DOAD: Decoupled One Stage Action Detection Network

Shuning Chang[1*] Pichao Wang[3†] Fan Wang[3] Jiashi Feng[2] Mike Zheng Shou[1†]
[1]Showlab, National University of Singapore [2]National University of Singapore [3]Alibaba Group

Abstract

Localizing people and recognizing their actions from videos is a challenging task towards high-level video understanding. Existing methods are mostly two-stage based, with one stage for person bounding box generation and the other stage for action recognition. However, such two-stage methods are generally with low efficiency. We observe that directly unifying detection and action recognition normally suffers from (i) inferior learning due to different desired properties of context representation for detection and action recognition; (ii) optimization difficulty with insufficient training data. In this work, we present a decoupled one-stage network dubbed DOAD, to mitigate above issues and improve the efficiency for spatio-temporal action detection. To achieve it, we decouple detection and action recognition into two branches. Specifically, one branch focuses on detection representation for actor detection, and the other one for action recognition. For the action branch, we design a transformer-based module (TransPC) to model pairwise relationships between people and context. Different from commonly used vector-based dot product in self-attention, it is built upon a novel matrix-based key and value for Hadamard attention to model person-context information. It not only exploits relationships between person pairs but also takes into account context and relative position information. The results on AVA and UCF101-24 datasets show that our method is competitive with two-stage state-of-the-art methods with significant efficiency improvement.

1. Introduction

The objective of action detection is to localize and recognize human actions in video clips along space and time. Unlike general action recognition, the actions in this task emphasize on actors' interactions with the context. As a fundamental and challenging task in video understanding, it has been widely applied to various tasks, such as abnormal behavior detection [50, 21] and autonomous driving [28].

Spatio-temporal action detection usually consists of two sub-tasks, person detection and action recognition. Most existing methods typically adopt two-stage solutions. As shown in Figure 1, they typically follow the top-down strategy [15, 36, 51, 46, 48, 38, 26] that employs off-the-shelf detectors to localize person instances at first and then recognize their action categories with various backbones. Though with high performance, these methods are not efficient as they require two-stage processing for detection and action recognition separately. We observe that detection and action recognition have different desired properties for context representation in this task. Due to the existence of interaction categories, action recognition requires to fuse corresponding entity (*e.g.*, other people or objects) features from context to construct various interaction relationships. In contrast, detection also benefits from context, but it tends to incorporate the context features that support the bounding box regression and is not sensitive to interaction. Take Figure 2a as a toy example. We slightly change the pose of the interested person from the left image to the right image. The aggregated context features for detection are nearly unchanged, but for action recognition they shall be learned from scratch again due to the change of interaction objects. The different objectives of the two sub-tasks need different context supports for optimal learning. Moreover, entangled modeling of the two sub-tasks leads to difficulty in opti-

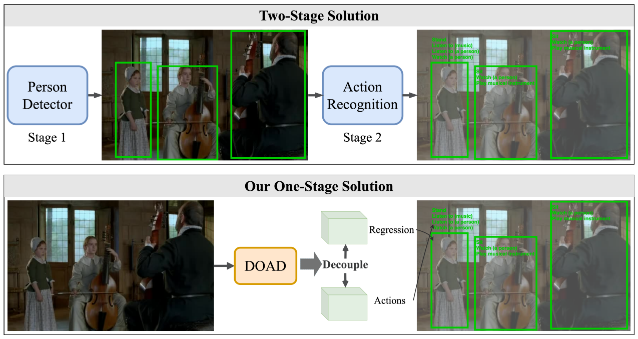


Figure 1: Comparison between two-stage solution and our one-stage solution to spatio-temporal action detection. Traditional two-stage methods use an off-the-shelf detector to generate person bounding boxes suffering low efficiency. The proposed DOAD model decouples detection representation and action recognition representation to accomplish different sub-tasks in a single stage.

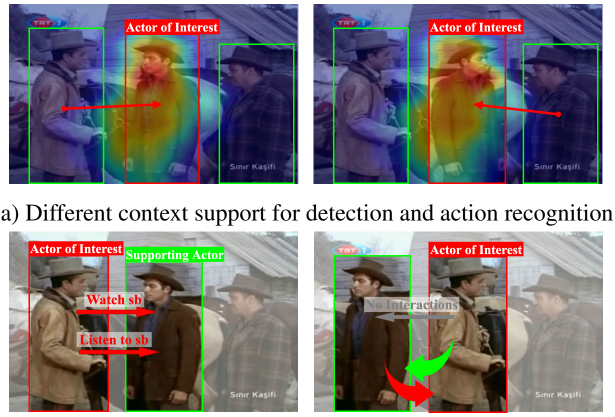
[*] Work done during an internship at Alibaba Group.
[†] Equal corresponding authors.

mization, especially in the case of limited video data with annotations. Due to these reasons, it is difficult to integrate them into entangled one-stage framework to achieve strong performance.

