Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization

Junting Pan[1*] Siyu Chen[4*] Mike Zheng Shou[2]
Yu Liu[4] Jing Shao[4] Hongsheng Li[1,3]

[1]CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong [2]Columbia University
[3]School of CST, Xidian University [4] SenseTime Research

Abstract

Localizing persons and recognizing their actions from videos is a challenging task towards high-level video understanding. Recent advances have been achieved by modeling direct pairwise relations between entities. In this paper, we take one step further, not only model direct relations between pairs but also take into account indirect higher-order relations established upon multiple elements. We propose to explicitly model the ***Actor-Context-Actor Relation****, which is the relation between two actors based on their interactions with the context. To this end, we design an Actor-Context-Actor Relation Network (ACAR-Net) which builds upon a novel High-order Relation Reasoning Operator and an Actor-Context Feature Bank to enable indirect relation reasoning for spatio-temporal action localization. Experiments on AVA and UCF101-24 datasets show the advantages of modeling actor-context-actor relations, and visualization of attention maps further verifies that our model is capable of finding relevant higher-order relations to support action detection. Notably, our method ranks first in the AVA-Kinetics action localization task of ActivityNet Challenge 2020, outperforming other entries by a significant margin (+6.71 mAP). The code is available online.*[1]

1. Introduction

Spatio-temporal action localization, which requires localizing persons and recognizing their actions from videos, is an important task that has drawn increasing attention in recent years [15, 12, 8, 46, 35, 58, 52, 54, 41, 29, 55, 17, 20]. Unlike object detection which can be accomplished solely by observing visual appearances, activity recognition usually demands for reasoning about the actors' interactions with the surrounding context, including environments, other people and objects. Take Fig. 1 as an example. To recognize the action "ride" of the person in the red bounding box, we need to observe that he is inside a car, and there is a driver next to him. Therefore, most recent progress in spatio-temporal action detection has been driven by the success of relation modeling. These approaches focus on modeling relationships in terms of pairwise interactions between entities.

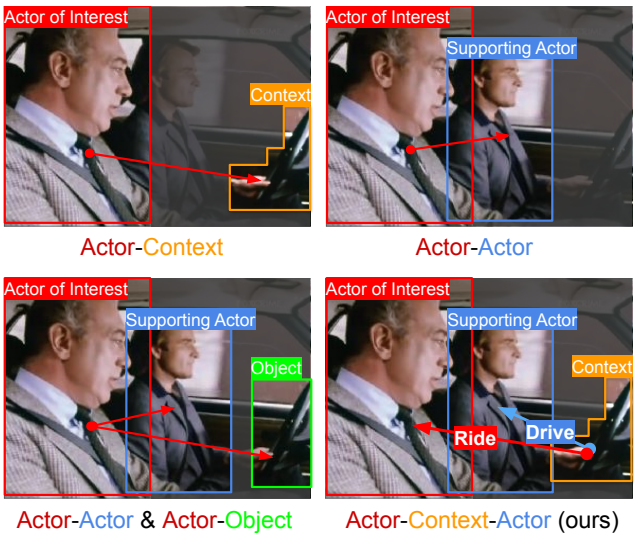

Figure 1. We contrast our Actor-Context-Actor relation modeling with existing relation reasoning approaches for action localization. Reasoning relations between pairs of entities may not always be sufficient for correctly predicting the action labels of all individuals. Our method not only reasons relations between actors, but also models connections between different actor-context relations. As an illustration, the relation between the blue actor and the steering wheel (drive) serves as a crucial clue for recognizing the action being performed by the red actor (ride).

However, it is not always the case that relations between elements can be formulated in terms of pairs; often, higher-order relations provide crucial clues for accurate action detection. In Fig. 1, it is difficult to infer the action of the red actor given only its relation with the blue actor, or only with the scene context (steering wheel). Instead, in order to identify that the red actor performs the action "ride", one has to reason over the interaction between the blue actor and the context (drive). In other words, it is necessary to capture the

*Equal contribution
[1]https://github.com/Siyu-C/ACAR-Net

implicit *second-order* relation between the two actors based on their respective *first-order* relations with the context.

There were previous works that employ Graph Neural Networks (GNNs) to implicitly model higher-order interactions between actors and contextual objects [45, 58, 38, 57, 10]. However, in these approaches, an extra pre-trained object detector is required, and only located objects are used as context. Since bounding-box annotations of objects in spatio-temporal action localization datasets are generally not provided, the pre-trained object detector is limited to its original object categories and may easily miss various objects in the scenes. In addition, the higher-order relations in these methods are limited to be inferred solely from contextual objects, which might miss important environmental or background cues for action classification.

To tackle the above issues, we propose an Actor-Context-Actor Relation Network (ACAR-Net) which focuses on modeling second-order relations in the form of Actor-Context-Actor relation. It deduces indirect relations between multiple actors and the context for action localization. The ACAR-Net takes both actor and context features as inputs. We define actor features as the features pooled from the actor regions of interest, while for context features, we directly use spatio-temporal grid feature maps from our backbone network. The context that we adopt does not rely on any extra object detector with predefined categories, thus making our overall design much simpler and flexible. Moreover, grid feature maps are capable of representing scene elements of various levels (*e.g*. instance level and part level) and types (*e.g*. background, objects and object parts), which is useful for fine-grained action discrimination. The proposed ACAR-Net first encodes first-order actor-context relations, and then applies a *High-Order Relation Reasoning Operator* to model interactions between the first-order relations. The High-Order Relation Reasoning Operator is fully convolutional and operates on first order relational features maps without losing spatial layouts . For supporting actor-context-actor relation reasoning between actors and context at different time periods, we build an *Actor-Context Feature Bank*, which contains actor-context relations from different time steps across the whole video.

