Actor and Context Attentions for Spatio-Temporal Action Localization

Manuel Sarmiento
msarmientocalder@apple.com
David Varas

Elisenda Bou-Balust

Apple
Cupertino,
California, USA

Abstract

Most recent advances in spatio-temporal action localization involve modeling different types of interactions between actors and their context to improve actor features. This approach may not be sufficient specially when the context contains essential information of the action. In this work, we propose a novel system in which both actor and context features are enriched taking into account each other for understanding relations. To this end, we contextualize actor information, and use it to adapt the context for each specific actor. We achieve this by using two attention mechanisms. First, an Actor Attention employs multi-head attention layers to enrich actors with contextual features. Then, a Context Attention extracts accurate scene information relevant to each actor. Finally, we introduce a new Context Memory Bank that models long-term temporal information to improve the previous modules.

Experiments on the AVA, AVA-Kinetics and UCF101-24 datasets show the advantages of our approach to model the actor-context interaction, outperforming previous methods by 1.1, 0.5 and 2 points in mAP respectively.

1 Introduction

Spatio-temporal action localization is the task of localizing people and recognizing the actions they are performing in videos. In recent years, this video understanding task has gained popularity because the main research focus has shifted from classifying short clips to understanding long untrimmed videos. This shift has also contributed to considering this task as relevant for industrial applications, such as surveillance, robotics and autonomous driving.

Most of the recent efforts in action localization are focused on reasoning using the relations between actors and their surroundings [27, 35, 41]. These reasoning modules have shown the importance of contextual information for action recognition systems. Many types of relations have been considered in previous works, like relations with the whole scene [47], other actors [29], objects [17], audio events [48], distant temporal relations [44, 45, 52] and even relations between multiple elements of the scene [14, 29, 36, 50].

These works mostly model actor relations through a single feature combination method that enhances the actor feature [34, 41]. However, this may not be sufficient when the context

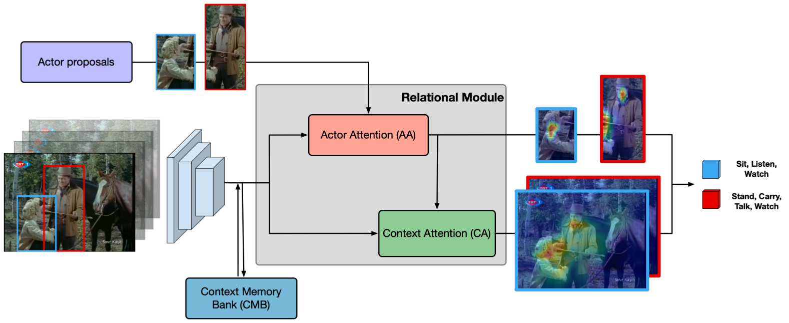


Figure 1: Our system consists of two different attention mechanisms to improve action understanding. The Actor Attention highlights close details of the actors using context features, while the Context Attention emphasizes regions from the scene relevant to each actor. Moreover, the Context Memory Bank enriches the context with adjacent temporal information

contains essential information about the action. Thus, we show that it is possible to further improve relation modeling through an approach that improves both features involved in the interaction. Specifically, our system consists of a Relational Module on top of a backbone architecture that takes actor and context information as inputs. This context information is the feature map output by the backbone, while the RoI pooled feature from the actor proposals are referred to as actor features, following common state of the art nomenclature.

As it can be observed in Figure 1, the Actor Attention (AA) creates an actor enriched feature using the context encoding the actor-context relation. Then, the Context Attention (CA) enhances the context feature for each actor with information from the previous actor-context relation. Using this method, we are able to create a refined actor and context features that gather information from their relation. In contrast to other methods, we use both features for action classification. In addition, we introduce a new Context Memory Bank (CMB) to expand the temporal receptive field of our system. This memory bank enriches context features with long-term temporal information. This module is introduced before AA and CA to exploit the potential impact of the expanded temporal information.

Our main contributions are two-fold:

- We propose a novel relational module for spatio-temporal action localization on videos that enriches both actor and context features to use them for classification, resulting in better action understanding.
- We introduce a new Context Memory Bank to model long-term temporal relations using contextual information. We place this memory bank before our relational module to propagate this information through all the system.

