برچسب زنی سیگنال ویدیو حاصل از دوربین مبتنی بر رخداد
Labeling video signal obtained from an event camera

استاد راهنما : آقای دکتر سعید باقری شورکی

دانشجو : مهدی قاسم زاده

Outline

- Introduction
  - What is an event camera
  - How Event Based Cameras Works
  - Event Cameras vs Frame Cameras
  - Event Representation
- Related Work
  - EV-SegNet(2019) : Semantic Segmentation for Event-based Cameras(Alonso)
  - ESS(2022) : Learning Event-based Semantic Segmentation from Still Images
- Our work
  - Introducing new dataset
  - Introducing new Event based semantic segmentation Model
  - Introducing new Event-Frame based semantic segmentation Model
- Experiments and Results
- Conclusion and future work
- References

Introduction

What is an Event Camera?

Human Eyes

- The photoreceptors in our eyes only report back to the brain when they detect change in some feature of the visual scene.

Event Based Camera

- Novel bio-inspired sensors that capture motion in the scene
- Focusing Only On What Changes

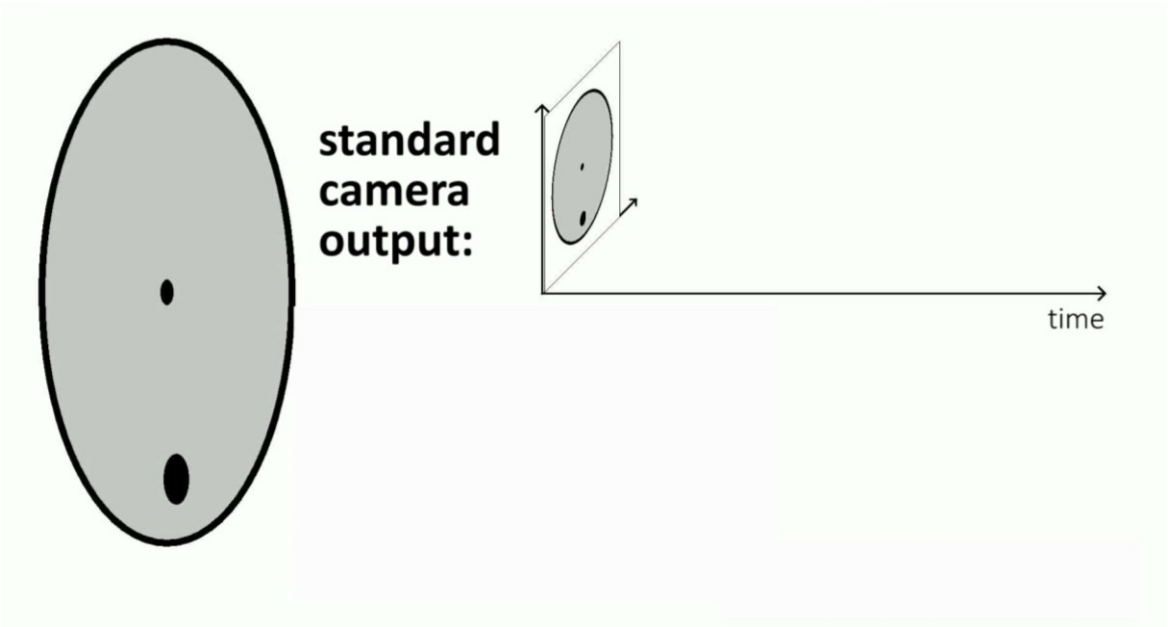
Introduction

- First commercialized in 2008 by T. Delbruck (UZH&ETH) under the name of Dynamic Vision Sensor (DVS)
- How Event-Based Cameras Works
  - Recording changes in the scene instead of recording the entire scene at regular intervals
  - Each pixel in an event-based sensor works independently
  - Data format of events
    - $e_k = (x_k, y_k, t_k, p_k)$

      e_0 = (35,66,1,0)

      e_1 = (100,27,12,1)

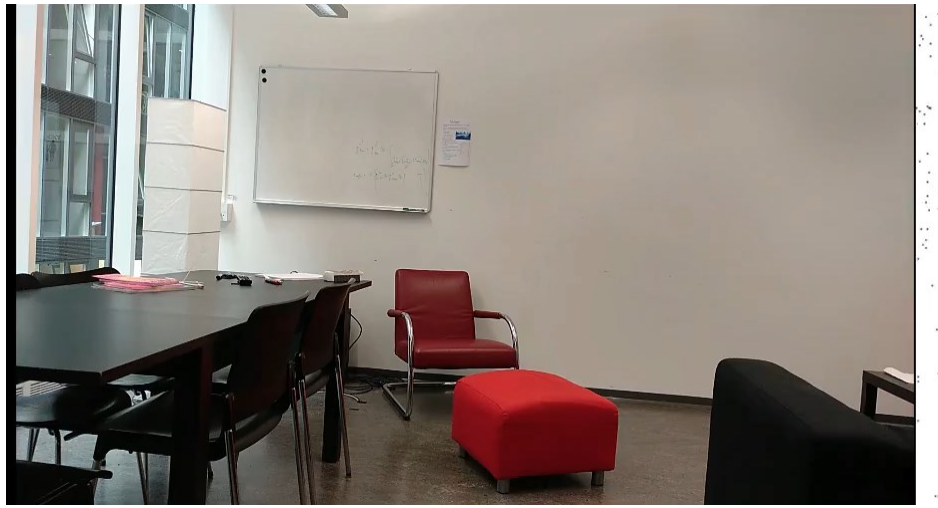
      e_2 = (79,51,12,0)



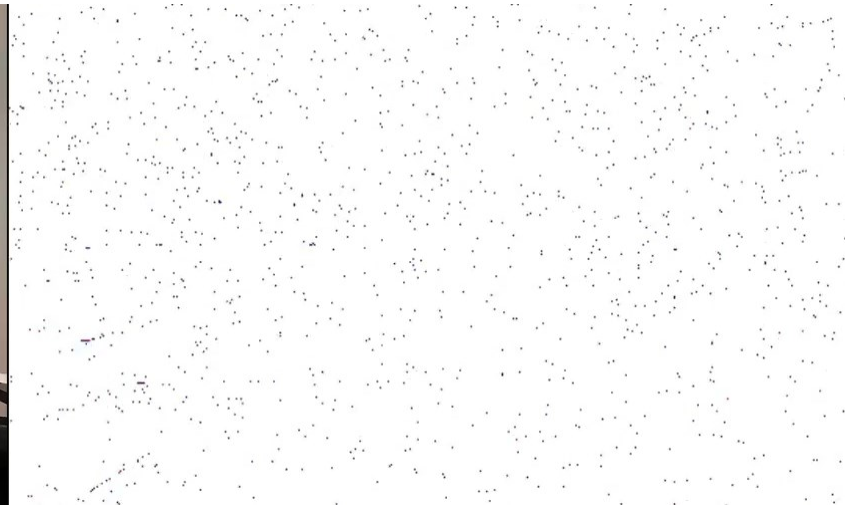
Introduction

How Event-Based Cameras Works

Frame Based Camera



Event Based Camera (ON , OFF events)



Introduction

Benefits

- Low latency (~ 1 microsecond)
- No motion blur
- High dynamic range (120 dB)
- Ultra-low power (1mW)

