STADIE-Net :Stagewise Disparity Estimation from Stereo Event-based Cameras

Abhishek Tomy, Anshul Paigwar, Alessandro Renzaglia, Christian Laugier

***Abstract*— Event-based cameras complement the frame-based cameras in low-light conditions and high dynamic range scenarios that a robot can encounter for scene understanding and navigation. Apart from that, the comparatively cheaper cost in relation to a LiDAR sensor makes this a viable candidate when designing a sensor suite for a robot designed to operate in a dynamic environment. One of the challenges that the sensor suite needs to address is the ability to provide a 3D scene understanding of the environment that would enable the robot to localize obstacles in the environment. This work evaluates the accuracy with which an event-based camera can support this task by providing the disparity estimate between left and right camera frame which can be utilized to calculate the depth of surrounding. A new deep network architecture, named STADIE-Net is proposed that takes advantage of stage-wise refinement and prediction of disparity using events from 2 neuromorphic cameras in a stereo setup. The method utilizes voxel grid representation for events as input and proposes a 4 stage network going from coarse to finer disparity prediction. The model is trained and evaluated on the publicly released DSEC dataset that has data recorded from multiple cities using event-based and frame-based cameras mounted on a moving vehicle. Experimental results show comparable accuracy with baseline method provided for DSEC dataset.**

I. Introduction

A neuromorphic or event-based camera is a bio-inspired sensor that detects the changes in intensity at pixel level and compared to a frame-based camera where light is sampled at fixed time-intervals, event-based camera samples the incoming light continuously. Event cameras rely on contrast Detector (CD) to detect the changes in intensity of incoming light at each pixel and when that exceeds a predefined threshold, an event signal is recorded and transmitted asynchronously. Event-based cameras are particularly efficient in recording changes in the environment and hence well suited for detecting dynamic objects even in low-light conditions, such as nighttime driving or situations involving high dynamic ranges (up to 120dB), and can track motions at high speed and temporal resolution. Progress in the learning-based approaches has meant that a lot of interesting application scenarios have been explored using an event-based vision system. Event-based sensor have been used for detection of objects in dynamic traffic [1], visual odometry [2], optical flow estimation [3] and depth prediction [4].

Depth estimation from stereo event camera opens a lot of applications in the automotive and drone sectors. A sensor suite that can provide the 3D scene understanding and localization of obstacles will help a robot to operate in a dynamically changing environment. Both frame-based cameras and event-cameras individually lack intrinsic information that can be used to estimate depth (there are learning-based methods that can output depth even with a monocular camera [5]), however in a stereo setup when the two cameras are placed close to each other they behave similarly to a human binocular vision system. Given the camera properties such as the focal length and stereo baseline distance, disparity values obtained by matching the features in the left and right image frame can be used to estimate the depth. Learning-based methods and especially the convolutional neural networks (CNN) which have shown to work well with images have made the stereo matching problem trainable in an end-to-end way.

Disparity estimation in frame based cameras is a well explored area and popular approaches have utilized stacked hourglass [6] structure in PSMNet or the stagewise estimation in AnyNet [7]. In this work, we are exploring the stagewise approach as it can output a full disparity at each stage and by restricting to residuals at higher resolutions, a lot of computational time is saved in real-time application.

The method presented in this work builds upon the stage-wise refinement of disparities and the 3D convolutional network in matching the cost tensor between two images from frame-based cameras proposed in [7]. We do our training and evaluation on the DSEC dataset [8] that contains data from event cameras in stereo setup and is recorded from a moving vehicle in multiple cities in Switzerland.

The key contributions of this study can be summarized as follows:

- We present a U-net based deep neural network architecture called STADIE-Net for stereo disparity estimation of event-camera. The network operates only on the information from the event camera and is end-to-end trainable.
- We demonstrate the network's ability to produce disparity at each stage that is closer to ground truth values and the ability of the network to output disparities using lower resolution which reduces the computation time.