To alleviate these issues, different from previous methods, we propose a novel decoupled one-stage architecture to unify detection and action recognition into one backbone. Our method decouples detection and action recognition into two separate branches to make the two sub-tasks learn their own optimal context support. In our architecture, person detection is performed in the detection branch, which aggregates context information from temporal supporting proposals, and action recognition is performed in the action branch for person-person or person-context relationship mining.

Specifically, in the detection branch, a region proposal network (RPN) is adopted to generate person bounding boxes, and an ROI pooled person feature is served for bounding box regression. Considering the detection is performed on video frames, we employ a temporal aggregation module to enhance the keyframe features by aggregating its adjacent frame features. In the action branch, we design a transformer-based structure which unifies person and context features to capture interaction representation. In this task, action recognition shall understand actors' interactions with surrounding context, including environments, other actors, and objects. Prior works [48, 38] focus on building various interaction relationships, such as person-person, person-object, and long-term temporal interactions. However, they just model corresponding entity features in each relationship and stack them independently, which neglects the correlation among person features, context and relative position information. Take Figure 2b as an example. There are two person-person interaction categories, "watch sb" and "listen to sb", between the actor-of-interest in the red bounding box and the supporting actor in the green bounding box. Assuming the two actors exchange their positions, their appearance features are nearly unchanged, but there are no interactions between them any more. If we do not consider the context and position information, wrong results might be obtained. Therefore, we argue that the relationships among people, context, and position shall be considered simultaneously. Inspired by the vanilla transformer [42], we design a TransPC (Transformer for Person-Context) layer which models pairwise person relationships upon their holistic spatial structure to retain the context and relative position information. We construct matrices of pairwise person-context features as key and value. Different from vanilla self-attention that deals with sequence input, we propose the Hadamard product to compute attention map between the sequence query and the matrix key. Our TransPC is able to incorporate features from more informative entities and produce more accurate action prediction.

Our contributions are summarized as three-fold:



(a) Different context support for detection and action recognition.

(b) The importance of the context and position information for action recognition.

Figure 2: (a) shows different context support for person detection and action recognition. The heatmap represents the informative context for detecting actor-of-interest. Changing his pose from the left image to the right image does not change the context for detection, but does change the context (supporting person) for action recognition to another person in the image. (b) illustrates that the context and position information are crucial clues for action recognition. In the right image, we exchange the positions of two people, then they have no interaction.

- We propose a one-stage spatio-temporal action detection model, which decouples detection representation and action representation for person detection and action recognition, respectively, ensuring that they have optimal context feature aggregation.
- We propose a novel transformer-based method, TransPC, to explicitly integrate person and context features with relative position information for action recognition.
- We demonstrate the effectiveness of our method on the mainstream datasets, AVA and UCF101-24. Our method outperforms well established state-of-the-art one-stage methods significantly and is comparable to the two-stage methods with significant improvement of efficiency.

2. Related work

Action recognition. Action recognition is a fundamental task of video understanding. Convolutional networks have long been the standard for this task. They can be roughly separated into two groups, *i.e.*, two-stream networks and 3D CNNs. Two stream networks [33, 12, 43, 47] use 2D CNNs to extract frame features from RGB and optical flow sequences, while 3D CNNs [39, 4, 27, 11, 22] adopt 3D convolutional layers to model the temporal information from the original videos. Since 3D convolutional networks consume more computation, many methods explore to decouple spatial and temporal dimensions or use grouped convolutions [37, 40, 41, 49, 10]. With the significant success of Vision Transformer (ViT) [8], a shift in action recognition from CNNs to transformers emerges recently. Benefit from the self-attention mechanism which broadens wider receptive field with fewer parameters and lower computation costs, those methods [3, 1, 25, 9, 52, 30] present state-of-the-art performance and impressive potential. In this work, besides traditional CNN-based networks, we also try to adopt transformer-based networks as our backbone to extract video features.

Spatio-temporal action detection. Action recognition processes well-trimmed videos, where the models only need to classify short video clips to action labels. However, most videos are untrimmed and long in practical applications. Recent works explored temporal action localization [32, 55, 6, 23, 2, 54, 5] and spatio-temporal action detection [13, 15, 36, 51, 20, 46, 48, 38, 26] on untrimmed videos. Spatio-temporal action detection is more difficult than action classification and temporal action detection because these models need to not only predict the action categories but also localize the action in time and space. Most recent works focus on capturing various interaction relationships between actors and context. They normally adopt a two-stage framework where actor boxes are first generated by an off-the-shelf detector and then classified. Wu et al. [48] present a strong baseline network by simply expanding actor bounding boxes and incorporating global feature, which demonstrates the importance of the context information. Tang et al. [38] explore nearly all the main interactions including person-person, person-object, and long-term temporal interaction. They model each interaction by the self-attention mechanism and then stack them to improve the performance. Moreover, an Asynchronous Memory Update (AMU) algorithm is proposed to estimate the memory features dynamically for long-term temporal interaction capture. Pan et al. [26] show an Actor-Context-Actor Relation model to uncover the deeper level relationship between actors and context by applying a high-order relation reasoning to build the actor-context-actor relations. Those methods achieve significant performance. However, two-stage approaches are not efficient, which limits their application in the real world. Girdhar et al. [13] adopt an RPN network to generate bounding box proposals and use a transformer head to generate classification and bounding box regression results. Their method is a one-stage method but lacks finer interaction relationship construction and does not consider the difference of optimal context support for two subtasks. In this work, we propose a one-stage method that contains our novel person-context interaction relationship mining module and addresses these issues. Experiments demonstrate that our method outperforms competitive baselines.