Challenges

- Data format of events

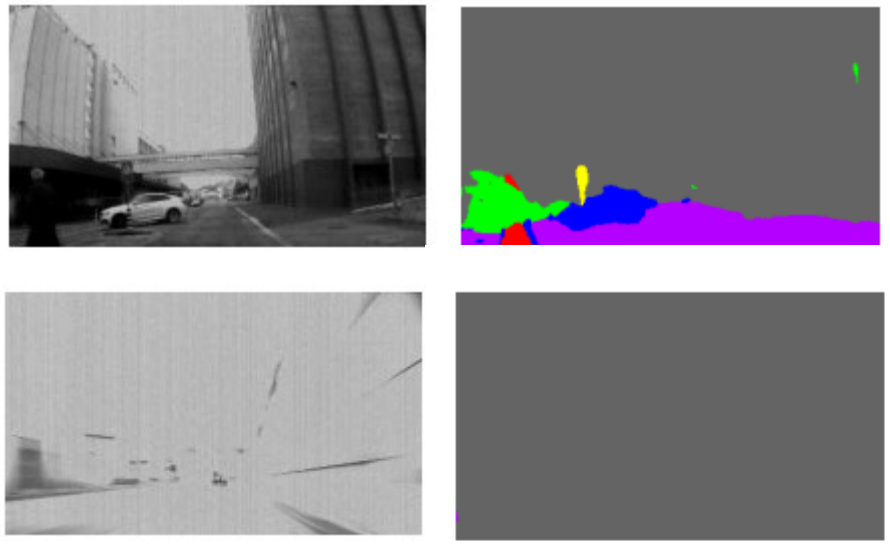
$$e_k = (x_k, y_k, t_k, p_k)$$

- Low resolution
- Dataset

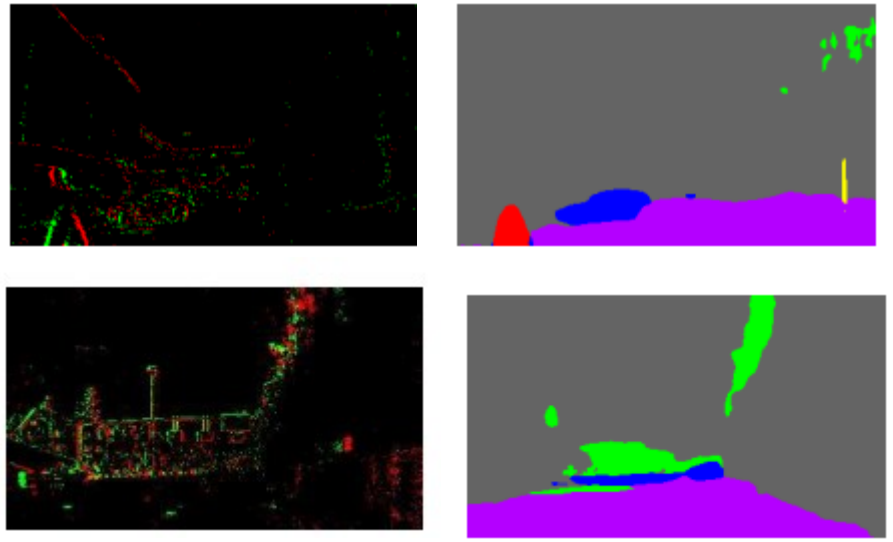
Introduction

Frame Based Cameras VS Event Based Cameras

- Motion blur
- Low dynamic range (60 dB)



- No motion blur
- High dynamic range (120 dB)



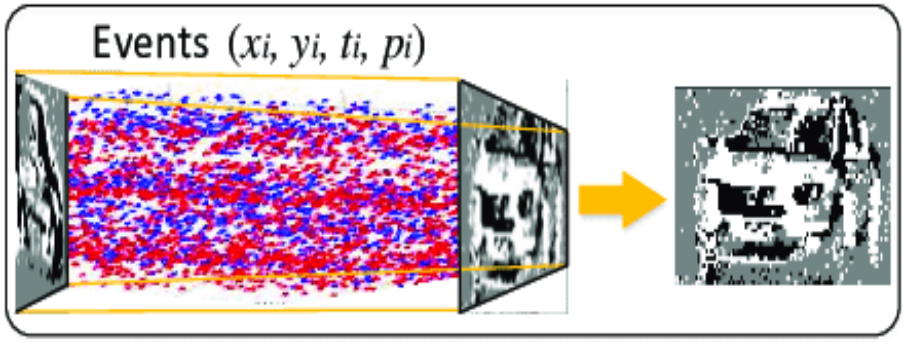
Introduction

Event Representation

- Event cameras are very different from conventional *RGB* cameras , Instead of encoding the appearance of the scene within three color channels, they only capture the changes in intensities for each pixel.
- The output of an event camera is not a 3-dimensional image (height, width and channels) it is a stream of events.

Event Representation methods

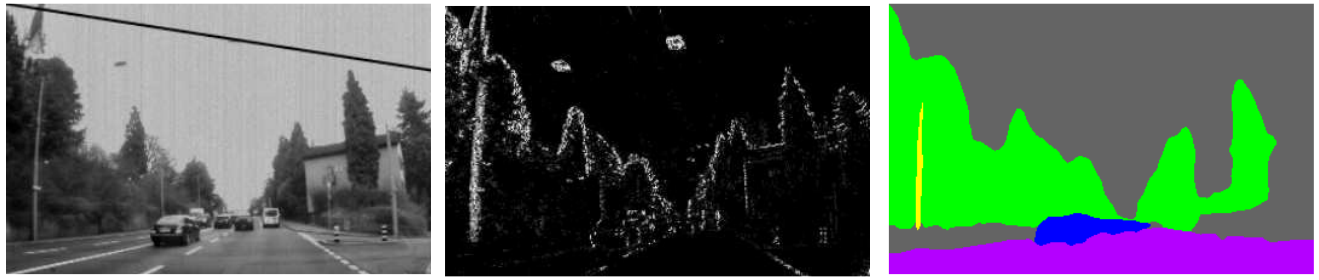
- Histogram
  - $e_k = (x_k, y_k, \boldsymbol{t_k}, p_k)$
- Timesurface
  - $e_k = (x_k, y_k, t_k, \boldsymbol{p_k})$
- Event Volume
  - $e_k = (x_k, y_k, t_k, p_k)$



Related Work

Event Based Semantic Segmentation Models

- EV-SegNet(2019): Semantic Segmentation for Event-based Cameras(Alonso)



- ESS(2022): Learning Event-based Semantic Segmentation from Still Images

EV-SegNet

Alonso Dataset

- Resolution : (200,346)
- 6 Classes
- Pseudo Label

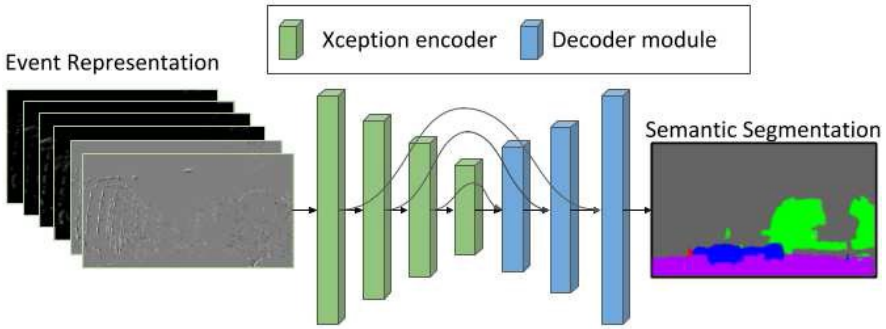
Train Images	15950
Test Images	3890

Class	Label
Flat	0
Construction , Sky , Ignore	1
Poles , Traffic lights	2
Nature	3
Human	4
Vehicle	5

EV-SegNet

Details

- Event Representation: 6 channels (Histogram , Timesurface)
- Encoder-Decoder Architecture
- Backbone: Xception



