Interpretable Spatio-temporal Attention for Video Action Recognition

Lili Meng, Bo Zhao, Bo Chang
University of British Columbia
{menglili@cs, bzhao03@cs, bchang@stat}.ubc.ca

Gao Huang
Tsinghua University
gaohuang@tsinghua.edu.cn

Wei Sun, Frederich Tung, Leonid Sigal
University of British Columbia
{wei.sun@alumni, ftung@cs, lsigal@cs}.ubc.ca

Abstract

Inspired by the observation that humans are able to process videos efficiently by only paying attention where *and* when *it is needed, we propose an interpretable and easy plug-in spatial-temporal attention mechanism for video action recognition. For spatial attention, we learn a saliency mask to allow the model to focus on the most salient parts of the feature maps. For temporal attention, we employ a convolutional LSTM based attention mechanism to identify the most relevant frames from an input video. Further, we propose a set of regularizers to ensure that our attention mechanism attends to coherent regions in space and time. Our model not only improves video action recognition accuracy, but also localizes discriminative regions both spatially and temporally, despite being trained in a weakly-supervised manner with only classification labels (no bounding box labels or time frame temporal labels). We evaluate our approach on several public video action recognition datasets with ablation studies. Furthermore, we quantitatively and qualitatively evaluate our model's ability to localize discriminative regions spatially and critical frames temporally. Experimental results demonstrate the efficacy of our approach, showing superior or comparable accuracy with the state-of-the-art methods while increasing model interpretability.*

1. Introduction

An important property of human perception is that one does not need to process a scene in its entirety at once. Instead, humans focus attention selectively on parts of the visual space to acquire information *where* and *when* it is needed, and combine information from different fixations over time to build up an internal representation of the scene [23], which can be used for interpretation or prediction.

In computer vision and natural language processing, over the last couple of years, attention models have proved important, particularly for tasks where interpretation or explanation requires only a small portion of the image, video or sentence. Examples include visual question answering [19, 43, 45], activity recognition [6, 18, 26], and neural machine translation [1, 38]. These models have also provided certain interpretability, by visualizing regions selected or attended over for a particular task or decision. In particular, for video action recognition, a proper attention model can help answer the question of not only *where* but also *when* it needs to look at the image evidence to draw a decision. It intuitively explains which part the model attends to and provides easy-to-interpret rationales when making a particular decision. Such interpretability is helpful and even necessary for some real applications, *e.g.*, medical AI systems [52] or self-driving cars [14].

Visual attention for action recognition in videos is challenging as videos contain both spatial and temporal information. Previous work in attention-based video action recognition either focused on spatial attention [6, 18, 26] or temporal attention [35]. Intuitively, both spatial attention and temporal attention should be integrated together to help make and explain the final action prediction. In this paper, we propose a novel spatio-temporal attention mechanism that is designed to address these challenges. Our attention mechanism is efficient, due to its space- and time- separability, and yet flexible enough to enable encoding of effective regularizers (or priors). As such, our attention mechanism consists of spatial and temporal components shown in Fig. 1. The spatial attention component, which attenuates frame-wise CNN image features, consists of the saliency mask, regularized to be discriminative and spatially smooth. The temporal component consists of a uni-modal soft attention mechanism that aggregates information over the nearby attenuated frame features before passing it into a convolutional LSTM for class prediction.

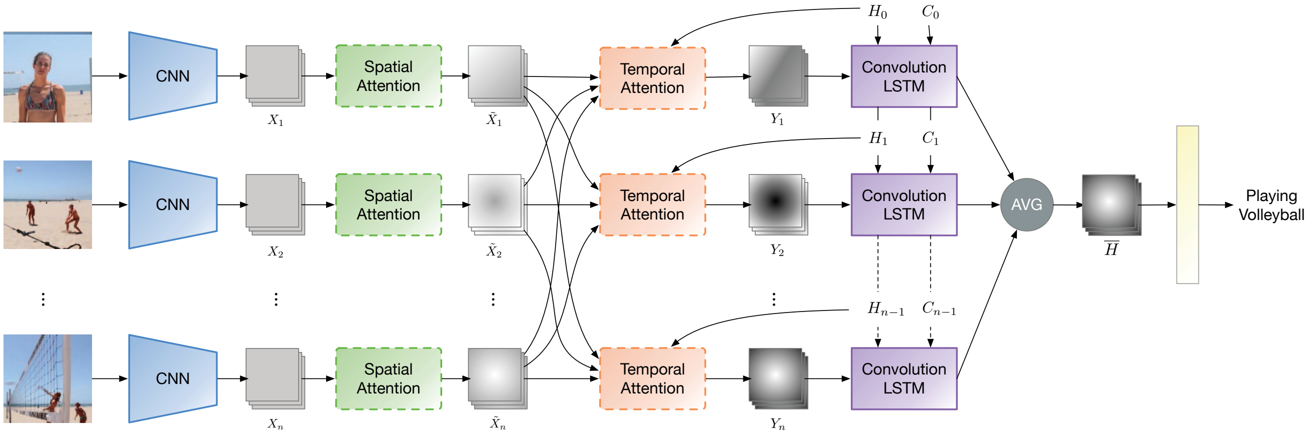

Figure 1. **Spatio-temporal attention for video action recognition.** The convolutional features are attended over both spatially, in each frame, and subsequently temporally. Both attentions are soft, meaning that the effective final representation at time t of an RNN, used to make the prediction, is a spatio-temporally weighted aggregation of convolutional features across the video along with the past hidden state from $t-1$. For details please refer to Sec. 3.

