

LiDAR-Event Stereo Fusion with Hallucinations

Supplementary Material

This document provides additional details regarding ECCV 2024 paper “LiDAR-Event Stereo Fusion with Hallucinations”. Specifically, we report:

- **page 1:** the impact of our framework on event streams distinctiveness
- **pages 2-5:** the composition of the datasets split used in our experiments, as well as the pre-processing pipeline necessary to obtain data suitable for our purposes from both DSEC [7] and M3ED [5]
- **pages 6-7:** detailed description of the LiDAR-stereo fusion strategies inherited from classical deep stereo literature [6, 13, 16]
- **page 8:** additional ablation studies
- **pages 9-12:** more qualitative results on the M3ED dataset [5]

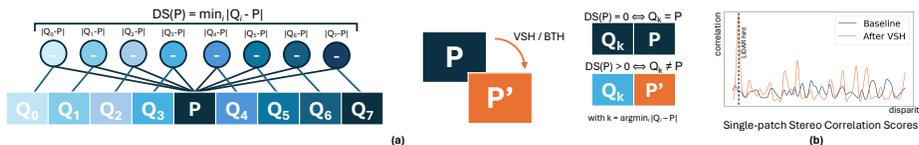


Fig. I: Distinctiveness before/after VSH/BTH (a), correlation scores (b).

1 Impact on Distinctiveness

In this section, we discuss the reasons behind the effectiveness of our hallucination strategies from a probabilistic perspective. In the stereo matching tasks, ambiguities arise when the patches to match are not distinctive – i.e., multiple patches along the horizontal epipolar line are identical. For a single patch P and its neighbors Q_i , we can define its **distinctiveness** [10] $DS(P)$ as $\min_i |P - Q_i|$. Such distinctiveness is 0 if at least one Q_i is identical to P – see Fig. I (a). This is very likely to occur for patches for which no event at all is triggered by the camera, as for the $\approx 30\%$ of the patches on the M3ED dataset.

As VSH puts a random P' on both left and right frames, it will be sufficient for P' to have $DS(P') > 0$ to ease the matching of those empty patches. A specific P' is generated with probability $p(P') = \frac{1}{V^{N^2 \times B}}$, being V the possible per-pixel values (e.g., 2^8 for uint8), with patch size $N \times N$ and B stack channels. Accordingly, the probability $p(DS(P') = 0)$ equals the probability of having another patch on the horizontal scanline identical to P' , i.e., $W \times p(P')$ if we assume patches to be independent (in the worst case scenario), being W the image width. On M3ED, $p(DS(P') = 0) = 1280 \times p(P)$, that is $\approx 5e^{-41}$ and $\approx e^{-257}$ for Histogram and Voxel Grid respectively. This means VSH has probability ≈ 1 to ease matching for empty patches if a LiDAR hint is available for them. We confirm this expected behavior in Fig. I (b), showing the correlation curve computed by the stereo network on an empty patch, peaked on the LiDAR value after VSH. Similar derivation can be done for BTH.

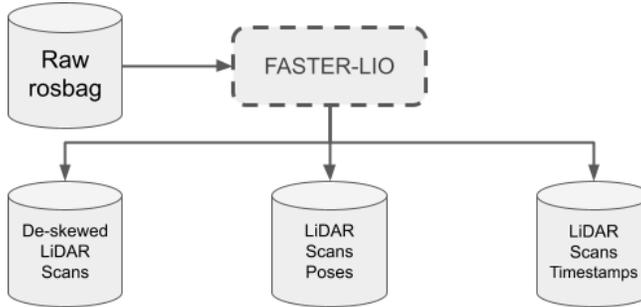


Fig. II: DSEC [7] processing scheme (1).

2 Evaluation Datasets and Pre-processing

In this section, we provide further details concerning the datasets used in our work and the pre-processing we carried out. In particular, we describe the DSEC [7] search split we used for the hyperparameters search concerning our proposals VSH and BTH, the M3ED [5] evaluation split we selected, and how we managed to process both datasets to extract raw LiDAR and, on M3ED dataset [5], for obtaining misaligned LiDAR measurements with respect to the timestamp at which we estimate disparity maps – and thus, at which we have ground-truth depth for evaluation.

2.1 DSEC [7] Dataset

We start with DSEC [7], which we use for i) tuning the hyper-parameters in our solutions, and ii) training the models involved in our experiments.

Search split. We select three sequences from the training set: *zurich_city_00_b*, *interlaken_00_c* and *zurich_city_09_c*.

