FusionSeg: Motion Segmentation by Jointly Exploiting Frames and Events

Lin Wang[1], Zhe Liu[1], Yi Zhang[2], Shaowu Yang[1], Dianxi Shi[2], and Yongjun Zhang[2(✉)]

[1] College of Computer, National University of Defense Technology, Changsha, China
wanglin12@nudt.edu.cn

[2] Artificial Intelligence Research Center, National Innovation Institute of Defense Technology, Beijing, China
yjzhang@nudt.edu.cn

Abstract. Segmentation of independently moving objects is an important stage in scene comprehension tasks like tracking and recognition. Frame-based cameras employed for dynamic scenes suffer from motion blur and exposure artifacts due to the sampling principle. In contrast, event-based cameras sample visual information based on scene dynamics and have the advantages of microsecond temporal resolution, high dynamic range, and more. Inspired by the complimentary of frame-based cameras and event-based cameras, we propose a cross-domain motion segmentation method, named FusionSeg, for fusing visual signals from frames and events to improve motion segmentation performance. To solve motion segmentation problem on the multi-objects scenario, we use the identification mechanism to embed multiple objects into the same feature space. In addition, to solve the feature matching and propagation problem, we design a long and short-term temporal-spatial attention. Our FusionSeg is evaluated on public datasets and outperforms the state-of-the-art by 4.7% in terms of detection rate. Experiments also demonstrate our method's robustness in situations with varying motion patterns and numbers of moving objects.

Keywords: Motion segmentation · Robot vision · Event camera

1 Introduction

Humans can easily perceive a complex scene as a set of distinct objects, a phenomenon known as perceptual grouping [23]. Robotic applications, such as autonomous driving and AR/VR, require the perception of dynamic scenes to interact effectively with the environment. In computer vision, perceptual grouping is closely related to the segmentation problem. That is, extracting objects with arbitrary shapes from a cluttered scene.

S. Khanna et al. (Eds.): PRICAI 2022, LNCS 13631, pp. 267–280, 2022.
https://doi.org/10.1007/978-3-031-20868-3_20

Much of the work in the visual segmentation research field involves optical flow computing as the initial stage [2,27]. However, precise optical flow calculation is difficult due to problems such as discontinuity and occlusion of moving objects. In the optimization field, various approaches use the idea of contrast maximization to accomplish the segmentation task [19]. In a multi-objects scenario, the motion-compensated images of each object need to be computed, which increases the computational cost. Feature point tracking allows for long-term estimation of pixel motion trajectories, which can resolve ambiguities in motion by analyzing pixel matches over larger time intervals. However, from a perceptual point of view, challenging visual effects (such as motion blur and underexposure/overexposure) make problem-solving with frame-based cameras more difficult. Therefore, inspired by biological visual motion processing mechanisms, neuromorphic engineers have developed a kind of sensor which is known as the event-based camera. It is not driven by a common clock because each pixel acts as an independent circuit, i.e., each pixel responds to motion independently and is therefore able to perceive dynamic changes in the scene efficiently and accurately. In addition, it can tolerate different lighting conditions and is sparsely encoded. The advantages in terms of temporal resolution, low latency ,and low bandwidth are enormous compared to frame-based cameras.

While event-based cameras have many benefits, they cannot measure absolute light intensity and difficult to capture slow motion and fine-grained texture information, which are important for high-performance segmentation. Frame-based cameras can just compensate for this. This unique complementarity leads us to propose a visual segmentation method based on the fusion of frames and events, called FusionSeg. In this paper, we use a simple and effective events aggregation method to discrete the time domain of asynchronous events. Thus it can be more easily processed based on CNN models. Another challenge is to efficiently obtain meaningful cues from the events domain and frames domain for different scenes. To this end, we introduce a new feature fusion method to efficiently fuse visual cues from both events and frames. The adaptive nature of our approach is maintained by a weighting scheme specifically designed to balance the contributions of both domains.

To make effective use of the motion information in the sequence, we propose a feature matching and propagation method. Firstly, we use an identification mechanism that assigns a unique recognition identity to each object and embeds multiple objects into the same feature space. The network can learn the association between all objects. The long and short-term temporal-spatial attention is then designed to implement feature matching and propagation. We demonstrate that our method outperforms other approaches on public datasets.

