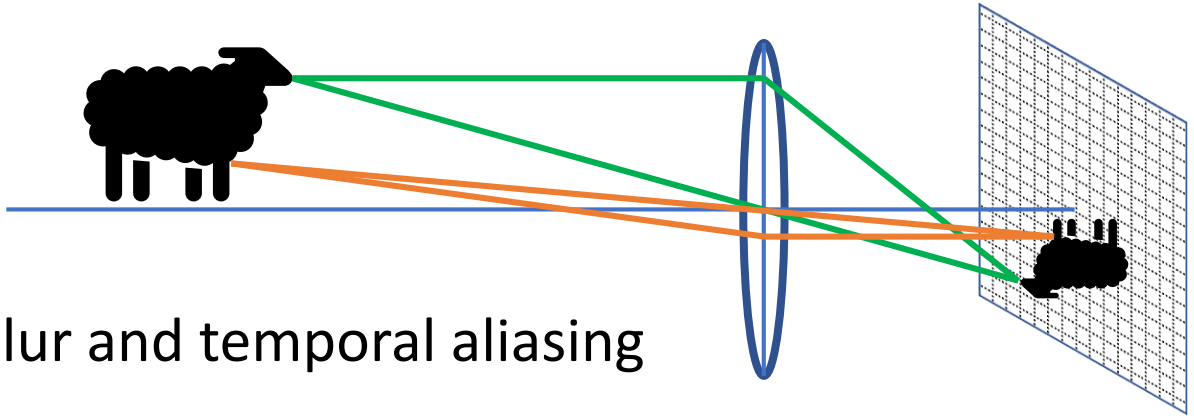
Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

Unsupervised Learning of Optical Flow and Camera Motion from Event Data

Alex Zihao Zhu
Kostas Daniilidis
University of Pennsylvania

Traditional Cameras

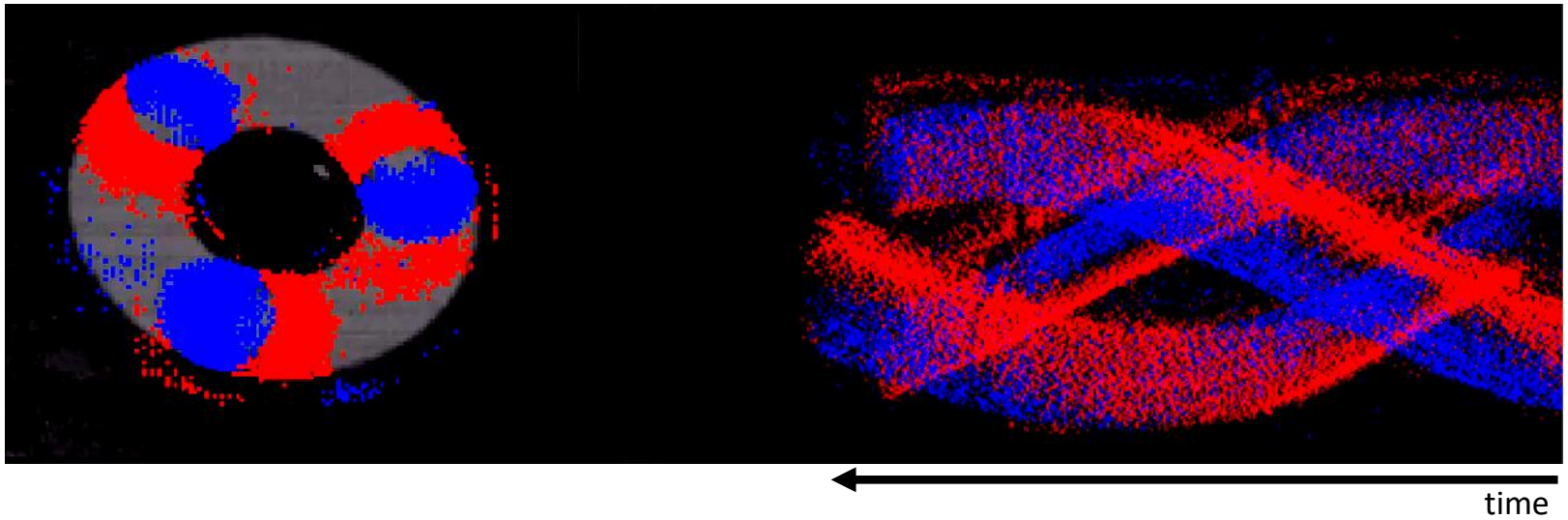
- Traditional cameras were designed for humans, not machines
- Images are generated with **fixed exposure times** and measure **absolute intensity**
  - Spatial and temporal relationships can be disentangled
  - Intensity values provide useful information for data association
  - Dynamic range is limited
  - Cameras are blind in the time between frames
  - Images are susceptible to motion blur and temporal aliasing



Event Cameras

Novel asynchronous sensor that tracks changes in log light intensity.

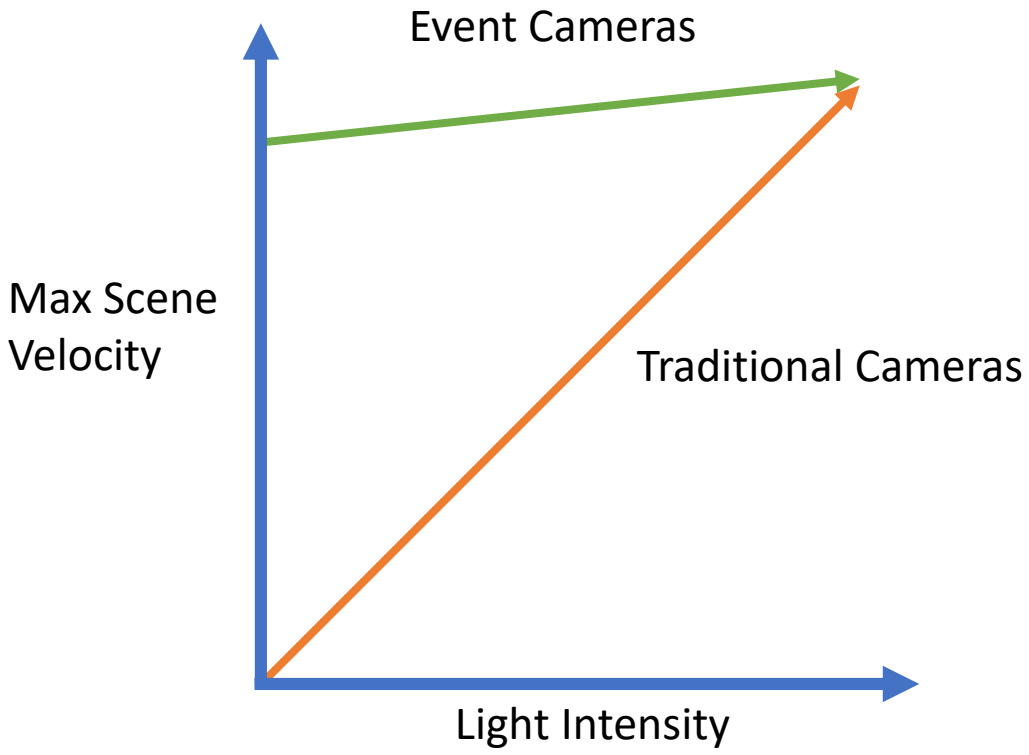
$$e_i = \{x_i, y_i, t_i, p_i\}$$
$$\left|\log\left(I_{t_i}(x,y)\right) - \log\left(I_{t_{i-1}}(x,y)\right)\right| > \theta$$



