Bounding box-based object detection with event-based data

Pascal Benschop [1]
Nergis Tömen [1], Ombretta Strafforello [1], Xin Liu [1]

[1]EEMCS, Delft University of Technology, The Netherlands

1. Background

Events are represented as a tuple (x,y,ts,p)
- x and y: coordinates of an event
- ts: timestamp of an event
- p: polarity of an event (positive or negative)

Convolutional Neural Networks (CNNs) are used for object detection

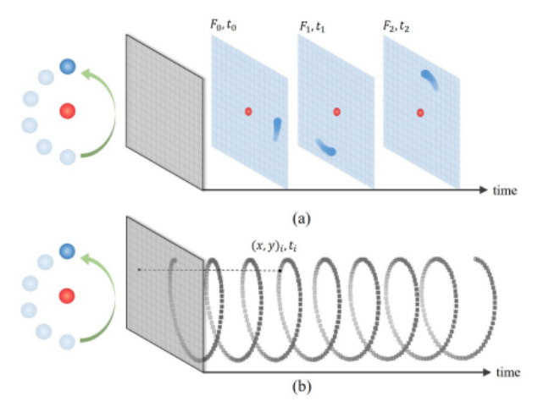


Figure 1. (a) RGB camera (b) Event camera, Source: [1]

2. Research question & Hypotheses

- **What is the accuracy-efficiency trade-off of an object detection convolutional neural network for using sparse event-based data instead of dense image-based data?**

1. Using event-based data is more efficient and similar in accuracy compared to using images as input for an object detection CNN
2. Using event-based data can lead to a better accuracy for object detection than image-based data at a similar efficiency.

3. Method

Event-based data can be represented in multiple ways [2]. The representations used are listed below:

- Image representation: 128 by 128 pixels in 3 channels: RGB
- Time Frame representation: 128 by 128 pixels in 1 channel: greyscale
- Point cloud representation: N points in 2 channels: X and Y
- Point cloud representation: N points in 3 channels: X, Y and time

All event-based data is taken from the Neuromorphic Caltech101 dataset [3]. 2 experiments are carried out:

1. Compare models on different event-based and image data
   - Model from figure 2
   - Model from figure 2 with sparse layers
   - Data selection: 4 classes (Car, Helicopter, Airplane, Motorbike)
2. Compare event-based and image data on YOLOv3 [4]
   - Time frame of different windows of time
   - Images
   - Data selection: 4 classes and entire dataset