II. STADIE-Net

An input from an event camera consists of asynchronous events corresponding to specific pixels that records a change in brightness signal if it exceeds a certain threshold value. An event $e_i = (x_i, y_i, t_i, p_i)$ contains spatio-temporal information in terms of pixel location (x_i, y_i) and the time t_i at which the event is triggered. Polarity of the event ($p_i = \pm 1$) denotes the direction of change of the brightness. Traditional network architectures designed to work with the RGB images

[1] Univ. Grenoble Alpes, Inria, 38000, Grenoble, France; e-mail: firstname.lastname@inria.fr

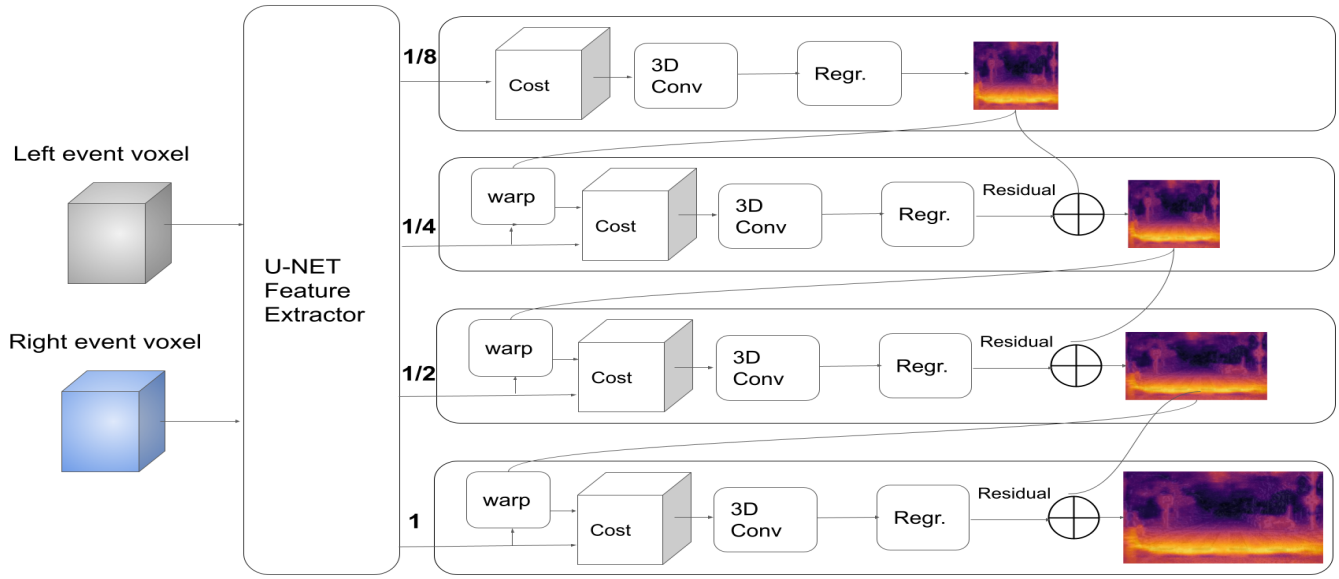

Fig. 1. STADIE-Net: As seen from the architecture, disparity is estimated stagewise going from coarse estimation at lower resolution to subsequent refinement by the addition of residuals in stage 2-4. Features at four resolutions is extracted from the decoder stage of the U-Net which is used as input for respective stages. At each stage a cost volume is constructed by comparing the feature vectors from left and right event cameras. The estimated cost volume is further refined by passing it through a 3D convolutional network.

can be directly reinstated to work with the event data with a few changes including utilization of an appropriate event representation as input.

This section provides a detailed description of our approach starting with the choice of event representation in subsection II-A. Then, in subsection II-B and subsection II-C, the architecture is elaborated in detail by dividing the disparity prediction network into two parts: 1) The U-net feature extractor 2) The construction of disparity cost and disparity estimation stages, which is inspired by [7].

