

# Multi-Label Activity Recognition using Activity-specific Features and Activity Correlations

Yanyi Zhang,<sup>1</sup> Xinyu Li,<sup>1,2</sup> Ivan Marsic<sup>1</sup>

<sup>1</sup> Rutgers University–New Brunswick, Electrical and Computer Engineering Department

<sup>2</sup> Amazon Web Services

## Abstract

*Multi-label activity recognition is designed for recognizing multiple activities that are performed simultaneously or sequentially in each video. Most recent activity recognition networks focus on single-activities, that assume only one activity in each video. These networks extract shared features for all the activities, which are not designed for multi-label activities. We introduce an approach to multi-label activity recognition that extracts independent feature descriptors for each activity and learns activity correlations. This structure can be trained end-to-end and plugged into any existing network structures for video classification. Our method outperformed state-of-the-art approaches on four multi-label activity recognition datasets. To better understand the activity-specific features that the system generated, we visualized these activity-specific features in the Charades dataset. The code will be released later.*

## 1. Introduction

Activity recognition has been studied in recent years due to its great potential in real-world applications. Recent activity recognition researches [31, 46, 18, 33] focused on single-activity recognition assuming that each video contains only one activity, without considering a multi-label problem where a video may contain multiple activities (concurrent or sequential). Multi-label activity recognition is an understudied field but has more general real-world use cases (e.g., sports activity recognition [47, 5], or daily life activity recognition [44]). Most of the recent multi-label activity recognition methods are derived from structures for single activities that generate a shared feature vector and apply sigmoid as the output activation function [35, 55, 7, 57, 12, 59]. Although these approaches enable the network to provide multi-label outputs, the features are not designed for multi-label activities.

We introduce our mechanism to recognize multi-label activities from another angle by generating independent fea-

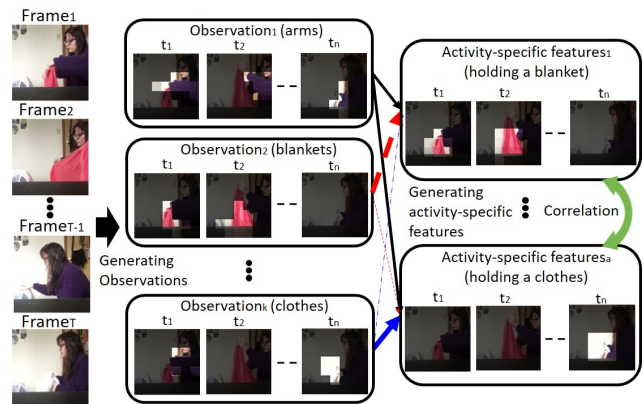


Figure 1: System overview using an example from Charades. The system first generates  $k$  independent feature snippets (“observations”) that focus on different key regions from the video (arms, blankets, and clothes). The activity-specific features are then generated by independently combining these observations. The weights of the observations that contribute to activity-specific features are represented as lines with different colors (black, red, and blue). The thicker lines denote higher weights. For example, the *activity-specific features*<sub>1</sub> (holding a blanket) are obtained by combining information from *observation*<sub>1</sub> (focuses on arms) and *observation*<sub>2</sub> (focuses on clothes). The system finally learns correlations between activity-specific features and provide multi-label activity predictions.

ture descriptors for different activities. We named these feature descriptors “**activity-specific features**”. This mechanism generates activity-specific features in two stages. The first-stage network (Figure 1, middle) summarizes the feature maps extracted by the backbone network (3D convolution layers) and generates a set of independent feature snippets by applying independent spatio-temporal attention for each snippet. We name these feature snippets “**observations**”. In the second stage (Figure 1, right), the network learns activity-specific features from different combinations of observations for different activities. In this way, each activity is represented as an independent set of feature descriptors (activity-specific features). The multi-label activity predictions can then be made based on the activity-specific features. Unlike most of the previous approaches [35, 55, 7, 57, 12, 59] that generate a shared feature vec-

tor to represent multiple activities by pooling feature maps globally, our network produces specific feature descriptors for each activity.

Label dependencies have proven important for multi-label image classification [26, 20, 53], and we argue that it is also important to consider the label-wise correlation for activities. We generate an activity label-wise **correlation map** to model co-existing patterns (e.g., walking and talking) and exclusive patterns (e.g., sitting and standing). The correlation map is applied on the activity-specific features to predict activities based on the highly correlated (or exclusive) activities’ descriptors. In multi-label activity videos, different activities might have different duration and need to be recognized using video clips with different lengths. To address this issue, we further introduced a **speed-invariant tuning** method for generating activity-specific features and recognizing multi-label activities using inputs with different downsampling rates.

We evaluated our model on both large-scale multi-label activity datasets (Charades [44] and AVA [19]), and real-world multi-label sports datasets (Volleyball [47] and Hockey [5]) to show that our model performs well on multi-label datasets and is applicable to real-world tasks. Our introduced model outperformed the recent state-of-the-art networks on all four datasets without using additional information (e.g., optical flow [24] or object information [29]) other than RGB frames, which demonstrated the efficiency of our introduced method. We also provide detailed ablation experiments on the model structure to show that the introduced activity-specific features and activity correlation work as expected. We further visualized the activity-specific features by applying the learned attention maps on the backbone features (feature maps after the last 3D convolution layer) to represent the activity-specific feature maps. Our contributions can be summarized as:

- A network structure that generates activity-specific features for multi-label activity recognition.
- An activity correlation map that learns correlations between different activity-specific features.
- A speed-invariant tuning method that produces multi-label activity predictions using different temporal-resolution inputs.