Processing Scheme. We managed to extract raw LiDAR disparity maps directly from the rosbag files provided by the authors. Our extraction pipeline is detailed in Figs. II and III. Differently from M3ED [5], this dataset does not provide any ground-truth pose. Consequently, as the first step detailed in Fig. II, we deployed a LiDAR inertial odometry framework – FASTER-LIO [2] – to obtain point clouds without rolling shutter effects (*de-skewed*), their corresponding poses and timestamps.

Assuming a shared clock between ground-truth and raw timestamps, we aim to synchronize each ground-truth disparity map at time t_d with the nearest de-skewed LiDAR scan \mathbf{M}_d . As shown in Fig. III, we can achieve this goal using a mapping function that links raw with ground-truth data using timestamps. Next, we can use the estimated poses to align \mathbf{M}_d to its correlated ground-truth frame, using linear interpolation. Unfortunately, we empirically noticed that this step is insufficient to guarantee an acceptable alignment. To further refine this alignment, we deployed point-to-plane ICP algorithm [14] with L1 robust kernel [1]

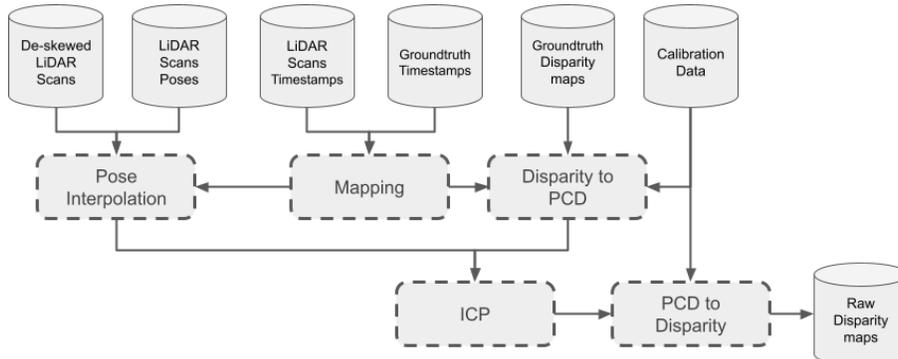


Fig. III: DSEC [7] processing scheme (2).

between \mathbf{M}_d and the ground-truth point cloud (PCD) obtained by reprojection of the ground-truth disparity map to 3D. To ensure the best alignment possible, we repeated the process described in Fig. III also using the previous and the next LiDAR scans $\mathbf{M}_{d-1}, \mathbf{M}_{d+1}$ and skipping the pose interpolation sub-step. As a result, for each ground-truth frame we have six possible candidates – *i.e.*, $\{(\mathbf{M}_{d-1}, \mathbf{I}), (\mathbf{M}_{d-1}, \mathbf{P}_{d-1}), (\mathbf{M}_d, \mathbf{I}), (\mathbf{M}_d, \mathbf{P}_d), (\mathbf{M}_{d+1}, \mathbf{I}), (\mathbf{M}_{d+1}, \mathbf{P}_{d+1})\}$, where \mathbf{I} (identity rigid transformation) and $\mathbf{P}_{d-1}, \mathbf{P}_d, \mathbf{P}_{d+1}$ (interpolated rigid transformation respectively for ground-truth frame $d-1, d, d+1$) represent the initial state for ICP algorithm. We select the one that best reduces the mean reprojection error (MAE). As the last step, raw point clouds are roto-translated into the left rectified event camera point of view and then projected into the rectified image plane to obtain depth maps, that are converted into disparity maps. Finally, we measure the MAE between each raw LiDAR depth map and its corresponding ground-truth depth map and discard those with an error exceeding 0.5 meters from training/search/testing splits.

2.2 M3ED [5] Dataset.

In this second dataset, we evaluate the generalization performance by any of the models involved in our evaluation, as well as we measure the robustness of LiDAR-event stereo frameworks when processing misaligned LiDAR data.

Evaluation Split. We select both some outdoor (*car_forest_tree_tunnel* and *car_urban_day_penno_small_loop*, with the former featuring a fully-static set of frames) and indoor (*falcon_indoor_flight_1*, *falcon_indoor_flight_2* and *falcon_indoor_flight_3*) sequences.

Processing Scheme. Figs. IV to VI sketch the main stage of our pre-processing pipeline. For each M3ED sequence, the authors provide *data* file and *depth_gt* file. The former provides raw LiDAR measurements, left and right distorted, unrectified stereo events, calibration parameters for each sensor, and other data not meaningful for our scope. The latter store ground-truth depth maps aligned to the left event camera and ground-truth poses for LiDAR (*i.e.*,

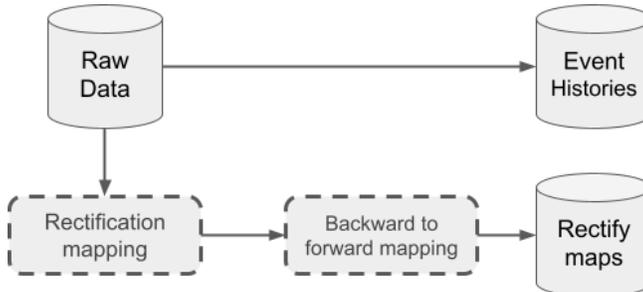


Fig. IV: M3ED [5] processing scheme (1). Extraction of stereo histories and rectification maps from the *data* file.