We conduct extensive experiments on the challenging Atomic Visual Actions (AVA) dataset [15, 22] as well as the UCF101-24 dataset [34] for spatio-temporal action localization. Our proposed ACAR-Net leads to significant improvements on recognizing human-object and human-human interactions. Qualitative visualization shows that our method learns to attend contextual regions that are relevant to the action of interest.

Our contributions are summarized as the three-fold:

- We propose to model actor-context-actor relations for spatio-temporal action localization. Such relations are mostly ignored by previous methods but crucial for achieving accurate action localization.
- We propose a novel Actor-Context-Actor Relation Network for improving spatio-temporal action localization by explicitly reasoning about higher-order relations between actors and the context.
- We achieve state-of-the-art performances with significant margins on the AVA and UCF101-24 datasets. At the time of submission, our method ranks first on the ActivityNet leaderboard [7].

2. Related Work

Action Recognition. Research works on action recognition generally fall into three categories: action classification, temporal localization and spatio-temporal localization. Early works mainly focus on classifying a short video clip into an action class. 3D-CNN [40, 1, 8], two-stream network [33, 43, 9] and 2D-CNN [56, 6, 24] are the three dominant network architectures adopted for this task. While progress has been made for short trimmed video classification, the main research stream also moves forward to understand long untrimmed videos, which requires not only to recognize the category of each action instance but also to locate its start and end times. A handful of works [32, 53, 3, 60] consider this problem as a detection problem in 1D temporal dimension by extending object detection frameworks.

Spatio-Temporal Action Localization. Recently, the problem of spatio-temporal action localization has drawn considerable attention from the research community, and datasets (such as AVA [15, 22]) with atomic actions of all actors in the video being continuously annotated are introduced. It defines the action detection problem into a finer level, since the action instances need to be localized in both space and time. Typical approaches used by early works apply R-CNN detectors on 3D-CNN features [15, 11, 54, 50, 23]. Wu *et al*. [47] show that actor features obtained by running 3D-CNN backbone on top of the cropped and resized actor region from the original video preserve better spatial details than RoI-pooled actor features. Nevertheless, it has the limitation that computational costs and inference time almost increase linearly with the number of actors. Several more recent works have exploited graph-structured networks to leverage contextual information [35, 12, 58, 38, 41, 39].

Relational Reasoning for Video Understanding. Relational reasoning has been studied in the domain of video understanding [44, 45, 61, 36, 58, 35, 46, 19, 27, 38]. This is natural because recognizing the action of an actor depends on its relationships with other actors and objects. Zhou *et al*. [61] extend Relation Network [31] for modeling relations between video frames over time. Non-local Networks [44] leverage self-attention mechanisms to capture long range dependencies between different entities. Wang *et al*. [45] show

that representing videos with Space-time Region Graph improves action classification accuracy. In the context of spatio-temporal localization, there are many traditional approaches that are dedicated to capturing spatio-temporal relationships in videos [59, 28, 37, 18]. For deep neural networks based methods, Sun *et al*. [35] propose Actor-Centric Relation Network that learns to aggregate actor and scene features. Girdhar *et al*. [12] re-purpose the Transformer network [42] for encoding pairwise relationships between every two actor proposals. Concurrently, Wu *et al*. [46] use long-term feature banks (LFB) to provide temporal supportive information up to 60s for computing long range interaction between actors. Zhang *et al*. [58] propose to explicity model interactions between actors and objects. However, their approach focuses on modeling actor-object and actor-actor relations separately. When deducing the action of a person, the interactions of other persons with contextual objects are ignored. In other words, they do not explicitly model the actor-context-actor relations. In contrast, our method emphasizes modeling those higher-order relations. Perhaps the most similar work to ours is [38], which aggregates multiple types of interactions with stacked units akin to Transformer Networks [42]. Nonetheless, while this approach also supports actor-context-actor interactions, it treats object detection results as context, which requires extra pre-trained object detectors with fixed object categories and ignores other important types of contexts (such as background, objects not in the predefined categories, and specific parts of some objects).

3. Method

In this section, we provide detailed descriptions of our proposed Actor-Context-Actor Relation Network (ACAR-Net). Our ACAR-Net aims at effectively modeling and utilizing higher-order relations built upon the basic actor-actor and actor-context relations for achieving more accurate action localization.

3.1. Overall Framework

We first introduce our overall framework for action localization, where the proposed actor-context-actor relation (ACAR) modeling is the key module. The framework is designed to detect all persons in an input video clip ($\sim$2s in our experiments) and estimate their action labels. As shown in Fig. 2, following state-of-the-art methods [46, 8, 49], the framework is built based on an off-the-shelf person detector (*e.g.* Faster R-CNN [30]) and a video backbone network (*e.g.* I3D [2]). Person and context features are then processed by the proposed ACAR module with a long-term *Actor-Context Feature Bank* for final action prediction.

In details, the person (actor) detector operates on the center frame (*i.e.* key frame) of the input clip and obtains N detected actors. The detected boxes are duplicated to neighboring frames of the key frame in the clip. In the meantime, the backbone network extracts a spatio-temporal feature volume from the input video clip. We perform average pooling along the temporal dimension to save follow-up computational cost, which results in a feature map $X \in \mathbb{R}^{C\times H\times W}$, and C, H, W are channel, height and width respectively. We apply RoIAlign [16] (7×7 spatial output) followed by spatial max pooling to the N actor features, producing a series of N actor features, $A^1, A^2, \ldots, A^N \in \mathbb{R}^C$, each of which describes the spatio-temporal appearance and motion of one Region of Interest (RoI).

