Unsupervised optical flow estimation of event cameras

The influence of training sets on model performance

Mark van den Berg

Supervisor(s): Hesam Araghi, Nergis Tömen

EEMCS, Delft University of Technology, The Netherlands

Draft of Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Mark van den Berg
Final project course: CSE3000 Research Project
Thesis committee: Hesam Araghi, Nergis Tömen, Guohao Lan

Abstract

Event cameras are cameras that capture events asynchronously based on changes in lighting. They offer multiple benifits, but pose challenges in computer vision due to their asynchronous nature and hard to capture ground truth values to compare against. This paper shows the effects training of a state of the art unsupervised learning algorithm Taming Contrast Maximisation for predicting optical flow on a new dataset BlinkFlow which promises improvements in performance of supervised algorithms. This paper aims to see if these improved performances also happen for unsupervised models. Results of this research were inconclusive for the effectiveness of training unsupervised models, but it was shown that pretrained models on DSEC and MVSEC datasets did not perform well on this new dataset.

1 Introduction

Event-based cameras are cameras inspired by biology, and respond to changes in brightness, instead of capturing a set amount of frames every second[2]. Whenever a pixels brightness threshold is crossed it signals either a positive or negative change in brightness, and its location [3]. Pixel brightness usually works on a logarithmic scale [2] and event based cameras have multiple advantages compared to regular cameras, namely lower latency, high event rate compared to frame rate, and high dynamic range. [17]. However, the different representation of events compared to frames means that event-based cameras are not compatible with classical frame based computer vision algorithms.[13] This different representation of vision compared to regular cameras mean that regular computer vision algorithms that compare frames do not work on this data. Multiple computer vision problems are also problems for event cameras, this paper focuses specifically on optical flow estimation, the estimation of motion in a scene.

1.1 Prior Work

There are multiple method to estimate optical flow for event based cameras, and they can be grouped in different ways [3][5][13]. There are model based methods that use an algorithm to calculate optical flow, like [3] or there are machine learning models. This paper focuses on the machine learning models and separates the models into supervised and unsupervised methods. Supervised methods learn on a dataset with ground truth. Unsupervised methods do not rely on ground truth datasets. Recent research has focused more on unsupervised methods because of a lack of ground truth datasets for supervised methods [13].

Generalisability

A well known problem for these machine learning models is that models trained on one data set sometimes do not carry over to different data sets. This is called generalisability. On average unsupervised methods generalise better than supervised methods [15]. However, in the literature, quite little attention is drawn to generalisability of models. In a survey of deep learning methods [17] we find a table that shows that almost all methods evaluated are also trained on at least some part of the same dataset. In fact, the supervised methods more commonly are trained on other datasets than MVSEC. This draws into question the claim by [15] that unsupervised methods generalise better.

However, a new ground truth dataset BlinkFlow has recently been released claiming improvement for the accuracy of deep learning methods [8]. To support this claim the authors only train three different supervised models and show that the performance of these models improves by over 80%.

Contrast Maximisation

Contrast Maximisation is a method based model to estimates optical flow [3]. It works by splitting videos into small parts in which it assumes motion to be linear. In every split, it looks to flow all events back to a single time frame t_{ref} according to equation 1.

$$\mathbf{x}'_k \doteq \mathbf{W}(\mathbf{x}_k, t_k; \boldsymbol{\theta}) \doteq \mathbf{x}_k - (t_k - t_{\text{ref}})\boldsymbol{\theta} \tag{1}$$

In this equation θ is the proposed flow. This flow is chosen by taking several possible proposed flows and creating an image of warped events (IWE) from the events in split according to equation 2.

$$H(\mathbf{x}; \Theta) \doteq \sum_{k=1}^{N_e} \delta(\mathbf{x} - \mathbf{x}'_k) \tag{2}$$

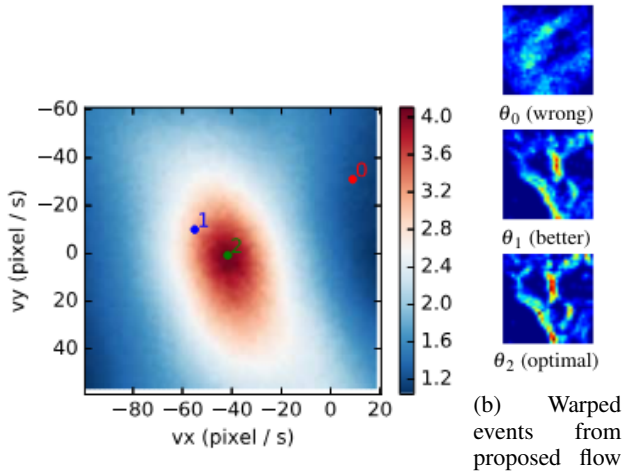
We then calculate the contrast of these different IWE's with equation 3. The higher this is the better, since it means a lot of events can be explained by a movement, so most likely this is a movement that is happening in the scene.

