Symbiotic Attention for Egocentric Action Recognition with Object-centric Alignment

Xiaohan Wang, Linchao Zhu, Yu Wu, Yi Yang

Abstract—In this paper, we propose to tackle egocentric action recognition by suppressing background distractors and enhancing action-relevant interactions. The existing approaches usually utilize two independent branches to recognize egocentric actions, i.e., a verb branch and a noun branch. However, the mechanism to suppress distracting objects and exploit local human-object correlations is missing. To this end, we introduce two extra sources of information, i.e., the candidate objects' spatial location and their discriminative features, to enable concentration on the occurring interactions. We design a **S**ymbiotic **A**ttention with **O**bject-centric feature **A**lignment framework (SAOA) to provide meticulous reasoning between the actor and the environment. First, we introduce an object-centric feature alignment method to inject the local object features to the verb branch and noun branch. Second, we propose a symbiotic attention mechanism to encourage the mutual interaction between the two branches and select the most action-relevant candidates for classification. The framework benefits from the communication among the verb branch, the noun branch, and the local object information. Experiments based on different backbones and modalities demonstrate the effectiveness of our method. Notably, our framework achieves the state-of-the-art on the largest egocentric video dataset.

Index Terms—Egocentric Video Analysis, Action Recognition, Deep Learning, Symbiotic Attention

1 Introduction

Video recognition is an important task in the computer vision community. With the emerging of deep convolutional neural networks [1], [2], [3], [4] and large-scale video datasets [5], [6], [7], the action recognition performance has been prominently boosted [8], [9], [10], [11], [12]. However, most existing methods focus on recognizing videos captured from a third-person viewpoint. The progress in the first-person video has been relatively slow. Recently, egocentric action recognition has attracted increasing attention with the widespread applications of wearable cameras.

Compared to third-person videos, egocentric videos contain more complex scenes. Egocentric action recognition requires to distinguish the object that human is interacting with from various small distracting objects [13], [14]. Action recognition in egocentric videos provides a uniquely naturalistic insight into how a person or an agent interacts with the world. To enable the recognition of more complex videos, a challenging large-scale first-person dataset, *i.e.*, EPIC-Kitchens [14], was recently introduced for egocentric daily human activities understanding. This dataset provides rich interactions, covering adequate objects and natural actions. The intense camera motion, occlusion, and first-person viewpoint make it even more challenging to recognize fine actions.

In EPIC-Kitchens, the actions are defined by the combination of verb and noun, *e.g.*, "open door" and "cut potato". Due to the large action vocabulary, the verb and the noun classifiers are usually trained separately [14], [15]. The verb branch focuses on classifying actions (verbs) that the actor is performing, *e.g.*, "cut" and "open". The main obstacles for verb classification are large camera motion and subtle occurring action locations. The noun branch is to identify the object that the actor is interacting with. As shown in Figure 1, distracting objects in oblique view decrease the prediction score of the interacting object.

Damen *et al.* [16] evaluated several video models on EPIC-Kitchens that were not specially designed for the egocentric action recognition such as TSN [11], TSM [17] and TRN [18]. These models failed to achieve high classification score due to the absence of location-aware guidance for the complex scenes in first-view videos.

Recently, Wu *et al.* [15] leveraged object detection features to introduce longer context information for the noun classification network in egocentric action recognition. The long-term feature bank is aggregated via a simple max pooling or average pooling operation, while the more sophisticated non-local operator is found to be not that effective. The verb branch and the noun branch are optimized independently. They only consider the interaction between the noun branch and the object features but fail to enable the communication between the verb branch and the noun branch. Baradel *et al.* [19] designed an object relation network for high-level object reasoning, where the relation modeling branch facilitates object masks to generate local object features. The object reasoning only performs on the object branch. It lacks interactions with the activity branch. These works ignore the mutual communication between the standalone verb and noun branches. They only focus on contextual modeling and relation reasoning on a *single* branch. However, an action is determined by both the interacting object and the motion that the actor is performing. It could

• *Xiaohan Wang and Yu Wu are with Baidu Research, Beijing, China and the ReLER Lab, Australian Artificial Intelligence Institute, University of Technology Sydney, NSW, Australia.*
Linchao Zhu and Yi Yang are with the ReLER Lab, Australian Artificial Intelligence Institute, University of Technology Sydney, NSW, Australia.
Email: xiaohan.wang-3@student.uts.edu.au; linchao.zhu@uts.edu.au; yu.wu-3@student.uts.edu.au; yi.yang@uts.edu.au

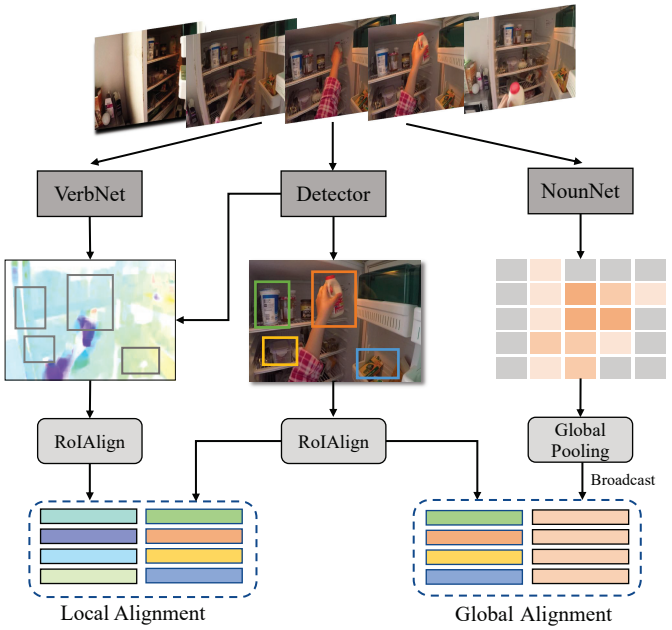


Fig. 1. The illustration of the object-centric feature alignment. For verb classification, the spatial location provided by the detector can possible reduce the object-irrelevant motions. Local motion features aligned with object features serve as possibly action candidates. For noun classification, global alignment inject the local object features into the context-aware global feature. These location-aware feature candidates from the two branches are beneficial to the subsequent meticulous reasoning.

be difficult even for a human to recognize an action by only looking at the objects while ignoring the actor's intention, or only understanding motion changes without the awareness of the interacting object.

To better exploit the local object guidance and the mutual benefits of the interactions between different branches, we make the following contributions.

We first propose an object-centric feature alignment method to dynamically integrate location-aware information to the verb and the noun branches. Our object-centric feature alignment encourages the meticulous reasoning between the actor and the environment. The object-centric features are extracted by an object detection model, providing finer local information that is beneficial to the attendance of an on-going action. The noun branch and the verb branch integrate location-aware information by two different approaches (Fig. 1). We introduce **global alignment for noun classification.** The noun features and the detection features are complementary to each other, and proper integration of these two features produces more accurate identification of the interacted object. In this global alignment, we concatenate each detection feature with the global noun feature. The generated features incorporate both local relevant features and global contextual features, which restrain the features of irrelevant objects. We introduce **local alignment for verb classification.** The verb feature contains motion information, which is quite different from the appearance information in noun feature and object features. The semantic gap between verb features and detection features is larger than the gap between noun features and detection features. It may not be straightforward to integrate global verb features with detection features directly. When we use the aforementioned global alignment for verb classification, it may generate indistinct features due to the accompanying background motion noises. We propose to integrate spatially-aligned verb features with object features. In this way, the most relevant verb features will be generated for better alignment with local object features. It eases the difficulties of the integration between verb features and local object features. We extract regional verb features from the verb branch by pooling from the spatial feature map with the given candidate spatial location. The regional motion feature is then combined with the corresponding detection feature.

After the object-centric alignment, we obtain a set of candidate verb features and noun features. A symbiotic attention mechanism is then introduced to enable mutual interactions between the two branches and select the most action-relevant features. It consists of two parts, *i.e.*, cross-stream gating mechanism and action-attended relation module. The fused object-centric features contain useful local details. However, due to the existence of inaccurate detection regions, there are quite a few disturbing background noises in the features. To this end, we propose a cross-stream gating mechanism to normalize the aligned features. This normalization process suppresses the action-irrelevant noises and enables mutual communication between the verb branch and the noun branch. To further uncover the relationships among the object-centric features and identify the most action-relevant information, we develop an action-attended relation module to examine each potential motion-object pair and then generate the final representation for classification. The proposed Symbiotic Attention with Object-centric Alignment (SAOA) method dynamically integrates three sources of information towards better action recognition.

