TORE-Based Disparity Estimation In Stereo Event-Only Vision

Ruixu Liu, *Member, IEEE,* R. Wes Baldwin, *Member, IEEE,* Vijayan Asari, *Senior Member, IEEE* and Keigo Hirakawa, *Senior Member, IEEE*

Abstract—We developed an event-based stereo vision system referred to as TSTORE (Temporal-Stereo Time-Ordered Recent Events) Network. With an Adaptive Aggregation Network at its core, we incorporated several novel ideas into this design, including the use of past and present Time-Ordered Recent Event (TORE) volumes, encoding of pixel positional prior, a set of plausible pseudo ground truth disparity maps computed from interpolating lidar data, and a loss function designed to ignore the interpolation errors in pseudo ground truth. We achieve MAE accuracy of 0.57 in the CVPR 2021 competition DSEC dataset.

1 Executive Summary

A system-level diagram of the **TSTORE** Network is shown in Figure 1. The proposed method is summarized below:

- **Network:** We modified a convolutional neural network called Adaptive Aggregation Network. It generates intermediary predicted coarse disparity maps using feature extraction, cost aggregation, and disparity computation. The disparity map is further refined with the help of TORE volumes and averaged ground truth disparity to yield the final high-resolution event-based disparity map (eDM).
- **Event Representation:** Data recorded by the two event cameras is represented using TORE volumes with a depth of $k = 3$. Representations were generated per polarity for each event camera.
- **Temporal TORE Stacks:** We concatenate the present TORE volumes with prior volumes from 100ms and 200ms in the past. Temporal stacking further enhances temporal consistency within the network.
- **Averaged Ground Truth (aGT):** The lidar-based disparity map was averaged across all scenes to generate a positional prior. The positional prior was concatenated with TORE volumes to create the input for the disparity estimation network.
- **Pseudo Ground Truth (pGT):** The network was trained on dense disparity maps obtained by interpolating the sparse lidar measurements. We provide three versions of pGT, yielding a set of *plausible* disparity values at each pixel.
- **Loss Function:** We developed a smooth L1 loss function that penalizes the event-based estimation of the disparity map, but ignores interpolation errors in pGT.
- **Results:** We achieve MAE performance of 0.57.

TSTORE Network implementation did not use any frame image data and relied only on causal event data.

- *R. Liu, R. Baldwin, V. Asari, and K. Hirakawa are with the Department of Electrical and Computer Engineering, University of Dayton, Dayton, OH, US, 45469.*
Email: (liur05,baldwinr2,vasari1,khirakawa1)@udayton.edu.

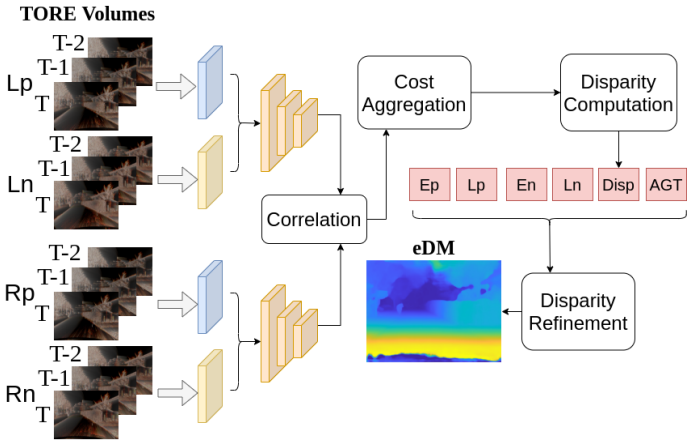


Fig. 1: Temporal-Stereo TORE Network (TSTORE). TORE volumes are generated from current (T) and past (T-1, T-2) timestamps (Lp=left positive, Ln=left negative, Rp=right positive, Rn=right negative). AGT=average ground truth disparity from training dataset. Disp=intermediary predicted coarse disparity. TORE volume errors are computed by the difference of left TORE and the right TORE translated to the left by DISP (Ep=positive TORE error, En=negative TORE error). eDM=event-based disparity map (final output).