In summary, our contributions are:

(1) We introduce a feature fusion method that adaptively fuses visual cues from events and frames, and thus makes full use of both data for segmenting scene objects.
(2) We introduce feature matching and propagation methods to make effective use of motion information from time sequences. To our knowledge, this is

the first attempt to introduce Transformer into event-based motion segmentation.
(3) Our extensive experimental results show that our method has significant advantages over other state-of-the-art methods.

In the rest of the paper, we first review related work (Section II) and then explain our approach (Sections III). Finally, the model is validated (Section IV).

2 Related Work

2.1 Motion Segmentation

In the last decade, event-based motion segmentation has been solved for different scene complexities. For a moving event camera, events are triggered by static background and moving objects. The goal of motion segmentation is to infer the causal classification of each event. However, the amount of information carried by each event is small, and classifying each event is extremely challenging. Since moving objects produce different events trajectories in the image plane, event-based segmentation algorithms achieve the segmentation task primarily by inferring the trajectories of moving objects.

Assuming that the shape of the object to be segmented is known, Glover et al. [21]. Extracted the optical flow from the time window of the event stream based on the Hough transform, which in turn enables the segmentation and tracking of the ball. Later, they extended the method using particle filtering to improve tracking robustness, i.e.by dynamically selecting the duration of the observation window to accommodate sudden changes in object acceleration [4].

Some recent work has proposed the idea of using motion-compensated event images [3] to solve the problem of motion segmentation. Essentially, the technique associates events that produce sharp edges based on motion assumptions. The simplest assumption is a linear motion model, where the scene can be described as a collection of objects over a short period time, producing events that fit multiple linear motion models. Timo [20] et al. first fited a camera motion compensation model to the main events, then eliminated these events and finally greedily fited the remaining events to another linear model to produce motion compensated images with clear object contours. They later proposed an iterative clustering algorithm [19] that jointly estimated the motion parameters of the event-objects association and the objects that produced the sharpest motion-compensated event images. It allows a generic parametric motion model to describe each object and produces relatively good results. Immediately afterward, Anton et al. [10] segmented moving objects by fitting a motion-compensated model to events caused by the background and then detecting events that were inconsistent with the background, and they tested the method in challenging scenes (HDR, high speed) that are difficult to capture with frame-based cameras and published the dataset.

More and more machine learning methods have recently been used for motion segmentation tasks. Anton et al. [13] proposed a motion segmentation framework

to jointly estimate optical flow, 3D motion and objects segmentation tasks and published the dataset for motion object segmentation. Later on, they used graph convolutional neural networks to segment the 3D events cloud [12], effectively improving problems such as occlusion and demonstrating that larger temporal slices can produce better results. In the same year Daniel et al. [7] proposed a visual motion network that predicts more accurate local visual motion and confidence levels as a way to achieve motion segmentation and camera pose estimation.

2.2 Visual Transformer

Through the attention mechanism, the Transformer architecture excels at modeling long-term relationships in input sequences. The Transformer module, like non-local neural networks, computes correlations with all input elements and aggregates their information using an attention mechanism. Transformer networks, when compared to RNNs, model global correlations in parallel, improving memory efficiency. They were originally used for language tasks but have since been applied to popular computer vision problems such as object segmentation [22] and object detection [1,9].

We use an attention mechanism in this work to match object features and pass segmentation masks from previous frames to the current frame. A long and short-term temporal-spatial attention is also designed to allow for efficient feature matching and propagation.

3 Methodology

3.1 Input Representation

From the perspective of perception principles, the frame-based camera records the intensity of all pixels by means of frames to capture the global scene. In contrast, the event-based camera asynchronously measures the light intensity changes in the scene. When the change in light intensity is greater than a threshold value, the pixel triggers an event independently. The polarity of the event reflects the direction of the change. As shown in Eq. 1, an event can be defined as

$$\varepsilon = \{e_k\}_{k=1}^{N} = \{[x_k, y_k, t_k, p_k]\}_{k=1}^{N} \tag{1}$$

where e_k denotes an event. (x_k, y_k) denotes the pixel location of the event. t_k denotes the timestamp of the event. $p_k \in \{-1, +1\}$ is the polarity of the event, with a positive polarity indicating an increase in light intensity and a negative polarity indicating a decrease in light intensity.

