Deep Event Stereo Leveraged by Event-to-Image Translation

Soikat Hasan Ahmed*, Hae Woong Jang*, S M Nadim Uddin, Yong Ju Jung†
College of Information Technology Convergence, Gachon University, Seongnam, South Korea
{soikat, leaping, nadim}@gc.gachon.ac.kr, yjung@gachon.ac.kr

Abstract

Depth estimation in real-world applications requires precise responses to fast motion and challenging lighting conditions. Event cameras use bio-inspired event-driven sensors that provide instantaneous and asynchronous information of pixel-level log intensity changes, which makes them suitable for depth estimation in such challenging conditions. However, as the event cameras primarily provide asynchronous and spatially sparse event data, it is hard to provide accurate dense disparity maps in stereo event camera setups - especially in estimating disparities on local structures or edges. In this study, we develop a deep event stereo network that reconstructs spatial image features from embedded event data and leverages the event features using the reconstructed image features to compute dense disparity maps. To this end, we propose a novel image reconstruction sub-network with a cross-semantic attention mechanism. A feature aggregation sub-network is also developed for accurate disparity estimation, which modulates the event features with the reconstructed image features by a stacked dilated spatially-adaptive denormalization mechanism. Experimental results reveal that our method outperforms the state-of-the-art methods by significant margins both in quantitative and qualitative measures.

Introduction

Stereo matching or finding corresponding pixels from two viewpoints for depth estimation has been regarded as one of the core problems in computer vision. Most of the recent deep learning-based stereo matching frameworks depend on passive frame-based cameras that provide intensity images at a fixed rate (e.g. 10ms latency). The traditional frame-based cameras are prone to motion blur and sudden intensity changes in high-speed scenarios. Thus, new camera sensors such as event-based cameras are emerging as alternatives to traditional frame-based cameras in stereo camera setups.

Event cameras use bio-inspired data-driven sensors that provide asynchronous and instantaneous per-pixel intensity change information and inherits several advantageous characteristics over traditional cameras, such as low latency (0.001ms), high dynamic range, high temporal resolution, no motion blur, and capability to handle different lighting conditions (Gallego et al. 2019). However, most of the current deep learning-based stereo models are not capable of taking full advantage of the temporally asynchronous and spatially sparse nature of the event data provided by event cameras. The main reasons behind this are difficulties in handling asynchronous event data and lack of spatial structure information.

Early pioneering studies have reported that event cameras can be used for depth estimation (Zhou et al. 2018; Tulyakov et al. 2019; Zhu, Chen, and Daniilidis 2018; Xie, Chen, and Orchard 2017; Dikov et al. 2017; Schraml and Belbachir 2010; Schraml, Schön, and Milosevic 2007). Prior studies introduced event data representations using hand-crafted methods (Sironi et al. 2018; Zhu et al. 2018b, 2019) to produce a dense event feature map from the sparse event data. (Kogler et al. 2014; Xie, Chen, and Orchard 2017) employed Belief Propagation on a Markov Random Field. (Xie, Zhang, and Wang 2018) used a traditional semi-global matching approach (Hirschmuller 2007). (Dikov et al. 2017; Piatkowska, Belbachir, and Gelautz 2013; Firouzi and Conradt 2016) extended the cooperative stereo algorithm using iterative non-linear operations to extract disparities. (Zhou et al. 2018; Zhu, Chen, and Daniilidis 2018) explicitly used camera motion information to improve depth estimation. (Zou et al. 2017) produced disparity at every location based on interpolation while (Tulyakov et al. 2019; Gehrig et al. 2019; Chen et al. 2020) focused on learning-based event representations instead of hand-crafted or rule-based event accumulation for deep learning-based stereo matching.

Nevertheless, though the existing representations encode the asynchronous event stream to a dense event feature, they do not contain any spatial intensity information of the scenes. To our best knowledge, despite the fact that the spatial structural information in intensity images is particularly important for stereo matching, all of the existing deep event-based stereo algorithms have used only event features directly extracted from the input event data without explicitly reconstructing spatial image features.

In this study, we propose a novel end-to-end deep event stereo architecture to generate spatial image features from input event data and use them as a guidance for the accurate stereo event matching. Inspired by the recent studies (Kalia, Navab, and Salcudean 2019; Rebecq et al. 2019; Scheerlinck

*Contributed equally.
†Corresponding author.

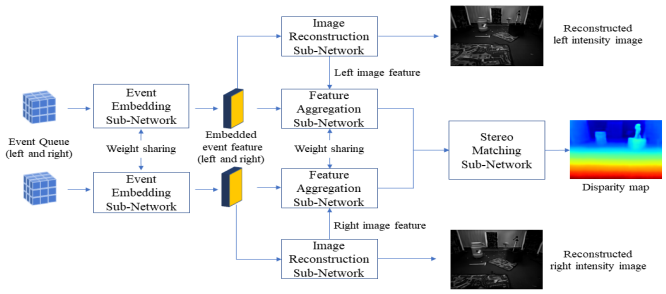

Figure 1: Overall architecture of the proposed deep event stereo network. The network consists of four sub-networks: event embedding, image reconstruction, feature aggregation, and stereo matching sub-networks. The proposed network takes event streams as input and embeds them to features, which are then fed to the image reconstruction sub-network for event-to-image translation. The reconstructed image features along with embedded event features are then fused and aggregated by the feature aggregation sub-network. The aggregated feature is then fed to the stereo matching sub-network for accurate dense depth estimation.

