

# Symbiotic Attention with Privileged Information for Egocentric Action Recognition

Xiaohan Wang<sup>1,2</sup>, Yu Wu<sup>1,2</sup>, Linchao Zhu<sup>1</sup>, Yi Yang<sup>1\*</sup>

<sup>1</sup>ReLER, University of Technology Sydney <sup>2</sup>Baidu Research  
 {xiaohan.wang-3,yu.wu-3}@student.uts.edu.au  
 {linchao.zhu,yi.yang}@uts.edu.au

## Abstract

Egocentric video recognition is a natural testbed for diverse interaction reasoning. Due to the large action vocabulary in egocentric video datasets, recent studies usually utilize a two-branch structure for action recognition, *i.e.*, one branch for verb classification and the other branch for noun classification. However, correlation studies between the verb and the noun branches have been largely ignored. Besides, the two branches fail to exploit local features due to the absence of a position-aware attention mechanism. In this paper, we propose a novel Symbiotic Attention framework leveraging Privileged information (SAP) for egocentric video recognition. Finer position-aware object detection features can facilitate the understanding of actor’s interaction with the object. We introduce these features in action recognition and regard them as privileged information. Our framework enables mutual communication among the verb branch, the noun branch, and the privileged information. This communication process not only injects local details into global features but also exploits implicit guidance about the spatio-temporal position of an on-going action. We introduce novel symbiotic attention (SA) to enable effective communication. It first normalizes the detection guided features on *one branch* to underline the action-relevant information from *the other branch*. SA adaptively enhances the interactions among the three sources. To further catalyze this communication, spatial relations are uncovered for the selection of most action-relevant information. It identifies the most valuable and discriminative feature for classification. We validate the effectiveness of our SAP quantitatively and qualitatively. Notably, it achieves the state-of-the-art on two large-scale egocentric video datasets.

## Introduction

We have witnessed a significant progress in tackling many computer vision problems, *e.g.*, image classification (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016; Huang et al. 2017), detection (Girshick et al. 2014; Girshick 2015; Ren et al. 2015; He et al. 2017), segmentation (Long, Shelhamer, and Darrell 2015; Chen et al. 2017; He et al. 2017). In video analysis, with the emerging of

\*This work was done when Xiaohan Wang and Yu Wu interned at Baidu Research. Yi Yang is the corresponding author. Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

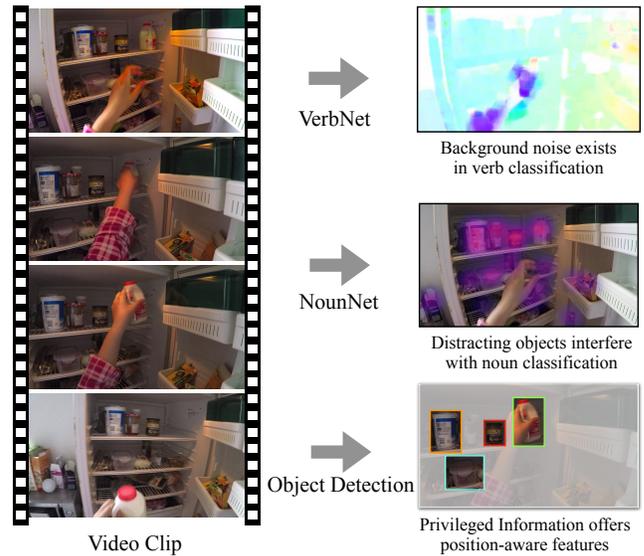


Figure 1: Three sources of information are leveraged. The VerbNet extracts motion information from raw videos, while background noise possibly degrades the recognition of target action. The NounNet recognizes the object in the scene. However, distracting objects interfere with accurate noun classification. Local position-aware object detection features serves as privileged information to enhance the communication between two branches.

deep convolutional neural networks and large-scale datasets, the action recognition performance has been prominently boosted (Simonyan and Zisserman 2014; Tran et al. 2015; Zhu et al. 2018; Wang et al. 2016; Xie et al. 2018; Wu et al. 2019b). In typical action recognition datasets (Carreira and Zisserman 2017; Goyal et al. 2017), the video duration is usually less than 10 seconds. These trimmed videos contain a single action that dominates the whole clip. As the background environment is not distracting, it is often not necessary to identify the interacting object. However, in many real-world applications, *e.g.*, a robot navigates in the mall, the surroundings are noisy with multiple actors and objects.

Human action recognition in videos has evolved from

classifying a single action in a clear background to understanding complex human-object interactions in a highly distracting environment. To enable the recognition of more complex videos, a challenging large-scale first-person dataset, *i.e.*, EPIC-Kitchens (Damen et al. 2018), was recently introduced for egocentric daily human activities understanding. This dataset provides rich interactions, covering adequate objects and natural actions. Compared to third-person action recognition, it requires to distinguish the object that human is interacting with from various small distracting objects. The intense camera motion, occlusion, and first-person viewpoint make it even more challenging to recognize fine actions.