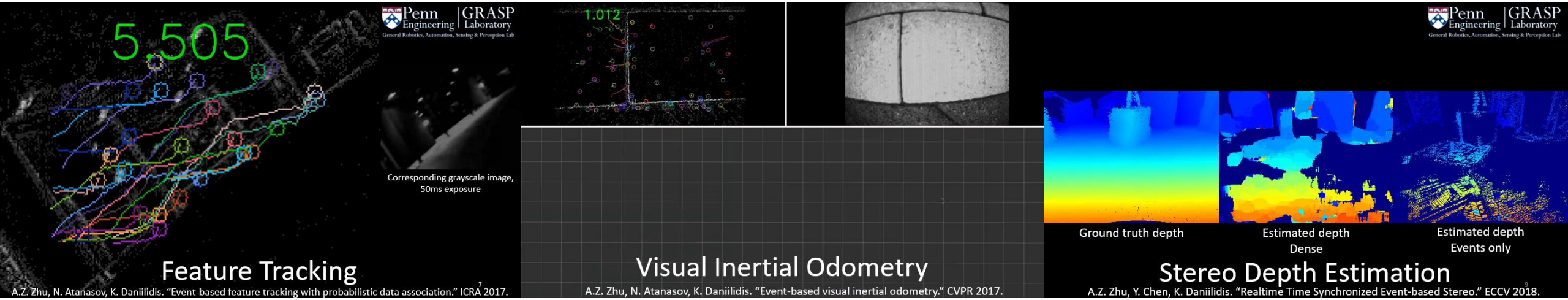
Positive events
Negative events

Benefits

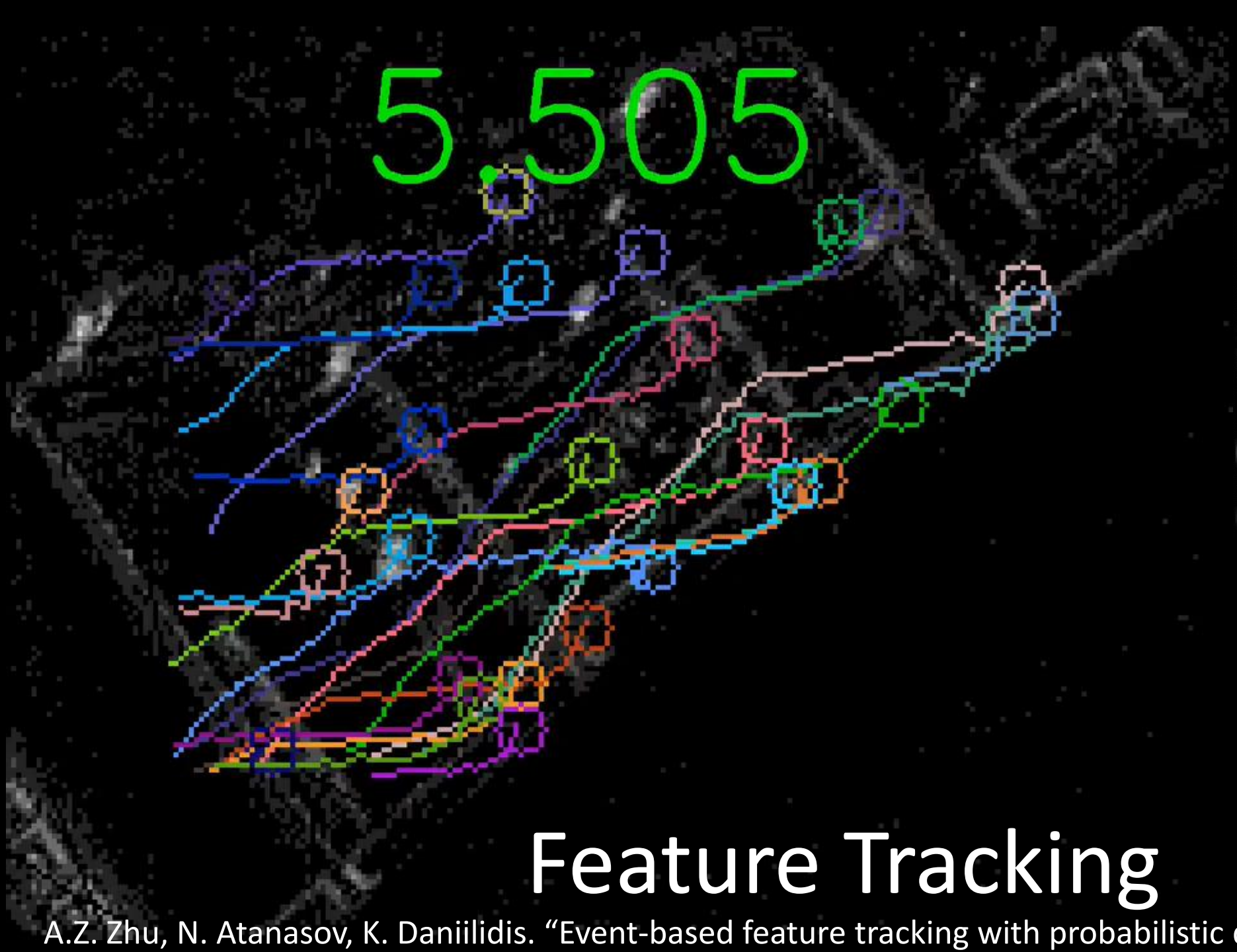
- Low latency
  - Allows tracking of very fast motions
- High dynamic range
  - Excellent low/challenging light performance
- Low power consumption



Prior Work



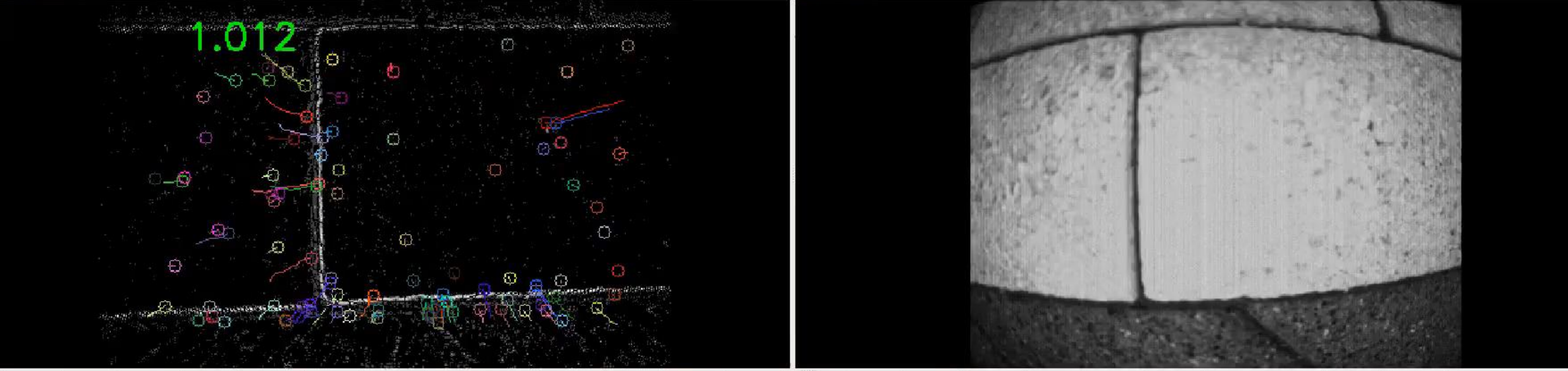
A.Z. Zhu, N. Atanasov, K. Daniilidis. "Event-based feature tracking with probabilistic data association." ICRA 2017.
A.Z. Zhu, N. Atanasov, K. Daniilidis. "Event-based visual inertial odometry." CVPR 2017.
A.Z. Zhu, Y. Chen, K. Daniilidis. "Realtime Time Synchronized Event-based Stereo." ECCV 2018.



Corresponding grayscale image, 50ms exposure

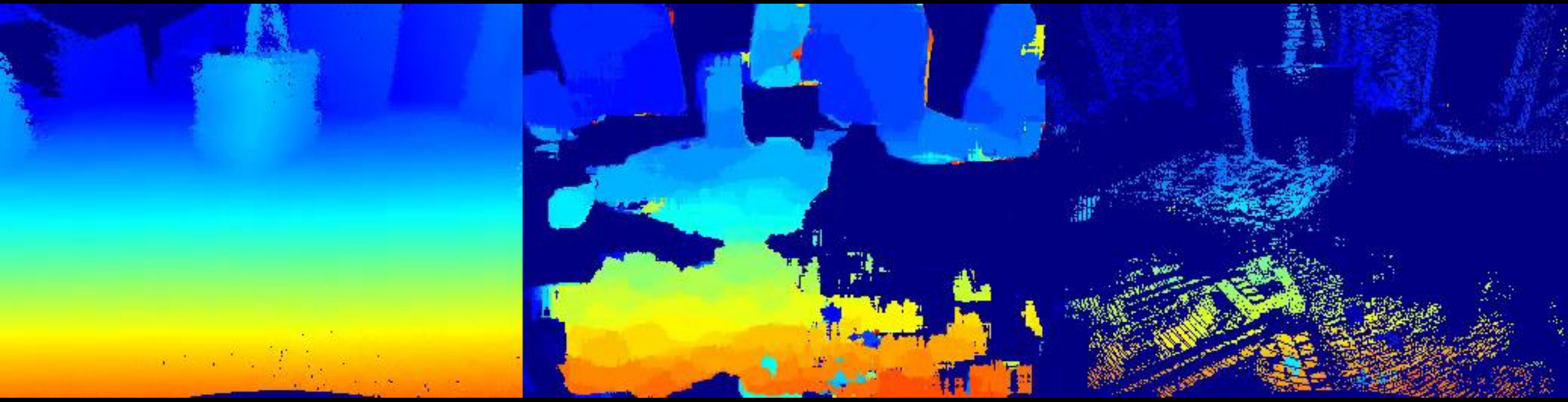
Feature Tracking

A.Z. Zhu, N. Atanasov, K. Daniilidis. “Event-based feature tracking with probabilistic data association.” ICRA 2017.