et al. 2020; VidalMata et al. 2019; Wang et al. 2019; Watkins et al. 2018), we explicitly reconstruct intensity images from the input event streams and use them as a guidance for event features. By doing so, we can use not only asynchronous event information but also spatial intensity image information. Our approach does not require additional input stereo images along with event streams for dense disparity estimation. Moreover, instead of directly fusing event streams and intensity images, we fuse the event features and reconstructed image features via a convolutional neural network (CNN) model. Specifically, an effective modulation mechanism for two different types of features is learned by the proposed feature aggregation sub-network.

In summary, the contributions of this work are as follows:

- We propose a deep event stereo network that extracts the event features leveraged by the reconstructed image features for dense disparity map estimation. To that end, the proposed network is trained with ground-truth disparity maps and intensity images in a supervised manner. In testing, the network outputs dense disparity maps using the spatio-temporal features extracted from the event features and the reconstructed image features.
- Inspired by recent studies in event-to-image translation, a novel image reconstruction sub-network is proposed to extract the image features reconstructed from the event features. The image reconstruction sub-network is based on a dual-path encoder-decoder network with a semantic attention mechanism.
- A feature aggregation sub-network is proposed to incorporate the reconstructed image features into the event features in a spatially adaptive modulation concept. The sub-network uses a stacked dilated SPatially-Adaptive DEnormalization (stacked dilated SPADE) mechanism (Schuster et al. 2019; Park et al. 2019) that modulates the event features using the reconstructed image features.

The proposed deep event stereo network has been evaluated using the public open dataset: the multi vehicle stereo event camera (MVSEC) dataset (Zhu et al. 2018a). The experimental results reveal that the proposed method outperforms the state-of-the-art methods in terms of the performance of depth estimation.

Deep Event Stereo Network

We construct our proposed architecture by adopting the stereo framework of (Tulyakov et al. 2019) as the baseline with several major architectural modifications. Our proposed network consists of four inter-linked sub-networks: event embedding sub-networks, novel image reconstruction sub-networks for event-to-image reconstruction, feature aggregation to fuse event features and image features, and a stereo matching sub-network, as shown in Figure 1.

The event embedding sub-network contains a kernel network with continuous fully connected layers, following (Tulyakov et al. 2019), for left/right event-to-feature embedding. The embedded event features are then fed to both the image reconstruction sub-network and feature aggregation sub-network as inputs. The proposed image reconstruction sub-network takes event features as input and uses a dual-path encoder-decoder network with a novel attention mechanism to reconstruct corresponding left and right images and also to obtain image features of the same shape of event features. The proposed feature aggregation sub-network takes embedded event features and reconstructed image features as inputs and fuses the features with a stacked dilated SPADE mechanism to obtain a final fused and aggregated feature, which is then fed into a stereo matching sub-network to obtain dense disparity maps. Note that the event embedding and stereo matching sub-networks are applied using the same methods as used in the previous study. Thus, the following subsections introduce the proposed sub-networks in detail (i.e., image reconstruction sub-network and feature aggregation sub-network).

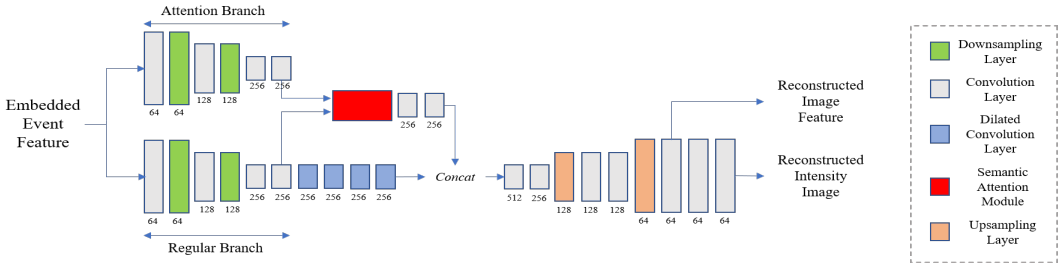

Figure 2: Architecture of the image reconstruction sub-network. It takes embedded event features as input and processes them with a dual-path encoder-decoder architecture with a semantic attention mechanism to generate a corresponding intensity image along with image feature.

Image Reconstruction Sub-Network

As shown in Figure 2, the image reconstruction sub-network consists of two separate branches, namely, a regular branch and an attention branch, in the encoder segment (Uddin and Jung 2020). The sub-network takes embedded event feature of shape $R^{c\times h\times w}$ as input, where c, h and w represent channels, height and width, respectively. The input is fed to both regular and attention branches, each of which consists of a set of convolutions and down-sampling layers in the encoder. The corresponding features from both branches are fed to the proposed semantic attention module, which calculates global context information.