Since the asynchronous events format is very different from synchronous frames, in order to accommodate the CNN input, previous approaches typically aggregate events into a frame-based representation. In this work we divide the event stream into successive time slices. Each time slice is projected onto a plane. The representation has three channels, two of which are accumulations of

positive and *negative* events, and the third is a *temporal image* [26]. Compared to the 3D learning approach, the 2D input representation has the advantage of reducing the sparsity of the data, thus improving computational efficiency.

3.2 Network Architecture

The overall architecture of the FusionSeg is illustrated in Fig. 1. The first part of our approach is the feature fusion module, which is used to balance the advantages of two domain features. To accommodate multi-objects scenarios, we use an identification mechanism to associate multiple targets. In addition, based on the identification mechanism, we design a long and short-term temporal attention for feature matching and propagation. Finally, the prediction mask is output by the MLP layer and decoder.

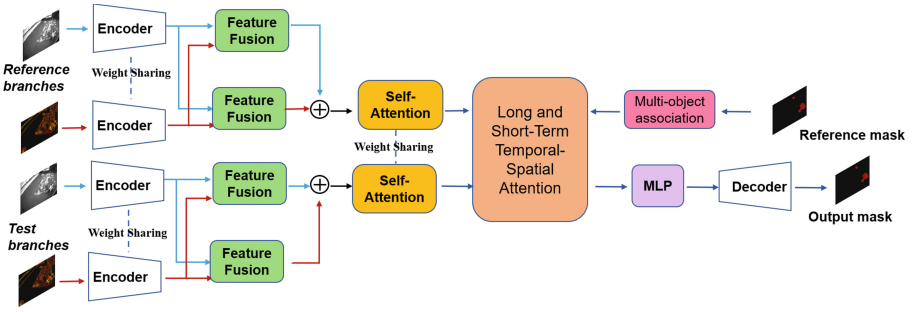

Fig. 1. An overview of our proposed Segmentation framework(FusionSeg) via collaboration of frames and events.

3.3 Feature Fusion Method

While frame-based cameras can easily capture rich textural and semantic cues, event-based cameras can easily capture edge information and have a high dynamic range. A feature fusion module is therefore designed to make effective use of both domain data. In the case of motion segmentation, objects are only detected when they are moving independently relative to the camera. Thus, previous work has attempted to compensate for camera motion. Instead of estimating the camera motion explicitly, we normalize the instances for each channel of each sample. Intuitively, the average activation tends to be controlled by motion in a large homogeneous region (usually the background). This normalization, combined with RELU, helps to separate background motion from foreground motion. As shown in the Fig. 2, the following feature enhancement scheme has been defined to generate enhanced features $\hat{F}_e$ for events.

$$\hat{F}_e = \hat{F}_{e\to e} \oplus \hat{F}_{v\to e} \oplus F_e \tag{2}$$

$$\hat{F}_{e\to e} = \sigma\left(\psi_{3\times 3}\left(F_e\right)\right) \otimes F_e \tag{3}$$

$$\hat{F}_{v\to e} = \sigma\left(\psi_{1\times 1}\left[\xi\left(\psi_{1\times 1}\left(F_v\right)\right), \xi\left(\psi_{3\times 3}\left(F_v\right)\right), \xi\left(\psi_{5\times 5}\left(F_v\right)\right)\right]\right) \otimes F_e \tag{4}$$

$$W_{F_e} = \sigma\left(\psi_{1\times 1}\left(\xi\left(\psi_{1\times 1}\left(\mathcal{A}\left(\hat{F}_e\right)\right)\right)\right)\right) \tag{5}$$

where $[\cdot]$ indicates channel-wise concatenation. ψ is convolutional layer. ξ is the instance normalization followed by a ReLU activation function. $\hat{F}_{e\to e}$ indicates event-based self-reinforcing features. $\hat{F}_{v\to e}$ indicates frame-based cross-domain reinforced features designed to enhance event-based features. W_{F_e} indicates the weight of event-enhanced features.

Inversely, the enhanced features of frames $\hat{F}_v$ and weight W_{F_v} can be generated. To balance the contributions of frames and events, inspired by [28], we propose an adaptive weighting balancing scheme:

$$X = W_{F_e}\hat{F}_e \oplus W_{F_v}\hat{F}_v \tag{6}$$

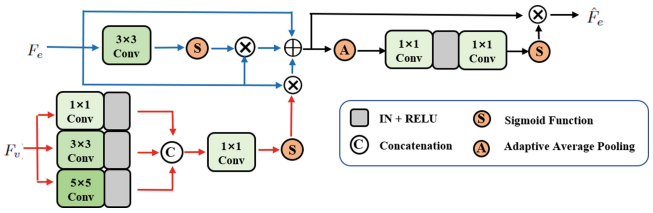

Fig. 2. The network structure of feature fusion network.