$$f(\Theta) = \sigma^2(H(\mathbf{x}; \Theta)) \doteq \frac{1}{N_p} \sum_{i,j} (h_{ij} - \mu_H)^2 \tag{3}$$

A visual explanation of this can be seen in figure 1. Where on the left we can see the contrast as calculated with equation (3) for different proposed flows. And on the right we can see the IWE's created by different proposed flows θ made with equation 2.

Research question

The research question that this paper aims to answer is *"What is the accuracy in terms of AEE, RSAT and FWL of a the unsupervised model Taming Contrast Maximisation trained on BlinkFlow on its accuracy in terms of AEE, RSAT and FWL on the DSEC dataset compared to the Taming Contrast Maximisation that is trained on the DSEC dataset."*. This research will contribute two parts to the current understanding of event camera optical flow. The first part is the generalisability of the Taming Contrast Maximisation algorithm. Since the paper in which this is introduced only uses a single DSEC trained model and compares it only to MVSEC a simpler dataset, we do not know the overall generalisability of the Taming Contrast Maximisation algorithm. The second part is that the overall knowledge



(a) Contrast f as a function of proposed flow θ. (b) Warped events from proposed flow θ

Figure 1: Optical flow prediction from contrast maximisation. The predicted flow is the one which maximises the contrast as in fig 1a to produce image of warped events as seen in fig 1b. From: [3]

about the quality and importance of BlinkFlow as a dataset will be further developed. If it turns out that BlinkFlow also improves unsupervised algorithms like Taming Contrast Maximisation it is important to do more research into it. If it turns out BlinkFlow does not improve the performance of Taming Contrast Maximisation it might not improve unsupervised algorithm performance, which would make it less important.

2 Methodology

The following chapter outlines the methodology and goal of this research. This research uses the code created in Taming Contrast Maximisation [10] and described below and uses it to train a new model on the BlinkFlow dataset. This model will then be evaluated against the DSEC and MVSEC datasets. The results of this model can then be compared to the pre-trained model on DSEC and on MVSEC.

Taming Contrast Maximisation

This section outlines the workings of the Taming Contrast Maximisation algorithm. Figure 2 represents the entire algorithm. Below parts of the algorithm will be described in more detail.

Input representation

Event cameras output events, these are given as a thruple of (coord: x, y, timestamp: ti, and polarity: +, -). Every time interval all events are collected and seperated into two channels positive and negative. This leaves us with two maps of all pixels and the events which have taken place there during the time interval. These two maps then are input into our neural network and the neural network estimates the flow and creates an IWE with equation 2.

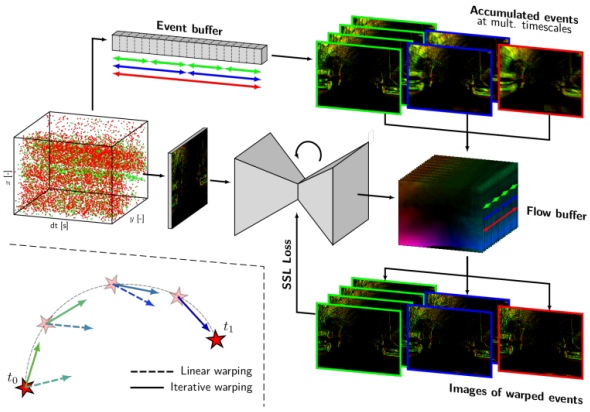


Figure 2: Outline of the Taming Contrast Maximisation algorithm. From: [10]

Loss function

To evaluate how well an IWE explains the events that have taken place in the current time frame we use a loss function. For every pixel in the IWE, temporal loss is calculated with equation 4 for both positive and negative polarity. The temporal loss is scaled by the time difference between the start of this time interval and the event time. This temporal loss only takes into account events which are sufficiently close to be considered caused by this optical flow with equation 5.

