CoTeRe-Net: Discovering Collaborative Ternary Relations in Videos

Zhensheng Shi[1][0000−0001−5793−9850], Cheng Guan[1][0000−0001−8176−9556],
Liangjie Cao[1][0000−0001−5573−152X], Qianqian Li[1][0000−0002−3866−8159],
Ju Liang[1][0000−0003−3484−1310], Zhaorui Gu[1][0000−0002−6673−7932],
Haiyong Zheng[1⋆][0000−0002−8027−0734], and Bing Zheng[1,2⋆][0000−0003−2295−3569]

[1] Underwater Vision Lab (ouc.ai), Ocean University of China
[2] Sanya Oceanographic Institution, Ocean University of China
{shizhensheng,guancheng,caoliangjie,liqianqian5957,liangjie8257}@stu.ouc.edu.cn
{guzhaorui,zhenghaiyong,bingzh}@ouc.edu.cn

Abstract. Modeling relations is crucial to understand videos for action and behavior recognition. Current relation models mainly reason about relations of invisibly implicit cues, while important relations of visually explicit cues are rarely considered, and the collaboration between them is usually ignored. In this paper, we propose a novel relation model that discovers relations of both implicit and explicit cues as well as their collaboration in videos. Our model concerns *Co*llaborative *Te*rnary *Re*lations (CoTeRe), where the ternary relation involves channel (C, for implicit), temporal (T, for implicit), and spatial (S, for explicit) relation (R). We devise a flexible and effective CTSR module to collaborate ternary relations for 3D-CNNs, and then construct CoTeRe-Nets for action recognition. Extensive experiments on both ablation study and performance evaluation demonstrate that our CTSR module is significantly effective with approximate 3% gains and our CoTeRe-Nets outperform state-of-the-art approaches on three popular benchmarks. Boosts analysis and relations visualization also validate that relations of both implicit and explicit cues are discovered with efficacy by our method. Our code is available at https://github.com/zhenglab/cotere-net.

Keywords: Video understanding · Action recognition · Relation model

1 Introduction

We carve our world into relations between things [36]. The ability to discover relations between entities and their properties is central to our cognition of the world [20,17]. Consider an action of "*something colliding with something and both come to a halt*", in contrast to the action of "*moving something and something closer to each other*", identifying "*colliding*" and "*halt*" requires to reason about invisibly implicit dependencies and interactions, while recognizing "*moving*" and "*something*" needs to exploit visually explicit temporal motions

[⋆] Corresponding authors

and spatial objects. Thus, we understand these two actions from the videos via fusing these two requirements in our mind, and we argue that they correspond to relations of implicit and explicit cues respectively.

Discovering relations between entities is crucial to understand action and behavior from videos [11,30,9,32]. Existing relation models [63,4,53] for recognizing actions from videos typically discover the relations by reasoning about invisibly implicit temporal or channel cues, like dependencies and interactions. While, many efforts have been devoted to detect visually explicit temporal motions [50,5,51] or spatial objects [47,55], such as optical flow and visual attention, due to their effectiveness to recognize human actions. However, discovering the relations of these visually explicit cues is rarely considered. In addition, the collaboration between relations of implicit and explicit cues is usually ignored.

In this work, going further in modeling relations on the implicit level, we discover relations via leveraging both implicit and explicit cues to represent videos for understanding actions better. Our proposed relation model discovers the collaborative ternary relations in videos, dubbed **CoTeRe**, where the ternary relation involves channel (**C**, for implicit), temporal (**T**, for implicit), and spatial (**S**, for explicit) relation (**R**). Specifically, the channel relations take in charge of reasoning about implicit cues among different perspectives of global information over spatiotemporal scope, and the temporal relations are responsible for reasoning about implicit temporal dependencies between video frames, while the spatial relations are in charge of exploiting explicit spatiotemporal topological information visually. Finally, we collaborate the ternary relation for fusing implicit and explicit cues to better understand actions from videos.

Our **contributions** include: (a) A novel relation model discovering relations of both implicit and explicit cues in videos. (b) A flexible and effective CTSR module to collaborate the ternary relation for 3D-CNNs. (c) CoTeRe-Nets achieving state-of-the-art performance with a significant gain on action recognition especially in densely-labeled and fine-grained situations.

2 Related Works

2.1 Video Representation

Early contributions in video representation have focused on developing hand-designed spatiotemporal features [48,49]. Since the breakthrough of Convolutional Neural Networks (CNNs) [24] for image representation [21,42,14,12,43,41], many works have tried to design effective architectures for spatiotemporal representation in videos [18,51,38,44,2,52]. Karpathy *et al.* [18] first introduced CNN to represent videos. Then, two-stream [38,6] and 3D-CNN [15,44] led two mainstreams of video representation. Two-stream methods mainly used video RGB data and motion features like optical flow to learn representation [60,51,6,28]. C3D [44] devised a 3D convolutional filter and I3D [2] inflated 2D convolutional filters into their 3D counterparts to learn spatiotemporal representation. The recent 3D-CNN methods, such as P3D [33], S3D [57], and R(2+1)D [46], gained superior performance under better video representation by factorizing the 3D convolutional filter into separate spatial and temporal operations.

Some recent works on video representation focused on better leveraging temporal information to improve the performance [52,62,5,50,63,37,3]. TrajectoryNet [62] incorporated trajectory convolution for integrating features along temporal dimension to replace the existing temporal convolution. SlowFast networks [5] proposed a SlowFast concept with a slow pathway and a fast pathway to capture spatial semantics and finer motions respectively.

2.2 Relation Models

Recently, relation models have been adopted in the area of visual question answering [36,25], object detection/recognition [13,8], and intuitive physics [1,54]. In the case of action recognition, a lot of efforts have been made on modeling pairwise human-object and object-object relations[11,58,59,23,35].

The latest works attempted to employ relational structures [36] for video representation and manifested that exploiting spatiotemporal relations is significant for video analysis [50,63,53,4]. ARTNet [50] decoupled spatial and temporal modeling into two parallel branches. TRN [63] was designed to learn and reason about temporal relations between video frames at multiple time scales. Wang *et al.* [53] proposed to represent videos as space-time region graphs connected by similarity relations and spatial-temporal relations. STC [4] modeled correlations between channels of a 3D-CNN with respect to temporal and spatial features.