Results

Accuracy	89.76
MIoU	54.81

ESS

ESS Dataset

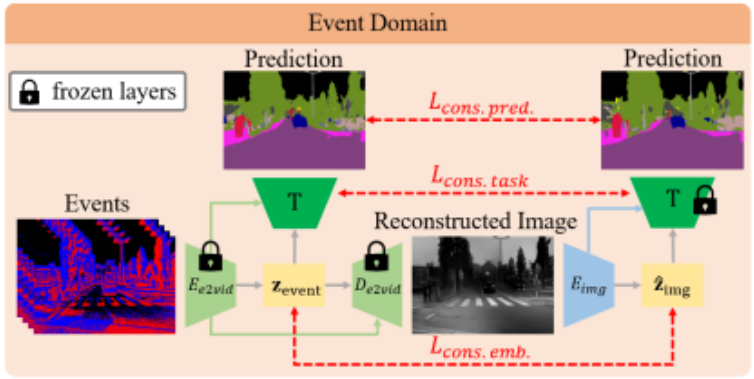
- Resolution (480,640)
- 11 Classes
- Pseudo Label

Class	Label
Background	0
Building	1
Fence	2
Person	3
Pole	4
Road	5
Sidewalk	6
Vegetation	7
Car	8
Wall	9
Traffic sign	10

ESS

Details

- Introducing unsupervised and supervised methods
- Event Representation: 5 channels (Event Volume)
- Encoder-Decoder and recurrent Architecture
- Backbone: ResNet18



Results

- Alonso Dataset

Method	Training Data	Accuracy	MIoU
unsupervised	Event	87.86	52.46
supervised	Event	91.08	61.37
supervised	Event + Frame	90.37	60.43

- ESS Dataset

Method	Training Data	Accuracy	MIoU
unsupervised	Event	84.02	44.87
supervised	Event	89.25	51.57
supervised	Event + Frame	89.37	53.29

Our works

- Introducing new dataset with more classes and exact label
- Introducing new Event based semantic segmentation Model
  - Evaluation of our model on our dataset and Alonso dataset
- Introducing new Event-Frame based semantic segmentation Model
  - Introducing new training method for achieving robustness and high accuracy
  - Evaluation of our model on our dataset and Alonso dataset

Our dataset

Our dataset was produced using autonomous driving simulator

- Autonomous driving simulator
  - Recent advancements in computer graphics technology allow more realistic rendering of car driving environments. They have enabled self-driving car simulators such as DeepGTA-V and CARLA (Car Learning to Act) to generate large amounts of synthetic data that can complement the existing real world dataset in training autonomous car perception.

Our dataset

Steps of producing dataset

- We produced a large dataset using Carla , this dataset contains Events , RGB image and semantic segmentation labels
- We added EventScape dataset (1) to our dataset

Final dataset:

	Town(s)	Number of sequences	Number of samples
Train data	Town [1,2,3,4,6,7]	600	130000
Test data	Town 5	100	30000