2 Related Work

In this section we review action recognition methods, focusing on spatio-temporal action localization methods that include relational reasoning to improve its performance.

Video Classification. The development of 3D CNNs [4, 37] opened the door to a large increase in performance in action classification. Early works on 3D CNNs consisted on expanding successful 2D architectures in the temporal axis [4, 30, 49]. Later, the SlowFast architecture [11] went one step forward and added a second pathway to model temporal and spatial dimensions at different frame rates. With [10], efficiency was taken into account in the design of 3D architectures obtaining competitive results and reducing the amount of computation needed by the network. More recently, with the appearance of video transformers [1, 2, 9, 23, 24, 25, 26, 28, 32, 40] we have witnessed a new leap in video understanding. The research in video transformers has been driven towards developing robust backbones for action classification. Moreover, in the last year, some efforts are made towards developing better pretraining strategies [12, 19, 43], to boost network performance. However, as our work is mainly focused on modeling actor and context relations, we use a 3D CNN backbone to compare our system against other methods modeling relations.

Spatio-Temporal Action Localization. The development of the action recognition field has encouraged more complex tasks like spatio-temporal action localization. This task not only consists of recognizing the action, but also localizing it in space and time in long untrimmed videos. Works on spatio-temporal action localization [9, 10, 11, 15] follow a common pipeline with a backbone pretrained on a classification task and a person detector. These approaches use cropped features from bounding boxes of people to infer actions. However, these methods neglect contextual information outside those bounding boxes, lacking the reasoning on actor interactions.

Relational Reasoning. One of the first works that considers relations with all the scene instead of a neighborhood of actors is [41]. It captures long range interactions between the extracted features with Non Local Blocks (NLB). In [16], contextual information is aggregated with a transformer-style [39] architecture using actor RoIs as attention queries. Similarly, [34] computes and accumulates pair-wise relation information to enrich actor features using the context. These works [16, 34] focus on enriching actor features. In contrast, we propose to enrich both actor and context features using them for classification.

Actor context relations are explicitly modeled reasoning about interactions between actors and objects in [51]. This approach also models the interactions between actors separately from actor-object interactions. We propose to select relevant parts of the context for each actor, which implicitly considers objects and associates them with the corresponding actor. In [38], instead of using RoI features for classification, they are used for attending to the context information. Although we also perform context attention using actors, we leverage our enriched actor features to extract relevant information from context features and we use both actor and context features to predict the action.

In order to expand the temporal span using features from the past and future, [44] introduces Long Term Feature Banks to capture possible relations between actors at distant time instants, increasing the temporal support of the network. Other works as [3], propose to model spatial and temporal relations as two separate types of relations. Similar to [44], we use a memory bank to model temporal relations. However, we propose a Context Memory Bank that stores context features at an early stage of the system. This allows us to propagate long-term temporal information through the whole system.

In [29], authors go one step further and propose to reason on more than one type of relation at the same time. This work incorporates the reasoning on the relations between actors on top of actor-context interactions. Following the idea of using more than one type of relation, [36] combines actor features with other actor features, object features from a pretrained object detector and long-term temporal information. [31] also combines features from four different sources with a pyramidal structure to model relations. These works model multiple relations using different techniques. However, all of them improve actor features

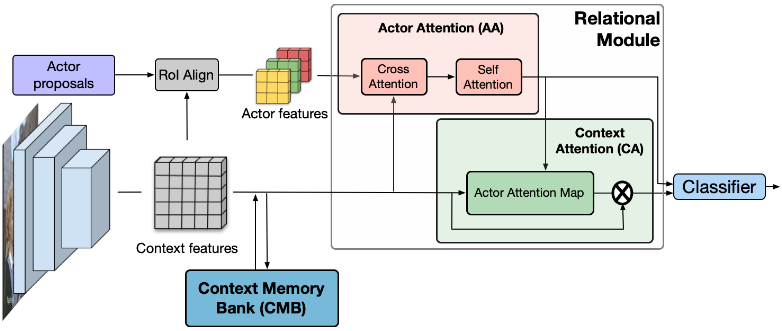


Figure 2: We use cross and self attention to enrich actor features with context. Then, we perform an attention of the context according to each actor. The Context Feature Bank introduces temporal information from adjacent video frames.

by adding information from contextual sources (e.g., scenes, objects, actors). Our approach aims at using both actor and context features for classification. These features implicitly gather information about different types of relations in the video clip.