2.3 Comparison to our approach

Compared to existing relation models for video representation, our approach aims to discover relations of both implicit and explicit cues for channel-temporal-spatial ternary collaboration, which significantly differs from previous works that capture only one or two scopes of relations with only implicit cues. We devise a novel CTSR module to discover collaborative ternary relations in videos, which is lightweight and flexible yet effective, and can be applied to any 3D-CNN architecture. Experiments demonstrate that our approach is able to not only outperform state-of-the-art on three action recognition datasets but also represent relations of implicit and explicit cues effectively (see Section 4).

3 Collaborative Ternary Relations Networks

We construct CoTeRe-Net by integrating CTSR modules (Fig. 1), which is designed on hierarchical mechanism with three levels: aggregation, relation and collaboration, for discovering collaborative ternary relations. CTSR module is lightweight and flexible, thus can be applied to any 3D-CNN architecture.

A CTSR module is a computational unit with the transformation mapping an input $\mathbf{X} \in \mathbb{R}^{C \times T \times H \times W}$ to collaborative relations $\mathbf{Z}^{\zeta} \in \mathbb{R}^{C \times T \times H \times W}$, as shown in Fig. 1. The input of CTSR module $\mathbf{X}$ is a set of feature maps:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_C] \in \mathbb{R}^{C \times T \times H \times W}, \tag{1}$$

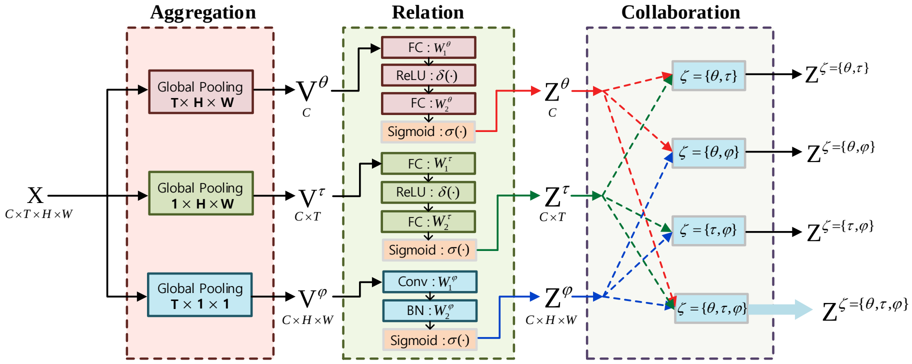


Fig. 1: **CTSR module.** A CTSR module is a computational unit designed on hierarchical mechanism with three levels: aggregation, relation, and collaboration. Aggregation level takes $\mathbf{X}$ as input and outputs relational descriptors $\mathbf{V}$. Relation level analyzes descriptors for yielding ternary relation $\mathbf{Z}$. Collaboration level builds upon relations to generate collaborative representation $\mathbf{Z}^\zeta$.

where $\mathbf{x}_c \in \mathbb{R}^{T\times H\times W}$ $(c = 1, 2, \cdots, C)$ denotes the c-th channel of feature maps, C, T, H, and W represent the channel number, temporal depth, height, and width of feature map, respectively.

We symbolize the ternary as channel θ, temporal τ, and spatial φ, and we define them relying on the scope of corresponding operation dimension, channel C, temporal $C \times T$, and spatial $C \times H \times W$, respectively. As shown in Fig. 1, aggregation level of CTSR module takes $\mathbf{X}$ as input and outputs three relational descriptors $\mathbf{V}^\theta$, $\mathbf{V}^\tau$, and $\mathbf{V}^\varphi$. These descriptors are then fed into relation level for yielding the ternary relation $\mathbf{Z}^\theta$, $\mathbf{Z}^\tau$, and $\mathbf{Z}^\varphi$. Finally collaboration level builds upon the ternary relation to generate channel-temporal-spatial collaborative representation $\mathbf{Z}^\zeta$ $(\zeta = \{\theta, \tau, \varphi\})$.

3.1 Aggregation Level

Aggregation level is at the first of CTSR module, and is designed to aggregate the channel, temporal, and spatial features separately from the input features $\mathbf{X}$, yielding the corresponding relational descriptor $\mathbf{V}$. We employ global pooling with different dimensions of kernel for aggregating different scopes of meaningful and non-linear relational descriptors.

Channel Aggregation. The input $\mathbf{X}$ is pooled on $T \times H \times W$ kernel over spatiotemporal scope, for aggregating channel descriptors $\mathbf{V}^\theta$:

$$\mathbf{V}^\theta = \left[\mathbf{v}_1^\theta, \mathbf{v}_2^\theta, \cdots, \mathbf{v}_C^\theta\right] \in \mathbb{R}^C, \tag{2}$$

and the c-th channel descriptors $\mathbf{v}_c^\theta$ are aggregated by:

$$\mathbf{v}_c^\theta = \frac{1}{T \times H \times W} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbf{x}_c\left(t, h, w\right). \tag{3}$$

where (t, h, w) represents the spatiotemporal position in volume.

Temporal Aggregation. Similarly, temporal descriptors $\mathbf{V}^{\tau}$ are aggregated by pooling the input $\mathbf{X}$ on $1 \times H \times W$ kernel over temporal scope:

$$\mathbf{V}^{\tau} = [\mathbf{v}_1^{\tau}, \mathbf{v}_2^{\tau}, \cdots, \mathbf{v}_C^{\tau}] \in \mathbb{R}^{C \times T}, \tag{4}$$

and the c-th temporal descriptors $\mathbf{v}_c^{\tau}$ are aggregated by:

$$\mathbf{v}_c^{\tau} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbf{x}_c\left(t, h, w\right). \tag{5}$$

Spatial Aggregation. In the same way, we aggregate spatial descriptors $\mathbf{V}^{\varphi}$ using $T \times 1 \times 1$ kernel to pool the input $\mathbf{X}$:

$$\mathbf{V}^{\varphi} = [\mathbf{v}_1^{\varphi}, \mathbf{v}_2^{\varphi}, \cdots, \mathbf{v}_C^{\varphi}] \in \mathbb{R}^{C \times H \times W}, \tag{6}$$

and the c-th spatial descriptors $\mathbf{v}_c^{\varphi}$ are aggregated by:

$$\mathbf{v}_c^{\varphi} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_c\left(t, h, w\right). \tag{7}$$

3.2 Relation Level

Relation level is designed to extract ternary relation $\mathbf{Z}$ from aggregated relational descriptor $\mathbf{V}$ based on gating mechanism. For channel-temporal-spatial ternary relational descriptors, we devise different operations to obtain corresponding ternary relation, involving implicit and explicit cues.