(1) https://arxiv.org/pdf/2102.09320

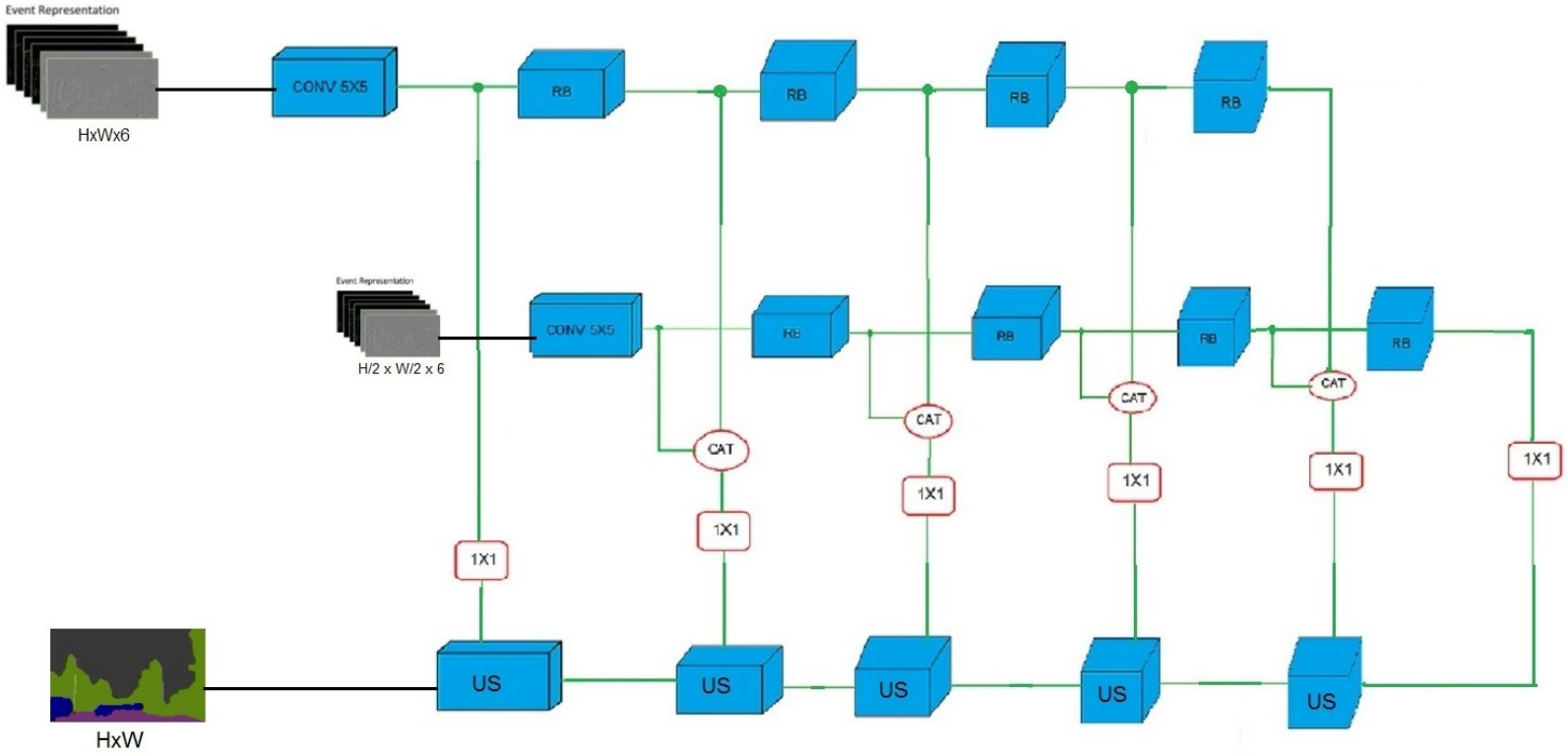
Our dataset

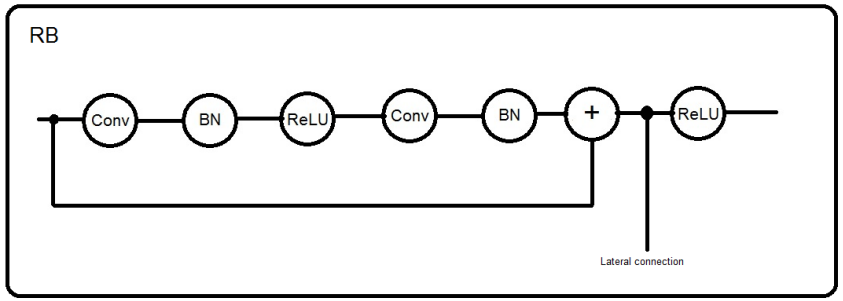
Two sequences from our dataset

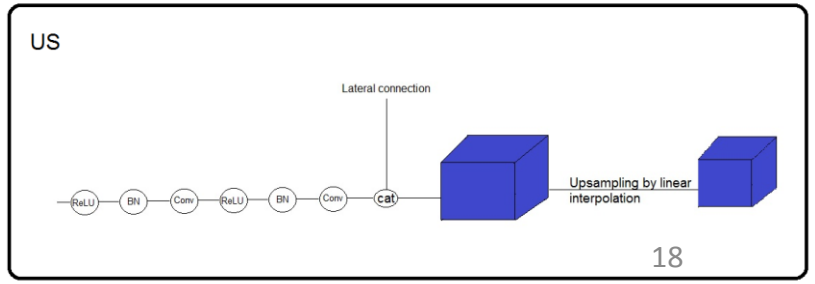
Our Event Based Model

Encoder-Decoder Architecture

Backbone Resnet 18





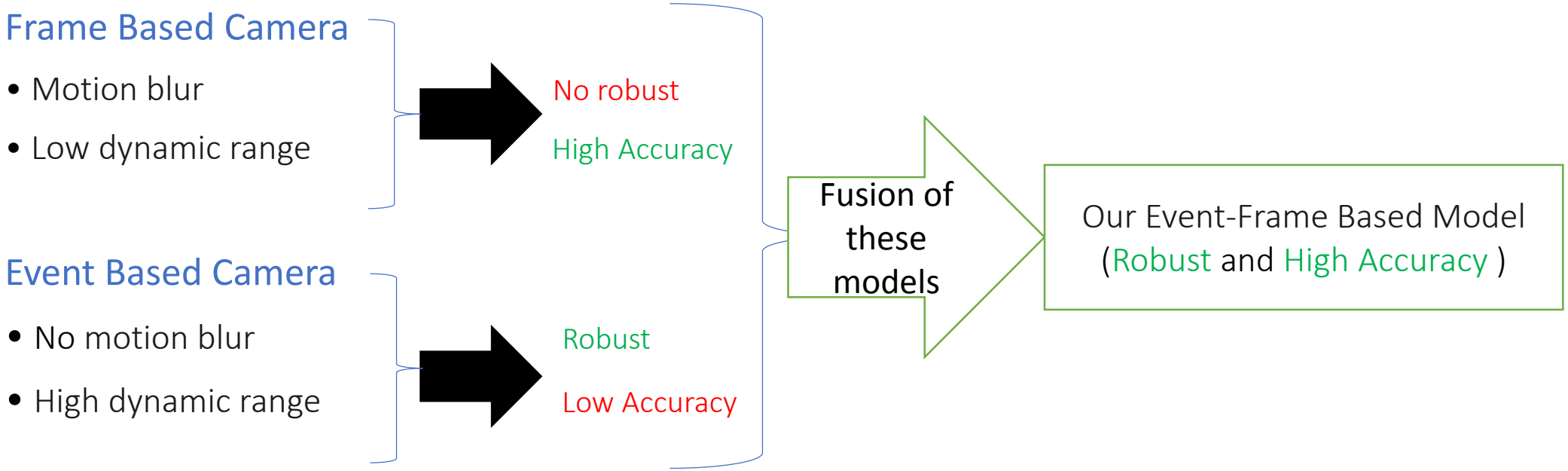


Our Event Based Model

Number of the parameters

Model	Number of parameters
Event-based Segmentation (R18 light)	11M
Event-based Segmentation (R18 original)	25M

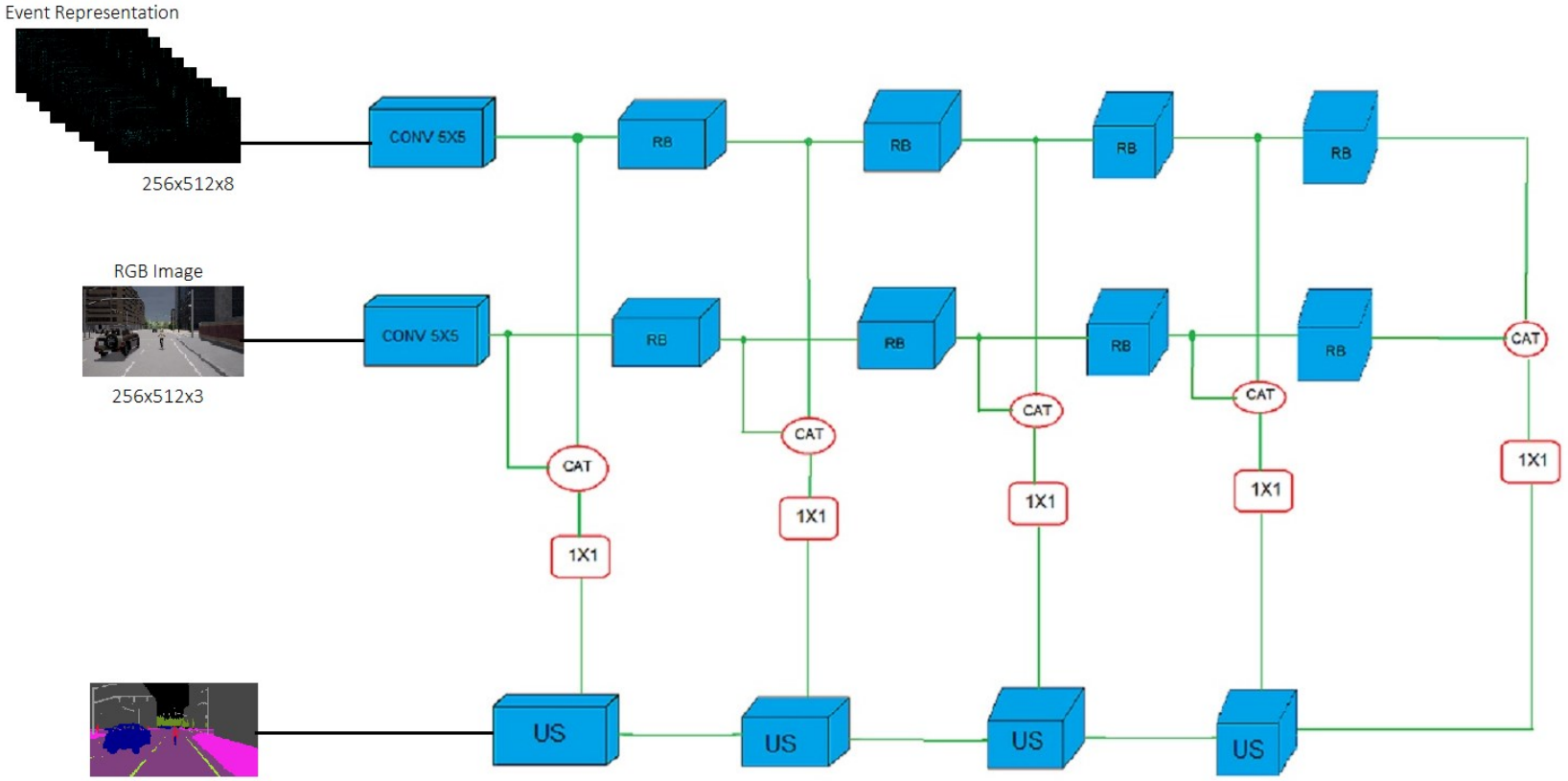
Our Event-Frame Based Model

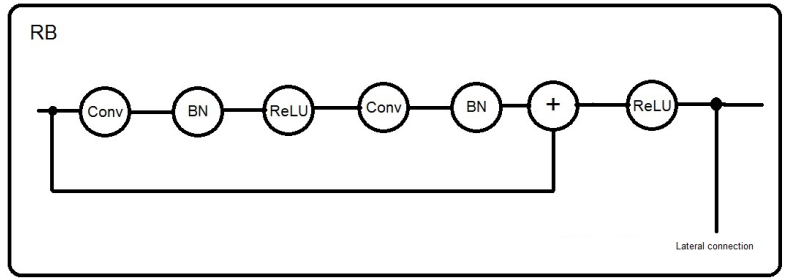


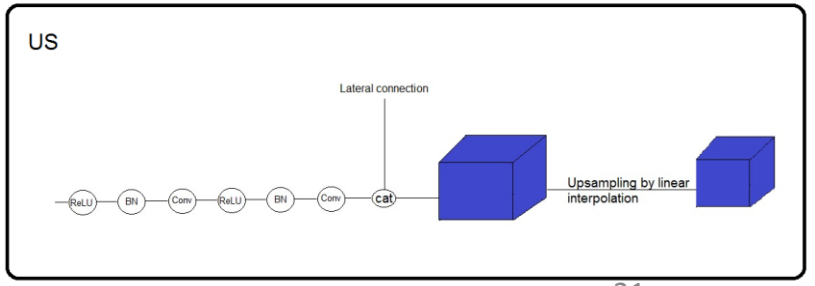
Our Event-Frame Based Model

Encoder-Decoder Architecture

Backbone Resnet 18







Our Event-Frame Based Model

Our Training Technique

- We used blurred image for achieving robustness (blurring module)