3 Method

3.1 System Overview

Following common spatio-temporal action localization frameworks [11, 13], our system is composed of a person detector and a video backbone. Given an input video clip I, we first use the backbone network from [11] to extract a feature map.

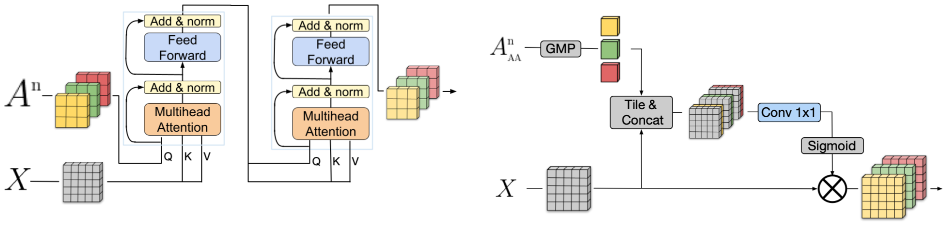
We perform a global temporal average pooling on this feature map to create a spatial context feature $X \in \mathbb{R}^{H \times W \times C}$, where H and W are the spatial dimensions and C is the number of channels. In parallel, we employ a pretrained person detector on the central frame of the input clip I to obtain N actor bounding boxes. Then, we apply a RoI Align operation on X using the extracted actor proposals to generate a feature $A^1, \dots, A^N \in \mathbb{R}^{7 \times 7 \times C}$ for each actor.

We propose to analyze actor and context features using a novel Relational Module composed of two attention mechanisms (Figure 2). The first block, the Actor Attention (AA), is used to understand actors in their context. This mechanism consists of a multi-head attention that enriches actor features A^n using contextual information X. Then, the Context Attention (CA), highlights context information according to the previously enriched actor features. This is done with a spatial reasoning system between X and the enhanced actor feature A^n_{AA}.

Finally, we introduce our Context Memory Bank (CMB) to expand the temporal receptive field of our system. It uses context from adjacent timestamps to improve context features from the current sample, expanding its temporal information.

3.2 Relational Module

In order to obtain rich relational features fusing actor and context information, we propose a Relational Module (RM) with two blocks, each one with a different attention method designed to highlight the most important part from each feature: Actor and Context Attentions.



(a) Actor Attention

(b) Context Attention

Figure 3: Detailed view of the two attention modules of our system. a) the Actor Attention is a Multi-head cross attention system focused on enriching actor features. b) the Context Attention enhances context features for each actor using a sigmoid function.

Actor Attention (AA). This attention mechanism is designed to enrich actor features with contextual information that is lost after the RoI Align operation (Figure 3a), in which we crop context features using actor proposals. To this end, we propose a novel structure composed of two multi-head attention layers. We design a first layer that incorporates context information and a second layer that highlights relevant actor features.

The first layer is a multi-head cross attention layer. We use actor features $\{A^n\}_{n=1}^N$ as Queries (Q), while both Keys (K) and Values (V) are the spatial context feature X. Formally, the resulting actor features Y obtained after this first layer are computed as follows:

$$Y = LN(Y' + FF(Y')), \tag{1}$$

$$Y' = LN(Q + Concat(h_1, ..., h_H)W^O), \tag{2}$$

$$h_i = Att(QW_i^Q, KW_i^K, VW_i^V), \tag{3}$$

where LN is a Layer Normalization, the Feed Forward (FF) consists of two fully connected layers, W^O, W_i^Q, W_i^K and W_i^V are projection matrices and the Att function follows [8].

Moreover, we concatenate the positional encoding introduced in [20] to the features before the multi-head attention layer to exploit the relative positional information between the actor and the context. Using this cross attention mechanism, we select relevant information from the context taking into account the interaction between actor and context and we fuse it into the resulting actor feature map generating an enhanced actor feature.

