Actor-Multi-Scale Context Bidirectional Higher Order Interactive Relation Network for Spatial-Temporal Action Localization

Jun Yu[1,2,3] , **Yingshuai Zheng**[3] , **Shulan Ruan**[1,2] , **Qi Liu**[1,2,*] , **Zhiyuan Cheng**[3] , **Jinze Wu**[3]

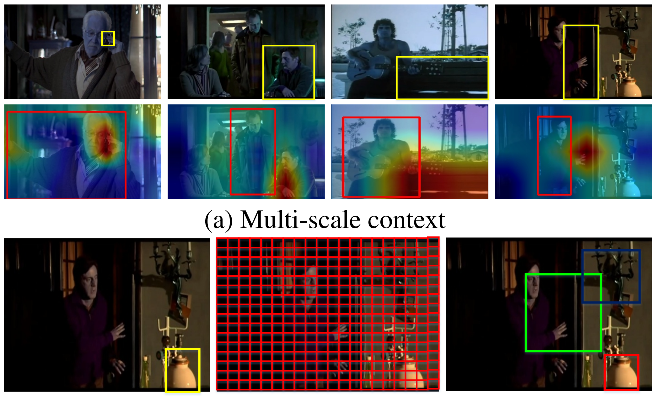
[1]Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China
[2]State Key Laboratory of Cognitive Intelligence, Hefei, China
[3]iFLYTEK Co., Ltd., Hefei, China
{junyu, yszheng2, zycheng14}@iflytek.com, {slruan, hxwjz}@mail.ustc.edu.cn, qiliuql@ustc.edu.cn

Abstract

The key to video action detection lies in the understanding of interaction between persons and background objects in a video. Current methods usually employ object detectors to extract objects directly or use grid features to represent objects in the environment, which underestimate the great potential of multi-scale context information (e.g., objects and scenes of different sizes). How to exactly represent the multi-scale context and make full utilization of it still remains an unresolved challenge for spatial-temporal action localization. In this paper, we propose a novel Actor-Multi-Scale Context Bidirectional Higher Order Interactive Relation Network (AMCRNet) that extracts multi-scale context through multiple pooling layers with different sizes. Specifically, we develop an Interactive Relation Extraction Module to model the higher-order relation between the target person and the context (e.g., other persons and objects). Along this line, we further propose a History Feature Bank and Interaction Module to achieve better performance by modeling such relation across continuing video clips. Extensive experimental results on AVA2.2 and UCF101-24 demonstrate the superiority and rationality of our proposed AMCRNet.

1 Introduction

Video action detection needs to locate each actor in a video and classify its action. It not only requires spatial-temporal features of the actor across the context before and after the keyframe, but also needs to make predictions based on other people, objects and scenes in the background. As Figure 1 (a) shows, the objects that the actor interacted with in different actions have different scales. Thus, comprehensive context including other persons and objects of different scales in the background as well as the background scene should be considered when detecting a person's action. We call the background information the multi-scale context, which was extracted by multiple pooling layers of different scales. The heat map visualization in Figure 1 (a) demonstrates that our AMCRNet can recognize the interaction between the subject and cellphone, person, bench and door in the background accurately, where each background objects have different sizes.



(a) Multi-scale context

(b) Background information extraction method

Figure 1: Visualization of multi-scale context and comparison of different background extraction methods. In (a), the top row represents the background object interacting with the subject person. The heat map in the bottom row shows the sensitivity of interaction with the subject person. In (b), we compare background information extracted by AIA (left), ACAR (middle) and our AMCRNet (right).

There are two methods to extract background objects, but both have the potential of missing background information. Figure 1 (b) demonstrates the output using a detector, grid feature and our method from left to right. The result of AIA [Tang *et al.*, 2020] represents the methods with a detector. Although it can accurately detect background information, it can only detect vases, while missing the wall lamp and the door due to the limitation of classification labels during training. ACAR [Pan *et al.*, 2021] can extract background object information from grid feature, but the area covered by the grid is very small, which can neither represent large objects such as door or scene information, nor small or medium objects such as vase and wall lamp. Thus, in order to extract background objects with different sizes, AMCRNet uses various-sized pooling layers to extract different-sized background objects. For example, a 3 × 3 pooling layer can extract the vase, a 5 × 5 pooling layer can extract the wall lamp, and a 7 × 7 pooling layer can extract the door.

[*] Corresponding Author.

Despite the outstanding progress of recent interactive relation-based models, most of them have two limitations: (1) Background information leakage. [Sun *et al.*, 2018; Pan *et al.*, 2021; Girdhar *et al.*, 2019; Zhao *et al.*, 2022; Feng *et al.*, 2021] use various methods to model the interactive relation between actor feature and grid feature to generate actor-context relation. Due to the limited coverage area of single grid feature, a large amount of background information is lost, especially the information on medium and large-scale objects and scenes. LFB[Wu *et al.*, 2019] does not consider background information. AIA relays on a pre-trained object detector to discover background objects as context. Since spatial-temporal action localization datasets generally do not provide bounding-box annotations of objects, the pre-trained object detector may easily miss various background objects. (2) Failing to construct complete information about an actor in a long video. Both LFB and AIA directly model the current actor feature and the history actor feature, ignoring the interaction information between actor and the context. ACAR [Pan *et al.*, 2021] misses the interaction between person in the current video clip. TubeR misses movement information in history video clips. We consider both movement information of actor and actor-context interaction in current and history video clips important while detecting current action.