We evaluate our framework with different backbones and modalities on the largest egocentric video dataset, *i.e.*, EPIC-Kitchens. We conducted experiments on two backbones (*i.e.*, I3D [5] and ResNet-50 [3]), two modalities (*i.e.*, RGB and optical flow). It can consistently improve the performance over the baselines by a large margin with different backbone and input modalities. The effectiveness of our framework is validated both quantitatively and qualitatively. Notably, our method outperforms the state-of-the-art method [20] by **6.7%** on the unseen test set and **2.9%** on the seen test set of Epic-Kitchens. The ensemble of the proposed method achieved first place in EPIC-Kitchens Action Recognition Challenge 2020.

This paper is an extension of [21]. In our previous work, global alignment is developed to integrate the object features for both the verb branch and the noun branch. And only the model with RGB frames as input and ResNet-3D [22] backbone were studied. We extend [21] by proposing a local alignment for verb classification to reduce the object-irrelevant motions and alleviate the large semantic gap between object features and motion features. Moreover, extensive experiments on different backbone and input modalities are conducted. That demonstrates our method is general, and the two-stream SAOA can significantly improve the recognition performance over the previous model SAP. In summary, our main contributions are as follows:

First, we develop an object-centric alignment method to inject local details into the verb branch and noun branch.

The alignment allows the model to take advantage of the location-aware object information and prevent it from confusing with background noise.

Second, we propose a novel symbiotic attention mechanism to enable the mutual interaction between the verb branch and the noun branch. It provides the meticulous reasoning between the actor and the environment. The experiment shows that the symbiotic attention is beneficial to distinguish the action-relevant motion and object.

Third, extensive experiments demonstrate the effectiveness and superiority of the proposed SAOA. Our results outperform the state-of-the-art by a large margin on the largest egocentric video dataset.

2 Related Work

2.1 Deep Video Recognition

Deep learning methods have achieved promising performance on the video classification task. 3D convolution kernel was introduced in [9] to model the spatio-temporal relation in videos. I3D [5] proposed to initialize 3D CNN with the inflated weights of 2D CNN pre-trained on ImageNet [23]. Hara *et al.* [22] evaluated various 3D CNN architectures on Kientics [5] and demonstrated the effectiveness of 3D models. More recently, P3D [24], S3D [12] and R(2+1)D [25] proposed to decompose the 3D kernel to a spatial 2D convolution and temporal 1D convolution. The decomposition of 3D convolution makes the network easier to be optimized and can boost video recognition performance. In this work, we utilize two typical 3D CNNs, *i.e.*, ResNet-3D [22], and I3D [5], as the backbones of VerbNet and NounNet to extract global motion and appearance features.

There are also many methods utilizing 2D CNN to tackle the video recognition task. Simonyan *et al.* [8] proposed a two-stream 2D CNN with both RGB frames and optical flow as input. TSN [11], [26] proposed to process the frames sampled from multiple temporal segments and aggregated all the predictions for the entire video classification. TRN [18] developed a temporal relation module to further enhance the temporal modeling in videos. TSM [17] proposed a temporal shift module to capture the motion information and temporal relations.

Moreover, Recurrent Neural Networks (RNNs) are effective architectures for temporal modeling and have been found useful for video classification in [6], [27], [28], [29]. Donahue *et al.* [28] utilized LSTM [30] to aggregate the frame features extracted by 2D CNN. Zhu *et al.* [27] proposed a multi-rate bi-directional GRU to deal with motion speed variance in videos.

These deep models are designed for third-person video recognition. They are able to capture motion and scene information but are not sufficient to locate various small objects in egocentric videos accurately.

2.2 First-Person Action Recognition

Compared to third-person video recognition, egocentric action recognition is more dependent on the modeling of the interactions between the actor and the environment. It is important to locate the interacted object and distinguish the motion of hands.

There are a number of previous works [15], [19], [31] proposed to extract object features in the videos to provide a better understanding of local details. Fathi *et al.* [31] proposed to learn a hierarchical model that exploits the consistent appearance of objects, hands, and actions and refines the object prediction based on action context. Ma *et al.* [32] located the object of interest by a hand segmentation net and fed the cropped regions and optical flow images to two CNNs for object and action classification, respectively. Baradel *et al.* [19] proposed to perform object-level visual reasoning about spatio-temporal interactions in videos through the integration of object detection networks. More recently, Wu *et al.* [15] combined Long-Term Feature Banks that contains detection features with 3D CNN to improve the accuracy of object recognition. These methods take the advantage of the local object information for egocentric action recognition. However, compared to our method, they overlook the interactions between the motion information and the object.

The attention mechanism is efficient in locating the region of interest on the feature map. Sudhakaran *et al.* [33] proposed a Long Short-Term Attention model to focus on features from relevant spatial parts. They extended LSTM with a recurrent attention component and an output pooling component to track the discriminative area smoothly across the video sequence. Li *et al.* [34] proposed to generate an attention map of the hand-object interaction by the guidance of the gaze information. These methods apply the attention mechanism on the feature maps produced by 2D/3D CNN. In this work, we design an attention mechanism for the object-centric features which fuses different information.

In addition, recently, Kazakos *et al.* [20] developed an egocentric action recognition model with multi-modal temporal-binding. Besides, some researchers [35], [36] focus on the egocentric action prediction task, which predicts the near future egocentric action before it occurs. Differently, in this paper, we focus on egocentric action recognition, which is to recognize an action given full observations.

2.3 Human-Object Interaction

Reasoning the interaction between human and objects is relevant to our task because it also requires to find out the interacting object. Most methods in this field are based on detection models. For example, Gkioxari *et al.* [37] predicted a density map to locate the interacted object and calculated the action score, with a modified Faster RCNN architecture. Qi *et al.* [38] proposed Graph Parsing Neural Networks that incorporates structural knowledge and deep object detection model. Fang *et al.* [39] developed a pairwise body-part attention model that can learn to focus on crucial parts for human-object interaction (HOI) recognition. Besides, some works use human-object interactions to help recognize actions. Wang *et al.* [40] proposed to represent videos as space-time region graphs, which models shape dynamics and relationships between actors and objects. Sun *et al.* [41] developed an Actor-Centric Relation Network for spatio-temporal action localization.

Most of these HOI techniques rely on the appearance of the actors, which is absent in egocentric videos. Instead of the use of the detection features of humans, we pay attention to the interactions between the motion and the objects.

2.4 Visual Attention

Attention mechanism can highlight visual regions or linguistic words that are important to the task predictions. It has been widely used in both computer vision [42], [43], [44], [45], [46], [47] and natural language processing [48], [49], [50]. Non-local networks [44] leveraged the non-local attention operation in spatio-temporal dimension for video recognition. Squeeze-and-excitation network (SENet) [43] developed the squeeze-and-excitation block, which introduced the channel-wise attention inside the residual block. Recently, Linsley *et al.* [45] improved the SE module by the global-and-local attention (GALA), which combined global contextual guidance with local saliency. In addition, they also introduced a large-scale dataset containing human-derived attention maps, which can be used to supervise the attention mechanism to be more accurate and interpretable. These methods are designed with a self-attention operation. The feature map used to generate attention weights is also the target feature where the weights are applied. Differently, our work concentrates on the interactions in egocentric videos. We apply a cross attention mechanism between the motion feature and the object appearance feature.

3 Proposed Method

3.1 Overview

In this section, we illustrate our network architecture for egocentric video recognition. We develop three backbone networks to extract features from the input video: (1) VerbNet is a 3D CNN and takes a video clip as input. It is designed to capture the motion information. (2) NounNet shares the same architecture with VerbNet. It is trained to produce a feature that represents object appearance. (3) Object detection model takes sampled individual frames as input. We use Faster R-CNN [51] as our detector to generate object features and location proposals. The output features and location proposals of the three base models are fed to the subsequent SAOA module. We aim to enable effective communication among VerbNet, NounNet, and object features. The SAOA module generates two feature vectors, which can be used to predict verb class and noun class. The overall framework is illustrated in Fig. 2.

3.2 Preliminaries

For each input egocentric video $X = \{x^1, ..., x^t\}$ with t frames, its verb and noun label is y^v and y^n, respectively. The action $y = (y^v, y^n)$ is a combination of the verb and noun. We use two individual 3D CNNs as the backbones in our framework, with one for the verb feature extraction and the other for the noun feature extraction. The extracted verb feature $f^v \in \mathbb{R}^{T\times H\times W\times C}$ contains the motion information, where T is the temporal size, H is the height, W is the width, and C is the channel size of the extracted feature, respectively. The noun feature $f^n \in \mathbb{R}^{T\times H\times W\times C}$ contains the global appearance information.