We use the output of this multi-head cross attention as input to a multi-head self attention layer. This mechanism allows us to weight the contribution of the spatio-temporal information associated to each actor. The resulting features obtained by this second layer are also computed using Equations 1, 2, 3. However, in this case, the queries (Q), keys (K) and values (V) of the attention are the features Y previously computed. We obtain a set of features $\{A^n_{\text{AA}}\}_{n=1}^N \in \mathbb{R}^{7\times7\times C}$ that gather relevant information of the actor-context relation.

Context Attention (CA). In this module, we use the relevant spatial information from the context associated to each actor to improve action classification. We propose to enrich context features according to the actor-context relation. To this end, we use each actor feature to weight relevant parts of the context, obtaining N enriched context features. As this weighting is performed by conditioning the context to each actor, we use the previously enhanced actor features in this attention instead of the RoI pooled features for better performance.

First, we spatially reduce the actor feature $\{A^n_{\text{AA}}\}$ through a Global Max Pooling (GMP) operation. Then, we replicate and concatenate this $1x1$ feature to all the locations of the context feature X (Figure 3b). We use a 2D-convolution to reduce the number of channels of this concatenated feature, obtaining a feature representation $\{Z^n\} \in \mathbb{R}^{H\times W\times C}$ with the same number of channels as the actor feature. Formally, this feature is obtained as:

$$Z^n_{j,k} = Concat(GMP(A^n_{\text{AA}}), X_{j,k})W_{j,k} \tag{4}$$

where j and k indexes correspond to the spatial dimensions and $W_{j,k}$ is a learnable weight.

Finally, we propose to transform this feature $\{Z^n\}$ into an actor attention map that highlights relevant regions of the context for each actor. Then, we apply a sigmoid function on $\{Z^n\}$ to attend to the spatial context feature map X with an element-wise multiplication to obtain A^n_{CA}. Using this method, we create an enhanced context feature for each actor that contains information from the actor-context relation.

Classifier. Finally, we perform a max pooling over the two features obtained for each actor (A^n_{AA} and A^n_{CA}), concatenate them and classify the resulting vector with two linear layers.

3.3 Context Memory Bank

In a video clip, the interactions between actors and context may occur at different time intervals. Thus, in order to understand actor-context relations, it is necessary to take into account a large temporal window around the actors.

We introduce a memory bank to model long-term temporal relations leveraging information from previous and future time steps as it is done by other works [13, 29, 36, 44]. Previous methods, extract actor features from adjacent timesteps. However, these approaches make it impossible to extract information from timesteps where there are no actors detected. Instead, we propose a novel actor-agnostic Context Memory Bank in which global context features are used to model temporal relations. As our CMB features are not associated with actors we can extract temporal information from adjacent frames even if there are no actors in the scene.

We propose to use temporal features at the start of the Relational Module which allows our system to take into account this information to model further relations, in contrast to previous systems where temporal relations are modeled at the end of the system. Our approach has two main advantages. On the one hand, we enrich actor features with context information from a wide temporal window in AA. On the other hand, the context feature used in CA is enhanced with spatial and temporal information of the actor-context relation.

To create our actor-agnostic CMB, we first train an instance of our network without the CMB, following the same training schedule. Then, the bank is created by doing inference of video clips sampled every second through this network. We store the context features spatially max pooled for computational reasons. We denote these features as $M \in \mathbb{R}^C$ to differentiate them from the context feature of the current time step.

During the training of our network with the CMB, at time T, we sample context features from the bank for time steps from $[T-\omega, T+\omega]$ obtaining $\{M^t\}_{t=1}^{2\omega+1}$ temporal features. Then, current context and adjacent temporal features relations are computed by adding a NLB [41], where X are the queries, while keys and values are M^t. This layer is initialized following [41] strategy to avoid breaking the behavior of the pretrained network.

	mAP
Baseline	25.17
AA	27.95
CA	28.10
AA → CA	**28.59**

(a) Modules Combination

	mAP
CA → AA	28.04
AA → CA	**28.59**

(b) Relation Order

	mAP
Parallel	28.27
Serial	**28.59**

(c) Aggregation Structure

	mAP
w/o CMB	28.59
Before RM	**29.71**
After RM	28.25

(d) Memory Bank Position

	mAP
1L	**29.71**
2L	29.10
3L	29.01

(e) CMB Depth

	mAP
ω 15	29.71
ω 20	29.99
ω 25	**30.07**
ω 30	29.99

(f) Temporal Support

Table 1: Ablation study of different aspects of our system on the AVA v2.2 validation set.