The proposed semantic attention module consists of two separate attention mechanisms - spatial context attention and cross-semantic attention, as shown in Figure 3. For the spatial context, the features are fed to 1×1 convolutions followed by an element-wise addition and $ReLU$ operation, which is then fed to another 1×1 convolution. Let f_{br_1} and f_{br_2} be features from regular and attention branch respectively, then the operation can be expressed as

$$f_{spatial} = w_3(ReLU(w_1(f_{br_1}) + w_2(f_{br_2}))), \quad (1)$$

where w_1, w_2 and w_3 denote 1×1 convolutions.

For the cross-semantic attention, we modify the *Squeeze and Excitation Network* (Hu, Shen, and Sun 2018) to perform a re-calibration of the global context between the features in a reciprocal manner, as shown in Figure 4. More specifically, first, we perform a global average pooling similar to the *SE* to squeeze spatial information into a channel descriptor which generates channel-wise statistics. For f_{br_1} with height H and width W, we obtain the channel descriptors z_{c_1} as

$$z_{c_1} = S(\frac{1}{HW}\sum_{i=1}^{H}\sum_{i=1}^{W}(f_{br_1}(i,j))), \quad (2)$$

where S denotes the sigmoid function. To obtain z_{c_2} for f_{br_2}, we follow the same process as z_{c_1}. After obtaining the channel descriptors, we perform a feature re-calibration in a cross manner by multiplying z_{c_1} with f_{br_2} and z_{c_2} with f_{br_1}. This re-calibration facilitates the fusion of the global context of the features. We then perform a gating mechanism (Dauphin et al. 2017) for normalizing and learning semantic correspondences. Specifically, the final feature $f_{context}$ from the cross-semantic attention can be expressed as

$$f_{context} = S(z_{c_1} \odot f_{br_2}) \odot ReLU(z_{c_2} \odot f_{br_1}), \quad (3)$$

where $\odot$ denotes an element-wise product. The final output $f_{attention}$ from the attention module is obtained by multiplying the spatially attended $f_{spatial}$ and $f_{context}$, which can be expressed as

$$f_{attention} = f_{spatial} \odot f_{context}. \quad (4)$$

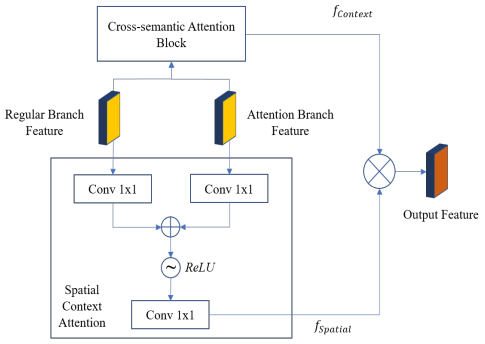

Figure 3: Semantic attention module. The module takes the input from the dual-path encoder in the image reconstruction sub-network and calculates both spatial context and cross-semantic context among the features.

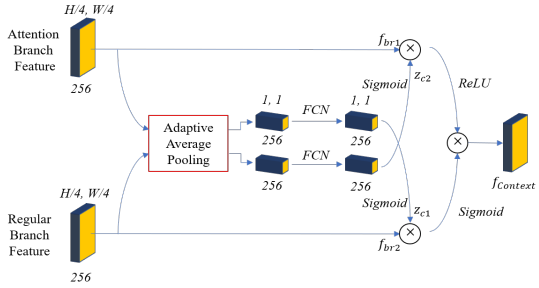

Figure 4: Cross-semantic attention. The module takes two features from the dual-path encoder and calculates respective channel descriptors. Then it reweighs the features in reciprocal and gated manners to integrate the cross-semantic global context.

The regular branch of the encoder uses 3×3 convolution layers with dilated convolution layers with a kernel size of 3×3 and rates of 2, 4, 8, and 16 to achieve large receptive fields that contribute to better feature extraction. The output from both the regular and attention branches are concatenated channel-wise and fed into a single decoder. The decoder consists of several convolutions and up-sampling layers (i.e., up-sampling operation followed by a convolution operation) which outputs a reconstructed intensity image and associated image features.

Feature Aggregation Sub-Network

The feature aggregation sub-network takes the embedded event feature f_{event} and the reconstructed image feature f_{image} as inputs, as shown in Figure 5. f_{image} has more structure information due to the influence of the reconstructed image feature. The goal of this aggregation sub-network is to map the structure information of f_{image} to f_{event}. To this end, we use a conditional denormalization method based on the stacked dilated convolution depicted in Figure 6.

The proposed aggregation sub-network is based on the spatially-adaptive denormalization (SPADE) method (Park et al. 2019) that modulates the existing feature using the conditional feature with learned scale and shift parameters. Let $X^{c\times h\times w}$ be the input feature and $Y^{c\times h\times w}$ be the conditional feature, then the SPADE is performed in a channel-wise manner as follows:

$$X' = \gamma(Y)(\frac{X - \mu(X)}{\sigma(X)}) + \beta(Y), \tag{5}$$

where $\mu(X)$ and $\sigma(X)$ are the mean and standard deviation calculated from the spatial dimension of each feature and $\gamma(Y)$ and $\beta(Y)$ are the learned modulation parameters (i.e. mean scale and shift) of the denormalization layer, respectively. In our case, X is the event feature and Y is the reconstructed image feature. Note that, in the SPADE, γ and β are tensors, not scalar values. That is, it denormalizes the normalized feature with the spatially varying learned scale and shift modulation parameters. The γ and β values are obtained through the convolution layers.