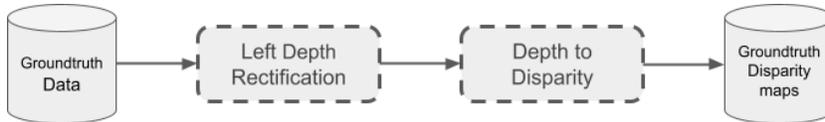


Fig. V: M3ED [5] processing scheme (2). After rectification, ground-truth depth maps are converted to disparity maps.

Ln_T_L0) and left event camera (*i.e.*, Cn_T_C0). This eases the LiDAR to ground-truth alignment with respect to DSEC dataset. All sensors, ground-truth depths, and poses are time-synchronized using a global clock. For more details, please refer to the original paper [5].

We extract the stereo event histories directly from the *data* file, as displayed in Fig. IV. Undistortion and rectification are managed at run-time using a forward mapping, stored in a single *rectify_map* file for each event camera. To create those mappings, we fed intrinsic, extrinsic, and distortion coefficients to the OpenCV functions. Furthermore, since OpenCV returns backward mapping functions, we used an iterative strategy to obtain left and right forward mappings. We maintain intermediate and final products of this step in memory, as they will be required by subsequent steps.

After the previous step, we process raw ground-truth depth maps from the *depth_gt* file (Fig. V). We rectify depth maps using the same left-forward mapping obtained in the previous step. For each depth map, we assume $t_d = t_z - i.e.$, the timestamp t_d at which we want to estimate disparity is the same as the timestamp t_z at which the depth map is captured. Finally, depth maps are converted into disparity maps using stereo camera parameters obtained previously from stereo calibration.

The final step consists of processing raw LiDAR scans from the *data* file (Fig. VI). For each LiDAR scan, we linearly interpolate the pose at which the scan starts and the pose at which the scan ends as we need them to obtain a point cloud without rolling shutter effects (*de-skewed*). We aim to acquire both in-sync LiDAR disparity maps (*i.e.*, $t_d = t_z, \Delta t_o = 0$) and out-of-sync (or misaligned)

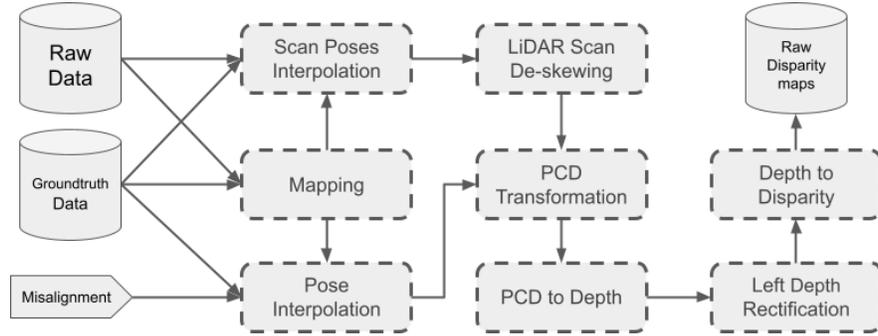


Fig. VI: M3ED [5] processing scheme (3). We used interpolated poses to correct raw LiDAR point clouds affected by rolling shutter effects. Then point clouds are roto-translated, projected into image plane, rectified and converted into disparity maps.

LiDAR disparity maps – *i.e.*, $t'_z = t_d - \Delta t_o$, $\Delta t_o > 0$, where Δt_o is the temporal misalignment chosen arbitrarily (*e.g.*, 3, 13, 32, 61 and 100 ms) – these latter are used for experiments in Section 5.6. Given \mathbf{M}_d as the closest de-skewed LiDAR scan to t_d (mapping step in Fig. VI) and the temporal misalignment Δt_o , the former goal is achieved by roto-translating \mathbf{M}_d as if it had been captured at $t'_z = t_d - \Delta t_o$. After that, we linearly interpolate the pose at time t'_z using the two temporally nearest LiDAR poses (*i.e.*, $L_n_T_L0$). Next, point clouds are roto-translated into the left event camera point of view, and then projected into the image plane to obtain depth maps. Finally, depth maps are rectified and converted into disparity maps. Before the evaluation, we calculate the MAE between each raw LiDAR depth map and its corresponding ground-truth depth map and discard them if the error exceeds 0.1 meters, as well as discard the first and last 50 frames where raw LiDAR data fully overlaps with ground-truth.