The proposed Actor-Context-Actor Relation (ACAR) module is illustrated on the right side of Fig. 2. This module takes the aforementioned video feature map X and RoI features $\{A^i\}_{i=1}^N$ as inputs, and outputs the final action predictions after relation reasoning. The ACAR module has two main operations. (1) It first encodes first-order actor-context relations between actors and spatial locations of the spatio-temporal context. Based on the actor-context relations, we further integrate a *High-order Relation Reasoning Operator* (HR^{2}O) for modeling the interactions between pairs of first-order relations, which are indirect relations mostly ignored by previous methods. (2) Our reasoning operation is extended with an Actor-Context Feature Bank (ACFB). The bank contains actor-context relations at different time stamps, and can provide more complete spatio-temporal context than the existing long-term feature bank [46] which only consists of features of actors. We will elaborate the two parts in the following sections. Notably, our high-order relation reasoning only requires action labels as supervision.

3.2. Actor-Context-Actor Relation Modeling

First-order actor-context relation encoding. We adopt the Actor-Centric Relation Network (ACRN) [35] as a module for encoding the first-order actor-context relations by combining RoI features $A^1, \ldots, A^N$ with the context feature X. More specifically, it replicates and concatenates each actor feature $A^i \in \mathbb{R}^C$ to all $H \times W$ spatial locations of the context feature $X \in \mathbb{R}^{C\times H\times W}$ to form a series of concatenated feature map $\{\tilde{F}^i\}_{i=1}^N \in \mathbb{R}^{2C\times H\times W}$. Actor-context relation features for each actor i can then be encoded by applying convolutions to this concatenated feature map $\tilde{F}^i$.

High-order relation reasoning. We now discuss how to compute high-order relations between two actors based on their first-order interactions with the context. Let $F^i_{x,y}$ record the first-order features between the actor A^i and the scene context X at the spatial location (x, y). We propose to model the relationship between first-order actor-context relations, which are high-order relations encoding more informative scene semantics. However, since there are a large number of actor-context relation features, $F^i_{x,y} \in \mathbb{R}^{C\times 1\times 1}$, $i \in \{1, \ldots, N\}, x \in [1, H], y \in [1, W]$, the number of their possible pairwise combinations is generally overwhelming. We therefore design a *High-order Relation Reasoning Oper-*

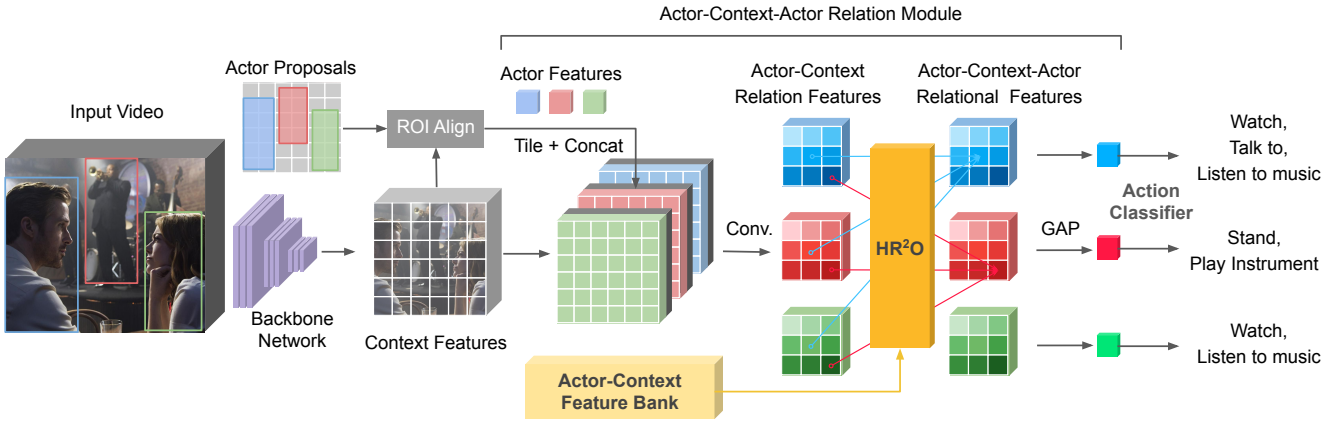

Figure 2. **Action Detection Framework.** Videos are processed with a Backbone Network to produce spatio-temporal context features. For each actor proposal (person bounding box), we extract actor features from the context features by RoIAlign. Given the actor and context features, the ACAR-Net computes second-order relation between every two actors based on their interaction with the context.

ator (HR^{2}O) that aims at learning the high-order relations between pairs of actor-context relations at the same spatial location (x, y), *i.e.*, $F^i_{x,y}$ and $F^j_{x,y}$. In this way, the proposed relational reasoning operator limits the relation learning to second-order actor-context-actor relations, *i.e.* two actors i and j can be associated via the same spatial context, denoted as $i \leftrightarrow (x, y) \leftrightarrow j$, to help the estimation of their actions.

Our proposed HR^{2}O takes as input a set of first-order actor-context relation feature maps $\{F^i\}_{i=1}^N$. The operator outputs $\{H^i\}_{i=1}^N = \text{HR}^2\text{O}(\{F^i\}_{i=1}^N)$ that encode second-order actor-context-actor relations for all actors. The operator is modeled as stacking several modified non-local blocks [44]. For each non-local block, convolutions are used to convert the input first-order actor-context relation feature maps F^i into query Q^i , key K^i and value V^i embeddings of the same spatial size as F^i. All feature maps are of dimension $d = 512$ in our implementation. It is worth noting that the use of convolutions is not only useful for aggregating local information but also makes the operator position and order-sensitive. The attention vectors are computed separately at every spatial location, and the Actor-Context-Actor Relation feature H^i is given by the linear combination of all value features $\{V^j\}_{j=1}^N$ according to their corresponding attention weights $Att^{i,j}$. The overall process can be summarized by the following equations,

$$\begin{aligned} Q^i, K^i, V^i &= \text{conv2D}(F^i) \\ Att^i_{x,y} &= \text{softmax}_j\left(\frac{\langle Q^i_{x,y}, K^j_{x,y}\rangle}{\sqrt{d}}\right), \\ \tilde{H}^i_{x,y} &= \sum_j Att^{i,j}_{x,y} V^j_{x,y}. \end{aligned} \tag{1}$$

