereo event camera dataset (MVSEC) [45] consists of data recorded using two DAVIS event cameras with a resolution of 346×260 pixels in a stereo rig, together with measurements from a LiDAR for ground truth depth. It features several day and night-time driving sequences as well as indoor sequences recorded on a quadcopter. The depth maps were recorded at 20Hz while the grayscale images were recorded at 10 Hz for night time and 45 Hz for daytime sequences.

B. Baselines

To understand the impact of multimodality and the asynchronous update scheme employed by RAM Net we implement three recurrent baselines termed (i) purely event-based (E), (ii) purely frame-based (I) and (iii) event and frame-based (E+I). Of these baselines E and I are designed to understand

the impact of multimodality, while E+I ablates the proposed asynchronous update scheme. They have identical residuals and decoders but instead of recurrent state combination operators, they feature recurrent convLSTM encoders at each level. While E only receives voxel grids as input, I receives only gray-scale frames and E+I receives stacks of voxel grids and frames, similar to [9]. For E+I when a new voxel grid arrives, we stack it with a copy of the last seen image. Of the described baselines E+I has access to the same amount of information as our method, but there is a temporal mismatch between event and image data. Supervision is performed sparsely, indicated with a blue arrow in Fig. 4. Finally, to understand the effect of recurrency we compare against a simple event- and frame-based feed-forward network, based on [9] which we term E+I no recurrency. Here we replace the convLSTM blocks with simple convolutions and train on 200 ms of events stacked with the corresponding images at 5 Hz.

V. RESULTS

A. Effect of Input Modality and Architecture

As a first step, we train all baselines and RAM Net with 5 Hz images and depth labels, and voxel grids constructed at 25 Hz on EventScape. We construct sequences from 10 sub-sequences, each containing 5 voxel grids followed by one frame and one depth label. We then pass these sub-sequences at each step of the sequential model unrolling, resulting in 10 frames and 50 voxel grids per sequence. A loss is then computed after each depth label, i.e. at 5Hz. After each sub-sequence, RAM Net predictions after the last voxel grid and after the subsequent image are temporally aligned. We therefore average the per-pixel loss for both predictions, which we found improved training. By contrast, the baselines only see a single loss.

Results: The results after computing the absolute relative depth error [31] can be seen in Fig. 4 (b). On the x-axis, a frame index 0 indicates that the metric was averaged over image predictions. For E we evaluate the prediction from the most recent voxel grid. The E+I baseline sees the concatenation of the new image and the current events. For frame index n the metric was computed on predictions after n event tensors. As the I and E+I recurrent baselines are only able to predict at image frame rate, it is the only configuration that is represented as a single data point in Fig. 4. Baselines E and E+I predict an output after each event voxel grid or combined event-image input. As RAM Net outputs separate

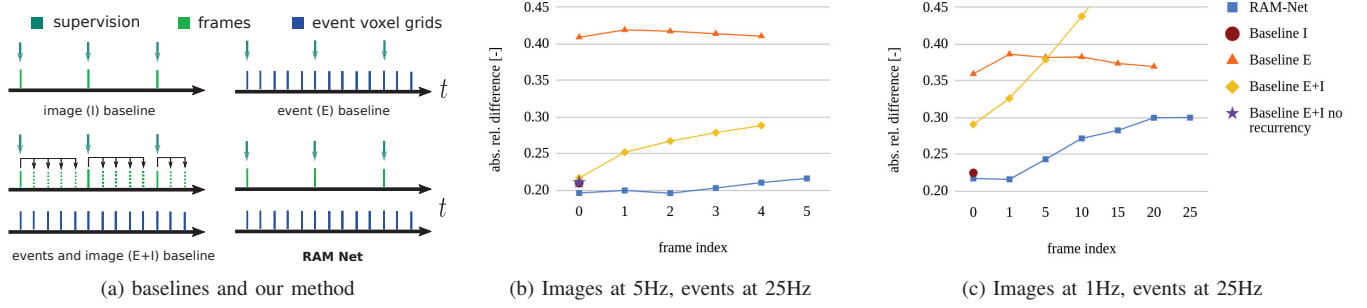


Fig. 4. Comparison of RAM Net and baselines (a). RAM Net outperforms the baselines in all data points. The x-axis (b-c) depicts the input to the network before the prediction that was used to calculate the corresponding data point. (b) Images occur at 5 Hz, depth labels and event data is available at 25Hz, which we use to train network. (c) Generalization study with images at 1Hz and events at 25 Hz, as during training.

predictions for the events and image inputs, it contains an additional data point in Fig. 4, indicated as frame index 5.