To enhance the global representation through the communication between two branches and enable meticulous reasoning, we use a pre-trained detection model to provide detailed locations of objects in the video. Considering the efficiency, for each video clip, we only use M sampled frames for detection inference. These frames are sampled around the center of the input clip for 3D CNNs within a fixed time duration. The duration is longer than the input clip to provide more context information. Given a feature map and a spatial location, $RoIAlign$ [52] first crops the feature map based on the location and then performs pooling operation to produce a fix-size feature map. In this work, we use max-pooling in $RoIAlign$ layer to produce a 1D feature vector. For object detection model, the output of the $RoIAlign$ layer is regarded as the feature for each detected object. To save memory usage and reduce the noisy information, we only keep top-K object proposals according to their confidence scores for each sampled frame. Thus, for each input clip of the 3D CNNs, we have a auxiliary object feature matrix $f^o \in \mathbb{R}^{N\times C_1}$, which contains $N = M \times K$ object features around the center of the short video clip. For each object detection feature $f_i^o, i \in [1 \dots N]$, we have an spatial detection location $l_i \in \mathbb{R}^4$. $l_i = (x_i^0, y_i^0, x_i^1, y_i^1)$ representing a rectangular in 2D space. The object feature matrices f^o are fused with the verb feature f^v and noun feature f^n by the following object-centric alignment method. After that, the verb branch and noun branch are interacted with each other to produce more discriminative features for action recognition with a symbiotic attention mechanism.

3.3 Object-centric Feature Alignment

The verb branch and noun branch produce two feature maps f^v and f^n by passing a video clip to each backbone. Due to the intensive camera motion and various distracting objects in egocentric videos, the useful interaction information in these features is hard to distinguish from the global feature map without any other guidance. To this end, we develop an object-centric feature alignment method to generate potential motion and object candidates, which disentangle the local information from the global feature maps. Specifically, we leverage object feature matrix f^o and corresponding locations as location-aware information to inject the local details into the global features. Considering the different semantic properties of the verb branch and noun branch, we introduce two mechanisms to integrate f^o with f^v and f^n:

Global alignment for the noun branch. The object features and the noun feature both represent the appearance of the objects in the videos. Considering the small semantic gap and the complementarity between the local object features and the global noun feature, we introduce a direct global alignment for the noun branch. Note that we have $f^o \in \mathbb{R}^{N\times C_1}$ and $f^n \in \mathbb{R}^{T\times H\times W\times C}$. We first leverage a global average pooling (GAP) on f^n, and the generated global feature vector f_g^n is of shape $1 \times C$. Each detection feature in f^o is then concatenated with the global feature vector, followed by a nonlinear activation. Formally, the global alignment operation can be presented as follow:

$$f_i^{\hat{n}} = \texttt{ReLU}(W^n f_g^{n\mathrm{T}} + W_o^n f_i^{o\mathrm{T}} + b^n), i \in [1 \dots N], \quad (1)$$

where $W^n \in \mathbb{R}^{C\times C}, W_o^n \in \mathbb{R}^{C\times C_1}, b_n \in \mathbb{R}^{1\times C}$, f_i^o denotes i-th detection feature in f^o and $f_i^{\hat{n}}$ is the aligned noun feature. We obtain the final noun feature $f^{\hat{n}} \in \mathbb{R}^{N\times C}$ by concatenating all $f_i^{\hat{n}}$ where $i \in [1 \dots N]$. Each row in $f^{\hat{n}}$ represent a object-centric feature, which integrates the global noun appearance with an explicit local object information.

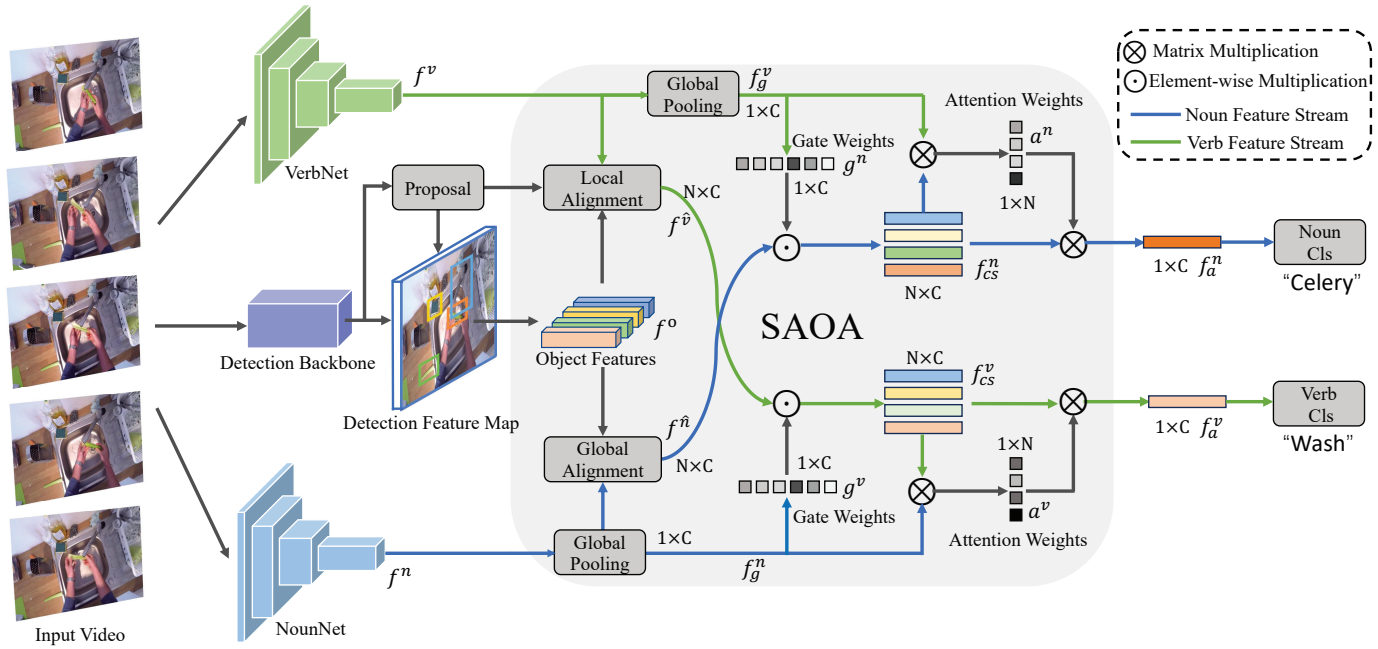


Fig. 2. The proposed SAOA framework. Our framework consists of three feature extractors and one interaction module. The detection model generates a set of local object features and location proposals. This location-aware information is injected to the two branches by an object-centric alignment method For the Verb branch, the feature map is locally aligned with the objects by combining the local motion features with corresponding object detection features. For the Noun branch, the object features are aligned with the global noun representation. Subsequently, the fused features from each branch interact with the global feature from the other branch by a symbiotic attention mechanism. The two object-centric feature matrices are first normalized by a cross-stream gating operation. After that, the matrices are attended by the other branch to select the most action-relevant information. The outputs of SAOA are used to classify the verb and noun, respectively.

Local alignment for the verb branch. Different from global alignment for the noun branch, we leverage a local alignment that integrates the verb feature map and the object detection features based on their spatial locations. The verb feature represents the motion information in the videos, which is quite different from the object features. Global alignment might not well integrate the two features due to the large semantic gap. The proposed local alignment can decompose the global motion information to object-centric local details. Note that we have $f^o \in \mathbb{R}^{N\times C_1}$ and $f^v \in \mathbb{R}^{T\times H\times W\times C}$. For each object detection feature $f_i^o, i \in [1 \dots N]$, we have a spatial detection location $l_i \in \mathbb{R}^4$. $l_i = (x_i^0, y_i^0, x_i^1, y_i^1)$ representing a rectangular in 2D space. We extract the locally aligned verb feature from f_i^v by the ROIAlign operation, *i.e.*, $f_i^v = ROIAlign(f^v, l_i)$. The final verb feature can be obtained via:

$$f_i^{\hat{v}} = \texttt{ReLU}(W^v f_i^{v\mathrm{T}} + W_o^v f_i^{o\mathrm{T}} + b^v), i \in [1 \dots N], \quad (2)$$

where $W^v \in \mathbb{R}^{C\times C}, W_o^v \in \mathbb{R}^{C\times C_1}, b_v \in \mathbb{R}^{1\times C}$ and $f_i^{\hat{v}}$ is the aligned verb feature. The final verb feature $f^{\hat{v}} \in \mathbb{R}^{N\times C}$ is obtained by concatenating all $f_i^{\hat{v}}$ where $i \in [1 \dots N]$. The final motion-object paired feature incorporates local detection features and location-aware motion features.

3.4 Symbiotic Attention

The object-centric alignment integrates the object features to the verb branch and noun branch. The fused object-centric feature matrices contain useful local details and provide potential action-relevant candidates for verb and noun classification. We propose a symbiotic attention mechanism to encourage mutual communication between the two branches. It further generates a better representation for classification. As illustrated in Fig. 2, symbiotic attention includes two stages. First, the fused object-centric features are re-calibrated by the other branch utilizing a cross-stream gating mechanism. After that, the normalized feature matrix is attended by the other branch to aggregate the most action-relevant information within an action-attended relation module.