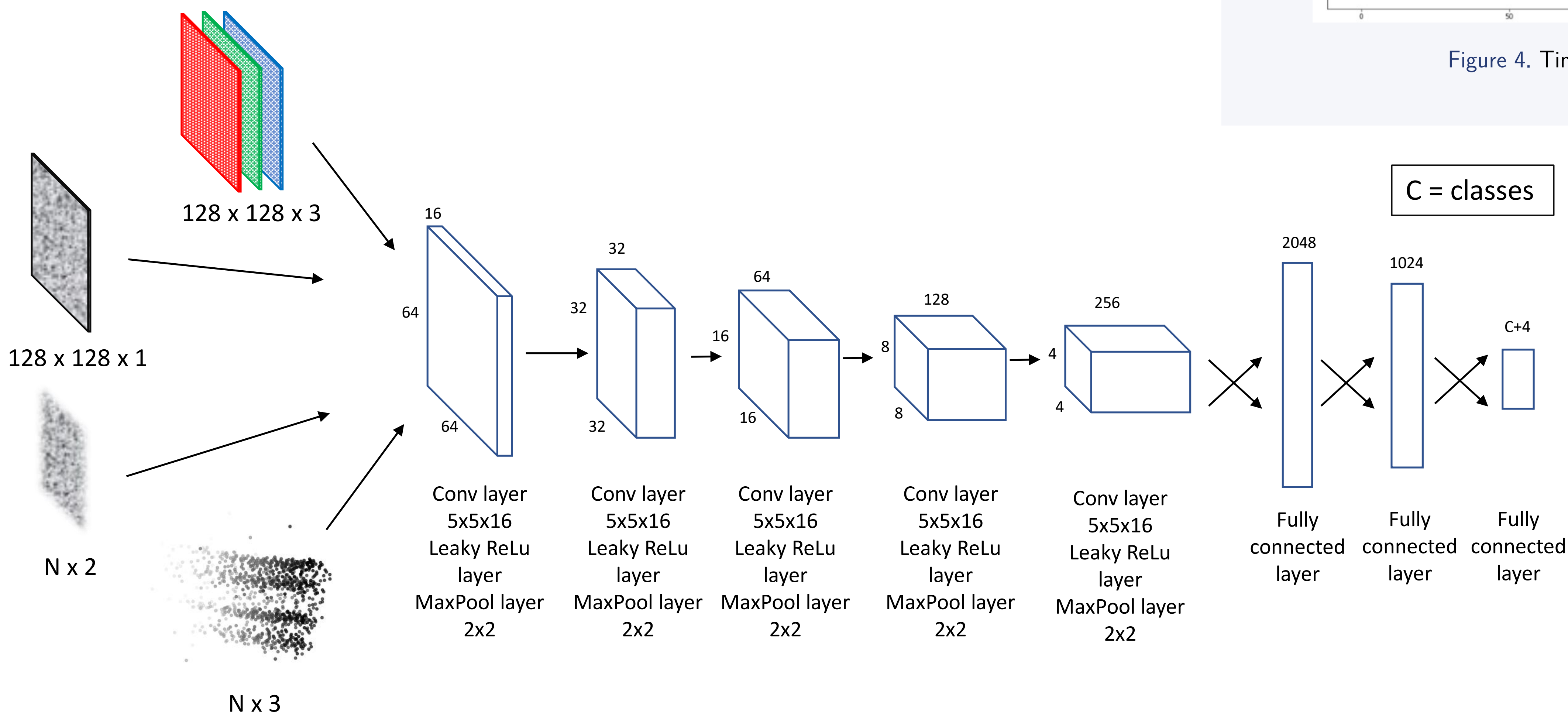

128 x 128 x 3
128 x 128 x 1
N x 2
N x 3
C = classes
16
64
64
32
32
32
64
16
16
128
8
8
256
4
4
2048
1024
C+4
Conv layer 5x5x16 Leaky ReLu layer MaxPool layer 2x2
Conv layer 5x5x16 Leaky ReLu layer MaxPool layer 2x2
Conv layer 5x5x16 Leaky ReLu layer MaxPool layer 2x2
Conv layer 5x5x16 Leaky ReLu layer MaxPool layer 2x2
Conv layer 5x5x16 Leaky ReLu layer MaxPool layer 2x2
Fully connected layer
Fully connected layer
Fully connected layer

Figure 2. CNN model used, input: data in representations, output: detected object