Channel Relations. Channel descriptors $\mathbf{V}^{\theta} \in \mathbb{R}^{C}$ consist of C descriptors. Thus, we feed these C descriptors into multi-layer perceptron (MLP) with one hidden layer to obtain non-linear channel relations, and then they are passed through the sigmoid for activating the final channel relations $\mathbf{Z}^{\theta}$. Channel relations are expressed as:

$$\mathbf{Z}^{\theta} = \sigma\left(\mathbf{W}_2^{\theta} \delta\left(\mathbf{W}_1^{\theta} \mathbf{V}^{\theta}\right)\right) = \left[\mathbf{z}_1^{\theta}, \mathbf{z}_2^{\theta}, \cdots, \mathbf{z}_C^{\theta}\right] \in \mathbb{R}^{C}, \tag{8}$$

where δ and σ refer to `Sigmoid` and `ReLU` functions respectively, $\mathbf{W}_1^{\theta} \in \mathbb{R}^{\frac{C}{r} \times C}$ and $\mathbf{W}_2^{\theta} \in \mathbb{R}^{C \times \frac{C}{r}}$ represent MLP weights and r is reduction ratio of MLP. In such a way, MLP is essentially implemented for implicit relation reasoning [36].

Temporal Relations. Temporal descriptors $\mathbf{V}^{\tau} \in \mathbb{R}^{C \times T}$ are T descriptors for each channel. Thus, we obtain temporal relations $\mathbf{Z}^{\tau}$ in the similar way as channel relations $\mathbf{Z}^{\theta}$, that is, we also employ MLP followed by sigmoid activation on temporal descriptors $\mathbf{V}^{\tau}$. Temporal relations are expressed by:

$$\mathbf{Z}^{\tau} = \sigma\left(\mathbf{W}_2^{\tau} \delta\left(\mathbf{W}_1^{\tau} \mathbf{V}^{\tau}\right)\right) = [\mathbf{z}_1^{\tau}, \mathbf{z}_2^{\tau}, \cdots, \mathbf{z}_C^{\tau}] \in \mathbb{R}^{C \times T}, \tag{9}$$

where $\mathbf{W}_1^{\tau} \in \mathbb{R}^{\frac{(C \times T)}{r} \times (C \times T)}$ and $\mathbf{W}_2^{\tau} \in \mathbb{R}^{(C \times T) \times \frac{(C \times T)}{r}}$ are MLP weights. Here MLP actually performs implicit relation reasoning temporally.

Spatial Relations. Spatial descriptors $\mathbf{V}^{\varphi} \in \mathbb{R}^{C\times T\times W}$ are $T \times W$ spatial representations for each channel, which are different from channel and temporal descriptors. Thereby we adopt spatial 3×3 convolution and batch normalization (BN) on spatial descriptors $\mathbf{V}^{\varphi}$ to extract spatial relations. The final spatial relations $\mathbf{Z}^{\varphi}$ are also obtained through sigmoid activation, and can be expressed as:

$$\mathbf{Z}^{\varphi} = \sigma\left(\mathbf{W}_2^{\varphi}\left(\mathbf{W}_1^{\varphi}\mathbf{V}^{\varphi}\right)\right) = \left[\mathbf{z}_1^{\varphi}, \mathbf{z}_2^{\varphi}, \cdots, \mathbf{z}_C^{\varphi}\right] \in \mathbb{R}^{C\times H\times W}, \tag{10}$$

where $\mathbf{W}_1^{\varphi} \in \mathbb{R}^{C\times 3\times 3}$ and $\mathbf{W}_2^{\varphi} \in \mathbb{R}^{C}$ are weights of convolutional and BN layers respectively. By spatial aggregation and relation, in essence, the simple convolution plays a role in exploiting explicit cues.

3.3 Collaboration Level

Based upon ternary relation $\mathbf{Z}$, we collaborate channel-temporal-spatial relations at the last of CTSR module. The designed collaboration level will discover collaborative ternary relations $\mathbf{Z}^{\zeta}$ among channel-temporal-spatial relations $\mathbf{Z}$:

$$\mathbf{Z}^{\zeta} = \left[\mathbf{z}_1^{\zeta}, \mathbf{z}_2^{\zeta}, \cdots, \mathbf{z}_C^{\zeta}\right] \in \mathbb{R}^{C\times T\times H\times W}, \tag{11}$$

where $\zeta \subseteq \{\theta, \tau, \varphi\}$ is the collaborative set of ternary relation $\{\theta, \tau, \varphi\}$.

Spatiotemporal features in volume are essentially channel-level in CNN architecture. We thereby employ channel-wise relation on each spatiotemporal relation for collaboration, and the c-th collaborative ternary relations $\mathbf{z}^{\zeta\subseteq\{\theta,\tau,\varphi\}}$ can be computed by:

$$\mathbf{z}_c^{\zeta\subseteq\{\theta,\tau,\varphi\}} = \mathbf{z}_c^{\theta} \cdot \left(\tilde{\mathbf{z}}_c^{\tau}\left(t,h,w\right) + \tilde{\mathbf{z}}_c^{\varphi}\left(t,h,w\right)\right), \tag{12}$$

$$\begin{cases} \mathbf{z}_c^{\theta} = \mathbb{1}^{C}, & \text{if } \theta \notin \zeta, \\ \tilde{\mathbf{z}}_c^{\tau} = \mathbb{0}^{C\times T\times H\times W}, & \text{if } \tau \notin \zeta, \\ \tilde{\mathbf{z}}_c^{\varphi} = \mathbb{0}^{C\times T\times H\times W}, & \text{if } \varphi \notin \zeta, \end{cases} \tag{13}$$

where $\tilde{\mathbf{z}}_c^{\tau} \in \mathbb{R}^{C\times T\times H\times W}$ and $\tilde{\mathbf{z}}_c^{\varphi} \in \mathbb{R}^{C\times T\times H\times W}$ denote that $\mathbf{z}_c^{\tau} \in \mathbb{R}^{C\times T}$ and $\mathbf{z}_c^{\varphi} \in \mathbb{R}^{C\times H\times W}$ are broadcasted to the size of $C \times T \times H \times W$, respectively.