Visual Inertial Odometry

A.Z. Zhu, N. Atanasov, K. Daniilidis. “Event-based visual inertial odometry.” CVPR 2017.



Ground truth depth

Estimated depth
Dense

Estimated depth
Events only

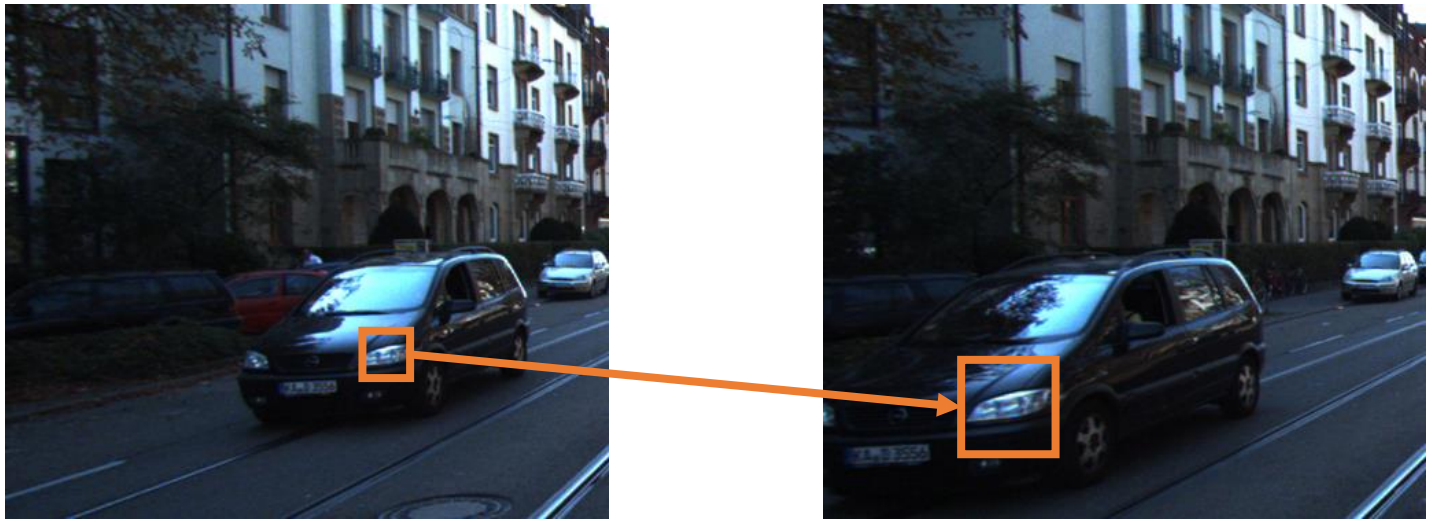
Stereo Depth Estimation

A.Z. Zhu, Y. Chen, K. Daniilidis. “Realtime Time Synchronized Event-based Stereo.” ECCV 2018.

Algorithm Development is Hard!

Have to develop new models for events. We need:

- Substitute for photometric loss
- Model of event noise.
- Solve complex optimization problems.



Deep Learning for Events

Neural networks allow us to solve complex nonlinear problems.

They can also learn the underlying noise models.

But: getting data is hard.

Self-supervised learning can provide the power of neural networks without the need for expensive labeled data.

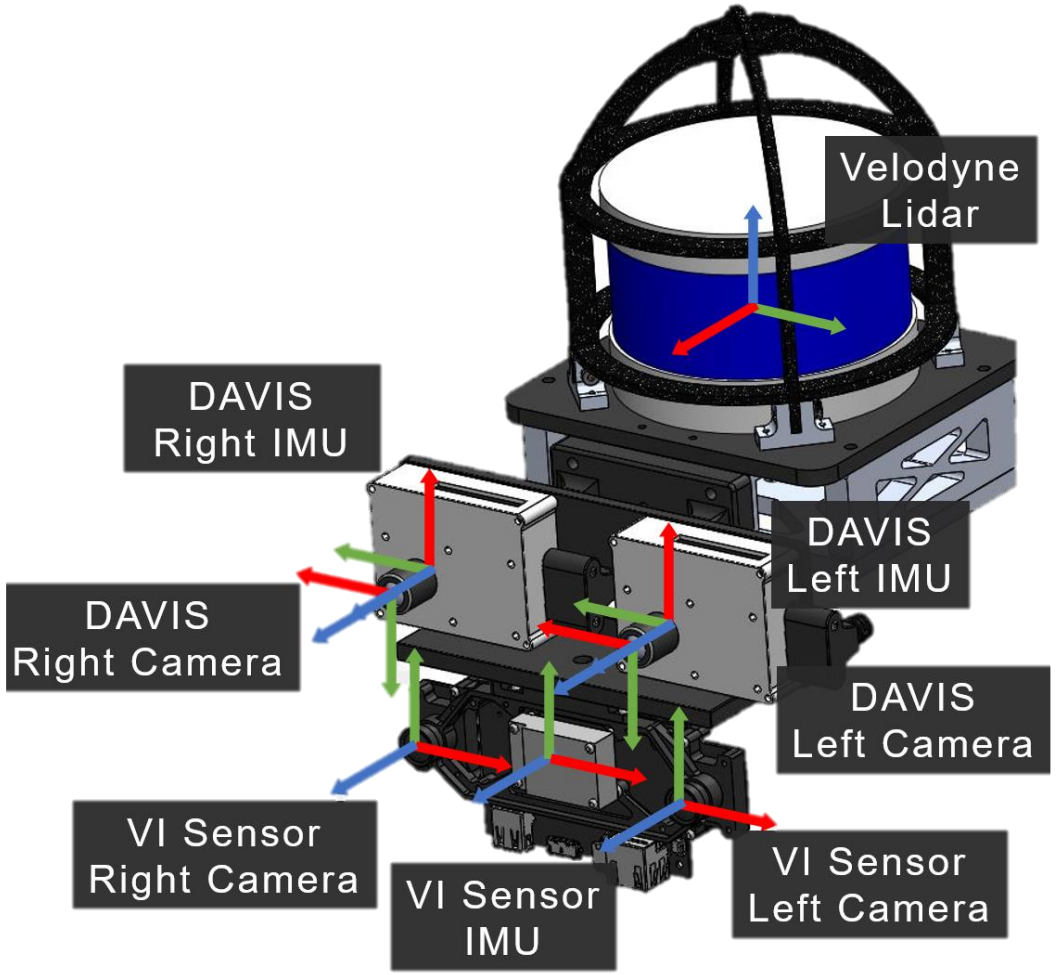
The Multi Vehicle Stereo Event Camera Dataset

Alex Zihao Zhu, Dinesh Thakur, Tolga Ozaslan, Vijay Kumar, Kostas Daniilidis

A.Z. Zhu, D. Thakur, T. Ozaslan, V. Kumar, K. Daniilidis. “The Multi Vehicle Stereo Event Camera Dataset: An Event Camera Dataset for 3D Perception.” RA-L/ICRA 2018.

Sensors

- Stereo DAVIS 346 event cameras
- VLP-16 Velodyne Puck
- VI-Sensor
- GPS
- Vicon/Qualisys Mocap



Data Collection

Data was collected from a car, hexacopter and motorbike, in a variety of environments, speeds and lighting conditions.