To solve the above problems, we propose our action detection method based on extracting a higher-order interactive relation between actor and multi-scale context. The two major components are 1) Interactive Relation Extraction Module; and 2) History Feature Bank and Interaction Module, which are responsible for modeling the interaction between actor and multi-scale context in the current video clip and the interactive relationship between actors in current video clip and actors in historical video clip respectively. Bidirectional Higher Order Interactive Relation Extraction (BHOI) Module is the core of Interactive Relation Extraction Module and is composed of multiple multi-head self-attention (MHSA) layers, which can model the interaction between actor and multi-scale context, and provide enough information for action detection. In order to extract comprehensive background information, multi-scale context represents persons, objects and scenes extracted by pooling layers of various scales. Therefore, the performance of our method would not be limited by object detector and our method can be easily migrated to other scenarios. History Feature Bank and Interaction Module are designed for storing interactive feature from BHOI and movement feature from ROIAlign [Ren *et al.*, 2015] as same as [Gu *et al.*, 2018]. Thus, we can model the interaction between features from viewed videos and current features to achieve long-term interaction for each actor.

We experiment with our method on AVA2.2 and UCF101-24 [Soomro *et al.*, 2012]. Compared to the baseline, our method increases the mAP by 4.0%. In addition, the visualization for the case study demonstrates that our proposed method could pay more attention to useful context.

The main contributions can be summarized as follows:

- We observe the great potential of multi-scale context information, and propose a novel AMCRNet to represent it for better spatial-temporal action localization.
- We develop an Interactive Relation Extraction Module to model the higher-order relation between the target person and the context.
- We design a History Feature Bank and Interaction Module to model the higher-order relation across continuing video clips.
- Extensive experiments demonstrate the superiority and rationality of our proposed method.

2 Related Work

2.1 Action Classification

With the successful accomplishments of deep learning, deep networks have achieved impressive performance in various computer vision tasks, such as video classification [Wang *et al.*, 2020], sentiment analysis [Ruan *et al.*, 2021a] and image synthesis [Ruan *et al.*, 2021b]. Action Classification algorithms try to extract video features containing motion information. Earlier algorithms [Lin *et al.*, 2018; Karpathy *et al.*, 2014; Donahue *et al.*, 2015; Yue-Hei Ng *et al.*, 2015; Shi *et al.*, 2015] used 2D classification network to extract semantic features from each frame, then merged image features to video features by clustering, pooling or recurrent neural network (RNN). [Simonyan and Zisserman, 2014; Feichtenhofer *et al.*, 2016] adopted a two-stream architecture process spatial semantic and temporal optical flow separately, then fused them together as the overall video feature to get a decent result. [Zolfaghari *et al.*, 2018; Tran *et al.*, 2019; Feichtenhofer, 2020; Feichtenhofer *et al.*, 2019] 3D convolution layer takes advantage of temporal receptive field, which helps the model learn from motions. It performs well under replicating weights of 2D classification networks as well as pretraining on large-scale video datasets. [Yan *et al.*, 2022; Bertasius *et al.*, 2021; Liu *et al.*, 2022; Fan *et al.*, 2021] could learn from long-term dependencies, and help model the overall temporal feature, which makes transformer-based methods take the lead comparing to 3D convolution methods when classifying an action video.

2.2 Action Detection

Earlier methods [Gu *et al.*, 2018; Girdhar *et al.*, 2018] are directly based on Faster RCNN [Ren *et al.*, 2015] structure, which utilized ROIAlign to get the temporal feature for each person. Since the actual perception field is rather limited, actor features extracted from ROIAlign might miss the perception and connection to persons and objects in the background. ACRN [Schumann and Stiefelhagen, 2017] suggested concatenating actor feature and the whole video feature together to generate the relationship between actor and each background object, while Video Action Transformer leveraged cross attention to generate the relation. [Zhang *et al.*, 2019] proposed to model actor-actor relation and actor-context relation separately using two independent networks. [Pan *et al.*, 2021; Feng *et al.*, 2021] constructed one order actor-context interaction first as in ACRN. Then adopted Non-Local [Wang *et al.*, 2018] or transformer [Vaswani *et al.*, 2017] to form the higher order actor-context-actor relation. AIA piled up multiple cross attention block to model