## 2. Related Work

**Activity Recognition.** Video-based activity recognition has been developing rapidly in recent years due to the success of deep learning methods for image recognition [32, 49, 23]. Compared to image classification, activity recognition depends on spatio-temporal features extracted from consecutive frames instead of spatial-only features from static images. Two-stream networks apply two-branch

convolution layers to extract motion features from consecutive frames as well as spatial features from static images and fuse them for activity recognition [45, 13, 55]. Others proposed 3D-convolution-based networks for extracting spatio-temporal features from the videos instead of using manually designed optical flow for extracting motions between frames [7, 50]. The nonlocal neural network [57] and the long-term feature bank (LFB) [59] extended the 3D ConvNet by extracting long-range features. The SlowFast network [12] introduced a two-pathway network for learning motion and spatial features separately from the videos. The recent X3D [10] and AssembleNet [42, 41] build a searching network to find the best multi-branch architecture for the target dataset. The GTA [22] and HDC [64] proposed decoupled spatio-temporal attention networks for self-supervised video understanding.

**Multi-label Activity Recognition.** Multi-label activity recognition is designed for recognizing multiple activities that are performed simultaneously or sequentially in each video [44, 63, 39]. Most of the video networks focus on single-activity recognition and use sigmoid as the output activation function to provide multi-label predictions [7, 57, 12]. Other systems use weak-supervised learning and Zero-shot learning methods to help recognize weakly labeled activities [37, 56]. We introduce our network that improves the performance of multi-label activities at the feature-level by extracting independent feature descriptors for each activity instead of shared feature representations.

**Multi-label Image Classification.** Multi-label image classification has been studied for several decades. Previous approaches using graphical models to represent the label-wise dependency, such as Dependency Network [20], and Correlative Model [61]. These networks are not end-to-end trainable and require to input the entire dataset into the model, which is not generalizable to large-scale datasets. Others proposed end-to-end CNN-RNN architectures to model label-wise correlations [8, 53, 62]. These networks are still using shared features repeated for predicting each label, and RNN-based networks are too slow for training large-scale datasets, such as video datasets [31, 44].

## 3. Methodology

Recent state-of-the-art networks recognize activities by learning a shared feature representation for all the activities. These features are designed for recognizing single activities instead of multi-label activities, which requires to extract features for multiple activities in each video as well as the activity correlations. Our network focuses on multi-label activity recognition from a different angle: that of learning independent representations for each activity (activity-specific features) as well as their correlations (activity correlation map). Given a video clip  $V \in \mathbb{R}^{3 \times 32 \times 224 \times 224}$  with

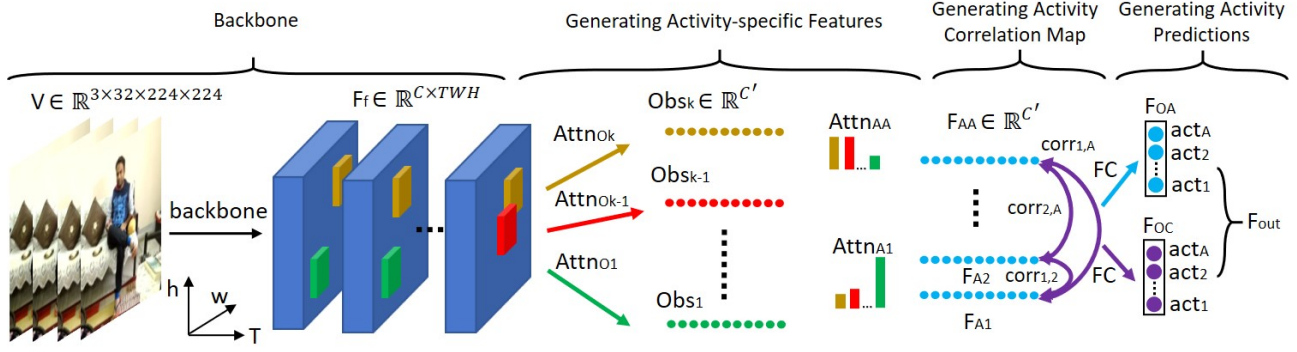


Figure 2: Method overview, showing the detail dimension transformation when generating activity-specific features and providing predictions. Attention (red, green, and brown) focus on different spatio-temporal regions of the backbone feature ( $F_f$ ) for generating observations ( $Obs$ ), and generating activity-specific features ( $F_A$ ) by combining observations using  $Attn_A$ . The purple lines ( $corr_{j,k}$ ) are the correlations between pairs of activity-specific features.

32 consecutive frames, our model provides activity predictions in three steps:

1. **Generating activity-specific features:** we generate independent feature representations for  $A$  different activities. This step consists of two sub-steps: we first generate  $K$  spatio-temporally independent feature snippets (observations),  $Obs \in \mathbb{R}^{K \times C'}$ , that focus on different spatio-temporal regions of the video. We then apply attention  $Attn_A$  on the observations to generate feature descriptors  $F_A$  (activity-specific features) that are independent for each activity using independent weighted combinations of observations (Figure 2, Generating Activity-specific Features).
2. **Generating activity correlation map:** we then generate an activity label-wise correlation map to represent the correlation between pairs of activities (Figure 2, Generating Activity Correlation Map).
3. **Generating activity predictions:** we finally provide multi-label activity predictions using their corresponding activity-specific features and the features of correlated activities (Figure 2, Generating Activity Predictions).