Attention mechanism for video context capture. Context information capture plays a pivotal role in video understanding. Attention mechanism is one of the most effective and common technique to solve it. Attention mechanism is to compute the response at a position in a sequence by accessing all positions and taking their weighted average in the embedding space. Vaswani et al. [42] first introduce a self-attention mechanism, called transformer, capturing long-range context among words in one sentence to address the machine translation task. Girdhar et al. [13] first utilize transformer in video tasks to aggregate features from the spatio-temporal context for recognizing and localizing human actions; after that, many related works [56, 31, 44] apply transformer in various video tasks. There are some other ways besides transformers to utilize attention mechanism to capture long-range dependencies. Wang et al. [45] embed non-local structure into the action recognition network to capture spatio-temporal context dependencies. Wu et al. [46] introduce long-term temporal context feature banks to compute interactions between the current short-term features and the long-term features for video analysis. Recent works [48, 38, 26] explore the interactions between people and all kinds of context in this task, *e.g.*, person-person and person-scene. However, most works model each context independently and are short of exploring the associations between different contexts. Although [26] proposes the concept of actor-context-actor relation, we still think it does not pay attention to the relative position relationship of different entities.

3. Method

In this section, we present the proposed method which targets to construct an effective and efficient one-stage model. The overall architecture of our method is shown in Figure 3. We first employ a video backbone to extract features of the input video. Then, the video representation is decoupled into two branches, *i.e.*, detection branch and action branch. The detection branch (Section 3.2) is based on Faster-RCNN structure and deploys a temporal aggregation module to generate person bounding boxes. The action branch (Section 3.3) mines person-context interaction by TransPC module and long-term temporal interaction by a memory feature bank to achieve action recognition.

3.1. Overall framework

Our method deals with a short video clip centered on the center frame $\mathcal{F}_k$ ("keyframe"). Following the pipeline of previous spatio-temporal action detection methods [13, 15, 36, 51, 46, 48, 38, 26], it generates a set of person bounding boxes for all the people in the keyframe, and recognizes all the actions for each person in this short time.

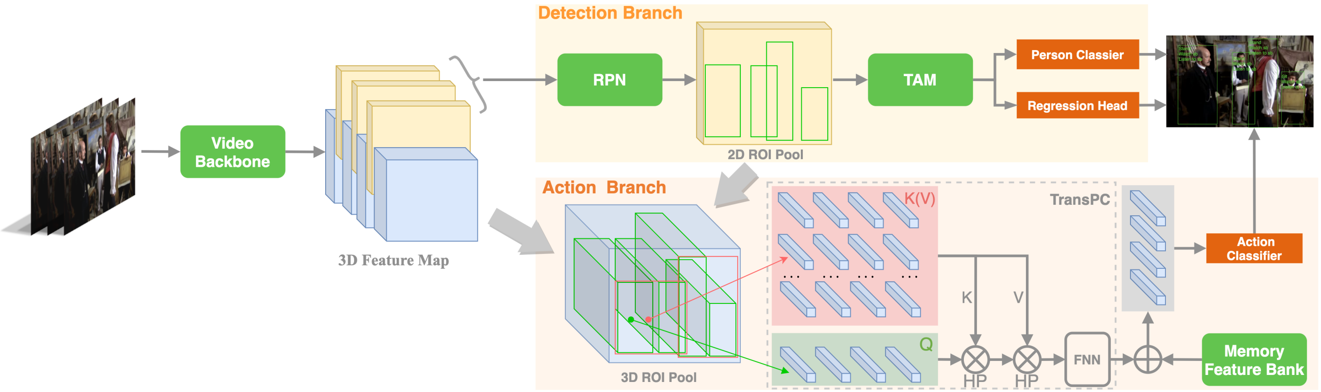


Figure 3: Illustration on the architecture of our method. It first employs a video backbone network to extract a 3D video feature. Then the detection and action recognition is decoupled into two branches. In the detection branch, we apply a similar Faster-RCNN framework incorporating video temporal context by temporal aggregation module (TAM) to generate bounding boxes and conduct pose action estimation. In the action branch, we adopt a TransPC to integrate person and context features to capture interaction relationships. HP is short for Hadamard product.

We begin by extracting a T-frame short clip centered on the given keyframe $\mathcal{F}_k$ from a video. We encode this input using a backbone network (*e.g.* Video Swin Transformer [25] or SlowFast [11]) to obtain a $T \times H \times W$ feature map X. We feed the feature map into the detection and action branches. The detection branch aggregates adjacent frame features to enhance the keyframe person features to generate person bounding boxes. For the action branch, we propose a TransPC module combining person and context features to capture interaction relationships. Finally, we obtain the generated bounding box coordinates and action categories as our final action detection results. We will illustrate our detection branch and action branch in Section 3.2 and Section 3.3, respectively.