$$T_p\left(x;u|t_{\text{ref}}\right) = \frac{\sum_j \kappa(x-x_j)\kappa(y-y_j)\bar{t}_j(t_{\text{ref}},t_j)}{\sum_j \kappa(x-x_j)\kappa(y-y_j)+\epsilon} \tag{4}$$

$$\kappa(a) = \max(0, 1-|a|) \tag{5}$$

$$j = \{i \mid p_i = p'\}, \quad p' \in \{+,-\}, \quad \epsilon \approx 0 \tag{6}$$

$$\mathcal{L}_{CM}(t_{\text{ref}}) = \frac{\sum_x T_+(x;u|t_{\text{ref}})^2 + T_-(x;u|t_{\text{ref}})^2}{\sum_x [n(x')>0]+\epsilon} \tag{7}$$

$$\mathcal{L}_{CM}^{R} = \frac{1}{R+1}\sum_{t_{\text{ref}}=0}^{R} \mathcal{L}_{CM}(t_{\text{ref}}) \tag{8}$$

$$\mathcal{L}_{CM}^{\text{multi}} = \frac{1}{S}\sum_{s=0}^{S-1}\frac{1}{2^s}\sum_{p=0}^{2^s-1}\mathcal{L}_{CM,p}^{R/2^s} \tag{9}$$

$$\bar{t}_i(t_{ref},t_i) = 1 - \frac{|\,t_{ref}-t_i\,|}{R}, t_i \in [0,R] \tag{10}$$

Total loss for this time interval is then calculated with equation 7 which goes over every pixel in the IWE, squares the temporal loss for both positive and negative polarity and then normalises by the amount of events that were mapped to this particular pixel.

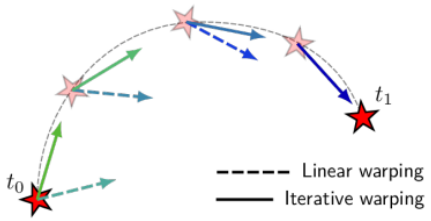


Figure 3: Comparison of iterative and linear warping. Linear warping would not accurately describe the path from t0 to t1, but would assume linear movement, compared to iterative warping which tracks the movement better. From: [10]

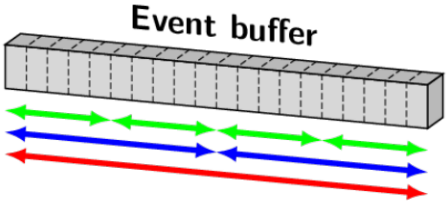


Figure 4: Splitting of event windowns with R=20 and S=3. The 20 windows are first fully evaluated in red and then split in half 2 more times to calculate the blue and green subwindows because of S. From: [10]

Iterative warping

A limitation of contrast maximisation is that the optical flow predictions are always linear. This is not always representative of actual movements. The Taming Contrast Maximisation algorithm deals with this by iterative warping. This means we apply contrast maximisation over multiple different time frames and combine their outcomes. This means we get a more accurate view of the optical flow as represented in figure 3. Iterative warping accuracy is decided by adjusting two hyperparameters, namely R and S. Parameter R decides how many time intervals are chained in a row when calculating optical flow. S decides how far we subdivide the time intervals for intermediate optical flow. R gets divided into 2^S pieces at the smallest. In figure 4 we can see the subdivisions for R=20 and S=3, and in equation 9 we can see that all larger subdivisions are still taken into account. According to [10] S is added to make the model more robust and reduce its dependency on hyper parameter tuning. Because with a higher S lower subdivisions are still taken into account, you are less dependant on getting your R exactly right.

Hyper parameters

Hyper parameters are parameters that are set by us before training a model. The three hyper parameters that are especially important in this are the time interval over which we collect events, R, and S. They decide for a large part how well our model performs as seen in table 1. According to [10] the best settings for hyper parameters are mostly dependant on the dataset on which we evaluate. We can slightly lower this dependency by increasing S, but at a currently unknown cost in terms of computing. As seen in table 1, R=5, S=4 is the best performing model, however, the paper goes on to use the R=10, S=1 model. Which although not stated we hypothesise to be because of the increased computing cost of running the model with a large S.

	EPE↓	$\%_{3PE}$↓	FWL↑	RSAT↓
dt = 0.01s, R = 2, S = 1	9.66	86.44	**1.91**	1.07
dt = 0.01s, R = 5, S = 1	7.40	50.62	1.39	0.96
dt = 0.01s, R = 10, S = 1	<u>4.52</u>	<u>24.08</u>	1.51	<u>0.92</u>
dt = 0.01s, R = 10, S = 4	16.63	73.13	<u>1.59</u>	1.02
dt = 0.02s, R = 5, S = 1	8.52	35.04	1.36	1.03
dt = 0.02s, R = 5, S = 4	**2.73**	**23.73**	1.43	**0.91**

