Learning Spatiotemporal Filters to Track Visual Saliency

Khaled Aboumerhi – ECE Ph.D. Candidate

Ralph Etienne-Cummings, Ph.D. – Professor of ECE

Johns Hopkins Whiting School of Engineering

Agenda

- Visual Saliency Overview
- Prophesee.ai Dataset
- Handcrafting Filters
- Pre-processing for Visual Saliency Model
- Spatiotemporal Filters
- Applications of Proposed Unsupervised Visual Saliency Model

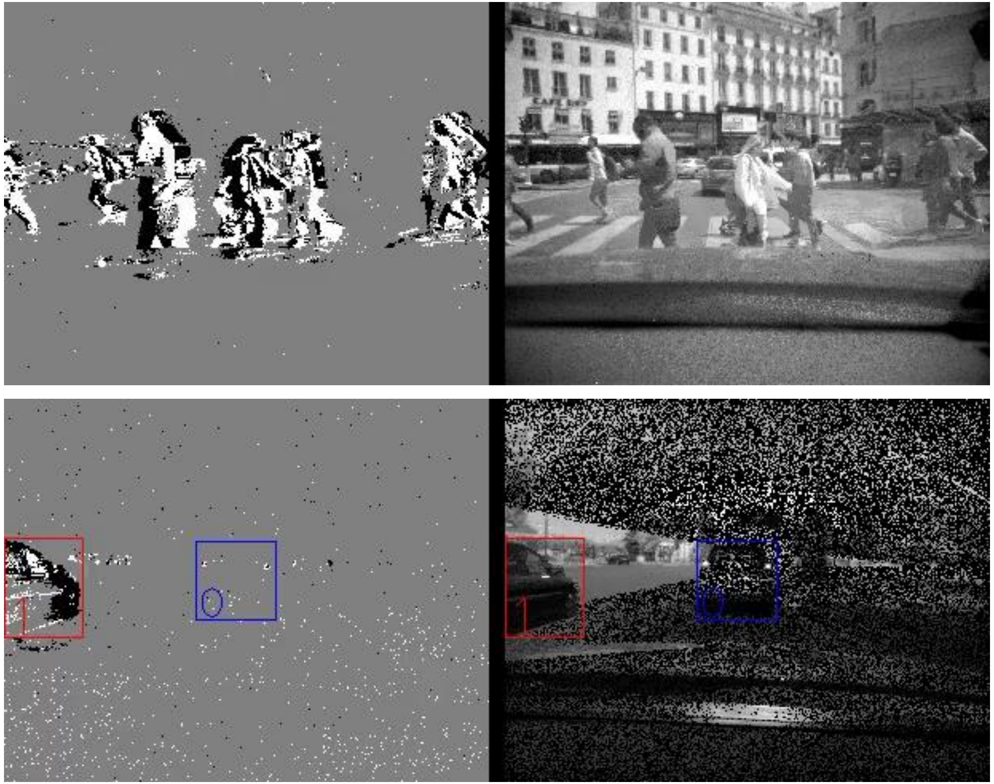
Visual Saliency

- Tendency to gaze in a particular direction or toward an interesting feature
- Applications in high-accuracy drone cameras, real-time traffic, criminal investigation, tele-tourism and more
- Why is it humans gaze toward specific features?

(Kadam et. al, 2020)