### 3.1. Generating Activity-specific Features

Given a feature set  $F_f \in \mathbb{R}^{C \times TWH}$  from the backbone network (e.g., i3D [7]), the activity-specific features can be generated as:

$$F_A = \{Attn_{A1}Obs, Attn_{A2}Obs, \dots, Attn_{AA}Obs\} \quad (1)$$

where  $F_A \in \mathbb{R}^{A \times C'}$  denotes  $A$  independent feature descriptors for their corresponding activities (activity-specific features),  $C'$  is the channel number of  $F_A$ ,  $Obs \in \mathbb{R}^{K \times C'}$  denotes  $K$  independent feature snippets (observations) that are extracted from the backbone features  $F_f$ , and  $Attn_{Ai}$  ( $i \in 1, 2, \dots, A$ ) are the attentions that independently combine the  $K$  observations to generate activity-specific features for the  $i^{th}$  activity. We create these observations in-

stead of directly generating  $Attn_A$  from the backbone features  $F_f$  to reduce redundant information. Each observation is an independent spatio-temporal feature snippet that focuses on a specific key region in a video. The  $Obs$  are generated by applying  $K$  independent spatio-temporal attentions on  $F_f$  as:

$$Obs_k = Attn_{Ok}[g_k^\alpha(F_f)]^T \quad (2)$$

$$Obs = \{Obs_1, Obs_2, \dots, Obs_k\} \quad (3)$$

where  $Obs_k \in \mathbb{R}^{C'}$  is the  $k^{th}$  observation that focuses on a specific key region of the video,  $Attn_{Ok} \in \mathbb{R}^{TWH}$  denotes the spatio-temporal attention for generating the  $k^{th}$  observation. The method for generating  $Attn_A$  and  $Attn_O$  will be introduced later. The  $g_k^\alpha$  is the linear function to integrate channels from  $F_f$ , which is represented as:

$$g_k^\alpha(F_f) = W_k^\alpha F_f \quad (4)$$

where  $W_k^\alpha \in \mathbb{R}^{C' \times C}$  are the weights for the linear function  $g_k^\alpha$ . The activity-specific set  $F_A$  can finally be written as:

$$F_A = \{Attn_{A1}\{Attn_{O1}[g_1^\alpha(F_f)]^T, \dots, Attn_{Ok}[g_k^\alpha(F_f)]^T\}, \\ Attn_{A2}\{Attn_{O1}[g_1^\alpha(F_f)]^T, \dots, Attn_{Ok}[g_k^\alpha(F_f)]^T\}, \\ \dots \\ Attn_{AA}\{Attn_{O1}[g_1^\alpha(F_f)]^T, \dots, Attn_{Ok}[g_k^\alpha(F_f)]^T\}\} \quad (5)$$

#### 3.1.1 Generating Attentions

The attention mechanism was introduced for capturing long-term dependencies within sequential inputs and is commonly used in natural language processing systems [52, 43]. We implemented the dot-product attention method [52] for generating  $Attn_O$  as:

$$Attn_{Ok} = softmax([g_k^\beta(F_f)_{-1}]^T g_k^\gamma(F_f)) \quad (6)$$

where  $Attn_{Ok} \in \mathbb{R}^{TWH}$  denotes the attention for the  $k^{th}$  observation,  $g_k^\beta$ ,  $g_k^\gamma$  are the linear functions same as the  $g_k^\alpha$  in equation 4, and  $g_k^\beta(F_f)_{-1} \in \mathbb{R}^{C' \times 1}$  denotes selecting the last row of the  $g_k^\beta(F_f) \in \mathbb{R}^{C' \times TWH}$  to make the produced attention  $Attn_{Ok} \in \mathbb{R}^{TWH}$ . The  $Attn_A$  was generated using a similar approach by replacing  $Obs$  with  $F_f$  in equation 6. Other attention methods (e.g., additive attention [3])

could be used for generating attentions, but we selected the dot-product attention method because previous research has shown that it is more efficient and works well for machine translation [52, 43].

Applying linear functions in equation (4) requires a large number of weights. Inspired by the channel-separated network [51], we applied a group linear function to reduce the number of weights for  $g_k^\alpha$  by splitting  $F_f$  into  $n$  groups based on the channel dimension and applying independent linear functions to each group as:

$$g_k^\alpha(F_f) = \text{Concat}(W_{k1}^\alpha F_{f1}, W_{k2}^\alpha F_{f2}, \dots, W_{kn}^\alpha F_{fn}) \quad (7)$$

where  $F_{fn} \in \mathbb{R}^{\frac{C}{n} \times TWH}$  is the  $n^{\text{th}}$  group of  $F_f$ , and  $W_{kn} \in \mathbb{R}^{\frac{C'}{n} \times \frac{C}{n}}$  are the weights of the linear function for  $F_{fn}$ . Using a larger number of groups ( $n$ ) results in fewer parameters ( $n$  times fewer). We set  $n = 32$  empirically to minimize the number of weights without affecting the model performance.

### 3.2. Generating Activity Correlation Maps

Label-wise correlations have proven important to be learned for multi-label classes [61]. Representing co-existing and exclusive patterns between activities is also necessary for recognizing multi-label activities. We apply an activity label-wise correlation map onto the activity-specific feature set  $F_A$  as:

$$F_{AC} = \text{Corr} F_A \quad (8)$$

where  $F_{AC} \in \mathbb{R}^{A \times C'}$  is the result of matrix multiplication between  $\text{Corr}$  and  $F_A$ . We applied the matrix multiplication between  $\text{Corr}$  and  $F_A$  to help recognize activities using their correlated activity-specific features.  $\text{Corr} \in \mathbb{R}^{A \times A}$  is the correlation map generated as:

$$\text{Corr} = \text{Attn}_C + \text{Mask} \quad (9)$$

$$\text{Mask}(j, k) = \frac{N_{j,k}}{N_j} \quad (10)$$

The correlation map  $\text{Corr}$  is composed from  $\text{Mask} \in \mathbb{R}^{A \times A}$  and  $\text{Attn}_C \in \mathbb{R}^{A \times A}$ ,  $\text{Mask}(j, k)$  is the frequency of activity  $j$  co-existing with activity  $k$  in the training ground truth,  $N_{i,j}$  is the number of samples that includes both activity  $j$  and  $k$ , and  $N_j$  is the number of samples in which activity  $j$  occurs.  $\text{Attn}_C$  is the attention generated from  $F_A$  to adjust the correlation map based on given different inputs.  $\text{Corr}$  is obtained by adding the  $\text{Mask}$  and the learned attention  $\text{Attn}_C$ . The model is able to adjust the correlation map if there is a mistake in  $\text{Mask}$ .  $\text{Attn}_C$  is generated using equation 6 by replacing  $F_{AC}$  with  $F_f$ .