In EPIC-Kitchens, due to the large action vocabulary, The verb and the noun classifiers are usually trained separately. The verb branch focuses on classifying verbs, *e.g.*, put, open, that the actor is performing. Large camera motion and subtle occurring action positions are the main obstacles for verb classification. The noun branch is to classify the object the actor is interacting with. As shown in Figure 1, distracting objects in oblique view decrease the prediction score of the interacting object. The predictions from the two branches are usually merged without interactions for action classification. (Wu et al. 2019a) utilized 3D convolution neural networks (CNN) for the standalone verb and noun classification. They leverage object detection features for longer context modeling. However, 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 not very effective. (Baradel et al. 2018) introduced object relation network for high-level object reasoning, where the relation modeling facilitates object recognition. However, due to the existence of distracting objects, the learned object relatedness without guided supervision might be not useful to identify the object that human is interacting with. These works ignore the mutual communication between the standalone branches. Instead, they only focus on contextual modeling and relation reasoning on a single branch. Even for a human, it can be difficult to recognize an action by only looking at objects while ignoring the actor’s intention, or only understanding motion changes without the awareness of the interacting object. To better exploit the mutual benefits of the interactions among different sources, we make the following contributions.

First, privileged local features are dynamically integrated to encourage the learning of action-relevant representations. For verb classification, the negative effect of background noise can be suppressed when target object information is effectively leveraged. However, noun representation loses finer spatial information. It fails to provide detailed position information to be exploited for the attendance of an ongoing action. Position-aware object detection features offer detailed local understanding of the objects. These features can serve as privileged information to possibly reduce the object-irrelevant motions. In noun classification, the introduced privileged information enhances the significance of correlated objects. The feature of action-relevant objects will be reinforced, thanks to the finer object presentations from the object detector.

Second, we propose symbiotic attention to enable mutual interactions among the three sources. The privileged information is first merged with the global feature from *one branch*. It is then normalized by a gated channel attention mechanism, which reweights the merged feature. This normalization process underlines the action-relevant information from the feature in *the other branch*. After this, each spatial position has integrated all three sources, *i.e.*, the two branches information, and the privileged information. We leverage a spatial relation module to further catalyze the communication between the verb and the noun features. Most action-relevant information is identified to generate a discriminative feature for final classification. This symbiotic attention mechanism dynamically integrates three sources of information towards better action recognition.

The effectiveness of our SAP framework is validated quantitatively and qualitatively. We achieve the state-of-the-art performance on two large-scale egocentric video datasets, *i.e.*, EPIC-Kitchens and EGTEA (Li, Liu, and Rehg 2018). Notably, we outperform the state-of-the-art (Wu et al. 2019a) by 2.7% on EPIC-Kitchens test unseen set.

## Related Work

### Deep Video Recognition

Deep learning methods have achieved promising performance on the video classification task. (Simonyan and Zisserman 2014) proposed to utilize both RGB frames and optical flow as the 2D CNN input to modeling appearance and motion, respectively. TSN (Wang et al. 2016) extended the two-stream CNN by extracting features from multiple temporal segments. (Tran et al. 2015) proposed a 3D CNN to learn the spatial-temporal information. I3D initializes 3D CNN with the inflated weights of 2D CNN. (Hara, Kataoka, and Satoh 2018) evaluated various 3D CNN architectures on a large-scale video dataset and demonstrated the effectiveness of 3D models. More recently, (Xie et al. 2018) and (Tran et al. 2018) proposed to decompose the 3D kernel to spatial and temporal convolution. Moreover, Recurrent Neural Networks (RNNs) are effective architectures for temporal modeling and have been found useful for video classification in (Abu-El-Haija et al. 2016; Zhu, Xu, and Yang 2017). 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.

### First Person Action Recognition

Compared to third-person video recognition, egocentric action recognition is more dependent on the modeling of the hand-object interaction. (Fathi, Farhadi, and Rehg 2011) proposed to learn a hierarchical model which exploits the consistent appearance of objects, hands, and actions and refines the object prediction based on action context. (Ma, Fan, and Kitani 2016) utilized a hand segmentation net to locate the object of interest. After that, the cropped regions and optical flow images are fed to a two-stream CNN to learn the action and object representation jointly. (Baradel et al. 2018) 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. 2019a) combined Long-Term Feature Banks that contains object-centric detection features with 3D CNN to improve the accuracy of object recognition. The attention mechanism is efficient to locate the region of interest on the feature map. (Sudhakaran, Escalera, and Lanz 2019) 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, Liu, and Rehg 2018) proposed to generate attention map of the hand-object interaction by the guide of the gaze information. (Kazakos et al. 2019) developed a egocentric action recognition model using three modalities.

## Human-Object Interaction

Reasoning the interaction between human and objects is relevant to our task. Most methods in this field are based on detection models. For example, (Gkioxari et al. 2018) predicted a density map to locate the interacted object and calculated the action score, with a modified Faster RCNN architecture. (Qi et al. 2018) proposed Graph Parsing Neural Networks that incorporates structural knowledge and deep object detection model. (Fang et al. 2018) developed a pairwise body-part attention model which 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 and Gupta 2018) proposed to represent videos as space-time region graphs, which models shape dynamics and relationships between actors and objects. (Sun et al. 2018) 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.