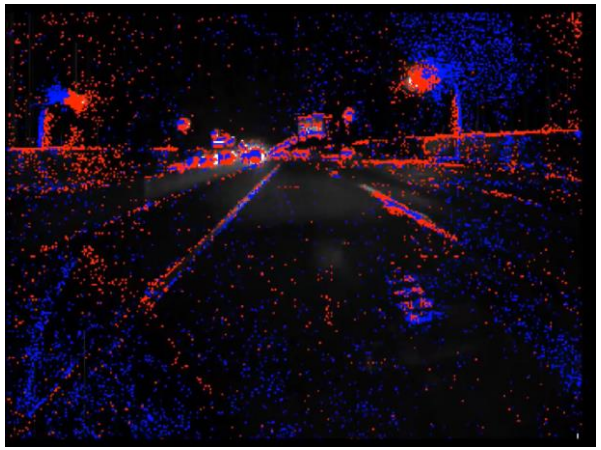
Ground truth poses and depths were collected from Vicon, Qualisys and Lidar odometry.

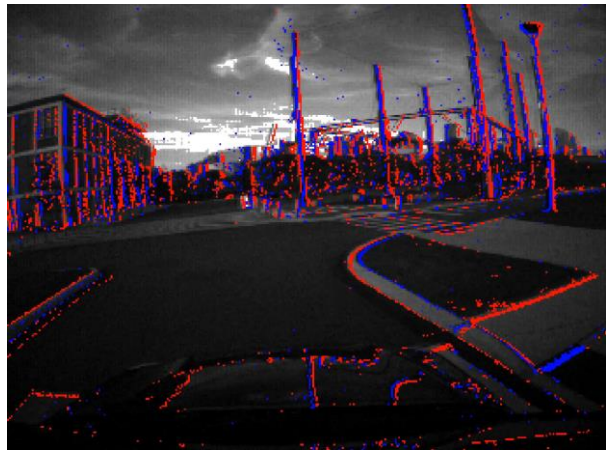
Sequences

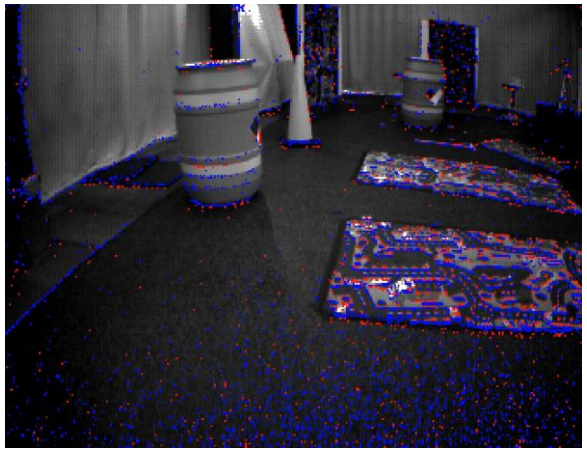
Multiple sequences over a variety of scenes were collected:

- Indoor flying
- Driving day
- Driving night
- Motorcycle

https://daniilidis-group.github.io/mvsec/
Files now available in hdf5 format

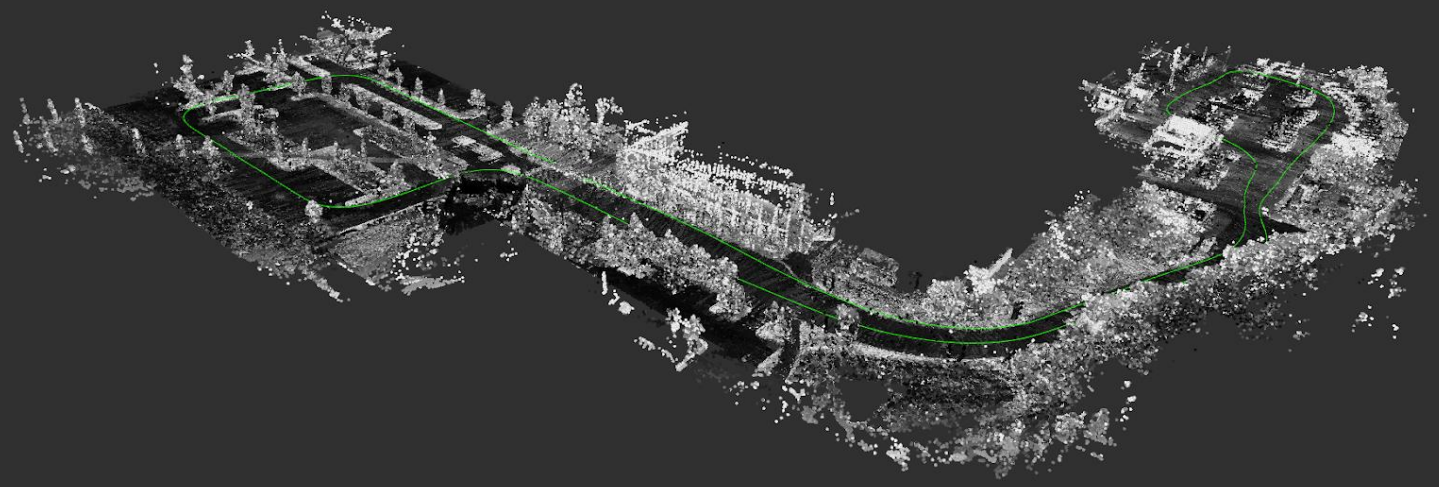


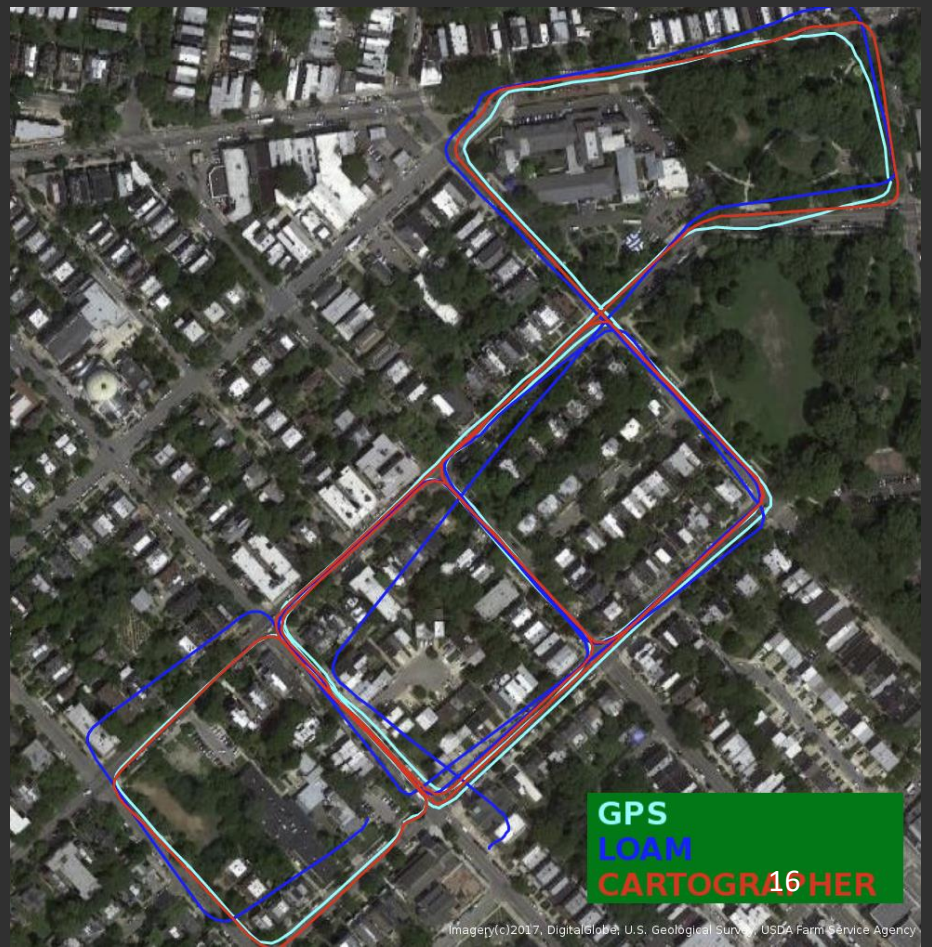




Ground Truth

Where possible, we provide ground truth pose, depth and optical flow.