2 Network

As shown in Figure 1, TSTORE Network is comprised of several stages. Its initial stage is a pyramid structure feature extraction module used by Guided Aggregation Network (GANet) to increase the global context information [1]. The inputs to this feature extraction module are the positive and negative TORE volumes from each camera (Lp, Ln, Rp, Rn in Figure 1) at current and past times, described in Section 3 below. This is followed by a cost volume aggregation module developed for Adaptive Aggregation Network (AANet) to carry out stereo matching through multi-scale cost aggregation [2]. We adopted the soft argmin mechanism in [3] to obtain the sub-pixel coarse disparity (DISP) in the subsequent disparity computation module. Marked DISP in Figure 1, this intermediary disparity map is 1/3 resolution of the original.

The latter half of TSTORE Network is designed to enhance and improve the quality of DISP by a disparity refinement module developed in [4]. Input to this module is DISP as well as the positive and negative TORE volume error (instead of the photometric error in [5]) by taking the difference of left TORE and the right TORE translated to the left by DISP (Ep and En in Figure 1). Drawing on Dilated Residual Stereo Net (DRSNet) in [5], we added TORE volumes of current stamp belonging to the left camera (Lp and Ln) as an additional input to the disparity refinement module. To this disparity refinement module, we also introduced the notion of positional prior by concatenating averaged ground truth (aGT) disparity map, described in Section 4. The refinement module is implemented as two-stage stack hourglass structure [4], yielding a high-resolution event-based disparity map (eDM) as the final output of TSTORE Network.

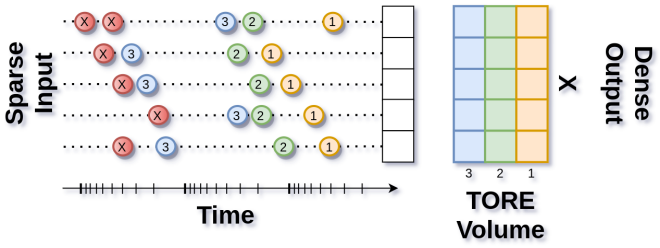


Fig. 2: 2D representation of TORE volume generation. TORE volumes use a FIFO buffer to retain the most recent events at each location. Events beyond the buffer k (shown here as $K = 3$) are forgotten. A convolution on the TORE volume can approximate the output of a single neuron.

3 Event Representation

Sparse events are processed into event representations using TORE volumes [6]. Unlike many event representations that generate tensors with relatively little information, TORE volumes use a FIFO queue per pixel to retain the latest k events per pixel (see Figure 2). FIFO is useful as a priority queue since the last several events tend to be the most important for making decisions about the current state. This queue helps to encode maximum information in each volume and avoids sparsity in the tensor when possible. TORE volume code and results can be found at the project GitHub site: https://github.com/bald6354/tore_volumes.

The fact that TORE encodes an ordered list of past events means that historical activity is encoded into the event representation. By having a history embedded into the representation, it avoids having to generate overly-complex networks that must learn history from fixed-sized temporal windowed representations and enforce temporal consistency. TORE representation is causal, as disparity estimation from a vehicle is a time-bound task requiring minimal latency. In our implementation, each representation channel was spatially rectified with the provided rectification maps using a nearest neighbor interpolation.

TORE volumes were generated per camera at each GT timestep (10Hz) with a depth of $k = 3$. Since TORE volumes treat positive and negative polarities separately, this yielded a total of 12 channels from event data (sensors×polarity×TORE depth=12), labeled Lp, Ln, Rp, and Rn in Figure 1. Furthermore, we concatenate this 12-channel TORE volume at the current timestamp to past TORE volumes (100ms and 200ms prior). This helps to further strengthen temporal consistency (sensors×polarity×TORE depth×(current+past TORE)=36 channels). The resulting 36 dimensional event representations were used as an input to the pyramid-structure, feature-extraction module.

TORE volumes are also used in the disparity refinement module. The differences between Lp (or Ln) and Rp (or Rn) TORE volumes translated towards left by DISP is recorded as TORE volume error Ep (or En). Ep and En along with Lp and Ln of current timestamp are used as inputs to the disparity refinement module.

