E-GMFLOW: TIME GRANULARITY FOR TRANSFORMER ARCHITECTURES IN EVENT-BASED OPTICAL FLOW

Author
Anca Badiu

Supervisors
Hesam Araghi
Nergis Tömen

TU Delft

01 INTRODUCTION

Research question:
Does a higher time granularity in the event representation for a transformer-based model improve accuracy for event-based optical flow?

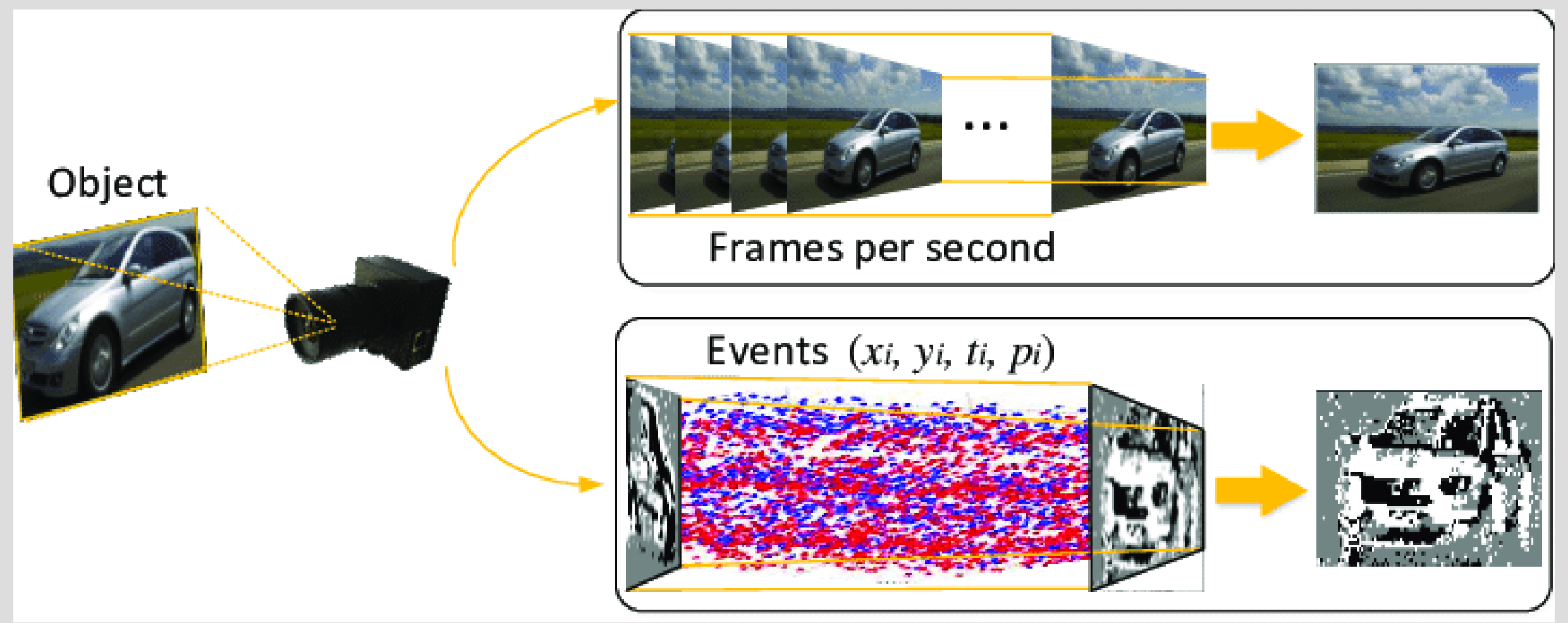


Fig 1. Comparison between standard and event cameras [2]

Optical flow estimation is a computer vision task designed to estimate the motion of objects in a video and generally formulated as estimating a displacement field between frames.

Event cameras are designed to mimic the functionality of the human eye's retina and offer significant advantages over standard cameras such as high temporal resolution (> 10K fps), but their output is more challenging to process.