Table 1: Table comparing hyper parameters, best in **bold**, runner up is <u>underlined</u>. ↓ means lower is better, ↑ means higher is better. From: [10]

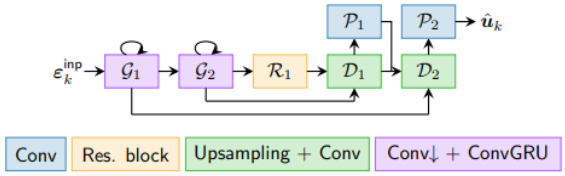


Figure 5: Neural network of Taming Contrast Maximisation. From: [10]

Network design

The neural network design of Taming Contrast Maximisation is based on a recurrent version of EV-FlowNet that was proposed by Gehrig et al. in [5] and is shown in figure 5. It consists of four encoders as described in [1], then it is followed by two residual blocks as described in [6] and then it is followed by four decoders. Lastly, the P blocks in the figure are used to produce estimates at lower time scales, so for higher values for S.

3 Experimental setup

Datasets

This paper uses three datasets, namely MVSEC, DSEC and BlinkFlow. These three sections are described below.

MVSEC

MVSEC or Multi Vehicle Stereo Event Camera dataset is a dataset from 2018 that consists of multiple videos of a set of stereo DAVIS event cameras mounted on different vehicles and in different lighting conditions [18]. They consists of videos from a car, motorcycle, and drone. They have videos in indoor and outdoor, day and night conditions. Apart from the events, the dataset also supplies grayscale images, intertial measurement and a second event camera's events for depth estimation. The dataset has relatively little motion compared to DSEC, with 80% of pixel displacements magnitude smaller then 4 pixels, which is only 18% of DSECs pixel displacement magnitude of 22 pixels [5].

Figure 6: A sample picture of MVSEC video Outdoor Driving Day. From: [12]

DSEC

DSEC is a dataset from 2021 designed for autonomous driving vehicles and therefore consists of only driving scenarios. [4]. Apart from the event cameras the dataset also supplies RGB images of the driving scenes, as well as GPS and inertial measurement. DSEC has a lot more motion at a pixel displacement magnitude of 22. [5].

Figure 7: A sample picture of DSEC video thun_00_a.

BlinkFlow

BlinkFlow is a dataset that is completely generated. It consists of videos of different objects drawn from a pool of object moving thru a scene according to some predefined rules [8], an example of what this might look like is seen in figure 8. BlinkFlow promises better generalisability then DSEC and MVSEC [8]. However, the paper only shows this on 3 methods and all of them are supervised, although the models did have very significant improvements in performance of over 84% in EPE [8]. Which leads to the question, how well does an unsupervised model trained on BlinkFlow perform on DSEC and MVSEC.

Evaluation metrics

Results will be evaluated with three evaluation metrics. Firstly average end point error or AEE, where lower is better. It is the average euclidean distance between where the flow was predicted versus the ground truth, or the average EPE. Second is flow warp loss, or FWL, where higher is better. FWL is a proxy for accuracy, introduced to help when there is no ground truth available [14].

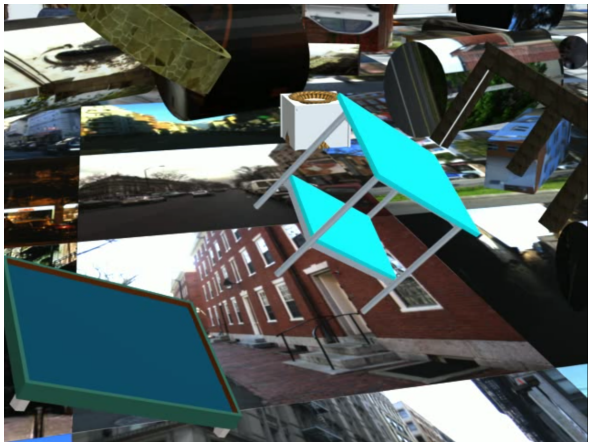


Figure 8: A sample picture of BlinkFlow video A 300 consisting of multiple table like objects floating in front of a background of pictures of buildings

The metric works by first compensating for the flow in the time frame in equation 11 and then normalising compared to the no flow image with equation 12.

$$I(E,\phi) = \begin{pmatrix} x_i' \\ y_i' \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} + (t_{ref} - t_i)\begin{pmatrix} u(x_i, y_i) \\ v(x_i, y_i) \end{pmatrix} \quad (11)$$

$$FWL = \frac{\sigma^2 I(E,\phi)}{\sigma^2 I(E,0)} \quad (12)$$

The metric RSAT or ratio of squared average timestamps. In this we compare the contrast after forward propagation of the flow and divide this by the image with no flow propagation. We expect contrast to go down after forward flow. So lower is better in this case.