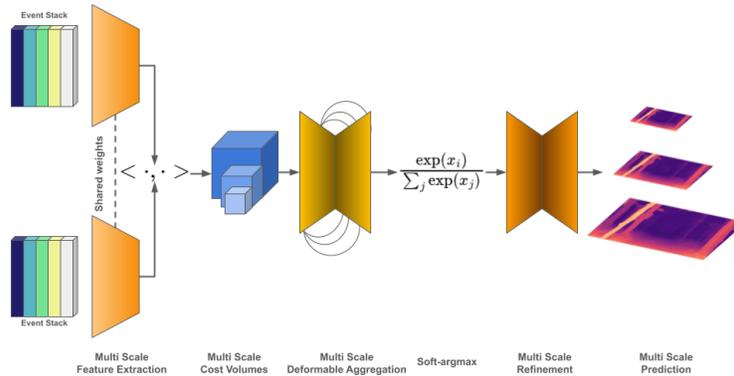


Fig. VII: SE-CFF architecture.

3 LiDAR-Stereo Matching Fusion Architectures

We report additional details about the deep architectures used in our experiments, starting from the baseline network, SE-CFF [12], and then showing the fusion strategies ported from classical deep stereo literature [6, 13, 15]

Baseline Model: SE-CFF. Fig. VII provides an overview of the SE-CFF [12] architecture assumed as the baseline in our work: i) rectified event streams are organized into MDES representations and then processed by a concentration network, producing single-channel stacks; ii) these are processed by two features extractor with shared weights, producing outputs respectively at $\frac{1}{3}$, $\frac{1}{6}$, $\frac{1}{12}$ of the original resolution; iii) these are used to build a multi-scale cost-volume computing the correlation between left and right features along the epipolar line; iv) a multi-scale, 2D network made with deformable convolutions is used to refine the cost volumes; v) an initial disparity map is obtained through a soft-argmax operator; vi) a multi-scale refinement network produces a set of refined disparity maps. We deploy eight variants of this architecture, one for each stacked representation considered in our experiments – i.e., stage i) is replaced with the different representations.

Guided Stereo [13]. The first among the strategies inherited from classical deep stereo literature is depicted in Fig. VIII. Specifically, LiDAR data are used to modulate any of the multi-scale cost volumes \mathcal{F} , according to a Gaussian function

$$\mathcal{G} = \left(1 - v_{ij} + v_{ij} \cdot k \cdot e^{-\frac{(d-g_{ij})^2}{2c^2}} \right) \cdot \mathcal{F} \quad (1)$$

with k, c being the height and width of the Gaussian. Following [13], LiDAR points are downsampled through nearest-neighbor interpolation to act at the different resolutions.

Concat [16]. An alternative strategy consists of providing the raw LiDAR data as an input to the stereo model. Fig. IX gives an overview of this lat-

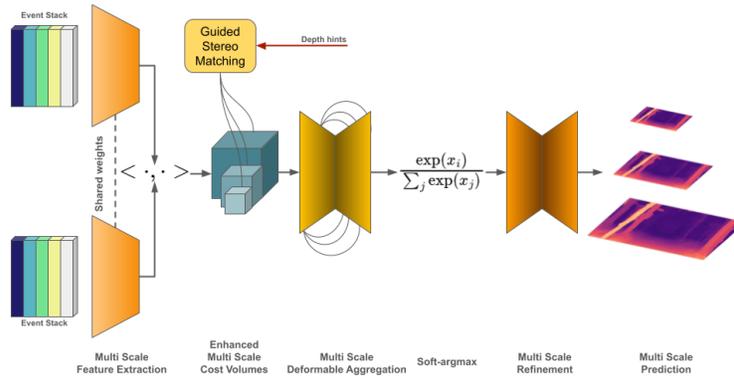


Fig. VIII: Guided Stereo Matching [13] Framework.

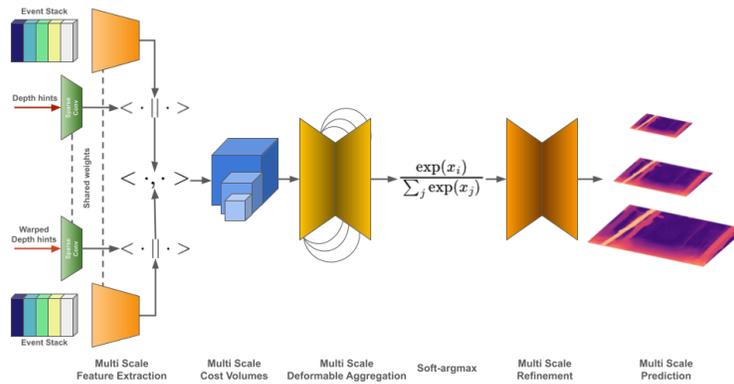


Fig. IX: Concat [6] architecture.

ter approach. LiDAR data are projected on left and right camera frames and processed by two feature extractors with shared-weights and made of Sparsity-Invariant Convolutions. These features are concatenated to those extracted from images – or stacked events, in our case – before the cost volumes are computed.