Noting that, we have empirically evaluated the performance of addition on the ternary relations, and it does perform worse (about 0.3% lower) than our way formulated in Eq.12. Except for collaborative ternary relations $\mathbf{Z}^{\zeta=\{\theta,\tau,\varphi\}}$, we can choose arbitrary two elements from the set $\{\theta, \tau, \varphi\}$ to acquire corresponding collaborative dual relations in videos, $\mathbf{Z}^{\zeta=\{\theta,\tau\}}$ refers to collaborative channel and temporal relations, $\mathbf{Z}^{\zeta=\{\theta,\varphi\}}$ concerns collaborative channel and spatial relations, and $\mathbf{Z}^{\zeta=\{\tau,\varphi\}}$ involves collaborative temporal and spatial relations.

Finally, the discovered collaborative ternary relations $\mathbf{Z}^{\zeta}$ are used to render the input features $\mathbf{X}$ via element-wise product, yielding the output features:

$$\mathbf{Y} = \mathbf{Z}^{\zeta} \cdot \mathbf{X} \in \mathbb{R}^{C\times T\times H\times W}. \tag{14}$$

Table 1: **Network details of CoTeRe-ResNet-18 architecture.** CoTeRe-ResNet-18 equips 3D ResNet-18 backbone with our CTSR modules.

layer name	output size	3D ResNet-18	CoTeRe-ResNet-18
conv_1	$32 \times 56 \times 56$	$3 \times 7 \times 7$, 64, stride $1 \times 2 \times 2$	
res_{2_x}	$32 \times 56 \times 56$	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \\ \texttt{CTSR}, 64 \end{bmatrix} \times 2$
res_{3_x}	$16 \times 28 \times 28$	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \\ \texttt{CTSR}, 128 \end{bmatrix} \times 2$
res_{4_x}	$8 \times 14 \times 14$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \\ \texttt{CTSR}, 256 \end{bmatrix} \times 2$
res_{5_x}	$4 \times 7 \times 7$	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 2$
pool_5	$1 \times 1 \times 1$	spatiotemporal avg pool, fc layer with softmax	

3.4 Network Architecture

Plug-in CTSR module. To render collaborative ternary relations from learned 3D feature representations, we plug our CTSR module into 3D residual block following 3D convolutions, constructing our CoTeRe-ResNet for videos.

An exemplar: CoTeRe-ResNet-18. We construct CoTeRe-Net by equipping 3D-CNN backbones with our CTSR modules. In current implementation, we develop a CoTeRe-Net by plugging CTSR modules into 3D ResNet-18 architecture, and the resulted architecture is coined as CoTeRe-ResNet-18. Since motion modeling may be partially useful in the early layers, while it might be not necessary at higher levels of semantic abstraction (late layers) [46], thus we integrate CTSR modules on res_2, res_3, and res_4 layers, and leave res_5 layer unchanged. In this way, we can also better balance between model capacity and processing efficiency.

Implementation Details. Table 1 lists the details of our CoTeRe-ResNet-18 architecture, which takes $32 \times 112 \times 112$ volumes as input. We adopt one spatial downsampling at conv_1 implemented by convolutional striding of $1 \times 2 \times 2$, and three spatiotemporal downsamplings at res_{2_1}, res_{3_1}, and res_{4_1} with convolutional striding of $2 \times 2 \times 2$ respectively. We then apply a spatiotemporal average pooling with kernel size of $4 \times 7 \times 7$ on the final convolution at res_5, followed by a FC layer predicting the classification.

Variant CoTeRe-Nets. As illustrated in Section 3.3 and Fig. 1, different choices of set $\{\theta, \tau, \varphi\}$ refer to different collaborations of ternary relation, corresponding to different variants of CTSR modules as well. Thereby we can construct different variants of CoTeRe-Nets equipped by different variants of CTSR modules. These variant CoTeRe-Nets can be used to well study the efficacy of the ternary relation and their collaborations. Thus, in our implementation, we also build these variant CoTeRe-Nets for ablative study to explore the proposed collaborative ternary relations.

4 Experiments

4.1 Datasets and Setups

Something-Something V1 and V2 [10]**.** V1 contains 108,499 short video clips in 174 action labels with simple textual descriptions, which is densely-labelled and fine-grained. V2 is the update of V1, and it contains 220,847 short video clips and also 174 same action labels with V1.
UCF101 [40] **and HMDB51** [22]**.** UCF101 contains 13,320 videos divided into 101 action categories, ranging from daily life activities to unusual sports. HMDB51 contains 6,766 videos divided into 51 action categories.
Training Details. We perform data augmentation on both temporal and spatial scopes. We randomly sample 32 consecutive frames with sampling step 1 for Something-Something V1 and V2, 2 for UCF101 and HMDB51. The input frames are cropped via multi-scale random cropping and then resized to 112×112. The cropping window size is $d \times d$, where d is the multiplication of input shorter side length and scale factors in $[0.8, 1]$ for Something-Something V1 and V2, $[0.7, 0.875]$ for UCF101 and HMDB51. We train and evaluate our models on a computer with 4 NVIDIA RTX 2080Ti GPUs, and set batch size to 8 with freezing BatchNorm parameters in training procedure for studying variant network settings (the batch size can be acctually set to 32 and it will gain within 0.5% performance compared to batch size 8). The network is trained by SGD with momentum 0.9 and weight decay 0.0001. The detailed training procedures for different experiments are explained in the specific sections. All the experiments are implemented by PyTorch framework (version 1.3).
Evaluation Metric. We report top-1 accuracy for all the experiments. We perform multiple clips testing for the evaluation at test time, temporal clips are uniformly sampled from each video, and spatial crops are then sampled from each frame of these clips. For UCF101 and HMDB51, we uniformly sample 10 temporal clips from the full length of the video, and use 3 spatial crops (two sides and the center). For Something-Something V1 and V2, temporal clips are uniformly sampled with the start frame in $[0, L-32]$ (L is the full length of the video), and are uniformly sampled 5 spatial crops (from left to right). We also perform spatial fully-convolutional inference [52,39] by scaling the shorter side of each video frame to 128 while maintaining the aspect ratios. The final prediction is the average softmax scores of all clips.

4.2 Ablation Study

We conduct ablative experiments on Something-Something V1 dataset [10]. We use 3D ResNet-18 as backbone, and construct variant CoTeRe-ResNet-18 models for analysis. Models are trained from scratch, and the training procedure takes 50 epochs total, with an initial learning rate 0.01 and reduces by a factor 0.1 at 40 and 45 epochs.