3.4 Multi-Object Association

The main challenge in propagating and decoding multi-objectes mask information in an end-to-end network is to use the network to accommodate different numbers of objects. To overcome this problem, inspired by [25], we use an identification mechanism consisting of **identity embedding** and **identity decoding**. For a multi-objects scenario, the identity embedding is constructed by assigning different identification vectors to different object regions. Specifically, we initialize a vector bank $D \in R^{M\times C}$. M denotes the maximum number of objects. $Y \in \{0,1\}$ is the mask of objects. Suppose there are $N\,(N < M)$ objects in the scenario. Then the identity embedding $E \in R^{THW\times C}$ can be expressed as

$$E = ID\left(Y, D\right) = YPD \tag{7}$$

where $P \in \{0,1\}^{HW\times C}$ is the random permutation matrix. After ID assignment, different objects have different recognition embedding vectors.

Identity decoding: The convolution's decoding network is used to predict the probability of each identity in the identity bank before selecting the specified identity and calculating the probability. Common multi-classes segmentation losses, such an cross-entropy losses, are used during training to optimize multiple objects with respect to ground-truth labels. The identity bank is trainable and is randomly initialized at the start of training. We re-initialize the random permutation matrix at each sequence sample and optimization iteration to ensure that all identity vectors have the same chance of competing with each other.

3.5 Feature Matching and Propagation

Based on the identification mechanism, we elaborate the long and short-term temporal-spatial attention for feature matching and propagation. The network

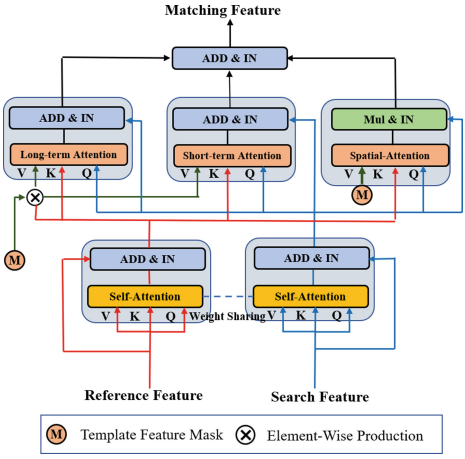

Fig. 3. The network structure of long and short-term temporal-spatial attention

starts with a self-attention layer that learns the correlation between objects in template frames and search frames, respectively. Then, long-term attention is introduced to aggregate features of template frames, short-term attention is introduced to learn memory frame features, and spatial attention is introduced to propagate memory frame object location information. Figure 3 illustrates the structure of our approach.

We define $Q \in R^{HW \times C}$, $K \in R^{HW \times C}$, $V \in R^{HW \times C}$ as the query embedding of the search frame, the key embedding of the template frame and the value embedding of the template frame, respectively. Where T, H, W, C denote the time, height, width and channel dimensions, respectively. The following is the attention-based matching and propagation equation.

$$Att\left(Q, K, V\right) = Corr\left(Q, K\right) V = softmax\left(\frac{QK^T}{\tau}\right) V \tag{8}$$

where $Corr(\cdot,\cdot)$ is the correlation function. τ is a temperature parameters controlling the softmax distribution, which is inspired by [5].

It is beneficial to propagate a representation of the objects when the camera changes dramatically in the scene. To suppress the background region, we multiply the template frame features by the mask recognition vector pixel by pixel, which propagates the template frame object features to the search frame.

$$V' = Att(Q, K, V \otimes ID(Y, D)) = Att(Q, K, V \otimes E) \tag{9}$$

where $\otimes$ is the pixel multiplication.

The self-attention operation is designed to enhance the features of the template frame and the search frame.

$$AttSA\left(X^t, X^t, X^t\right) = Att\left(X^t W^Q, X^t W^K, X^t W^V\right) \tag{10}$$

$$\hat{S}_{sa} = Ins.Norm\left(AttSA\left(X^t, X^t, X^t\right) + X^t\right) \tag{11}$$

where X^t is the template feature or search feature. $Ins.Norm(\cdot)$ indicates instance normalization. W^Q, W^K, W^V represent the projection matrix for matching and propagation.