## Proposed Method

### Overview

In this section, we illustrate our network architecture for egocentric video recognition. We develop three base 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 has the same architecture as VerbNet. It is trained to produce a feature representing object appearance. (3) Object detection model takes sampled individual frames as input. We use Faster RCNN as our detector and utilize *RoIAlign* operation to obtain the local object features as privileged information. The output features of the three base models are fed to the subsequent SAP module. We aim to enable effective communication among VerbNet, NounNet, and Privileged Information. The SAP module generates two feature vectors which can be used to predict verb class and noun class. The overall framework is illustrated in Fig. 2.

### Preliminaries

For each input egocentric video  $X = \{x^1, \dots, x^t\}$  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  $f^v \in \mathbb{R}^C$  contains the motion information, where  $C$  is the dimension of the extracted feature. Differently, the noun feature  $f^n \in \mathbb{R}^C$  contains the global appearance information.

To enhance the global representation through the communication between two branches, we use a pre-trained detection model to provide detailed information of objects in the video. Considering the efficiency, for each video, we only use  $M$  sampled frames for detection inference. The output of the *RoIAlign* layer of the detection model is regarded as the feature for each detected object. To save memory usage and reduce the noisy information, we only keep top- $K$  object features according to their confidence scores for each sampled frame. Thus, we have the auxiliary object feature matrix  $f^o \in \mathbb{R}^{N \times C}$ , which contains  $N = M \times K$  object features of the video. The verb feature  $f^v$ , noun feature  $f^n$ , and object feature matrix  $f^o$  are interacted with each other to produce more discriminative features for action recognition with the following SAP module.

### Symbiotic Attention with Privileged Information

As illustrated in Fig. 2, SAP includes three stages. First, the privileged information is integrated into the global feature from one branch. Second, the fused object-centric features are recalibrated by the other branch utilizing 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. Considering the symmetry of SAP, we only formulize the noun branch as an example in the following section.

**Privileged Information Integration.** The separated verb branch and noun branch produce two feature vectors  $f^v$  and  $f^n$  by global average pooling. The local information in these features is indistinct. We aim to leverage object feature matrix  $f^o$  as privileged information to inject the local details into the global features. Moreover, the position-aware object features can also guide the model to attend salient area. Thus, we need to fuse  $f^o$  with  $f^v$  and  $f^n$ , respectively. Besides, it's necessary to avoid jumbling the object feature matrix  $f^o$ . To these ends, we perform a concatenation operation on the object feature matrix and the broadcasting global feature vectors. After that, we use a nonlinear transformation to enhance the fused feature. Formally, this operation can be presented as follow:

$$f^{\hat{n}} = \text{ReLU}(W_f^n f^n + W_f^o f^o + b_f), \quad (1)$$

where  $W_f^n, W_f^o \in \mathbb{R}^{C \times C}, b_f \in \mathbb{R}^C$  and  $f^{\hat{n}} \in \mathbb{R}^{N \times C}$ . Each column in  $f^{\hat{n}}$  represent a object-centric feature, which integrates the global noun appearance with a explicit local object information.