4 Experiments on AVA

In this section, we assess the proposed method using the Atomic Visual Actions (AVA) dataset [18] [1], for training and evaluating our method. This is the main video dataset for spatio-temporal action localization. AVA is composed of 430 movie clips of 15 minutes each. Its annotation consists of bounding boxes and labels for every person in the video provided at one fps. We use AVA v2.2 version, which consists of 184k annotated frames for training and 50k for validation. We also use Kinetics [4, 5, 6] to pretrain the video backbone network. In addition, we use the recent AVA-Kinetics v1.0 [22] for evaluation. This dataset is an expansion of AVA using videos from [6] annotated like AVA. All the results are obtained using the 60 most common classes, following the standard evaluating protocol [18]. We use mean Average Precision with a frame-level IoU threshold of 0.5 (mAP@0.5).

4.1 Implementation Details

Person Detector. To obtain actor region proposals we use the precomputed detections from [44]. Following [11], we only use person detections with a confidence greater than 0.8.
Backbone Network. The video backbone selected for this task is a SlowFast ResNet-50 network and input sampling $T \times \tau = 8 \times 8$. Following [11], we increase the spatial resolution of res_5 layer $2\times$. Unless stated otherwise, the network is pretrained with Kinetics-400.
Training and Inference. We use a batch of 32 video clips during training and a single clip per GPU at inference. We train for 6 epochs, with a learning rate of 0.08 using linear warm-up during the first epoch. We decay the learning rate by a factor of 10 at epochs 5.6 and 5.8. We set a weight decay of 10^{-7} and Nesterov momentum of 0.9 for the optimizer. During training we use ground truth annotations and bounding boxes from the person detector that match an annotation with an IoU greater than 0.9 as actor proposals. We resize frames to have short size of 256 pixels and use binary cross entropy loss for training.

4.2 Ablation Experiments

In this section, we present experiments to assess the performance of each the proposed systems. The *Baseline* method consists of the backbone with two classification layers.
Modules Combination. In this experiment, we evaluate the contribution of each attention block to the overall performance. To this end, we conduct experiments removing these attention mechanisms from our system. We consider three possible structures: 1) Only use the AA block, 2) Only use the CA block, or 3) Combine the two previous modules (AA+CA). The results from these three structures are shown in Table 1a. We can observe that both

[1] AVA is made available by Google Inc. under a Creative Commons Attribution 4.0 International license

Model	Pretrain	mAP
SlowFast, R50, 8 × 8 [11]	K400	24.8
ACAR, R50, 8 ×8 [29]	K400	28.8
MViTv2-B, 32×3 [23]	K400	29.0
Ours, R50, 8 ×8	K400	**30.1**
SlowFast, R101+NL, 8 × 8 [11]	K600	27.4
MViTv2-B, 32×3 [23]	K600	30.5
Object Transformer [46]	K600	31.0
ACAR, R101+NL, 8 × 8 [29]	K600	31.4
Ours, R101+NL, 8 ×8	K600	**32.1**

Table 2: Results on AVA v2.2 validation set.

Model	mAP
SlowFast, R101, 8 × 8 [7]	33.0
ACAR*, R101, 8 × 8 [7]	35.8
Ours, R101, 8 × 8	**36.3**

Table 3: Comparison on AVA-Kinetics v1.0 validation set. * refers to a better person detector.

systems 1 and 2 achieve competitive results. However, the combination of both attention mechanisms obtains the best results. This shows that the enhanced context features provide additional information to the enriched actor features for classifying actions.