$$RSAT = \frac{\mathcal{L}_{contrast}(t_{ref}^{fw} \mid u)}{\mathcal{L}_{contrast}(t_{ref}^{fw} \mid 0)} \quad (13)$$

4 Results

4.1 Evaluation on BlinkFlow

In table 2 the quantitative results of the evaluation on the BlinkFlow dataset is shown. The training settings are shown in table 3. As seen the amount of epochs differ between the different videos. This is because the amount of time that could be used for training and evaluation was limited. The longest training run was about 14 hours for R=10 S=1, the shortest was around 6 hours for R=2 S=1. The amount of epochs was at a maximum 50, but also only 1 or 2 at the other models. This is compared to the 174 that the Taming Contrast paper used to train their best DSEC model. The learning rate of all methods was 0.00001. In figure 9 we can see the best performing of our own trained models R=2 S=1. Qualitatively the model seems to detect edges very well. Filling

in larger figures with the correct flow goes wrong most often, this also goes for the background which sometimes does not have that much texture, in which case the model has a hard time determining the flow. The DSEC model had more trouble with finding the edges of objects, but was in general much better at predicting the background flow. As seen in figure 9b the DSEC model often applies the flow not only exactly to the object moving in front, but also in varying degrees to the background.

4.2 Evaluation on DSEC

Quantitative results are shown in table 4, the DSEC and MVSEC models outperform our trained models in most aspects. In terms of AEE our models have around 200%-300% higher average endpoint error then DSEC and MVSEC. Interestingly the R=2 S=1 model has the best FWL by around 35%. Qualitatively there is a large difference between the flow estimation between our models and the DSEC and MVSEC models which look quite similar. Our models have sharper edges and a lot of space where very little flow is predicted. The DSEC model predicts flow almost everywhere, but does not have sharp edges. All our trained models except R=2 S=1 predicted almost all flow in exactly the same direction for every part of the screen. Although not all models predicted the same direction. R=10 S=1 predicted mostly red and orange flow direction top-left as seen in figure 10b, but dt=0.02 R=5 S=1 predicted mostly pink flow, bottom-left flow. Two of our models have RSAT and FWL around 1, which means their flow prediction did not improve contrast compared to a predicted flow of 0.

5 Responsible Research

This section outlines the ethical aspects of this research as well as a reflection on the reproducability of this research and this field of research in general. In our opinion, the field of event camera optical flow has an issue of not being nearly reproducable enough. A lot of papers get published with no code, for example [7] [11] [16] [19] . Although the algorithms are usually explained quite well in the papers itself. The amount of work required for a researcher to actually verify the results by writing an entire project based on the paper are infeasible. There are also papers that claim certain results for models without supplying the supposed model [9], and even the Taming Contrast Maximisation paper only supplies two pretrained models, whilst they have trained and show results for a lot more models like in table 1. This means there is no way to verify the claims without training the model yourself which is very expensive, both in terms of time for the researcher as well as in energy consumption as models could take multiple days of training on power heavy setups.
To combat these issues, this paper has published both the code used, as well as all models that were used to generate the results of this paper. Which models belong to which results can be found in appendix table 5.

6 Discussion

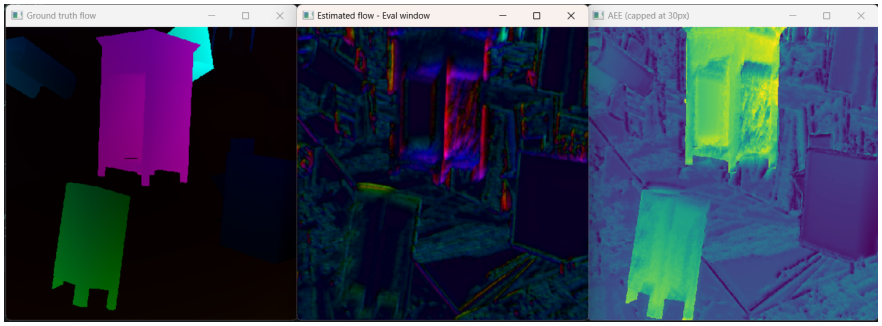
The training time for the models that we trained was too low to draw any definitive conclusions about the performance of Taming Contrast Maximisation as a whole. This was also visible in figure 10b where all flow pointed to one side. This was first thought to be a result of overfitting, leading us to up the amount of videos in the subset of the data that was used for training. However, this made the problem worse as it lead to a decrease in epochs trained.
This lead to us training our best performing model R=10 S=1 on only 10 videos in BlinkFlow, for overall 10 seconds of video. This is only a very small part of the around 3300 videos in the BlinkFlow training dataset. All 10 videos were also of subsection A of blinkflow, the subsections of BlinkFlow did not seem to have any visible distinction between them, so this might not matter.
Another shortcoming of this method was the lack of equal settings throughout the training. The amount of video's every model was trained on was based on the runtime. A fairer comparison might be to compare models trained on an equal amount of video footage, however, this was not possible due to runtime constrictions.
A question left unresolved is the question about the effect of S on runtime. S had compatibility issues with our torch and therefore cuda version. Which means we could not test the effect of S on runtime.
Lastly more evaluation metrics could be added. As the FWL metric seemed to not predict accuracy very well, it would be good to add more metrics for a more rigorous evaluation.