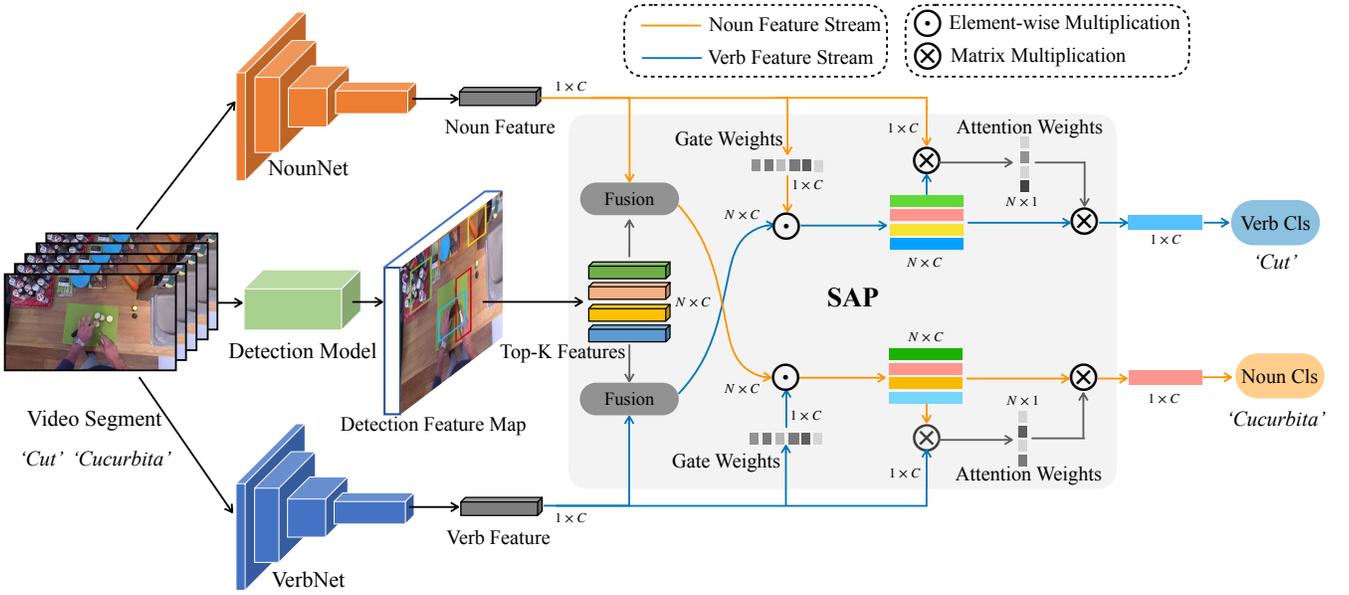


Figure 2: The proposed method. Our framework consists of three feature extractors and one interaction module SAP. VerbNet and NounNet produce global features. Detection Model generates a set of local object features as privileged information. The privileged information is integrated into the global features to obtain object-centric feature matrices. These feature matrices are normalized by a cross stream gating mechanism. After that, the object-centric matrices are attended by the other branch to select the most action-relevant information. The outputs of SAP are used to classify the verb and noun, respectively.

**Cross Stream Gating.** The fused object-centric feature matrix contains useful local details. However, due to the existence of inaccurate detection regions, there are quite a few disturbing background noises in the features. To address this problem, we propose a gated channel attention to underline the action relevant information. Furthermore, to enhance the interaction between the verb stream and noun stream, we utilize a cross gating mechanism. For an input noun feature matrix  $f^n$ , we generate gating weights for it using the verb feature  $f^v$ :

$$g^n = \text{Sigmoid}(W_g^n f^v + b_g), \quad (2)$$

where  $W_g^n \in \mathbb{R}^{C \times C}$ ,  $b_g \in \mathbb{R}^C$  and  $g^n \in \mathbb{R}^C$ . The output is produced by rescaling the noun feature matrix with the gating weights:

$$f_g^n = g^n \odot f^n, \quad (3)$$

where  $f_g^n \in \mathbb{R}^{N \times C}$  and  $\odot$  denotes the element-wise multiplication. After re-calibrating the object-centric noun feature by the verb feature, the action-unrelated noise can be suppressed. Moreover, the cross gating mechanism enables mutual communication between the two branches, which exploits the correlations of verbs and nouns adaptively.

**Action-attended Relation Module** The calibrated object-centric feature matrix contains the action-relevant information and implicit guidance about the spatio-temporal position of an on-going action. To make full use of the information, we consider uncovering the relationships among the features. First, we propose to assess the relevance between the global feature and position-aware object-centric

feature. Second, we sum the object-centric features weighted by the relevance coefficients. Specifically, we perform attention mechanism on the normalized object-centric noun features  $f_g^n$  and the original verb feature  $f^v$  by

$$f_a^n = \text{Softmax}(f_g^n f^v) f_g^n, \quad (4)$$

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

## Training and Objectives

We use Faster R-CNN with ResNeXt-101-FPN backbone as our object detector. Following the training procedure in (Wu et al. 2019a), we first pre-train the detector on Visual Genome and then finetune it on EPIC-Kitchens object detection set. For VerbNet and NounNet, we adopt 3D Resnet-50 (Hara, Kataoka, and Satoh 2018) as our backbones. The two nets are both initialized with Kinetics pretrained weights. We first individually train the VerbNet and NounNet with the corresponding Cross-Entropy Loss:  $\mathcal{L}^v$  and  $\mathcal{L}^n$ . After the base training stage, we cascade our SAP module and fine-tune the entire model in an end-to-end manner. The objective for the fine-tuning is the sum of  $\mathcal{L}^v$  and  $\mathcal{L}^n$ .

## Action Re-weighting

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 EPIC-Kitchens dataset (Damen et al. 2018). Following the approach in (Wu et al. 2019a), we re-weight the final action probability by a prior, *i.e.*

$$P(\text{action} = y) = \mu(y^v, y^n)P(\text{verb} = y^v)P(\text{noun} = y^n), \quad (5)$$

where  $\mu$  is the occurrence frequency of action in training set.

## Experiments

### Datasets

We evaluate our method on two large-scale egocentric datasets: EPIC-Kitchens (Damen et al. 2018) and EGTEA (Li, Liu, and Rehg 2018).

**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 activities performed 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 and 321 noun classes. We split the original training set to new training and validation set following (Baradel et al. 2018). We focus on the recognition task on EPIC-Kitchens, which is to predict the verb, noun, and the combination pair in each video segment.

**EGTEA** is a large-scale egocentric video dataset which consists of 10,321 video clips annotated with 19 verb classes, 51 noun classes, and 106 action classes. There are no bounding boxes annotations in this dataset.

### Experiment Settings

We implement and test our method using Caffe2, PaddlePaddle and Pytorch. We observe similar performance from these deep learning frameworks. Specifically, we pretrain the base models (VerbNet and NounNet) individually and then fine-tune the entire model in an end-to-end manner. We find in our experiments that end-to-end training consistently yields better performance than two-stage training, which only updates the SAP model at the second stage. Next, we first illustrate the details on how to pre-train the backbones (Backbone details) and how to extract privileged information (Detection details). Finally, we show the details of the end-to-end fine-tuning (SAP details).