Analysis of Channel Relations. We first ablatively analyze the efficacy of channel relations by setting $\zeta = \{\theta\}$ (see Section 3.4). For channel relations

Table 2: **Ablation study of 12 variant CoTeRe-Nets.** Each of our ternary relation contributes to accuracy gain of action recognition, and collaborative dual relations contribute more than the single one of ternary relation with further accuracy gains. Our CoTeRe-Net model performs the best, validating the efficacy of CTSR module for discovering relations of both implicit and explicit cues.

model	ζ	r	top-1	params	FLOPs
3D ResNet-18	-	-	43.1	1×	1×
CoTeRe-ResNet-18	$\{\theta\}$	4	43.3	1.011×	1×
CoTeRe-ResNet-18	$\{\theta\}$	2	43.5	1.021×	1×
CoTeRe-ResNet-18	$\{\theta\}$	1	44.2	1.042×	1×

(a) Analysis of channel relations.

model	ζ	r	top-1	params	FLOPs
3D ResNet-18	-	-	43.1	1×	1×
CoTeRe-ResNet-18	$\{\tau\}$	128	43.3	1.012×	1×
CoTeRe-ResNet-18	$\{\tau\}$	64	43.7	1.024×	1×
CoTeRe-ResNet-18	$\{\tau\}$	32	44.5	1.048×	1×

(b) Analysis of temporal relations.

model	ζ	Conv	top-1	params	FLOPs
3D ResNet-18	-	-	43.1	1×	1×
CoTeRe-ResNet-18	$\{\varphi\}$	1×1	43.4	1.005×	1.001×
CoTeRe-ResNet-18	$\{\varphi\}$	3×3	44.7	1.047×	1.009×

(c) Analysis of spatial relations.

model	ζ	top-1	params	FLOPs
3D ResNet-18	-	43.1	1×	1×
CoTeRe-ResNet-18	$\{\theta, \tau\}$	44.9	1.090×	1×
CoTeRe-ResNet-18	$\{\theta, \varphi\}$	45.0	1.089×	1.009×
CoTeRe-ResNet-18	$\{\tau, \varphi\}$	45.4	1.094×	1.009×
CoTeRe-ResNet-18	$\{\theta, \tau, \varphi\}$	**45.8**	1.136×	1.009×

(d) Analysis of collaborative relations.

$\mathbf{Z}^{\theta} \in \mathbb{R}^{C}$, we design a MLP-based gate to reason about relations of implicit cues among different perspectives of global information over spatiotemporal scope (see Section 3.2). We thereby study the impact of reduction ratio r of MLPs. Due to only C dimension of channel relations, we employ three different sizes of $r = \{4, 2, 1\}$ for evaluation, and the results of these CoTeRe-ResNet-18 models are reported in Table 2a. As it can be observed, by decreasing reduction ratio of MLPs, top-1 accuracy of CoTeRe-ResNet-18 models increases, but at the cost of more parameters as well. Considering the balance between accuracy gain and computational cost, we use $r = 1$ for MLPs contributing to channel relations. Moreover, in contrast to the baseline, we can see that channel relations do help to improve the performance of recognizing actions from videos.

Analysis of Temporal Relations. Similarly, we conduct ablative study to analyze the efficacy of temporal relations by setting $\zeta = \{\tau\}$. As to temporal relations $\mathbf{Z}^{\tau} \in \mathbb{R}^{C \times T}$, we adopt similar MLP-based gate design to reason about relations of implicit dependencies between video frames at multiple time scales. Thus we also study the impact of reduction ratio r of MLPs. Differing from channel relations, temporal relations are $C \times T$ dimensional, such that we employ three different sizes of $r = \{128, 64, 32\}$, and evaluation results of these CoTeRe-ResNet-18 models are reported in Table 2b. Similar conclusion can be drawn according to the analysis of r for channel relations. Considering that we can only obtain 0.6% gains with $r = 64$ while the number of parameters only increases 0.048 times with $r = 32$, we thus set $r = 32$ for MLPs conducing to temporal relations. Comparing to baseline, our temporal relations are also beneficial to gain accuracy of action recognition.

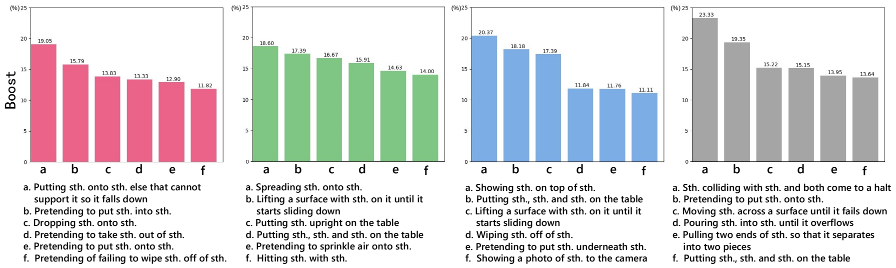
Analysis of Spatial Relations. Then, we investigate the efficacy of spatial relations by setting $\zeta = \{\varphi\}$. Spatial relations concern explicit spatiotemporal topological information that differs from implicit cues, hence we devise a different convolution-based gate for relation discovery (see Section 3.2). For the setting of convolution, we study the impact of kernel size with commonly used 1×1 and 3×3, and evaluation results are reported in Table 2c. Since 1×1 convolution only contributes 0.3% gains, we choose 3×3 convolution with only 0.047 times parameters and 0.009 times FLOPs increase but 1.6% gains on top-1 accuracy. Furthermore, CoTeRe-ResNet-18 with spatial relations involved outperforms baseline 3D ResNet-18 model, indicating efficacy of spatial relations.

Analysis of Collaborative Relations. Except for analyzing each single one of our ternary relation, we also study collaborative dual relations, for validating the efficacy of our collaboration level (see Section 3.3). By setting ζ to $\{\theta, \tau\}$, $\{\theta, \varphi\}$, and $\{\tau, \varphi\}$, we can get three variants of CoTeRe-ResNet-18 models with collaborative dual relations. While, the full CoTeRe-ResNet-18 model with collaborative ternary relations is built by $\zeta = \{\theta, \tau, \varphi\}$. The evaluation results of these four models are reported in Table 2d.

Observing the whole evaluation of Table 2, it is obvious that each of ternary relation makes a contribution to accuracy gain, and collaborative dual relations contribute more than single one of relations with further gains on top-1 accuracy. Overall, the model with collaborative ternary relations performs the best among these variants, verifying the effectiveness of our CTSR module for discovering relations of both implicit and explicit cues for better video representation.