Dataset – Prophesee.ai

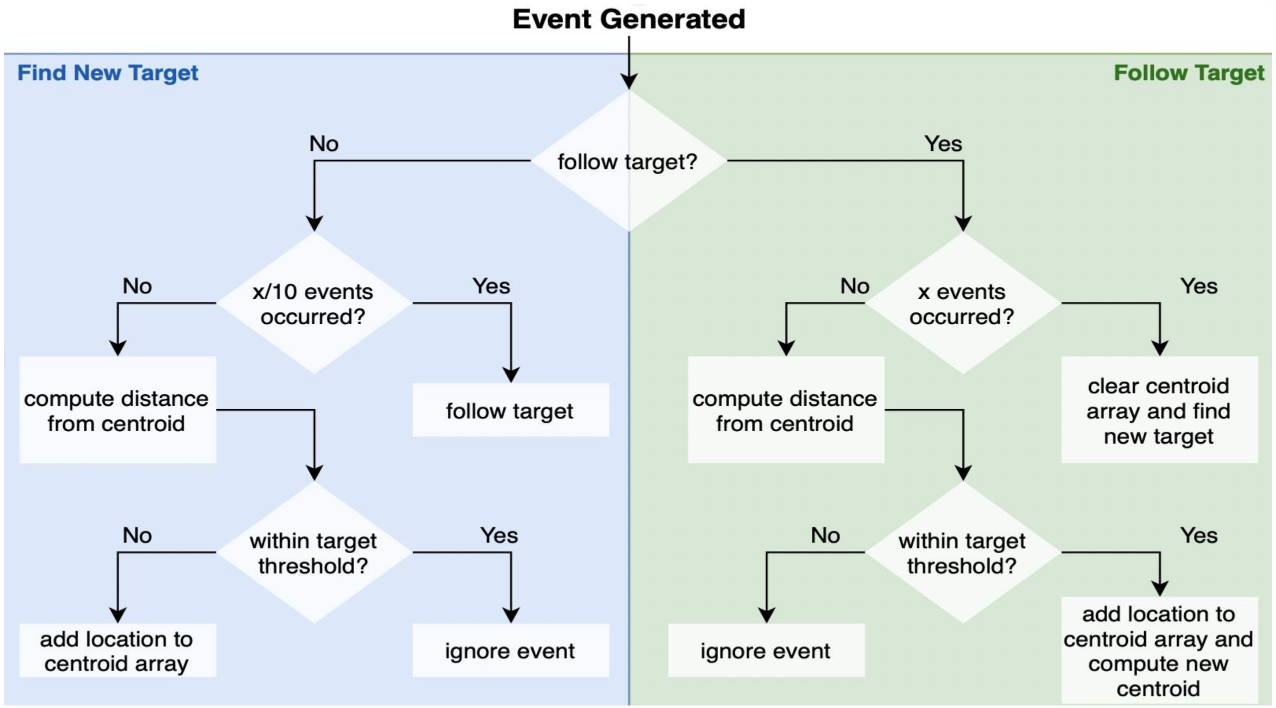
- ATIS Camera
- APS events (left) and real visualization (right)
- 10 μs resolution



Streetcar Dataset

Motorway Dataset

Preprocessing Technique



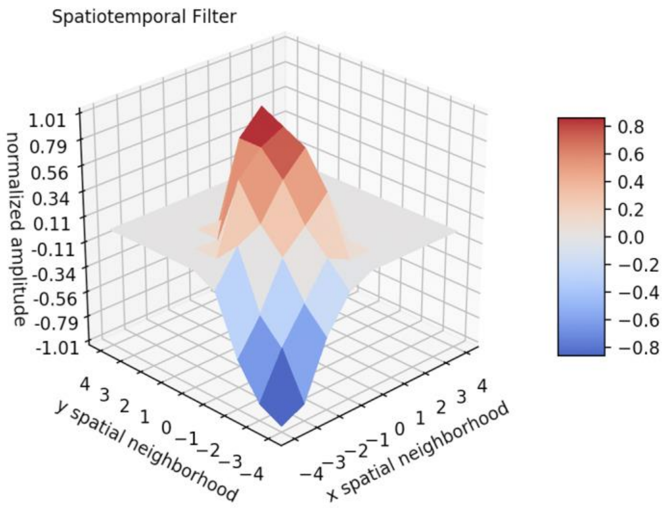
Spatiotemporal Filters

a	a	b	b	a	a
a	a	b	b	a	a
a	a	b	b	a	a
a	a	b	b	a	a
a	a	b	b	a	a
a	a	b	b	a	a

Vertical (handcrafted) Filter

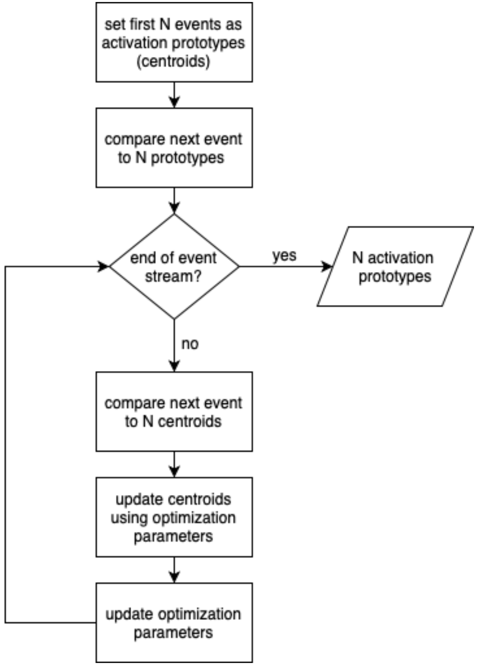
-1	-0.9	-0.4	-0.1	0	0
-0.9	-0.5	-0.2	0.1	0	0
-0.6	-0.3	0.1	0.29	0.5	0
-0.1	0.2	0.3	0.6	0.8	0
0	0	0.6	0.8	0.9	0
0	0	0	0.8	1	0