After the frame prediction, RAM Net outperforms all baselines. As we move further from the initial frame, the prediction error gracefully increases but remains at lower levels than methods E and E+I (49% and 38% reduction respectively in terms of absolute relative difference at frame index 4 in Fig. 4 (b)). Compared to RAM Net, the I baseline shows a 5% higher difference, but cannot generate predictions at a high temporal resolution as the other methods. Furthermore, while E+I has a strong prediction after a new image input the prediction error increases due to the gradual misalignment between events and copied frames. The E baseline has the worst performance but has constant metrics over time since it is independent of the last seen image. Finally, E+I no recurrency at 0.211 exhibits a slightly lower error at frame index 0, than its recurrent counterpart, E+I, at 0.216. This small difference likely comes from the fact that, for E+I, events and images become more temporally misaligned, as the frame index increases, leading to a higher error (Fig. 4 b). Upon seeing a new image, the network learns to correct this error but fails to do so completely. By contrast, the E+I no recurrent baseline does not suffer from this misalignment and thus performs slightly better. However, crucially, this comes at the cost of a lower prediction rate compared to the E+I recurrent baseline (5 Hz vs. 25 Hz). RAM Net outperforms both methods with an error of 0.198 by leveraging the advantages of both. It does not suffer from temporal misalignment of inputs since images are never copied and is also able to make predictions between frames. These experiments show (i) the benefit of combining frames and events which highlights their complementarity, (ii) the effect of temporal alignment for recurrent methods and (iii) the importance of recurrency which enables high-quality depth predictions in the blind-time between image frames.

B. Generalization to Different Data Rates

To further study the network behavior, we analyze its ability to generalize to new frame rates. For the experiment in Fig. 4 (c) we use the networks trained on 5Hz image data and tested them on 1Hz image data. Event data was still available at 25 Hz. This task is much more challenging since the gaps between images become larger which means that networks need to be able to retain a longer memory.

Results: The I baseline performs slightly worse, which can be explained by the increased time interval between seen images. The small decrease in performance indicates that I does not heavily depend on recurrency. The prediction quality after the image input also decreases for RAM Net. As before, the accuracy of RAM Net's predictions gracefully degrades for increasing frame index, reaching a plateau at index 20. Nonetheless, RAM Net retains a low error compared to E+I and E. This time, the error for E+I grows drastically, even surpassing E. This is because for large time gaps the events and copies of the previous frame show significant misalignment. This shows that RAM Net retains additional information from the image even after a large time-span. Furthermore, this shows that RAM Net can generalize well to data inputs at different frequencies.

C. Real World Experiments

Next we evaluate our method on MVSEC. Since ground truth depth and images are no longer synchronized, labels and network predictions need to be temporally aligned. To do this, we split the events between depth labels into equal packets of 10'000 events, distributing the surplus of events equally, if required. Next, we generate voxel grids from packets, with the timestamp corresponding to that of the last event in the packet. We feed this irregular sequence of voxel grids to the network together with frame sequence, alternating between frames and voxel grids as they appear.

To test the generalization of our network to real data we first test on four *outdoor_night* sequences and the *outdoor_day1* sequence, without retraining and report the mean absolute depth error between the prediction and ground truth depth maps, indicated by [S] in Tab. I. Methods fine-tuned on the MVSEC dataset (as was done in [21]) are indicated with [S→R]. We compare against different event-based [21], [6] and frame-based [31] state-of-the-art methods as well as our image-based baseline. We evaluate the error for ground truth depth up to 10 m, 20 m and 30 m, only considering pixels with valid depth measurements. An example of the ground truth depth map is shown in Fig. 5. For the event-based baselines, events are grouped into a single packet between depth labels. While [6] processes them in a feed-forward network, [21] uses a recurrent architecture. Similar to our method, [21] pretrains on synthetic events from the DENSE dataset, recorded in CARLA. Since both

TABLE I
AVERAGE ABSOLUTE DEPTH ERRORS (IN METERS) AT DIFFERENT MAXIMUM CUT-OFF DEPTHS.