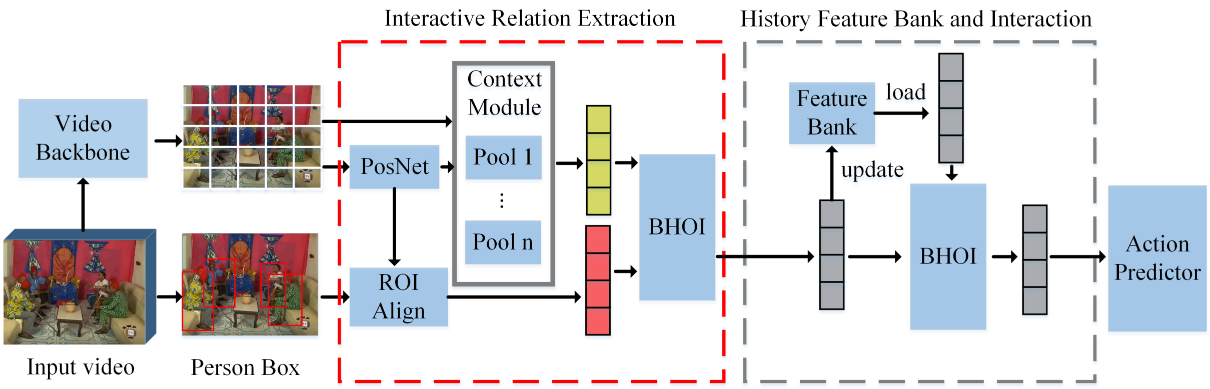


Figure 2: Overall framework of AMCRNet.

actor-actor, actor-context and actor-history actor interactions. ORBNet [Herzig *et al.*, 2022] constructed actor-context interaction on each multi-head attention block and outperforms the baseline model. [Arnab *et al.*, 2021; Li *et al.*, 2022; Wu *et al.*, 2022] used the transformer structure to extract video features, and automatically generated the high-order relationship between actor and background context, since each position has the global perception of the spatial and temporal dimension. Self-supervised learning-based methods such as [Wei *et al.*, 2022; Tong *et al.*, 2022] achieved the best result. TuBeR utilized tublet query regression to get actor position and form accurate cuboids, then further extracted motion information for each person based on cuboids. LFB proposed to store actor features to gain support from a longer context. The non-local mechanism was used to model the relation between persons in the current video clip and persons in the past video clips. [Zhao *et al.*, 2022] directly adopted cross-attention layer to do computation on the background features in the current video clip and past clips to fertilize the current background feature. Memvit [Wu *et al.*, 2022] stored the key and value matrix in the attention layer, so the perception of history information could be reached.

3 Method

In this section, we provide detailed descriptions of our proposed AMCRNet. AMCRNet aims at effectively modeling bidirectional higher-order relations between actor and multi-scale context for achieving more accurate action localization.

3.1 Overall Structure

Figure 2 demonstrates the overall structure of AMCRNet. Similar to the most advanced action detection algorithm, AMCRNet is built on top of an actor detection network and a video feature extraction network. Interactive Relation Extraction Module is the core component, which establishes interaction between actor and context in longer videos to provide rich information that can be used to support action detection.

A uniformly sampled frame sequence within 2 seconds range of a keyframe from a video clip is used as the input. The frame sequence($input\_frames$) is sent through a video feature extraction network (BackboneNet) to extract spatial and temporal video feature, the equation follows:

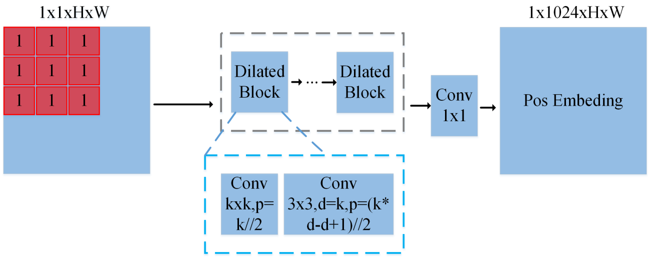


Figure 3: Architecture of PosNet.

$$\boldsymbol{X\_slow}, \boldsymbol{X\_fast} = BackboneNet(input\_frames), \quad (1)$$

where $X\_slow \in R^{C \times T\_1 \times H \times W}$ and $X\_fast \in R^{C \times T\_2 \times H \times W}$ denote the video features.

The dimension of a video feature is $\mathrm{C} \times \mathrm{T} \times \mathrm{H} \times \mathrm{W}$, which represents channel, time, height and width respectively. Then we convert X_slow,X_fast to $X^1 \in R^{\mathrm{C} \times \mathrm{H} \times \mathrm{W}}$,

Meanwhile, the person detector would detect N person on the keyframe. While extracting movement information, we use ROIAlign to get a fixed-sized (8×8) actor feature from the video feature and then finalize the actor feature $A \in \mathrm{R}^C$ by average pooling on space. As Figure 2 shows, the dotted red box represents Interactive Relation Extraction Module, which takes video feature X^1 and actor feature A as the input, and outputs an action label for each person by extracting the relation between actor and multi-scale context.

3.2 Interactive Relation Extraction Module

Interactive Relation Extraction Module is used to model the interaction between actor and multi-scale contexts such as other persons, objects and scenes in the background, so it can provide global information that action prediction needs. As shown in Figure 2, it has three components: Position Embedding Network (PosNet), Multi-Scale Context Module (Context Module) and Bidirectional Higher Order Interactive Interaction Module (BHOI Submodule).

