WESTERN SYDNEY UNIVERSITY

Unsupervised Motion Segmentation for Neuromorphic Aerial Surveillance

Sami Arja, Alexandre Marcireau, Saeed Afshar, Bharath Ramesh, Gregory Cohen

Scan the QR code:
To access the Project page, code and paper

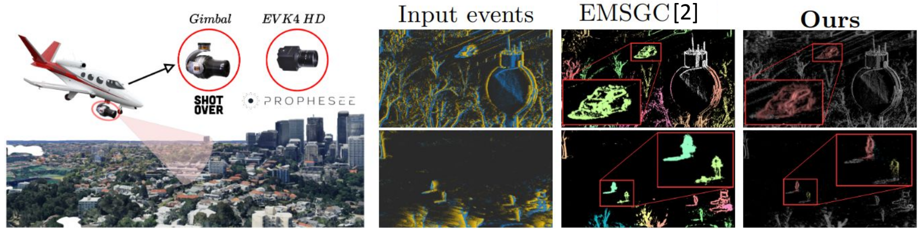
TL;DR

How can we detect any moving target from an aerial platform based on motion? We develop a network that detect any salient object regardless its shape and appearance and estimate its motion without the need of any human labelling, relying purely on self-supervised features from Vision Transformers (ViT) and optical flow.

Background

Human operators face considerable challenges in interpreting data from aerial surveillance videos due to the overwhelming amount of visual information. This often results in fatigue and reduced effectiveness. This work highlights the critical need for automated scene understanding to alleviate the workload of these operators, with motion segmentation identified as a pivotal task in achieving this objective.

Event-based aerial surveillance setup



An event camera with an HD resolution (1280x720 pixels) is attached to a stabilised gimbal on a small airplane looking either downward or in an oblique view.

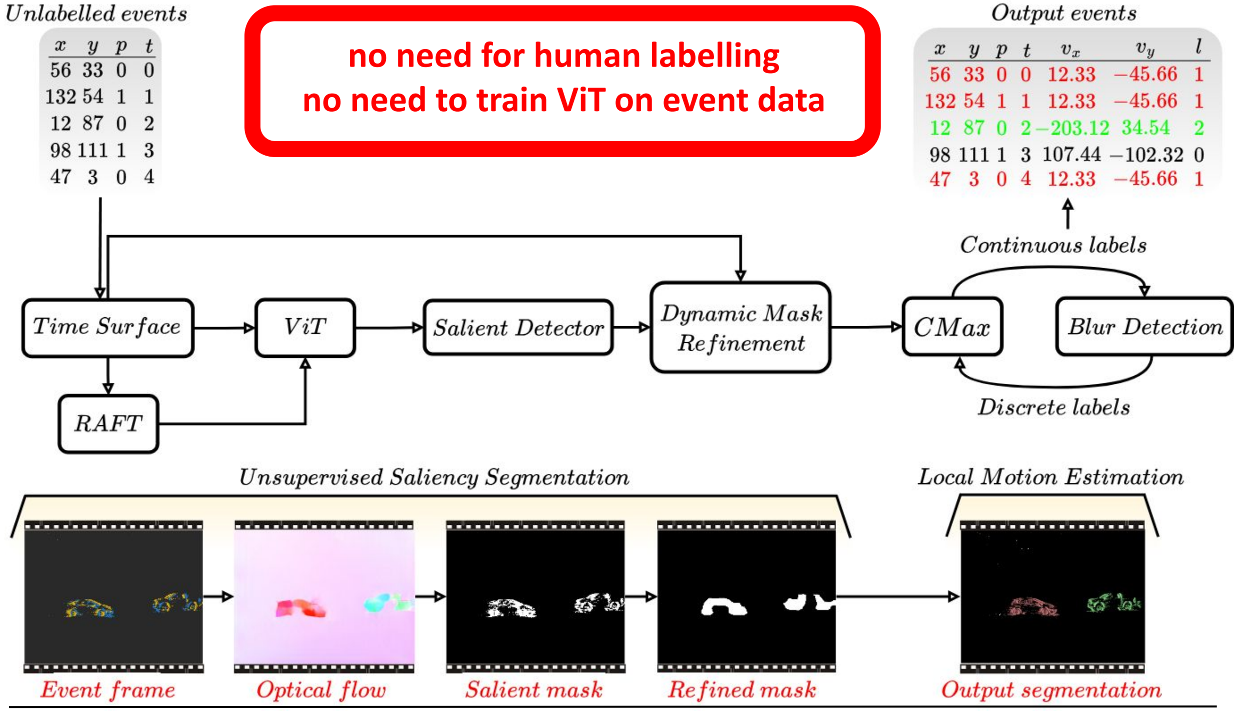
Contributions of this work

- An unsupervised method for discovering and segmenting moving objects using motion cues from an aerial platform, eliminating the need for manual annotations.
- A dynamic mask refinement process that integrates appearance information from a self-supervised DINO [1] model to improve the accuracy of the saliency masks.
- The introduction of the Ev-Airborne dataset, providing high-resolution data from an aerial platform, complete with ground truth annotations.
- Superior segmentation performance on major benchmarks, showcasing strong generalization capabilities compared to existing event-based motion segmentation methods.