Dataset	Cutoff	frame-based			event-based		event and frame-based	
		MegaDepth [31]	I baseline [S]	I baseline [S \rightarrow R]	Zhu et al. [6]	E2Depth [21]	RAM Net [S]	RAM Net [S \rightarrow R]
outdoor day1	10 m	2.37	13.02	1.74	2.72	1.85	3.77	1.39
	20 m	4.06	11.37	2.55	3.84	2.64	4.27	2.17
	30 m	5.38	11.29	3.07	4.40	3.13	5.17	2.76
outdoor night1	10 m	2.54	7.93	2.72	3.13	3.38	4.00	2.50
	20 m	4.15	8.81	3.35	4.02	3.82	4.53	3.19
	30 m	5.60	9.29	3.99	4.89	4.46	5.18	3.82
outdoor night2	10 m	3.92	4.79	1.36	2.19	1.67	2.41	1.21
	20 m	5.78	4.52	2.42	3.15	2.63	3.24	2.31
	30 m	7.05	5.34	3.47	3.92	3.58	3.47	3.28
outdoor night3	10 m	4.15	4.64	1.20	2.86	1.42	2.21	1.01
	20 m	6.00	4.08	2.44	4.46	2.33	3.31	2.34
	30 m	7.24	4.93	3.64	5.05	3.18	4.72	3.43

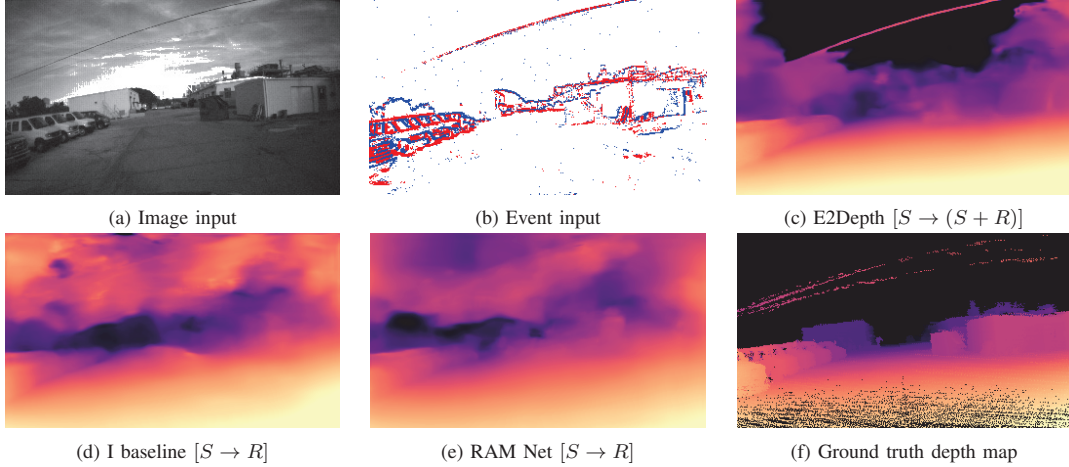


Fig. 5. Qualitative performance on the MVSEC dataset. (a) and (b) show the image and event input. (c) Results of E2Depth [21] when trained on both simulation and real-world data simultaneously. This training configuration allows to achieve qualitative better results for the sky. However, it was not applicable to our network as EventScape and MVSEC are recorded with different intrinsic camera parameters. (d) & (e) I baseline and RAM Net retrained on real-world data. The sky is predicted wrong, however, objects on the ground are correct which is reflected in the superior metrics compared to E2Depth.

event-based methods are trained on *outdoor_day2* we did not include it for evaluation. The frame-based methods on the other hand only process the images and [31] uses a feed-forward method. Since the method [31] is trained on images from the internet (the MegaDepth Dataset [31]) we include the recurrent image-baseline from Sec. IV-A. For image-based methods, we evaluate only on predictions with a temporal misalignment with depth maps of at most 10 ms.

To improve the prediction quality, we fine-tune our network and the I baseline on the *outdoor_day2* sequence from MVSEC, as was done in [21]. We choose different values $\alpha = 3.7$ and D_{\max} to normalize depth and unroll the sequence a by 8 steps due to memory limitations. Apart from these changes, we train with the same data configurations as during testing. Again one step is defined as the data measured between two depth labels. However, while in EventScape two losses could be used per label due to image and depth map alignment, here we only use a single loss.