PosNet. [Islam *et al.*, 2020] exposed that padding operation during convolution helps the network learn positional information. Transformer Network such as [Chu *et al.*, 2021;

Yuan *et al.*, 2021] makes use of this to learn the hidden positional embedding. However, we discover that a single convolution layer cannot form a global perception field nor transfer padding signal to all positions due to its limited perception field, which generates inaccurate positional embedding. [Islam *et al.*, 2020] also proved that later modules and deeper layers in a network predict more accurate positional embedding. Thus, we construct a lightweight positional embedding network in which the perception field covers input size by stacking multiple narrow (channel size of 2) convolution layers. As shown in Figure 3, PosNet consists of multiple dilated blocks that increase perception filed size. When generating positional embedding, we first randomly initialize a tensor $Input\_pos \in \mathrm{R}^{1\times1\times\mathrm{H}\times\mathrm{W}}$ which has the same scale as the video feature, then pass it to PosNet for the positional embedding of the whole video feature:

$$Pos = \mathrm{PosNet}(Input\_pos), \tag{2}$$

where $Pos \in R^{1\times C\times H\times W}$ denotes the positional embedding of video feature.

The theoretical perception field size of the output layer is greater than the size of the input tensor, so the embedding of each position can represent the relevant position of video feature. ROIAlign uses average pooling ($AvgPooling$) to extract positional embedding for actor based on its detection bound box (Box) by the following equation:

$$Pos\_person = \mathrm{AvgPooling}(\mathrm{ROIAlign}(Pos, Box)), \tag{3}$$

where $Pos\_person \in R^C$ is the positional embedding for the actor.

Context Module. We consider that action detection need support from multi-scale context, such as background objects of different sizes or background scene.Therefore, we construct a multi-scale context extraction module to extract multi-scale context and corresponding positional embedding. As Figure 2 shows, Context Module is built by a series of self-adapted pooling layer or convolution layer of different sizes running in parallel. By inputting video feature and positional embedding feature into Context Module, multi-scale context feature and corresponding position embedding can be obtained as follows:

$$Context, Pos\_context = \mathrm{CM}(X^1, Pos). \tag{4}$$

where $Context \in \mathrm{R}^C, Pos\_context \in \mathrm{R}^C$ denote the multi-scale context feature and the corresponding position embedding, respectively. Here, CM means Contex Module. The sum of context feature and positional embedding is used as the input of BHIO Module.

BHOI Module. BHOI Module is formed by stacking multiple multi-head self-attention (MHSA) layers. We first construct the input feature by adding the actor feature and context feature with their corresponding positional embedding: $I = \{\{A_i^1 + Pos\_person_i\}_{i=1}^{N}, \{Context_j + Pos\_context_j\}_{j=1}^{M}\}$, while N is the number of actors and M is the number of multi-scale contexts. BHOI Submodule is used to compute the self-attention $\{Attn_{i,j}\}_{j=1}^{N+M}$of each feature, while Q_i, K_i, V_i is generated by passing the input feature sequence $\{I_i\}_{i=1}^{N+M}$ through a 1×1 convolution layer, and the output H_i of each position is calculated by the sum of V_j linearly weighted by $Attn_{i,j}$. Thus, the overall bidirectional interaction is constructed with in person, object and scene, shown by the following equations:

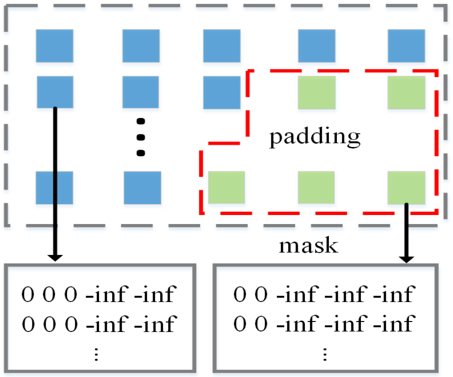


Figure 4: Computation graph of attention mask. By adding masks and attention, the disruption of padded elements can get removed.

$$\begin{aligned} Q_i, K_i, V_i &= \mathrm{conv2d}(I_i), \\ Attn_{i,j} &= \mathrm{softmax}(\frac{\langle Q_i, K_i \rangle}{\sqrt{C}}), \\ H_i &= \sum_{j}^{N+M} Attn_{i,j}. \end{aligned} \tag{5}$$

In order to process multiple videos concurrently, we add a mask in the multi-head attention. As shown in Figure 4, since the number of actors is different in each video, the length of input feature sequence within the same batch for BHOI Submodule is also different. Thus, all input sequences in the same batch need to be padded to same length with a mask and add an infinite constant to the attention correspondingly.

By repeating the above computational process, we managed to construct different levels of bidirectional interaction between actor features and multi-scale context.

3.3 History Feature Bank and Interaction Module

For the purpose of supporting the modeling of interaction between actor and multi-scale context on a longer video clip, we use a History Feature Bank to store higher-order interactive relation feature and original movement feature for actors in each video clip, then construct the interaction between actor and context and the movement information of actor in parallel during the inference over longer video clip.