**Backbone details.** We take the Kinetics pre-trained ResNet50-3D model as the initialization of our backbone model. We then train the backbone models (VerbNet and NounNet) individually on the target dataset. The input of the two models is the trimmed video clips with 64 frames. The targets for the VerbNet and NounNet are the verb label and noun label, respectively. Videos are decoded at 60 FPS for the EPIC-Kitchens dataset, and 24 FPS for the EGTEA dataset. We adopt the stochastic gradient descent (SGD) with momentum 0.9 and weight decay 0.0001 to optimize the parameters for 40 epochs. The overall learning rate is initialized to 0.003 and then changed to 0.0003 in the last 10 epochs. The batch size is 32. During training, the frame size is  $224 \times 224$  pixels, randomly cropped from a random scaled video whose side is randomly sampled in [224, 288]. We sample 64 frames with stride=2 for both datasets. During

testing, for each input video segment, we use the center clip and scale it to the size 256. Then we use the center cropped frames with size 224 as the input.

**Detection details.** Following (Wu et al. 2019a), we use the same Faster R-CNN to detect objects and extract object features. The detector is first pre-trained on Visual Genome (Krishna et al. 2016) and then fine-tuned on the training split of the EPIC-Kitchens dataset. For fine-tuning, we use a batch size of 12 and train the model for 15k iterations. We use an initial learning rate of 0.005, which is decreased by a factor of 10 at iteration 116k and 133k. Finally, our object features are extracted using RoIAlign from the detector’s feature maps. Considering the efficiency, for a video clip, we extract object features on the center window with a size of 6 seconds. The sample rate is two frames per second. For each frame, we keep the top five features according to the confidence scores. Therefore, we have 60 detection features for a video clip. For EGTEA, we use the detection model pretrained on EPIC-Kitchens to extract object features.

**SAP details.** With the pre-trained backbone models and the detection results, we fine-tune the backbone models with our SAP module in an end-to-end manner. We use the same SGD to optimize the parameters for 40 epochs with batch-size of 32. The learning rate is initialized to 0.0001 and then reduced to 0.00001 in the last 20 epochs. The rest training details are the same as the backbone details.

### Comparison with State-of-the-art Results

We compare our model with the following state-of-the-art methods. **TSN** (Price and Damen 2019) is a two-stream model for video recognition. The performance is provided by the dataset authors. **ORN** (Baradel et al. 2018) introduces object relation reasoning upon detection features, while the interactions between verb and noun branches are largely ignored. **LFB** (Wu et al. 2019a) combines Long-Term Feature Banks (detection features) with 3D CNN to improve the accuracy of object recognition. “LFB Max” denotes their best operation in EPIC-Kitchens, which leverages max pooling for feature bank aggregation. **LSTA** (Sudhakaran, Escalera, and Lanz 2019) is an attention-based method, they only report the top-1 action accuracy on the test set. Table 1 summarizes the top-1 and top-5 accuracy for verb, noun, and action predictions on the EPIC-Kitchens dataset.

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 (S1) set and the test unseen (S2) set. On the validation set, compared to our baseline model (“Ours Baseline”), our SAP on the noun prediction significantly improves the top-1 accuracy from 23.8% to 35.0%. Compared to “LFB Max”, which also utilizes the detection features, our method outperforms them by 3.2% at top-1 accuracy. For the verb prediction, our SAP obtains 1.3% (from 54.6% to 55.9%) top-1 accuracy improvement compared to our baseline model. For the final action classification, our method achieves 25.0% top-1 accuracy, which is higher than “LFB Max” by 2.2%. The significant improvement mainly benefits from the interactions between the verb branch, noun branch, and privileged information. Similar performance improve-