Learned Filter



$$S_{u,i} = e^{(-t_i - T_{u,i})/\tau} \qquad (1)$$

Learning Spatiotemporal Filters



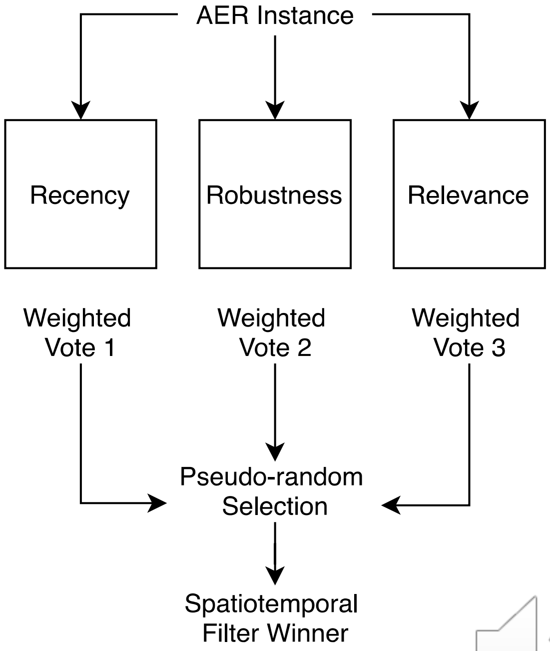
$$\eta_k = \det\left|S_{u,i} - S_{u,i-1}\right| \quad (2)$$

$$C_k = argmin|\eta, k| \quad (3)$$

$$F_A(\delta) = 0 \quad (4)$$

$$\varepsilon_A = \det|\delta F_A| \quad (5)$$

$$\sigma = rand(\Sigma\epsilon) \quad (6)$$



Decision Trees for Lifelong Learning
"wisdom of crowds"

But what is lifelong learning?

- Human brains use lifelong learning to keep important information based on many parameters, including
  - Recency
  - Relevance
  - Intense/reactive response
- How can we properly manage distinct, unique information over a timeless process without losing different types of important filters?
- Decision trees are great for simple decision making, but how can we incorporate many parameters for lifelong learning?

Constraint	Definition
$\varepsilon_n(t) \to 0$	consistency
$\varepsilon_n(t > t') = \varepsilon_n(t)$	no catastrophic forgetting
$\varepsilon_{n+m}(t > t') \to 0$	continual learning
$\#(f_{n+m}) - \#(f_n) \in \vartheta(\log(m))$	space efficient

(Vogelstein et. al, 2019)

Lifelong learning has few, yet strong parameters, of which no catastrophic forgetting and continual learning are incredibly important for timeless and online learning

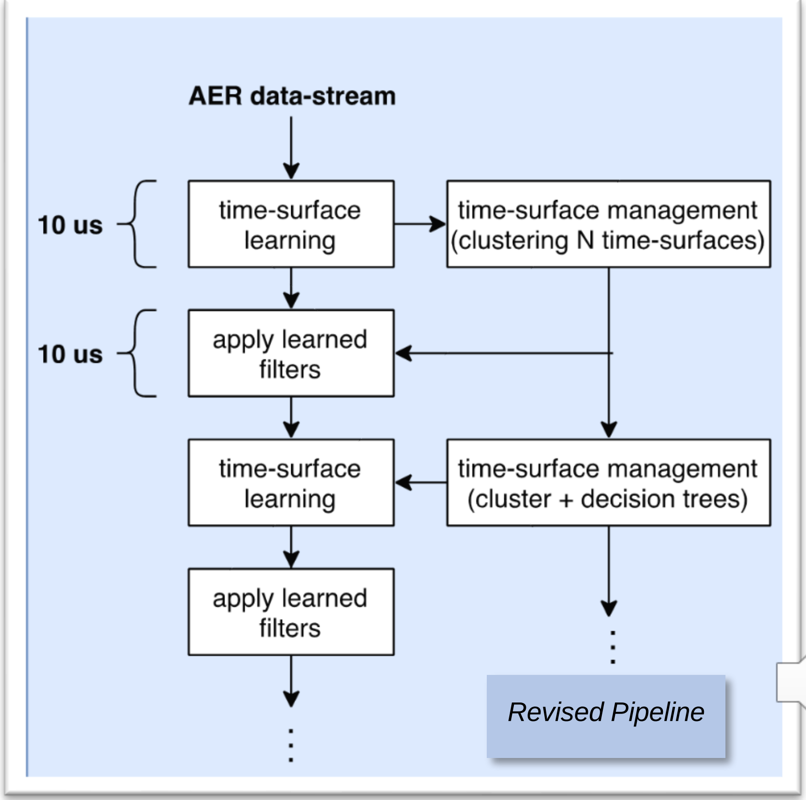
Creating stronger and less correlated trees using relaxation constraints clears up with computational resources

