Active Object Segmentation: A New Modality for Egocentric Action Recognition

Jian Ma
Tianjin University
China
jianma@tju.edu.cn

Bin Zhu
Singapore Management University
Singapore
binzhu@smu.edu.sg

Kun Li
Tianjin University
China
lik@tju.edu.cn

Dima Damen
University of Bristol
United Kingdom
dima.damen@bristol.ac.uk

Abstract

Egocentric actions typically exhibit Human-Object Interactions (HOIs), involving the transformation of objects (e.g., "*cutting*" an "*onion*") using various tools and utensils (e.g., "*knife*" and "*chopping board*"). Recognising these actions requires networks to model object transformations by understanding the presence, relative positioning, and motion of the objects involved. However, current methods for **e**gocentric **a**ction **r**ecognition (EAR) generally do not explicitly model active objects. In this paper, we introduce "active object segmentation" as a new modality. Leveraging EPIC-VISOR annotations from EPIC-KITCHENS 100, we: (a) present an indirect method for obtaining RGB-like representations from segmentations via a generative model; (b) propose a direct method that directly learns presence, relative positioning, and motion representations from segmentations via a multi-stream network; (c) introduce an analogous network to gain regions of active objects and hands from RGB frames rather than active object segmentation input to enhance EAR. Experiments show that active object segmentation is more robust in recognising transformed objects during interactions. The proposed networks (a and b) with active object segmentation inputs boost action recognition accuracy by at least 3%, with network (b) improving noun recognition by around 10%. Method (c), using RGB frame input, enhances noun and action recognition accuracies while helping action networks focus more on interactions.

CCS Concepts

• **Computing methodologies** → **Activity recognition and understanding**.

Keywords

Egocentric Vision, Semantic Segmentation, Action Recognition

MMASIA '24, December 03–06, 2024, Auckland, New Zealand

ACM ISBN 979-8-4007-1273-9/24/12
https://doi.org/10.1145/3696409.3700164

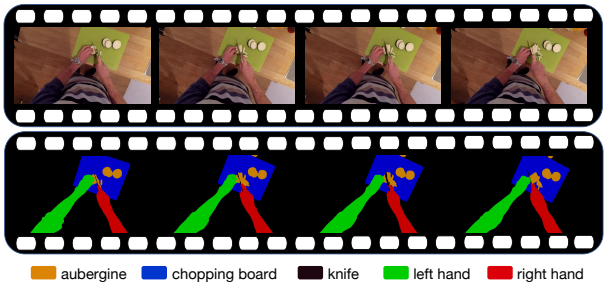

Figure 1: A demonstration of active object segmentation. RGB frames are shown at the top while segmentations of active objects and interacting hands are illustrated at the bottom. This clip showcases an action of "*cut aubergine*".

1 Introduction

Action recognition is a widely studied task in video understanding, with the aim of identifying the actions that appear in a short clip. A video can be viewed as a data structure composed of a set of image frames (in RGB) arranged in temporal order. However, key interacting elements appearing in actions, such as active objects, are highly coupled with non-interacting information in RGB. Hence, it would cater to non-interacting information in action recognition [3].

Recently, many novel works have used other modalities to study videos. Optical flows are particularly efficient for capturing moving objects and detecting active objects. Some works [9, 30] use only optical flow as input, while others [1, 26] use it as supplementary information to address RGB's sensitivity limitations to moving objects. However, optical flow is not ideal for egocentric videos due to the prevalence of head motion over active object motion. Additionally, although some works [32, 33] enhance spatial-temporal features through learned motion-relevant regions, these regions only indicate the location of interacting objects, not their states. In egocentric videos, objects often undergo transformations, such as cutting "sliced aubergine" into "diced aubergine" (Fig. 1). These transformations are common in egocentric interactions, posing a challenge for accurate action recognition. This requires methods to understand the presence, relative positioning, and motion of the objects involved in interactions.

In this work, we introduce a novel modality termed "active object segmentation" to enhance EAR. We build on EPIC-KITCHENS VISOR [5], which provides fine-grained segmentation annotations

of interacting hands and active objects, detailing object transformations during interactions. To evaluate this, we propose: (a) an indirect method, using a generative model (SPADE [22]) to obtain RGB-like representations from segmentations for action recognition models. Averaging the prediction scores with a baseline (Video Swin Transformer [19]) improves accuracy by 3%. (b) A direct method adds a segmentation-aware stream (based on TSN [29]), showing that object segmentation is more robust for recognizing transformed objects, boosting noun accuracy by 10% and overall recognition by 7.6% when fused with RGB and optical flow. (c) To address the lack of ground-truth segmentations, we use SPADE to infer active object regions from RGB frames, enhancing the RGB representation in models like SlowFast [8] and Video Swin Transformer [19], improving noun and verb recognition by capturing object presence, positioning, and motion.