Long-term attention is in charge of aggregating the features of the template frame objects to the current frame. Temporal smoothness is difficult to achieve because the time interval between the search frame and the template frame is variable. Therefore, a non-local attention approach is used to implement the long-term attention module. The following is the equation of long-term attention.

$$AttLT\left(X^t, X^m, Y^m\right) = Att\left(X^t W^Q, X^m W^K, X^m W^V \otimes E^m\right) \tag{12}$$

$$\hat{S}_{lt} = Ins.Norm\left(AttLT\left(X^t, X^m, X^m\right) + X^t\right) \tag{13}$$

where X^t is the search feature. $t \in \{2, \ldots, T\}$ is the feature index of sequence. $m \in \{1\}$ is the index of first frame. X^m and Y^m are the template feature and masks, respectively.

Short-term attention is in charge of aggregating the memory frame's objects features to the current frame. Changes between successive time slices appear to be smooth and continuous. As a result, object matching and propagation can be limited to a small spatiotemporal domain, providing greater efficiency than non-local attention.

$$AttST\left(X^t, X^n, Y^n \mid p\right) = Att\left(X_p^t W^Q, X_{N(p)}^n W^K, X_{N(p)}^n W^V \otimes E_{N(p)}^n\right) \tag{14}$$

$$\hat{S}_{st} = Ins.Norm\left(AttST\left(X^t, X^n, Y^n \mid p\right) + X^t\right) \tag{15}$$

where X_p^t is the search feature at location p. $N(p)$ is 15×15 spatial neighbourhood centered at location p. $n \in \{t-1\}$ is the memory feature index of sequence. $X_{N(p)}^n$ and $E_{N(p)}^n$ are the memory feature and masks of the spatial-temporal neighbourhood, respectively.

Spatial attention is in charge of aggregating the location information of memory frame objects to the current frame. The mutual attention establishes a pixel-to-pixel relationship between the two frames so that it supports object location propagation.

$$AttSP\left(X^t, X^n, E^n\right) = Att\left(X^t W^Q, X^n W^K, E^n W^V\right) \tag{16}$$

$$\hat{S}_{sp} = Ins.Norm\left(AttSP\left(X^t, X^n, E^n\right) \otimes E^n\right) \tag{17}$$

4 Experiment and Results

To explore the effectiveness of our proposed algorithm for the motion segmentation task, we validate it on two widely used and public datasets. Firstly, we present the implementation details. And then describe the use of the datasets. Finally, we show the performance of our method in both qualitative and quantitative terms.

4.1 Implementation Details

For feature extraction, we used MobileNet v3 [6] as the encoder and FPN [8] as the decoder to generate the objects' mask. The AdamW optimizer and a sequential training strategy [24] with a sequence length of 5 are used. The loss function is a combination of bootstrapped cross-entropy loss and soft Jaccard loss [15]. We used an exponential moving average (EMA) [17]. The initial learning rate was set to 0.0002. To reduce overfitting, the encoder's initial learning rate was reduced to 0.1 of the other network parts.

4.2 Overview of Datasets

The EV-IMO dataset [14] is a real dataset captured indoors with the DAVIS 346 camera that includes backgrounds like *box*, *floor*, *table*, and *wall* as well as multiple independently moving objects. It is one of the most challenging open-source datasets for segmenting independent motion objects. The authors provide dense segmentation masks of independently moving objects for quantitative evaluation. Two standard metrics are used in the quantitative evaluation, including detection rate and Intersection over Union (IoU). Details about the metrics can be found in [11,29].

The Extreme Event Dataset (EED) [11] is one of the first open-source datasets used for independent moving objects detection and tracking research. There are independent moving objects in addition to the camera's ego-motion. All sequences were captured in a laboratory setting to demonstrate the superior performance of event cameras in HDR scenes.

Table 1. Segmentation performance on the EVIMO dataset, measured by IoU.