outdoor_day2

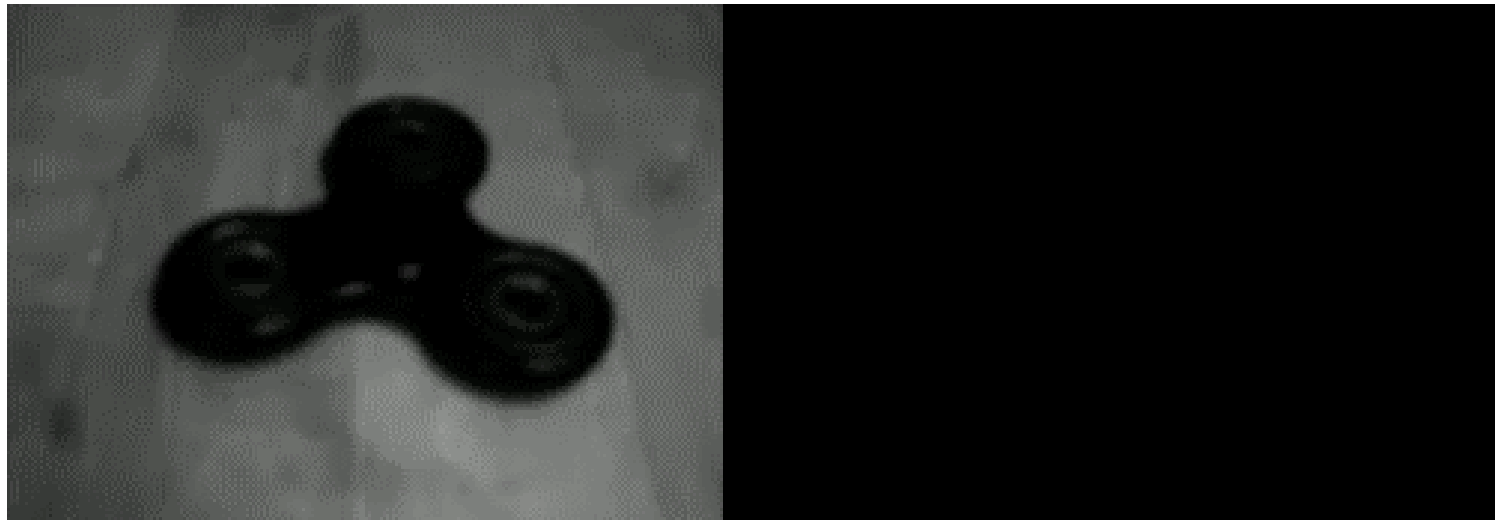
Unsupervised Event-based Learning of Optical Flow, Depth and Egomotion

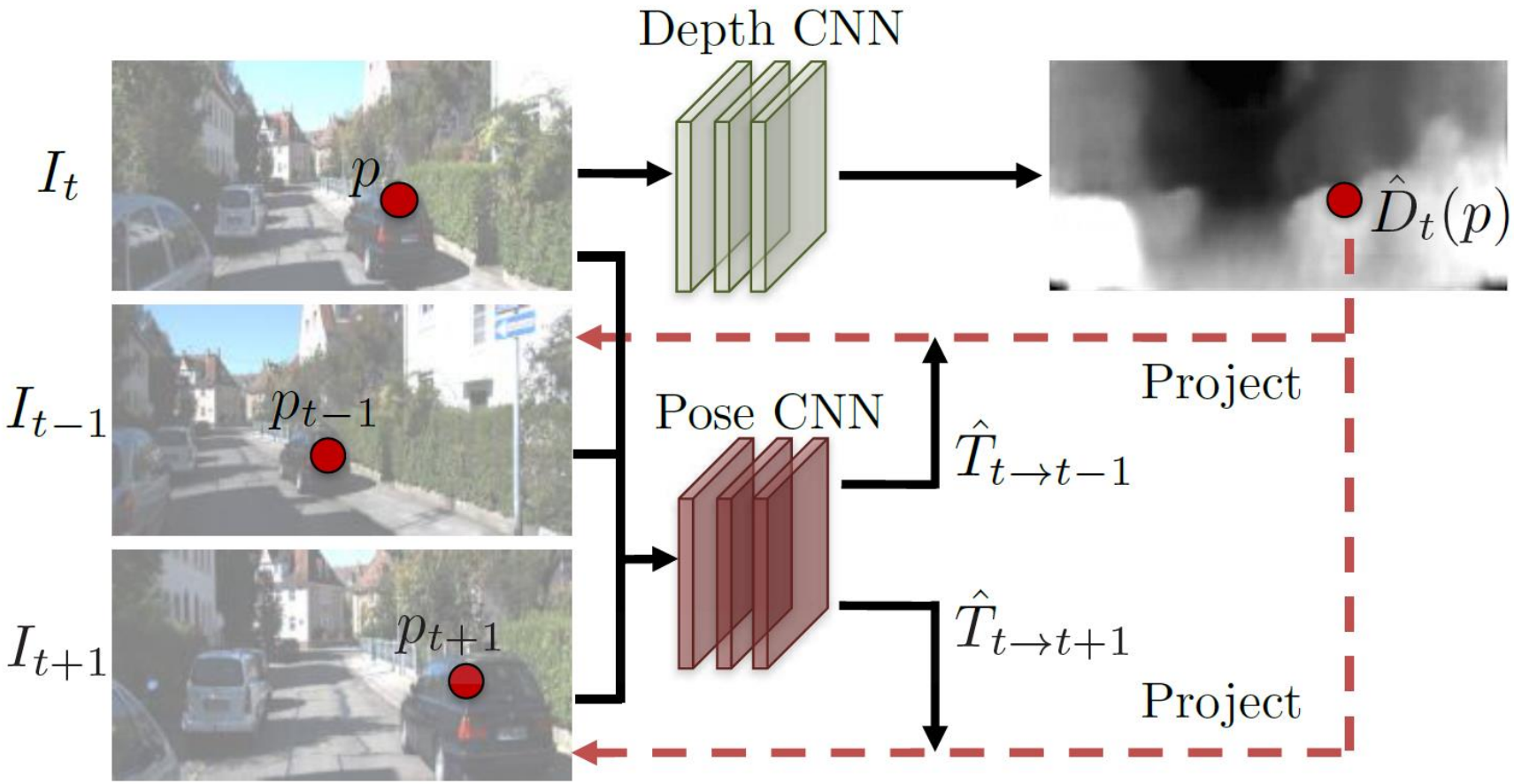
Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, Kostas Daniilidis

A.Z. Zhu, L. Yuan, K. Chaney, K. Daniilidis. “EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras.” RSS 2018. Best Student Paper Finalist (1 of 3).
A.Z. Zhu, L. Yuan, K. Chaney, K. Daniilidis. “Unsupervised Event-based Learning of Optical Flow, Depth and Egomotion.” CVPR 2019.

Contributions

- Self and unsupervised learning frameworks for event-only optical flow, depth and egomotion estimation.
- Novel input representations of events for CNNs.
- Supervision from the grayscale frames from the same camera, or a motion blur loss.





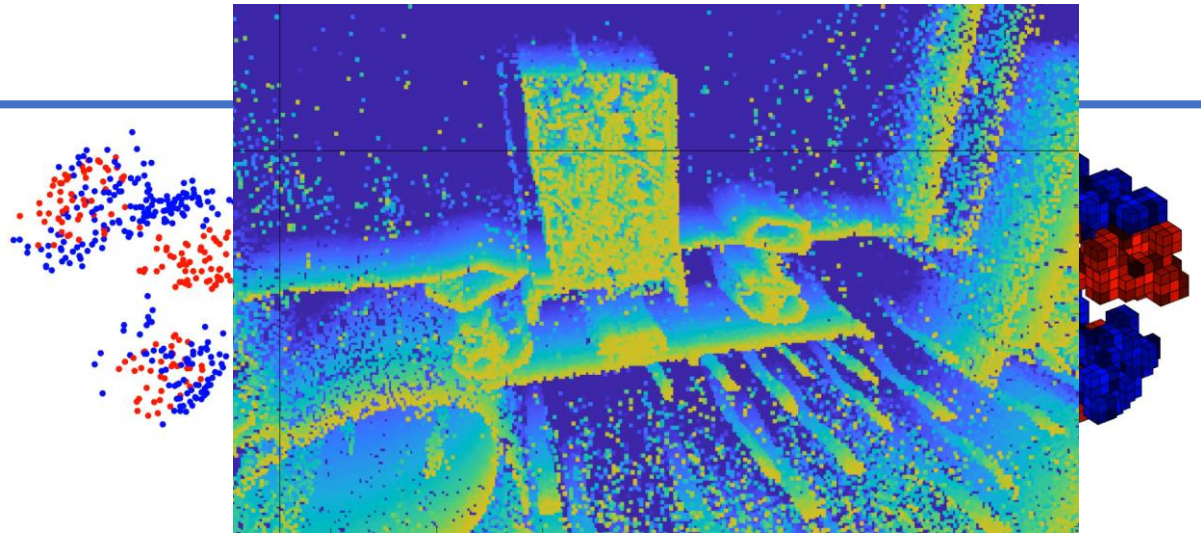
Zhou, Tinghui, et al. "Unsupervised learning of depth and ego-motion from video." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

Input Representations

1. Events are encoded as a 4-channel image, consisting of the last positive and negative timestamp at each pixel and the number of positive and negative events at each pixel.