3.2. Detection branch

Our detection branch is similar to the Faster R-CNN object detection framework [29]. We slice out the keyframe feature $X_k \in \mathbb{R}^{H \times W}$ from feature map X and feed it into a region proposal network (RPN). The RPN generates person bounding box proposals with scores. We then select N proposals ($N = 300$) according to the interaction of overlap (IoU) with ground truths. After that, person features are extracted by align ROI pooling operation from the selected proposals. These features are applied to classify proposals into *Person* and *Background*, two categories, and regress to a 4D vector of offsets to predict a more accurate bounding box. Finally, we use NMS to remove redundant boxes and set a threshold to filter out boxes with low confidence scores. The final bounding boxes are regarded as actor spatial location, and they are also served for action recognition during inference.

Temporal aggregation module. Above operations belong to a general image detection pipeline. However, our input is a video clip. Only using static images ignores temporal context information and makes it hard to deal with challenging situations in videos, *e.g.*, occlusion and motion blur. Therefore, we design a temporal aggregation module to enhance the features of the keyframe proposals. Besides keyframe feature X_k, we select two frame features X_{k-s} and X_{k+s} with a distance of s from the keyframe as reference frames. Both keyframe feature and reference frame features are fed into RPN and ROI pooling to generate keyframe proposal features and reference proposal features, notated as $F^k = \{f_1^k, f_2^k, ..., f_N^k\}$ and $F^r = \{f_1^r, f_2^r, ..., f_N^r\}$ respectively. Transformer is adopted here to aggregate features of reference proposals to generate more informative keyframe proposal. The transformer block is composed of self-attention layer and feed-forward network (FFN). The attention map produced by self-attention layer is computed by matching the transformed keyframe proposals F^k (a.k.a. the queries) $Q = \phi(F^k)$ to another transformation of the reference proposals (a.k.a. the keys) $K = \theta(F^r)$, with ϕ and θ being learnable linear transformation.

$$A_d = \text{Softmax}(\frac{\phi(F^k) * \theta(F^r)^\top}{\sqrt{d}}), \tag{1}$$

where $A_d \in \mathbb{R}^{N \times N}$ is the generated attention map, d is the dimension for F^k and F^r, and $*$ is the dot product. Considering that our goal is to enhance the features of the keyframe proposals with the reference proposals, we use the original reference proposal features as values instead of projecting them with a linear transformation. In the FFN layer, we apply a linear projection on the aggregated feature and add it to the keyframe proposals. Namely,

$$F^{k'} = F^k + \text{Linear}(A_d * F^r), \tag{2}$$

where $F^{k'}$ is the final aggregated proposal features, Linear is the linear projection, and $*$ is the dot product.

3.3. Action branch

In the action branch, we emphasize on the variant interactions for action recognition. We propose a TransPC module to construct person-context interaction and it considers entity features, context and position information synchronously. Furthermore, we employ a memory feature bank to capture long-term temporal interaction.

TransPC. In this module, we adopt transformer to integrate person feature and context feature to explore the interaction relationships. For a keyframe $\mathcal{F}_k$, its person bounding box set $P = \{p_1, p_2, ..., p_n\}$ is obtained from our detection branch, where n is the number of proposals. We aim to compute the correlation between each person (target person $p_i^t \in P$) and other person (supporting person $p_j^s \in P$). We use align ROI pooling to crop 3D person feature from 3D feature map X. Then, the 3D person feature is converted to 1D person feature f^t via temporal and spatial pooling. We obtain the sequence of person feature set $F^p = [f_1^p, f_2^p, ..., f_n^p]$. Different from using F^p as queries and keys in the conventional way, here we use new person-context features as keys to embed context features and position information. Specifically, for pairwise target person p^t and supporting person p^s, we select a rectangle box which encloses two person bounding boxes with a minimum area as our interested region. The four coordinates of this new box, *i.e.*, top-left and bottom-right coordinates (x_1, y_1, x_2, y_2), are computed as:

$$\begin{aligned} x_1 &= \min(x_1^t, x_1^s), \quad y_1 = \min(y_1^t, y_1^s), \\ x_2 &= \max(x_2^t, x_2^s), \quad y_2 = \max(y_2^t, y_2^s), \end{aligned} \tag{3}$$

where $(x_1^t, y_1^t, x_2^t, y_2^t)$ and $(x_1^s, y_1^s, x_2^s, y_2^s)$ are the coordinates of target person and supporting person, respectively. This box keeps the most critical context and their relative position. We also use aligned ROI pooling following two convolution layers with zero padding to crop our person-context box from feature map X and project it to generate the key and value. The convolution operation benefits retaining the spatial position and adding zero padding can further strengthen this effect. Finally, a max pooling is employed to transform it as a 1D feature f^{pc}. Because each target person $p^t \in P$ need to compute the f^{pc} with each supporting person $p^s \in P$, our person-context feature set F^{pc} is a $n \times n$ matrix but not a sequence. The F^{pc} is defined as:

$$F^{pc} = \begin{pmatrix} f_{11}^{pc} & f_{12}^{pc} & \cdots & f_{1n}^{pc} \\ f_{21}^{pc} & f_{22}^{pc} & \cdots & f_{2n}^{pc} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1}^{pc} & f_{n2}^{pc} & \cdots & f_{nn}^{pc} \end{pmatrix}. \tag{4}$$