Guided+Concat [6]. This strategy is inspired by CCVNorm architecture [6] and re-adapted to be deployed with SE-CFF. Specifically, CCVNorm deploys LiDAR data both as the input to the model, as well as to modulate the cost-volume at different stages during inference. Fig. X shows how we tailor this strategy to SE-CFF, basically by combining the two previous strategies.

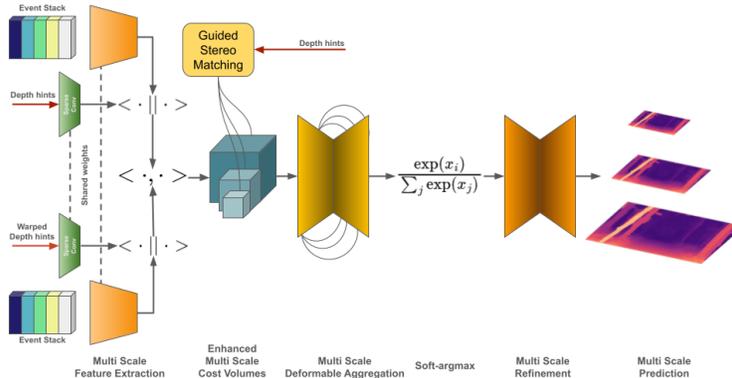


Fig. X: Guided+Concat [15] architecture.

Table I: Ablation on VSH (top) and BTH (bottom). We report 1PE for stereo backbones with different stacked representations on the DSEC search split.

Event Representations							
Configuration	Histogram [11]	MDES [12]	Voxel Grid [16]	TORÉ [3]	T. Surface [9]	ERGO-12 [17]	Tencode [8]
Baseline	13.33	12.07	12.75	12.23	11.69	11.58	11.15
(A) VSH	12.29	10.33	11.71	10.49	10.20	10.07	9.84
(B) (A)+Occlusion handling [4]	12.28	10.33	11.72	10.50	10.19	10.06	9.83
(C) (B)+Splatting [4]	12.00	10.33	11.70	10.51	10.14	10.04	9.82

Event Representations							
Configuration	Histogram [11]	MDES [12]	Voxel Grid [16]	TORÉ [3]	T. Surface [9]	ERGO-12 [17]	Tencode [8]
Baseline	13.33	12.07	12.75	12.23	11.69	11.58	11.15
(A) BTH	10.99	10.53	10.52	10.34	10.13	10.09	9.78
(B) (A)+Occ. disc. [4]	10.99	10.52	10.53	10.33	10.12	10.08	9.77
(C) (B)+Splatting	11.00	10.54	10.53	10.34	10.13	10.08	9.77

4 Ablation Study – Additional Experiments

We complement the ablation studies already shown in the main paper. Specifically, Tab. I shows the results achieved by VSH and BTH on the search split, respectively on top and bottom.

Starting from VSH, we show how handling occlusions according to [4] (B) allows for slightly improving the results, with further improvements achieved by applying sub-pixel splatting [4] (C). When it comes to BTH, we observed empirically that discharging occlusions yields the best results (B). We also implemented a revised version of pixel splatting applied to the event streams (C), yet without noticeable improvements.

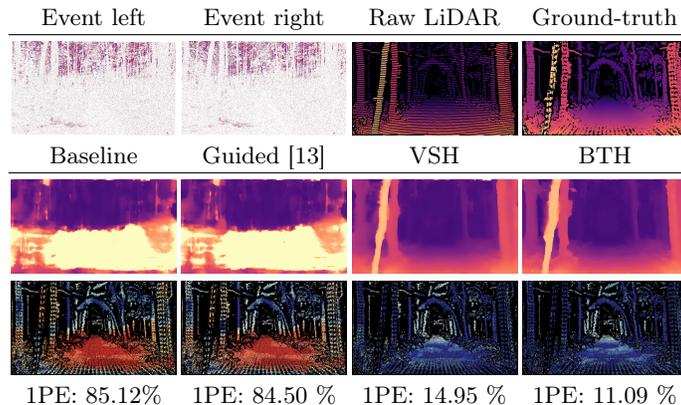


Fig. XI: Qualitative results – baseline stereo backbone. Results on *car_forest_tree_tunnel*, with MDES [12] representation.