Transformers [1] are a type of deep learning model centered on the mechanism of self-attention, designed to process sequential data. Transformer architectures are achieving SOTA performance for standard camera optical flow estimation, but fewer such architectures are used for event cameras. Given their ability to capture long-term dependencies, transformers could take better advantage of the high temporal resolution of event cameras.

02 BACKGROUND

Volumetric voxel grid

A relatively common approach to representing events is a volumetric voxel grid representation [3]. Given a batch of N events $\{ (x_i, y_i, t_i, p_i) | i \in [1, N] \}$, we discretize the time domain into B bins. To improve the amount of encoded time information, events are inserted into the volume using the bilinear interpolation function:

$$t_i^* = (B-1)(t_i - t_1)/(t_N - t_1)$$

$$V(x,y,t) = \sum_i p_i \cdot k_b(x - x_i) \cdot k_b(y - y_i) \cdot k_b(t - t_i^*)$$

$$k_b(a) \; = \; max(0, 1 \; - \; |a|)$$

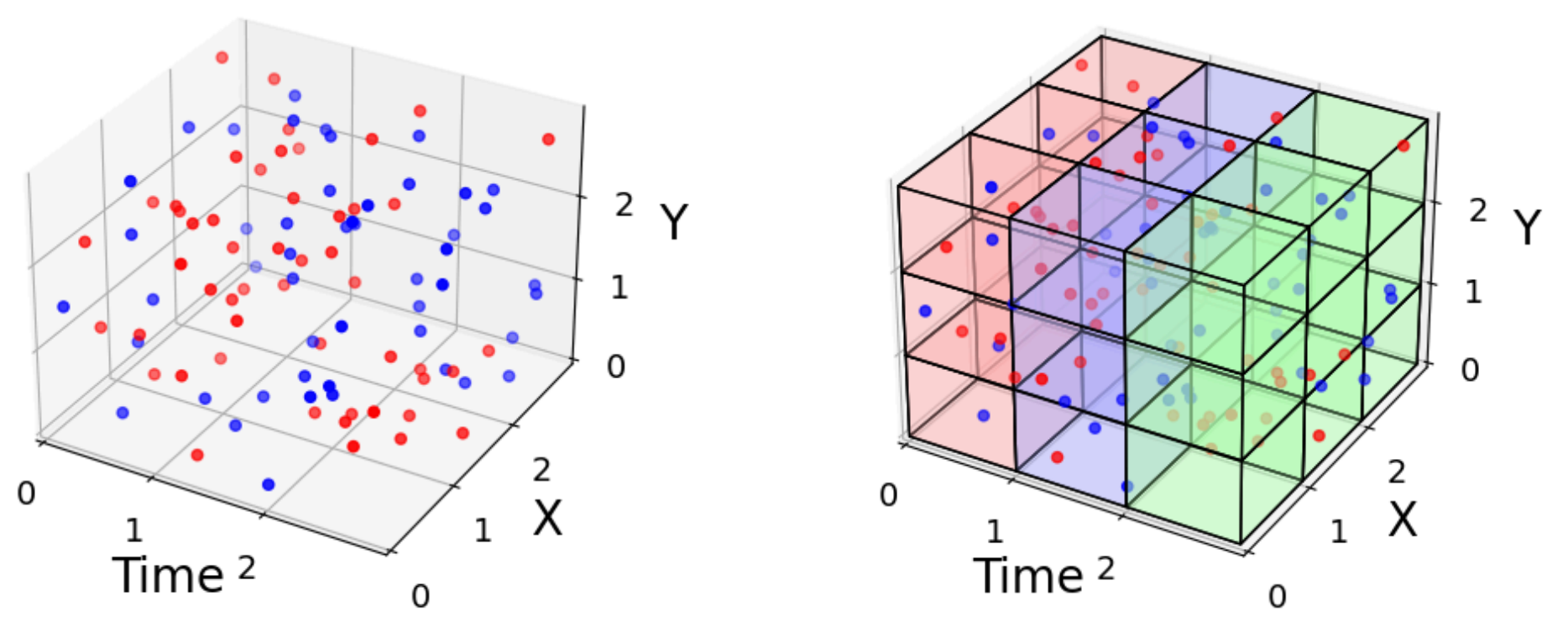


Fig 2. Event stream before and after converting it to a volumetric voxel grid

GMFlow

GMFlow [4] is a optical flow algorithm for standard cameras with a global matching approach that is effective at dealing with large displacements.