M002

Large Dataset Management

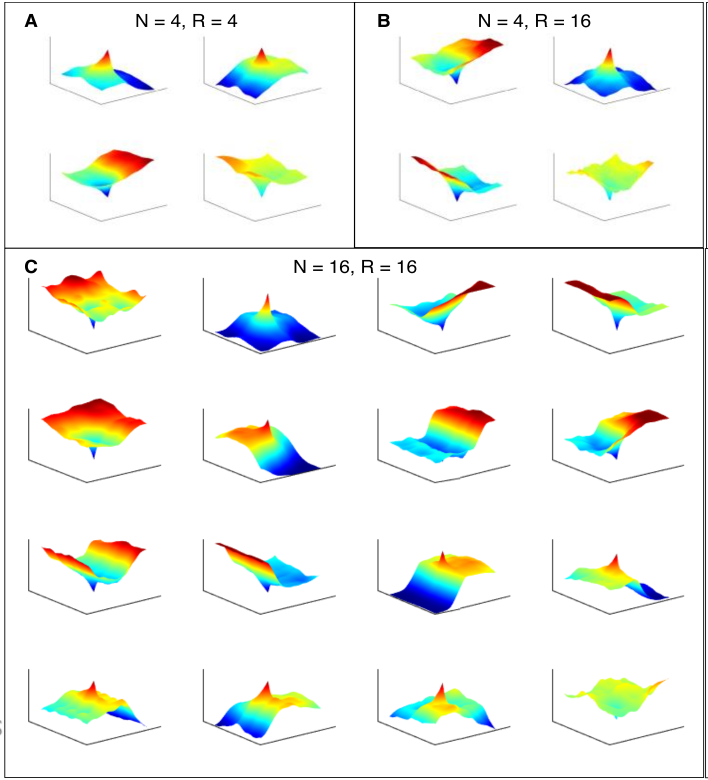
- Prophesee data set is large ($\sim 10^9 \times 1$ vector)
- Parsed to 10 μs blocks of online learning – online application
- Note: after a full 20 μs period of learning, time surface management becomes complicated
  - How do we incorporate previous filters in future learning/application?
  - How can we implement desired filter inputs?

20-bit 10ns sensor data interface ATIS AER Data Blocks
note: 10us with 10ns sensor data interface -> 1e3 x [240 x 304] -> 7.296e7 dimensionality