3.4.1 Cross-Stream Gating

Due to the existence of inaccurate detection regions, there are quite a few disturbing background noises in the features. Besides, it is important to introduce information from one branch to guide discrimination in the other branch. For example, given a video clip that presents the action "cut potato" but also contains the object "bowl", the motion information of "cut" can provide extra guidance for more accurate recognition that the interacted object is "potato" rather than "bowl". To this end, we develop a cross-stream gating operation to underline the action-relevant information from the verb branch and the noun branch.

In noun classification, we generate a gating weight to normalize the input noun feature matrix $f^{\hat{n}}$. The gating weight g^n is obtained from the global verb feature:

$$f_g^v = GAP(f^v), \quad (3)$$

$$g^n = \texttt{Sigmoid}(W_g^n f_g^{v\mathrm{T}} + b_g), \quad (4)$$

$$f_{cs}^n = g^n \odot f^{\hat{n}}, \quad (5)$$

where $W_g^n \in \mathbb{R}^{C\times C}$, $f_g^v \in \mathbb{R}^{1\times C}$, $b_g \in \mathbb{R}^{1\times C}$, $g^n \in \mathbb{R}^{1\times C}$, $f_{cs}^n \in \mathbb{R}^{N\times C}$, and $\odot$ denotes the element-wise multiplication. g^n is the scaling vector to rescale the noun feature

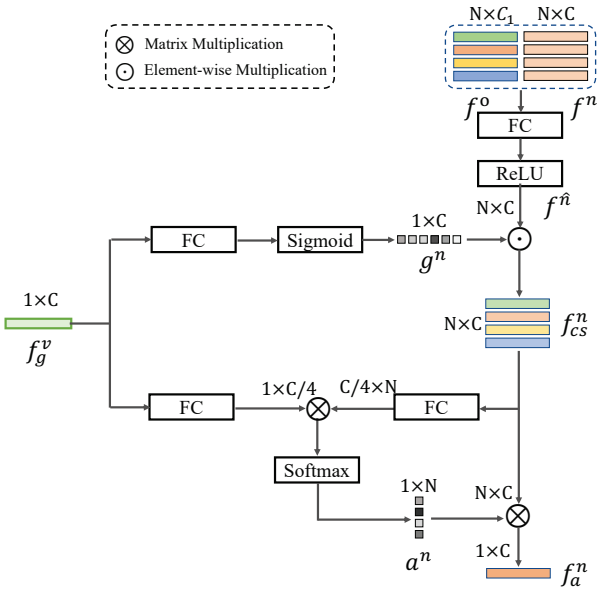


Fig. 3. The illustration of symbiotic attention on the noun branch. The object-centric noun feature matrix is first normalized by the global verb feature. After that, the feature matrix interacts with the global verb feature to generate attention weights. The final noun representation is the weighted sum of the normalized object-centric features.

matrix. After re-calibrating the object-centric noun feature by the verb feature, the distracting noises can be suppressed while the action-relevant channels can be enhanced. Similarly, the gated verb feature f_{cs}^{v} can be obtained by:

$$f_g^n = GAP(f^n), \tag{6}$$

$$g^v = \texttt{Sigmoid}(W_g^v f_g^{n\mathrm{T}} + b_g^v), \tag{7}$$

$$f_{cs}^v = g^v \odot f^{\hat{v}}, \tag{8}$$

where $W_g^v \in \mathbb{R}^{C\times C}$, $f_g^n \in \mathbb{R}^{1\times C}$, $b_g^v \in \mathbb{R}^{1\times C}$, $g^v \in \mathbb{R}^{1\times C}$, $f_{cs}^v \in \mathbb{R}^{N\times C}$. Our cross-stream gating mechanism enables mutual communication between the verb branch and the noun branch, and it adaptively exploits the correlations of verbs and nouns. We illustrate on the noun branch in Fig. 3.

3.4.2 Action-attended Relation Module

The calibrated object-centric feature matrix contains the action-relevant information and implicit guidance about the spatio-temporal location of an on-going action. To uncover the relationships among the object-centric features and identify the most action-relevant information, more meticulous reasoning is required. Therefore, we develop an action-attended relation module to examine each potential motion-object pair and then generate the final representation for classification.

Specifically, we first propose to assess the relevance between the global feature and location-aware object-centric features. Taking the noun branch for example. The global verb feature and the object-centric noun features are projected to the same dimension space. The distance between each noun feature and the verb feature is calculated to represent their relevance score. After that, we sum the object-centric features weighted by the relevance coefficients. Formally, we perform attention mechanism on the normalized object-centric noun features $f_{cs}^n \in \mathbb{R}^{N\times C}$ and the global verb feature $f_g^v \in \mathbb{R}^{1\times C}$,

$$a^n = \texttt{Softmax}(f_g^v W_v^a W_{cn}^a f_{cs}^{n\mathrm{T}}), \tag{9}$$

where $W_v^a \in \mathbb{R}^{C\times\frac{C}{4}}$, $W_{cn}^a \in \mathbb{R}^{\frac{C}{4}\times C}$ are projection matrices. We project the features to a low feature dimension $\frac{C}{4}$ to reduce the computational cost of matrix multiplication. We found $\frac{C}{4}$ performs well in our experiments. $a^n \in \mathbb{R}^{1\times N}$ is the generated attention weights. The final noun representation f_a^n is produced by the weighted sum of the object-centric features,

$$f_a^n = a^n f_{cs}^n. \tag{10}$$

Similarly, we select relevant action features from f_{cs}^v with query f_g^n,

$$f_a^v = \texttt{Softmax}(f_g^n W_n^a W_{cv}^a f_{cs}^{v\mathrm{T}}) f_{cs}^v. \tag{11}$$

The final noun feature $f_a^n \in \mathbb{R}^{1\times C}$ and the final verb feature $f_a^v \in \mathbb{R}^{1\times C}$. Through the interaction of global feature and object-centric features, our model selects the most action-relevant feature for classification.

3.5 Training and Objectives

We use Faster R-CNN with the ResNeXt-101-FPN backbone as our object detector. Following the training procedure in [15], we first pre-train the detector on Visual Genome [53] and then finetune it on EPIC-Kitchens object detection set. For VerbNet and NounNet, we adopt 3D Resnet-50 [22] and I3D [5] as our backbones. The two networks are both initialized with Kinetics pre-trained weights. In the first stage, we individually train the VerbNet and NounNet with the corresponding CrossEntropy Loss, *i.e.*, $\mathcal{L}^v$ and $\mathcal{L}^n$.

$$\mathcal{L}^n = CrossEntropy(f_a^n, y^n), \tag{12}$$

$$\mathcal{L}^v = CrossEntropy(f_a^v, y^v). \tag{13}$$

After the base training stage, we freeze the weights of the backbone and cascade our SAOA module. The objective for the second stage is the same as the base training stage, and only the weights of SAOA are optimized.

4 Experiments

4.1 Datasets

EPIC-Kitchens is the largest dataset in first-person vision so far. It consists of 55 hours of recordings capturing all daily activities in the kitchens. The performed activities are non-scripted, which makes the dataset very challenging and close to real-world data. The dataset contains 39,594 action segments which are annotated with 125 verb classes (*e.g.*,"cut","take") and 321 noun classes (*e.g.*,"potato","knife"). The action of each video segment is defined by the verb-noun pair (*e.g.*, "cut potato", "take knife"). We split the original training set to new training and validation set following [19]. All hyper-parameters are selected based on the performance on the validation set. We report the top-1 and top-5 accuracy of the verb, noun, and action.

4.2 Experiment Settings

We implement and test our method using PaddlePaddle and PyTorch. We train our framework in a two-stage optimization scheme. Specifically, we firstly pre-train the base models (VerbNet, NounNet, and the detector) individually. After that, we optimize the subsequent SAOA module using extracted features from the base models. Next, we illustrate the details on how to pre-train the backbones (Backbone details) and how to extract local object information (Detector details). Finally, we show the training details of the module (SAOA details).

Backbone details. We adopt two typical 3D CNNs as our backbones, *i.e.*, ResNet50-3D [22] and I3D [5]. ResNet50-3D is built with residual blocks and I3D is based on Inception architecture. We take the Kinetics [5] pre-trained weights as the initialization of our backbone model. We then train the backbone models (VerbNet and NounNet) individually on the target dataset using 64-frame input clips. The targets for the VerbNet and NounNet are the verb label and noun label, respectively. The videos are decoded at 60 FPS for the EPIC-Kitchens dataset. We adopt the stochastic gradient descent (SGD) with momentum 0.9 and weight decay 0.0001 to optimize the parameters for 35 epochs. The overall learning rate is initialized to 0.003, and then it is changed to 0.0003 in the last 5 epochs. The batch size is 32. During the first training stage, the input frame size is 224×224, and the input frame is randomly cropped from a random scaled video whose side is randomly sampled in [224, 288]. We sample 64 successive frames with stride=2 from each segment to constitute the input clip. The center index of the input clip is randomly chosen in the segment during training. For the testing, we sample a center clip per segment. We resize the clip to the size of 256×256 and use a single center crop of 224×224.