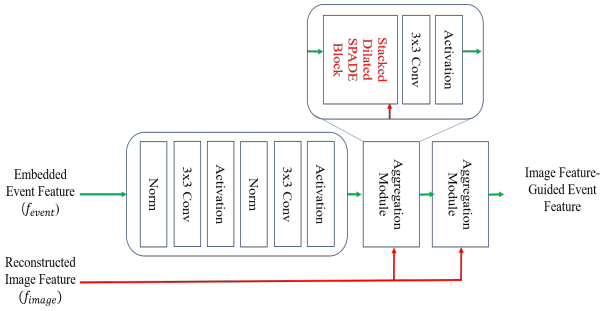

Figure 5: Feature aggregation sub-network. The sub-network takes both embedded event features and image features as inputs and processes them with a stacked dilated SPADE module to generate a fused and aggregated feature.

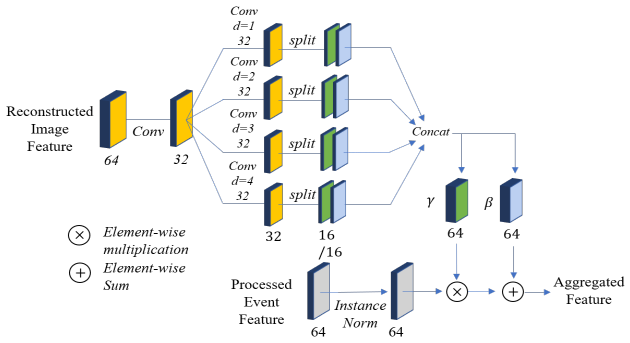

Figure 6: Stacked dilated SPADE block. The event features are conditionally denormalized by the scale γ and shift β. These scale and shift values are modulation parameters and extracted from the reconstructed image features through the stacked dilated convolutions.

Notably, the SPADE was proposed as a conditional denormalization method for image synthesis tasks. However, in this paper, we modify the SPADE to fit the stereo matching task. Specifically, to compensate for the insufficient structural information of the event feature, the proposed conditional denormalization uses the stacked dilated convolution.

As shown in Figure 6, the modulation parameters γ and β are computed with four different dilated convolutions (i.e., with the dilation factors d as 1, 2, 3, and 4). The stacked dilated convolution uses large receptive fields and hence provides discriminative spatial features in the dense stereo matching field. Increasing the receptive field of the conditional feature (i.e., reconstructed image feature) results in more accurate matching results, comparing with a single dilated convolution-based denormalization.

Stereo Matching Sub-Network

After obtaining the left and right aggregated features from the reconstructed image features and embedded event features, the aggregated features are fed to a stereo matching sub-network, following (Tulyakov et al. 2019). Note that, for the stereo matching sub-network, we applied the existing model used for the event stereo matching framework (Tulyakov, Ivanov, and Fleuret 2018). It should be further noted that any stereo matching sub-network (even for frame-based stereo models) can be independently combined with our proposed image reconstruction and feature aggregation sub-networks by simply modifying the interface between the sub-networks (as verified in the ablation studies).

In the stereo matching sub-network, after matching all features for all disparities, a resultant 4D feature of size $\frac{c}{8} \times \frac{d_{max}}{4} \times \frac{h}{4} \times \frac{w}{4}$ is constructed, where c, h and w and d_{max} denote channels, height and width and maximum disparity to be computed, respectively. Then, a matching cost feature C of size $\frac{d_{max}}{2} \times h \times w$ is constructed, which is then fed to a sub-pixel estimator. Then, an estimated disparity $\hat{D}$ is calculated as

$$\hat{D}_{y,x} = \sum_{\substack{j \\ j:|\hat{j}-j|\leq\delta}} d(j) \cdot softmin(C_{j,y,x}), \tag{6}$$

where $\hat{j} = \arg\min_j(C_{j,y,x})$, δ is an estimator support, and $d(j) = 2 \cdot j$ is a disparity corresponding to index j in the matching cost tensor.

Objective Function

For training the image reconstruction sub-network, we adopt the $l_1 + SSIM$ loss. Specifically, let the reconstructed image be I_{recon} and respective ground truth intensity image be I_{gt}, then the reconstruction loss l_R can be expressed as

$$l_R = l_1(I_{gt}, I_{recon}) + SSIM(I_{gt}, I_{recon}). \quad (7)$$

For training the stereo matching sub-network, we adopt sub-pixel cross entropy loss, following (Tulyakov, Ivanov, and Fleuret 2018). Specifically, the sub-pixel cross entropy loss l_Φ can be expressed as

$$l_\Phi = \frac{1}{HW} \sum_{y,x} \sum_j Laplace(d(j)|\mu = D^{gt}_{y,x}, b) \times \zeta, \quad (8)$$

$$\zeta = \log(\underset{j}{softmin}(C_{j,y,x})), \quad (9)$$

where $Laplace(d(j)|\mu = D^{gt}_{y,x}, b)$ is a discretized and normalized Laplace probability density function over disparities with a mean equal to the ground truth disparity and diversity b. In our setup, we set b as 2, following the baseline.