2. Time domain is discretized into bins to generate a 3D volume. Events are inserted into the volume using trilinear interpolation.

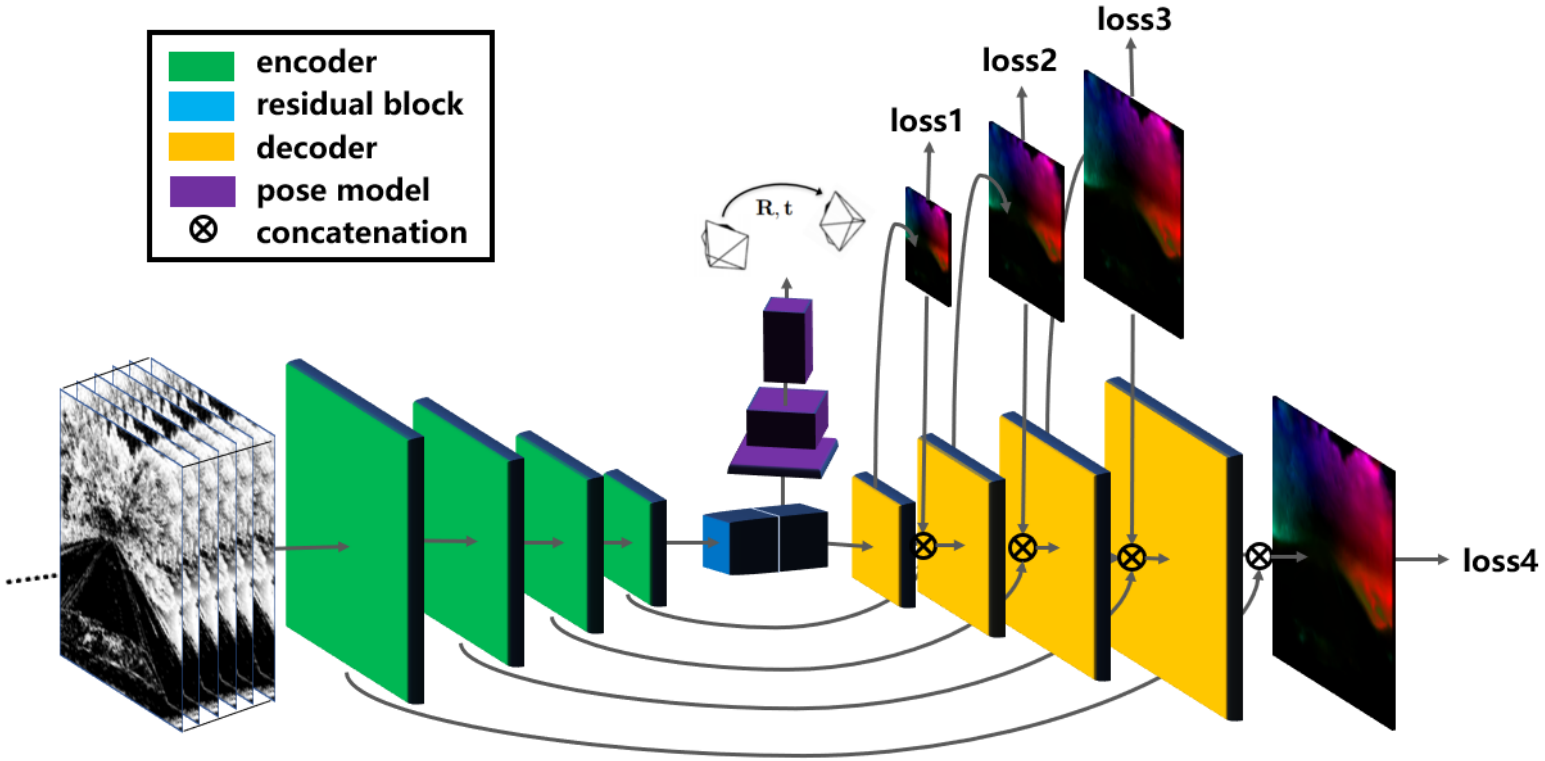
$$V(x,y,t) = \sum_i p_i \max(0, 1 - |x - x_i|) \max(0, 1 - |y - y_i|) \max(0, 1 - |t - t_i|)$$



Network Architecture

Encoder-decoder network with skip connections and a multi-scale loss.

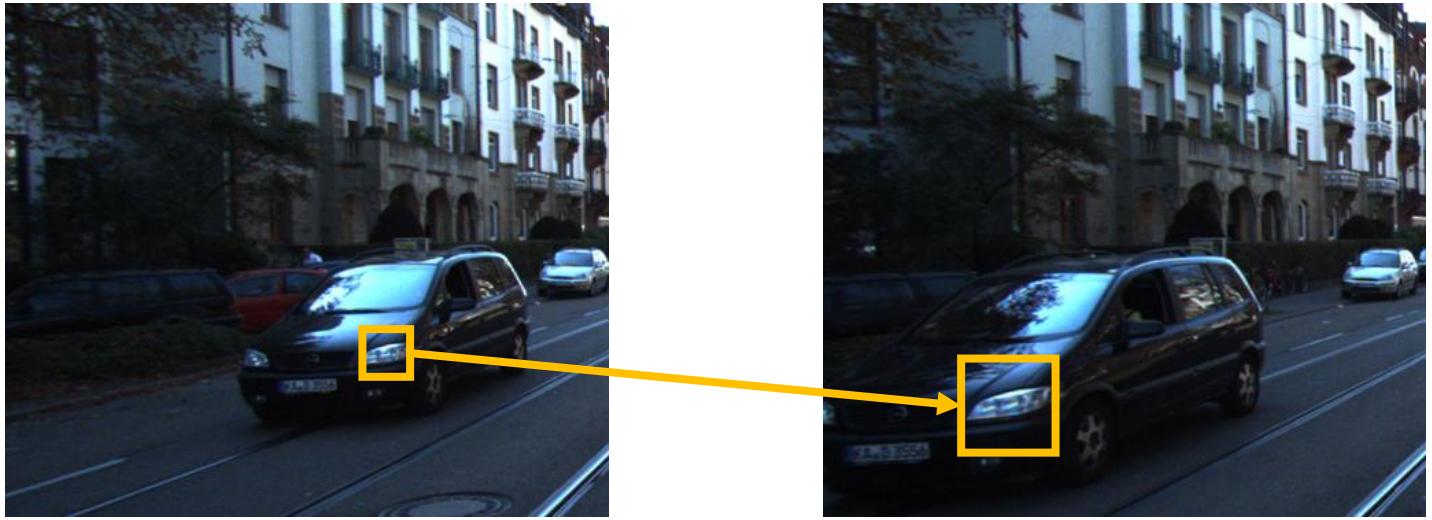
Runs at 20Hz on a NVIDIA GTX 960M, 75Hz on a Tesla V100.