Method

The method works in two steps:

1. Detecting where the moving targets are using Vision Transformers (ViT) and optical flow
2. Estimating the motion of each moving target along with the camera egomotion using contrast maximization (CMax) framework



no need for human labelling
no need to train ViT on event data

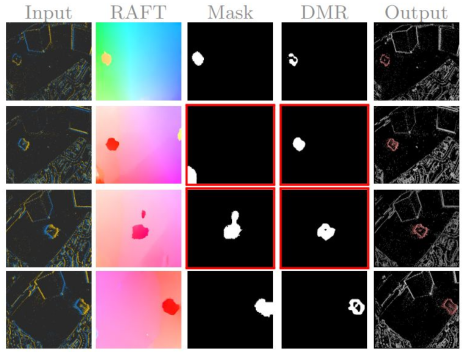
Unlabelled events

x	y	p	t
56	33	0	0
132	54	1	1
12	87	0	2
98	111	1	3
47	3	0	4

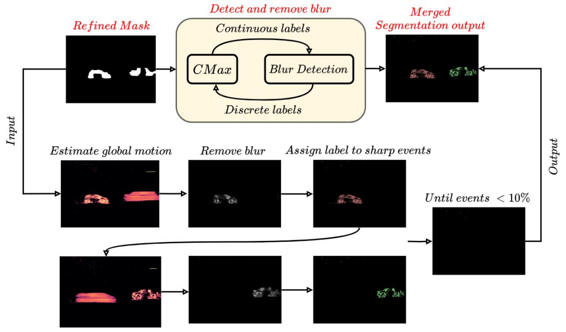
Output events

x	y	p	t	v_x	v_y	l
56	33	0	0	12.33	−45.66	1
132	54	1	1	12.33	−45.66	1
12	87	0	2	−203.12	34.54	2
98	111	1	3	107.44	−102.32	0
47	3	0	4	12.33	−45.66	1

Advantage of dynamic mask refinement

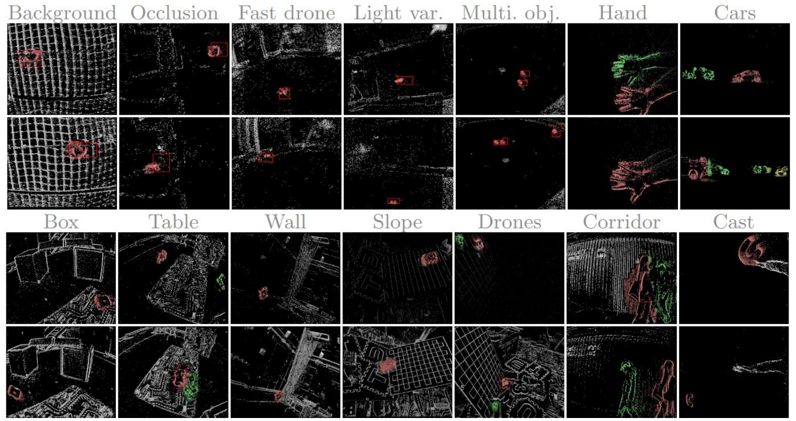


CMax-blur detection algorithm

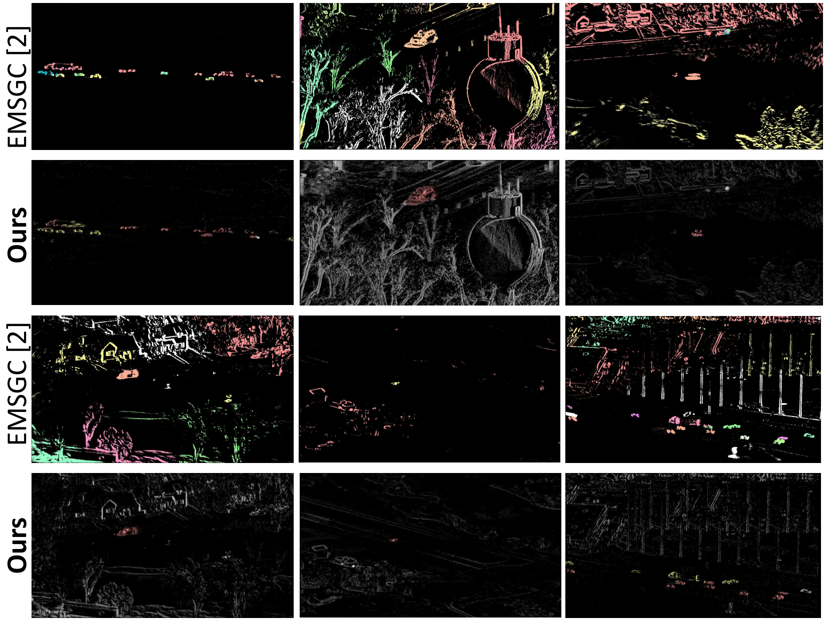


Results

Qualitative results on publicly available datasets



Qualitative results on EV-Airborne dataset compared with EMSGC [2] method



Reference

[1] Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9630–9640. IEEE, Montreal, QC, Canada (Oct 2021).

[2] Zhou Y, Gallego G, Lu X, Liu S, Shen S. Event-based motion segmentation with spatio-temporal graph cuts. IEEE Transactions on Neural Networks and Learning Systems. 2021 Nov 12;34(8):4868-80.