4 Averaged Ground Truth Disparity Map

Consider an average ground truth (aGT) disparity map based on the provided lidar data, as shown in Figure 3(b). Note the basic trends captured in this single image, such small disparity near the center of the image (road far away) and large disparity at the bottom of the image (road closer to car). It also captures attributes specific to the camera and vehicle configurations, such as disparity that is specific to the fixed baseline space between the two cameras, or the trend that pixels on the right half of the image have larger disparity distances than the left half of the image (because the car is driving on the right side of the road). One may regard aGT disparity map as a type of pixel positional *prior*.

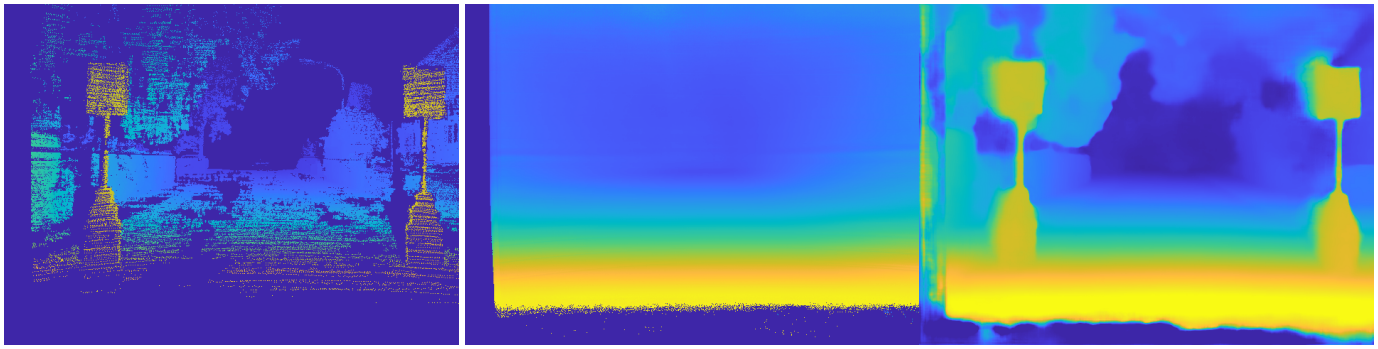
The disparity refinement module in Figure 1 and Section 2 above is implemented as two-stage U-Net [7] architecture. U-Net is comprised of the usual components such as convolution, activation, pooling, concatenation, and interpolation steps. Unlike fully connected layers, these operations are spatially invariant—in the sense that same operations are applied to every pixel in the same exact way. For this reason, U-Net has no built in mechanism to constrain or regularize the output disparity map based on *absolute* pixel coordinates, such as the positional attributes captured by aGT.

Thus, we improve the disparity refinement module by providing it with aGT image as an input in addition to the intermediary disparity (DISP), left TORE volumes (Lp, Ln) of current timestamp, and TORE volume error (Ep, En). aGT is a static input image, and we posit that it would promote extraction of features from TORE volumes that provide information above and beyond what is already encoded within aGT.

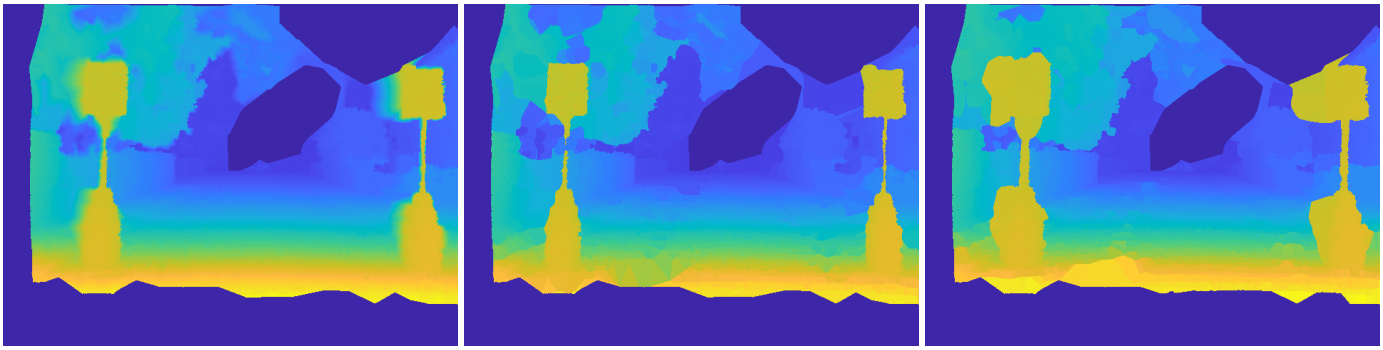
5 Loss Function with Plausible Pseudo Ground Truths