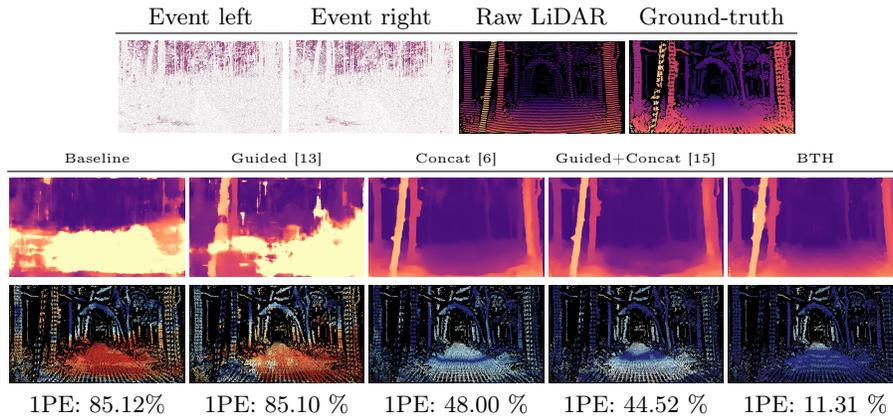


Fig. XII: Qualitative results – re-trained stereo backbone. Results on *car_forest_tree_tunnel*, with MDES [12] representation.

5 Qualitative Results

In conclusion, we present some qualitative results to support the efficacy of our proposal. We display eight (*i.e.*, Figs. XI to XVIII) different figures from the M3ED [5] dataset, using raw LiDAR measurements as guidance for all fusion frameworks.

Figs. XI and XII show an example from the *car_forest_tree_tunnel* sequence, respectively spotlighting fusion strategies applied without re-training the stereo backbone (the former) or when the network is trained from scratch to perform fusion (the latter), processing MDES [12] representations in both cases. In the former case, Guided [13] is nearly ineffective, whereas both VSH and BTH largely improve the results. In the latter case, Concat [6] and Guided+Concat [15] can reduce the error by about 40%, yet far behind the improvement yielded by BTH (more than 70% error rate reduction).

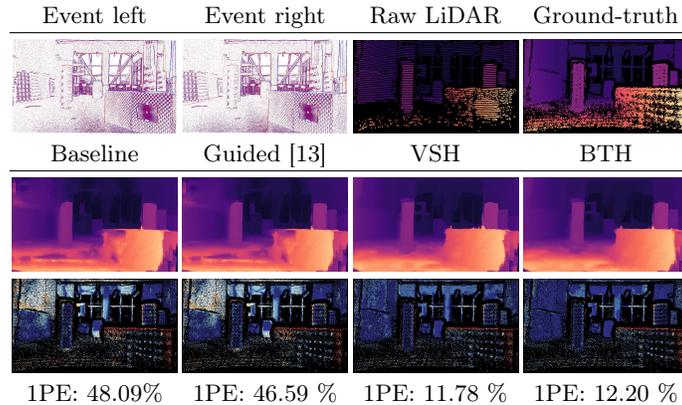


Fig. XIII: Qualitative results – baseline stereo backbone. Results on *spot_indoor_obstacles*, with Voxel Grid [16] representation.

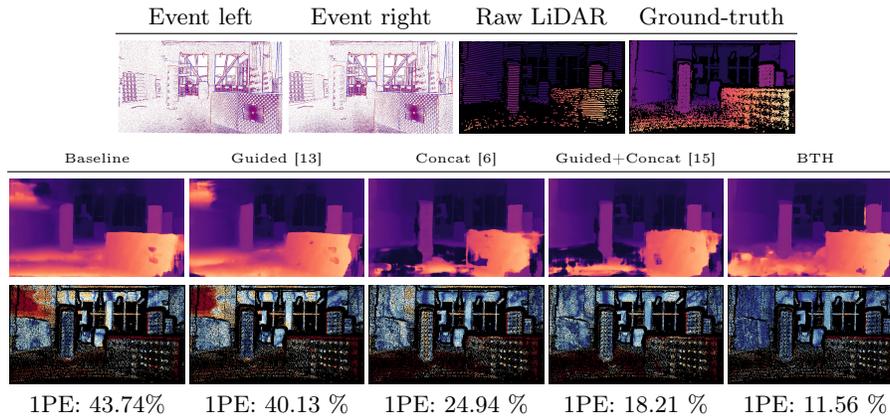


Fig. XIV: Qualitative results – re-trained stereo backbone. Results on *spot_indoor_obstacles*, with ERGO-12 [17] representation.