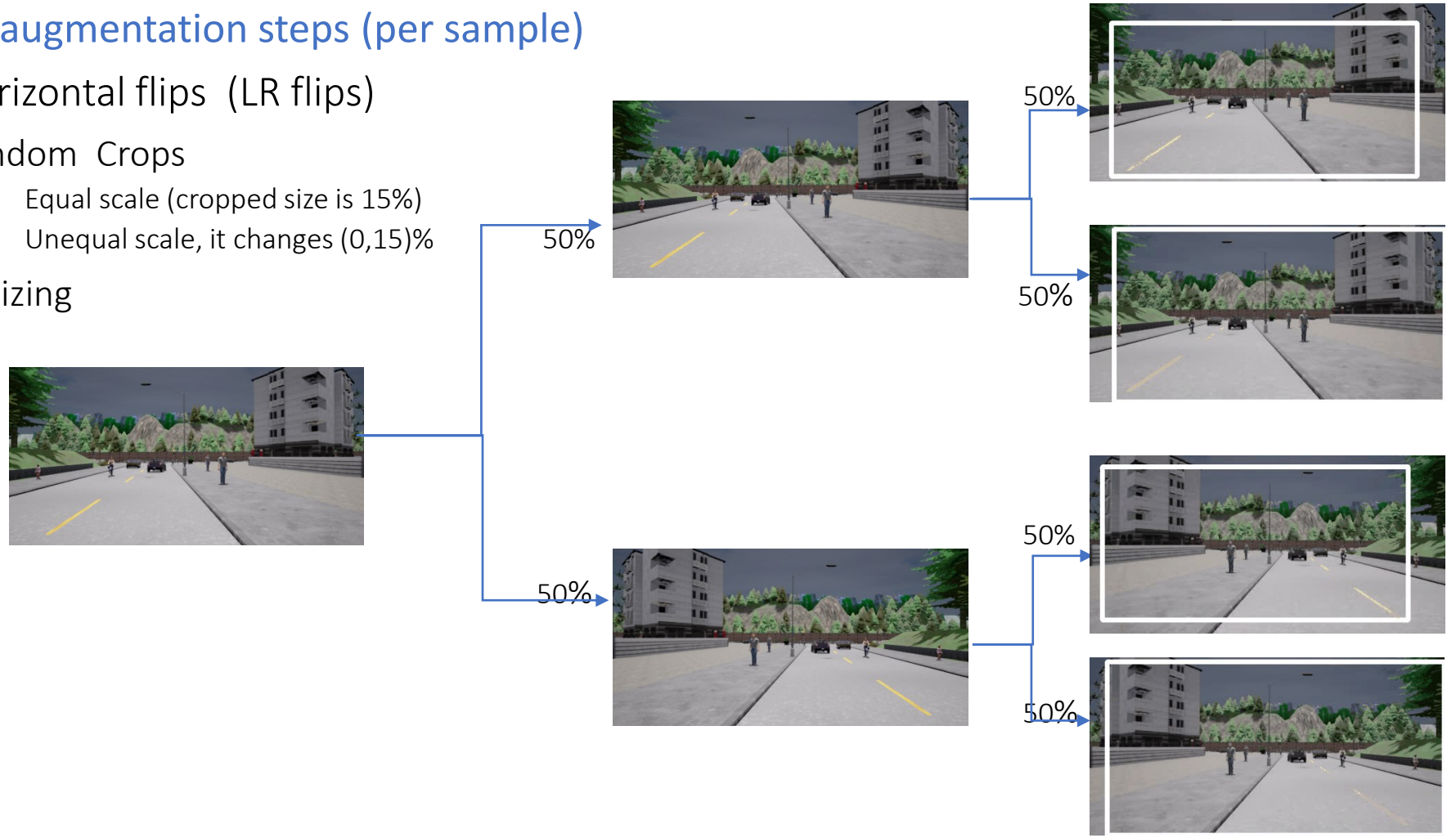
Blurring Module

- With probability of 50% , a gaussian filter is used for image blurring and kernel size is chosen randomly between 3 and 255

Data Augmentation

Data augmentation steps (per sample)

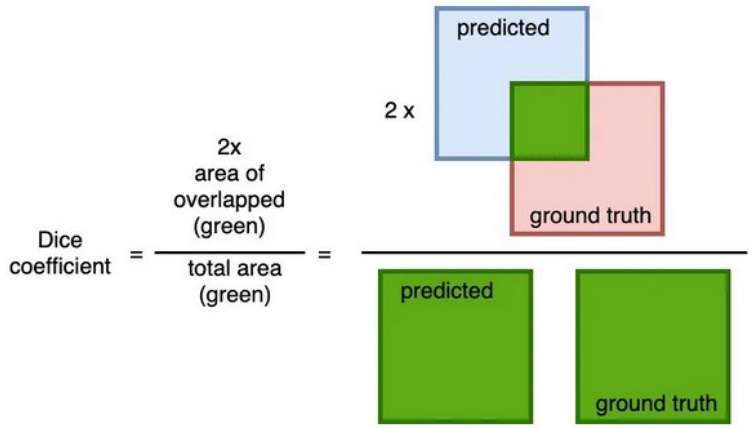
- Horizontal flips (LR flips)
- Random Crops
  - Equal scale (cropped size is 15%)
  - Unequal scale, it changes (0,15)%
- Resizing



Loss Function

Loss Function

- Loss = CrossEntropy + DiceLoss
- DiceLoss = 1 - Dice



Experiments on Alonso Dataset

Our Event Based Model

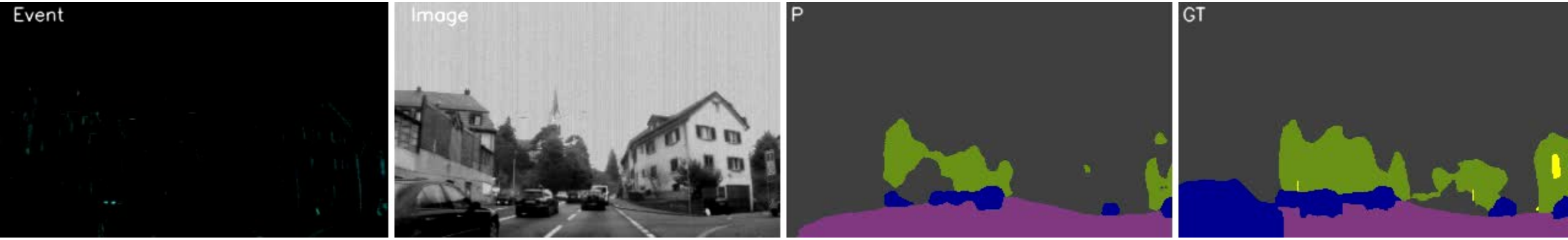
Training Details

- Event Representation : Alonso methods
- Event resolution : (200,346)
- Batch size : 16
- Learning Rate : 3e-4 ➡ 1e-5
- Epoch : 80 , iteration : 80k
- Training time : 20h on GPU T4 , Colab pro
- Prediction time on GPU mx130
  - Light model : 18ms
  - Original model : 22ms