Lidar provides a sparse depth measurements based on return time of the laser scans. As evidenced by Figure 3(a), the "ground truth" (GT) disparity map generated using lidar data have missing values at pixels that the laser scan did not occur, or in parts of the scene where the laser was occluded. The lack of disparity values at majority of these pixels make the training process slower and less stable.

We draw inspiration from prior work [2], where dense pseudo ground truth (pGT) was shown to improve network accuracy and reduce training times. Unlike the prior work that relied on a combination of lidar, RGB images, and a pretrained disparity-estimation network, we created a series of plausible pGT using the lidar-only sparse GT as a source. Interpolation in smooth regions of the scene (e.g. road) can be especially beneficial and dense pGT can help normalize loss across the entire image. Instead of using



(a) Ground Truth (b) Averaged Ground Truth (aGT) (c) TSTORE Network Output (proposed)



(d) Linear Pseudo Ground Truth (pGT_lin) (e) Minimum Pseudo Ground Truth (pGT_min) (f) Maximum Pseudo Ground Truth (pGT_max)

Fig. 3: Ground truth v.s. averaged ground truth v.s. TSTORE Network output v.s. pseudo ground truth generated by linear, min, and max interpolation. We regard aGT as a pixel positional prior (sky above, pavement close towards bottom of the frame, etc.). Collectively, {pGT_lin, pGT_min, pGT_max} comprise a set of plausible disparity values. Each pGT has about three times the number of valid disparity values which helps improve network accuracy as well as normalize for non-uniform lidar sampling.

a single interpolated pGT, we propose to generate multiple "plausible" pseudo ground truth disparity maps. Specifically, missing disparity values at sharp depth boundaries stemming from laser occlusions (such as the one shown in Figure 4) are difficult to interpolate precisely. Event data is present at most of these depth transition regions, however, resulting in training loss function that inadvertently penalize pGT interpolation error (instead of the inaccuracy in event-based disparity estimation).

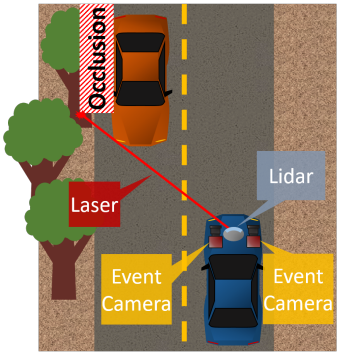
We addressed this issue by generating *three* pGT disparity maps that describe plausible depths in the scene, as follows:

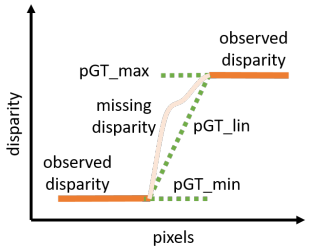
- pGT_lin uses Delaunay triangle-based linear interpolation.
- pGT_min uses Delaunay triangle-based minimum filter.
- pGT_max uses Delaunay triangle-based maximum filter.

Examples are shown in Figure 3. Intuitively, {pGT_lin, pGT_min, pGT_max} yield a set of *reasonable* disparity values consistent with the measured lidar samples, in the following sense. We expect that the missing disparity value would be bounded by pGT_min and pGT_max representing the disparity range of the nearby scene points. The linear interpolation would be accurate when we expect a smooth depth transition. At the onset, however, it is difficult to know *a priori* which of the three plausible scenarios is closest to the truth.