Our final objective function of the proposed model, l_{final} is given by

$$l_{final} = l_R + l_\Phi. \quad (10)$$

Experiments and Results

Experimental Setup

We evaluate our proposed method on the Multi Vehicle Stereo Event Camera Dataset (MVSEC) (Zhu et al. 2018a). The MVSEC consists of precise depth information recorded form Lidar sensors along with event streams from two event cameras and corresponding intensity images with a resolution of 346×260 pixels. We use the Indoor Flying dataset from the MVSEC and divide them into three split, following (Tulyakov et al. 2019; Zhu, Chen, and Daniilidis 2018). In the split one, we train the model using 3110 samples from the Indoor Flying 2-3 and for the validation and test, we use 200 and 861 samples from the Indoor Flying 1 sequence, respectively. In the split three, we train the model with 2600 samples from the Indoor Flying 1-2 and for the validation and test, we use 200 and 1343 samples from the Indoor Flying 3, respectively. We do not use the split two due to the difference in dynamic characteristics in the training and testing events, as mentioned in (Tulyakov et al. 2019).

We have evaluated our proposed model with the state-of-the-art methods for event-stereo matching. Specifically, we compare our method with the Semi-Dense 3D (Zhou et al. 2018), FCVF* (Hosni et al. 2012; Zhou et al. 2018), SGM* (Hirschmuller 2007; Zhou et al. 2018), TSES (Zhu, Chen, and Daniilidis 2018), CopNet (Piatkowska et al. 2017) and DDES (Tulyakov et al. 2019). We use the same evaluation protocol as (Tulyakov et al. 2019; Zhu, Chen, and Daniilidis 2018) and compute *mean depth error*, *median depth error* and *mean disparity error* for the sparse disparity ground truth and additional *one-pixel accuracy* (1 PA) for the dense disparity ground truth.

The proposed deep event stereo network was implemented using PyTorch. The model was trained in an end-to-end manner with the RMSprop optimizer using default settings. We used the default event queue length and kernel initialization procedure for the event embedding sub-network, following the baseline. A single NVIDIA TITAN XP GPU was used for the training. We trained the model up to 15 epoch and chose the best checkpoint based on the validation results for the testing.

Qualitative Results

Visual results for disparity estimation are shown in Figure 7 for the split 1 and split 3. For the proposed method, like the existing methods, the inference was performed with only event data.

It can be seen from Figure 7 that our proposed model can provide better estimates in edges and structures than the existing methods. In the case of the TSES (Zhu, Chen, and Daniilidis 2018), the events are accumulated via a hand-crafted manner (i.e. stack and sum) which are then processed into event disparity volumes for calculating the matching cost. However, due to the hand-crafted event accumulation and blurring-based volume generation, the TSES fails to preserve event information and hence fails to generate accurate disparity maps. Although the Semi-Dense 3D (Zhou et al. 2018) uses additional information such as known camera motion and works with continuous depth values, it fails to map the disparity in a dense manner due to relying on only temporal coherence among events via the forward-projection approach without any event matching mechanism. The DDES (Tulyakov et al. 2019) works well for the dense disparity estimation due to the learning-based event accumulation with a kernel network and explicit stereo matching framework. However, it fails to calculate plausible disparity maps in the cases of edges and texture-less areas.

Our method explicitly reconstructs intensity images from the event features and uses the reconstructed image features as a guidance for the embedded events. Moreover, our method uses the stacked dilated SPADE to aggregate the reconstructed image features and embedded event features to generate comparatively better dense disparity maps. In our case, due to the additional guidance, our method can predict more plausible structures and refined edges in the dense disparity maps.

Quantitative Results

We have compared our proposed method in two setups - sparse disparity estimation and dense disparity estimation. The performance on disparity estimation has been measured with respective sparse or dense ground truth data.

Table 1 shows the comparison of our method with the DDES in dense disparity map estimation. Note that, among the recent event stereo methods, only the DDES performs dense disparity estimation. It can be seen from the table that our method outperforms the DDES in mean depth error, median depth error, mean disparity error and one-pixel accu-