### 3.3. Generating Activity Predictions

We finally make activity predictions using their corresponding activity-specific features and the features of correlated activities. To ensure that the  $F_A$  represents activity-specific features, without mixing features of other activities

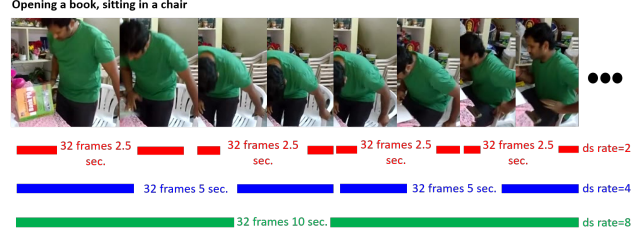


Figure 3: A video example in Charades shows multiple activities (opening a book and sitting in a chair) and the inputs of speed-invariant tuning that have different downsample rates (ds rate).

from the loss propagated by  $F_{AC}$ , we build a multi-output network that predicts activities from both  $F_A$  and  $F_{AC}$  as:

$$F_{OA} = \text{sigmoid}(W^\varphi F_A + b^\varphi) \quad (11)$$

$$F_{OC} = \text{sigmoid}(W^\theta F_{AC} + b^\theta) \quad (12)$$

$$F_{out} = [F_{OA}, F_{OC}] \quad (13)$$

where  $F_{out}$  is the final output of the network with  $F_{OA} \in \mathbb{R}^{A \times 1}$  and  $F_{OC} \in \mathbb{R}^{A \times 1}$ , two outputs for activity predictions.  $W^\varphi$ ,  $W^\theta$ ,  $b^\varphi$ , and  $b^\theta$  are the weights and bias for predicting activities from  $F_A$  and  $F_{AC}$ .

### 3.4. Comparison with Attention-based Networks

We compared our network with other attention-based networks to show two differences [16, 9, 34, 36, 16, 22, 64]. First, previous approaches apply attentions shared for all the channels in  $F_f$  and generate shared feature descriptors for all the activities. In our approach, observations are generated using independent attentions and focus on different key regions of the video. Second, attentions in previous methods only work through spatio-temporal dimensions, while our  $\text{Attn}_A$  works through observations to generate activity-specific features. In addition, our network also learns activity correlations, which is not included in the other attention-based methods.

### 3.5. Speed-invariant Tuning

In multi-label activity videos, different activities may have different duration in a large range compared with single activity videos. Figure 3 shows an example in Charades having two activities: opening a book (short) and sitting in a chair (long). The LFB method [59] proved that combining long-term features helps to improve the performance of recognizing multi-label activities. The LFB method requires extra FLOPs, and is not end-to-end trainable. We introduced our speed-invariant tuning method that ensures activities of different durations are properly covered in fixed frame inputs (32-frame) without requiring extra FLOPs.

**Training:** We first trained the complete model using the downsampling rate of 4 and froze the weights for all the 3D convolution layers. We then started finetuning the module after the 3D convolutions for  $I$  iterations by using 32-frame inputs obtained by randomly selecting downsampling rate  $r$



among 2, 4, and 8 (Figure 3, red, blue, and green segments). Using inputs with different downsampling rates makes the 32-frame-inputs able to cover video segments of different length. This training method enables the model to learn activities of different durations.

**Inference:** During the evaluation stage, we applied the 30-view test (followed [12]) by selecting views from different downsampling rates (12 from  $r = 4$ , and 9 from both  $r = 8$  and  $r = 2$ ) and summed the model predictions using inputs from the 30 views for the video-level activity prediction. The model can then recognize activities that have different duration by aggregating the results from branches that used different downsampling rates as the input.

We noticed that the previous multi-grid method [60] has a similar idea to speed up the model training process. Unlike the multi-grid method, we merged the predictions from multiple input speeds for the final prediction (without additional parameters and FLOPs) instead of making the model converge to a specific shape at later epochs. Our speed-invariant method had the performance improved based on the model trained with fixed-speed inputs, while the multi-grid method achieved similar performance as the baseline.

### 3.6. Implementation Details

We implemented our model with PyTorch [38]. We used batch normalization [28] and ReLU activation [21] for all the convolution layers. We used binary cross-entropy loss and the SGD optimizer with an initial learning rate  $3.5e - 2$  and  $1.25e - 5$  as the weight decay. Dropout (rate=0.5) was used after the dense layer to avoid overfitting [48]. We set the batch size to 12 and trained our model with 3 RTX 2080 Ti GPUs for 50k iterations. We applied the scale-jittering method in the range of [256, 320] and horizontal flipping to augment the frames [12]. During the evaluation, we used the 30-view test [12].

## 4. Experiments on Charades

Charades dataset [44] contains 9848 videos with average length of 30 seconds. This dataset includes 157 multi-label daily indoor activities. We used the officially provided train-validate split (7985/1863) to evaluate the network. We used the officially-provided 24-fps RGB frames as input and the officially-provided evaluation script for evaluating the validation set.