Results: For methods only trained on synthetic data, (indicated with [S]), RAM Net generalizes better than I. However, sim2real results are significantly worse than E2Depth [21]. The results for [S \rightarrow R] are significantly better for both I and RAM Net. The high performance of the I baseline shows the importance of image data for depth prediction by outperforming E2Depth [21]. However,

as noted, purely frame-based methods cannot generate predictions in the blind-time between frames. Finally, RAM Net outperforms all other methods on almost all datasets.

Compared to E2Depth [21], our method can reconstruct more details and we observe that there are artifacts in the sky. This is because during training these regions are masked since they do not have valid LiDAR depth. In fact, the I baseline also suffers from these artifacts. We thus also think that these artifacts are due to the difference of images in the training and test set. While in the training set the sky is clear, in the test set the sky is overcast, so the network mistakenly thinks it is structure. Purely event-based methods such as E2Depth [21] do not rely on images, and thus do not suffer from these issues, however, their accuracy in other areas of the image is reduced as indicated by Tab. I. Training on more realistic data could significantly improve sky predictions. Furthermore, the MVSEC depth labels could be improved by using semantic segmentation to detect the sky and adding maximum depth to the depth labels at these pixels. Another approach would be using multitask learning with depth estimation and semantic segmentation to mask the sky at test time.

D. Timing Results

We measure the processing speed of our architecture on a Quadro RTX 8000 GPU. While processing a voxel grid takes

3.08 ms, processing an image takes 3.14 ms, resulting in a throughput of up to 320 Hz.

VI. CONCLUSION

In this work, we introduce RAM Net, a novel network architecture which generalizes traditional RNNs to handle asynchronous and irregular data from multiple sensors. We applied RAM Net to the task of monocular depth estimation, where we showed that it can make high-quality predictions in the blind-time between image frames, as shown by experiments on simulated data from EventScape and real data from MVSEC. On both synthetic and real datasets, RAM Net outperforms unimodal baselines by leveraging the complementarity of events and frames. Furthermore, it also outperforms several baselines and state-of-the-art method based on classical RNNs. This is because our network is more robust to changes to the input frequency. Finally, to enable further research on multimodal learning and depth predictions for events we release EventScape, a synthetic large-scale dataset recorded in the CARLA simulator, comprising events, frames, semantic labels, depth maps, and vehicle navigation parameters, to enable further research on multimodal learning using events. This work makes an important step towards solving the problem of fusing asynchronous and multimodal sensory data for learning.