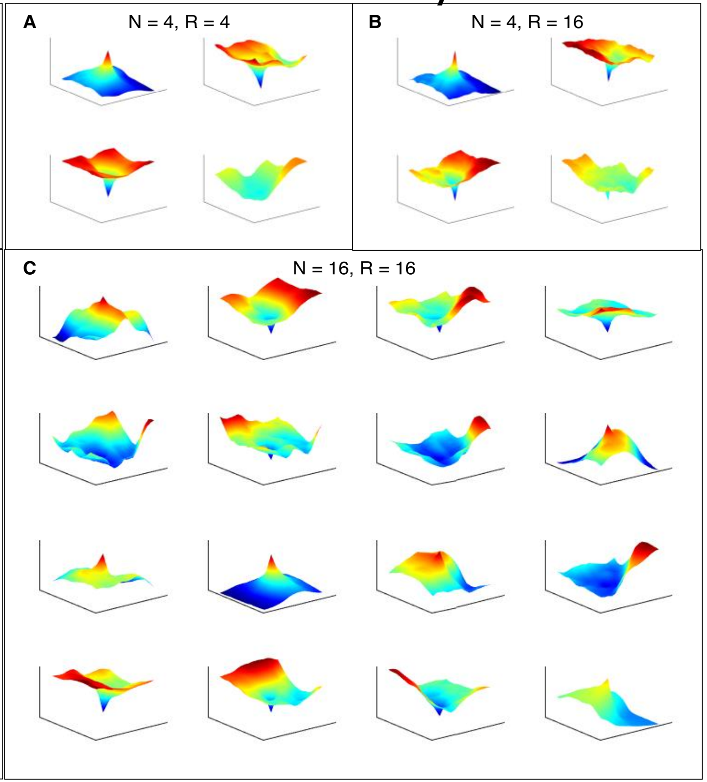


Learned Spatiotemporal Filters

Streetcar

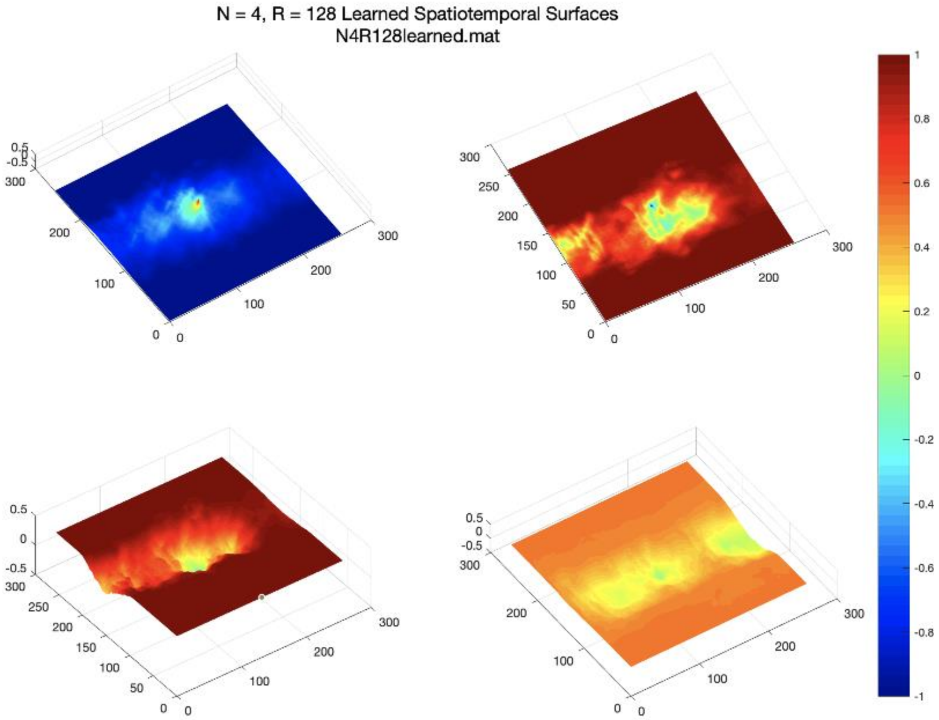


Motorway

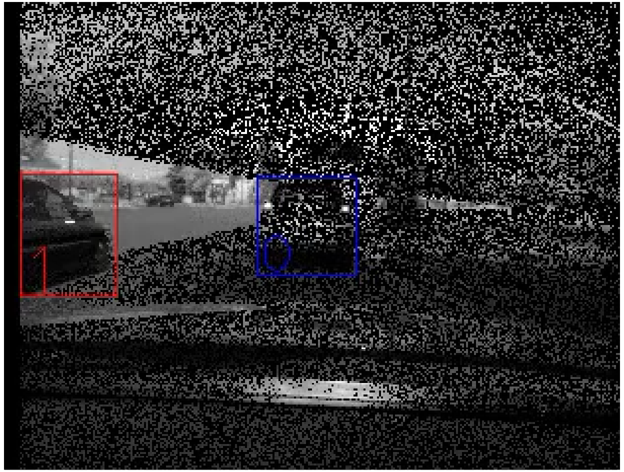


Learned Spatiotemporal Filters

- Large region (R) filters
  - Not used, but shows clearly the movement of event-based data through a field of vision

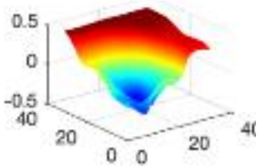


Implementations of Filters - Motorway

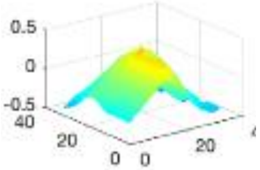


			E	IE
		7.1	2.3	7.83
			3.403	4.13
E				24.10

green



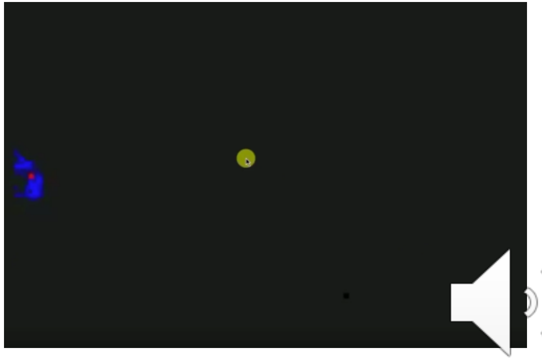
red



yellow

Abstract centroid placement comparison between lowest and highest generality filters