### 4.1. Single RGB Branch Results

We first compared our network with other state-of-art networks on Charades using a single RGB branch (only RGB videos as input and no ensemble). We compared our system with three baseline networks, CSN [51], pre-trained on IG-65M [15], Slowfast-101, pre-trained on Kinetics-600, and Nonlocal-101, pre-trained on Kinetics-400 as well as other state-of-the-art methods that work

Table 1: Comparison with other networks on Charades using a single RGB branch. The evaluation metric is mAP (mean-average-precision) in percentages, calculated using the officially provided script. The Slowfast\* and Nonlocal\* are the results of the baseline that we trained on Charades instead of the one reported in paper [12, 57].

method	backbone	pre-train	mAP
2D CNN [44]	Alexnet	ImageNet	11.2
MultiScale TRN [65]	Inception	ImageNet	25.2
I3D [7]	Inception	Kinetics-400	32.9
STRG [58]	Nonlocal-101	Kinetics-400	39.7
LFB [59]	Nonlocal-101	Kinetics-400	42.5
Multi-Grid [60]	SlowFast-50	Kinetics-400	38.2
Nonlocal [57]	Nonlocal-101	Kinetics-400	37.5
Nonlocal* (Baseline)	Nonlocal-101	Kinetics-400	40.3
Ours	Nonlocal-101	Kinetics-400	43.1
<b>Ours + speed-invariant</b>	Nonlocal-101	Kinetics-400	<b>44.2</b>
SlowFast [12]	SlowFast-101	Kinetics-600	45.2
X3D-XL [10]	X3D-XL	Kinetics-600	47.2
Slowfast* (Baseline)	SlowFast-101	Kinetics-600	44.7
Ours	SlowFast-101	Kinetics-600	46.8
<b>Ours + speed-invariant</b>	SlowFast-101	Kinetics-600	<b>48.1</b>
CSN [51] (Baseline)	CSN-152	IG-65M	46.4
Ours	CSN-152	IG-65M	49.2
<b>Ours + speed-invariant</b>	CSN-152	IG-65M	<b>50.3</b>

Table 2: Comparison with state-of-art networks on Charades using ensemble and additional inputs.

method	extra inputs	mAP	#backbones
Hallucinating [54]	IDT+Flow	43.1	1
AssembleNet ++ [17]	Flow+Object	59.9	4
Action-Genome [29]	Object+Person	60.1	1
CSN [51] (Baseline)	Object	63.1	1
<b>Ours</b>	Object	<b>65.5</b>	1

on Charades. Our method achieved around 4% higher mAP score on Charades over all the baseline networks pre-trained with different datasets. This shows that the generated activity-specific features and the activity label-wise correlations help the network to work better for multi-label activity recognition by plugging into any video networks. The model with speed-invariant tuning achieved around 1% mAP improvement on Charades by using all the backbone networks, which demonstrated that the speed-invariant tuning helps the model to learn activities with multiple duration. We also outperformed all the other methods [59, 25, 58, 12] including the recent state-of-the-art approach, X3D [10] that pre-trained on Kinetics-600 [6] on Charades. We note that learning activity correlations requires the ground truth of the dataset make sense in real-life. If the correlations in the ground truth of the dataset do not make sense in real-life, we can freeze the feature extractor as is, and learn the new activity correlations quickly in a small number of epochs for cross-dataset experiments.

### 4.2. Data Reinforced Model Results

Previous research showed that using additional inputs (e.g., flows and object information) improves the performance on Charades. The recent AssembleNet++ [42, 41] built a searching algorithm to fuse a 4-branch network that included optical flow and object information beside RGB

Table 3: Ablation experiments on the Charades dataset. We show the mAP scores, parameter numbers, and GFLOPs by using different hyper-parameters, backbone networks, and removing different modalities from our network.

group size	mAP	Params	GFLOPs
1	47.9	50.4M	$9.87 \times 30$
8	48.3	6.3M	$1.24 \times 30$
<b>32</b>	<b>48.4</b>	1.6M	$0.32 \times 30$
64	48.1	0.8M	$0.16 \times 30$

(a) Group size: performance on Charades using different group sizes by setting observation number to 64 and sample rate to 4.

obs num	mAP	Params	GFLOPs
16	46.1	0.4M	$0.08 \times 30$
32	47.5	0.8M	$0.16 \times 30$
<b>64</b>	<b>48.4</b>	1.6M	$0.32 \times 30$
128	48.4	3.2M	$0.62 \times 30$

(b) Observation number: performance on Charades using different number of observations by setting group size to 32 and sample rate to 4.

model structures	mAP
baseline	46.4
no activity-specific	46.6
no activity correlation	48.4
<b>complete</b>	<b>49.2</b>

(c) Model structure ablation: performance on Charades after removing each module.

sample rate	mAP	GFLOPs
4 only	49.2	$(\text{backbone} + 0.32) \times 30$
2 + 4	49.9	$(\text{backbone} + 0.32) \times (15 + 15)$
<b>2 + 4 + 8</b>	<b>50.3</b>	$(\text{backbone} + 0.32) \times (9 + 12 + 9)$

(d) Sample rate for speed-invariant tuning: performance and FLOPs of the model when using different sample rates by setting observation number to 64 and group size to 32.

method	mAP
Baseline	46.4
Temporal self-attention	46.3
Nonlocal block	46.7
Decoupled spatio-temporal attention	46.4
<b>Ours</b>	<b>48.4</b>

(e) Other attention-based models: compare the performance on Charades between our network and other attention-based models.