Method	Pre-training	Verbs		Nouns		Actions	
		top-1	top-5	top-1	top-5	top-1	top-5
<b>Validation</b>							
ORN (Baradel et al. 2018)	ImageNet	40.9	-	-	-	-	-
I3D GFA (Wang et al. 2019)	Kinetics+ImageNet	-	-	34.1	60.4	-	-
R(2+1)D 34 (Ghadiyaram, Tran, and Mahajan 2019)	Kinetics	46.8	79.2	25.6	47.5	15.3	29.4
LFB Max (Wu et al. 2019a)	Kinetics+ImageNet	52.6	81.2	31.8	56.8	22.8	41.1
Ours Baseline	Kinetics	54.6	80.9	23.8	45.1	19.5	36.0
Ours SAP	Kinetics	<b>55.9</b>	<b>81.9</b>	<b>35.0</b>	<b>60.4</b>	<b>25.0</b>	<b>44.7</b>
<b>Test seen (S1)</b>							
TSN RGB (Price and Damen 2019)	ImageNet	48.0	87.0	38.9	65.5	22.40	44.8
TSN Flow (Price and Damen 2019)	ImageNet	51.7	84.6	26.8	50.6	16.8	33.8
TSN Fusion (Price and Damen 2019)	ImageNet	54.7	87.2	40.1	65.8	25.4	45.7
R(2+1)D 34 (Ghadiyaram, Tran, and Mahajan 2019)	Kinetics	59.1	87.4	38.0	62.7	26.8	46.1
LSTA (Sudhakaran, Escalera, and Lanz 2019)	ImageNet	-	-	-	-	30.2	-
LFB Max (Wu et al. 2019a)	Kinetics+ImageNet	60.0	<b>88.4</b>	45.0	<b>71.8</b>	32.7	55.3
Ours SAP	Kinetics	<b>63.2</b>	86.1	<b>48.3</b>	<b>71.5</b>	<b>34.8</b>	<b>55.9</b>
<b>Test Unseen (S2)</b>							
TSN RGB (Price and Damen 2019)	ImageNet	36.5	74.4	22.6	46.9	11.3	26.3
TSN Flow (Price and Damen 2019)	ImageNet	47.4	77.0	21.2	42.5	13.5	27.5
TSN Fusion (Price and Damen 2019)	ImageNet	46.1	76.7	24.3	49.3	14.8	29.8
R(2+1)D 34 (Ghadiyaram, Tran, and Mahajan 2019)	Kinetics	48.4	77.2	26.6	50.4	16.8	31.2
LSTA (Sudhakaran, Escalera, and Lanz 2019)	ImageNet	-	-	-	-	15.9	-
LFB Max (Wu et al. 2019a)	Kinetics+ImageNet	50.9	77.6	31.5	57.8	21.2	39.4
Ours SAP	Kinetics	<b>53.2</b>	<b>78.2</b>	<b>33.0</b>	<b>58.0</b>	<b>23.9</b>	<b>40.5</b>

Table 1: The comparison with the state-of-the-art methods on the EPIC-Kitchens dataset.

ment is also observed on the test seen (S1) set and the test unseen (S2) set. For the final action prediction, Our SAP outperforms the state-of-the-art method “LFB Max” by 2.1% on S1 and 2.7% on S2. Although (Ghadiyaram, Tran, and Mahajan 2019) use much more videos (65M videos) and extreme deep 3D CNN to train the model, we still outperform their best model on noun prediction and action prediction on S1.

The results on the EGTEA dataset is shown in Table 3. The results of Two Stream, I3D and TSN, are provided by the dataset developer (Li, Liu, and Rehg 2018). Ego-RNN (Sudhakaran and Lanz 2018) and LSTA (Sudhakaran, Escalera, and Lanz 2019) utilize RNNs and attention mechanism for egocentric video recognition. Our method achieves higher accuracy on all splits than the state-of-the-art.

## Ablation Studies

We conduct extensive ablation studies to evaluate the effectiveness of each component in our SAP model. Table 2 shows the *noun* prediction accuracy of several variants of our method. The first row “CNN Baseline” indicates we only use the noun branch, without the help of the verb and detection feature. “Noun + Verb” is the model that we take the verb feature as the gate to enhance the noun feature. Specifically, the implementation of the gating operation is the same as the single branch in CSG. “Avg Pooling” and “Max Pooling” take only the privileged information as input. We apply the corresponding pooling operation on the  $M * N$  detection features and use a fully connected layer to output the noun prediction. The last six rows show the effectiveness of

Inputs			Methods	Top-1
Noun	Verb	Pril		
✓	-	-	CNN Baseline	23.8
✓	✓	-	Noun + Verb	24.2
-	-	✓	Avg Pooling	24.5
-	-	✓	Max Pooling	25.6
-	✓	✓	Ours w/o CSG & ARM	30.4
✓	✓	✓	Ours w/o Gating	33.6
✓	✓	✓	Ours w/o Cross Stream	33.2
✓	✓	✓	Ours w/o CSG	32.6
✓	✓	✓	Ours w/o ARM	32.7
✓	✓	✓	Ours	<b>35.0</b>

Table 2: Ablation Study based on noun prediction on the EPIC-Kitchens validation set. “Pril” denotes the privileged information from finer object detection features, ARM denotes the action-attended relation module, CSG denotes the cross stream gating.

our component CSG and ARM. Specifically, we decompose CSG to two part: Cross Stream and Gating. We also investigate the impact of each part. In the table, “Ours w/o Cross Stream” indicates using the same stream to gate and attend the object-centric matrix.

**Importance of the privileged information.** The first two rows in Table 2 shows the importance of the privileged information guidance. Without the detection feature, both the noun branch (“CNN Baseline”) and the two branches with