Following [46], we also add layer normalization and dropout to our modified non-local block,

$$\begin{aligned} H^i &= \text{Dropout}(\text{Conv2D}(\text{ReLU}(\text{norm}(\tilde{H}^i)))), \\ F'^i &= F^i + H^i, \end{aligned} \tag{2}$$

where H^i and the input actor-context features F^i are fused via residual addition to obtain the actor-context-actor feature F'^i, which can be further processed by the following non-local block again.

We also exploit another instantiation, which directly obtains second-order actor-context-actor interaction features from actor features $\{A^i\}_{i=1}^N$ and the context feature X by a Relation Network [31]. More specifically, we obtain the relation feature between actors A^i, A^j and context $V_{x,y}$ as

$$H^{i,j}_{x,y} = f_\theta([A^i, A^j, V_{x,y}]), \tag{3}$$

where $[\cdot, \cdot, \cdot]$ denotes concatenation along the channel dimension and $f_\theta(\cdot)$ is a stack of two convolutional layers. The high-order relation of an actor i is calculated as the average of all relation features related to that actor,

$$H^i = \frac{1}{N}\sum_j H^{i,j}_{x,y}. \tag{4}$$

It is also fused with the input features to obtain actor-context-actor features via residual addition, *i.e.* $F'^i = F^i + H^i$. This method is computationally expensive when the number of actors N is large, since the number of feature triplets is proportional to N^2.

Action classifier. After the actor-context-actor feature maps $\{F'^i\}_{i=1}^N$ are obtained for all actors, a final action classifier is introduced as a single fully-connected layer with a non-linearity function to output the confidence scores of each actor belonging to different action classes.

3.3. Actor-Context Feature Bank

In order to support actor-context-actor relation reasoning between actors and context at different time periods in a long video, we propose an Actor-Context Feature Bank (ACFB), in which we store contextual information from both past and future. This is inspired by the Long-term Feature Bank (LFB) proposed in [46]. Yet instead of providing relational

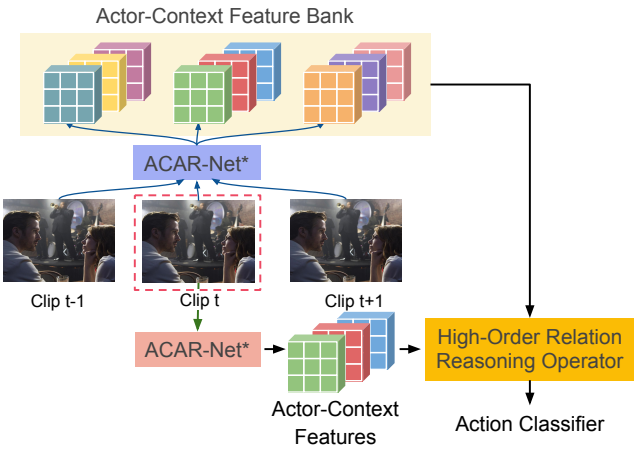

Figure 3. Illustration of ACAR-Net equipped with Actor-Context Feature Bank, where ACAR-Net* refers to the first-order relation extraction part of our proposed module.

features for long-term higher-order reasoning, the previous LFB only stores actor features for facilitating first-order actor-actor interaction recognition.

As is illustrated in Fig. 3, clips are evenly sampled (every 1 second) from an input video, and the clips (∼2s) could overlap with each other. We first train a separate ACAR-Net without any feature bank following the descriptions in Section 3.2. First-order actor-context relation features F^i of each actor in all clips of the entire video would be extracted by the separately pre-trained ACAR-Net and archived as the feature bank. To avoid confusion, we re-denote these acquired first-order features in the bank as L^i.

To train a new ACAR-Net with the support of the long-term actor-context feature bank to conduct high-order relation reasoning at some current time step t, we retrieve all M archived actor-context relation features $\{L^i\}_{i=1}^M$ from the frames within a time window $[t-w, t+w]$. Actor-context-actor interactions between short-term features (encoding first-order interactions at current time t) and long-term ones from the archived bank can be computed as $\{H^i\}_{i=1}^N = \text{HR}^2\text{O}(\{F^i\}_{i=1}^N, \{L^j\}_{j=1}^M)$. Note that, the HR^2O is the same as before, but the self-attention mechanism is replaced with the attention between current and long-term actor-context relations, where query features Q are still computed from short-term features $\{F^i\}_{i=1}^N$, but key and value features, K and V, are calculated with the long-term archived features $\{L^i\}_{j=1}^M$, *i.e.*,

$$\begin{aligned} Q^i &= \text{conv2D}(F^i), \\ K^j, V^j &= \text{conv2D}(L^j). \end{aligned} \tag{5}$$

Consequently, for any actor i at current time t, our ACAR-Net is now capable of reasoning about its higher-relations with actors and context over a much longer time span, and thus better captures what is happening in the temporal context for achieving more accurate action localization.