Table 4: Complexity analysis of our model. We compare the additional FLOPs and parameters that our model require to the other backbones.

method	Params	GFLOPs
Nonlocal-101 [44]	54.3 M	$359.0 \times 30$
Slowfast-101 [17]	59.9 M	$234.0 \times 30$
CSN-152 [51]	32.8 M	$109.0 \times 30$
X3D-XL [10]	11.0 M	$48.4 \times 30$
Our's-157 activities	1.6 M	$0.32 \times 30$
Our's-1024 activities	1.7 M	$0.33 \times 30$

frames as additional inputs. The Action-Genome builds a graph network to learn the relationship between objects and persons for activity prediction [29]. For a fair comparison, we evaluated our network on Charades by including the object information provided by [29] as inputs. Table 2 show that our network also outperformed the baseline and other state-of-the-art networks using additional inputs. These results demonstrate that the activity-specific features and activity correlations also help for multi-label activity recognition with additional inputs included.

### 4.3. Ablation Experiments

We next ablated our system with various hyper-parameters (group size, observation number, and sampling rate for speed-invariant inputs). We used the CSN-152 [51] pre-trained on IG-65M as backbone for all the experiments. **Group sizes.** Table 3a shows the system performance for using different group sizes ( $n$ ) in equation 7 when generating observations (64 observations and the downsampling rate is 4). The performance on Charades stayed at 48% mAP when the group size increased from 1 to 32 but dropped for  $n = 64$ . A larger group size results in using a smaller subset of channels from  $F_f$  for generating observations, which requires fewer parameters but may cause a performance drop because of information loss.

**Observations number.** Table 3b compares the system performance for different numbers of observations (group size is 32 and the downsampling rate is 4). The best-performing number of observations is 64, which also

requires the fewest weights. Using a larger number of observations helps cover more key parts from the videos but the performance saturates for more than 64 observations.

**Model structure ablation.** We next evaluated the system on Charades by removing each component from our network. Table 3c shows that the model without activity-specific features achieved similar performance as the baseline model although with the correlation map. Without learning activity-specific features, the system only learns correlations between different feature channels that have already have been learned in convolutions. The network with activity correlation achieved 1% mAP score improvement on Charades that helped to predict co-existed activities and eliminate exclusive activities. These results show that our network achieved a performance boost by the combination of key components of our network.

**Sampling rates for speed-invariant tuning.** We evaluated our speed-invariant tuning method by merging predictions using different downsampling rates at inputs. Table 3d shows that speed-invariant models achieved better performance compared to the model using a single downsampling rate of 4 because the speed-invariant tuning method makes the features better represent activities of different duration. The system achieved the best performance by merging predictions based on 2, 4, and 8 downsampling rates and this method did not require extra FLOPs by using the same view numbers (30 views) for evaluation (Table 3d row 3).

**Other attention-based models.** We then compared our network with other attention-based models including the models applying temporal attention, nonlocal block, and decoupled spatio-temporal attention (two layers) after the backbone to show the difference. The attentions are generated using equation 6 and the  $F_f$ . We then applied a 3D global average pooling after the attention layers that average the features over the spatio-temporal dimensions and a fully-connected layer to make predictions. Table 3e shows that all the two-stage transformer-based models did not increase the performance compared to the baseline

Table 5: Comparison with other networks on AVA. The evaluation metric is mean Average Precision using 0.5 IoU as threshold. The result of Slowfast is evaluated using official released parameters.

method	backbone	pre-train	mAP
ATR [30]	Nonlocal-50	Kinetics-400	21.7
LFB [59]	Nonlocal-101	Kinetics-400	27.7
Slowfast-101 [12]	Slowfast-101	Kinetics-600	29.1
X3D-XL [10]	X3D-XL	Kinetics-600	27.4
<b>Ours</b>	Slowfast-101	Kinetics-600	<b>30.2</b>
CSN-152 (Baseline)	CSN-152	IG-65M	27.9
<b>Ours</b>	CSN-152	IG-65M	<b>29.2</b>

network, while our network significantly outperformed the baseline network. The other attention-based models make all the channels share the same attention that also attempts to learn features shared by all the activities instead of activity-specific features.

**Complexity Analysis.** We finally calculated the additional complexity that our network required for learning activity-specific features. Table 4 shows that the additional parameters and FLOPs for our network are negligible compared to the complexity of the backbone networks. We also calculated the complexity of our network when the number of activities is 1024 (Table 4 last row). The complexity of our network is not linearly related to the activity number because of the design of observation generation to reduce redundant parameters. Our network only requires small amounts of additional computing resources based on the backbone networks even with a large number of activities.

## 5. Experiments on AVA

We next evaluated our network on another large-scale dataset, AVA [19], for multi-label activity detection. The AVA dataset includes 211k video clips for training and 57k clips for validating. Each clip has one keyframe with multiple persons localized using a bounding box and labeled with 80 multi-label activities. Following the official protocol [19], we evaluated the network on 60 activities using mean Average Precision (mAP) score and 0.5 IoU as the threshold. To make a fair comparison, we used the person bounding boxes generated by previous work and threshold the detected persons having confidence score  $> 0.8$  [12, 11]. We applied the ROI-Align algorithm [40] to extract the features of the persons using their corresponding bounding boxes. We set the spatial stride of the last stage of the backbone from 2 to 1 to increase the spatial resolution ( $14 \times 14$ ).

We compared our method with the baseline and other state-of-the-art networks on AVA in Table 5. We again outperformed the current state-of-the-art network (Slowfast-101, officially released parameters [11]) on AVA using center-crop views for evaluation. This demonstrated that our proposed activity-specific feature and activity correlation map generalizes well to different datasets and consistently improves multi-label classification problems by a sig-

Table 6: Experimental results on Volleyball. The “s” and “bb” in the last two columns denote using the whole scene and bounding boxes of persons as supplemental for recognizing group activities.