7 Conclusions and Future Work

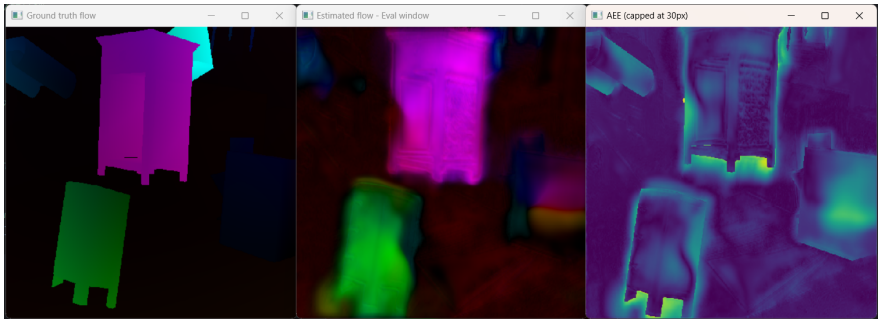
The main research question of this paper was: *"What is the accuracy in terms of AEE, RSAT and FWL of a the unsupervised model Taming Contrast Maximisation trained on BlinkFlow on its accuracy in terms of AEE, RSAT and FWL on the DSEC dataset compared to the Taming Contrast Maximisation that is trained on the DSEC dataset."*. Overall the DSEC trained model performs better than our trained models. This is most likely due to the fact that DSEC has had a lot more training than our own models. The amount of training of our models have had is too low to make any meaningful conclusions about the generalisability of the Taming Contrast Maximisation method when trained on BlinkFlow. Our best performing model performing model in terms of AEE on BlinkFlow R=10 S=1 does not have a significantly better AEE than our other models which were trained less. When looking at the data more qualitatively, the BlinkFlow trained models do seem to have some upsides compared to the DSEC and MVSEC models, in that it is a lot better at edge detection than DSEC trained models.
When looking at the generalisability of DSEC to the BlinkFlow dataset we see a significant performance drop in AEE of about 400%. This shows that DSEC does not generalise well to BlinkFlow. The largest issues with this seem to happen around edges of objects. This is not unexpected as the DSEC dataset does not have a lot of object moving across our plane of vision. DSEC is a driving dataset and the types of cross camera movement in BlinkFlow does not happen often when driving.
Another interesting finding of this research is the unreliability of FWL as a measure in our results. When evaluating on

	AEE↓	FWL↑	RSAT↓	training time per video (s)	inference time (s)
dt = 0.01s, R = 2, S = 1	15.05	2.982	1.013	121	306
dt = 0.01s, R = 5, S = 1	16.06	0.997	0.999	485	286
dt = 0.01s, R = 10, S = 1	16.94	1.092	0.996	1504	290
dt = 0.02s, R = 5, S = 1	16.52	**4.402**	1.152	1293	296
DSEC best model	**6.17**	2.207	**0.753**	unknown	290
MVSEC best model	6.93	2.04	0.769	unknown	281

Table 2: Table comparing hyper parameters of models trained on BlinkFlow subsection A video's 0-10, evaluated on subsection A video's 300-309, best in **bold**, runner up is underlined. ↓ means lower is better, ↑ means higher is better. Runtime training calculated on HP Zbook with Quadro P2000. Runtime evaluation calculated on PC with GTX 1060TI



(a) BlinkFlow A 309 evaluated by model R=2 S=1



(b) BlinkFlow A 309 evaluated by DSEC

Figure 9: BlinkFlow A 309 evaluation. Left to right: ground truth, predicted flow, AEE

	epochs trained	videos per epoch
dt = 0.01s, R = 2, S = 1	50	10
dt = 0.01s, R = 5, S = 1	1	20
dt = 0.01s, R = 10, S = 1	1	80
dt = 0.02s, R = 5, S = 1	2	10