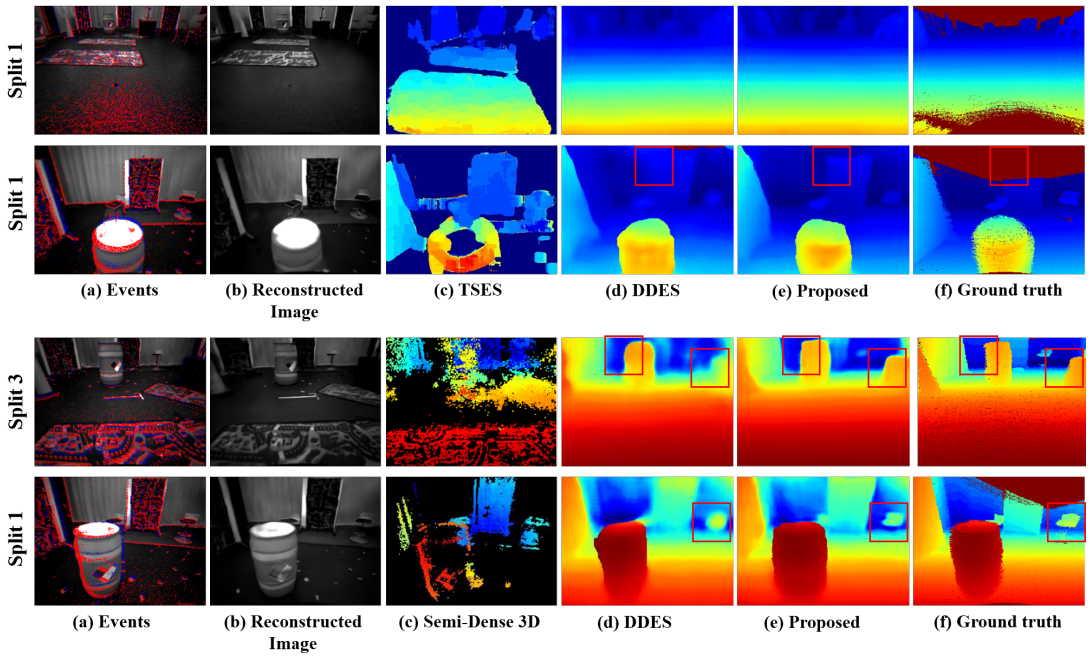

Figure 7: Visual comparison results for event stereo methods. Note that the results for TSES (Zhu, Chen, and Daniilidis 2018), Semi-Dense 3D (Zhou et al. 2018), and DDES (Tulyakov et al. 2019) are borrowed from the original papers. For visual comparison, the similar frames are selected for the proposed method. The first row is frame #100 from split 1, the second row is frame #340 from split 1, the third row is frame #1700 from split 3, and the fourth row is frame #980 from split 1, respectively. In (a), the most recent $15,000$ events from the left camera are overlaid in an intensity image for easier visualization (red color represents positive events and blue color represents negative events).

Methods	Mean depth error [cm] ↓		Median depth error [cm] ↓		Mean disparity error [pix] ↓		One-pixel accuracy [%] ↑	
	Split 1	Split 3	Split 1	Split 3	Split 1	Split 3	Split 1	Split 3
Baseline	16.6	23.5	6.8	14.7	0.59	0.94	89.8	82.5
Proposed	**14.2**	**19.4**	**5.9**	**10.4**	**0.55**	**0.75**	**92.1**	**89.6**

Table 1: Results for dense disparity estimation. Note that the baseline method is DDES (Tulyakov et al. 2019).

racy. This is mainly due to the additional guidance from the reconstructed image features and stacked dilated feature aggregation mechanism which contributes to estimating better disparities on local structures or edges than those of the existing method (as seen from Figure 7).

Moreover, we have evaluated our method in sparse disparity estimation. Table 2 shows the comparison results between our method and recent event stereo methods in sparse disparity estimation. Note that the FCVF* and SGM* methods are the frame-based methods but works on event images, implemented in (Zhou et al. 2018). To generate the event images, the FCVF* and SGM* use temporal event aggregation for feature accumulation, as used in Semi-Dense 3D (Zhou et al. 2018). However, the Semi-Dense 3D relies only on temporal information of the events and fails to preserve spatial information in events which leads to poor results in the FCVF* and SGM*. The CopNet and TSES relies only on spatial information of the events and fails to preserve the temporal information of the events, which is essential for finding event correspondence in time and leads to poor disparity estimation. The Semi-Dense 3D and DDES show a better tendency in disparity estimation due to the inherent ability to work with sparse disparity. However, it can be seen from the table that our method outperforms all the existing methods in terms of mean depth error, median depth error, and mean disparity error by large margins. Our method is capable of handling both sparse and dense disparity scenarios and performs comparatively better than all the existing event stereo methods.

Ablation Studies

We perform the ablation studies on the effectiveness of the proposed sub-networks. Table 3 summarizes the effectiveness of the proposed sub-networks and modules. For the ablation study, we have trained four variants of our proposed model. Model 1 is only trained with the reconstruction sub-network and the feature aggregation sub-network is replaced with a simple concatenation and convolution (i.e. the event features and the image features are concatenated, followed by a convolution). Model 2 is trained with the reconstruction

Methods	Mean depth error [cm] ↓		Median depth error [cm] ↓		Mean disp. error [pix]↓	
	split 1	split 3	split 1	split 3	split 1	split 3
FCVF* (Hosni et al. 2012)	99	103	25.0	11	-	-
SGM* (Hirschmuller 2007)	93	119	31.0	20	-	-
CopNet (Piatkowska et al. 2017)	61	64	-	-	1.03	1.01
TSES (Zhu, Chen, and Daniilidis 2018)	36	36	-	-	0.89	0.88
DDES (Tulyakov et al. 2019)	13.6	18.4	5.9	9.9	0.54	0.69
Semi-Dense 3D (Zhou et al. 2018)	13	33	5.0	11	-	-
Proposed	**11.3**	**15.2**	**4.6**	**7**	**0.49**	**0.63**

Table 2: Results for sparse disparity estimation. Note that the blank entries in the table denote the unavailability of the respective values from the associated papers.