4. Experiments on AVA

AVA [15] is a video dataset for spatio-temporally localizing atomic visual actions. For AVA, box annotations and their corresponding action labels are provided on key frames of 430 15-minute videos with a temporal stride of 1 second. We use version 2.2 of AVA dataset by default. In addition to the current AVA dataset, Kinetics-700 [1] videos with AVA [15] style annotations are also introduced. The new AVA-Kinetics dataset [22] contains over 238k unique videos and more than 624k annotated frames. However, only a single frame is annotated for each video from Kinetics-700. Following the guidelines of the benchmarks, we only evaluate 60 action classes with mean Average Precision (mAP) as the metric, using a frame-level IoU threshold of 0.5.

4.1. Implementation Details

Person Detector. For person detection on key frames, we use the human detection boxes from [46], which are generated by a Faster R-CNN [30] with a ResNeXt-101-FPN [51, 25] backbone. The model is pre-trained with Detectron [13] on ImageNet [5] as well as the COCO human keypoint images [26], and fine-tuned on the AVA dataset.
Backbone Network. We use SlowFast networks [8] as the backbone in our localization framework and increase the spatial resolution of res_5 by $2\times$. We conduct ablation experiments using a SlowFast R-50 8×8 instantiation (without non-local blocks). The inputs are 64-frame clips, where we sample $T = 8$ frames with a temporal stride $\tau = 8$ for the slow pathway, and $\alpha T (\alpha = 4)$ frames for the fast pathway. The backbone is pre-trained on the Kinetics-400 dataset[2].

Training and inference. In AVA, actions are grouped into 3 major categories: poses (e.g. stand, walk), human-object and human-human interactions. Given that poses are mutually exclusive and interactions are not, we use softmax for poses and sigmoid for interactions before binary cross-entropy loss for training. We train all models end-to-end (except for the feature bank part) using synchronous SGD with a batch size of 32 clips. We train for 35k iterations with a base learning rate of 0.064, which is then decreased by a factor of 10 at iterations 33k and 34k. We perform linear warm-up [14] during the first 6k iterations. We use a weight decay of 10^{-7} and Nesterov momentum of 0.9. We use both ground-truth boxes and predicted human boxes from [46] for training. For inference, we scale the shorter side of input frames to 256 pixels and use detected person boxes with scores greater than 0.85 for final action classification.

4.2. Ablation Study

We conduct ablation experiments to investigate the effect of different components in our framework on AVA v2.2.

[2]The pre-trained SlowFast R-50 and SlowFast R-101+NL (in the following section) are downloaded from SlowFast's official repository.

	mAP
Baseline + STO [46]	26.10
Baseline + ACRN [35]	26.71
Baseline + AIA [38]	26.79
Baseline + HR^{2}O	**27.83**

(a) **Relation Modeling**

	mAP
Baseline	25.39
Baseline + HR^{2}O	27.83
ACAR	**28.84**

(b) **Component Analysis**

	mAP
Avg	26.97
RN	27.18
NL	**27.83**

(c) **HR^{2}O Design**

	mAP
Actor First	27.62
Context First	**27.83**

(d) **Relation Order**

	mAP
HR^{2}O-1L	27.63
HR^{2}O-2L	**27.83**
HR^{2}O-3L	27.25

(e) **Relation Depth**

	mAP
HR^{2}O	27.83
HR^{2}O + LFB [46]	27.75
HR^{2}O + ACFB	**28.84**

(f) **Feature Bank**

Table 1. **Ablation study on AVA dataset**. The "Baseline" of our framework only consists of the video backbone, actor detector and one-layer action classifier. HR^{2}O: High-order Relation Reasoning Operator. ACFB: Actor-Context Feature Bank.

The baseline of our framework only consists of the video backbone (SlowFast R-50), the actor detector and the single-layer action classifier (denoted as "Baseline" in Table 1).

Relation Modeling - Comparison. In order to show the effectiveness of our actor-context-actor relation reasoning module, we compare against several previous approaches that leverage relation reasoning for action localization based on our baseline. Here we focus on validating the effect of relation modeling only, thus we disable long-term support in this study. We adapt their reasoning modules such that all methods use the same baseline as our ACAR-Net in order to fairly compare only the impact of relation reasoning. We evaluate ACRN that focuses on learning actor-context relations; STO [46] (a degraded version of LFB) that only captures actor interactions within the current short clip; AIA (w/o memory) [38] that aggregates both actor-actor and actor-object interactions. As listed in Table 1a, our proposed actor-context-actor relation modeling ("Baseline + HR^{2}O" in Table 1a) significantly improves over the compared methods. We observe that AIA with both actor and context relations performs better than ACRN and STO which only model one type of first-order relations, yet our method based on high-order relation modeling outperforms all compared methods by considerable margins.

We further break down the performances of different relation reasoning modules into three major categories of the AVA dataset, which are poses (*e.g.* stand, sit, walk), human-object interactions (*e.g.* read, eat, drive) and human-human interactions (*e.g.* talk, listen, hug). Fig. 4 compares the gains of different approaches with respect to the baseline in terms of mAP on these major categories. We can see that our HR^{2}O gives more performance boosts on two interaction categories compared to the pose category, which is consistent with our motivation to model indirect relations between actors and context. Once equipped with ACFB, our framework can further improve on the pose category as well.

Finally, we contrast our ACAR with existing relation reasoning approaches in AVA. We visualize attention maps from different reasoning modules over an example key frame in Fig. 5. Without needing object proposals, ACAR is capable of localizing free-form context regions for indirectly establishing relations between two actors (the actor of interest is listening to the supporting actor reading a report). In comparison, the attention weights of STO as well as AIA are distributed more diversely and do not have a clear focus point. Note that we do not show the attention map of ACRN since it assigns equal weights to all context regions.

Component Analysis. To validate our design, we first ablate the impacts of different components of our ACAR as shown in Table 1b. We can observe that both HR^{2}O and ACFB lead to significant performance gains over baseline.