Experiments on Alonso Dataset

Qualitative Results

Our Event Based Model



Experiments on Alonso Dataset

Our Event-Frame Based Model

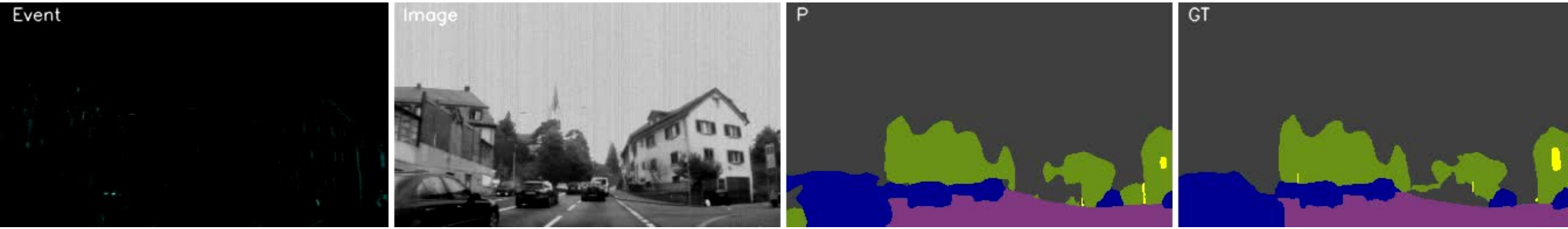
Training Details

- Event Representation : Alonso methods
- Event resolution : (200,346)
- Batch size : 8
- Learning Rate : 5e-4 ➡ 1e-5
- Epoch : 80 , iteration : 160k
- Training time : 40h on GPU T4 , Colab pro
- Prediction time on GPU mx130: 23ms

Experiments on Alonso Dataset

Qualitative Results

Our Event-Frame Based Model



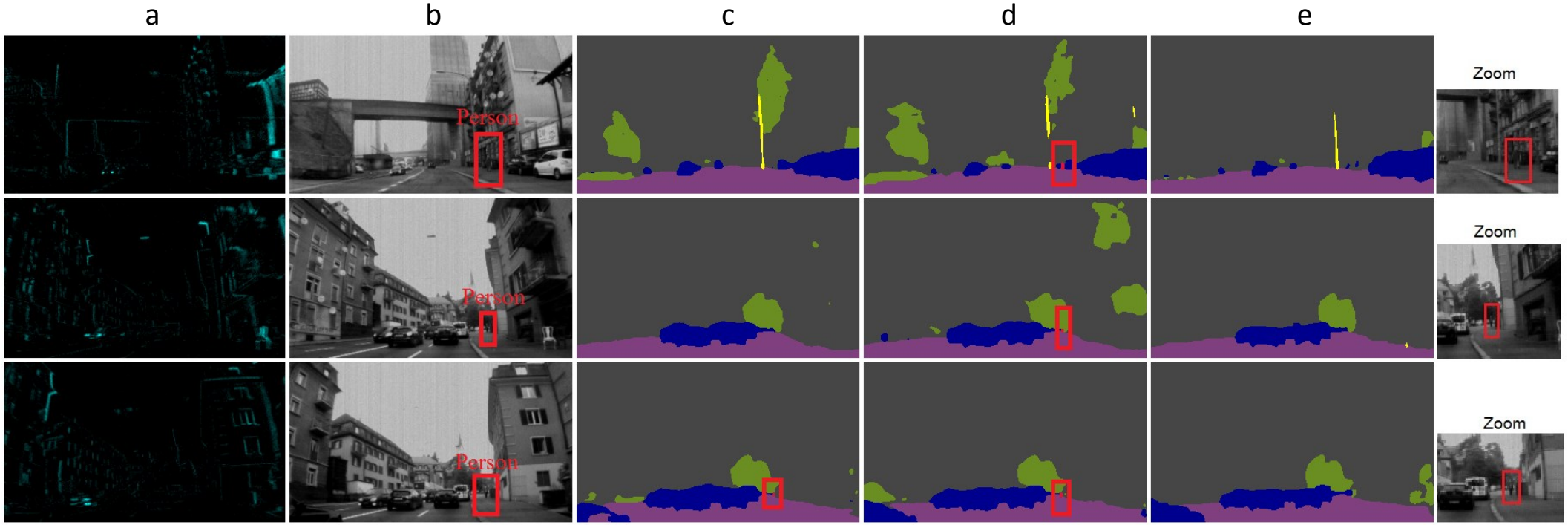
Experiments on Alonso Dataset

Results

Model	Input Data	Accuracy	MIoU	Parameters (M)
EV-Segnet [1]	Event	89.76	54.81	29
EV-Segnet [1]	Event + Frame	95.22	**68.36**	29
Vid2E [6]	Event	-	45.48	-
EvDistill (2ch) [7]	Event	-	57.16	59
EvDistill (Mch) [7]	Event	-	58.02	59
DTL [9]	Event	-	58.80	60
ESS [10]	Event	91.08	**61.37**	12
ESS [10]	Event + Frame	90.37	60.43	12
HALSIE [16]	Event + Frame	92.50	60.66	1.82
Our light Event network	Event	88.21	59.28	11
Our original Event network	Event	89.16	**61.11**	25
Our Event-Frame network without blurring module	Event + Frame	95.72	**72.20**	26
Our Event-Frame network with blurring module	Event + Frame	94.76	**70.50**	26

Experiments on Alonso Dataset

Impact of blurring module on the performance



a: events, b: images, c: the network without blurring module, d: the network with blurring module, e: Ground Truth

Experiments on Our Dataset

Dataset

We use only $\frac{1}{7}$ of dataset for training and test our models

- 17700 samples for training
- 3777 samples for test

- Segmentation classes for Event Model

Sky, Unlabeled	0
Buildings, Fences, Walls	1
Pedestrians	2
Poles, Traffic Signs	3
Roads	4
Sidewalks	5
Vegetation	6
Vehicles	7

- Segmentation classes for Event-Frame Model

Sky, Unlabeled	0
Buildings, Fences, Walls	1
Pedestrians	2
Poles, Traffic Signs	3
Roads	4
Sidewalks	5
Vegetation	6
Vehicles	7
Road Lines	8

