EV-LayerSegNet: Self-supervised Motion Segmentation using Event Cameras

Youssef Farah[1], Federico Paredes-Vallés[2], Guido C.H.E. De Croon[2], Muhammad Ahmed Humais[1], Hussain Sajwani[1], Yahya Zweiri[1]
[1]Khalifa University, [2]Delft University of Technology

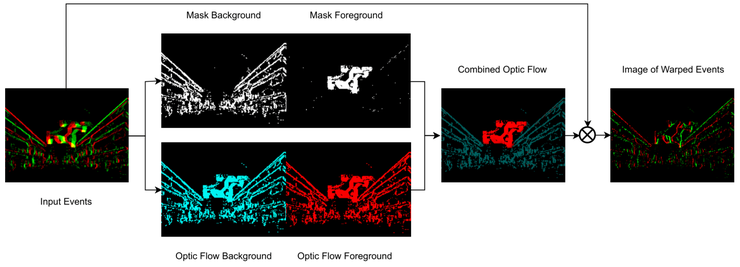
Introduction

Motivation

- Current event-based motion segmentation that do not rely on ground truth or pretrained weights work in an iterative fashion, limiting their deployment since iterations take considerable time
- Learning-based algorithms demonstrate good results and can be used in real-time, but are limited by the ground truth labels, which are approximative and expensive to obtain
- Consequently, there is the need of learning-based approaches that enable motion segmentation using self-supervised learning

Overview

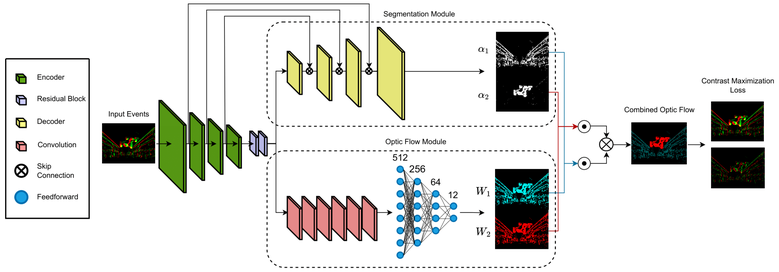
- We solve for these challenges by using a layered representation of the scene to segment moving objects, and leverage motion compensation as learning loss for self-supervised learning



- Our method takes as input an event volume that results in a blurry scene. It then generates a background and foreground mask
- Independently, it estimates affine optical flow for both layers, and combines them together using the masks
- Finally, it warps the events according to the combined flow. A successful motion deblur leads to accurate segmentation

Method

Overall architecture of EV-LayerSegNet



Optical Flow Module

6 convolutional layers, each followed by leaky ReLU activation. We then flatten the output of the last convolution layer and pass it to a feed-forward network consisting of 4 layers (512, 256, 64 and 12 output units), followed by tanh activation except the last layer. We then use the output of the feedforward network and split it to two sets of 6 affine motion parameters, and we compute the two flow maps W_1 and W_2

$$W_i(x,y) = \mathbf{A}_i \begin{bmatrix} 1 \\ x \\ y \end{bmatrix} = \begin{bmatrix} a_i^1 & a_i^2 & a_i^3 \\ a_i^4 & a_i^5 & a_i^6 \end{bmatrix} \begin{bmatrix} 1 \\ x \\ y \end{bmatrix}, \quad i \in \{1,2\}$$

Segmentation Module

The output of the residual blocks is bilinearly upsampled by 4 decoding layers. Each decoding layer is connected to the respective encoding layer by skip connection and is followed by leaky DoReLU activation with γ = 100. At the last layer, softmax is applied instead to ensure that the channel values are bounded [0,1] and sum up to 1

$$\mathcal{L}_{\text{contrast}}\ (t_{\text{ref}}) = \frac{\sum_{\boldsymbol{x}} T_{\pm}\left(\boldsymbol{x}; \boldsymbol{u} \mid t_{\text{ref}}\right)^2}{\sum_{\boldsymbol{x}} \left[n\left(\boldsymbol{x}'\right) > 0\right] + \epsilon}$$

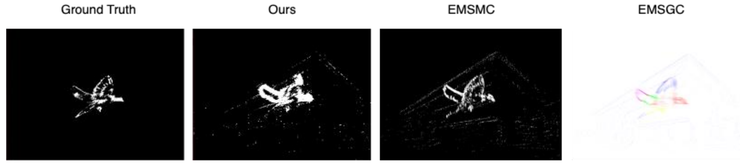
Experimental Results

Quantitative Results

We train and test the network on a simulated dataset with only affine motion, achieving IoU and detection rate up to 71% and 87% respectively

	Mean IoU			Mean Detection Rate		
Sequence name	**EMSMC**	**EMSGC**	**Ours**	**EMSMC**	**EMSGC**	**Ours**
Drone above playground	0.20	0.05	0.71	0.00	0.00	0.87
Plane over city	0.06	0.03	0.67	0.00	0.00	0.77
Bird above playground	0.26	0.00	0.48	0.00	0.00	0.71
Second drone above playground	0.36	0.02	0.52	0.00	0.00	0.78
Bird in front of building	0.20	0.41	0.10	0.00	0.21	0.00
Helicopter over city	0.07	0.00	0.55	0.00	0.00	0.83

Qualitative Results



Conclusion

- We propose a novel self-supervised network for learning event-based motion segmentation by introducing a novel optical flow module that enables self-supervised learning of affine optical flow and a segmentation module that learns separately the masks corresponding to the independently moving objects
- We contribute a new event-based dataset of several simulated backgrounds and objects moving according to affine motion.
- We demonstrate through experiments that our method shows superior performance in comparison with the state-of-the-art in unsupervised motion segmentation.