HR^{2}O Design. We test different instantiations of the High-order Relation Reasoning Operator on top of our baseline in Table 1c. Our modified non-local (denoted as "NL") mechanism works better than simply designing HR^{2}O as an average function (denoted as "Avg"), *i.e.* $H^i = \frac{1}{N}\sum_i F^i$,. In addition, the instantiation with relation network (RN) described in Section 3.2 also works alright. Nonetheless, the modified non-local attention is computationally more efficient than RN with feature triplets and has better performance.

Relation Ordering. There are two possible orders for reasoning actor-context-actor relations: 1) aggregating actor-actor relations first, or 2) encoding actor-context relations first. Note that our ACAR-Net adopts the latter one. We implement the former order by performing self-attention between actor features with the modified non-local attention before incorporating context features in our baseline. The results in Table 1d validate that context information should be aggregated earlier for better relation reasoning.

HR^{2}O Depth. In Table 1e, we observe that stacking two modified non-local blocks in HR^{2}O has higher mAP than the one-layer version, yet adding one more non-local block produces worse performance, possibly due to overfitting. We therefore adopt two non-local blocks as the default setting.

Actor-Context Feature Bank. In this set of experiments, we validate the effectiveness of the proposed ACFB. We set the "window size" $2w + 1$ to 21s due to memory limitations, and longer temporal support is expected to perform better [46]. As presented in Table 1f, adding long-term support with ACFB significantly improves the baseline (HR^{2}O's 27.83 $\rightarrow$ HR^{2}O + ACFB's **28.84**). We also test replacing the ACFB in our framework with the long-term feature bank (LFB) [46] (denoted as "HR^{2}O + LFB"). However, LFB even fails to match the baseline performance. This drop might be because LFB encodes only "zeroth-order" actor features, which cannot provide enough relational information from neighboring frames for assisting interaction recognition.

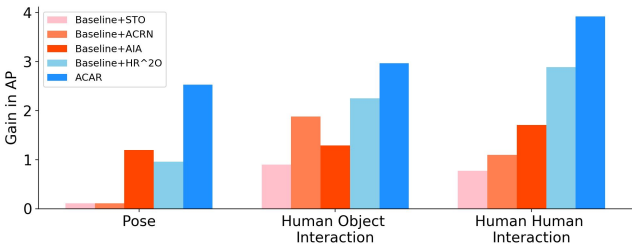

Figure 4. **Gains of mAP on three major categories of the AVA dataset with respect to Baseline.** Our ACAR consistently outperforms other relation reasoning methods, and achieves larger performance gains on the two interaction categories.

model	inputs	val mAP
ACRN, S3D [35]	V+F	17.4
Zhang *et al*. [58], I3D	V	22.2
Action TX, I3D [12]	V	25.0
LFB, R-50+NL [46]	V	25.8
LFB, R-101+NL [46]	V	27.4
SlowFast, R-50, 8×8 [8]	V	24.8
SlowFast, R-101, 8×8 [8]	V	26.3
Ours, R-50, 8×8	V	28.3
Ours, R-101, 8×8	V	**30.0**

Table 2. **Comparison with state-of-the-arts on AVA v2.1.** All models are pre-trained on Kinetics-400. V and F refer to visual frames and optical flow respectively.

model	pre-train	val mAP
SlowFast, R-101+NL [8]	Kinetics-600	29.0
AIA, R-101+NL [38]	Kinetics-700	32.3
Ours, R-101+NL	Kinetics-600	31.4
Ours, R-101	Kinetics-700	**33.3**

Table 3. **Comparison with state-of-the-arts on AVA 2.2.** We do not conduct testing with multiple scales and flips. All models use $T \times \tau = 8 \times 8$.

model	val mAP	test mAP
AIA++, ensemble [48]	-	32.91
MSF, ensemble [62]	-	31.88
SlowFast, R-101, 8×8 (our impl.)	32.98	-
Ours, R-101, 8×8	35.84	-
Ours++, R-101, 8×8	36.36	-
Ours++, ensemble	**40.49**	**39.62**

Table 4. **AVA-Kinetics results**. "++" refers to inference with 3 scales and horizontal flips. Models submitted to the test server are trained on both training and validation sets.

4.3. Comparison with State-of-the-arts on AVA

We compare our ACAR-Net with state-of-the-art methods on the validation set of both AVA v2.1 (Table 2) and v2.2 (Table 3). Note that we also provide results with more advanced video backbones, *i.e*. two SlowFast R-101 instantiations (with / without NL). On AVA v2.1, our framework achieves 30.0 mAP and outperforms all prior results with pre-trained Kinetics-400 backbone. On AVA v2.2, our ACAR-Net reaches 33.3 mAP with only single-scale testing, establishing a new state-of-the-art. Note that our method surpasses AIA [38] with only $1/3$ of temporal support. The results indicate that with proper modeling of higher-order relations, our approach extracts more informative cues from the context.

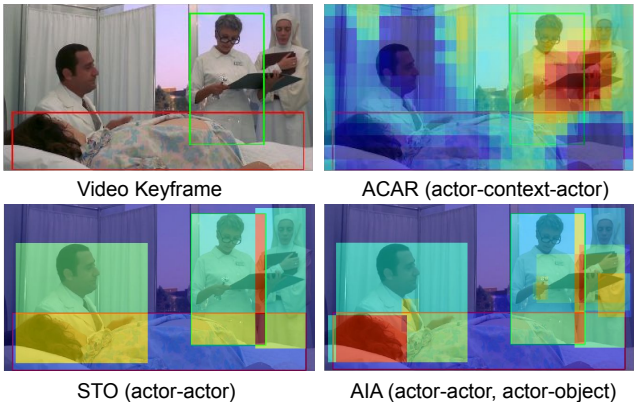

Figure 5. **Comparison of attention maps from different approaches of relation modeling for action detection.** Our method is able to attend the contextual region (some document) that relates the actor of interest marked in red (performing "listen to") and the supporting actors in the green box (performing "read"), while other methods fail to achieve similar effects.