Detector details. Following [15], we use the same Faster R-CNN to detect objects and extract object features. The detector is first pre-trained on Visual Genome [53] and then fine-tuned on the training split of the EPIC-Kitchens dataset. We use a batch size of 12 and train the model on EPIC-Kitchens for 180k iterations for the trainval/test split. We use an initial learning rate of 0.005, which is decreased by a factor of 10 at iteration 140k and 160k. For the train/val split, we train the model for 150k iterations, and the learning rate decays at iteration 116k and 133k. Finally, our object features are extracted using *RoIAlign* from the detector's feature maps. For each video clip, we perform object detection on a set of frames that are sampled around the clip center within a fixed time duration. The time duration is set to 6 seconds for global alignment and 4 seconds for local alignment. The sample rate is at two frames per second. For each frame, we keep the top five features and proposals according to the confidence scores. Therefore, given a video clip, we obtain 60 detection features during global alignment. In local alignment, we obtain 40 detection features and corresponding locations.

SAOA details. We leverage the pre-trained backbone models and the detection models as the feature extractors. During the second-stage training, only the weights of SAOA are updated. We use SGD with momentum 0.9 and weight decay 0.0001 to optimize the parameters with batch-size of 32. For the model equipped with the I3D backbone, we train the model for 15 epochs. The learning rate is initialized to 0.001 and then reduced to 0.0001 in the last 5 epochs. For the models based on R-50, we train the model for 15 epochs, and the learning rate is set to a constant value 0.0001. Notably, since the detection features have different scales from the I3D features, the features from the I3D backbone need to be normalized before concatenation with detection features in the alignment modules. However, the feature from the R-50 backbone can be directly fed to the SAOA module without normalization. The main reason is the different network types between the detection backbones (based on residual block) and the I3D model (based on Inception block). Specifically, the features produced by the I3D backbone and detection model are l_2-normalized before concatenation. The combined feature is then multiplied by the l_2-norm of the I3D feature to scale the amplitude. A similar normalization strategy is introduced in [41]. During the training and testing of SAOA, we utilize the same temporal sampling strategy during the training and testing of the backbone. For each input video clip, we resize it to the size of 256. Then we feed the 64-frame clip to the network without spatial cropping.

Action calculation The actions are determined by the pairs of verb and noun. The basic method of obtaining the action score is to calculate the multiplication of verb probability and noun probability. However, there are thousands of combinations and most verb-noun pairs that do not exist in reality, *e.g.*, "open the knife". In fact, there are only 149 action classes that have more than 50 samples in the EPIC-Kitchens dataset [14]. Following the approach in [15], we re-weight the final action probability by a prior, *i.e.*

$$P(action = y) = \mu(y^v, y^n)P(verb = y^v)P(noun = y^n), \tag{14}$$

where μ is the occurrence frequency of action in training set.

TABLE 1
The effectiveness of Symbiotic Attention (SA) for **verb prediction** and **noun prediction** on the EPIC-Kitchens validation set. "ARM" denotes the Action-attended Relation Module. "CSG" denotes the Cross-Stream Gating.

Methods	Verb Top-1	Noun Top-1
Baseline	54.6	23.8
SA w/o CSG	57.0	32.6
SA w/o Gating	57.2	33.6
SA w/o Cross-Stream	57.4	33.2
SA w/o ARM	56.6	32.7
SA	**57.7**	**34.8**

4.3 The Effectiveness of SAOA

In this section, we focus on investigating the effectiveness of the proposed SAOA model. We conduct extensive ablation studies to evaluate the contributions of each component and the benefits of different input modalities.

4.3.1 The effectiveness of the symbiotic attention

Ablation studies of SA. The symbiotic attention (SA) consists of two modules, *i.e.*, Cross-Stream Gating (CSG), and Action-attended Relation Module (ARM). We evaluate each component on the Epic-Kitchens validation set for both verb

and noun classification. We use R-50 as the backbone and RGB data as the input. The results are shown in Table 1.

"*Baseline (noun)*" uses a single branch backbone for noun classification. The cross-stream gating module enables mutual communication between the verb branch and the noun branch and re-calibrates the fused features. We implement "*SA w/o CSG*" by performing ARM with the single stream. Specifically, we utilize the global noun feature to attend the object-centric matrix produced by the global alignment module. "*SA w/o CSG*" obtained 32.6% top-1 accuracy, which is 2.2% worse than the unified symbiotic attention. The performance comparison between symbiotic attention and "*SA w/o CSG*" validates the effectiveness of the CSG module. Furthermore, we decompose CSG into two parts, *i.e.*, Cross-Stream and Gating. We aim to investigate the impact of each component. "*SA w/o Cross-Stream*" indicates using the same stream to gate and attend the noun features. "*SA w/o Gating*" indicates utilizing the feature from the verb stream to attend the object-centric matrix. Specifically, without the cross-stream operation, the performance drops from 34.8% to 33.2%, which confirms the importance of the interaction between the two branches. Without the gating operation, the performance drops from 34.8% to 33.6%, which shows the benefits of our gating mechanism in feature normalization.

We now study the effectiveness of the ARM module. ARM can select the most action-relevant information from the object-centric features and explore the relationships in the spatio-temporal context. The performance drops from 34.8% to 32.7% when ARM is not used, which demonstrates the effectiveness of ARM in action-relevant information selection.

For verb classification, the unified SA outperforms the baseline model by 3.1%. Without the Cross-Stream Gating (CSG), the performance drops by 0.7%. This demonstrates the effectiveness of CSG for verb classification. Specifically, without the gating operation, the performance drops from 57.7% to 57.2%. The performance drops by 0.3% without the cross-stream operation. Moreover, when ARM is not used, the performance drops from 57.7% to 56.6%, which shows the benefits of the action-attended reasoning for verb classification.

SA outperforms other aggregation operations. We first study the effectiveness of our symbiotic attention only using the object detection feature. We directly apply average pooling and max pooling on the object detection features for noun classification. We denote the two pooling methods as "*Det Feat+Avg Pooling*" and "*Det Feat+Max Pooling*", respectively. The results are shown in Table 2. "*SA (Det Feat only)*" performs symbiotic attention on object features without the integration of the global noun feature. "*SA (Det Feat only)*" achieves 30.4% on top-1 accuracy, which outperforms the average pooling baseline and the max-pooling baseline by 5.9% and 4.8%, respectively. The result confirms the superiority of our attention mechanism.

"*Noun+Det Feat*" is one of the baselines to integrate noun features and object detection features, which utilizes the concatenated global noun feature and the max-pooled object feature for classification. "*Noun+Det Feat*" introduces the location-aware object information and uses a simple fusion method to incorporate the location-aware object information. Our symbiotic attention outperforms "*Noun+Det Feat*" by 3.6% on top-1 accuracy (34.8% v.s 31.2%), which demonstrates our symbiotic attention is more effective than the simple aggregation method.

TABLE 2
Comparisons between our symbiotic attention and other aggregation methods for **noun prediction** on the EPIC-Kitchens validation set. "Noun" denotes the global feature from NounNet. "Det Feat" is the location-aware object features.

Methods	Top-1 Accuracy
Det Feat+Avg Pooling	24.5
Det Feat+Max Pooling	25.6
SA (Det Feat only)	30.4
Noun + Det Feat	31.2
SA + Local Alignment	33.6
SA + Global Alignment	**34.8**

4.3.2 The effectiveness of the global alignment for noun classification

We first conduct the experiment of performing local alignment for the noun branch. The results are shown in the last two rows in Table 2. Compared to the model using global alignment for noun classification, the model with local alignment on the noun feature is 1.2% lower in top-1 accuracy on the EPIC-Kitchens validation set. The results show that the global alignment is more proper than local alignment for the noun classification. As the noun features and the local detection features are mutually complementary with less semantic gap, we use global alignment for the integration of the noun feature and the object detection feature.

4.3.3 The effectiveness of the local alignment for verb classification.

In our SAOA framework, the location-aware object information is globally aligned with the noun feature for noun classification. The location-aware information is locally aligned with the verb feature for verb classification. In this section, we quantitatively compare our SAOA with the previous work SAP [21] and demonstrate that the local alignment approach is better than the global alignment for *verb classification*. We use **RGB** data as inputs, and evaluate the performance on both R-50 and I3D. The top-1 results are shown in Table 3.