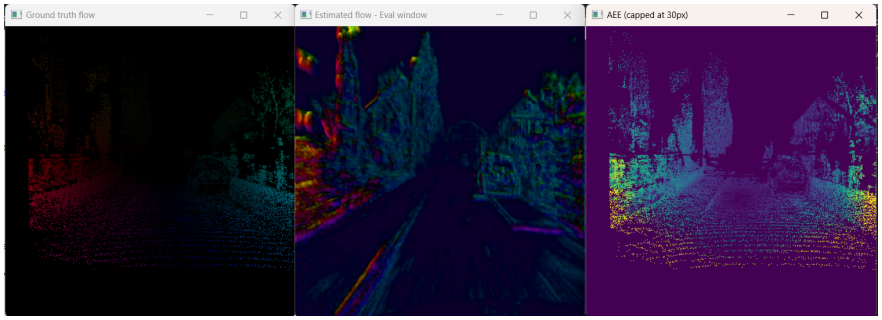
Table 3: Amount training in epochs and video's per epoch per model.

	AEE↓	FWL↑	RSAT↓
dt = 0.01s, R = 2, S = 1	9.31	1.962	1.102
dt = 0.01s, R = 5, S = 1	9.08	0.998	0.999
dt = 0.01s, R = 10, S = 1	9.94	1.022	0.989
dt = 0.02s, R = 5, S = 1	9.629	**2.547**	1.159
DSEC best model	1.758	1.188	0.870
MVSEC best model	**1.468**	1.280	**0.851**

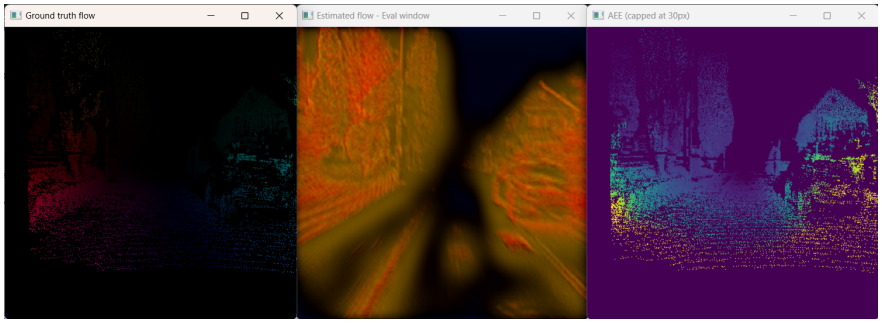
Table 4: Table comparing models trained on BlinkFlow subsection A video's 0-10, evaluated on subsection DSEC thun_00_a, best in **bold**, runner up is underlined. ↓ means lower is better, ↑ means higher is better.

Blinkflow the FWL of two of our models are the highest of all, but when looking at AEE, which takes into account ground truth data, they do not perform nearly as well as the MVSEC and DSEC model. This repeats again on the DSEC evaluation where the same two models have the highest FWL with a much lower AEE accuracy. This indicates that FWL is not a good predictor of AEE, although it does claim to be a proxy for accuracy [14]. Though it must be said that the sample size of this result is rather small and not enough to make any definitive claims about the validity of FWL as a metric.

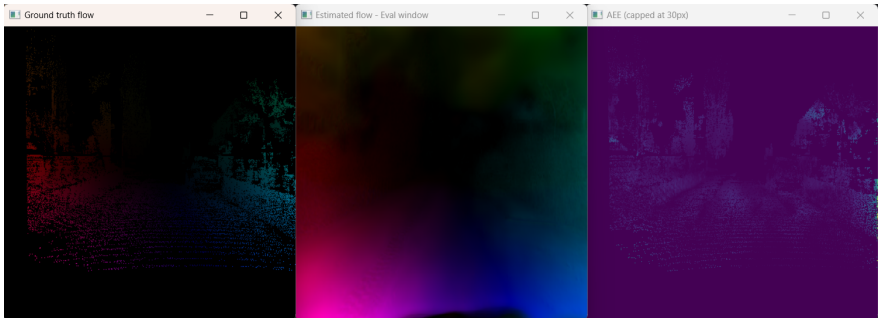
Future work could try to train models on a larger part of the BlinkFlow dataset, and for more epochs to try and



(a) Optical flow prediction trained on dt=0.01, R=2, S=1



(b) Optical flow prediction trained on dt=0.01, R=10, S=1



(c) Optical flow prediction trained on DSEC

Figure 10: DSEC thun_00_a evaluation. Left to right: ground truth, predicted flow, AEE

answer the question posed in this paper. Another question to research could be to see how well FWL can be used to predict the accuracy of different models, both supervised and unsupervised.