4.3 Boosts Analysis

We show class-wise boosts over baseline with our ternary relation in Fig. 2. It can be seen that, **(1)** channel relations help to gain performance of recognizing actions that require implicit reasoning, such as classes with *pretending* behavior, indicating the efficacy of discovered channel relations with implicit cues; **(2)** temporal relations contribute more to understand actions with implicit temporal cues (dependencies and interactions), *e.g.*, recognizing *spreading*, *putting* and *hitting* actions needs temporal reasoning between video frames; **(3)** spatial relations benefit action recognition with explicit spatiotemporal topological information, for instance, *showing* and *wiping* actions mainly concern relations of objects and motions; **(4)** some fine-grained actions, like *lifting a surface with sth. on it until it starts sliding down* and *putting sth., sth. and sth. on the table*, contain both implicit temporal dependencies and explicit visual objects, thus require both temporal and spatial relations for boosting; **(5)** while CTSR takes advantages of the ternary relations for both implicit reasoning and explicit discovery spatiotemporally, such as, identifying *something colliding with something and both come to a halt* from *moving something and something closer to each*



(a) Channel relations (b) Temporal relations (c) Spatial relations (d) CTSR

Fig. 2: **Class-wise boosts of our ternary relation with respect to the baseline.** Refer to Section 4.3 for further details.

other needs relations of temporal dependencies and visual objects with reasoning information (for *colliding* and *halt*). Therefore, these boosts demonstrate that our ternary relation facilitates the model to discover relations of both implicit and explicit cues which do help to understand actions in videos.

Actually, relations in videos for recognizing actions are much more complex and elusive than that we can imagine, thus ternary relation might still not be elaborate enough to discriminate them, *e.g.*, *pretending* is such an elusive action, and it appears to be boosted by each of ternary relation and CTSR as well. So more effort is still needed for further exploring, and we hope that our work opens up new avenues for video understanding.

4.4 Relations Visualization

We visualize ternary relation on res_2 layer of CoTeRe-ResNet-18 in Fig. 3. For channel relations $\mathbf{Z}^{\theta} \in \mathbb{R}^{C}$ (Eq. 8), we use 2D chart to show scores of channels. For temporal relations $\mathbf{Z}^{\tau} \in \mathbb{R}^{C\times T}$ (Eq. 9), we also use 2D chart to show scores of temporal frames under certain channels. For spatial relations $\mathbf{Z}^{\varphi} \in \mathbb{R}^{C\times H\times W}$ (Eq. 10), we thereby use 3D chart to show scores of spatial widths and heights.

The top of Fig. 3 represents channel descriptors (red curve) and our discovered channel relations (blue curve), it's clear to see the changes indicating relation discovery, and high relation score reflects rich relation while low relation score implies poor relation. We take a closer look at two obvious positive changes at 8^{th} and 18^{th} channel, and two obvious negative changes at 29^{th} and 32^{th} channel, to observe their corresponding temporal and spatial relations, shown in the middle and bottom of Fig. 3 respectively. We can see that the trend and direction of changes for temporal relation discovery are the same as those for channel relation discovery, demonstrating that both of the 8^{th} and 18^{th} channel discover significant relations of implicit cues while both of the 29^{th} and 32^{th} channel discover insignificant relations of implicit cues. By contrast, visualization of spatial relations indicates that, the 8^{th} and 18^{th} channel don't have more relations of visually explicit motion and object information, but the 29^{th} and 32^{th} channel

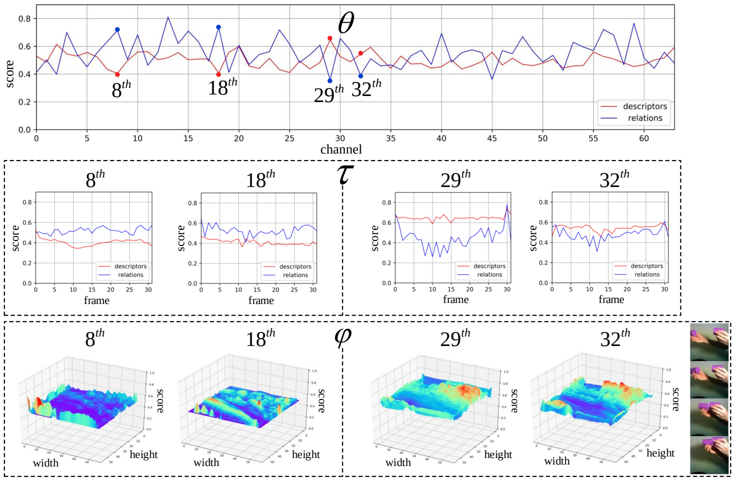


Fig. 3: **Visualization of our ternary relation.** From top to bottom: channel relations, temporal relations, and spatial relations. All scores are obtained by adopting sigmoid functions. Refer to Sections 4.4 and 3 for further details.

have, and high relation scores match with the objects and motion trajectory in the right clip spatially. These visualizations further interpret that our channel and temporal relations refer to implicit cues while our spatial relations concern explicit cues, thus they are supplementary for representing videos better.

4.5 Experiments on Something-Something V1 and V2

We use ResNet-34 backbone network with separable R(2+1)D [46] to construct CoTeRe-ResNet-34, by implementing CTSR modules on res$_2$, res$_3$, and res$_4$ layers. We implement our CoTeRe-ResNet-34 on Kinetics-400 [19] to produce the pre-train models: Kinetics (top-1: 75.3, train Kinetics from scratch for 200 epochs), IG-65M+Kinetics (top-1: 80.2, finetune Kinetics on the released IG-65M [7] pre-trained model for 40 epochs). Besides, we also similarly build CoTeRe-ResNet-18 pre-trained on Kinetics (top-1: 70.6) for fair comparison with previous works. For the experiments on Something-Something V1 and V2, the training procedure is different from ablation study in Section 4.2 when uses IG-65M+Kinetics pre-train, which takes 30 epochs with initial learning rate 0.001 and reduces by a factor 0.1 at 20 and 25 epochs.