4. Predictions

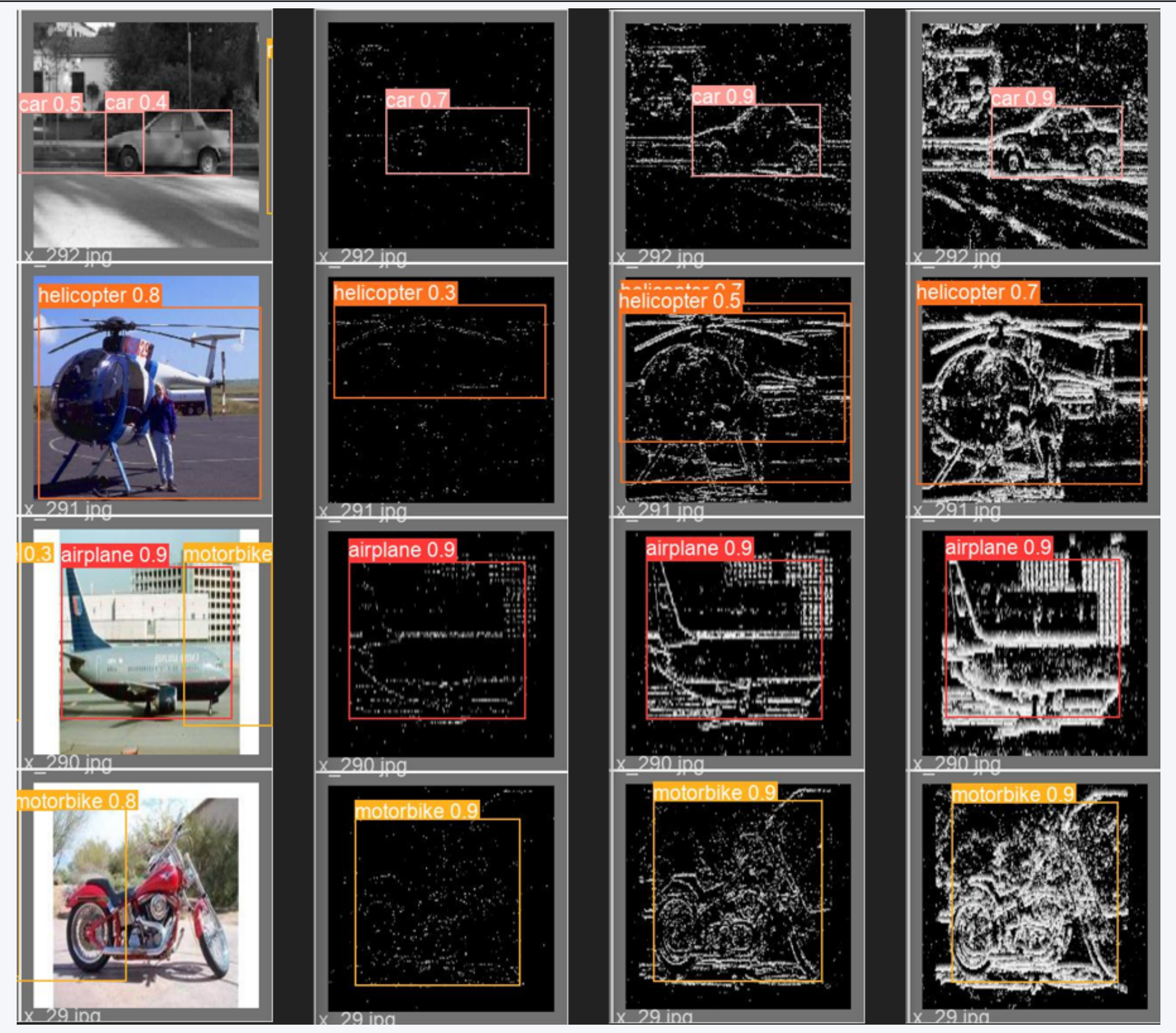


Figure 3. YOLOv3 predictions: image, TF 