History Feature Bank. When training the second stage AMCRNet to store the history feature, we used two methods to set up the history feature bank. The first method is similar to ACAR, which constructs a history feature bank offline. Then train a full AMCRNet with a history feature bank, the actor's high-order interactive relation features and movement features of 2W (W=7) video clips were extracted from the feature bank interact with the actor's current high-order interactive relation features and movement features calculated by AMCRNet respectively. The second method is similar to

Model	Backbone	Pre-train	Params(M)	mAP (%)
Slowfast R50 [Feichtenhofer *et al.*, 2019]	SlowFast R50 8 × 8	Kinetics-400	35	24.8
Mvit-B [Wu *et al.*, 2022]	Mvit-B	Kinetics-400	53.21	27.5
AIA R50 [Tang *et al.*, 2020]	SlowFast R50 4 × 16 + NL	Kinetics-700	75.8	29.8
ACAR R50 [Pan *et al.*, 2021]	SlowFast R50 8 × 8	Kinetics-400	NA	28.84
ORViT [Herzig *et al.*, 2022]	MViT-B 32 × 3	Kinetics-400	NA	28.0
TubeR [Zhao *et al.*, 2022]	CSN-50	IG + Kinetics-400	NA	29.2
AMCRNet-slim R50(Ours)	SlowFast R50 4 × 16	Kinetics-400	73.4	28.77
AMCRNet R50(Ours)	SlowFast R50 4 × 16	Kinetics-400	149	29.17
AMCRNet-slim R50(Ours)	SlowFast R50 8 × 8	Kinetics-400	73.2	29.85
AMCRNet R50(Ours)	**SlowFast R50 8 × 8**	**Kinetics-400**	**148**	**30.23**
Slowfast R101 [Feichtenhofer *et al.*, 2019]	SlowFast R50 8 × 8	Kinetics-400	65	29.8
AIA R101 [Tang *et al.*, 2020]	SlowFast R101 8 × 8 + NL	Kinetics-700	104.1	32.26
ACAR R101 [Pan *et al.*, 2021]	SlowFast R101 8 × 8	Kinetics-700	NA	33.3
TubeR [Zhao *et al.*, 2022]	CSN-152	IG + Kinetics-400	NA	33.6
AMCRNet-slim R101(Ours)	**SlowFast R101 8 × 8**	**Kinetics-700**	**101.7**	**34.2**

Table 1: Comparing to advanced models on AVA2.2.

AIA, which is constructed online. During training, a zero vector would be used to represent the unprocessed video. The latest features will be updated to the history feature bank if it has a lower loss. Although the ways that offline history feature bank and online history feature bank have been constructed are different, the interaction between the history feature and current feature is the same. Here, we denote the feature bank that stores both the high-order interaction feature and movement feature as High Relation-Action Feature Bank (HR-AFB), and the one only stores the high-order interaction feature as High Relation Feature Bank (HRFB).

History Feature Interaction. We first take the sum of high-order interactive feature from current video clip (H), features from history feature bank (F), and temporal positional embedding of video clip ($temporal\_pos$) as the input sequence send into a BHOI Submodule to get the high-order interactive feature (O) of long video. The equation follows:

$$\begin{aligned} Q_i, K_i, V_i &= \mathrm{conv2d}(\{\{H, F\} + temporal\_pos\}_i), \\ Attn_{i,j} &= \mathrm{softmax}(\frac{\langle Q_i, K_i \rangle}{\sqrt{C}} + HMask(i,j)), \\ O_i &= \sum_{j}^{N+M} Attn_{i,j}. \end{aligned} \tag{6}$$

Similar to Mask in Figure 4, $HMask$ here is used to mark the padding of the feature sequence. The movement feature is computed in a similar fashion, where the current movement feature (A^1) and history movement feature (M) are used to get the final movement feature (P) for the long video.

4 Experiments

AVA is a video dataset of spatial-temporally localized atomic visual actions, which contains 430 video clips from movies, with 235 training clips, 64 validation clips and 131 test clips. Each video clip is within 15 to 30 minutes and labels were added on each keyframe, which is extracted at 1 frame per second. Those labels include person bounding box and 80 action labels cover action, person-person interaction and person-object interaction. Since our method is designed for action detection, we use the AVA2.2 dataset for ablation experiments and performance comparison. During testing, we use an IOU threshold greater than or equal to 0.5 to calculate mAP on 60 action labels as the final performance metric. Our codes are publicly available online. [*].

4.1 Implementation Details

Actor Detector. For the actor detector, we use Faster RCNN with the same configuration following ACAR-Net, which is pre-trained on the COCO dataset and fine-tuned on the AVA dataset. For a fair comparison, we do not use the latest detector that has better performance.

Backbone Network. SlowFast network is used as the backbone for feature extraction, and we use the version pre-trained on Kinetics-400 in our model. SlowFast takes two video sequences of different lengths as the input, and it can effectively learn temporal information and rich context information.