The vanilla transformer deals with sequence key and value. We cannot directly adopt it to compute our attention map. To enable attention map calculation between our F^p and F^{pc}, we first repeat the person feature sequence F^p for n times along row dimension to produce a $n \times n$ matrix F^{p*} which is represented as:

$$F^{p*} = \begin{pmatrix} f_1^p & f_1^p & \cdots & f_1^p \\ f_2^p & f_2^p & \cdots & f_2^p \\ \vdots & \vdots & \ddots & \vdots \\ f_n^p & f_n^p & \cdots & f_n^p \end{pmatrix}. \tag{5}$$

Then, We compute the Hadamard product of F^{p*} and F^{pc} to obtain attention map A_a:

$$\begin{aligned} A_a &= \sigma(F^{p*} \odot F^{pc}) \\ &= \sigma(\begin{pmatrix} f_1^p * f_{11}^{pc} & f_1^p * f_{12}^{pc} & \cdots & f_1^p * f_{1n}^{pc} \\ f_2^p * f_{21}^{pc} & f_2^p * f_{22}^{pc} & \cdots & f_2^p * f_{2n}^{pc} \\ \vdots & \vdots & \ddots & \vdots \\ f_n^p * f_{n1}^{pc} & f_n^p * f_{n2}^{pc} & \cdots & f_n^p * f_{nn}^{pc} \end{pmatrix}), \end{aligned} \tag{6}$$

where $\odot$ is Hadamard product, σ is a softmax function, and $*$ is the dot product of two vectors.

The person feature aggregation is performed as a weighted summation of the person-context feature values with the attention map as summation weights. Thus, we compute the Hadamard product of attention map A_a and value matrix, and sum the output along the row dimension to generate the sequence of the aggregated features. Finally, we apply a residual connection to sum the person features and the aggregated features. Similar to [38], we also adopt a dense serial structure to integrate our TransPC blocks.

Memory feature bank. Long-term memory features can provide effective temporal information to assist recognizing actions. Inspired by the Long-term Feature Bank (LFB) proposed in [46], we build a memory feature bank to store both past and future person features for the long-term temporal interaction capture. During training, we store the person features according to the ground truth bounding boxes. During inference, we use the bounding boxes provided by our detection branch. We insert our memory feature bank after our TransPC module.

4. Experiments on AVA

The Atomic Visual Actions (AVA) [15] dataset is collected for spatio-temporal action detection. In this dataset, each person on keyframes is annotated with a bounding box and corresponding action labels at 1 FPS. There are 80 atomic action categories including 14 pose categories and 66 interaction categries. This dataset contains 430 15-minute videos splitting into 235 training videos, 64 validation videos, and 131 test videos.

Since our method is designed for spatio-temporal action detection, we adopt AVA dataset as the main benchmark to conduct detailed ablation experiments. The results are evaluated with the official metric of frame level mean average precision (mAP) at spatial IoU $\geq$ 0.5, and 60 categories with at least 25 instances in validation and test splits are used for evaluation following the conventional setup [15].

4.1. Implementation details

Backbone. With the modeling shift from CNNs to transformers in the vision community, pure transformer architectures have achieved top accuracy on the major video recognition benchmarks [3, 1, 25, 9, 52, 30]. In this work, we adopt state-of-the-art Video Swin Transformer [25] as our backbone. Its base version (Swin-B) is selected, which consists of 4 stages, each containing 2, 2, 18, and 2 Swin Transformer Blocks. The original Swin-B performs 32× spatial downsampling for the input videos. To maintain a larger spatial resolution of feature map X, we remove the last stage with the last patch merging layer to make the downsampling rate to be 16. All other settings follow the recipe in [25]. Our backbone is pre-trained on Kinetics-600 [4] dataset for action recognition task. To the best of our knowledge, this is the first work to explore the performance of transform-based backbone for this task. Besides the transformer-based backbone, we also use the 3D CNN backbone for fair comparison. We choose SlowFast [11] network with ResNet-101 structure which is pre-trained on Kinetics-700 dataset.