10ms, TF 25ms, TF 50ms

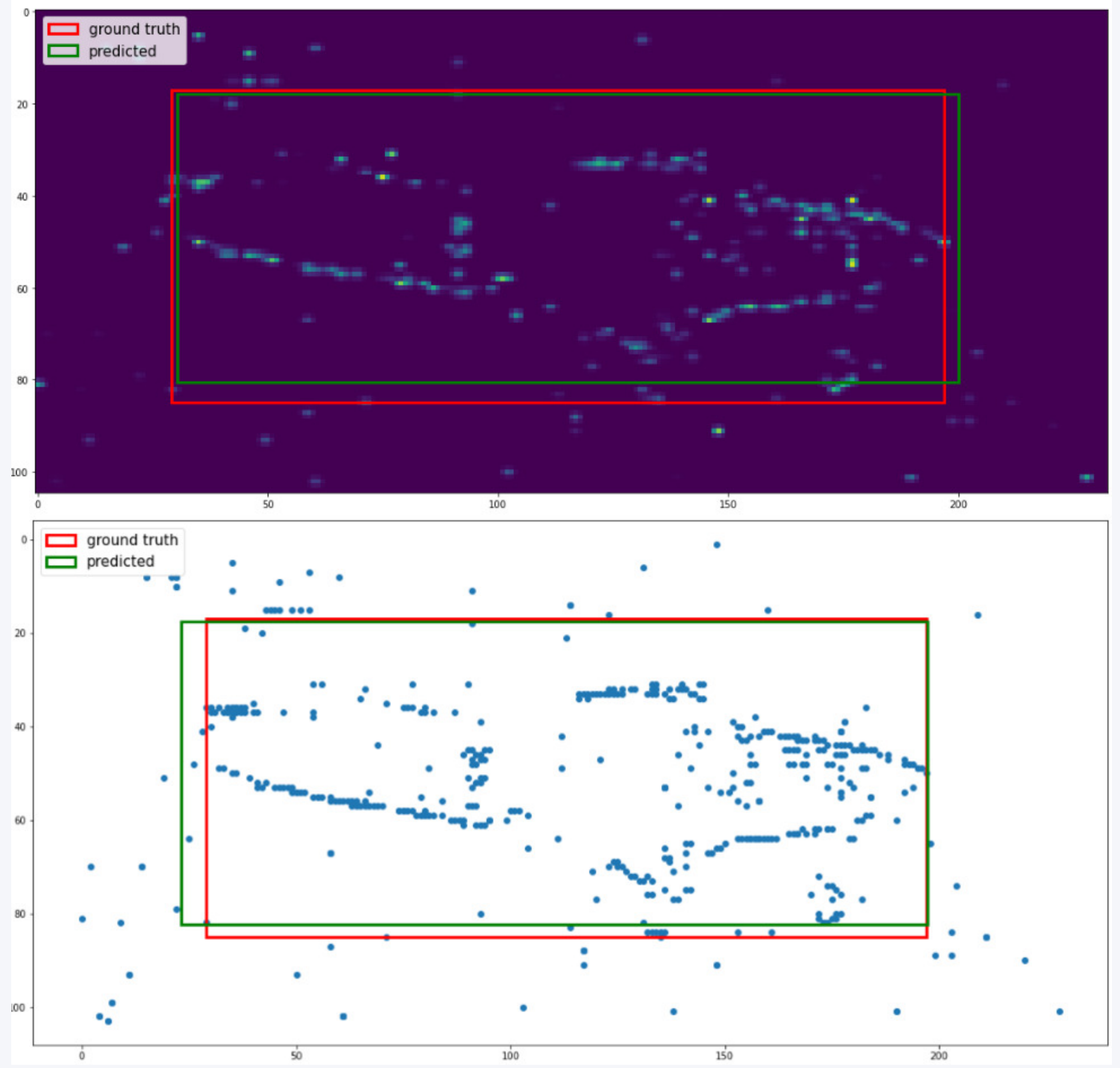
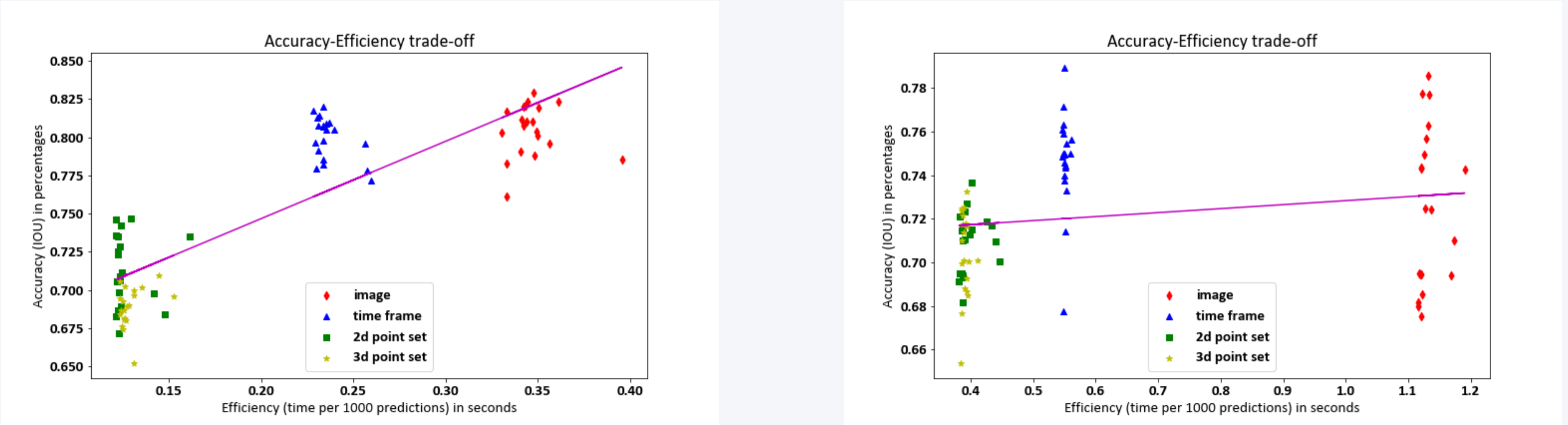