Table 3 reports comparison results of top-1 accuracy on Something-Something V1 and V2 datasets, from which we make the following observations: First, our approach outperforms baseline model considerably, by improving top-1 accuracy from 47.2 to 49.7 (V1) and 60.8 to 63.2 (V2) using ResNet-34 without pre-train, which also beats the latest ir-CSN [45] (49.7 vs. 49.3 on V1); Second, our approach with a ResNet-18 backbone pre-trained on Kinetics improves over previous state-of-the-art (with the same settings) by 1% (V1) in top-1 ac-

Table 3: Comparison with state-of-the-arts on Something-Something V1 and V2.

model	backbone	pre-train	FLOPs	top-1(V1)	top-1(V2)
3D-CNN [10]	C3D	Sports1M	N/A	11.5	N/A
TRN* [63]	BNInception	ImageNet	N/A	42.0	55.5
NL I3D+GCN [53]	ResNet-50	ImageNet +Kinetics	158G	46.1	N/A
TrajectoryNet* [62]	ResNet-18	Kinetics	N/A	47.8	N/A
ECO* [64]	ResNet-18	Kinetics	N/A	49.5	N/A
S3D-G [57]	Inception	ImageNet	71.4G	48.2	N/A
GST [29]	ResNet-50	ImageNet	59G	48.6	62.6
TRN Dual Attn.* [56]	BNInception	ImageNet	N/A	N/A	58.4
CPNet [27]	ResNet-34	ImageNet	N/A	N/A	57.7
ir-CSN [45]	ResNet-152	-	96.7G	49.3	N/A
Ghadiyaram *et al.* [7]	ResNet-152	IG-65M	252G	51.6	N/A
STM [16]	ResNet-50	ImageNet	66.5G	50.7	64.2
TSM* [26]	ResNet-50	Kinetics	65.8G	52.6	66.0
MARS* [3]	ResNeXt-101	Kinetics	N/A	53.0	N/A
Martinez *et al.* [31]	ResNet-152	ImageNet	N/A	53.4	N/A
Our baseline	ResNet-34	-	76.3G	47.2	60.8
Our CoTeRe-Net	ResNet-34	-	77.9G	49.7	63.2
Our CoTeRe-Net	ResNet-18	Kinetics	41.1G	50.5	63.9
Our CoTeRe-Net	ResNet-34	Kinetics	77.9G	52.8	66.2
Our CoTeRe-Net	ResNet-34	IG65M +Kinetics	77.9G	**53.9**	**67.1**

*more complicated models with extra information (trajectory features or optical flow). "N/A" means that the paper didn't report the corresponding evaluation value.

curacy (50.5 vs. 49.5, ECO [64]); Third, we further improve our performance by training on deeper backbone ResNet-34 and larger pre-training datasets IG-65M+Kinetics, and substantially increase top-1 accuracy by 6.7% (V1) and 6.3% (V2) against baseline model, achieving state-of-the-art performance on both V1 and V2 datasets. Also note that Martinez *et al.* [31] uses a much deeper ResNet-152 backbone to achieve competitive top-1 accuracy (53.4), while we have not tried it, we expect a similar improvement, referring to boosts of 50.5 to 52.8 (V1) and 63.9 to 66.2 (V2) by only changing backbone from ResNet-18 to ResNet-34, which can further boost our performance. Our CoTeRe-Nets also show competitive computational cost via FLOPs comparison.

Our state-of-the-art results on these two challenging datasets demonstrate the strength of our CoTeRe-Net and highlight the importance of discovering collaborative ternary relations for action recognition from videos.

4.6 Experiments on UCF101 and HMDB51

We also conduct experiments on two classic action recognition benchmarks: UCF101 [40] and HMDB51 [22]. The training procedure takes 40 epochs to-

Table 4: Comparison with state-of-the-arts on UCF101 and HMDB51.

model	backbone	pre-train	UCF101	HMDB51
IDT [49]	-	-	86.4	61.7
C3D-RGB [44]	-	Sports1M	85.2	N/A
Two-stream [38]	-	ImageNet	88.0	59.4
TSN [51]	BNInception	ImageNet	94.2	69.4
P3D [33]	ResNet-152	ImageNet	93.7	N/A
ARTNet with TSN [50]	ResNet-18	Kinetics	94.3	70.9
Attention Cluster [28]	ResNet-152	ImageNet	94.6	69.2
I3D-RGB [2]	Inception	ImageNet+Kinetics	95.6	74.8
STC-Net [4]	ResNeXt-101	ImageNet+Kinetics	95.8	72.6
Zhao *et al.* [61]	BNInception	ImageNet+Kinetics	95.9	N/A
R(2+1)D-RGB [46]	ResNet-34	Sports1M+Kinetics	96.8	74.5
S3D-G [57]	Inception	ImageNet+Kinetics	96.8	75.9
LGD-3D-RGB [34]	ResNet-101	ImageNet+Kinetics	97.0	75.7
Our CoTeRe-Net	ResNet-18	Kinetics	94.5	71.3
Our CoTeRe-Net	ResNet-34	Kinetics	96.4	75.0
Our CoTeRe-Net	ResNet-34	IG-65M+Kinetics	**97.6**	**76.0**

tal, with an initial learning rate 0.001 for Kinetics pre-train and 0.0001 for IG-65M+Kinetics, and reduces by a factor 0.1 at 15 and 30 epochs.

We compare against single RGB models, and Table 4 reports results of top-1 accuracy on these two datasets. Compared to the models with similar or deeper backbone and same pre-train, our CoTeRe-ResNets pre-trained on Kinetics achieve superior performance (top-1 accuracy) on UCF101 and HMDB51. Also our CoTeRe-ResNet-34 pre-trained on IG-65M+Kinetics achieves state-of-the-art performance over existing models with similar settings but much deeper backbones.

5 Conclusion

We propose a novel relation model for discovering collaborative ternary relations in videos. Both boosts analysis and relations visualization validate the efficacy of our CTSR module for representing relations from videos. Both ablation study and evaluation comparison verify the effectiveness of our CoTeRe-Net models on action recognition. To the best of our knowledge, our work is one of the first to model relations involving both implicit and explicit cues for video representation. Nevertheless, relations between things are much more complex and elusive than that we can imagine, thus more effort is still needed for further exploring, and we hope that our work opens up new avenues for video understanding.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under grant numbers 61771440 and 41776113.