Our contributions can be summarised as follows:

- We introduce "active object segmentation" as a new modality for EAR, which is superior in capturing object transformation and interaction than RGB and optical flow.
- We present an indirect method using generative models to obtain RGB-like representation from active object segmentation. This representation enhances action recognition accuracy by 3%.
- We propose a direct method with a segmentation-aware stream to directly model active object segmentation input. This stream boosts noun recognition by 10% and action accuracy by 7.6%.
- We adopt SPADE to model object presence, positioning, and motion from RGB frames. The latent could improve recognition without ground-truth active object segmentation.

2 Related Work

2.1 Modalities in Action Recognition

The need for higher accuracy and the development of affordable sensors have led to alternative approaches utilising various modalities alongside RGB. Skeleton data [18] efficiently predicts non-object-interacting actions like jumping or hugging by reflecting key human body nodes. Point cloud [24] and depth [34] capture 3D information and geometric shapes, making them useful in autonomous vehicles. Infrared data [13] visualises dark environments in low-light conditions. Event streams [9] extract action-related foregrounds to mitigate messy backgrounds. Invisible modalities, such as audio [17], radar [16], and WiFi [28], offer privacy protection. Audio locates actions in temporal sequences, while radar and WiFi efficiently explore scenarios obstructed by walls.

2.2 Fusion Methods in Action Recognition

Other modalities can enhance video understanding accuracy, often by fusing with RGB frames through concatenation. For instance, [36] proposes a two-stream network combining RNN and CNN for skeleton and RGB data, outperforming score fusion. [11] extracts temporal features from RGB, skeleton, and depth modalities, concatenating them to capture time-varying information. [6] uses a 3D CNN and a 2D CNN to process skeleton, infrared, and RGB data, with a multi-layer perceptron for human action recognition. Non-visual modalities are also employed: [21] designs a dual-stream network for audio and RGB frames, while [27] and [15] use three-stream networks for audio, RGB, and optical flow.

Transformer-based networks also advance multi-modality representation learning for action recognition. Perceiver [12] employs cross-attention to fuse multiple modalities. Fluxformer [10] uses a duplex attention mechanism on optical flow and RGB to improve temporal information. [14] develops a Two-Pathway Vision Transformer to integrate RGB and skeletal data, with separate paths for spatial and temporal information. [35] introduces an audio adaptive encoder to enhance visual features using sound. Large language models like VideoLLM [2] show improvements in multimodal models for continuous video streams. Despite these advancements, perceiving object transformations remains challenging across modalities like audio, depth maps, and optical flow. This paper introduces "active object segmentation" to enhance EAR, demonstrating greater robustness in recognising transformed objects during interactions.

3 Method

3.1 EAR from Active Object Segmentation

3.1.1 Indirect Manner: RGB-like Embedding. As we all know, active object segmentation includes relative positioning, and shapes of objects relevant to the action, but excludes colours and textual details. Inspired by the success of RGB frames in action modelling, we adopt SPADE [22], a generative model, to model actions from the segmentations. Fig. 2 shows a two-stage framework and it firstly trains SPADE with active object segmentation inputs to reconstruct fine-grained images, learning RGB-like representations. The encoder learns a representation of the segmentation map using convolutional layers. The decoder comprises five upsampling layers and six SPADE Res-blocks. Each SPADE Res-block consists of spatially-adaptive normalization, activation function, and convolutional layers. These blocks are stacked and connected through residual connections, enabling the model to effectively learn and represent the relationship between images and their corresponding semantic segmentations. Subsequently, we freeze SPADE and only train the action network (Video Swin Transformer [19] for example) based on RGB-like representations to predict the action.

The RGB-like representations generated by SPADE [22] can be integrated into any network designed for RGB-based action recognition. In this paper, we adopt the Video Swin Transformer [19], using the RGB-like representations from SPADE as input. As illustrated in the red part of Fig. 2, these representations are fed into the 3D patch partition module, followed by Video Swin Transformer blocks with linear embedding. In addition to recognizing actions from active object segmentation alone, the proposed network can be extended to multi-modality input through early and late fusion strategies. In early fusion, features from Video Swin with RGB frames are combined with SPADE's representations. For late fusion, we average the scores from two separate Video Swin networks—one trained on RGB frames and the other on active object segmentation—to enhance overall action recognition.

3.1.2 Direct Manner: Semantic Segmentation Embedding. Unlike the indirect manner introduced above, we directly model active object segmentations via a one-stage framework. Inspired by the success of multi-stream networks modelling multi-modality in action recognition, we adopt TSN [31] to model egocentric action