Figure 4. Time Frame & 2D point cloud predictions

5. Results

+ The accuracy-efficiency trade-off for using events is better with the sparse model
+ The time frame of events input performed better than images when using a larger time window
– The accuracy of using images is better when the entire dataset is taken as input



(a) Trade-off results from conventional model

(b) Trade-off results from sparse model

Figure 5. Comparing results of conventional vs sparse model

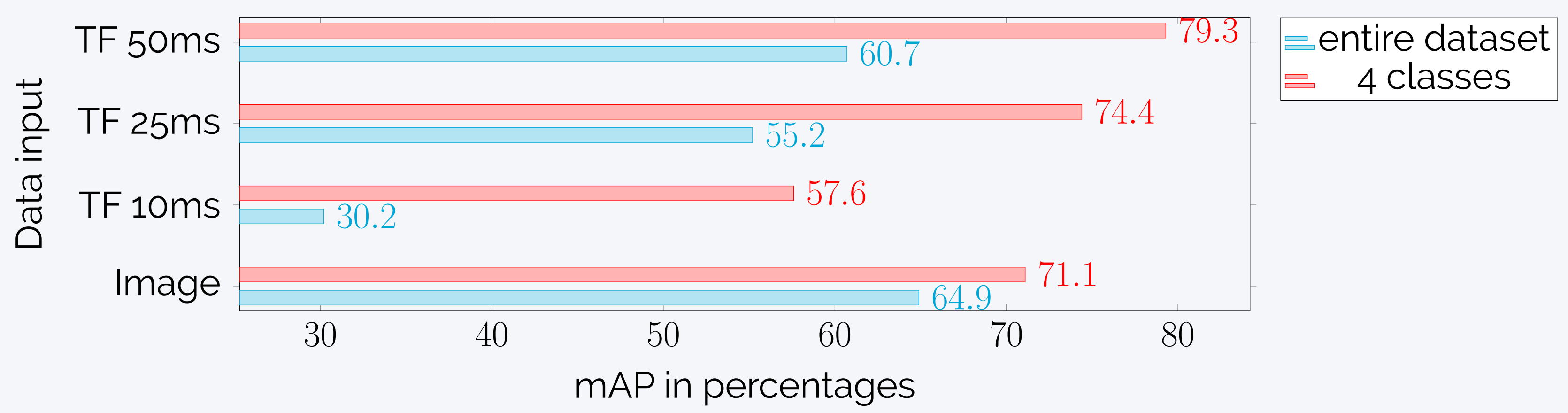


Figure 6. YOLOv3 accuracy results, TF = Time Frame of events

6. Conclusion

1. The accuracy-efficiency trade-off for using event-based data is: a small loss in accuracy and large gain in efficiency.
2. With the best model and the best event-based data representation the accuracy-efficiency trade-off can be even better.

7. Future work

- Use a model that fully exploits the sparsity of events to test whether the accuracy-efficiency trade-off can be improved, an example is [5].
- Find the best event-based data representation as input to a neural network.
- Test whether using events with color values can increase the accuracy for object detection.
- Use a more realistic event-based datasets like the Prophesee 1 megapixel automotive detection dataset [6].

8. References

[1] Qinyi Wang, Yexin Zhang, Junsong Yuan, and Yilong Lu. Space-time event clouds for gesture recognition: From rgb cameras to event cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1826–1835, 2019.

[2] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022.

[3] Garrick Orchard, Ajinkya Jayawant, Gregory K. Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9, 2015.

[4] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018. cite arxiv:1804.02767 Comment: Tech Report.

[5] Nico Messikommer, Daniel Gehrig, Antonio Loquercio, and Davide Scaramuzza. Event-based asynchronous sparse convolutional networks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 415–431, Cham, 2020. Springer International Publishing.

[6] Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *CoRR*, abs/2009.13436, 2020.

p.benschop@student.tudelft.nl