Thus, we designed a novel loss function that would compare the event-based CNN generated disparity map (eDM) to pGT_lin, pGT_min, and pGT_max, as follows:

$$\text{Loss} = \min\{\|\text{pGT\_lin} - \text{eDM}\|_1, \|\text{pGT\_min} - \text{eDM}\|_1, \|\text{pGT\_max} - \text{eDM}\|_1\}, \quad (1)$$





(a) Lidar with occlusion (b) Disparity with missing values

Fig. 4: (a) Lidar samples are missing in areas of the scene where the laser is occluded. (b) An example of disparity map (cross section) and the three interpolation algorithms used to fill the missing samples.

where $\|\cdot\|_1$ refers to the L1 norm. In other words, the loss function above is designed to penalize eDM values only when *none* of the three plausible scenarios are consistent with eDM. The flexibility afforded by the loss function above ensures that the pGT inaccuracies would not increase the training loss.

In practice, there are large regions of the scene where no lidar measurements are present (e.g. sky). Since pGT_lin, pGT_min, and pGT_max will all fail to yield meaningful disparity in these re-

	validation MAE
Positive TORE volumes only	0.77
+ Both positive and negative TORE volumes	0.60
+ Pseudo ground truth	0.57
+ Average disparity map	0.54
+ Current and past TORE volumes	0.52

TABLE 1: Performance increase by introducing different components of TSTORE Network. Tested on 5-fold cross-validation on the DSEC training dataset [10].

gions, we mask out interpolated disparity values at pixel locations further than ten pixels away from any valid lidar measurements (i.e. does not figure into the loss function).

6 Training and Evaluation for Competition

6.1 Network Training Procedure

We implemented the TSTORE Network in PyTorch using the Adam (β_1 = 0.9, β_2 = 0.999) optimizer. Training took place on 4× NVIDIA TITAN RTX GPUs. The network was initially trained using only "current TORE volumes." We began with pre-trained GANet and AANet models (replicating the RGB weights to each of TORE volumes of depth 3) obtained using Scene Flow [8] and KITTI2015 [9] datasets. This network was trained for 60 epochs with batch size 32 and a starting learning rate of 0.001 that was decreased by half every 10 epochs using the provided DSEC dataset [10]. Once trained, "past TORE volumes" were added to TSTORE Network, and it was fine-tuned from the "current TORE volumes"-only network. The initial learning rate for the fine-tuning is 0.0001 and decreased by half at the 15th, 25th, 35th and 40th epochs until we reached 50 additional epochs.

The training was performed using a set of three plausible pseudo ground truth (pGT) disparity maps described in Section 5. The loss function compared the estimated disparity map to all of the pseudo ground truths using Eq. (1). In order to take maximum advantage of any learnable pixel position-specific attributes of the DSEC dataset, we did not use any data augmentation methods. We used the 5-fold cross-validation on the training dataset to evaluate our TSTORE model. DSEC testing dataset was never used to train or fine-tune the network.

6.2 Results

We studied the effectiveness of various parts within the TSTORE Network. In Table 1, we show a comparison of the model trained by different configurations, tested using the 5-fold cross-validation on the DSEC training dataset [10]. We report the performance of TSTORE Network in terms of mean absolute error (MAE) between the ground truth disparity map computed from lidar and the event-based disparity map (eDM) yielded by the TSTORE Network.

7 Conclusion

We presented TSTORE (Temporally-Stereo Time-Ordered Recent Events) Network. We generate event representations from the stereo cameras using TORE volumes for current and past timestamps. These features are subsequently used to perform stereo matching via cost aggregation, and the resultant disparity map is further refined using TORE volume error and an average ground truth disparity map. The network was trained using pseudo ground truth disparity maps, and tested on training sets of DSEC dataset. Our tests showed that TSTORE Network yields reliable and high quality disparity maps.

Acknowledgments

This work was made possible in part by funding from Ford University Research Program.

References

[1] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 185–194.

[2] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1959–1968.

[3] 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.

[4] 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.

[5] R. Chabra, J. Straub, C. Sweeney, R. Newcombe, and H. Fuchs, "Stereodrnet: Dilated residual stereonet," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 786–11 795.

[6] R. Baldwin, R. Liu, M. Almatrafi, V. Asari, and K. Hirakawa, "Time-ordered recent event (tore) volumes for event cameras," *arXiv preprint arXiv:2103.06108*, 2021.

[7] 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.

[8] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048.

[9] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3061–3070.

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