Training and inference. The inputs of our network are 16 RGB frames, uniformly sampled from a 32-frame raw clip centered on a keyframe. All the video clips are scaled such that the shorter side becomes 256 and the longer side becomes 464, and then fed into backbone network initialized from Kinetics pre-trained weights. Random flipping is used during training. We train our network using the SGD optimizer with batch size 16. We train for 110k iterations with a base learning rate of 0.001, which is then decreased by a factor of 10 at 70k and 90k iteration. A linear warm-up scheduler [14] is applied for the first 2k iterations. On AVA dataset, pose categories are mutually exclusive and interaction categories are not, so we use cross-entropy loss function for pose categories classification and binary cross-entropy loss function for interaction categories. To alleviate the deficiency of training data for person detector, we first use the data with "person" labels from MSCOCO [24] to pre-train the detection branch. Since the data in MSCOCO are static images, the same images are stacked repeatedly to form video clips. During training, the person bounding boxes produced by RPN with IoU greater than 0.8 predicted will be fed into our action branch for action recognition. During inference, predicted person bounding boxes with a confidence score larger than 0.8 are used. In our memory feature bank, both 30 past and future clips are used.

4.2. Ablation study

To verify the effectiveness of our method, we conduct ablation experiments on the validation set of AVA v2.2. The backbone we used is the modified Swin-B (more detail in Section 4.1).

Decoupled *vs*. coupled. In our method, we use a decoupled structure to decouple detection representation and action representation. Further, we experiment with decoupled structure and coupled structure. The results are shown in Table 1. The coupled structure refers to the structure of [13]. After the RPN module generates person proposals, the detection branch integrates into the action branch. Both bounding box regression loss and action classification loss are upon the person features produced by the action branch. Our decoupled structure outperforms the coupled structure with a large margin, which demonstrates the effectiveness of our method.

Detection branch. Our detection branch is responsible for generating person bounding boxes, which is the basis of our framework. We evaluate our variants in our detection branch by ablating its temporal aggregation module (TAM) and MSCOCO data pre-training in Table 2. *Baseline* represents the model containing a basic faster-rcnn structure and a completed action branch. We can see that the temporal aggregation module can improve the results, which demonstrates that the temporal information is effective for our detection task but has been missed in previous methods. With more complicated video object detection methods [16, 7] having been developed, we believe that the performance of detection branch still has lots of potential. One of the advantages of two-stage methods is that the off-the-shelf person detector can benefit from large-scale datasets in the person detection community. Similarly, in our one-stage method, extra pre-training data like MSCOCO dataset can also be employed, which brings in a significant gain.

The effectiveness of TransPC. Experiments in Table 3 verify the effectiveness of our TransPC and compare TransPC with its counterpart method. In this table, *Baseline* represents the model without TransPC module, and *Person-person* represents the general person-person interaction module (similar to [38]) without considering the context information. Our TransPC can largely enhance the mAP of *Baseline* because the interaction relationships play a core role in this task. Our TransPC is also much better than the general person-person interaction method, demonstrating that the interaction relationship benefits from context information.

To further verify the effectiveness of our TransPC, we visualize attention weights from our TransPC and the general person-person interaction module in Figure 4. The actors of

Structure	mAP
Coupled	26.16
Decoupled (ours)	28.82

Table 1: Performance evaluation on coupled structure and decoupled structure.

Variants of model	mAP
Baseline	27.34
Baseline + TAM	27.75
Baseline + TAM + MSCOCO	28.82

Table 2: Performance evaluation on different components of detection branch.

Variants of model	mAP
Baseline	26.50
Baseline + TransPC	28.82
Baseline + Person-person	28.29

Table 3: Performance evaluation on the effectiveness of TransPC.

Number of TransPC blocks	mAP
1	28.35
2	28.68
3	28.82
4	28.84

Table 4: Performance evaluation on different number of TransPC blocks.

interest are in the red boxes and the other actors are in the green boxes. In these multiple people cases, the general interaction module cannot distinguish the importance of other actors due to the lack of context and position information during building person-person relationship. In contrast, our TransPC can pay more attention to the actual supporting actors.

Number of TransPC blocks. We adopt dense serial structure [38] to arrange our TransPC blocks. The effects of different numbers of TransPC blocks are shown in Table 4. Considering the trade-off between accuracy and time consumption, we use three TransPC blocks in our method.

Our TransPC *vs.* other schemes. Furthermore, we explore different structures which could be an alternative to our TransPC, including: (i) directly use Eq. 4 to generate attention map, (ii) use general person sequence feature as key to produce attention map and add the result of Eq. 4. Note that, only one block is used here for simplicity. Table 5 compares all these variants, with our choice outperforming other two variations.

Scheme	mAP
Scheme i	28.01
Scheme ii	28.10
TransPC	28.35

Table 5: Comparison of our TransPC with other schemes.

Pipeline	Model	Modalities	Input	mAP
2-stage	ACRN [36]	V+F	32 × 2	17.4
	SlowFast [11]	V	32 × 2	27.3
	LFB [46]	V	32 × 2	27.7
	CA-RCNN [48]	V	32 × 2	28.8
	AIA [38]	V	32 × 2	31.2
	ACAR-Net [26]	V	64 × 2	30.0
1-stage	YOWO [20]	V	16 × 1	19.2
	Jiang et al. [18]	V+F	20× -	21.7
	VAT [13]	V	64× -	25.0
	Ours (SlowFast)	V	16 × 2	27.5
	Ours (Swin-B)	V	16 × 2	27.7

Table 6: Comparison on AVA v2.1. V and F refer to visual frames and optical flow respectively. The input is shown as the frame number and corresponding sample rate.