A. Input Event Representation

In this field of work, multiple representations have been used to convert the asynchronous data into a dense representation that can be fed to a deep-neural architecture. Earlier work had focused on converting the events to a greyscale image [9] enabling easy adaptation of the network architecture developed for frame-based methods. Also, some of the previous works have tried to exploit the sparse nature of events by utilizing spiking neural networks [10]. In this work, we have utilized the voxel grid representation proposed in [4]. The events within a time window ΔT are converted into a $B \times H \times W$ voxel grid where H and W are the height and width of the image and B is the number of temporal bins.

$$V(x,y,t) = \sum_i p_i \delta_b(x - x_i)\delta_b(y - y_i)\delta_b(t - t_i^*) \quad (1)$$

where,

$$t_i^* = \frac{B-1}{\Delta T}(t_i - t_0) \quad (2)$$

$$\delta_b(a) = max(0, 1 - |a|) \quad (3)$$

Our models use a time window $\Delta T = 50ms$ and $B = 8$ temporal bins.

B. Feature extractor

The U-net type architecture [11] is utilized as the shared feature extractor for the left and right event voxel grid. The U-net architecture consists of an initial processing layer followed by a 4-layer encoder network and a 2-layer residual network that maintains the same dimensions. This is followed by a 4-layer decoder network through which the feature maps at various resolution (1/8, 1/4, 1/2 and 1) are extracted. Also, the mirrored decoder layers are connected by skip connections from the symmetric encoder stages [5].

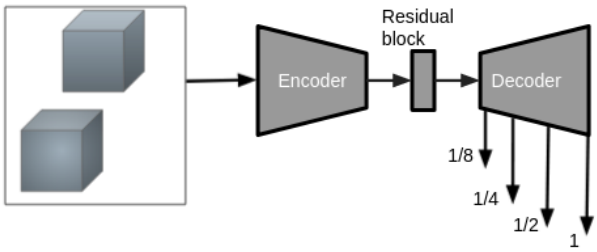

Fig. 2. **UNet architecture:** Consists of 4 encoder stages, 2 residual layers and a symmetric 4 stage of decoder. Feature maps at various resolution are extracted from the decoder layers.

C. Stage-wise Disparity Estimation

The disparity estimation is done in 4 stages by taking the input feature maps from the decoder at various resolutions as shown in Figure 1. In this model, from stage 2-4 only

residuals are added to the initial estimate from stage 1, thus reducing the computational load of calculating the cost over N pixels [7]. In each stage, the input feature map and the output disparity from the previous scale (for stages 2-4) are passed through a disparity estimation network that constructs a 3D cost volume of size $H \times W \times M_s$, where M_s corresponds to the maximum disparity at each scale.

In stage 1, the feature map at $1/8$ resolution is passed through the disparity estimation network which constructs a cost volume by measuring the similarity of pixel (i, j) in left event camera frame with rectified right frame at location $(i, j - k)$. The similarity is measured by the L1 distance between the output feature vector from the decoder. Further, this cost function is refined to remove any artifacts by passing them through a 3D convolutional network which results in a cost for each disparity value from 0 to M in the cost volume denoted by $S_{disp}(i, j, d)$. The disparity of pixels is not the location with the least cost rather it is the weighted average of the cost with the corresponding pixel from the right image as proposed in [12]. This forces the network to output a probability distribution over the possible disparity values instead of a single disparity prediction [13] given by:

$$D(i,j) = \sum_{d=0}^{M} softmax\big(-S_{disp}(u,v,d)\big) \times d \qquad (4)$$

Residuals in stages 2-4 are calculated by first interpolating the coarse estimation from the previous stage and then using that disparity to map the features at the current scale from the left and right event camera frame. The mismatch in the mapping is corrected by adding the residuals to the coarse disparity estimate from previous stages.

III. Training and Experimental results

A. Training Settings

Our model is trained only on the DSEC dataset [8] and is implemented in PyTorch. The model utilizes only the information coming from left and right event cameras. The model is trained to minimize the smooth-L1 loss of the predicted disparity from all the stages jointly. The losses are weighted as following: Stage 1 ($\lambda_1 = 0.25$), Stage 2 ($\lambda_2 = 0.5$), Stage 3 ($\lambda_3 = 0.75$) and Stage 4 ($\lambda_1 = 1$). The model is trained by using the Adam optimizer [14] with an initial learning rate of $3e^{-4}$ and a batch size of 40. The dataset with ground truth disparities is divided into training (90%) and validation set (10%) randomly for evaluation.