	EVDodgeNet [18]	MOMS [16]	EMSGC [29]	EV-IMO [14]	AOT [25]	**Ours**
EVIMO	65.76	74.82	76.81	77.00	31.82	**77.49**

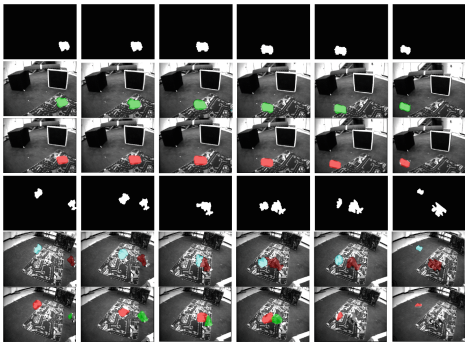

Fig. 4. Segmentation results on the EVIMO dataset, on sequences *Box*(rows 1–3) and *Table*(rows 4–6).Time runs from left to right. Our segmentation masks(rows 2 & 5) are shown on the frames. The masks from [29] (rows 3 & 6) are also shown on the frames. The first and fourth row are ground truth labels

4.3 Discussion of Results

As can be seen from the quantitative results in Table 1 , our method outperforms other state-of-the-art solutions. Moreover, due to the drastic changes in the scene recorded in the EVIMO dataset, the AOT [25] using only frames could not achieve satisfactory results, suggesting that exploiting the complementary of events and frames can improve the robustness of the model under degraded conditions.

Table 2. Segmentation performance on the EVIMO dataset, measured by detection rate.

Algorithm	Mitrokhin [10]	MOMS [16]	**Ours**
Detection rate	64.79	77.06	**80.74**

As shown in Table 2, our method outperforms other methods using the detection rate metric. Due to the lack of open source code, the numbers for the baseline method were obtained directly from the corresponding publications [16]. Figure 4 shows the example results of our method on EVIMO. In the *Box* sequence, The toy car moves from right to left on a highly textured carpet with multiple stationary interfering objects in the scene, and our model can still continuously detect it as a moving object. In the *Table* sequence, there are two Independent moving objects, which hit each other and meet in the middle. The toy plane moves slowly at the end of the sequence and [29] marks it as background. Even though

they partially overlap, our method successfully segments out the independently moving plane.

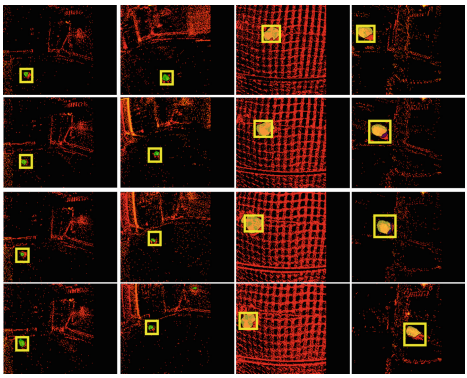

Fig. 5. Segmentation results on EED dataset. Time runs from top to bottom. The ground truth bounding boxes (in yellow) show the 2D location of the independent moving objects. column 1 and column 2 are recorded in low light conditions. column 3 and column 4 are recorded in the presence of obscuration. (Color figure online)

We ablate the main modules of our method. The key components of our method are feature fusion, long-term attention, short-term attention, and spatial attention. To verify their effectiveness, we modified the original model by removing each component and retrained the modified model, accordingly, we obtained four models. Table 3 reports the results of the four modified models. It shows that the feature fusion module is the key to our excellent results. The long and short-term spatio-temporal attention module does effectively match and propagate the target features and location information. This reflects the fact that our method can improve the model's segmentation ability by combining events and frames and using the attention mechanism to learn the motion information in the sequence, demonstrating that our method is robust to the shape, size, and the number of objects.

In addition to the quantitative results above, we also show example results from the EED dataset in Fig. 5. Note that it is sometimes difficult to detect the objects in the corresponding frames, which motivates us to combine frames and events and use motion cues for independent moving objects detection.

Table 3. Ablation studies on feature fusion(FF), long-term attention(LT), short-term attention(ST) and spatial-attention(SP).

Method	FF	LT	ST	SP	IoU
Ours-A	×	✓	✓	✓	75.99
Ours-B	✓	×	✓	✓	74.97
Ours-C	✓	✓	×	✓	75.94
Ours-D	✓	✓	✓	×	76.94
FusionSeg	✓	✓	✓	✓	**77.49**

5 Conclusions and Future Work

This paper presents a multi-objects motion segmentation method using frames and events. We design a feature fusion scheme that effectively fuses the information obtained from the frames and events. In addition, we introduce the Transformer architecture to make full use of the motion cues for motion segmentation of multi-objects scenes.