"*Baseline*" denotes the single branch verb classification model. In "*Verb+Noun Fusion*", we train a verb classifier with the concatenation of the global noun feature and the verb feature. *SAP* [21] utilizes the global alignment for verb classification. We observe that "*Verb+Noun Fusion*" slightly improves the "*Baseline*" classification model. It shows a simple fusion method does not help to improve verb classification. Compared to "*Baseline*", *SAP* obtain a 1.3% improvement on R-50 and a 1.1% improvement on I3D. This clearly shows that *SAP* well integrates all three sources of information. Compared to the combination of the *global* motion feature and object features in *SAP*, our SAOA leverage a *local alignment* method that can alleviate the semantic gap between detection features and motion features. Our SAOA R-50 model outperforms the SAP R-50 model by 1.8% on top-1 accuracy. Our SAOA I3D model

TABLE 3
Ablation study for **verb prediction** using **RGB** data as inputs. We evaluate the comparisons two backbones, *i.e.*, R-50 and I3D. The top-1 results are reported on the EPIC-Kitchens validation set. "Det Feat" denotes the object detection feature. "Det Box" denotes the location of the object detection proposal.

Methods	Verb	Noun	Det Feat	Det Box	R-50	I3D
Baseline (RGB)	✓	-	-	-	54.6	53.2
Verb+Noun Fusion (RGB)	✓	✓	-	-	54.7	53.7
SAP (RGB)	✓	✓	✓	-	55.9	54.3
SAOA (RGB)	✓	✓	✓	✓	**57.7**	**55.1**

TABLE 4
Two-stream SAOA for both verb classification and noun classification.

Methods	Verb Top-1	Noun Top-1
Our SAOA (RGB+Obj)	55.1	34.7
Our SAOA (Flow+Obj)	56.9	35.0
Our SAOA (RGB+Flow+Obj)	**60.4**	**37.4**

also consistently outperforms the SAP I3D model by 0.8% on top-1 accuracy. Our SAOA significantly outperforms the "*Baseline*" model. We obtained 3.1% and 1.9% improvements for R-50 and I3D, respectively. Compared to the SAP R-50 results on the test sets in Table 5, our SAOA R-50 boosts the verb top-1 accuracy from 63.2% to 64.0% and from 53.2% to 55.1% on the test seen set and unseen set, respectively. This demonstrates that our local alignment is effective for verb classification.

4.3.4 Benefit of the multi-modal fusion

Inspired by the two-stream network [8], [54], we aim to leverage a late fusion of the predictions from the SAOA RGB model and the SAOA Flow model to further boost the performance. We conduct the experiments using I3D as the backbone, and we report the results for both verb classification and noun classification on the EPIC-Kitchens validation set in Table 4. Our SAOA based on the I3D backbone with the "RGB+Flow+Obj" inputs achieves the highest performance. "SAOA (RGB+Flow+Obj)" outperforms "SAOA (RGB+Obj)" and "SAOA (Flow+Obj)" by 3.5% and 4.7% on top-1 accuracy for verb classification, respectively. For the noun classification scenario, we observe a 2.7% and a 1.6% improvements when comparing "SAOA (RGB+Flow+Obj)" with "SAOA (RGB+Obj)" and "SAOA (Flow+Obj)", respectively. This shows that our two-stream SAOA framework is capable of integrating benefits from both RGB and Flow inputs.

4.4 Comparison with State-of-the-art Results

We compare our model with the following state-of-the-art methods. ***TSN*** [56] is a 2D CNN model for video recognition. The performance is provided by the dataset authors. ***ORN*** [19] introduces object relation reasoning upon detection features, while the interactions between the verb and noun branches are largely ignored. ***R(2+1)D 34*** [55] indicates the CNN model pre-trained at a very large scale dataset IG-Kinetics (over 65 million videos). ***LFB*** [15] combines Long-Term Feature Banks (detection features) with 3D CNN to improve the accuracy of object recognition. "LFB Max" denotes their best operation on EPIC-Kitchens, which leverages max pooling for feature bank aggregation. ***LSTA*** [33] is an attention-based method, they only report the top-1 action accuracy on the test set. ***TBN*** [20] takes the RGB, Flow, and Audio modalities as input and performs mid-level fusion instead of late fusion. ***SAP*** [21] is our previous work which utilizes global alignment to integrate object features for both verb and noun branches. It also benefits from the symbiotic attention mechanism, and is trained on the R-50 backbone with the RGB input modality.

Table 5 summarizes the top-1 and top-5 accuracy for action, verb and noun predictions on the EPIC-Kitchens dataset. We develop our approach with R-50 and I3D backbone, and we leverage both RGB and optical flow as the input types for I3D and only RGB frames for R-50. In the Pre-training column, "Kinetics" indicates the backbone is pre-trained on Kinetics [5] directly. "Kinetics+ImageNet" indicates the backbone is pre-trained using the I3D [5] strategy, which first initializes the 3D CNN with the inflated weights of the 2D CNN pre-trained on ImageNet [23] and then trains the 3D CNN on Kinetics. "IG-Kinetics" indicates the backbone is pre-trained on a large-scale dataset, *i.e.*, IG-Kinetics [55], with weak supervision.

The first part of Table 5 shows the results of our method and the baselines on the EPIC-Kitchens validation set. The proposed SAOA outperforms the baselines by a large margin on all modalities with both two backbones. Specifically, with RGB frames as input, our SAOA R-50 significantly boosts the top-1 accuracy from 19.5% to 25.7% on action classification. With optical flow as input, our SAOA R-50 outperforms the baseline by 8.1% in top-1 accuracy on action recognition, demonstrating that SAOA can incorporate optical flow input effectively. Compared to I3D baselines, our SAOA I3D consistently improves the action recognition performance with different modalities. The remarkable performance gains mainly benefit from the symbiotic attention mechanism and the integration of the location-aware object information. Our SAOA R-50 achieves higher top-1 accuracy than the original SAP on verb prediction and action prediction. The improvement is owing to the integration of the location-aware alignment for the verb feature. Compared to the model with single modality input, the two-stream SAOA achieves higher accuracy, which demonstrates the effectiveness of the proposed multi-modal fusion strategy.

Our model outperforms the state-of-the-art methods by a large margin on all three evaluation splits, *i.e.*, the validation set, the test seen set and the test unseen set. On the validation set, compared to "LFB Max", which also utilizes the detection features, our two-stream SAOA (I3D) on the action prediction significantly improves the top-1 accuracy from 22.8% to 28.8%. With the same type of input (RGB+Obj), our SAOA (R-50) still outperforms them by 3.0%. The sig-

TABLE 5
The comparison with the baseline models and state-of-the-art methods on the EPIC-Kitchens dataset. "Obj" indicates the method leverages the information from the object detection model. ↑ indicates the improvement of our method compared to the baseline.