First, each of the frames gets passed through a CNN to extract relevant features. For enhancing this feature extraction transformer architecture is used to account for the mutual relationships between the previously extracted features.

A 4D-correlation volume is constructed out of the feature vectors:

$$C = \frac{\hat{F}_1 \hat{F}_2^{\,T}}{\sqrt{D}} \in \mathbb{R}^{H \times W \times H \times W}$$

Now, for each pixel in the first image, we would like to find a matching pixel in the second image such the pair has a high correlation value. To do this we can use a softmax layer on the correlation volume. We use a self-attention layer to account for occlusions.

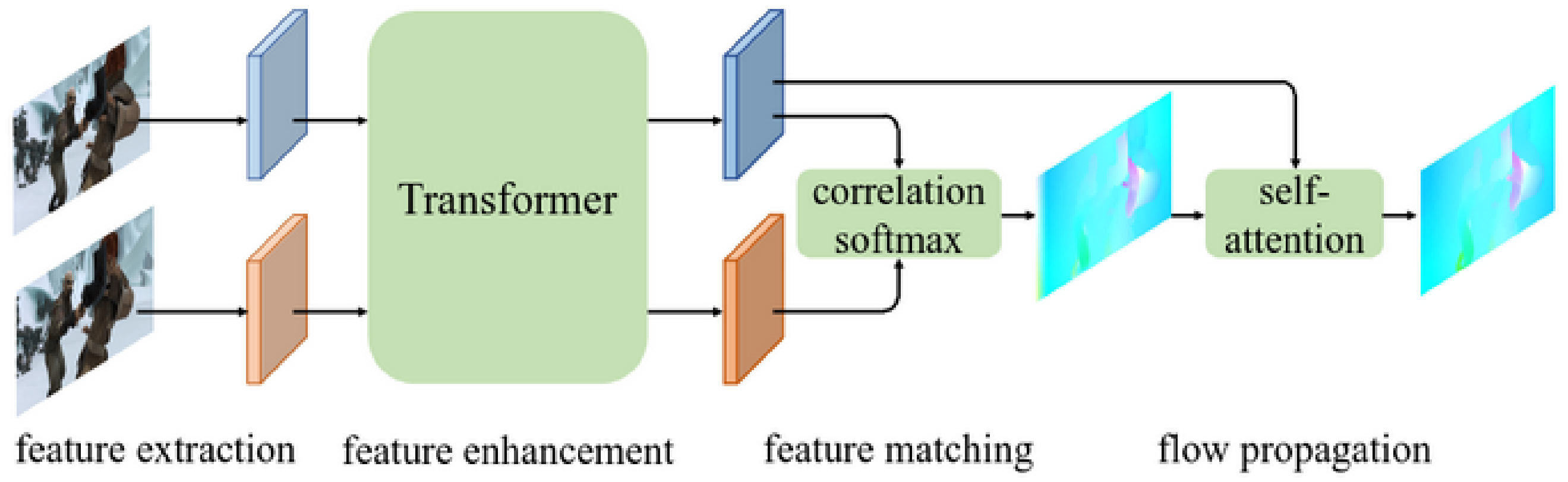


Fig 3. Overview of the GMFlow framework [4].

03 PROPOSED METHOD & EXPERIMENTS

Approach:
- To compute the optical flow from time t_i to time t_{i+1}, we choose event sequences from time (t_{i-1}, t_i) and (t_i, t_{i+1}) and process each into a volumetric voxel grid.
- We adapted the first layer of the CNN feature extractor to process more than 3 channels.