We present our results on AVA-Kinetics in Table 4. Our baseline is already highly competitive (∼33 mAP). Yet integrating our ACAR modeling still leads to a significant gain of +2.86 mAP. This demonstrates that performance enhancement brought by high-order relation modeling can generalize to this new dataset. With an ensemble of models, we achieve **39.62 mAP** on the test set, ranking first in the AVA-Kinetics task of ActivityNet Challenge 2020 and outperforming other entries by a large margin (+6.71 mAP). More details on our winning solution are provided in the technical report [4].

4.4. Qualitative Results

Our proposed ACAR operates fully convolutionally on top of spatio-temporal features, and this allows us to visualize the actor-context-actor relation maps $\{Att^{i,j}\}$ generated by our High-order Relation Reasoning Operator. As shown in Fig. 6, the first two columns include the key frame as well as the corresponding relation map from the same clip, and the last three columns show the relation map denoting interactions with actors and context from a neighboring clip. We can observe that the attended regions usually include the actor of interest, supporting actors' body parts (*i.e*. head, hands and arm) and objects being in interaction with the actors. Take the first example on the left as an example. The green supporting actor A^j is taking a package from the red actor of interest A^i. Such information is well encoded by our ACAR-Net in the form of actor-context-actor relations: packages, hands and arms of both actors are highlighted.

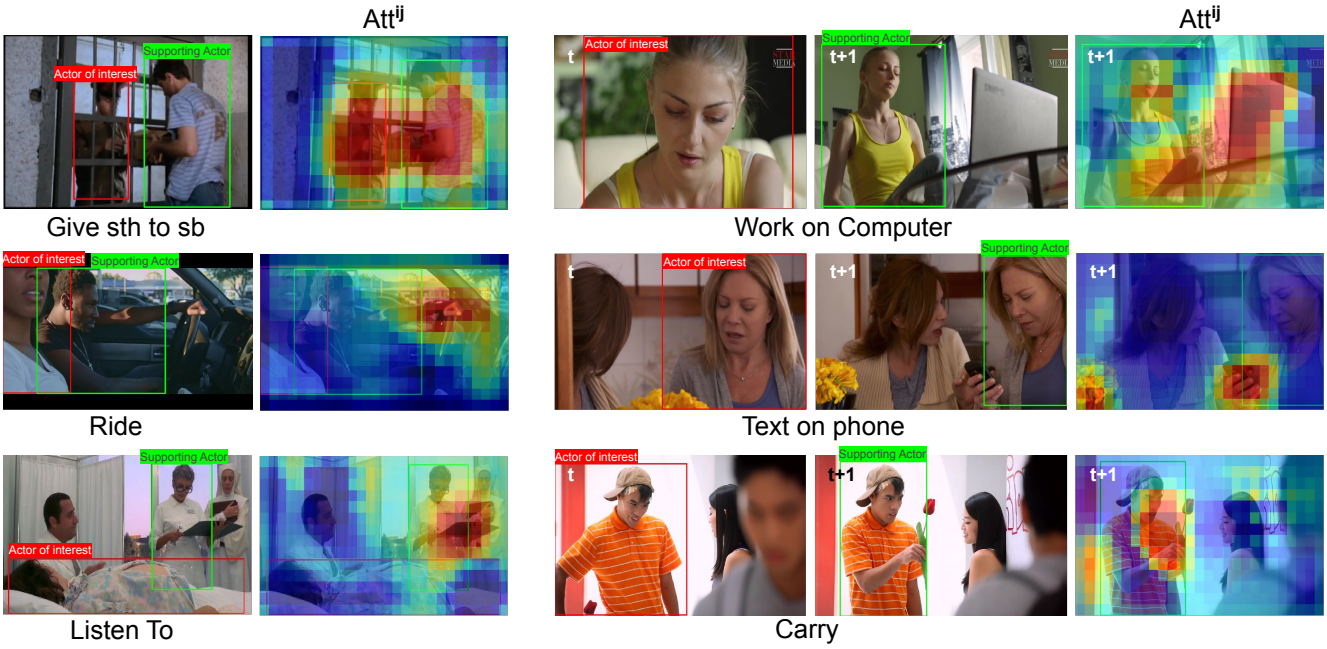

Figure 6. **Visualization of actor-context-actor attention maps on AVA.** Actors of interest are marked in red and supporting actors in green. Heat maps illustrate the context regions' attention weights $Att^{i,j}$ from actor-context-actor relation reasoning. We observe that our model has learned to attend to useful relations between actors and context, and the context serves as the bridge for connecting actors.

5. Experiments on UCF101-24

UCF101-24 is a subset of UCF101 [34] that contains spatio-temporal annotations for 3,207 videos on 24 action classes. Following the evaluation settings of previous methods [20, 54]. We experiment on the first split and report frame-mAP with an IoU threshold of 0.5.

Implementation Details. We also use SlowFast R-50 pretrained on Kinetics-400 as the video backbone, and adopt the person detector from [21]. The temporal sampling for the slow pathway is changed to 8×4 and the fast pathway takes as input 32 continuous frames.

For training, we train all the models end-to-end for 5.4k iterations with a base learning rate of 0.002, which is then decreased by a factor of 10 at iterations 4.9k and 5.1k. We perform linear warm-up during the first quarter of the training schedule. We only use ground-truth boxes for training, and use all boxes given by the detector for inference. Other hyper-parameters are similar to the experiments on AVA.