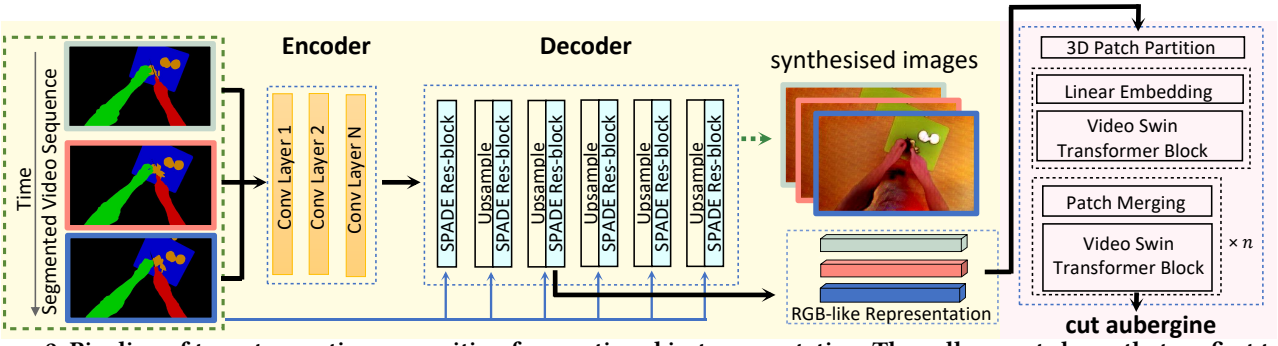

Figure 2: Pipeline of two-stage action recognition from active object segmentation. The yellow part shows that we first train SPADE [22] with active object segmentation inputs to reconstruct fine-grained images, learning RGB-like representations. Subsequently, we freeze SPADE and only train the action network (Video Swin Transformer [19] for example) based on RGB-like representation inputs to predict the action as the red part shown. The black bold arrows demonstrate the main data flow.

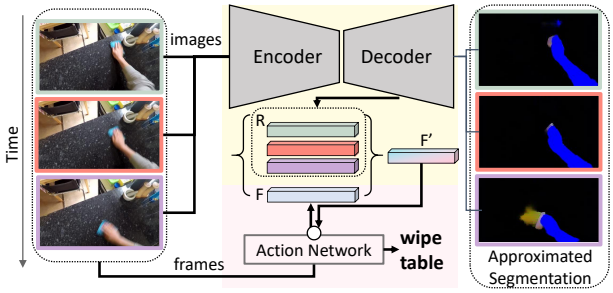

Figure 3: Pipeline of action recognition by approximated segmentations through RGB frames Input. The white circle in the action network refers to the layer from which the frame features are extracted and where these features are concatenated with the representation obtained from SPADE [22].

from RGB frames, optical flows and active object segmentations, respectively. We keep the standard structure of RGB stream and flow stream in TSN while designing a new stream called segmentation-aware stream to recognise actions from active object segmentations. The structure of the new stream is similar to the flow stream and the main difference is the input size. The flow stream takes optical flows whose channel size is 2 while the channel size of segmentation-aware stream input is 305. We average the scores of predictions from different brunches, yielding final predictions.

3.2 EAR from Approximated Active Object Segmentation

Obtaining real segmentations can be quite resource-intensive. We attempt to obtain the approximated active object segmentation from RGB frames. In order to ensure the uniformity of the framework, we adopt SPADE to predict approximated segmentations from RGB inputs. The input to this network consists of RGB images, while the output is the approximated segmentations. This network is trained using ground-truth segmentations but is evaluated solely using RGB images. By incorporating this network, we aim to leverage the power of RGB input while benefiting from the additional information provided by approximated segmentations for improved action recognition performance.

Fig. 3 illustrates the overall architecture of fusing the approximated segmentations with RGB frame inputs. The latent $R \in \mathbb{R}^{\dot{C}\times\dot{W}\times\dot{H}\times T}$ from the decoder are extracted and stacked over time, followed by max pooling to adjust the dimension sizes. Moreover, we extract the RGB frame representation $F \in \mathbb{R}^{\dot{C}'\times\dot{W}'\times\dot{H}'\times T}$ from the selected layer of the action network. We borrow the idea of concatenating different modalities for the fusion from previous works [20, 23]. Following this, we concatenate the RGB frame representation with the latent from SPADE to form $F' \in \mathbb{R}^{(\dot{C}+\dot{C}')\times\dot{W}'\times\dot{H}'\times T}$ and adopt 1x1x1 3D CNN to linearly map F' to $F'' \in \mathbb{R}^{\dot{C}'\times\dot{W}'\times\dot{H}'\times T}$ for subsequent blocks of the action network.

4 Experiment

4.1 Dataset and Evaluation

In this paper, we introduce **V**ISOR-**C**LIP-**A**CTION (*VCA*) dataset containing active object segmentation from EPIC-KITCHENS VISOR dataset [5] and corresponding RGB frames and optical flows from EPIC-KITCHENS 100 [4]. *VCA* consists of 17,654 action clips with segmentations for training and 2,114 action clips with segmentations for evaluation. *VCA* dataset is for all proposed methods while we evaluate EAR from approximated active object segmentation on the entire EPIC-KITCHENS 100. The evaluation metrics used to assess the performance of the networks in this study are Top-1/5 accuracies for verbs, nouns, and actions.

4.2 Implementation Details

SPADE. Before training the proposed method, we need to train two distinct SPADE models [22] based on *VCA*. The input size of the first SPADE with segmentation input is $H\times W\times T\times 305$, which represents the number of active object classes. The input size of the second SPADE with RGB input is $H\times W\times T\times 3$. Whether segmentations or RGB frames are input to SPADE, their representation dimensions of decoders are the same, i.e. $latent \in \mathbb{R}^{1024\times 8\times 8}$, $up1 \in \mathbb{R}^{1024\times 16\times 16}$, $up2 \in \mathbb{R}^{512\times 32\times 32}$, $up3 \in \mathbb{R}^{256\times 64\times 64}$, $up4 \in \mathbb{R}^{128\times 128\times 128}$, $up5 \in \mathbb{R}^{64\times 256\times 256}$. We fully follow the original training setup described in [22] and once SPADE is trained, it is frozen.