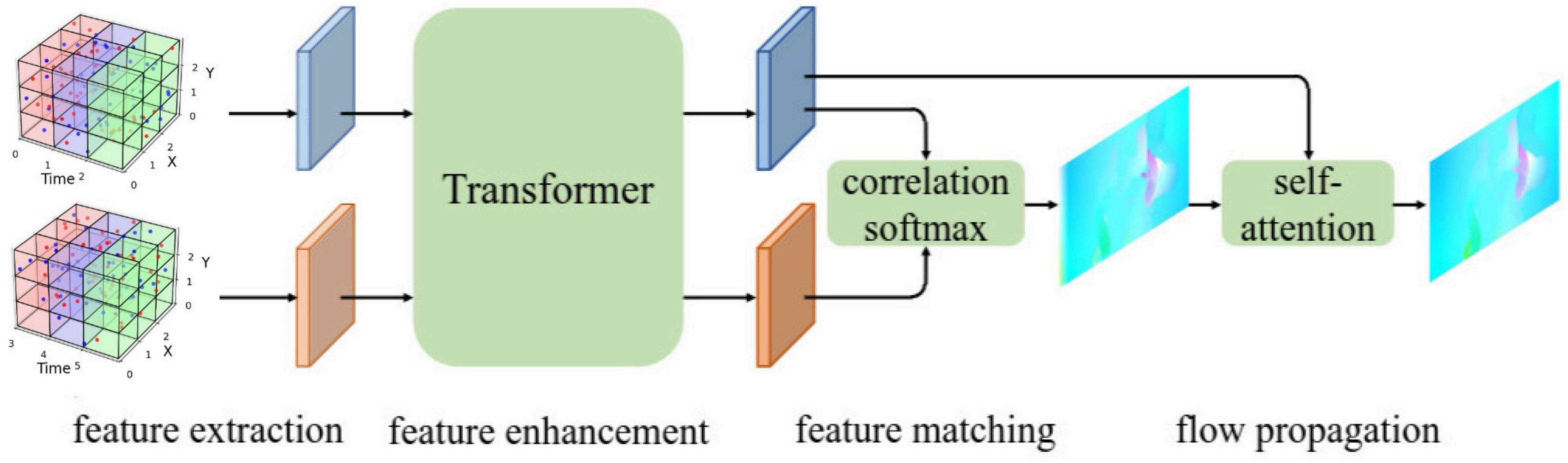
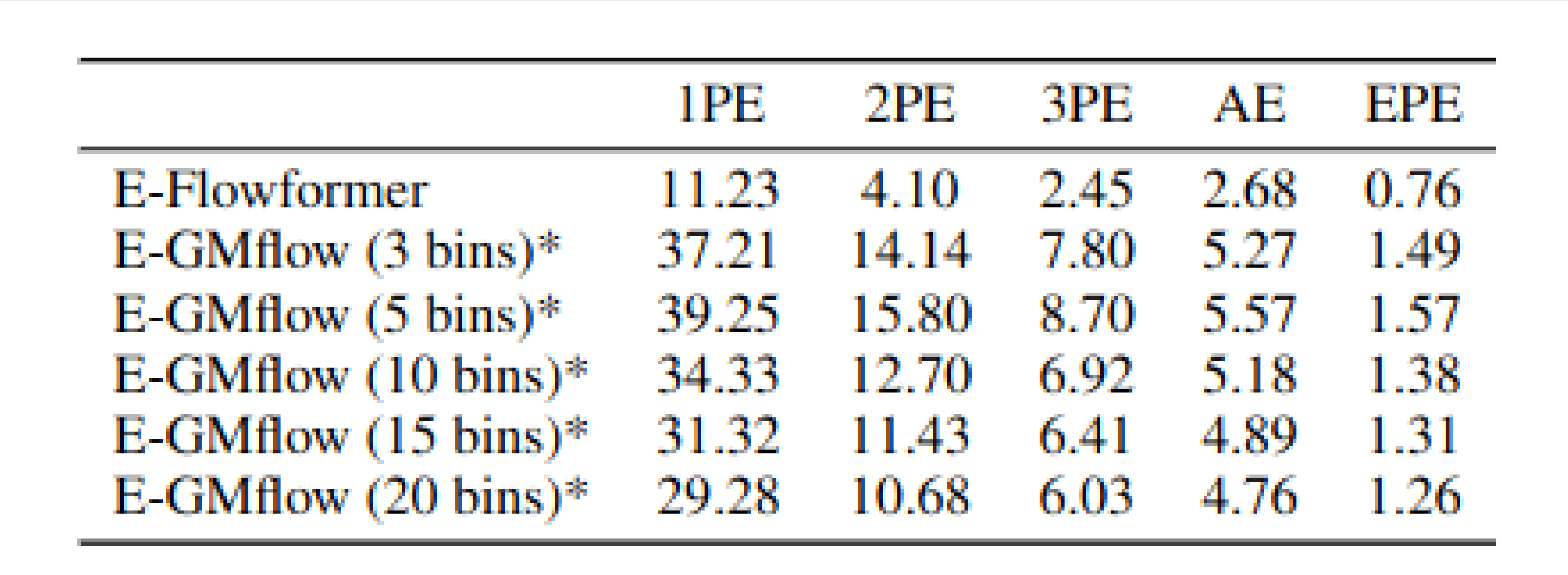