Experiments on Our Dataset

Our Event Based Model

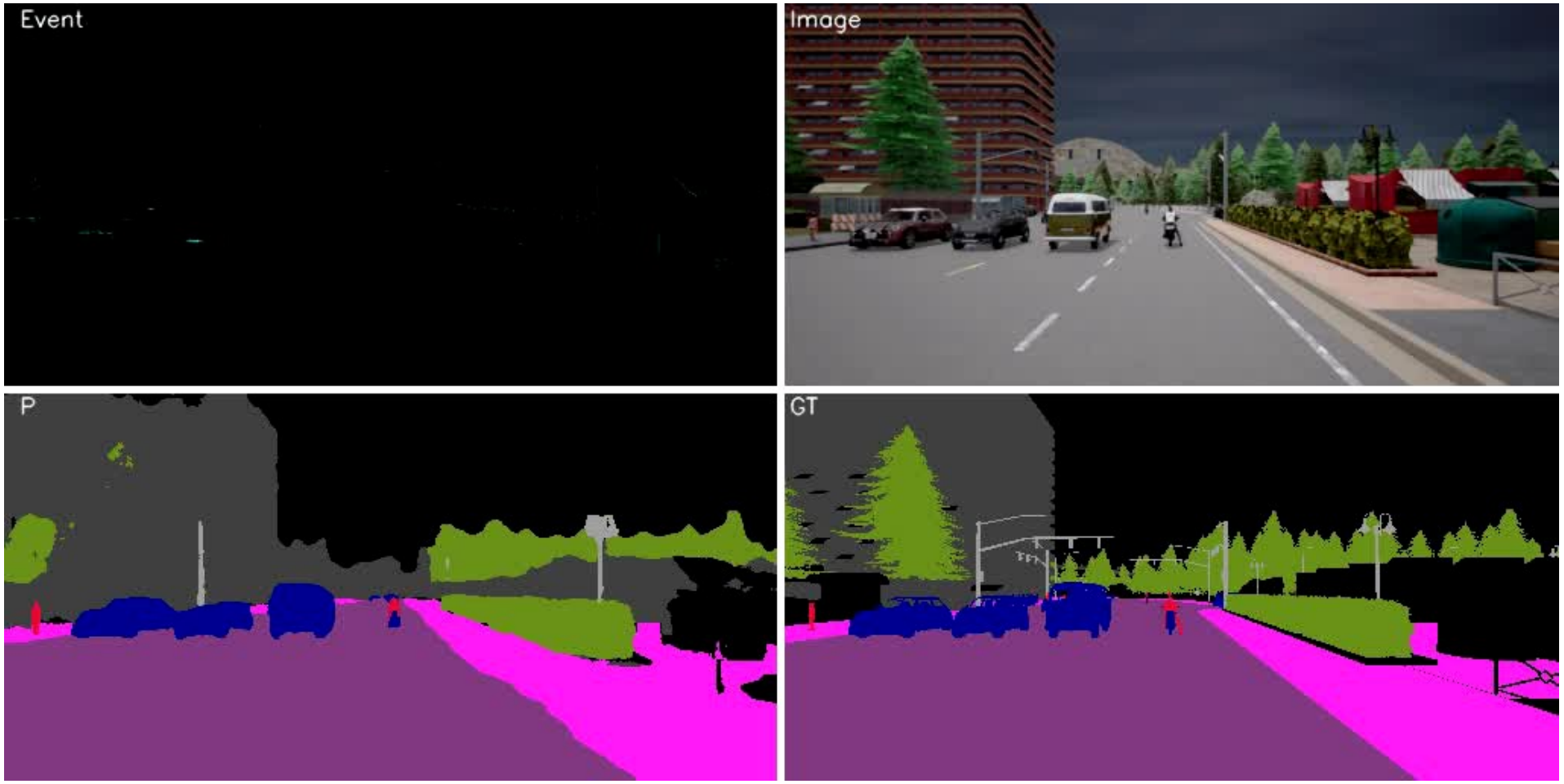
Training Details

- Event Representation: Histogram (2ch) , Timesurface(mean , recent) (4ch) , Event Volume (2ch)
- Event resolution: (256,512)
- Batch size : 8
- Learning Rate : 5e-4 ➡ 1e-5
- Epoch : 60 , iteration : 132k
- Training time : 60h on GPU T4 , Colab pro
- Prediction time on GPU mx130 : 24ms

Experiments on Our Dataset

Qualitative Results

Our Event Based Model



Experiments on Our Dataset

Our Event-Frame Based Model

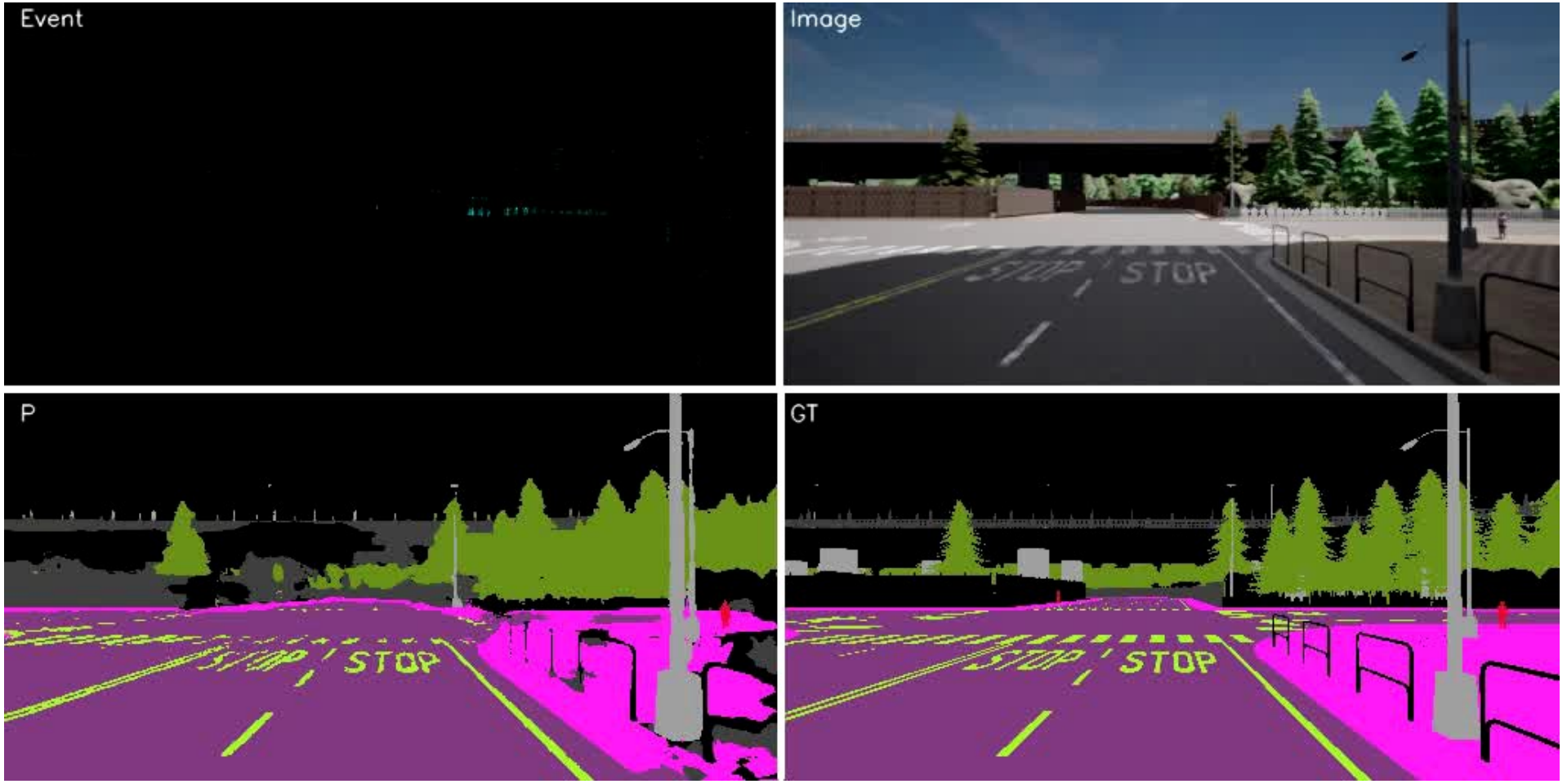
Training Details

- Event Representation : Histogram (2ch) , Timesurface(mean , recent) (4ch) , Event Volume (2ch)
- Resolution : (256,512)
- Batch size : 8
- Learning Rate : 5e-4 ➡ 1e-5
- Epoch : 60 , iteration : 132k
- Training time : 80h on GPU T4 , Colab pro
- Prediction time on GPU mx130 : 26ms

Experiments on Our Dataset

Qualitative Results

Our Event-Frame Based Model



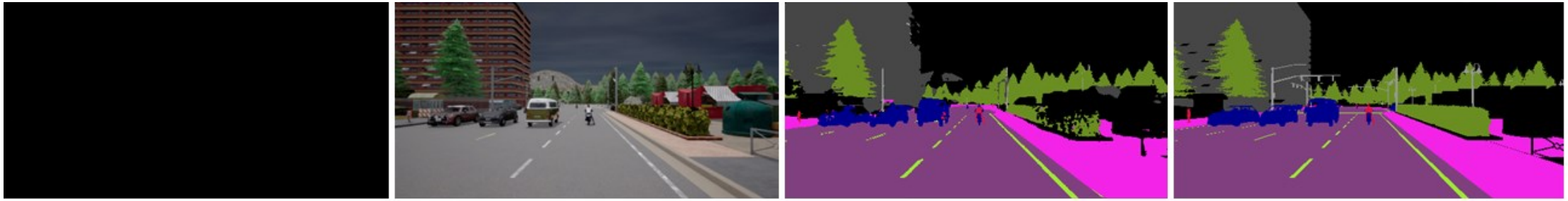
Experiments on Our Dataset

Robustness test for blurred images

Experiments on Our Dataset

Robustness test when Event Camera does not work!