Relation Order. In this experiment, we study different ways of combining AA and CA. We consider two configurations: 1) CA → AA uses the enhanced context output by CA as input context for AA whereas 2) AA → CA uses the enriched actor features extracted from AA as actor features for CA. As Table 1b shows, it is important the order in which these two blocks are included, rich actor information has a greater impact on context features than what an enhanced context can contribute to improve actor features. This result indicates that actor features contain more relevant information than context features.

Aggregation Structure. In this experiment, we analyze the impact of connecting the AA and CA attention blocks. On the one hand, the Parallel model uses actor features from the initial RoI pooling as input for both AA and CA. On the other hand, the Serial one leverages enriched actor features from AA for context enhancement in CA. In Table 1c we observe that the higher semantics from the Serial model obtain better performance than the raw individual blocks, showing that the second attention mechanism benefits from the first.

Context Memory Bank. In these experiments, we evaluate the effect of the CMB on our system. First, we study the position in which we can introduce a memory bank to exploit long-term temporal dependencies. We analyze 2 different locations: 1) Before and 2) After the RM. The main difference is that in 1 we store a global feature for each timestep while in 2 we store a feature associated with each actor. In Table 1d, we can see the performance of these two systems. Introducing a memory block after the RM performs worse than the system without long-term temporal information. This behaviour, which has been observed before in [29], may be related to the system needing a more complex reasoning structure to understand relations between the temporal information and our enriched features. The good performance of the memory block before the RM shows that introducing our actor-agnostic CMB allows the temporal information to be propagated through all the stages of the system.

We conducted an experiment to show the effect of stacking NLBs with a fixed value of $\omega = 15$ in our CMB. In Table 1e, we can observe that the system does not benefit from having more than one NLB. This may be related to the absence of spatial information in the temporal features. Finally, we experiment with different values of the temporal window. As shown in Table 1f, we achieve the best performance with $\omega = 25$. Using this length we obtain a result of 30.07 in mAP, a new state of the art in relation modeling to the best of our knowledge. However, the impact of expanding the temporal window is limited. This suggests that the

Model	Frame-mAP
AIA, R50 [36]	78.8
ACAR, R50 [29]	84.3
YOWO [21]	87.3
Ours, R50	**89.3**

Table 4: Results on UCF101-24 Split 1.

relevant temporal information is close to the instant where the action happens.

4.3 Comparison to State of the Art

In this section we assess our method using AVA v2.2 validation set and other relevant methods in this field of research. We present results pretraining the backbone with Kinetics-400 and Kinetics-600. Even though there are other methods like [12, 42, 45] that propose different pretraining techniques or use more powerful networks to obtain better results, we decided to compare our system to comparable methods for fairness.

As Table 2 shows, our method outperforms [23] by a margin of 1.1 and 1.6 mAP points when pretraining on K400 and K600. MViTv2 consists of a powerful video transformer but no relational module is considered after the backbone. These results show the importance of relation reasoning for the task of spatio-temporal action localization.

Comparing our method to other works that use relational modules after the backbone as ACAR-Net [29], it can be observed that we outperform this system by 1.3 and 0.7 mAP points when using R50 and R101 backbones. This result shows that our RM and CMB capture relevant information from the actor-context relation enriching actor and context features.

We also conducted experiments on the AVA-Kinetics dataset using a model pretrained with [6]. Table 3 shows that our system outperforms other models obtaining an improvement of 0.5 in mAP, even though the second best system is trained using a better person detector. This shows that our model can generalize well to other datasets.

5 Experiments on UCF101-24

In this section we present the results obtained by our system on the split 1 of the UCF101-24 [33], a dataset with 24 action classes and 3207 videos.

We use SlowFast R50 8 × 4 pretrained on Kinetics-400, the person detector from [21], and trained the system without the CMB. We used a base learning rate of 0.01. For training we used ground truth boxes while for inference we employ the detected boxes. The other experimental settings are the same used for the AVA dataset.

In Table 4 we report the Frame-mAP with an IoU threshold of 0.5. Our system outperforms previous methods, highlighting the importance of enriching both actor and context features Although in UCF101-24 there are no actor interactions like in the AVA dataset, our method exploits context information to improve the action detection results. This demonstrates again that our model can perform well in a different dataset.