Training. During training, we use the pre-trained SlowFast to initialize the backbone, then fine-tune the model on AVA. For the input image sequence, we scale the shorter side of input frames to 256 pixels. The training hardware is 2 Tesla V100, and the training is done under the batch size of 8 (4 sequences on each one) with SGD optimizer. Since the batch size is small, we freeze the batch normalization layer to ensure stability. The base learning rate is 0.02, and we totally train the model for 10 epochs, while reducing it by 10 times on the 7th and 9th epoch. To accelerate convergence, we established a warm-up schedule [Yue-Hei Ng *et al.*, 2015] for the first 3 epochs.

4.2 Comparison on AVA

We compare the performance of AMCRNet against advanced action detection models on the AVA2.2 dataset. For a comprehensive comparison, we experiment on various versions of backbone such as SlowFastR50 and SlowFastR101. As Table 1 shows, our method outperforms advanced detection models including ACAR-Net, AIA and TubeR. With the support

[*] https://github.com/manzhihuangnian/AMCRNet

Model	mAP (%)
Baseline	20.6
Baseline+ACRN	21.9
Baseline+AIA	22.77
Baseline+HR2O	23.46
Baseline+BHOI	23.88

(a) Relational modeling methods

Depth	mAP (%)
1	21.13
2	22.45
4	22.92
6	23.16
8	23.02

(b) Depth of BHOI module

Context module	mAP (%)
w / o pooling layer	23.16
kernel_size=3	23.62
kernel_size=5	23.65
kernel_size=7	23.46
kernel_size=3,5,7	23.88

(c) Depth of BHOI module

Spatial pos	mAP (%)
w / o pos	23.61
learned fix_pos	23.32
sin_pos	23.75
posnet(ours)	23.88

(d) Positional embedding

Feature bank	mAP (%)
LFB	24.21
ACFB	24.72
HRFB	25.64
HR-PFB	25.85

(e) Feature bank

Training method	mAP (%)
Offline	25.62
Online	25.64

(f) Training method of feature bank

Table 2: Ablation study on AVA2.2 dataset.

Model	Inputs	mAP (%)
AIA [Tang *et al.*, 2020]	V	78.8
ACAR [Pan *et al.*, 2021]	V	84.3
HIT [Faure *et al.*, 2023]	V+P	84.8
Ours	**V**	**84.9**

Table 3: Comparing to advanced models on UCF101-24.

of 15s video clips before and after the video split, we achieve 34.2 % mAP, which proves that our modeling method with the BHOI module can extract spatial and temporal context that action detection needed effectively. Here, AMCRNet-slim is a slim version that reduces the number of stacked MHSA from 6 to 4 in the BHOI Submodule and reduces the channel dilation multiplier on MLP from 4 to 1 in the MHSA module.

4.3 Comparison on UCF101-24

UCF101-24 is another action detection dataset with 3207 videos labeled with 24 different action types. We experiment on the first split and report frame-level mAP with an IoU threshold of 0.5. Similar to ACAR, we also use SlowFast R50 pre-trained on Kinetics-400 as our backbone, and the actor detector is the one from [Pan *et al.*, 2021]. During training, we freeze all batch normalization layers of the backbone. We train the models for 10 epochs with a base learning rate of 0.006, while decreasing it by 10 times at epoch 7 and 9. Also, to accelerate converges, we deploy a linear warm-up schedule for the first 3 epochs. As Table 3 shows, AMCRNet outperforms advanced methods such as AIA or ACAR, HIT [Faure *et al.*, 2023] which proves that the bidirectional higher-order interaction is effective when extracting interaction between actor and background environment. Here, V refers to visual frames and P refers to the pos extracted by pos estimation network.