Fig 4. Overview of the E-GMFlow framework. Adapted from [4].

Datasets:
- **DSEC** is a dataset for event-based vision containing driving sequences captured at various times of day.
- **KITTI** is a dataset containing traffic scenarios that can be used for standard camera optical flow prediction.

Training and Testing:
- The GMFlow pre-trained model with refinement trained on the KITTI dataset was used and fine-tuned. We only fine-tuned the weights required for the CNN encoder.
- Fine-tuned on the entirety of the training split of DSEC .
- Tested on the testing split of DSEC.
- Experiments were run for 5 bin sizes: 3, 5, 10, 15, 20.

04 RESULTS & CONCLUSION



	1PE	2PE	3PE	AE	EPE
E-Flowformer	11.23	4.10	2.45	2.68	0.76
E-GMflow (3 bins)*	37.21	14.14	7.80	5.27	1.49
E-GMflow (5 bins)*	39.25	15.80	8.70	5.57	1.57
E-GMflow (10 bins)*	34.33	12.70	6.92	5.18	1.38
E-GMflow (15 bins)*	31.32	11.43	6.41	4.89	1.31
E-GMflow (20 bins)*	29.28	10.68	6.03	4.76	1.26

Table 1. Accuracy comparison between E-Flowformer and E-GMFlow with 5 different numbers of time bins.