Video Swin Transformer [19]. For inferring actions from segmentations, we adopt $up3 \in \mathbb{R}^{256\times 64\times 64\times 32}$ from SPADE and this representation is processed to form patches $P \in \mathbb{R}^{128\times 56\times 56\times 16}$ by 3D CNNs. Patches are then input to Video Swin Transformer. For the inferring actions from approximated segmentations, another Video

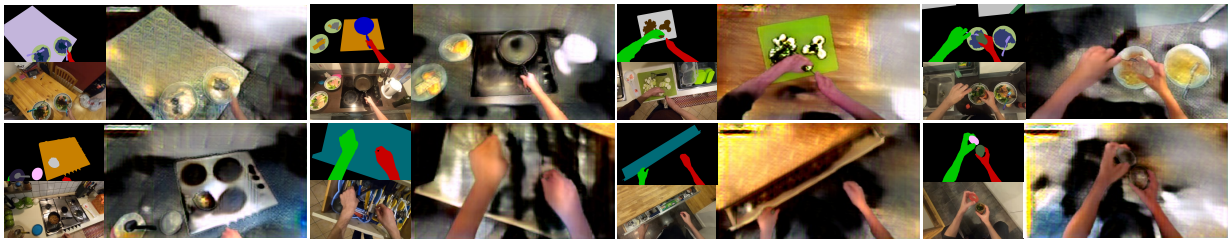

Figure 4: Visualisations of SPADE [22] on EPIC-KITCHENS VISOR [5]. We show the input segmentations (left top) used to produce synthesised images (right). These are compared to the corresponding real RGB frames (left bottom).

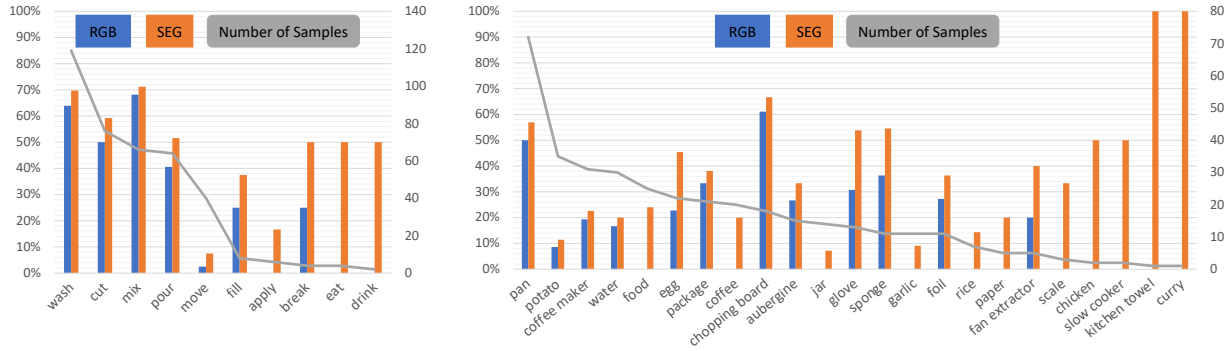

Figure 5: More accurate Verb (Left) and noun (Right) predictions in segmentations performance along with numbers of samples.

Swin Transformer is trained with RGB frames $\in \mathbb{R}^{3\times224\times224\times32}$ and we extract the features $F \in \mathbb{R}^{1024\times7\times7\times16}$ from its last layer. Meanwhile, the representation of predicting segmentations $latent \in \mathbb{R}^{1024\times8\times8\times32}$ is pooled to $R' \in \mathbb{R}^{1024\times7\times7\times16}$. Thus the concatenated features $F' \in \mathbb{R}^{(1024+1024)\times7\times7\times16}$ are fed into a 1x1x1 3D convolutional layer to form $F'' \in \mathbb{R}^{1024\times7\times7\times16}$ and then F'' is passed to a classifier for action recognition. Both networks, fine-tuned from Something-Something v2, are optimized with ADAMW, using a cosine schedule, and trained for 60 epochs.

SlowFast [7]. SlowFast [8] with Res-50 takes 8, 32 RGB frames with the spatial size 224, i.e. $T = 8$ or $T = 32$, $W = 224$ and $H = 224$, for the slow and fast pathways. The fusions happen in both slow and fast pathways. The representation of predicting segmentations $up3 \in \mathbb{R}^{256\times64\times64\times T}$ is pooled to $R' \in \mathbb{R}^{256\times56\times56\times T}$ and then it is concatenated with the RGB representation $F \in \mathbb{R}^{320\times56\times56\times T}$. Afterward, the fused features $F' \in \mathbb{R}^{(320+256)\times56\times56\times T}$ are then convoluted to $F'' \in \mathbb{R}^{320\times56\times56\times T}$ for the rest blocks of SlowFast. The trained setups are the same as the official.