Appendix

Setting	Model id
dt = 0.01s, R = 2, S = 1	b63b1c2c4f2b44748bf1b5819a78046b
dt = 0.01s, R = 5, S = 1	9be02248631f4297995f852ed8012b63
dt = 0.01s, R = 10, S = 1	5acdcb9f357f40258fd7a71c6380b457
dt = 0.01s, R = 10, S = 4	unavailable
dt = 0.02s, R = 5, S = 1	6a5bf0154a4440bb87da96a2d43c79a1
dt = 0.02s, R = 5, S = 4	unavailable

Table 5: Model and corresponding ID to run

References

[1] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving Deeper into Convolutional Networks for Learning Video Representations, March 2016. arXiv:1511.06432 [cs].

[2] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Joerg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, January 2022. arXiv:1904.08405 [cs].

[3] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A Unifying Contrast Maximization Framework for Event Cameras, with Applications to Motion, Depth, and Optical Flow Estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, June 2018. arXiv:1804.01306 [cs].

[4] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A Stereo Event Camera Dataset for Driving Scenarios, March 2021. arXiv:2103.06011 [cs].

[5] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense Optical Flow from Event Cameras, October 2021. arXiv:2108.10552 [cs].

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.

[7] Adarsh Kumar Kosta and Kaushik Roy. Adaptive-SpikeNet: Event-based Optical Flow Estimation using Spiking Neural Networks with Learnable Neuronal Dynamics, March 2023. arXiv:2209.11741 [cs].

[8] Yijin Li, Zhaoyang Huang, Shuo Chen, Xiaoyu Shi, Hongsheng Li, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. BlinkFlow: A Dataset to Push the Limits of Event-based Optical Flow Estimation, March 2023. arXiv:2303.07716 [cs].

[9] Haotian Liu, Guang Chen, Sanqing Qu, Yanping Zhang, Zhijun Li, Alois Knoll, and Changjun Jiang. TMA: Temporal Motion Aggregation for Event-based Optical Flow, August 2023. arXiv:2303.11629 [cs].

[10] Federico Paredes-Vallés, Kirk Y. W. Scheper, Christophe De Wagter, and Guido C. H. E. de Croon. Taming Contrast Maximization for Learning Sequential, Low-latency, Event-based Optical Flow, September 2023. arXiv:2303.05214 [cs].

[11] Xin Peng, Ling Gao, Yifu Wang, and Laurent Kneip. Globally-Optimal Contrast Maximisation for Event Cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. arXiv:2206.05127 [cs].

[12] Hao Qu, Lilian Zhang, Xiaoping Hu, Xf He, Xianfei Pan, and Changhao Chen. Self-supervised Egomotion and Depth Learning via Bi-directional Coarse-to-Fine Scale Recovery. November 2022.

[13] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of Event-Based Optical Flow. volume 13678, pages 628–645. 2022. arXiv:2207.10022 [cs].

[14] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the Sim-to-Real Gap for Event Cameras, August 2020. arXiv:2003.09078 [cs].

[15] Zhexiong Wan, Yuchao Dai, and Yuxin Mao. Learning Dense and Continuous Optical Flow from an Event Camera, November 2022. arXiv:2211.09078 [cs].

[16] Yaozu Ye, Hao Shi, Kailun Yang, Ze Wang, Xiaoting Yin, Yining Lin, Mao Liu, Yaonan Wang, and Kaiwei Wang. Towards Anytime Optical Flow Estimation with Event Cameras, October 2023. arXiv:2307.05033 [cs, eess].

[17] Xu Zheng, Yexin Liu, Yunfan Lu, Tongyan Hua, Tianbo Pan, Weiming Zhang, Dacheng Tao, and Lin Wang. Deep Learning for Event-based Vision: A Comprehensive Survey and Benchmarks, April 2024. arXiv:2302.08890 [cs].

[18] Alex Zihao Zhu, Dinesh Thakur, Tolga Ozaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The Multi Vehicle Stereo Event Camera Dataset: An Event Camera Dataset for 3D Perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, July 2018. arXiv:1801.10202 [cs].

[19] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised Event-Based Learning of Optical Flow, Depth, and Egomotion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 989–997, June 2019. ISSN: 2575-7075.