REFERENCES

- [1] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, "A 240x180 130dB 3 μ s latency global shutter spatiotemporal vision sensor," *IEEE JSSC*, vol. 49, no. 10, pp. 2333–2341, 2014.
- [2] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE T-PAMI*, 2020.
- [3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [4] D. Neil, M. Pfeiffer, and S.-C. Liu, "Phased LSTM: Accelerating recurrent network training for long or event-based sequences," in *NeurIPS*, 2016.
- [5] C. R. T. Q., Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *NeurIPS*.
- [6] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion," in *CVPR*, 2019.
- [7] H. Rebecq, R. Ranfil, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE T-PAMI*, 2019.
- [8] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. Mahony, and D. Scaramuzza, "Fast image reconstruction with an event camera," in *IEEE WACV*, 2020.
- [9] S. Pini, G. Borghi, R. Vezzani, and R. Cucchiara, "Video synthesis from intensity and event frames," in *ICIAP*, 2019.
- [10] Y. Hu, J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, "DDD20 End-to-End event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction," in *ITSC*, 2020.
- [11] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *CoRL*, 2017.
- [12] A. Rosinol Vidal, H. Rebecq, T. Horstschäfer, and D. Scaramuzza, "Ultimate SLAM? combining events, images, and IMU for robust visual SLAM in HDR and high speed scenarios," *IEEE RA-L*, 2018.
- [13] B. Kueng, E. Mueggler, G. Gallego, and D. Scaramuzza, "Low-latency visual odometry using event-based feature tracks," in *IROS*, 2016.
- [14] D. Gehrig, H. Rebecq, G. Gallego, and D. Scaramuzza, "Asynchronous, photometric feature tracking using events and frames," in *ECCV*, 2018, pp. 766–781.
- [15] D. Gehrig, H. Rebecq, G. Gallego, and D. Scaramuzza, "EKLt: Asynchronous photometric feature tracking using events and frames," *IJCV*, 2019.
- [16] C. Scheerlinck, N. Barnes, and R. Mahony, "Continuous-time intensity estimation using event cameras," in *ACCV*, 2018.
- [17] L. Pan, C. Scheerlinck, X. Yu, R. Hartley, M. Liu, and Y. Dai, "Bringing a blurry frame alive at high frame-rate with an event camera," in *CVPR*, 2019.
- [18] G. Gallego, J. E. A. Lund, E. Mueggler, H. Rebecq, T. Delbruck, and D. Scaramuzza, "Event-based, 6-DOF camera tracking from photometric depth maps," *IEEE T-PAMI*, 2018.
- [19] S. Tulyakov, F. Fleuret, M. Kiefel, P. Gehler, and M. Hirsch, "Learning an event sequence embedding for dense event-based deep stereo," in *ICCV*, October 2019.
- [20] D. Gehrig, A. Loquercio, K. G. Derpanis, and D. Scaramuzza, "End-to-end learning of representations for asynchronous event-based data," in *ICCV*, 2019.
- [21] J. Hidalgo-Carrió, D. Gehrig, and D. Scaramuzza, "Learning monocular dense depth from events," *3DV*, 2020.
- [22] I. Alonso and A. C. Murillo, "EV-SegNet: Semantic segmentation for event-based cameras," in *CVPRW*, 2019.
- [23] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *Proc. IEEE*, vol. 41, pp. 423–443, 2019.
- [24] S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition - Volume 2*. ACM and Morgan & Claypool, 2018, vol. 21.
- [25] L. Chen, W. Tang, N. W. John, T. R. Wan, and J. J. Zhang, "Augmented reality for depth cues in monocular minimally invasive surgery," *CoRR*.
- [26] X. Liu, A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath, "Dense depth estimation in monocular endoscopy with self-supervised learning methods," *TMI*, 2020.
- [27] T. Laidlow, J. Czarnowski, and S. Leutenegger, "Deepfusion: Real-time dense 3d reconstruction for monocular slam using single-view depth and gradient predictions," in *ICRA*, 2019, pp. 4068–4074.
- [28] P. R. Palafox, J. Betz, F. Nobis, K. Riedl, and M. Lienkamp, "Semanticdepth: Fusing semantic segmentation and monocular depth estimation for enabling autonomous driving in roads without lane lines," *Sensors*, vol. 19, p. 3224, Jul 2019.
- [29] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," *CoRR*, 2014.
- [30] C. Godard, O. M. Aodha, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," *CoRR*, 2018.
- [31] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *CVPR*, 2018.
- [32] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," *CoRR*, 2018.
- [33] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," *CoRR*, 2017.
- [34] H. Rebecq, T. Horstschäfer, and D. Scaramuzza, "Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization," in *BMVC*, 2017.
- [35] A. Z. Zhu, N. Atanasov, and K. Daniilidis, "Event-based visual inertial odometry," in *CVPR*, 2017, pp. 5816–5824.
- [36] H. Kim, S. Leutenegger, and A. J. Davison, "Real-time 3D reconstruction and 6-DoF tracking with an event camera," in *ECCV*, 2016.
- [37] H. Rebecq, T. Horstschäfer, G. Gallego, and D. Scaramuzza, "EVO: A geometric approach to event-based 6-DOF parallel tracking and mapping in real-time," *IEEE RA-L*, vol. 2, no. 2, pp. 593–600, 2017.
- [38] M. Siam, S. Valipour, M. Jagersand, and N. Ray, "Convolutional gated recurrent networks for video segmentation," in *IEEE ICIP*, 2017.
- [39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.
- [40] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *CoRR*, 2014.
- [41] H. Rebecq, G. Gallego, and D. Scaramuzza, "EMVS: Event-based multi-view stereo," in *BMVC*, 2016.
- [42] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: an open event camera simulator," in *CoRL*, 2018.
- [43] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Workshops*, 2017.
- [44] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.
- [45] A. Z. Zhu, D. Thakur, T. Ozaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3D perception," *IEEE RA-L*, 2018.