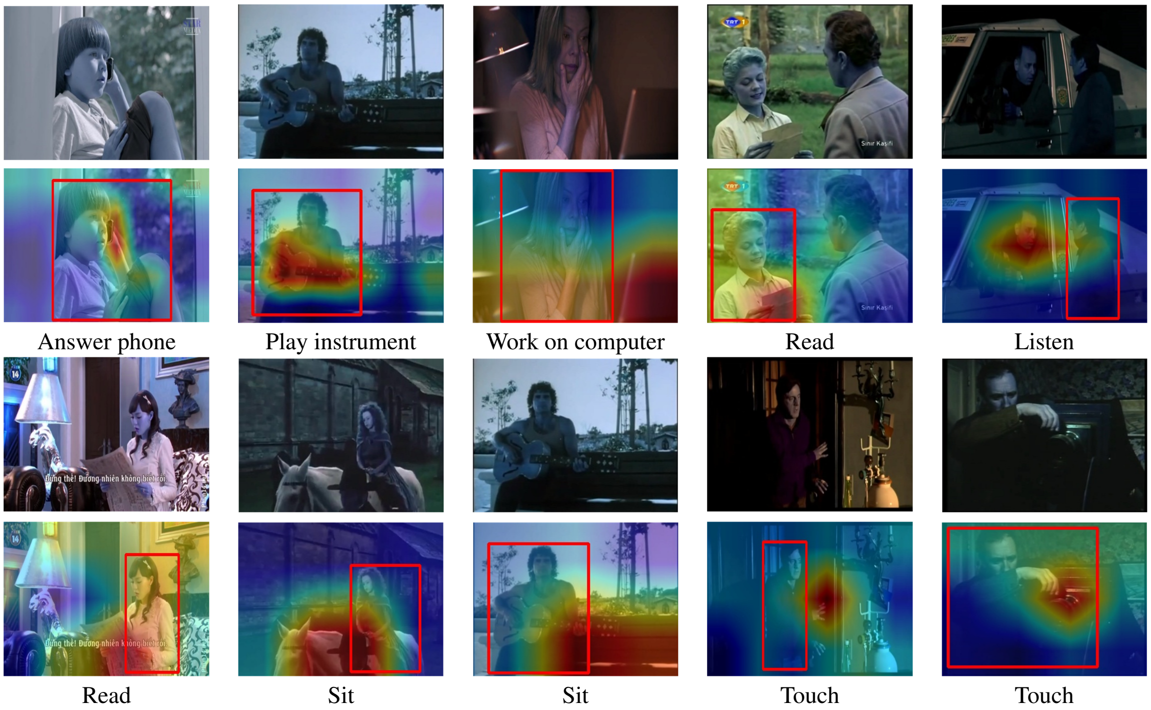
4.4 Ablation Experiments on AVA2.2

To explore the effectiveness of each module of AMCRNet, we perform ablation experiments on AVA2.2. Due to the limitation of training resources, SlowOnly R50 4×16 pre-trained on Kinetics-400 is used as the backbone. The baseline model only includes a backbone, a detector and a one-layer classifier. ACFB is the Actor-Context Feature Bank in ACAR-Net, whereas HR2O is the High-order Relation Reasoning Operator in ACAR-Net.

Compare relational modeling methods. To prove our method that model the bidirectional higher-order relation is effective, we compare against several previous approaches. Thus, the history feature bank is not included in this specific experiment. In the comparing methods, ACRN-Net focuses on learning the interaction between actor and background, AIA extracts actor-actor and actor-context interaction in serial, and ACAR-Net constructs actor-context-actor interaction using the background information at the same position, while our method proposes BHOI. BHOI can directly generate the bidirectional higher-order interaction between persons, objects, and background scenes. From Table 2 (a), we notice that model using BHOI performs the best. In addition, AIA outperforms ACRN-Net since it models multiple interactions. Since ACAR-Net includes all background information, it performs better than AIA.BHOI outperforms ACAR-Net with the support of bidirectional higher-order interaction and multi-scale context.

Compare depth of BHOI module. In Table 2 (b), we compare the performance change with a different number of stacks of MHSA module. We notice that, with more MHSA modules stacked together, the overall performance increase. While stacking 8 MHSA modules, the accuracy decrease presumably because of over-fitting. Thus, the default number of stacks of MHSA is 6.

Compare pooling layer of context module. In Table 2 (c), we compare the performance on different scales of pooling layer that is used to extract context information. We discover that adding a pooling layer would boost the performance regardless of kernel size according to table 1c. Although using all scales performs the best, to avoid increasing computational cost too much as a smaller pooling size result in a longer sequence, the default kernel size of the pooling layer



Answer phone, Play instrument, Work on computer, Read, Listen

Read, Sit, Sit, Touch, Touch

Figure 5: Visualization of correlation between actor and the multi-scale context in the AVA2.2 test set. The red bound box marks the actor, and heat maps represent the level of correlation between the actor and the multi-scale background context.

is 3, 5, and 7. Without using kernel size 1, we reduce the computation cost of multi-head attention by 6 times.

Compare positional embedding. Since the height and width of different video varies, the scale of video feature also varies. Traditional methods that interpolate fixed-sized positional embedding to different scales would decrease the performance. We compared it to our positional embedding method, and Table 2 (d) shows that our method can achieve higher accuracy, even exceeding the sin_pos.

Compare feature bank. We use 2 × W + 1 consecutive video clips before and after the current one to achieve a 15s perception field when W is set to 7. We compare our history feature bank with other feature bank methods. From Table 2 (e), the Baseline gains a small boost with the support of LFB, as it uses movement information in the feature bank. After replacing LFB with ACFB, the one stores interaction between actor and background, the performance increases. A great boost to performance happens when using HRFB, but HR-AFB reaches the best accuracy. It proves that while expanding the perception over the temporal dimension, both higher-order interaction and movement information is effective.

Training method of feature bank. The cost to construct an offline history feature bank is expensive since we need to train twice. Therefore, we compare the performance while using an online history bank or an offline ban. For the offline method, an AMCRNet-lite without a feature bank has to be trained to obtain the history feature, then use the generated bank to train an AMCR-Net. Since two networks are optimized independently, the performance gain of two phrases may not be in a linear relationship, which indicates that a AMCRNet trained with offline history bank is unstable. Online history bank updates feature to the bank while training, which is efficient and achieves end-to-end. From Table 2 (f), we notice that the performance using an online history bank or an offline bank is similar, so the online history bank method is used in later experiments.