Self-Supervised Grayscale Loss

The optical flow from the network is used to warp to next grayscale image to the previous, and a loss is applied on the difference between the warped image and the previous image.

$$L = \left\| I_{t_1}(x + \dot{x}, y + \dot{y}) - I_{t_0}(x, y) \right\|$$



Focus Loss

We attempt to focus the events using the optical flow predictions from the network, and generate an image of the <u>average timestamp</u> at each pixel.

$$\begin{pmatrix} x_i' \\ y_i' \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} + (t' - t_i) \begin{pmatrix} u(x_i, y_i) \\ v(x_i, y_i) \end{pmatrix}$$

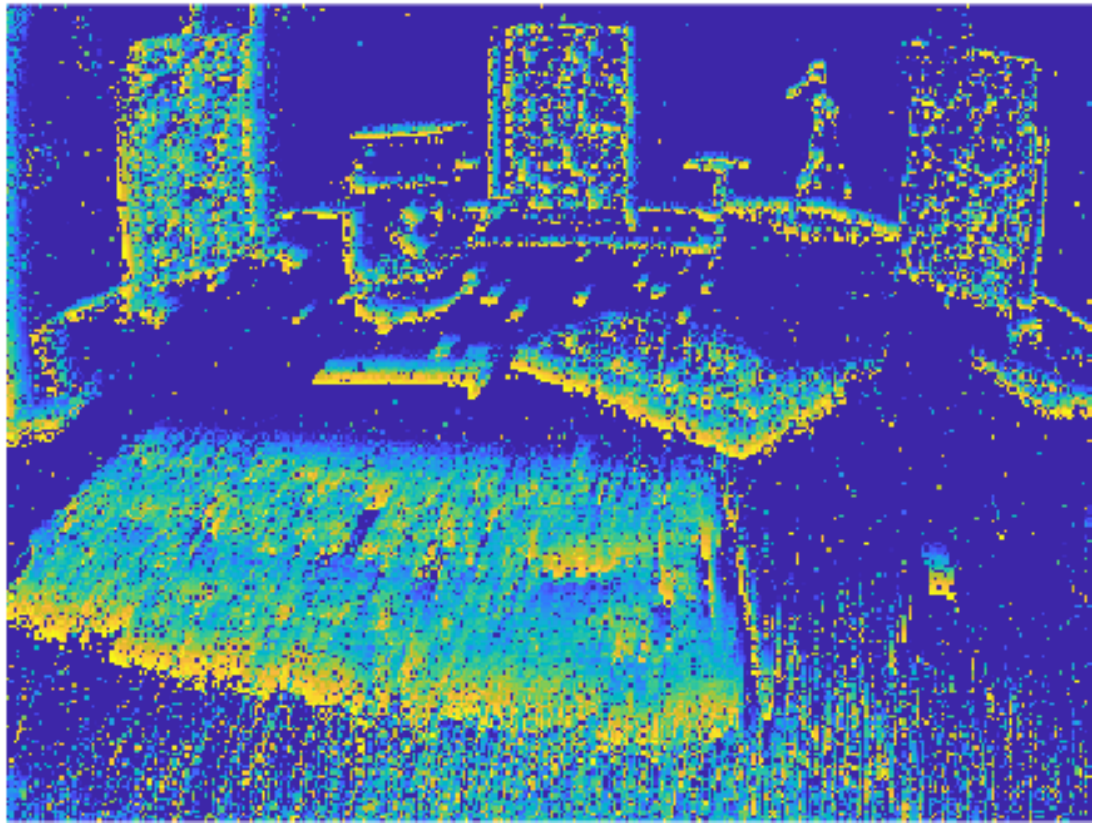
$$T(x, y, t) = \frac{\sum_i \max(0, 1 - |x - x_i'|) \max(0, 1 - |y - y_i'|)\, t_i}{\sum_i 1(|x - x_i'| < 1) 1(|y - y_i'| < 1)}$$

Our loss is the Charbonnier loss on the image, T.

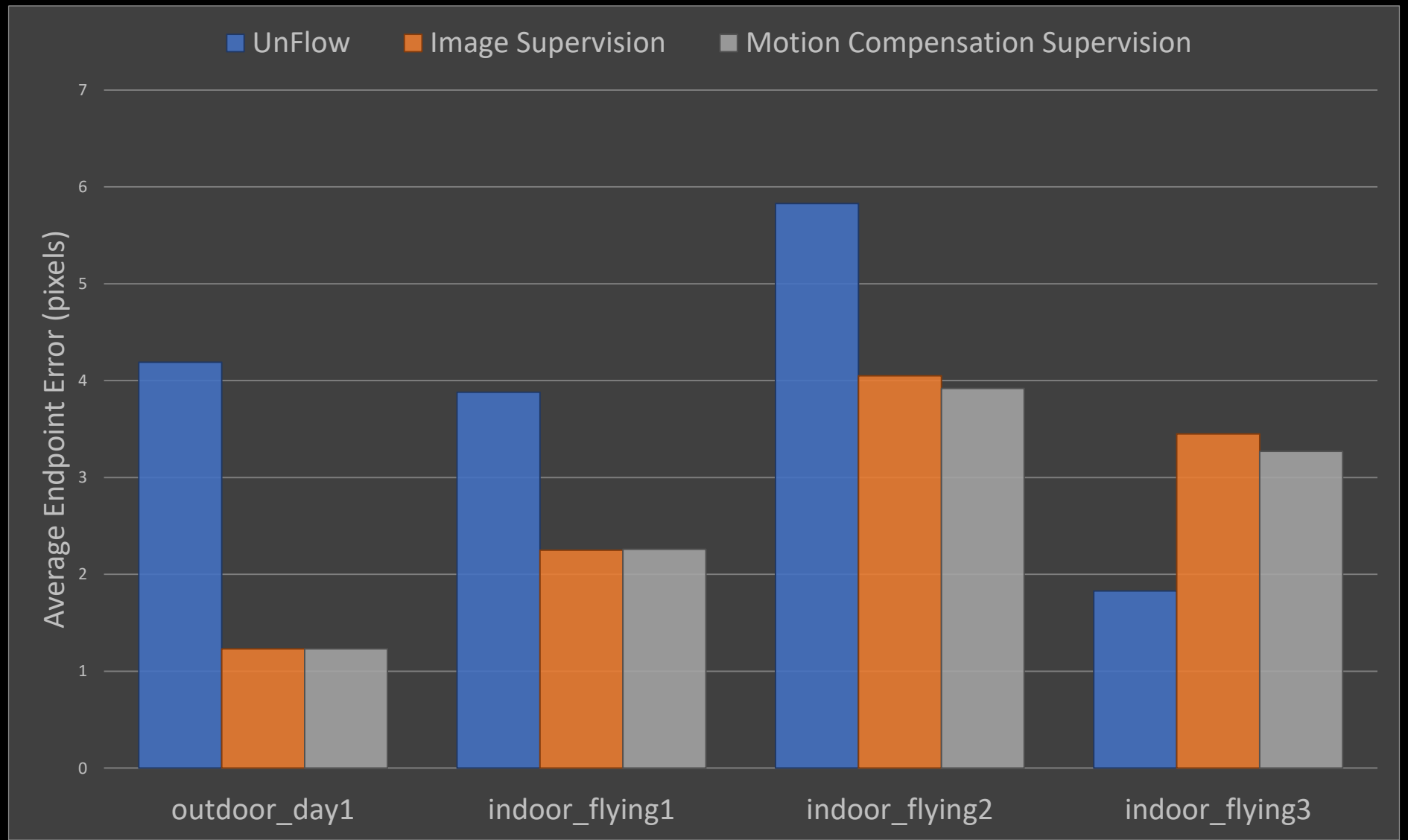
Mitrokhin, Anton, et al. "Event-based moving object detection and tracking." *IROS 2018.*

Focus Loss

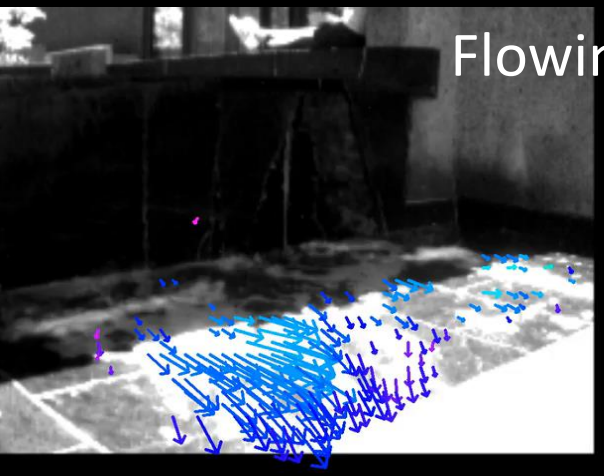
We attempt to focus the events using the optical flow predictions from the network, and generate an image of the <u>average timestamp</u> at each pixel.