Results. As shown in Table 5, ACAR surpasses the strong baseline with a considerable margin, which again indicates the importance of high-order relation reasoning.

model	inputs	mAP
T-CNN [17]	V	67.3
ACT [20]	V	69.5
STEP, I3D [54]	V+F	75.0
I3D [15]	V+F	76.3
Zhang *et al.* [58], I3D	V	77.9
S3D-G [52]	V+F	78.8
AIA, R-50 [38]	V	78.8
SlowFast R-50, 8×4 (ours)	V	82.4
Ours w/o ACFB, R50, 8×4	V	**84.3**

Table 5. **Comparison with previous works on UCF101-24.** We evaluate frame-mAP on split 1. V and F refer to visual frames and optical flow respectively.

6. Conclusion

Given the high complexity of realistic scenes encountered in the spatio-temporal action localization task which involve multiple actors and a large variety of contextual objects, we observe the demand for a more sophisticated form of relation reasoning than current ones which often miss important hints for recognizing actions. Therefore, we propose Actor-Context-Actor Relation Network for explicitly modeling higher-order relations between actors based on their interactions with the context. Extensive experiments on the action detection task show our ACAR-Net outperforms existing methods that leverage relation reasoning, and achieves state-of-the-art results on several challenging benchmarks of spatio-temporal action localization.

Acknowledgements. We thank Charlie W., Jiajun T. and Daixin W. for helpful discussions. This work is supported in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants (Nos. 14208417, 14207319, 14202217, 14203118, 14208619), in part by Research Impact Fund Grant No. R5001-18, in part by CUHK Strategic Fund.

References

[1] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

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

[3] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] Siyu Chen, Junting Pan, Guanglu Song, Manyuan Zhang, Hao Shao, Ziyi Lin, Jing Shao, Hongsheng Li, and Yu Liu. 1st place solution for ava-kinetics crossover in acitivitynet challenge 2020. *arXiv preprint arXiv:2006.09116*, 2020.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[7] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

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

[9] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[10] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020.

[11] Rohit Girdhar, João Carreira, Carl Doersch, and Andrew Zisserman. A better baseline for ava. *arXiv preprint arXiv:1807.10066*, 2018.

[12] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.

[13] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron, 2018.

[14] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[15] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[17] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5822–5831, 2017.

[18] Stephen S Intille and Aaron F Bobick. A framework for recognizing multi-agent action from visual evidence. 1999.

[19] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020.

[20] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4405–4413, 2017.

[21] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*, 2019.

[22] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020.

[23] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 303–318, 2018.

[24] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7083–7093, 2019.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[27] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 1049–1059, 2020.

[28] Vlad I Morariu and Larry S Davis. Multi-agent event recognition in structured scenarios. In *CVPR 2011*, pages 3289–3296. IEEE, 2011.

[29] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream r-cnn for action detection. In *European conference on computer vision*, pages 744–759. Springer, 2016.

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[31] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017.

[32] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058, 2016.

[33] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[34] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[35] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 318–334, 2018.

[36] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Rahul Sukthankar, Kevin Murphy, and Cordelia Schmid. Relational action forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 273–283, 2019.

[37] Eran Swears, Anthony Hoogs, Qiang Ji, and Kim Boyer. Complex activity recognition using granger constrained dbn (gcdbn) in sports and surveillance video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 788–795, 2014.

[38] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. *arXiv preprint arXiv:2004.07485*, 2020.

[39] Matteo Tomei, Lorenzo Baraldi, Simone Calderara, Simone Bronzin, and Rita Cucchiara. Stage: Spatio-temporal attention on graph entities for video action detection. *arXiv preprint arXiv:1912.04316*, 2019.

[40] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[41] Oytun Ulutan, Swati Rallapalli, Mudhakar Srivatsa, Carlos Torres, and BS Manjunath. Actor conditioned attention maps for video action detection. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 527–536, 2020.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[43] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[45] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 399–417, 2018.

[46] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.

[47] Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, and Gangshan Wu. Context-aware rcnn: A baseline for action detection in videos. *arXiv preprint arXiv:2007.09861*, 2020.

[48] Jin Xia, Wei Li, Jie Shao, Zehuan Yuan, Jiajun Tang, Cewu Lu, and Changhu Wang. Multiple attempts for ava-kinetics challenge 2020 https://static.googleusercontent.com/media/research.google.com/es//ava/2020/ByteDance-SJTU_AVA_report_2020.pdf, 2020.

[49] Jin Xia, Jiajun Tang, and Cewu Lu. Three branches: Detecting actions with richer features. *arXiv preprint arXiv:1908.04519*, 2019.

[50] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[51] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[52] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018.

[53] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, pages 5783–5792, 2017.

[54] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz. Step: Spatio-temporal progressive learning for video action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2019.

[55] Yuancheng Ye, Xiaodong Yang, and Yingli Tian. Discovering spatio-temporal action tubes. *Journal of Visual Communication and Image Representation*, 58:515–524, 2019.

[56] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[57] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018.

[58] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. A structured model for action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9975–9984, 2019.

[59] Yongmian Zhang, Yifan Zhang, Eran Swears, Natalia Larios, Ziheng Wang, and Qiang Ji. Modeling temporal interactions with interval temporal bayesian networks for complex activity recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2468–2483, 2013.

[60] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017.

[61] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. *arXiv preprint arXiv:1711.08496*, 2017.

[62] Xiantan Zhu, Xuan Tao, Lu Shi, Shaoqi Chen, Rui Yin, Lan Ding, Yuya Obinata, Takuma Yamamoto, and Zhiming Tan. Multi-scale spatiotemporal features for action localization. https://static.googleusercontent.com/media/research.google.com/es//ava/2020/Multi-scale_Spatiotemporal_Features_for_Action_Localization.pdf, 2020.