B. Experimental Results

The performance of various stages of our method on the validation set is shown in Table III. The disparity estimation goes from coarse to fine as stages increase. It can be seen that from Stage 1 to Stage 3 there is a significant drop between stages, however between stage 3 and 4 the drop is smaller indicating that a further addition of stages would add to the computational complexity without considerable gain in accuracy.

The method was also evaluated by submitting the results to the evaluation server. The performance of our method in comparison to the DSEC baseline is shown in Table I. Our method comes close to the performance of DSEC baseline in terms of mean average pixel error however the baseline performs better in terms of refinement of disparity estimation.

The evaluation is mainly done on the mean absolute error (MAE) between the predicted and ground truth disparities on pixels where the disparity is known. Apart from this, 1PE and 2PE pixel-error is evaluated to show the percentage of ground truth pixels with disparity error greater than 1 and 2 pixels respectively. Both the baseline and our method have shown that the percentage of pixels with an error greater than 2PE is less than 10% indicating a very accurate estimate of disparity using event-camera information.

TABLE I

Comparison of performance of disparity estimation on DSEC dataset

Methods	***MAE***	***1Px***	***2Px***	***RMSE***
DSEC baseline	0.57	10.9	2.9	1.36
Ours	1.01	26.76	9.62	2.02

TABLE II

Comparison of performance of disparity estimation on DSEC dataset on individual sequences

	MAE	***1PE***	***2PE***	***RMSE***
Interlaken				
DSEC baseline	0.57	10.67	3.12	1.36
Ours	0.95	25.13	9	1.92
Thun				
DSEC baseline	0.63	10.85	3.22	1.63
Ours	1.09	26.44	10.13	2.3
Zurich				
DSEC baseline	0.56	11.18	2.59	1.33
Ours	1.04	28.56	10.14	2.06

Our evaluation on the validation set had given us a MAE error of 1.21. Also, in Table III we have shown the error from each stage as it goes from a low-resolution coarse estimate to subsequent refinement. Stage 1 has a mean absolute error of 1.98 where the resolution is $1/8$ of the original image and can provide an estimate of disparity fairly faster. According to application scenarios such as drones where low latency and lesser computation time is a major requirement compared to a highly accurate estimate of the disparity or depth, the models can only use lower resolution input and stage for disparity estimation.

TABLE III

Comparison of MAE error at multiple stages of the network

Stages	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**
MAE	1.98	1.56	1.28	1.21