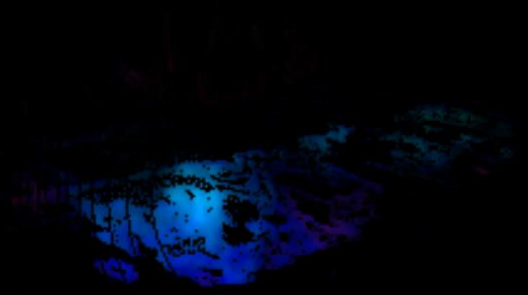
Quantitative Results



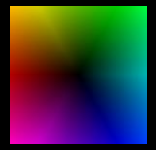
The network generalizes to a variety of challenging scenes!



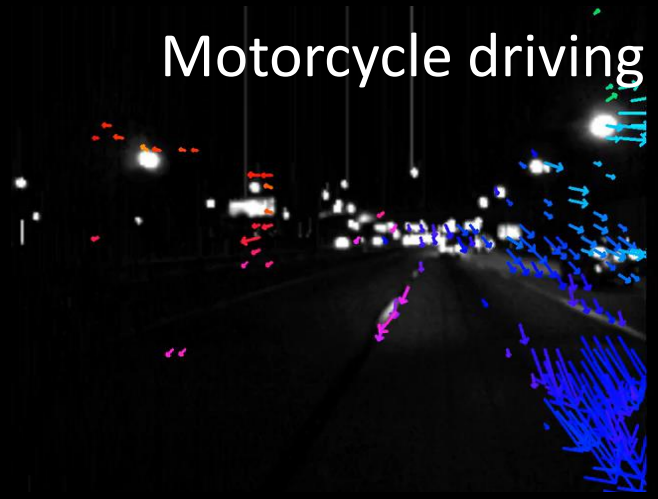
Flowing water



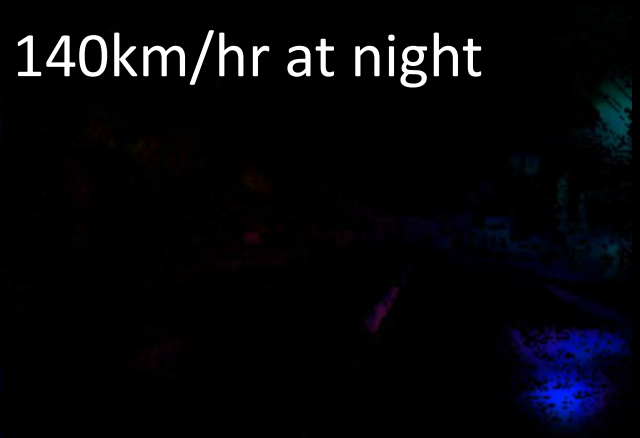
Fast ball passing (0.5x realtime)



Flow Direction



Motorcycle driving 140km/hr at night



Egomotion and Depth Estimation

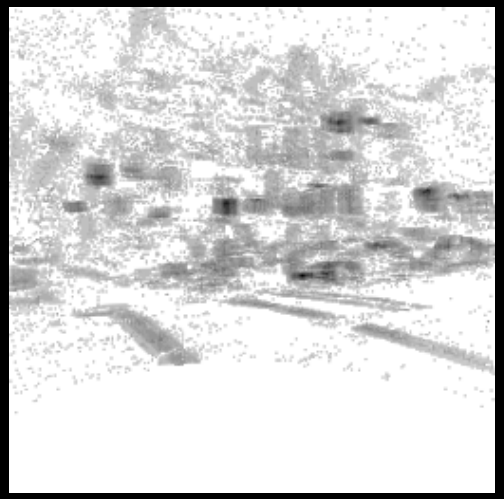
Given egomotion and depth predictions from the network, per pixel optical flow can be estimated using the motion field equations, assuming a static scene.

This optical flow can then be used with both the self and unsupervised losses.

For depth prediction, we can also incorporate stereo data by applying a photometric loss on the census transform of the deblurred images.

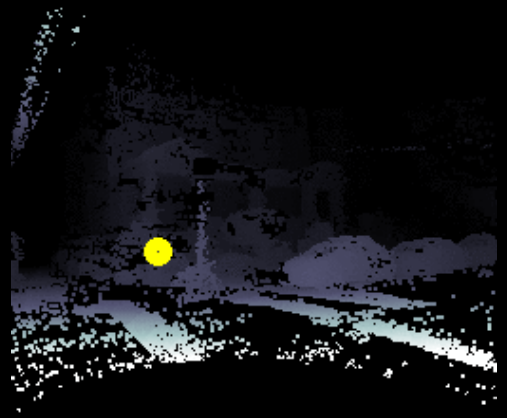
Egomotion and Depth Results

Grayscale Image

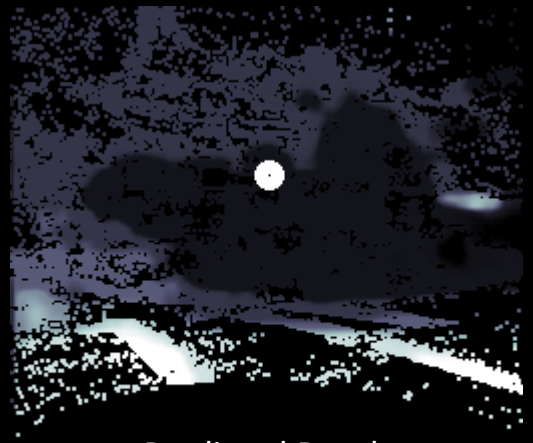


Event Image

Deblurred Event Image

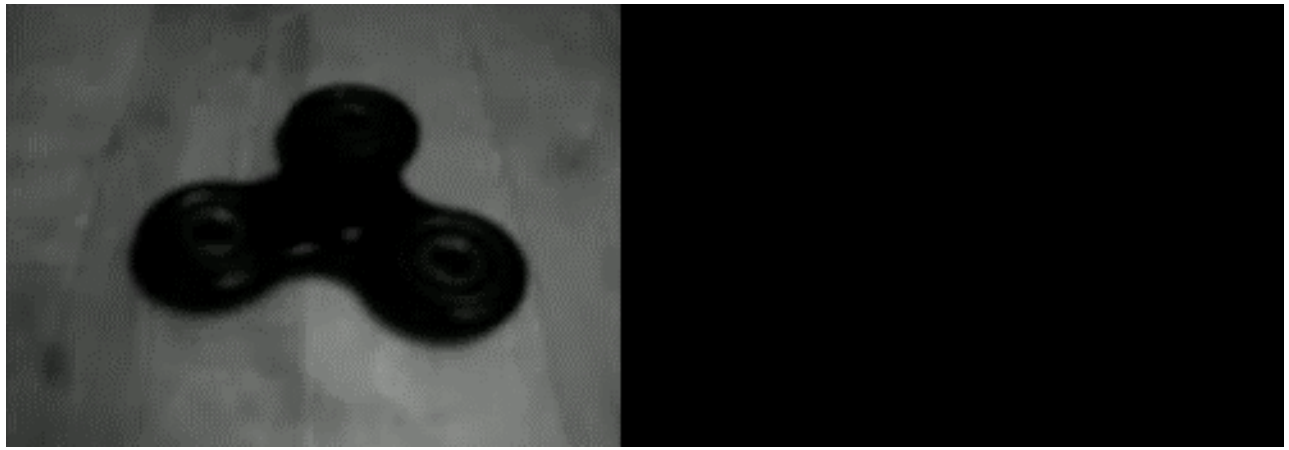


GT Depth and Heading Direction



Predicted Depth
and Heading Direction

Thank you. Any questions?



https://github.com/daniilidis-group/EV-FlowNet

A.Z. Zhu, N. Atanasov, K. Daniilidis. “Event-based feature tracking with probabilistic data association.” ICRA 2017.

A.Z. Zhu, N. Atanasov, K. Daniilidis. “Event-based visual inertial odometry.” CVPR 2017.

A.Z. Zhu, Y. Chen, K. Daniilidis. “Realtime Time Synchronized Event-based Stereo.” ECCV 2018.

A.Z. Zhu, D. Thakur, T. Ozaslan, V. Kumar, K. Daniilidis. “The Multi Vehicle Stereo Event Camera Dataset: An Event Camera Dataset for 3D Perception.” RA-L/ICRA 2018.

A.Z. Zhu, L. Yuan, K. Chaney, K. Daniilidis. “EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras.” RSS 2018. Best Student Paper Finalist (1 of 3).

A.Z. Zhu, L. Yuan, K. Chaney, K. Daniilidis. “Unsupervised Event-based Learning of Optical Flow, Depth and Egomotion.” CVPR 2019.