Our method can address slow object motion and highly textured scenes through feature matching and propagation, demonstrating that exploiting the complementary of events and frames can improve the robustness of motion segmentation under degraded conditions. All these allow us to perform motion segmentation in challenging scenes, thus unlocking the remarkable capabilities of the event camera. In the future, we will investigate the feasibility of exploiting the high measurement rate of the event camera to increase the segmentation frequency.

Acknowledgements. This work was partially supported by the National Natural Science Foundation of China(No. 91948303).

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
2. Charig Yang, H.L., Lu, E., Zisserman, A., Xie, W.: Self-supervised video object segmentation by motion grouping (2021)
3. Gallego, G., Rebecq, H., Scaramuzza, D.: A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3867–3876 (2018)
4. Glover, A., Bartolozzi, C.: Robust visual tracking with a freely-moving event camera. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3769–3776. IEEE (2017)

5. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 2(7) (2015)
6. Howard, A., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)
7. Kepple, D.R., Lee, D., Prepsius, C., Isler, V., Park, I.M., Lee, D.D.: Jointly learning visual motion and confidence from local patches in event cameras. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12351, pp. 500–516. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58539-6_30
8. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
9. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: a survey. Int. J. Comput. Vis. **128**(2), 261–318 (2020)
10. Mitrokhin, A., Fermüller, C., Parameshwara, C., Aloimonos, Y.: Event-based moving object detection and tracking. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1–9. IEEE (2018)
11. Mitrokhin, A., Fermüller, C., Parameshwara, C., Aloimonos, Y.: Event-based moving object detection and tracking. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1–9. IEEE (2018)
12. Mitrokhin, A., Hua, Z., Fermuller, C., Aloimonos, Y.: Learning visual motion segmentation using event surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14414–14423 (2020)
13. Mitrokhin, A., Ye, C., Fermüller, C., Aloimonos, Y., Delbruck, T.: Ev-imo: Motion segmentation dataset and learning pipeline for event cameras. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 6105–6112. IEEE (2019)
14. Mitrokhin, A., Ye, C., Fermüller, C., Aloimonos, Y., Delbruck, T.: Ev-imo: Motion segmentation dataset and learning pipeline for event cameras. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 6105–6112. IEEE (2019)
15. Nowozin, S.: Optimal decisions from probabilistic models: the intersection-over-union case. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 548–555 (2014)
16. Parameshwara, C.M., Sanket, N.J., Gupta, A., Fermuller, C., Aloimonos, Y.: Moms with events: Multi-object motion segmentation with monocular event cameras. arXiv preprint arXiv:2006.06158 **2**(3), 5 (2020)
17. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM J. Contr. Optimization **30**(4), 838–855 (1992)
18. Sanket, N.J., et al.: Evdodgenet: Deep dynamic obstacle dodging with event cameras. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 10651–10657. IEEE (2020)
19. Stoffregen, T., Gallego, G., Drummond, T., Kleeman, L., Scaramuzza, D.: Event-based motion segmentation by motion compensation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7244–7253 (2019)
20. Stoffregen, T., Kleeman, L.: Simultaneous optical flow and segmentation (sofas) using dynamic vision sensor. arXiv preprint arXiv:1805.12326 (2018)
21. Vasco, V., Glover, A., Mueggler, E., Scaramuzza, D., Natale, L., Bartolozzi, C.: Independent motion detection with event-driven cameras. In: 2017 18th International Conference on Advanced Robotics (ICAR), pp. 530–536. IEEE (2017)

22. Wang, Y., et al.: End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8741–8750 (2021)
23. Wertheimer, M.: Untersuchungen zur lehre von der gestalt. Psychologische forschung **1**(1), 47–58 (1922)
24. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by foreground-background integration. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 332–348. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_20
25. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
26. Ye, C., Mitrokhin, A., Fermüller, C., Yorke, J.A., Aloimonos, Y.: Unsupervised learning of dense optical flow, depth and egomotion from sparse event data. arXiv preprint arXiv:1809.08625 (2018)
27. Zhang, J., Shi, F., Wang, J., Liu, Y.: 3D motion segmentation from straight-line optical flow. In: Sebe, N., Liu, Y., Zhuang, Y., Huang, T.S. (eds.) MCAM 2007. LNCS, vol. 4577, pp. 85–94. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73417-8_15
28. Zhang, J., Yang, X., Fu, Y., Wei, X., Yin, B., Dong, B.: Object tracking by jointly exploiting frame and event domain. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13043–13052 (2021)
29. 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)