4.3. Comparison with state of the art

We compare our results with existing state-of-the-art one-stage and two-stage methods on the validation set of both AVA v2.1 (Table 6) and v2.2 (Table 7). For fair comparison, our experiments only use a single model and a single scale for testing. We provide results with SlowFast backbone and popular transformer-based Video Swin Transformer backbone. On both AVA v2.1 and v2.2, our results outperform all the results of one-stage methods by a large margin and are superior or comparable with two-stage methods, which indicates our method is effective. Comparing two backbones, Swin-B achieves a slightly better performance than SlowFast, which demonstrates that the transformer-based backbone is effective in this task. The per category results for our method are shown in our supplementary material.

As a one-stage method, another advantage of our method is the inference speed. We evaluate the average inference time of a single video clip on a typical two-stage method AIA [38] and ours in Table 8. Both use SlowFast backbone with ResNet-101 structure. Our inference time is only 64% of that of AIA, showing its efficiency.

5. Experiments on UCF101-24

Dataset. UCF101-24 is a subset of UCF101 [35] consisting of 3207 videos with spatio-temporal annotations on 24 action categories. We conduct experiments on the first split of this dataset following previous methods. We use the corrected annotations provided in [34].

Pipeline	Model	Modalities	Input	mAP
2-stage	SlowFast [11]	V	32 × 2	29.0
	AIA [38]	V	32 × 2	32.3
	ACAR-Net [26]	V	64 × 2	33.3
1-stage	YOWO [20]	V	16 × 1	20.2
	Ours (SlowFast)	V	16 × 2	28.5
	Ours (Swin-B)	V	16 × 2	28.8

Table 7: Comparison on AVA v2.2. V and F refer to visual frames and optical flow respectively. The input is shown as the frame number and corresponding sample rate.

Method	Detector	Action Recognition	Total time
AIA [38]	0.106	0.156	0.262
Ours	-	-	0.168

Table 8: Comparison of the average inference time of each video clip (s/video clip) between two-stage method AIA and ours.

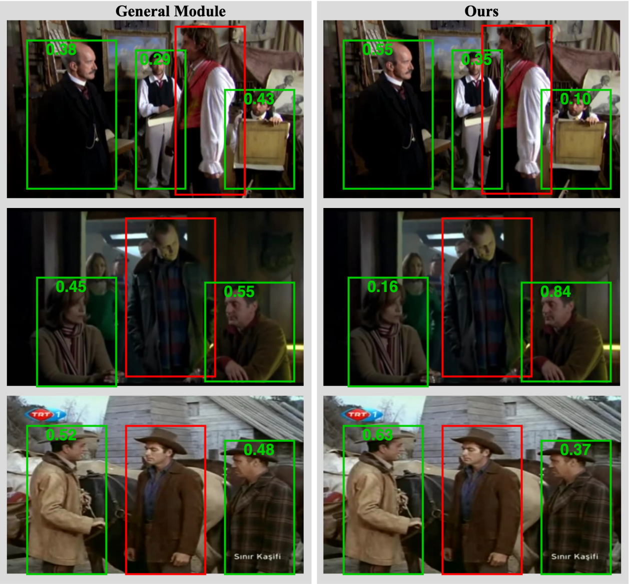


Figure 4: We visualize attention weights from our TransPC and the general person-person interaction module. The actors of interest are in the red boxes and the other actors are in the green boxes. In each image, we remove the weight of actor of interest and re-normalize the rest attention weights to 1.

Implementation Details. Following [26], we use SlowFast with ResNet-50 structure as our backbone. The temporal sampling for the slow pathway is 8 × 4, and the 32 frames as input are fed into the fast pathway. We pre-train it on the Kinetics-400 dataset. Other hyper-parameters are similar to the experiments on AVA.

Pipeline	Model	Modalities	mAP
2-stage	T-CNN [17]	V	67.3
	ACT [19]	V	69.5
	STEP [51]	V+F	75.0
	I3D [4]	V+F	76.3
	Zhang et al. [53]	V	77.9
	S3D-G [49]	V+F	78.8
	AIA [38]	V	76.8
1-stage	YOWO [20]	V	70.5
	Ours	V	74.8

Table 9: Comparison on UCF101-24 split 1. V and F refer to visual frames and optical flow respectively. The metric we used is frame-mAP.

Quantitative results. Table 9 shows the result of UCF101-24 test set in frame-mAP with 0.5 IoU threshold. Our method surpasses another one-stage method with a considerable margin and is also competitive with two-stage methods. This outstanding performance illustrate the effectiveness and generality of our method again. We argue that UCF101-24 is not very suitable for most recent methods, including ours, in this task because the categories in this dataset are not interactive. Thus, many interaction relationships exploited by these methods are not very beneficial. Moreover, the quality of frames in this dataset is lower than AVA and MSCOCO, which adversely influences our detection results.