Contributions: In summary, the main contributions of this work are: (1) We introduce an interpretable and easy plug-in spatial-temporal attention for video action recognition, which consists of the saliency mask for spatial attention learned by ConvNets and temporal attention learned by convolutional LSTM. (2) We introduce three different regularizers, two for spatial and one for temporal attention components, to improve performance and interpretability of our model; (3) We demonstrate the efficacy of our model for video action recognition on three public datasets and explore the importance of our modeling choices through ablation experiments; (4) Finally, we qualitatively and quantitatively show that our spatio-temporal attention is able to localize discriminative regions and important frames, despite being trained in a purely weakly-supervised manner with only classification labels.

2. Related work

2.1. Visual explanations of neural networks

Various methods have been proposed to provide explanations of neural networks by visualization [20, 22, 29, 34, 49, 50, 51], including visualizing the gradients, perturbing the inputs, and bridging relations with other well-studied systems. The class activation mapping (CAM) applies the global average pooling layer for identifying discriminative regions [55]. The gradient-weighted class activation mapping (Grad-CAM) utilizes the gradient flow and relaxes the architectural assumptions made by CAM [25]. The guided attention inference network (GAIN) makes the network's attention trainable in an end-to-end fashion [17]. Visual attention is one way to explain which part of the image is responsible for the network's decision [12, 14]. A visual attention model is used to train a ConvNet regressor from images to steering angle for self-driving cars, and image regions that potentially influence the network's output are highlighted by the model [14]. A semantically and visually interpretable medical image diagnosis network is proposed to explore discriminative image feature descriptions from reports [52]. However, few work focus on visual interpretation both spatially and temporally.

2.2. Visual attention for video action recognition

Video action recognition is one of the fundamental problems in computer vision and has been widely studied [2, 5, 7, 36, 30, 37]. A recent survey can be found in [15], so here we only focus on attention-based models [4, 6, 18, 26, 32, 35]. For video action recognition, visualizing which part of the frame and which frame of the video sequence that the model was attending to provides valuable insight into the model's behavior. An attention-driven LSTM [26] is proposed for action recognition and it can highlight important spatial locations. Attentional pooling [6] introduces an attention mechanism based on a derivation of bottom-up and top-down attention as low-rank approximations of bilinear pooling methods. However, these work only focus on the crucial spatial locations of each image, without considering temporal relations among different frames in a video sequence. To alleviate this shortcoming, visual attention is incorporated in the motion stream [4, 18, 41]. However, the motion stream only employs the optical flow frames generated from consecutive frames, and cannot consider the long-term temporal relations among different frames in a video sequence. Moreover, motion stream needs additional optical flow frames as input, which imposes burden due to additional optical flow extraction, storage and computation and is especially severe for large datasets. [35] proposes an attention based LSTM model to highlight frames in videos, but spatial information is not used for temporal attention. An end-to-end spatial and temporal attention model is proposed in [32] for human action recognition, but additional skeleton data is needed.

2.3. Spatial and temporal action localization

Spatial and temporal action localization have been long studied and here we constrain the discussion to some recent learning based methods [28, 31, 39, 42, 44, 47, 48]. For spatial localization, [42] proposes to detect proposals at the frame-level and scores them with a combination of static and motion CNN features, and then tracks high-scoring proposals throughout the video using a tracking-by-detection approach. However, these models are trained in a supervised manner in which bounding box annotations are required. For temporal localization, [47] employs REINFORCE to learn the agent's decision policy for action detection. UntrimmedNet [39] couples the classification and the selection module to learn the action models and reason about the action temporal duration in a weakly supervised manner. However, the above mentioned methods consider spatial and temporal localization separately. A spatio-temporal action localization is proposed to localize both spatially and temporally in [31], but both spatial and temporal labels are required. It is very expensive and time-consuming to acquire these per-frame labels in large-scale video datasets.

Our proposed method can localize discriminative regions both spatially and temporally besides improving action recognition accuracy, despite being trained in a weakly-supervised fashion with only classification labels.