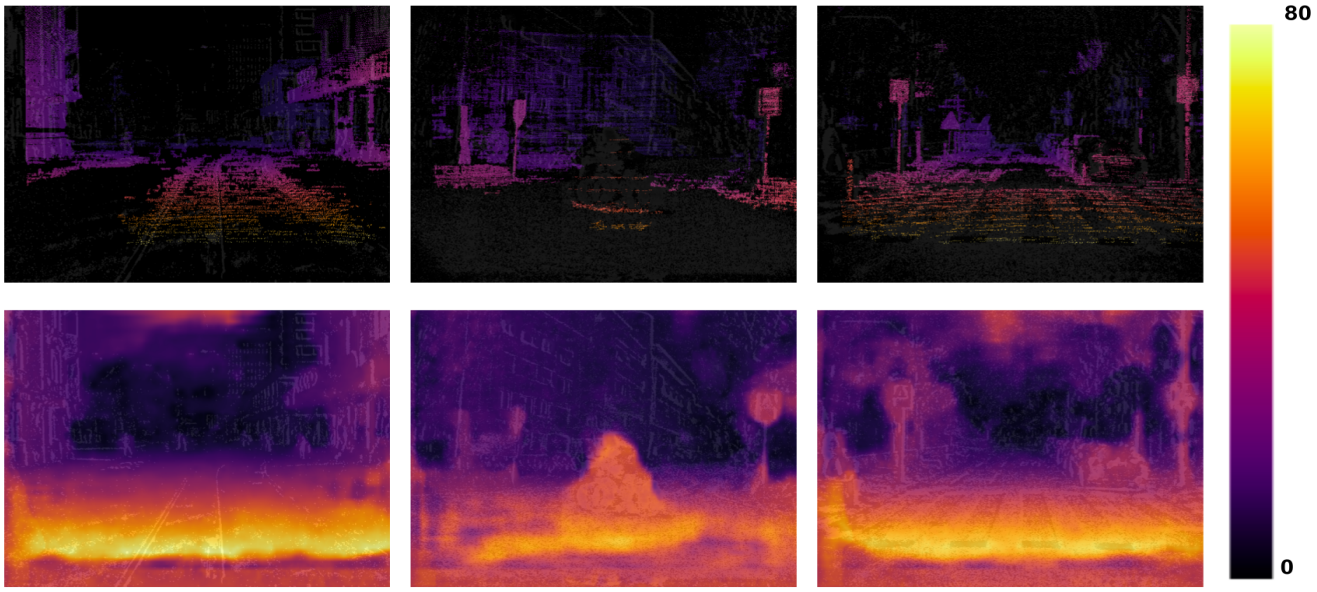

Fig. 3. **Visualization of predicted disparities:** In this figure, 3 of the scenarios and their corresponding ground truth and predicted disparities are plotted on the left event camera frame. In the top row, pixels for which the ground truth is available are shown overlayed on the top of the image. In the bottom row, the prediction from the model is plotted (Model outputs prediction for all the pixels). In the 2nd scenario (middle image), the disparity map shows a sharp discontinuity as a cyclist is observed in front of the vehicle. Also, it is to be noted that pixels in front of the vehicle has a disparity that does not match the depth in all 3 scenarios(this can be attributed to the lack of sufficient training data from this region)

Conclusions and Future work

Our current model would be worked upon and trained further to close the gap from the baseline method. It would be interesting to look at the results after running the model for a larger number of epochs. Also, STADIE-Net uses only the event camera as an input to the network for participating in this benchmark. One of the advantages of the stagewise estimation of disparity is that it can be adjusted to output prediction according to the computational power available at the cost of higher accuracy.

In the future, it would be interesting to explore the replacement of evaluation and creation of a cost volume with a neural network that can output probability distribution for the maximum disparity considered at each stage. This would further improve the speed at which the network can output predictions in a real-world application.

Acknowledgement

This work has been conducted within the scope of ES3CAP (Embedded Smart Safe Secure Computing Autonomous Platform) project.

References

[1] E. Perot, P. de Tournemire, D. Nitti, J. Masci, and A. Sironi, "Learning to detect objects with a 1 megapixel event camera," *arXiv preprint arXiv:2009.13436*, 2020.

[2] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 994–1001, 2018.

[3] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Ev-flownet: Self-supervised optical flow estimation for event-based cameras," *arXiv preprint arXiv:1802.06898*, 2018.

[4] A. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 989–997.

[5] J. Hidalgo-Carrió, D. Gehrig, and D. Scaramuzza, "Learning monocular dense depth from events," *arXiv preprint arXiv:2010.08350*, 2020.

[6] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[7] Y. Wang, Z. Lai, G. Huang, B. H. Wang, L. Van Der Maaten, M. Campbell, and K. Q. Weinberger, "Anytime stereo image depth estimation on mobile devices," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5893–5900.

[8] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "Dsec: A stereo event camera dataset for driving scenarios," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4947–4954, 2021.

[9] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[10] N. Kasabov, K. Dhoble, N. Nuntalid, and G. Indiveri, "Dynamic evolving spiking neural networks for on-line spatio-and spectro-temporal pattern recognition," *Neural Networks*, vol. 41, pp. 188–201, 2013.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[12] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 66–75.

[13] D. Garg, Y. Wang, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao, "Wasserstein distances for stereo disparity estimation," *arXiv preprint arXiv:2007.03085*, 2020.

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