4.5 Visualization

Since our proposed method makes use of stacked self-attention in the BHOI module to model the interaction between the actor and multi-scale context, we can show the attention by visualizing its weight. From Figure 5, odd rows show the keyframes whereas even row shows the heat map of attention weight. We discover that the attention is focused on objects of different scales interacting with actors in each video. All small objects such as cellphone and paper slip, median scale objects such as newspaper and laptop as well as large objects such as horse and door can be covered by attention. Thus, we prove that AMCRNet can pay attention to the context of various scale that is related to the actor.

4.6 Discussion

Although AMCR-Net is capable of extracting multi-scale objects in the background, it may not extract objects with irregular shapes effectively. As Figure 5 shows, regular shape objects such as paper slip, laptop or camera box can be extracted through a single pooling layer, whereas only some partial information can be extracted from door and guitar. For objects with unbalanced height and width like guitar, we can only extract its information by applying a 3×3 pooling layer multiple times to avoid introducing useless noise. We will add pooling layers with different scales and different aspect ratios to cover different shaped objects or use a network to predict the approximate range of persons and objects of different sizes and background scene associated with the people.

5 Conclusion

In this paper, we argued that the multi-scale context information is quite helpful for spatial-temporal action localization. To this end, we proposed a novel Actor-Multi-Scale Context Bidirectional Higher Order Interactive Relation Network (AMCRNet) to represent and utilize multi-scale context information in a long video. Specifically, we developed an Interactive Relation Extraction Module to extract bidirectional higher-order relations between target person and other persons and objects in the context. Along this line, to better model such relations and original movement information across multiple continuing video clips, we further proposed a History Feature Bank and Interaction Module. Extensive experimental results demonstrated the superiority and rationality of our proposed AMCRNet. In the future, we will explore to use deformable convolution or transformer for automatically multi-scale context representation.

Acknowledgments

This research was partially supported by grant from the National Natural Science Foundation of China (Grant No. 61922073), and the University Synergy Innovation Program of Anhui Province (GXXT-2021-002).

References

[Arnab *et al.*, 2021] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[Bertasius *et al.*, 2021] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[Chu *et al.*, 2021] Xiangxiang Chu, Bo Zhang, Zhi Tian, Xiaolin Wei, and Huaxia Xia. Do we really need explicit position encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 3(8), 2021.

[Donahue *et al.*, 2015] 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.

[Fan *et al.*, 2021] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.

[Faure *et al.*, 2023] Gueter Josmy Faure, Min-Hung Chen, and Shang-Hong Lai. Holistic interaction transformer network for action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3340–3350, 2023.

[Feichtenhofer *et al.*, 2016] 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.

[Feichtenhofer *et al.*, 2019] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[Feichtenhofer, 2020] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

[Feng *et al.*, 2021] Yutong Feng, Jianwen Jiang, Ziyuan Huang, Zhiwu Qing, Xiang Wang, Shiwei Zhang, Mingqian Tang, and Yue Gao. Relation modeling in spatio-temporal action localization. *arXiv preprint arXiv:2106.08061*, 2021.

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

[Girdhar *et al.*, 2019] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 244–253, 2019.

[Gu *et al.*, 2018] 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.

[Herzig *et al.*, 2022] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3148–3159, 2022.

[Islam *et al.*, 2020] Md Amirul Islam, Sen Jia, and Neil DB Bruce. How much position information do convolutional neural networks encode? *arXiv preprint arXiv:2001.08248*, 2020.

[Karpathy *et al.*, 2014] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[Li *et al.*, 2022] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022.

[Lin *et al.*, 2018] Rongcheng Lin, Jing Xiao, and Jianping Fan. Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[Liu *et al.*, 2022] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[Pan *et al.*, 2021] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 464–474, 2021.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[Ruan *et al.*, 2021a] Shulan Ruan, Kun Zhang, Le Wu, Tong Xu, Qi Liu, and Enhong Chen. Color enhanced cross correlation net for image sentiment analysis. *IEEE Transactions on Multimedia*, 2021.

[Ruan *et al.*, 2021b] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13960–13969, 2021.

[Schumann and Stiefelhagen, 2017] Arne Schumann and Rainer Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–28, 2017.

[Shi *et al.*, 2015] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[Soomro *et al.*, 2012] 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.

[Sun *et al.*, 2018] 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.

[Tang *et al.*, 2020] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *European Conference on Computer Vision*, pages 71–87. Springer, 2020.

[Tong *et al.*, 2022] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

[Tran *et al.*, 2019] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2018] 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.

[Wang *et al.*, 2020] Xin Wang, Wei Huang, Qi Liu, Yu Yin, Zhenya Huang, Le Wu, Jianhui Ma, and Xue Wang. Fine-grained similarity measurement between educational videos and exercises. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 331–339, 2020.

[Wei *et al.*, 2022] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.

[Wu *et al.*, 2019] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.

[Wu *et al.*, 2022] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022.

[Yan *et al.*, 2022] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3333–3343, 2022.

[Yuan *et al.*, 2021] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 579–588, 2021.

[Yue-Hei Ng *et al.*, 2015] 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.

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

[Zhao *et al.*, 2022] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, et al. Tuber: Tubelet transformer for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13598–13607, 2022.

[Zolfaghari *et al.*, 2018] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712, 2018.