6. Limitations

There are several limitations of this work. First, the proposed DOAD method is only evaluated on the commonly used AVA and UCF101-24 datasets. More evaluations on other datasets will be better. Second, limited by the more GPU memory consumption of our key matrix, our method is difficult to be applied in crowded scenes.

7. Conclusion

In this paper, we propose a new effective and efficient one-stage sptio-temporal action detection network, DOAD. We decouple the person detection and action recognition into two branches to alleviate the issue of different optimal context supports. Moreover, different from independently utilizing kinds of context, we present a novel TransPC module to integrate the person and context features to capture the interaction relationships. Our method significantly outperforms all the existing one-stage methods and is superior or comparable with two-stage methods on challenging benchmarks. Our method provides a new and strong one-stage framework which still has tremendous potential. In the future, we plan to further study how to improve the performance of person detection and capture more fine-grained detail features for action recognition.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. 2021. 3, 6

[2] Yueran Bai, Yingying Wang, Yunhai Tong, Yang Yang, Qiyue Liu, and Junhui Liu. Boundary content graph neural network for temporal action proposal generation. *arXiv preprint arXiv:2008.01432*, 2020. 3

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021. 3, 6

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 6, 8

[5] Shuning Chang, Pichao Wang, Fan Wang, Hao Li, and Jiashi Feng. Augmented transformer with adaptive graph for temporal action proposal generation. *arXiv preprint arXiv:2103.16024*, 2021. 3

[6] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1130–1139, 2018. 3

[7] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10337–10346, 2020. 6

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[9] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *arXiv preprint arXiv:2104.11227*, 2021. 3, 6

[10] 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. 3

[11] 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. 2, 4, 6, 7, 8

[12] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 2

[13] 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. 3, 6, 7

[14] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 6

[15] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 1, 3, 5, 6

[16] Liang Han, Pichao Wang, Zhaozheng Yin, Fan Wang, and Hao Li. Exploiting better feature aggregation for video object detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1469–1477, 2020. 6

[17] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 5822–5831, 2017. 8

[18] Jianwen Jiang, Yu Cao, Lin Song, Shiwei Zhang4 Yunkai Li, Ziyao Xu, Qian Wu, Chuang Gan, Chi Zhang, and Gang Yu. Human centric spatio-temporal action localization. In *ActivityNet Workshop on CVPR*, 2018. 7

[19] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4405–4413, 2017. 8

[20] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. 2019. 3, 7, 8

[21] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Transactions on Image Processing*, 29:2395–2408, 2019. 1

[22] Jun Li, Xianglong Liu, Wenxuan Zhang, Mingyuan Zhang, Jingkuan Song, and Nicu Sebe. Spatio-temporal attention networks for action recognition and detection. *IEEE Transactions on Multimedia*, 22(11):2990–3001, 2020. 2

[23] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3889–3898, 2019. 3

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[25] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. 3, 4, 6

[26] 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. 1, 3, 7, 8

[27] 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*, pages 5533–5541, 2017. 2

[28] Mohamed Ramzy, Hazem Rashed, Ahmad El Sallab, and Senthil Kumar Yogamani. Rst-modnet: Real-time spatio-temporal moving object detection for autonomous driving. *CoRR*, abs/1912.00438, 2019. 1

[29] 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*, 28:91–99, 2015. 4

[30] Michael S Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: What can 8 learned tokens do for images and videos? *arXiv preprint arXiv:2106.11297*, 2021. 3, 6

[31] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Video multitask transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[32] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1049–1058, 2016. 3

[33] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 2

[34] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3637–3646, 2017. 7

[35] 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. 7

[36] 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. 1, 3, 7

[37] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *ICCV*, pages 4597–4605, 2015. 3

[38] 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. 1, 2, 3, 5, 6, 7, 8

[39] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 2

[40] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, pages 5552–5561, 2019. 3

[41] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 3

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2, 3

[43] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 2

[44] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1571–1580, 2021. 3

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3

[46] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, 2019. 1, 3, 5, 7

[47] Hanbo Wu, Xin Ma, and Yibin Li. Convolutional networks with channel and stips attention model for action recognition in videos. *IEEE Transactions on Multimedia*, 22(9):2293–2306, 2019. 2

[48] Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, and Gangshan Wu. Context-aware rcnn: A baseline for action detection in videos. In *European Conference on Computer Vision*, pages 440–456. Springer, 2020. 1, 2, 3, 7

[49] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 3, 8

[50] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. In *British Machine Vision Conference 2015*, 2015. 1

[51] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz. Step: Spatio-temporal progressive learning for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2019. 1, 3, 8

[52] Xuefan Zha, Wentao Zhu, Tingxun Lv, Sen Yang, and Ji Liu. Shifted chunk transformer for spatio-temporal representational learning. *arXiv preprint arXiv:2108.11575*, 2021. 3, 6

[53] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. A structured model for action detection. In *CVPR*, pages 9975–9984, 2019. 8

[54] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *ECCV*, 2020. 3

[55] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017. 3

[56] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 3