Figs. XIII and XIV show an example from an indoor sequence – *i.e.*, *spot_indoor_obstacles* – again when not re-training or training the stereo backbone from scratch. In the former case, we report results by processing Voxel Grid representations [16], which confirm the trend observed in the previous example. Indeed, our proposal confirms again the best solution for exploiting raw LiDAR measurements and improve the accuracy of event-based stereo networks.

Figs. XV and XVI showcase another example from outdoor, with a person being framed by the event cameras – *i.e.*, *spot_outdoor_day_skatepark_1*. In this case, we report results obtained by processing TORE [3] representations when re-training the stereo backbones, confirming again how our proposal consistently yields the largest drop of the error rate independently of the chosen representation.

To conclude, we evaluate the effectiveness of different fusion strategies against partially filled raw LiDAR depth maps both in indoor (*i.e.*, *falcon_indoor_* -

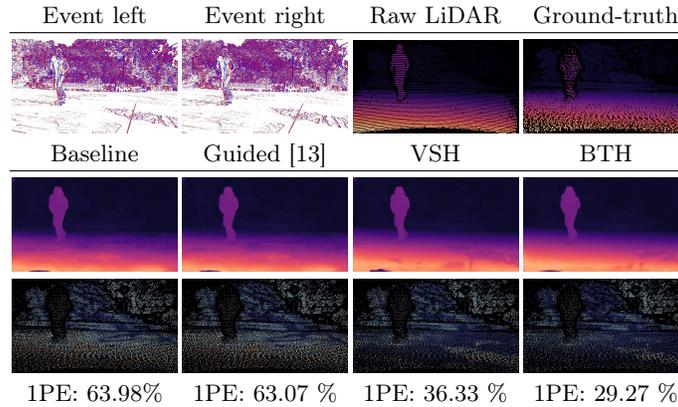


Fig. XV: Qualitative results – baseline stereo backbone. Results on *spot_outdoor_day_skatepark_1*, with MDES [12] representation.

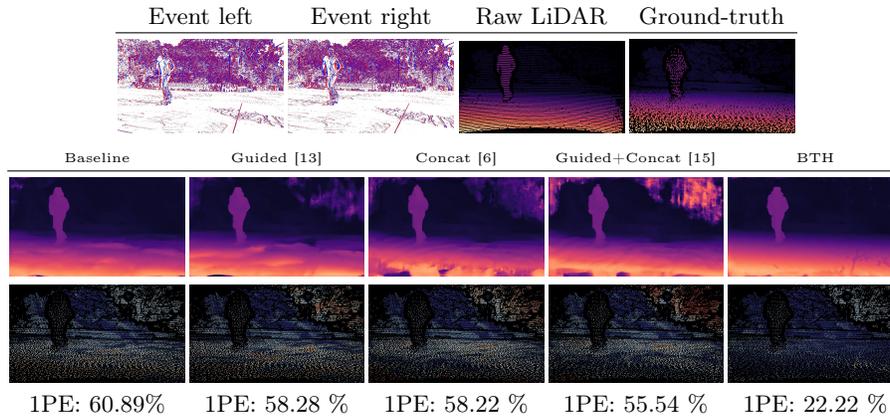


Fig. XVI: Qualitative results – re-trained stereo backbone. Results on *spot_outdoor_day_skatepark_1*, with TORE [3] representation.

flight_1) and outdoor (*i.e.*, *falcon_outdoor_day_penno_parking_1*) scenarios. The former case (Fig. XVII) highlights the behaviour of not-retrained stereo backbones at recovering fine details such as the tip of the cone and the upper part of the left-most foreground object. Compared to other methodologies, BTH manages to preserve more thin details. Furthermore, the latter case (Fig. XVIII) stress the performance of retrained stereo backbones in case of large uniform areas – *i.e.*, the road. Guided [13] is almost ineffective, while using both Concat [6] and Guided+Concat [15] leads to 20% error reduction. However, when dealing with large homogeneous regions where LiDAR coverage is limited, BTH clearly dominates alternative methods, achieving a remarkable 80% reduction in error.

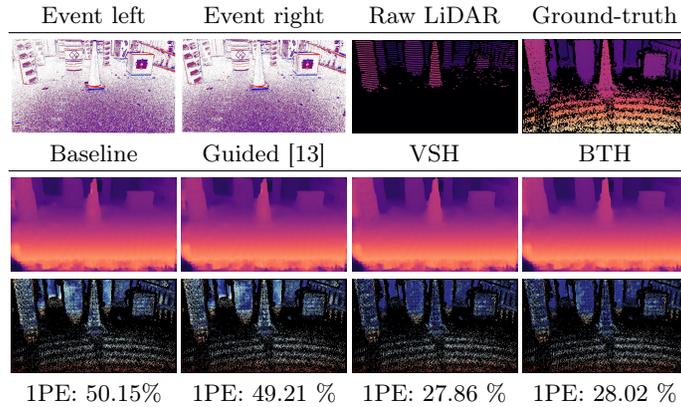


Fig. XVII: Qualitative results – baseline stereo backbone. Results on *falcon_indoor_flight_1*, with Voxel Grid [16] representation.