Ablation settings	Recon-struction sub-network	Cross-semantic attention	Spatial context attention	Feature aggregation sub-network	Stacked dilated conv.	Mean depth error [cm]↓	Median depth error [cm]↓	Mean disparity error [pix]↓
Baseline (DDES)	×	×	×	×	×	13.6	5.9	0.54
Model 1	✓	✓	✓	×	×	11.8	5.0	0.51
Model 2	✓	✓	✓	✓	×	11.7	4.9	0.51
Model 3	✓	✓	×	✓	✓	12.4	5.4	0.52
Model 4	✓	×	✓	✓	✓	12.2	5.1	0.53
Proposed	✓	✓	✓	✓	✓	**11.3**	**4.6**	**0.49**

Table 3: Ablation studies of the proposed sub-networks and attention mechanisms.

and feature aggregation sub-network without the stacked dilated convolution. Model 3 is trained with cross-semantic attention only (i.e., without spatial context attention), and model 4 is trained with spatial context attention only (i.e., without cross-semantic attention). In model 3 and 4, the feature aggregation sub-network is kept unchanged.

From Table 3, it can be seen that model 1 shows lower depth and disparity errors by significant margins from the baseline. This is due to the additional structure information obtained from the reconstructed image features. Model 2 shows slightly better performance than model 1 due to the reconstruction and feature aggregation sub-networks and also shows the effectiveness of the dilated convolution in the feature aggregation sub-network. Models 3 and 4 show the effectiveness of the proposed attention mechanisms (i.e., cross-semantic attention and spatial context attention, respectively) in the reconstruction sub-network. However, though each individual proposed sub-network and module contributes to the improvement of the depth and disparity estimation, the best performance is obtained when all the proposed sub-networks and blocks are present. It can be concluded that the image reconstruction with the semantic attention module and feature aggregation with a stacked dilated convolution improves the overall performance on disparity estimation. Moreover, an additional experiment was performed to validate the effectiveness of the proposed reconstruction mechanism. We removed the reconstruction loss and used the reconstruction sub-network only as a feature extractor. The model trained without the reconstruction loss was worse than the proposed method (12.9 vs 11.3 for mean depth error). This result reveals that the explicit reconstruction mechanism is required for better depth estimation.

In addition, we have validated the effect of using a different stereo matching sub-network in our architecture. We replaced the current stereo matching sub-network with a frame-based stereo matching framework, PSMNet (Chang and Chen 2018). We found that the frame-based stereo matching method show worse results than the current method (i.e. 12.2 vs 11.3 for mean depth error). This is due to the frame-based methods generally require comparatively larger datasets and consist of more training parameters, which lead to requiring more training for convergence. The ablation models are trained with the sparse disparity ground truth of split 1 (Indoor Flying 1).

Conclusion

We proposed a novel end-to-end deep event stereo architecture to generate spatial image features from the embedded event data using a novel image reconstruction sub-network and fuse the embedded event features and reconstructed intensity image features with a novel feature aggregation sub-network to perform accurate stereo event matching. For this, we proposed a dual-path encoder-based image reconstruction sub-network with a semantic attention mechanism. In addition, we proposed a novel feature aggregation sub-network based on a stacked dilated convolution-based SPADE module that modulates the event features with the reconstructed image features to be used as guidance for the accurate event stereo matching. We evaluated the proposed method with the state-of-the-art methods in both sparse and dense disparity estimation scenarios. The results show that our method performs comparatively better in both scenarios and improves the existing event stereo methods by significant margins. Ablation studies also validate the effectiveness of the proposed architecture in the event stereo matching.

Acknowledgments

This research was funded in part by the National Research Foundation of Korea (grant no. NRF-2020R1A2C1008753).

References

Chang, J.-R.; and Chen, Y.-S. 2018. Pyramid stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5410–5418.

Chen, H.; Suter, D.; Wu, Q.; and Wang, H. 2020. End-to-end learning of object motion estimation from retinal events for event-based object tracking. In *AAAI*, 10534–10541.

Dauphin, Y. N.; Fan, A.; Auli, M.; and Grangier, D. 2017. Language modeling with gated convolutional networks. In *International Conference on Machine Learning (ICML)*, 933–941.

Dikov, G.; Firouzi, M.; Röhrbein, F.; Conradt, J.; and Richter, C. 2017. Spiking cooperative stereo-matching at 2 ms latency with neuromorphic hardware. In *Conference on Biomimetic and Biohybrid Systems*, 119–137. Springer.

Firouzi, M.; and Conradt, J. 2016. Asynchronous event-based cooperative stereo matching using neuromorphic silicon retinas. *Neural Processing Letters* 43(2): 311–326.

Gallego, G.; Delbruck, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A.; Conradt, J.; Daniilidis, K.; et al. 2019. Event-based vision: A survey. *arXiv preprint arXiv:1904.08405* .

Gehrig, D.; Loquercio, A.; Derpanis, K. G.; and Scaramuzza, D. 2019. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 5633–5643.