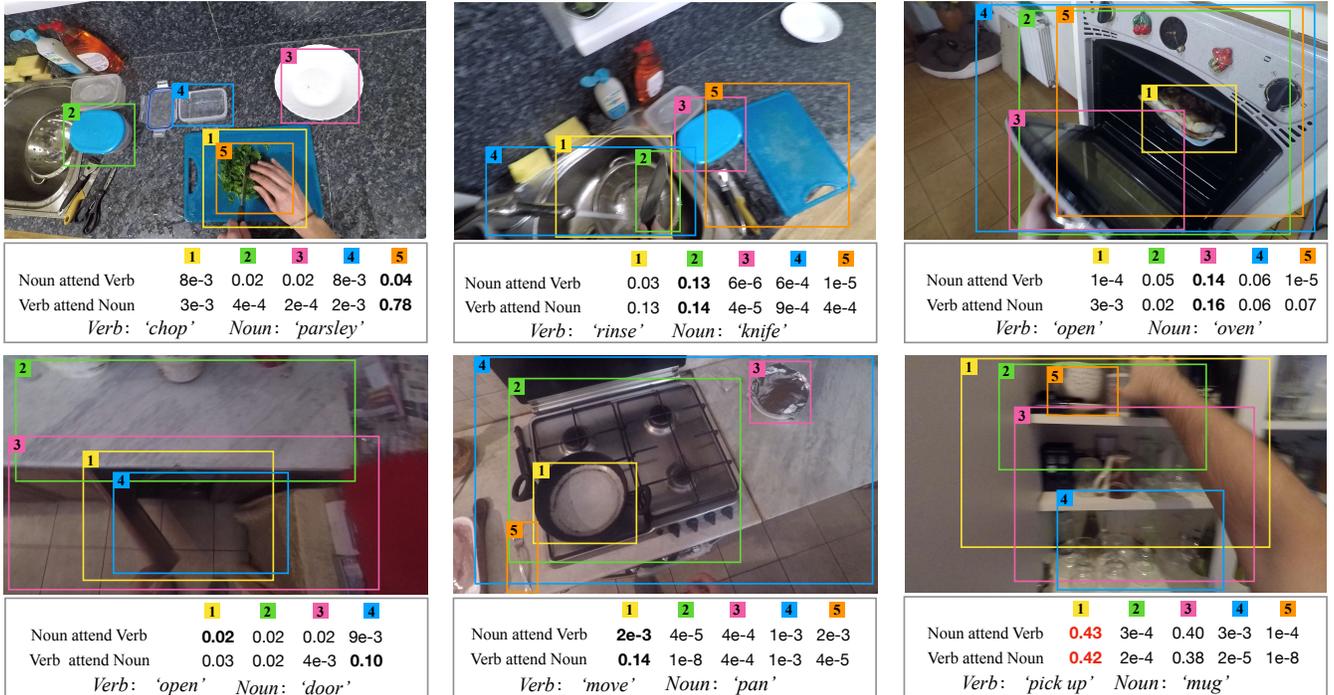


Figure 3: Qualitative results of our SAP model. The colored boxes show the top-5 detected regions and the numbers are the corresponding attention weights generated by our action-attended relation module. Red indicates the failure case.

Methods	Split1	Split2	Split3	Average
Two Stream	43.8	41.5	40.3	41.8
I3D	54.2	51.5	49.4	51.7
TSN	58.0	55.0	54.8	55.9
Ego-RNN	62.2	61.5	58.6	60.8
LSTA	-	-	-	61.9
Ours	<b>64.1</b>	<b>62.1</b>	<b>62.0</b>	<b>62.7</b>

Table 3: The comparison with the state-of-the-art methods on the EGTEA dataset.

communication of noun and verb branches (“Noun + Verb”) fail in the noun prediction. It is consistent with our motivation that the local information provided by the detection features is a critical clue in the noun prediction.

In addition, the third row (“Avg Pooling”) and the fourth row (“Max Pooling”) also achieve comparable performance compared to the “CNN Baseline” model, which indicates the detection features contains rich information for noun prediction. Even without the input of the whole video clip, the local details from the detection model are enough for predicting the interacted object.

**Importance of Cross Stream Gating.** The performance comparison between the model “Ours w/o CSG & ARM” and the model “Ours w/o ARM” validates the effectiveness of the CSG module. The CSG module enables mutual communication between the verb branch and the noun branch. Therefore, the CSG can improve the performance from 30.4% to 32.7%. Moreover, the results of “Ours w/o

Cross Stream” and “Ours w/o Gating” demonstrate the two components both benefits the noun prediction.

**Importance of Action-attended Relation Module.** The last rows of Table 2 shows the improvement of the proposed ARM. ARM can select the most action-relevant information from the object-centric features and explore the relationships in the spatio-temporal context. Therefore, the ARM further improves the performance from 32.7% to 35.0%.

## Visualization

In Fig. 3, we show some qualitative results on the EPIC-Kitchens dataset. The colored boxes in the figure indicate the top confident objects found by the pre-trained detection model. We do not use labels of detected objects since they are not accurate. Instead, we use the detection feature to guide the mutual communication of the verb and noun branch. The numbers below each image are the value of ARM attention weights for the five object-centric features. Taking the first one (the left-top one) as an example, the ground truth of this video clip is “chop parsley”. The pre-trained detection model generates five proposals. Our ARM module correctly finds the interacted area with “parsley” and generates a high value (0.78) that describes the contribution of the enhanced feature to the final noun classification.

## Conclusion

In this paper, we propose a novel framework named Symbiotic Attention with Privileged Information for egocentric action recognition. We introduce a new attention mecha-

nism called symbiotic attention that can interactively leverage sources from the verb branch, the noun branch, and the privileged information. Our experimental results demonstrate the effectiveness of our framework, and we outperform the state-of-the-art methods on large-scale egocentric video datasets. In the future, we will explore a hierarchical structure that can readily interpret the multi-step attention process. It is also promising to design new models to suppress background distractors directly.