Segmentation-aware Stream. The structure of the new stream is similar to the flow stream. We average the scores of predictions from different brunches, yielding final predictions. This network is trained from scratch using ADAM for 50 epochs, following the training setups of TSN [31].

4.3 Advances of EAR from Active Object Segmentation

4.3.1 Indirect Manner. In this section, we visualise the image synthesis performance of SPADE [22] and then analyse the performance of various modalities with different fusion strategies.

SPADE Visualisation. We qualitatively assess the performance of the trained SPADE [22] encoder-decoder by synthesising images from ground-truth segmentations, as shown in Fig. 4. We compare input segmentations from the test set with the corresponding real images. The model effectively synthesises segmented active objects with a reasonable background, consistently reconstructing hands and small objects like courgette pieces and knives. Overlapping objects are displayed hierarchically, such as food on a plate. However, objects not present in the segmentations receive limited supervision, and some images show overfitting elements from the training data, like a green chopping board or a wooden kitchen counter.

				Top-1 Accuracy (%)			Top-5 Accuracy (%)		
RGB	Seg	Early	Late	Verb	Noun	Act.	Verb	Noun	Act.
✓				55.72	42.95	30.84	86.14	65.09	59.41
	✓			45.69	31.03	21.10	82.26	53.41	48.11
✓	✓	✓		57.57	43.95	32.69	85.62	66.37	60.03
✓	✓		✓	**58.56**	**46.22**	**33.63**	**88.69**	**69.30**	**63.95**

Table 1: Quantitative results of early and late fusion of [19].

Quantitative Results. We compare Video Swin [19] using either RGB or segmentations as input, as shown in Tab.1. With segmentations, verb, noun, and action Top-1 accuracies are about 10% lower than with RGB, and verb Top-5 accuracy is 4% lower. However, segmentations outperform RGB for verbs like "*wash*", "*cut*", "*mix*", and "*pour*" (Fig.5 Left). Segmentations struggle more with nouns, showing a Top-5 accuracy gap of 12% compared to RGB, likely due to the absence of textures. Yet, segmentations better recognize objects like "*potato*", "*water*", "*egg*", and "*aubergine*", particularly for transformed or deformed objects (Fig.5 Right).

Although segmentation alone underperforms compared to RGB, fusing it with RGB significantly improves results. Early fusion boosts verb and noun accuracies by 1.8% and 1%, raising action accuracy by 1.8% to 32.69%. Late fusion shows even stronger gains, with verb accuracy increasing by 2.8%, noun accuracy by 3.3%, and action recognition reaching 33.63%, outperforming early fusion by 1% for verbs and actions, and 2.3% for nouns.

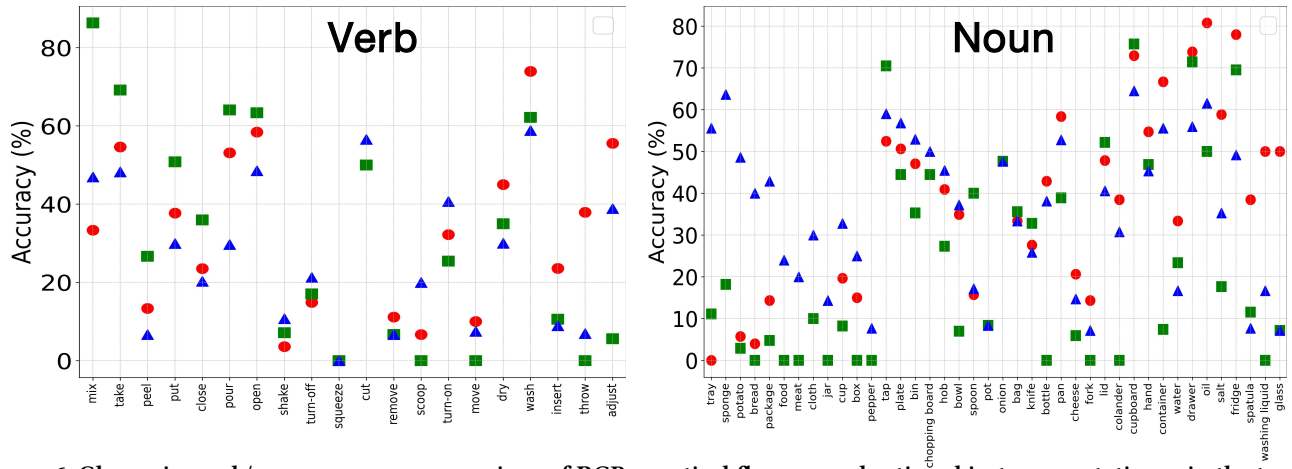

Figure 6: Class-wise verb/noun accuracy comparison of RGB ●, optical flows ■, and active object segmentation ▲ in the top-20 and top-40 frequency labels. This demonstrates while active object segmentation is limited in improving verb accuracy, it shows potential as a sole modality in recognising objects likely to undergo transformations, such as "*sponge*", "*potato*", *etc.*.