Volleyball Personal (multi-label)		Volleyball Group	
method	Acc.	Acc. (s)	Acc. (bb)
Hier LSTM [27]	72.7	63.1	81.9
SRNN [4]	76.6	-	83.4
So-Sce [2]	82.4	75.5	89.9
CRM [1]	-	75.9	93.0
Act-trans [14]	85.9	-	94.4
CSN-152 baseline	85.0	87.1	-
<b>Ours</b>	<b>86.6</b>	<b>87.6</b>	<b>95.5</b>

nificant margin.

## 6. Experiments on Hockey and Volleyball

We also run experiments on two small datasets, Hockey and Volleyball [27, 47]. The Volleyball Dataset contains 55 videos with 4830 annotated video clips. This dataset includes two sets of labels for group activity recognition task (8-class multi-class classification) and multi-label activity recognition task (9-label multi-label classification). We evaluated our method on both of these tasks. **The experimental results on Hockey is in the supplemental material because of the page limit.**

Table 6 shows that our system substantially outperformed all the existing approaches on Volleyball for multi-label activities [14]. We also compared our method with the baseline model using the latest backbone network (CSN baseline in Table 6) that works on activity recognition. Our system achieved roughly 2% higher accuracy score compared with the baseline model, which shows that the activity-specific features also improve multi-label activity recognition on small sports datasets.

We further evaluated our method on Volleyball for group activity recognition. The group activity is essentially a single-activity recognition problem: only one activity occurs during one video clip. Our method outperformed other state-of-the-art methods when using the whole scene (s, Table 6) as input [14] (RGB frames without using bounding boxes around people). This shows that our method generalizes to the single-activity recognition problem as well. Previous methods [1, 14] used bounding boxes around people (bb, Table 6) and their individual activities as supplemental information for group activity recognition. We tested our model by including this supplemental information, and our approach outperformed the recent state-of-the-art method [1] (95.5 for our system vs. 94.4 for the Act-trans in the last column of Table 6). Compared to the baseline network, our method slightly outperformed the baseline network (87.2 vs. 87.1 in the second-to-the-last column of Table 6). The activity-specific features do not help significantly in the single-activity problems, unlike the case of multi-label activities, because the feature maps will only focus on one region where the single-activity occurred.



(a) Visualization of the activity-specific features in video “2NXFV” from Charades.

(b) Visualization of the activity-specific features in video “6C0BK” from Charades.

(c) Visualization of the observation-56 in two videos from Charades.

Figure 4: Visualizing the activity-specific features and observations from Charades. The bounding boxes in the original frames of (a) and (b) correspond to the activated regions in the activity-specific feature maps.

## 7. Feature Visualization

To better understand what activity-specific features are learned, we visualized these features for the activities present in the video clips. Figure 4 shows two examples in the Charades, including the activity-specific feature maps (last two rows of each example in Figure 4) and their corresponding input frames. The activity-specific feature maps were generated by applying the learned  $Attn_O$  and  $Attn_A$  on the backbone features  $F_f$ . We normalized the feature maps between 0 and 1 and plotted these maps for the activities present in the video (last two rows of each example in Figure 4). To make the visualized maps more understandable, we applied the 0.5 threshold to the activity-specific feature maps and drew the bounding boxes using different colors (blue, red) for different activities in the original frames around the regions activated in the feature maps.

Based on the visualizations in Figure 4, we can make four points. First, unlike the backbone features (last row in Figure 4a and 4b) the activity-specific features will only focus on the spatial regions for the corresponding activity when multiple activities are performed simultaneously. The visualization of video “2NXFV” (Figure 4a) shows the activity-specific features #1 (holding a pillow) focusing the region of the left person, and activity-specific features #2 (eating something) focusing the right person who is performing the corresponding activity. Second, only the activity-specific features corresponding to the activity being performed will have high values when the video has one activity or activities performed in sequence. The visualization of video “6C0BK” (Figure 4b) shows the activity-specific features #1 (holding a blanket) having activated regions at the first two frames, while the activity-specific features #2 (playing with a phone) focused on the last frames. Third, all the activity-specific features will have low values for the frames in which no activities were performed (Figure 4, the third column of the right diagram). Finally, the same obser-

vation will focus on the same semantic region if it is present in both videos. Figure 4c shows the visualized observation 56 in two videos, the observation in both videos focused on the hands. These visualizations demonstrate that the observations focus on their specific semantic regions and activity-specific features will focus on the key regions from the video that are related to their corresponding activities.

## 8. Conclusion and Future Work

We introduced a network that focuses on multi-label activity recognition. The network generates spatio-temporally independent activity-specific features for each activity and learns activity correlations. We outperformed previous state-of-the-art methods on four multi-label activity recognition datasets. The visualizations showed that the activity-specific features are representative of their corresponding activities. One issue remains in the speed-invariant tuning method, where we simply summed the predictions by using different downsampling rates for the inputs. Extending the speed-invariant method to enable the model to learn to select features from appropriate speeds for different activities will be our future work.

**Acknowledgements.** This research has been funded under NIH/NLM grant 2R01LM011834-05 and NSF grant IIS-1763827. We thank Kaixiang Huang for insightful discussion.