## References

- [Abu-El-Haija et al. 2016] Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; and Vijayanarasimhan, S. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.
- [Baradel et al. 2018] Baradel, F.; Neverova, N.; Wolf, C.; Mille, J.; and Mori, G. 2018. Object level visual reasoning in videos. In *ECCV*.
- [Carreira and Zisserman 2017] Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.
- [Chen et al. 2017] Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *T-PAMI*.
- [Damen et al. 2018] Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; and Wray, M. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*.
- [Fang et al. 2018] Fang, H.-S.; Cao, J.; Tai, Y.-W.; and Lu, C. 2018. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*.
- [Fathi, Farhadi, and Rehg 2011] Fathi, A.; Farhadi, A.; and Rehg, J. M. 2011. Understanding egocentric activities. In *ICCV*.
- [Ghadiyaram, Tran, and Mahajan 2019] Ghadiyaram, D.; Tran, D.; and Mahajan, D. 2019. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*.
- [Girshick et al. 2014] Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.
- [Girshick 2015] Girshick, R. 2015. Fast r-cnn. In *ICCV*.
- [Gkioxari et al. 2018] Gkioxari, G.; Girshick, R.; Dollár, P.; and He, K. 2018. Detecting and recognizing human-object interactions. In *CVPR*.
- [Goyal et al. 2017] Goyal, R.; Kahou, S. E.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Freund, I.; Yianilos, P.; Mueller-Freitag, M.; et al. 2017. The "something something" video database for learning and evaluating visual common sense. In *ICCV*.
- [Hara, Kataoka, and Satoh 2018] Hara, K.; Kataoka, H.; and Satoh, Y. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*.
- [He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.
- [He et al. 2017] He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2017. Mask r-cnn. In *ICCV*.
- [Huang et al. 2017] Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*.
- [Kazakos et al. 2019] Kazakos, E.; Nagrani, A.; Zisserman, A.; and Damen, D. 2019. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*.
- [Krishna et al. 2016] Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.
- [Krizhevsky, Sutskever, and Hinton 2012] Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NeurIPS*.
- [Li, Liu, and Rehg 2018] Li, Y.; Liu, M.; and Rehg, J. M. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *ECCV*.
- [Long, Shelhamer, and Darrell 2015] Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.
- [Ma, Fan, and Kitani 2016] Ma, M.; Fan, H.; and Kitani, K. M. 2016. Going deeper into first-person activity recognition. In *CVPR*.
- [Price and Damen 2019] Price, W., and Damen, D. 2019. An evaluation of action recognition models on epic-kitchens. *arXiv preprint arXiv:1908.00867*.
- [Qi et al. 2018] Qi, S.; Wang, W.; Jia, B.; Shen, J.; and Zhu, S.-C. 2018. Learning human-object interactions by graph parsing neural networks. In *ECCV*.
- [Ren et al. 2015] Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *T-PAMI*.
- [Simonyan and Zisserman 2014] Simonyan, K., and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*.
- [Sudhakaran and Lanz 2018] Sudhakaran, S., and Lanz, O. 2018. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. In *BMVC*.
- [Sudhakaran, Escalera, and Lanz 2019] Sudhakaran, S.; Escalera, S.; and Lanz, O. 2019. Lsta: Long short-term attention for egocentric action recognition. In *CVPR*.
- [Sun et al. 2018] Sun, C.; Shrivastava, A.; Vondrick, C.; Murphy, K.; Sukthankar, R.; and Schmid, C. 2018. Actor-centric relation network. In *ECCV*.
- [Tran et al. 2015] Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.
- [Tran et al. 2018] Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*.
- [Wang and Gupta 2018] Wang, X., and Gupta, A. 2018. Videos as space-time region graphs. In *ECCV*.
- [Wang et al. 2016] Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*.
- [Wang et al. 2019] Wang, X.; Wu, Y.; Zhu, L.; and Yang, Y. 2019. Baidu-uts submission to the epic-kitchens action recognition challenge 2019. In *CVPR-W*.
- [Wu et al. 2019a] Wu, C.-Y.; Feichtenhofer, C.; Fan, H.; He, K.; Krahenbuhl, P.; and Girshick, R. 2019a. Long-term feature banks for detailed video understanding. In *CVPR*.
- [Wu et al. 2019b] Wu, Y.; Zhu, L.; Yan, Y.; and Yang, Y. 2019b. Dual attention matching for audio-visual event localization. In *ICCV*.

- [Xie et al. 2018] Xie, S.; Sun, C.; Huang, J.; Tu, Z.; and Murphy, K. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*.
- [Zhu et al. 2018] Zhu, Y.; Lan, Z.; Newsam, S.; and Hauptmann, A. 2018. Hidden two-stream convolutional networks for action recognition. In *ACCV*.
- [Zhu, Xu, and Yang 2017] Zhu, L.; Xu, Z.; and Yang, Y. 2017. Bidirectional multirate reconstruction for temporal modeling in videos. In *CVPR*.