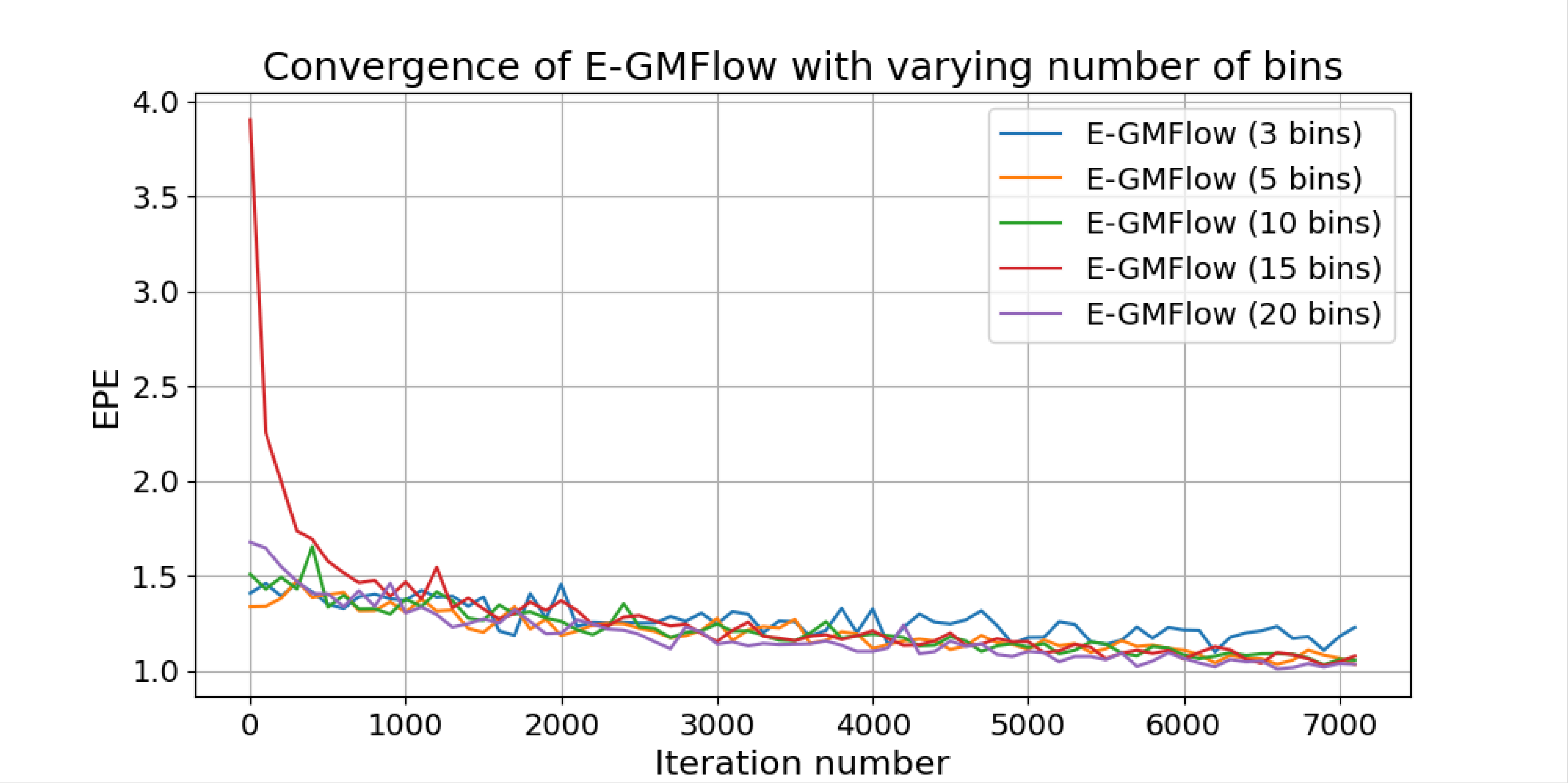


Fig 5. Training EPE for E-GMFlow with 3, 5, 10, 15 and 20 bins. The EPE is calculated during training every 100 iterations on a batch size of 6 samples.

The results suggests that the increase in the number of time bins for E-GMFlow corresponds to an increase in accuracy. However, due to the low number of iterations, it is not entirely clear whether this effect would still be observable if the model was trained for a longer amount of time.

Given the quality of the results for such a low number of iterations, it is likely that E-GMFlow would perform significantly better if trained for a longer time. Future research could investigate this potential by training the model for a longer duration.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[2] Gao, S., Guo, G., Huang, H., Cheng, X., & Chen, C. P. (2020). An end-to-end broad learning system for event-based object classification. IEEE Access, 8, 45974-45984.

[3] Zhu, A. Z., Yuan, L., Chaney, K., & 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 (pp. 989-997).

[4] Xu, H., Zhang, J., Cai, J., Rezatofighi, H., & Tao, D. (2022). Gmflow: Learning optical flow via global matching. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8121-8130).