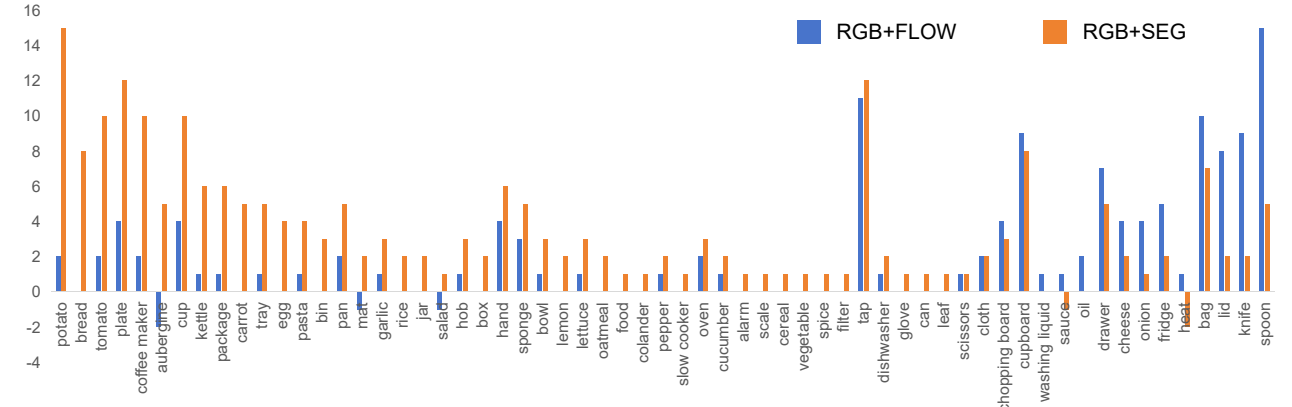

Figure 7: Increments of noun prediction by two different multimodal fusions in contrast to RGB single modality. This illustrates the fusion of RGB and active object segmentation predicts more correct nouns, particularly for objects prone to transformation, such as "*potato*", "*tomato*", *etc.*.

			Top-1 Accuracy (%)			Top-5 Accuracy (%)		
RGB	Flow	Seg	Verb	Noun	Act.	Verb	Noun	Act.
✓			40.35	32.97	**18.35**	**83.40**	**58.14**	**50.85**
	✓		**46.26**	24.97	15.70	82.31	43.90	39.88
		✓	34.11	**33.68**	16.79	75.31	53.45	44.18
✓	✓		**49.53**	37.84	23.75	**86.61**	60.12	54.45
✓		✓	42.24	42.05	22.47	84.15	66.32	58.23
	✓	✓	44.70	37.37	22.85	82.64	58.94	51.47
✓	✓	✓	48.53	**43.00**	**25.97**	85.24	**67.12**	**59.27**

Table 2: Quantitative results of different streams with different modalities.

4.3.2 Direct Manner. In this section, we compare the performance of active object segmentation with optical flows in aiding RGB-based action recognition in egocentric videos through the proposed multi-stream network.

Single Modality Comparison. The top three rows of Tab.2 show the accuracies for each stream. RGB frames achieve the best performance in action recognition, indicating their importance for this task. Active object segmentation underperforms in verb recognition compared to optical flows, which excel at identifying frequent verbs like "*mix*", "*take*", and "*put*" (Fig.6, left). However, active object segmentation outperforms optical flows in noun recognition, with a top-1 accuracy advantage of 9% over optical flows and 0.7% over RGB frames.

Interestingly, Fig.6 (right) reveals that active object segmentation excels in recognizing entities with mutable shapes, such as "*sponge*", "*potato*", "*food*", *etc.*. Notably, for "*potato*" segmentations outperform both RGB and optical flows by over 30%. These objects are particularly prone to transformation during interactions. The label "food" is unique, representing a mix of various processed objects, often resulting from multiple transformations. Fig.6 shows that active object segmentation predicts "food" more accurately than RGB or optical flow. Therefore, this experiment highlights that active object segmentation as a single input modality facilitates recognising transformed objects.

Multiple Modality Comparison. Fusion significantly enhances performance by combining representations from different modalities. Specifically, verb top-1 accuracy increases by 9%, and noun recognition jumps from 32.97% to 37.84% when RGB input is combined with optical flows. Additionally, integrating the proposed segmentation-aware stream with the spatial stream boosts noun