Hirschmuller, H. 2007. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 30(2): 328–341.

Hosni, A.; Rhemann, C.; Bleyer, M.; Rother, C.; and Gelautz, M. 2012. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35(2): 504–511.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7132–7141.

Kalia, M.; Navab, N.; and Salcudean, T. 2019. A real-time interactive augmented reality depth estimation technique for surgical robotics. In *2019 International Conference on Robotics and Automation (ICRA)*, 8291–8297.

Kogler, J.; Eibensteiner, F.; Humenberger, M.; Sulzbachner, C.; Gelautz, M.; and Scharinger, J. 2014. Enhancement of sparse silicon retina-based stereo matching using belief propagation and two-stage postfiltering. *Journal of Electronic Imaging* 23(4): 043011.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2337–2346.

Piatkowska, E.; Belbachir, A.; and Gelautz, M. 2013. Asynchronous stereo vision for event-driven dynamic stereo sensor using an adaptive cooperative approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 45–50.

Piatkowska, E.; Kogler, J.; Belbachir, N.; and Gelautz, M. 2017. Improved cooperative stereo matching for dynamic vision sensors with ground truth evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 53–60.

Rebecq, H.; Ranftl, R.; Koltun, V.; and Scaramuzza, D. 2019. High Speed and High Dynamic Range Video with an Event Camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 1–1.

Scheerlinck, C.; Rebecq, H.; Gehrig, D.; Barnes, N.; Mahony, R.; and Scaramuzza, D. 2020. Fast image reconstruction with an event camera. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 156–163.

Schraml, S.; and Belbachir, A. N. 2010. A spatio-temporal clustering method using real-time motion analysis on event-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 57–63.

Schraml, S.; Schön, P.; and Milosevic, N. 2007. Smartcam for real-time stereo visoin - Address-event based embedded system. In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - Volume 2: VISAPP,*, 466–471. INSTICC, SciTePress. ISBN 978-972-8865-74-0. doi:10.5220/0002057604660471.

Schuster, R.; Wasenmuller, O.; Unger, C.; and Stricker, D. 2019. SDC-Stacked dilated convolution: A unified descriptor network for dense matching tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2556–2565.

Sironi, A.; Brambilla, M.; Bourdis, N.; Lagorce, X.; and Benosman, R. 2018. HATS: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1731–1740.

Tulyakov, S.; Fleuret, F.; Kiefel, M.; Gehler, P.; and Hirsch, M. 2019. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1527–1537.

Tulyakov, S.; Ivanov, A.; and Fleuret, F. 2018. Practical deep stereo (pds): Toward applications-friendly deep stereo matching. In *Advances in Neural Information Processing Systems*, 5871–5881.

Uddin, S.; and Jung, Y. J. 2020. Global and local attention-based free-form image inpainting. *Sensors* 20(11): 3204.

VidalMata, R. G.; Banerjee, S.; RichardWebster, B.; Albright, M.; Davalos, P.; McCloskey, S.; Miller, B.; Tambo, A.; Ghosh, S.; Nagesh, S.; et al. 2019. Bridging the gap between computational photography and visual recognition. *arXiv preprint arXiv:1901.09482* .

Wang, L.; Ho, Y.-S.; Yoon, K.-J.; et al. 2019. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10081–10090.

Watkins, Y.; Thresher, A.; Mascarenas, D.; and Kenyon, G. T. 2018. Sparse coding enables the reconstruction of high-fidelity images and video from retinal spike trains. In *Proceedings of the International Conference on Neuromorphic Systems*, 1–5.

Xie, Z.; Chen, S.; and Orchard, G. 2017. Event-based stereo depth estimation using belief propagation. *Frontiers in Neuroscience* 11: 535.

Xie, Z.; Zhang, J.; and Wang, P. 2018. Event-based stereo matching using semiglobal matching. *International Journal of Advanced Robotic Systems* 15(1): 1729881417752759.

Zhou, Y.; Gallego, G.; Rebecq, H.; Kneip, L.; Li, H.; and Scaramuzza, D. 2018. Semi-dense 3D reconstruction with a stereo event camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 235–251.

Zhu, A. Z.; Chen, Y.; and Daniilidis, K. 2018. Realtime time synchronized event-based stereo. In *European Conference on Computer Vision (ECCV)*, 438–452. Springer.

Zhu, A. Z.; Thakur, D.; Özaslan, T.; Pfrommer, B.; Kumar, V.; and Daniilidis, K. 2018a. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. *IEEE Robotics and Automation Letters* 3(3): 2032–2039.

Zhu, A. Z.; Yuan, L.; Chaney, K.; and Daniilidis, K. 2018b. EV-FlowNet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898* .

Zhu, A. Z.; Yuan, L.; Chaney, K.; and Daniilidis, K. 2019. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 989–997.

Zou, D.; Shi, F.; Liu, W.; Li, J.; Wang, Q.; Park, P.-K.; Shi, C.-W.; Roh, Y. J.; and Ryu, H. E. 2017. Robust dense depth map estimation from sparse DVS stereos. In *Proceedings of The British Machine Vision Conference (BMVC)*, volume 1.