6 Conclusions

In this work, we present Actor and Context Attentions, a new method for relation modeling in spatio-temporal action localization. Our approach enriches both actor and context features considering each other and uses them for classification improving action understanding.

We conducted extensive experiments to show the necessity of robust relation modeling and the importance of both actor and context features for action recognition. We also showed the importance of an early modeling of temporal relations with our novel Context Memory Bank. Our method outperforms previous systems that model relations on the spatio-temporal action localization with a mAP improvement of 1.3 points in AVA, 0.5 points in AVA-Kinetics and 2 points in UCF101-24.

References

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

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *arXiv preprint arXiv:2102.05095*, 2021.

[3] Manuel Sarmiento Calderó, David Varas, and Elisenda Bou-Balust. Spatio-temporal context for action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2021.

[4] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[5] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. In *arXiv preprint arXiv:1808.01340*, 2020.

[6] Lucas Smaira Joao Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. In *arXiv preprint arXiv:2010.10864*, 2020.

[7] Siyu Chen, Junting Pan, Guanglu Song, Manyuan Zhang, Hao Shao, Ziyi Lin, Jing Shao, Hongsheng Li, and Yu Liu. Actor-context-actor relation network for spatio-temporal action localization. In *arXiv preprint arXiv:2006.09116*, 2020.

[8] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

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

[10] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 200–210, 2020.

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[12] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *Advances in Neural Information Processing Systems*, 2022.

[13] Yutong Feng, Jianwen Jiang, Ziyuan Huang, Zhiwu Qing, Xiang Wang, Shiwei Zhang, Mingqian Tang, and Yue Gao. Relation modeling in spatio-temporal action localization, 2021. URL https://static.googleusercontent.com/media/research.google.com/en//ava/2021/A1_CVPRW2021_AVA-Kinetics_Alibaba-Tsinghua.pdf,2021.

[14] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

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

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

[17] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8359–8367, 2018.

[18] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6047–6056, 2018.

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022.

[20] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664, 2021.

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

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

[23] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[25] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11999–12009, 2022.

[26] K. Mangalam, H. Fan, Y. Li, C. Wu, B. Xiong, C. Feichtenhofer, and J. Malik. Reversible vision transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10820–10830, 2022.

[27] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[28] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 3156–3165, 2021.

[29] 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 Conference on Computer Vision and Pattern Recognition*, pages 464–474, 2021.

[30] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[31] Carlos Roig, Manuel Sarmiento, David Varas, Issey Masuda, Juan Carlos Riveiro, and Elisenda Bou-Balust. Multi-modal pyramid feature combination for human action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 3742–3746, 2019.

[32] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. An image is worth 16x16 words, what is a video worth? In *arXiv preprint arXiv:2103.13915*, 2021.

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

[34] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Proceedings of the European conference on computer vision (ECCV)*, pages 335–351, 2018.

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

[36] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *Proceedings of the European conference on computer vision (ECCV)*, volume 12360, pages 71–87, 2020.

[37] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[38] Oytun Ulutan, Swati Rallapalli, Mudhakar Srivatsa, and B. S. Manjunath. Actor conditioned attention maps for video action detection. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 516–525, 2020.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[40] J. Wang and L. Torresani. Deformable video transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14033–14042, 2022.

[41] 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 (CVPR)*, pages 7794–7803, 2018.

[42] Chen Wei, Haoqi Fan, Saining Xie, Chaoxia Wu, Alan Loddon Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14648–14658, 2021.

[43] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *arXiv preprint arXiv:2304.03283*, 2023.

[44] C. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 284–293, 2019.

[45] C. Wu, Y. Li, K. Mangalam, H. Fan, B. Xiong, J. Malik, and C. Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13577–13587, 2022.

[46] Chao-Yuan Wu and Philipp Krähenbül. Towards long-form video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPRR)*, 2021.

[47] Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, and Gangshan Wu. Context-aware rcnn: A baseline for action detection in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 440–456, 2020.

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

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

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

[51] Yubo Zhang, Pavel Tokmakov, Cordelia Schmid, and Martial Hebert. A structured model for action detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9967–9976, 2019.

[52] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 831–846, 2018.