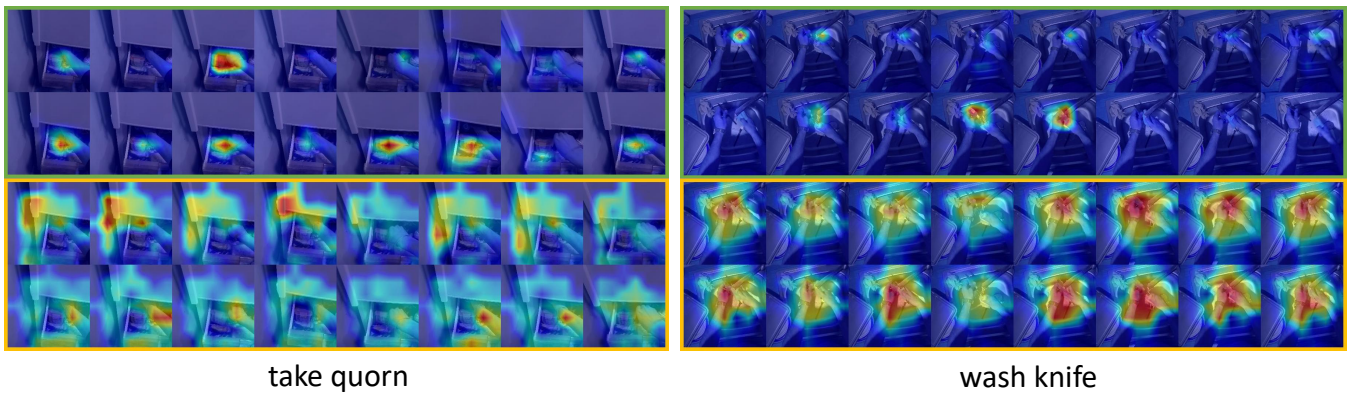

Figure 8: Visualisations of the heatmaps from SlowFast [8] (green box) and Video Swin [19] (yellow box). The top row displays networks without fused segmentations, while the bottom row features networks with concatenated segmentation inputs.

accuracy by 9% and verb recognition by 2%, leading to an overall action accuracy improvement from 18.35% to 22.47%.

Notably, in Fig. 7, we illustrate the gains in noun prediction from fusing RGB with active object segmentation and optical flow. This fusion notably improves predictions for objects like "*potato*", "*bread*", "*tomato*", "*aubergine*", which are often cut into pieces, and also enhances accuracy for less common objects such as "scale" and "alarm." Conversely, for objects like "*spoon*", "*knife*", "*lid*", "*fridge*", which undergo a minimal transformation, RGB and optical flow fusion yields more accurate predictions. This highlights how RGB and active object segmentation fusion improves noun recognition for transformed and less frequent objects.

Moreover, in Tab. 2, results of the fusion of the flow stream with the segmentation-aware stream outperforms the spatial streams, with top-1 accuracies increasing by about 4%. While optical flows excel in verb modelling and active object segmentation in noun prediction, their combination surpasses RGB alone by leveraging complementary information for enhanced action understanding in egocentric videos.

Furthermore, in the bottom row of Tab. 2, fusing all streams shows significant enhancements across all accuracies, with noun accuracy achieving the highest performance at 43%. While the verb prediction accuracy of the fusion of all streams is slightly lower compared to RGB and optical flow fusion, it still achieves robust action recognition, reaching an overall accuracy of 25.97%. This comprehensive fusion approach benefits from combining modalities to achieve more accurate and robust action recognition models.

4.4 Advances of Approximated Active Object Segmentation.

We evaluate different base networks with and without the fusion of the generated segmentations in Tab. 3. SlowFast [8] with approximated segmentations outperforms SlowFast [8] without fusion. In particular, top-1 verb accuracy is improved by 0.8% while the improvement of noun and action classification is about 1.4%, reaching 50.42% and 39.04%, respectively. Moreover, the fused Video Swin [19] shows better performance in noun prediction, increasing the top-1 noun, action accuracies by 0.26% and 0.03%. These results highlight the effectiveness of segmentation fusion in boosting egocentric action recognition across different backbones.

		Top-1 Accuracy (%)			Top-5 Accuracy (%)		
Network	Fusion	Verb	Noun	Act.	Verb	Noun	Act.
SlowFast [8]		65.13	49.02	37.61	89.84	74.57	59.29
SlowFast [8]	✓	65.92	50.42	39.04	90.04	74.78	59.50
Video Swin [19]		70.18	60.51	48.05	91.69	82.75	77.62
Video Swin [19]	✓	70.15	60.77	48.08	91.97	82.45	77.46

Table 3: The quantitative results of fusing segmentations to RGB-based networks. The input modality is RGB frames.

Visualisations. We visualise the heatmaps of "take quorn" and "wash knife" generated by Grad-CAM [25] for SlowFast [8] and Video Swin [19], as shown in Fig. 8. The integration of segmentation not only enables SlowFast to concentrate more precisely on the spatial locations where interactions occur but also allows it to focus more effectively on interactive objects over time. Notably, although the heatmaps of the vanilla Video Swin seem reasonable, approximated segmentations tend to correct the attention of Video Swin to the hands and corresponding objects. This could explain the slight improvement of the last two rows in Tab. 3.