events | image | prediction | Ground Truth



Experiments on Our Dataset

Results

Model	Input Data	Accuracy	MIoU	Parameters (M)
Our Event network	Event	84.96	52.74	25
Our Event-Frame network	Event + Frame	90.05	65.23	26

Robustness Test

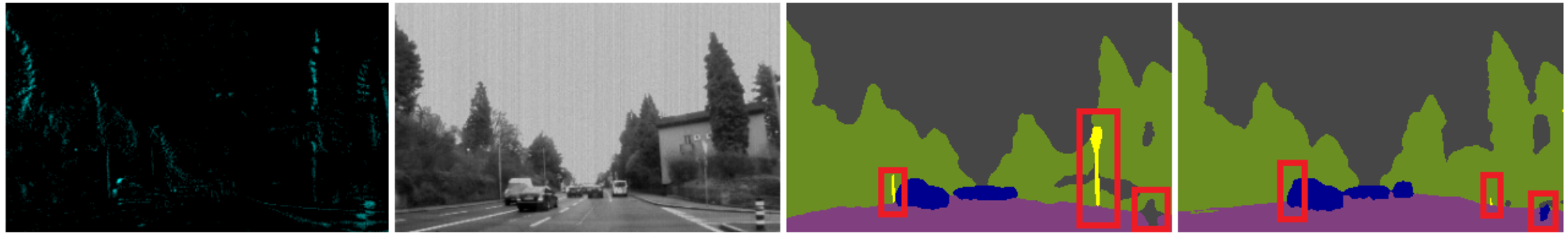
- For demonstrating the robust performance of the Event-Frame network, all test images are blurred by a Gaussian filter with a kernel size of k then they are applied to the network.

Model	Images are blurred with a kernel size of (k)	Using events	Accuracy	MIoU
Our Event-Frame network	K= 1	Yes	90.05	65.23
Our Event-Frame network	K= 11	Yes	88.70	62.06
Our Event-Frame network	K= 55	Yes	86.64	56.77
Our Event-Frame network	K= 111	Yes	85.57	54.50
Our Event-Frame network	K= 255	Yes	84.42	52.29
Our Event-Frame network	K= 1	No	82.25	53.90

Conclusion and future work

- Small objects in DDD17 dataset are sometimes missed, red rectangles show an example of this problem, low quality of DDD17 dataset and pseudo labels cause that. Therefore reported results (accuracy and MIoU) could not be precise and they could be a bit higher or lower.

Events *Images* *our event network* *Ground Truth*



Conclusion and future work

- In this work, we introduced an event-based network for semantic segmentation and we evaluated our network on (our dataset + Event-Scape dataset [5]) our experiments showed that event-based network has good performance in recognizing some classes such as pedestrians and cars, but in recognizing some classes such as road is not accurate as well as frame-based models. For improving the performance of the event-based model, we proposed to use events along with images in an event-frame-based semantic segmentation network.
- Our experiments revealed if common training methods are used for training event-frame-based networks, events data are just used for boosting accuracy, and expected robustness will not be obtained due to, we proposed to use the blurring module for achieving robustness to blurred images and experiments showed our method is efficient and the networks with it could be robust and accurate and also more reliable compared to the event-frame based network without the blurring module.

Conclusion and future work

- We consider the most common failure in images indeed image blurring, for future works other types of failure in images should be considered for achieving more effective robustness.
- Other types of the fusion methods and neural networks could be applied and evaluated in the future.

References

[1][Alonso, I., Murillo, A.C.: EV-SegNet: Semantic segmentation for event-based cameras. In: IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW) (2019)]

[2] [J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, "Ddd17:End-to-end davis driving dataset," *arXiv preprintarXiv:1711.01458*, 2017]

[3] [Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A 240x180 130dB 3μs latency global shutter spatiotemporal vision sensor. IEEE J. Solid-State Circuits **49**(10), 2333–2341 (2014). https://doi.org/10.1109/JSSC.2014.2342715]

[4] [Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 1800–1807 (2017). https://doi.org/10.1109/CVPR.2017.195]

[5] [Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. In: IEEE Robotics and Automation Letters (2021). https://doi.org/10.1109/LRA.2021.3068942]

[6] [D. Gehrig, M. Gehrig, J. Hidalgo-Carrio, and D. Scaramuzza, "Video to events: Recycling video datasets for event cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3586–3595,2020]

[7] [Wang, L., Chae, Y., Yoon, S.H., Kim, T.K., Yoon, K.J.: Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR) (2021)]

[8] [Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR) (2016)]

[9] [Wang, L., Chae, Y., Yoon, K.J.: Dual transfer learning for event-based end-task prediction via pluggable event to image translation. In: Int. Conf. Comput. Vis.(ICCV). pp. 2135–2145 (2021)]

[10] [Z. Sun, N. Messikommer, D. Gehrig, and D. Scaramuzza,"Ess: Learning event-based semantic segmentation from stillimages," arXiv preprint arXiv:2203.10016, 2022]

[11] [Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: Conf. on Robotics Learning (CoRL) (2017)]

[12] [Diederik Paul Moeys, Federico Corradi, Emmett Kerr, Philip Vance, Gautham Das, Daniel Neil, Dermot Kerr, and Tobi Delbrück. Steering a predator robot using a mixed frame/event-driven convolutional neural network. In 2016 Second International Conference on Event-based Control, Communication, and SignalProcessing (EBCCSP), pages 1–8. IEEE, 2016.]

References

[13] [Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. 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, pages 5419–5427, 2018.]

[14][X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B.Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. IEEE transactions on pattern analysis and machine intelligence, 39(7):1346–1359, 2017.]

[15] [Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event based learning of optical flow, depth, and egomotion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 989–997, 2019.]

[16] [Shristi Das Biswas, Adarsh Kosta, Chamika Liyanagedera, Marco Apolinario and Kaushik Roy. HALSIE – Hybrid Approach to Learning Segmentation by Simultaneously Exploiting Image and Event Modalities.https://arxiv.org/abs/2211.10754, 2022]

[17] [Event-based Vision: A Survey Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Joerg Conradt, Kostas Daniilidis, Davide Scaramuzza]

[18] [Bardow, P., Davison, A.J., Leutenegger, S.: Simultaneous optical flow and intensity estimation from an event camera. In: IEEE Conf. Comput. Vis. Pattern Recog.(CVPR). pp. 884–892 (2016). https://doi.org/10.1109/CVPR.2016.102]

[19] [Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: IEEE Conf. Comput. Vis. Pattern Recog.(CVPR) (2019)]

[20] [Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. IEEE Trans. Pattern Anal. Mach. Intell. (2019).https://doi.org/10.1109/TPAMI.2019.2963386]

[21] [Reinbacher, C., Graber, G., Pock, T.: Real-time intensity-image reconstruction for event cameras using manifold regularisation. In: British Mach. Vis. Conf. (BMVC) (2016). https://doi.org/10.5244/C.30.9]

[22] [Gehrig, D., R¨uegg, M., Gehrig, M., Hidalgo-Carrio, J., Scaramuzza, D.: Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. IEEE Robotic an Automation Letters. (RA-L) (2021)]

[23] [Marin Oršić, Ivan Krešo, Petra Bevandić, Siniša Šegvić : In Defense of Pre-trained ImageNet Architectures for Real-time Semantic Segmentation of Road-driving Images]

[24] [He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 770–778 (2016).https://doi.org/10.1109/cvpr.2016.90]


THANK
YOU !