## References

- [1] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2019.
- [2] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective



- activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4315–4324, 2017.
- [3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
  - [4] Sovan Biswas and Juergen Gall. Structural recurrent neural network (srnn) for group activity analysis. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1625–1632. IEEE, 2018.
  - [5] Marc-André Carboneau, Alexandre J Raymond, Eric Granger, and Ghyslaine Gagnon. Real-time visual play-break detection in sport events using a context descriptor. In *2015 IEEE International Symposium on Circuits and Systems (IS-CAS)*, pages 2808–2811. IEEE, 2015.
  - [6] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.
  - [7] 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.
  - [8] Shang-Fu Chen, Yi-Chen Chen, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Order-free rnn with visual attention for multi-label classification. *arXiv preprint arXiv:1707.05495*, 2017.
  - [9] Wenbin Du, Yali Wang, and Yu Qiao. Recurrent spatial-temporal attention network for action recognition in videos. *IEEE Transactions on Image Processing*, 27(3):1347–1360, 2017.
  - [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.
  - [11] Christoph Feichtenhofer and Haoqi Fan. Slowfast released code, <https://github.com/facebookresearch/slowfast>. 2019.
  - [12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019.
  - [13] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.
  - [14] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. *arXiv preprint arXiv:2003.12737*, 2020.
  - [15] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055, 2019.
  - [16] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *Advances in Neural Information Processing Systems*, pages 34–45, 2017.
  - [17] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 971–980, 2017.
  - [18] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The” something something” video database for learning and evaluating visual common sense. In *ICCV*, volume 1, page 3, 2017.
  - [19] 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.
  - [20] Yuhong Guo and Suicheng Gu. Multi-label classification using conditional dependency networks. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1300. Citeseer, 2011.
  - [21] Richard HR Hahnloser and H Sebastian Seung. Permitted and forbidden sets in symmetric threshold-linear networks. In *Advances in neural information processing systems*, pages 217–223, 2001.
  - [22] Bo He, Xitong Yang, Zuxuan Wu, Hao Chen, Ser-Nam Lim, and Abhinav Shrivastava. Gta: Global temporal attention for video action understanding. *arXiv preprint arXiv:2012.08510*, 2020.
  - [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
  - [24] Berthold KP Horn and Brian G Schunck. Determining optical flow. In *Techniques and Applications of Image Understanding*, volume 281, pages 319–331. International Society for Optics and Photonics, 1981.
  - [25] Nouredien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019.
  - [26] Dat Huynh and Ehsan Elhamifar. Interactive multi-label cnn learning with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9423–9432, 2020.
  - [27] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1980, 2016.
  - [28] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
  - [29] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020.

- [30] Jianwen Jiang, Yu Cao, Lin Song, Shiwei Zhang<sup>4</sup> 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.
- [31] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- [32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [33] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.
- [34] Dong Li, Ting Yao, Ling-Yu Duan, Tao Mei, and Yong Rui. Unified spatio-temporal attention networks for action recognition in videos. *IEEE Transactions on Multimedia*, 21(2):416–428, 2018.
- [35] Xinyu Li, Yanyi Zhang, Jianyu Zhang, Shuhong Chen, Ivan Marsic, Richard A Farneth, and Randall S Burd. Concurrent activity recognition with multimodal cnn-lstm structure. *arXiv preprint arXiv:1702.01638*, 2017.
- [36] Lili Meng, Bo Zhao, Bo Chang, Gao Huang, Wei Sun, Frederick Tung, and Leonid Sigal. Interpretable spatio-temporal attention for video action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [37] Ehsan Adeli Mosabbebi, Ricardo Cabral, Fernando De la Torre, and Mahmood Fathy. Multi-label discriminative weakly-supervised human activity recognition and localization. In *Asian conference on computer vision*, pages 241–258. Springer, 2014.
- [38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [39] Jamie Ray, Heng Wang, Du Tran, Yufei Wang, Matt Feiszli, Lorenzo Torresani, and Manohar Paluri. Scenes-objects-actions: A multi-task, multi-label video dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 635–651, 2018.
- [40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [41] Michael S Ryoo, AJ Piergiovanni, Juhana Kangasunta, and Anelia Angelova. Assemblenet++: Assembling modality representations via attention connections. *arXiv preprint arXiv:2008.08072*, 2020.
- [42] Michael S Ryoo, AJ Piergiovanni, Mingxing Tan, and Anelia Angelova. Assemblenet: Searching for multi-stream neural connectivity in video architectures. *arXiv preprint arXiv:1905.13209*, 2019.
- [43] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. Disan: Directional self-attention network for rnn/cnn-free language understanding. *arXiv preprint arXiv:1709.04696*, 2017.
- [44] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016.
- [45] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.
- [46] 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.
- [47] Konstantin Sozykin, Stanislav Protasov, Adil Khan, Rasheed Hussain, and Jooyoung Lee. Multi-label class-imbalanced action recognition in hockey videos via 3d convolutional neural networks. In *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 146–151. IEEE, 2018.
- [48] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- [49] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [50] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
- [51] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5552–5561, 2019.
- [52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [53] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2016.
- [54] Lei Wang, Piotr Koniusz, and Du Huynh. Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns. In *Proceedings of the 2019 International Conference on Computer Vision*. IEEE, Institute of Electrical and Electronics Engineers, 2019.
- [55] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition.

- In *European conference on computer vision*, pages 20–36. Springer, 2016.
- [56] Qian Wang and Ke Chen. Multi-label zero-shot human action recognition via joint latent ranking embedding. *Neural Networks*, 122:1–23, 2020.
  - [57] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.
  - [58] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 399–417, 2018.
  - [59] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.
  - [60] Chao-Yuan Wu, Ross Girshick, Kaiming He, Christoph Feichtenhofer, and Philipp Krahenbuhl. A multigrid method for efficiently training video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 153–162, 2020.
  - [61] Xiangyang Xue, Wei Zhang, Jie Zhang, Bin Wu, Jianping Fan, and Yao Lu. Correlative multi-label multi-instance image annotation. In *2011 International Conference on Computer Vision*, pages 651–658. IEEE, 2011.
  - [62] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartłomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2020.
  - [63] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2-4):375–389, 2018.
  - [64] Zehua Zhang and David Crandall. Hierarchically decoupled spatial-temporal contrast for self-supervised video representation learning. *arXiv preprint arXiv:2011.11261*, 2020.
  - [65] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018.