References

1. Battaglia, P., Pascanu, R., Lai, M., Rezende, D.J., et al.: Interaction networks for learning about objects, relations and physics. In: NIPS. pp. 4502–4510 (2016)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the Kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
3. Crasto, N., Weinzaepfel, P., Alahari, K., Schmid, C.: MARS: Motion-augmented RGB stream for action recognition. In: CVPR. pp. 7882–7891 (2019)
4. Diba, A., Fayyaz, M., Sharma, V., Mahdi Arzani, M., Yousefzadeh, R., Gall, J., Van Gool, L.: Spatio-temporal channel correlation networks for action classification. In: ECCV. pp. 284–299 (2018)
5. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV. pp. 6202–6211 (2019)
6. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR. pp. 1933–1941 (2016)
7. Ghadiyaram, D., Tran, D., Mahajan, D.: Large-scale weakly-supervised pre-training for video action recognition. In: CVPR. pp. 12046–12055 (2019)
8. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: CVPR. pp. 8359–8367 (2018)
9. Gkioxari, G., Girshick, R., Malik, J.: Actions and attributes from wholes and parts. In: ICCV. pp. 2470–2478 (2015)
10. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The "something something" video database for learning and evaluating visual common sense. In: ICCV. pp. 5842–5850 (2017)
11. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE TPAMI **31**(10), 1775–1789 (2009)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
13. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: CVPR. pp. 3588–3597 (2018)
14. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
15. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE TPAMI **35**(1), 221–231 (2013)
16. Jiang, B., Wang, M., Gan, W., Wu, W., Yan, J.: STM: Spatiotemporal and motion encoding for action recognition. In: ICCV. pp. 2000–2008 (2019)
17. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR. pp. 2901–2910 (2017)
18. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR. pp. 1725–1732 (2014)
19. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P.: The Kinetics human action video dataset. arXiv preprint arXiv:1409.1556 (2017)
20. Kemp, C., Tenenbaum, J.B.: The discovery of structural form. PNAS **105**(31), 10687–10692 (2008)

21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS. pp. 1097–1105 (2012)
22. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A large video database for human motion recognition. In: ICCV. pp. 2556–2563 (2011)
23. Kumar, M.P., Koller, D.: Efficiently selecting regions for scene understanding. In: CVPR. pp. 3217–3224 (2010)
24. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
25. Li, L., Gan, Z., Cheng, Y., Liu, J.: Relation-aware graph attention network for visual question answering. In: ICCV. pp. 10313–10322 (2019)
26. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: ICCV. pp. 7083–7093 (2019)
27. Liu, X., Lee, J.Y., Jin, H.: Learning video representations from correspondence proposals. In: CVPR. pp. 4273–4281 (2019)
28. Long, X., Gan, C., de Melo, G., Wu, J., Liu, X., Wen, S.: Attention clusters: Purely attention based local feature integration for video classification. In: CVPR. pp. 7834–7843 (2018)
29. Luo, C., Yuille, A.L.: Grouped spatial-temporal aggregation for efficient action recognition. In: ICCV. pp. 5512–5521 (2019)
30. Ma, C.Y., Kadav, A., Melvin, I., Kira, Z., AlRegib, G., Graf, H.P.: Attend and interact: Higher-order object interactions for video understanding. In: CVPR. pp. 6790–6800 (2018)
31. Martinez, B., Modolo, D., Xiong, Y., Tighe, J.: Action recognition with spatial-temporal discriminative filter banks. In: ICCV. pp. 5482–5491 (2019)
32. Ni, B., Yang, X., Gao, S.: Progressively parsing interactional objects for fine grained action detection. In: CVPR. pp. 1020–1028 (2016)
33. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with Pseudo-3D residual networks. In: ICCV. pp. 5533–5541 (2017)
34. Qiu, Z., Yao, T., Ngo, C.W., Tian, X., Mei, T.: Learning spatio-temporal representation with local and global diffusion. In: CVPR. pp. 12056–12065 (2019)
35. Russell, B.C., Freeman, W.T., Efros, A.A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: CVPR. vol. 2, pp. 1605–1614 (2006)
36. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. In: NIPS. pp. 4967–4976 (2017)
37. Shou, Z., Lin, X., Kalantidis, Y., Sevilla-Lara, L., Rohrbach, M., Chang, S.F., Yan, Z.: DMC-Net: Generating discriminative motion cues for fast compressed video action recognition. In: CVPR. pp. 1268–1277 (2019)
38. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS. pp. 568–576 (2014)
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
40. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
41. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: AAAI. pp. 4278–4284 (2017)
42. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. pp. 1–9 (2015)

43. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (2016)
44. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: ICCV. pp. 4489–4497 (2015)
45. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. In: ICCV. pp. 5552–5561 (2019)
46. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR. pp. 6450–6459 (2018)
47. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: CVPR. pp. 3156–3164 (2017)
48. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR. pp. 3169–3676 (2011)
49. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV. pp. 3551–3558 (2013)
50. Wang, L., Li, W., Li, W., Van Gool, L.: Appearance-and-relation networks for video classification. In: CVPR. pp. 1430–1439 (2018)
51. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV. pp. 20–36 (2016)
52. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (2018)
53. Wang, X., Gupta, A.: Videos as space-time region graphs. In: ECCV. pp. 399–417 (2018)
54. Watters, N., Zoran, D., Weber, T., Battaglia, P., Pascanu, R., Tacchetti, A.: Visual Interaction Networks: Learning a physics simulator from video. In: NIPS. pp. 4539–4547 (2017)
55. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: CBAM: Convolutional block attention module. In: ECCV. pp. 3–19 (2018)
56. Xiao, T., Fan, Q., Gutfreund, D., Monfort, M., Oliva, A., Zhou, B.: Reasoning about human-object interactions through dual attention networks. In: ICCV. pp. 3919–3928 (2019)
57. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: ECCV. pp. 305–321 (2018)
58. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR. pp. 17–24 (2010)
59. Yao, J., Fidler, S., Urtasun, R.: Describing the scene as a whole: Joint object detection. In: CVPR. Citeseer (2012)
60. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: CVPR. pp. 4694–4702 (2015)
61. Zhao, Y., Xiong, Y., Lin, D.: Recognize actions by disentangling components of dynamics. In: CVPR. pp. 6566–6575 (2018)
62. Zhao, Y., Xiong, Y., Lin, D.: Trajectory convolution for action recognition. In: NeurIPS. pp. 2208–2219 (2018)
63. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: ECCV. pp. 803–818 (2018)
64. Zolfaghari, M., Singh, K., Brox, T.: ECO: Efficient convolutional network for online video understanding. In: ECCV. pp. 695–712 (2018)