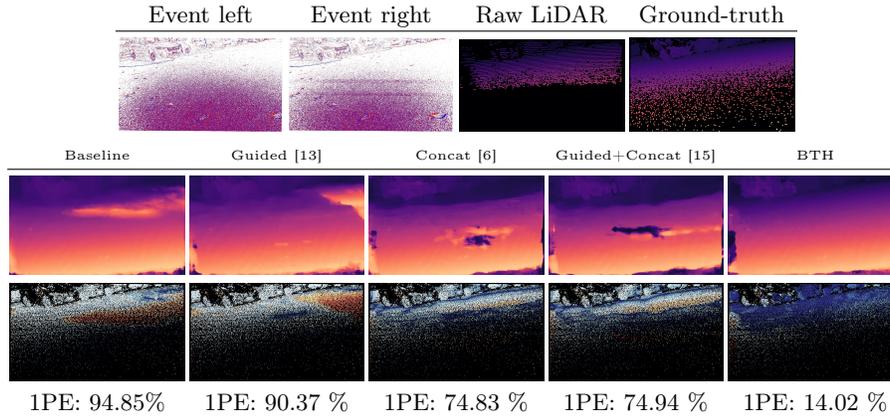


Fig. XVIII: Qualitative results – re-trained stereo backbone. Results on *falcon_outdoor_day_penno_parking_1*, with MDES [12] representation.

References

1. Babin, P., Giguere, P., Pomerleau, F.: Analysis of robust functions for registration algorithms. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 1451–1457. IEEE (2019)
2. Bai, C., Xiao, T., Chen, Y., Wang, H., Zhang, F., Gao, X.: Faster-lio: Lightweight tightly coupled lidar-inertial odometry using parallel sparse incremental voxels. IEEE Robotics and Automation Letters **7**(2), 4861–4868 (2022). <https://doi.org/10.1109/LRA.2022.3152830>
3. Baldwin, R.W., Liu, R., Almatrafi, M., Asari, V., Hirakawa, K.: Time-ordered recent event (tore) volumes for event cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 2519–2532 (2022)
4. Bartolomei, L., Poggi, M., Tosi, F., Conti, A., Mattoccia, S.: Active stereo without pattern projector. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18470–18482 (October 2023)

5. Chaney, K., Cladera, F., Wang, Z., Bisulco, A., Hsieh, M.A., Korpela, C., Kumar, V., Taylor, C.J., Daniilidis, K.: M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4015–4022 (June 2023)
6. Cheng, X., Zhong, Y., Dai, Y., Ji, P., Li, H.: Noise-aware unsupervised deep lidar-stereo fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6339–6348 (2019)
7. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters* (2021). <https://doi.org/10.1109/LRA.2021.3068942>
8. Huang, Z., Sun, L., Zhao, C., Li, S., Su, S.: Eventpoint: Self-supervised interest point detection and description for event-based camera. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5396–5405 (January 2023)
9. Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.B.: Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence* **39**(7), 1346–1359 (2016)
10. Manduchi, R., Tomasi, C.: Distinctiveness maps for image matching. In: Proceedings 10th International Conference on Image Analysis and Processing. pp. 26–31. IEEE (1999)
11. Maqueda, A.I., Loquercio, A., Gallego, G., García, N., Scaramuzza, D.: Event-based vision meets deep learning on steering prediction for self-driving cars. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5419–5427 (2018)
12. Nam, Y., Mostafavi, M., Yoon, K.J., Choi, J.: Stereo depth from events cameras: Concentrate and focus on the future. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6114–6123 (2022)
13. Poggi, M., Pallotti, D., Tosi, F., Mattoccia, S.: Guided stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 979–988 (2019)
14. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: Proceedings third international conference on 3-D digital imaging and modeling. pp. 145–152. IEEE (2001)
15. Wang, T.H., Hu, H.N., Lin, C.H., Tsai, Y.H., Chiu, W.C., Sun, M.: 3d lidar and stereo fusion using stereo matching network with conditional cost volume normalization. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5895–5902. IEEE (2019)
16. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–997 (2019)
17. Zubić, N., Gehrig, D., Gehrig, M., Scaramuzza, D.: From chaos comes order: Ordering event representations for object recognition and detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12846–12856 (October 2023)