5 Conclusion

In this work, we introduce a novel modality, termed "active object segmentation", to enhance egocentric action recognition. By leveraging fine-grained segmentation annotations from EPIC-KITCHENS VISOR [5], we evaluate its performance using three methods: (a) an indirect method with SPADE to generate RGB-like representations from segmentations, which improves action accuracy by 3%; (b) a direct method with the proposed segmentation-aware stream in Temporal Segment Network (TSN) to model active object segmentation, which increases noun accuracy by 10% and action recognition by 7.6%; and (c) using SPADE to perceive regions of active objects and hands from RGB frames, enhancing noun and verb recognition without ground-truth segmentations. Whether or not ground-truth active object segmentation is provided as input, our experiments demonstrate it enhances noun recognition and improves the ability to identify transformed objects, thereby increasing the robustness and versatility of video understanding systems.

Acknowledgments. This work was supported by the China Postdoctoral Science Foundation under Grant Number 2024M752365.

References

[1] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[2] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. 2024. VideoLLM-online: Online Video Large Language Model for Streaming Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18407–18418.

[3] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. 2019. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems* 32 (2019).

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2022. Rescaling Egocentric Vision. *International Journal of Computer Vision (IJCV)* (2022).

[5] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Ely Locke Higgins, Sanja Fidler, David Fouhey, and Dima Damen. 2022. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[6] Alban Main De Boissiere and Rita Noumeir. 2020. Infrared and 3d skeleton feature fusion for rgb-d action recognition. *IEEE Access* 8 (2020), 168297–168308.

[7] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. 2020. PySlowFast. https://github.com/facebookresearch/slowfast.

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 6202–6211.

[9] Rohan Ghosh, Anupam Gupta, Andrei Nakagawa, Alcimar Soares, and Nitish Thakor. 2019. Spatiotemporal filtering for event-based action recognition. *arXiv preprint arXiv:1903.07067* (2019).

[10] Younggi Hong, Min Ju Kim, Isack Lee, and Seok Bong Yoo. 2023. Fluxformer: Flow-Guided Duplex Attention Transformer via Spatio-Temporal Clustering for Action Recognition. *IEEE Robotics and Automation Letters* (2023).

[11] Jian-Fang Hu, Wei-Shi Zheng, Jiahui Pan, Jianhuang Lai, and Jianguo Zhang. 2018. Deep bilinear learning for rgb-d action recognition. In *Proceedings of the European Conference on Computer Vision*. 335–351.

[12] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*. PMLR, 4651–4664.

[13] Zhuolin Jiang, Viktor Rozgic, and Sancar Adali. 2017. Learning spatiotemporal features for infrared action recognition with 3d convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 115–123.

[14] Yanhao Jing and Feng Wang. 2022. Tp-vit: A two-pathway vision transformer for video action recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2185–2189.

[15] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. 2019. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 5492–5501.

[16] Youngwook Kim and Taesup Moon. 2015. Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters* 13, 1 (2015), 8–12.

[17] Dawei Liang and Edison Thomaz. 2019. Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–18.

[18] Jun Liu, Amir Shahroudy, Dong Xu, Alex C Kot, and Gang Wang. 2017. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 12 (2017), 3007–3021.

[19] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2021. Video Swin Transformer. *arXiv preprint arXiv:2106.13230* (2021).

[20] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. 2020. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1049–1059.

[21] Andrew Owens and Alexei A Efros. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision*. 631–648.

[22] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic Image Synthesis with Spatially-Adaptive Normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[23] Gorjan Radevski, Marie-Francine Moens, and Tinne Tuytelaars. 2021. Revisiting spatio-temporal layouts for compositional action recognition. *The British Machine Vision Conference (BMVC)* (2021).

[24] Hossein Rahmani and Ajmal Mian. 2016. 3d action recognition from novel viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1506–1515.

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.

[26] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Proceedings of Advances in neural information processing systems*.

[27] Cheng Wang, Haojin Yang, and Christoph Meinel. 2016. Exploring multimodal video representation for action recognition. In *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1924–1931.

[28] Fei Wang, Yunpeng Song, Jimuyang Zhang, Jinsong Han, and Dong Huang. 2019. Temporal unet: Sample level human action recognition using wifi. *arXiv preprint arXiv:1904.11953* (2019).

[29] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision*. 20–36.

[30] Qinyi Wang, Yexin Zhang, Junsong Yuan, and Yilong Lu. 2019. Space-time event clouds for gesture recognition: From RGB cameras to event cameras. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE, 1826–1835.

[31] Xiaolong Wang, Ali Farhadi, and Abhinav Gupta. 2016. Actions˜ transformations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2658–2667.

[32] Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. 2020. Symbiotic attention with privileged information for egocentric action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12249–12256.

[33] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. 2021. Interactive Prototype Learning for Egocentric Action Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 8168–8177.

[34] Yancheng Wang, Yang Xiao, Fu Xiong, Wenxiang Jiang, Zhiguo Cao, Joey Tianyi Zhou, and Junsong Yuan. 2020. 3dv: 3d dynamic voxel for action recognition in depth video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 511–520.

[35] Yunhua Zhang, Hazel Doughty, Ling Shao, and Cees GM Snoek. 2022. Audio-adaptive activity recognition across video domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 13791–13800.

[36] Rui Zhao, Haider Ali, and Patrick Van der Smagt. 2017. Two-stream RNN/CNN for action recognition in 3D videos. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*. IEEE, 4260–4267.