Method	Input Type	Pre-training	Actions		Verbs		Nouns	
			top-1	top-5	top-1	top-5	top-1	top-5
Validation								
ORN [19]	RGB+Obj	ImageNet	-	-	40.9	-	-	-
R(2+1)D-34 [55]	RGB	IG-Kinetics	22.5	39.2	56.6	**83.5**	32.7	55.5
LFB Max [15]	RGB+Obj	Kinetics+ImageNet	22.8	41.1	52.6	81.2	31.8	56.8
SAP (R-50) [21]	RGB+Obj	Kinetics	25.0	44.7	55.9	81.9	35.0	60.4
Baseline (R-50)	RGB	Kinetics	19.5	36.0	54.6	80.9	23.8	45.1
SAOA (R-50)	RGB+Obj	Kinetics	25.7 (6.2↑)	45.9	57.7	82.3	34.8	59.7
Baseline (R-50)	Flow	Kinetics	16.6	32.8	53.2	79.6	19.7	40.7
SAOA (R-50)	Flow+Obj	Kinetics	24.7 (8.1↑)	43.0	56.1	81.3	33.6	58.7
Baseline (R-50)	RGB+Flow	Kinetics	22.0	40.2	59.3	83.3	27.7	50.9
Our SAOA (R-50)	RGB+Flow+Obj	Kinetics	27.9 (5.9↑)	47.5	**61.0**	**83.8**	36.1	61.6
Baseline (I3D)	RGB	Kinetics+ImageNet	20.5	39.2	53.2	80.4	26.2	51.3
Our SAOA (I3D)	RGB+Obj	Kinetics+ImageNet	24.3 (3.8↑)	44.3	55.1	80.1	34.7	61.4
Baseline (I3D)	Flow	Kinetics+ImageNet	17.9	35.6	54.5	79.9	22.7	45.6
Our SAOA (I3D)	FLow+Obj	Kinetics+ImageNet	25.2 (7.3↑)	43.1	56.9	79.7	35.0	59.7
Baseline (I3D)	RGB+Flow	Kinetics+ImageNet	23.3	43.1	59.7	83.2	29.9	56.0
Our SAOA (I3D)	RGB+Flow+Obj	Kinetics+ImageNet	**28.8** (5.5↑)	**48.4**	60.4	82.8	**37.4**	**63.8**
Test seen								
TSN RGB [56]	RGB	ImageNet	22.4	44.8	48.0	87.0	38.9	65.5
TSN Flow [56]	Flow	ImageNet	16.8	33.8	51.7	84.6	26.8	50.6
TSN Fusion [56]	RGB+Flow	ImageNet	25.4	45.7	54.7	87.2	40.1	65.8
R(2+1)D-34 [55]	RGB	IG-Kinetics	34.4	54.2	63.3	87.5	46.3	69.6
LSTA [33]	RGB+Flow	ImageNet	30.2	-	-	-	-	-
LFB Max [15]	RGB+Obj	Kinetics+ImageNet	32.7	55.3	60.0	88.4	45.0	71.8
TBN [20]	RGB+Flow	Kinetics+ImageNet	30.3	51.8	60.9	89.7	42.9	68.6
TBN [20]	RGB+Flow+Audio	Kinetics+ImageNet	34.8	56.7	64.8	**90.7**	46.0	71.3
SAP R-50 [21]	RGB+Obj	Kinetics	34.8	55.9	63.2	86.1	48.3	71.5
Our SAOA (R-50)	RGB+Obj	Kinetics	37.0	58.3	64.0	88.0	**49.6**	**73.2**
Our SAOA (I3D)	RGB+Obj	Kinetics+ImageNet	33.8	55.3	63.6	87.4	46.1	70.0
Our SAOA (I3D)	Flow+Obj	Kinetics+ImageNet	33.4	54.7	63.8	86.8	45.7	69.2
Our SAOA (I3D)	RGB+Flow+Obj	Kinetics+ImageNet	**37.7**	**59.2**	**67.6**	89.2	47.8	71.8
Test Unseen								
TSN RGB [56]	RGB	ImageNet	11.3	26.3	36.5	74.4	22.6	46.9
TSN Flow [56]	Flow	ImageNet	13.5	27.5	47.4	77.0	21.2	42.5
TSN Fusion [56]	RGB+Flow	ImageNet	14.8	29.8	46.1	76.7	24.3	49.3
R(2+1)D-34 [55]	RGB	IG-Kinetics	23.7	39.1	55.5	80.9	33.6	56.7
LSTA [33]	RGB+Flow	ImageNet	15.9	-	-	-	-	-
LFB Max [15]	RGB+Obj	Kinetics+ImageNet	21.2	39.4	50.9	77.6	31.5	57.8
TBN [20]	RGB+Flow	Kinetics+ImageNet	16.8	32.6	49.6	78.4	25.7	50.9
TBN [20]	RGB+Flow+Audio	Kinetics+ImageNet	19.1	36.5	52.7	79.9	27.9	53.8
SAP R-50 [21]	RGB+Obj	Kinetics	23.9	40.5	53.2	78.2	33.0	58.0
Our SAOA (R-50)	RGB+Obj	Kinetics	23.3	41.2	55.1	79.9	32.3	57.1
Our SAOA (I3D)	RGB+Obj	Kinetics+ImageNet	21.9	42.1	52.9	79.9	31.7	58.5
Our SAOA (I3D)	Flow+Obj	Kinetics+ImageNet	23.2	42.4	55.5	80.1	32.6	58.1
Our SAOA (I3D)	RGB+Flow+Obj	Kinetics+ImageNet	**25.8**	**45.1**	**58.1**	82.6	**34.4**	**60.4**

nificant improvement mainly benefits from the interactions between the verb branch, noun branch, and the location-aware alignment with the location-aware object information. Although R(2+1)D 34 [55] uses much more videos to train the model, our best model still outperforms them by 6.3% in top-1 accuracy for action classification.

On the test seen set and the test unseen set, compared to the previous state-of-the-art method TBN, our two-stream SAOA (I3D) outperforms the recognition accuracy by a large margin. Specifically, the improvement of top-1 accuracy on the unseen set is 6.7%, 5.4%, and 6.5% for action, verb, and noun, respectively. Compared to our vanilla SAP model [21], our two-stream SAOA (I3D) achieves higher accuracy on all metrics. This demonstrates the effectiveness of the proposed location-aware alignment and the multi-modal fusion strategy.

4.5 EPIC-Kitchens Action Recognition Challenge 2020

We further verified the effectiveness of our framework on the EPIC-Kitchens Action Recognition Challenge. Our method achieved first place on both the seen set and unseen set. As shown in Table 6, we compare our approach with the top-3 submissions of Action Recognition Challenge and three published works on the leaderboard.

Notably, on the unseen test set, our single model (two-stream SAOA I3D) achieves higher performance than the TBN [20] Ensemble on all evaluation metrics. We also report the result of our final submission to the challenge, which fuses the predictions of six models trained with different backbones and input modalities. For the final action recognition, our ensemble achieves 42.6% top-1 accuracy on seen set and 28.0 % on unseen set, which are higher than the TBN ensemble by 5.9% and 7.0%, respectively. Our result outperforms the second place submission on seen set by

TABLE 6
Comparison with the methods on the leaderboard of EPIC-Kitchens Action Recognition Challenge. The results of Attention Clusters are borrowed from [20].

	Method	Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall		
		Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
Seen	Attention Clusters [57]	40.4	19.4	11.1	78.1	41.7	24.4	21.2	9.7	2.5	14.9	11.5	3.4
	TSN Fusion [14]	48.2	36.7	20.5	84.1	62.3	39.8	47.3	35.4	11.6	22.3	30.5	9.8
	LSTA Ensemble [58]	63.3	44.8	35.5	89.0	69.9	57.2	63.2	42.3	19.8	37.8	41.3	21.2
	TBN Ensemble [20]	66.1	47.9	36.7	91.3	72.8	58.6	60.7	44.9	24.0	46.8	43.9	22.9
	Sudhakaran (3rd place)	68.7	49.4	40.0	91.0	72.5	60.2	60.6	45.5	21.8	47.2	45.8	24.3
	action banks (2nd place)	66.7	49.6	41.6	90.1	77.0	64.1	59.4	45.6	25.4	41.7	46.3	27.0
	two-stream SAOA I3D	67.6	47.8	37.7	89.2	71.8	59.3	57.8	42.1	19.6	42.7	44.8	20.7
	Our SAOA (1st place)	**70.4**	**52.9**	**42.6**	90.8	76.6	63.6	60.4	**47.1**	24.9	45.8	**50.0**	26.9
Unseen	Attention Clusters [57]	32.4	12.0	5.6	69.9	31.8	15.7	17.2	3.9	1.8	11.6	7.9	2.6
	TSN Fusion [14]	39.4	22.7	10.9	74.3	45.7	25.3	22.5	15.3	6.2	13.1	17.5	6.5
	LSTA Ensemble [58]	49.4	27.1	20.3	77.5	52.0	37.6	31.1	21.1	9.2	18.7	21.9	14.2
	TBN Ensemble [20]	54.5	30.4	21.0	81.2	55.7	39.4	32.6	21.7	11.0	27.6	25.6	13.3
	action banks (3rd place)	54.6	33.5	27.0	80.4	61.0	46.4	33.6	30.5	15.0	25.3	28.4	18.0
	aptx4869lm (2nd place)	60.1	38.1	27.4	82.0	63.8	45.2	33.6	31.9	16.5	29.3	33.9	20.1
	two-stream SAOA I3D	58.1	34.4	25.8	82.6	60.4	45.1	38.9	28.7	14.8	28.7	30.1	17.5
	Our SAOA (1st place)	**60.4**	37.3	**28.0**	**83.1**	63.7	**46.8**	35.2	**32.6**	**17.4**	29.0	32.8	19.8