Experienced Observer

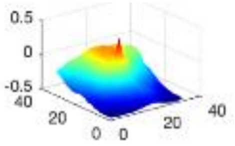


Inexperienced Observer

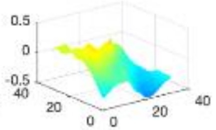
Implementations of Filters - Streetcar

			E	IE
		2.02	5.66	6.101
			8.8	5.5
E				19.23

green



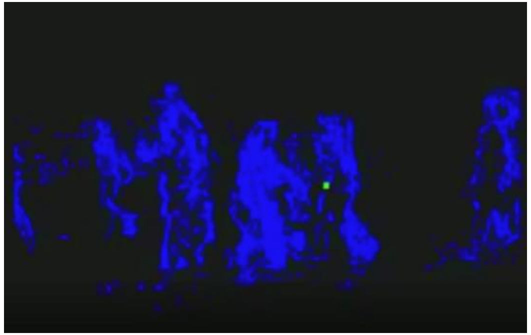
red

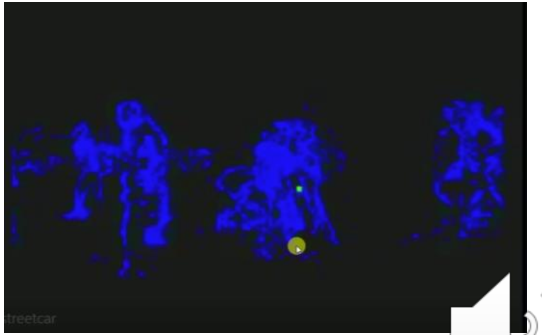


yellow

Abstract centroid placement comparison between lowest and highest generality filters

Experienced Observer





Inexperienced Observer

Future Work

- acquire the first spike-based saliency data-set
- Visual saliency algorithms that track and attempt to expose the brain's tendencies cannot be truly verified without a ground-truth.
- We can do this using eye-tracking devices in a closed environment, such as the HTC Vive or Google HoloLens.
- Breaking up the TD events in groups of two, four, and eight seconds, while applying the centroids during the intermittent latent phases, could prove to be a more accurate and quick process.

References

1. Q. Zhao and C. Koch, “Learning saliency-based visual attention: A review,” Signal Processing, vol. 93, no. 6, pp. 1401–1407, Jun. 2013, doi: 10.1016/j.sigpro.2012.06.014.Sam Anzaroot and Andrew McCallum. 2013. UMass Citation Field Extraction Dataset. Retrieved May 27, 2019 from http://www.iesl.cs.umass.edu/data/data-umasscitationfield

2. M M. Iacono, G. D’Angelo, A. Glover, V. Tikhanoff, E. Niebur, and C. Bartolozzi, “Proto-object based saliency for event-driven cameras,” 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Nov. 2019, doi: 10.1109/iros40897.2019.8967943.Chelsea Finn. 2018. Learning to Learn with Gradients. PhD Thesis, EECS Department, University of Berkeley.

3. G. Gallego et al., “Event-based Vision: A Survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2020, doi: 10.1109/tpami.2020.3008413.Matthew Van Gundy, Davide Balzarotti, and Giovanni Vigna. 2007. Catch me, if you can: Evading network signatures with web-based polymorphic worms. In Proceedings of the first USENIX workshop on Offensive Technologies (WOOT ’07) . USENIX Association, Berkley, CA, Article 7, 9 pages.

4. A. Andreopoulos, H. J. Kashyap, T. K. Nayak, A. Amir and M. D. Flickner, "A Low Power, High Throughput, Fully Event-Based Stereo System," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 7532-7542, doi: 10.1109/CVPR.2018.00786.David Harel. 1979. First-Order Dynamic Logic. Lecture Notes in Computer Science, Vol. 68. Springer-Verlag, New York, NY. https://doi.org/10.1007/3-540-09237-4