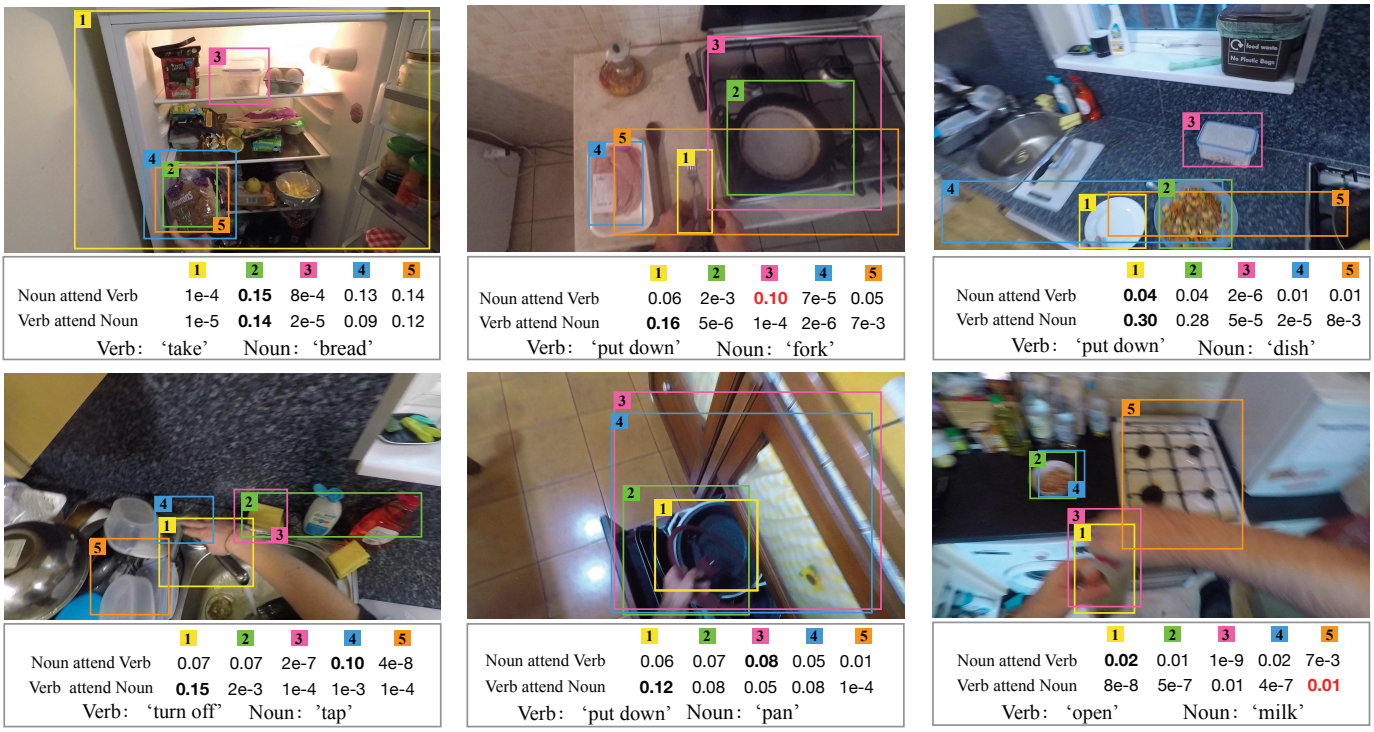


Fig. 4. Qualitative results of our SAOA I3D (Flow) model. The colored boxes show the top-5 detected regions and the numbers are the corresponding attention weights generated by our action-attended relation module. Red indicates the failure case.

1.0% and 0.6% on unseen set. The final rank is based on top-1 action accuracy.

4.6 Visualization

In Fig. 4, we show some qualitative results on the EPIC-Kitchens validation set.

The colored boxes in the figure indicate the top confident object proposals generated by the pre-trained detection model. We do not use labels of detected objects since they are not accurate. Instead, we use the object feature and location to guide the mutual communication of the verb and noun branch. The numbers below each image are the values of ARM attention weights for the five object-centric features.

As illustrated in the first video frame (the left-top one), the ground truth of this video clip is "take bread". Our ARM module generates the highest value for the feature corresponding to the second box where the interaction happened. The weights for the fourth and fifth box is similar to the second box since their locations are very close and also the boxes also contain the object "bread". The distracting objects with the first and third boxes obtain the lowest scores. For the last figure in the second row, our ARM failed to produce correct values for the boxes in the noun branch. This is owing to the large camera motion and occlusion of the objects. According to the qualitative analysis of the six examples, we can observe that the attention weights of the

noun branch ("Verb attend Noun") are more accurate than the values of the verb branch ("Noun attend Verb").

5 Conclusion

In this paper, we propose a novel framework named Symbiotic Attention with Object-centric feature Alignment (SAOA) for egocentric action recognition. We introduce a local and global alignment method to integrate the location-aware object information. Local motion features are produced to bridge the semantic gap between the motion feature and the object detection feature. We introduce a new attention mechanism called symbiotic attention that interactively leverages sources from the verb branch, the noun branch, and the location-aware object information. We evaluate SAOA on two backbones, two modalities, and the largest egocentric action recognition dataset. Our experimental results demonstrate the effectiveness of our framework, and we significantly outperform the state-of-the-art methods on the largest egocentric video dataset. In the future, we will explore to suppress background distractors in the convolutional backbones. It is promising to leverage other attention mechanisms to integrate multiple sources of information.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[4] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[6] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[7] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense." in *ICCV*, 2017.

[8] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, 2014.

[9] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[10] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *ACCV*, 2018.

[11] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Val Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*, 2016.

[12] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *ECCV*, 2018.

[13] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. W. Mayol-Cuevas, "You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video." in *BMVC*, vol. 2, 2014, p. 3.

[14] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *ECCV*, 2018.

[15] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick, "Long-term feature banks for detailed video understanding," in *CVPR*, 2019.

[16] D. Damen, H. Doughty, G. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "The epic-kitchens dataset: Collection, challenges and baselines," *IEEE T-PAMI*, 2020.

[17] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *ICCV*, 2019.

[18] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *ECCV*, 2018.

[19] F. Baradel, N. Neverova, C. Wolf, J. Mille, and G. Mori, "Object level visual reasoning in videos," in *ECCV*, 2018.

[20] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *ICCV*, 2019.

[21] X. Wang, Y. Wu, L. Zhu, and Y. Yang, "Symbiotic attention with privileged information for egocentric action recognition," in *AAAI*, 2020.

[22] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *CVPR*, 2018.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[24] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *ICCV*, 2017.

[25] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2018.

[26] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *PAMI*, vol. 41, no. 11, pp. 2740–2755, 2018.

[27] L. Zhu, Z. Xu, and Y. Yang, "Bidirectional multirate reconstruction for temporal modeling in videos," in *CVPR*, 2017.

[28] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE T-PAMI*, 2017.

[29] L. Zhu, L. Sevilla-Lara, D. Tran, M. Feiszli, Y. Yang, and H. Wang, "Faster recurrent networks for video classification," in *AAAI*, 2020.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[31] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *ICCV*, 2011.

[32] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in *CVPR*, 2016.

[33] S. Sudhakaran, S. Escalera, and O. Lanz, "Lsta: Long short-term attention for egocentric action recognition," in *CVPR*, 2019.

[34] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *ECCV*, 2018.

[35] A. Furnari and G. M. Farinella, "What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention," in *ICCV*, 2019, pp. 6252–6261.

[36] A. Furnari, S. Battiato, K. Grauman, and G. M. Farinella, "Next-active-object prediction from egocentric videos," *Journal of Visual Communication and Image Representation*, vol. 49, pp. 401–411, 2017.

[37] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *CVPR*, 2018.

[38] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *ECCV*, 2018.

[39] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu, "Pairwise body-part attention for recognizing human-object interactions," in *ECCV*, 2018.

[40] X. Wang and A. Gupta, "Videos as space-time region graphs," in *ECCV*, 2018.

[41] C. Sun, A. Shrivastava, C. Vondrick, K. Murphy, R. Sukthankar, and C. Schmid, "Actor-centric relation network," in *ECCV*, 2018.

[42] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

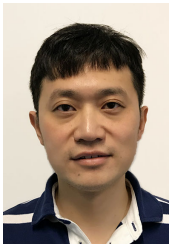
[44] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[45] D. Linsley, D. Shiebler, S. Eberhardt, and T. Serre, "Learning what and where to attend," in *International Conference on Learning Representations*, 2019.
[46] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *ICCV*, 2019.
[47] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.
[48] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *ICLR*, 2017.
[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
[50] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.
[51] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *T-PAMI*, 2015.
[52] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn." *IEEE TPAMI*, 2018.
[53] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, 2016.
[54] Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, and X. Xue, "Multi-stream multi-class fusion of deep networks for video classification," in *ACM Multimedia*, 2016, pp. 791–800.
[55] D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *CVPR*, 2019.
[56] W. Price and D. Damen, "An evaluation of action recognition models on epic-kitchens," *arXiv preprint arXiv:1908.00867*, 2019.
[57] X. Long, C. Gan, G. De Melo, J. Wu, X. Liu, and S. Wen, "Attention clusters: Purely attention based local feature integration for video classification," in *CVPR*, 2018.
[58] S. Sudhakaran, S. Escalera, and O. Lanz, "Fbk-hupba submission to the epic-kitchens 2019 action recognition challenge," *arXiv preprint arXiv:1906.08960*, 2019.

Xiaohan Wang received the B.E. degree from University of Science and Technology of China, Hefei, China, in 2017. He is currently a PhD student in Australian Artificial Intelligence Institute, University of Technology Sydney, Australia. His research interests include video analytics and visual reasoning.

Linchao Zhu received the Ph.D. degree in computer science from University of Technology Sydney, Australia, in 2019. He received the B.E. degree from Zhejiang University, China, in 2015. He is currently a lecturer in Australian Artificial Intelligence Institute, University of Technology Sydney, Australia. His research interests are video analysis and understanding.

Yu Wu received the B.E. degree from Shanghai Jiao Tong University, China, in 2015. He is currently a Ph.D. candidate in Australian Artificial Intelligence Institute, University of Technology Sydney, Australia. His research interests are video analysis and understanding.



Yi Yang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He is currently a professor with the University of Technology Sydney, Australia. He was a post-doctoral researcher in the School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania. His current research interests include machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, surveillance video analysis, and video content understanding.