3. Spatial-temporal attention mechanism

Our overall model is an Recurrent Neural Network (RNN) that aggregates frame-based convolutional features across the video to make action predictions as shown in Fig. 1. The convolutional features are attended over both spatially, in each frame, and subsequently temporally for the entire video sequence. Both attentions are soft, meaning that the effective final representation at time t is a spatio-temporally weighted aggregation of convolutional features across the video along with the past hidden state from $t-1$. The core novelty is the overall form of our spatio-temporal attention mechanism and the additional terms of the loss function that induce sensible spatial and temporal attention priors.

3.1. Convolutional frame features

We use the last convolutional layer output extracted by ResNet50 or ResNet101 [8], pretrained on the ImageNet [3] dataset and fine-tuned for the target dataset, as our frame feature representation. We acknowledge that more accurate feature extractors (for instance, network with more parameters such as ResNet-152 or higher performance networks such as DenseNet [10] or SENet [9]) and optical flow features will likely lead to better overall performance. Our primary purpose in this paper is to prove the efficacy of our spatial-temporal attention mechanism. Hence we keep the features relatively simple.

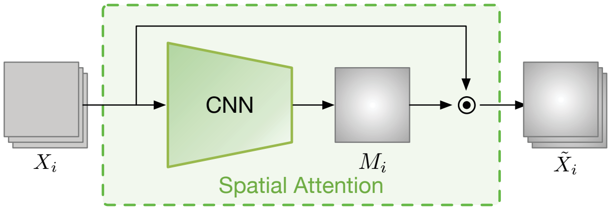

Figure 2. **Spatial attention component.** We use several layers of convolutional network to learn the importance mask $\boldsymbol{M}_i$ for the input image feature $\boldsymbol{X}_i$, the output is the element-wise multiplication $\tilde{\boldsymbol{X}}_i = \boldsymbol{X}_i \odot \boldsymbol{M}_i$. Details please refer to Sec. 3.2.

3.2. Spatial attention with importance mask

We apply an *importance mask* $\boldsymbol{M}_i$ to the image feature $\boldsymbol{X}_i$ of the i-th frame to obtain attended image features by element-wise multiplication:

$$\tilde{\boldsymbol{X}}_i = \boldsymbol{X}_i \odot \boldsymbol{M}_i, \tag{1}$$

for $1 \leq i \leq n$, where n is the number of frames, and each entry of $\boldsymbol{M}$ lies in $[0, 1]$. This operation attenuates certain regions of the feature map based on their estimated importance. Here we simply use three convolutional layers to learn the importance mask. Fig. 2 illustrates our spatial attention component and Table 1 shows the network architecture details. However, if left unconstrained, an arbitrarily structured mask could be learned, leading to possible overfitting. We posit that, in practice, it is often useful to attend to a few important larger regions (*e.g.*, objects, elements of the scene). To induce this behavior, we encourage smoothness of the mask by introducing total variation loss on the spatial attention, as will be described in Sec. 3.4.

3.3. Temporal attention based on ConvLSTM

We introduce the temporal attention mechanism inspired by attention for neural machine translation [1]. However,

Index	Operation
(1)	$\mathbf{X}_i$
(2)	CONV-(N1024, K3, S1, P1), BN, ReLU
(3)	CONV-(N512, K3, S1, P1), BN, ReLU
(4)	CONV-(N1, K3, S1, P1), Sigmoid

Table 1. **Architecture of our spatial attention module** (N: number of channels, K: kernel size, S: stride, P: padding).

conventional LSTM in neural machine translation uses full connections in the input-to-state and state-to-state transitions and cannot encode the spatial information in images. Therefore, to mitigate this drawback, we use Convolutional LSTM (ConvLSTM) [27] instead. For ConvLSTM, each input, cell output, hidden state, and gate are 3D tensors whose last two dimensions are spatial dimensions. These 3D tensors can preserve spatial information, which is more suitable for image inputs.

Our temporal attention mechanism generates *energy* for each attended frame $\tilde{\boldsymbol{X}}_i$ at each time step t,

$$e_{ti} = \Phi(\boldsymbol{H}_{t-1}, \tilde{\boldsymbol{X}}_i), \tag{2}$$

where $\boldsymbol{H}_{t-1}$ represents the ConvLSTM hidden state at time $t-1$ that implicitly contains all previous information up to time step $t-1$, $\tilde{\boldsymbol{X}}_i$ represents the i-th frame masked features, and $\Phi(\boldsymbol{H}_{t-1}, \tilde{\boldsymbol{X}}_i) = \Phi_H(\boldsymbol{H}_{t-1}) + \Phi_X(\tilde{\boldsymbol{X}}_i)$, where Φ_H and Φ_X are feed-forward neural networks which are jointly trained with all other components of the proposed system.

This temporal attention model directly computes a soft attention weight for each frame at each time t as shown in Fig. 3. It allows the gradient of the cost function to be backpropagated through. This gradient can be used to train the entire spatial-temporal attention model jointly.

The importance weight w_{ti} for each frame is:

$$w_{ti} = \frac{\exp(e_{ti})}{\sum_{i=1}^{n}(\exp(e_{ti}))}, \tag{3}$$

for $1 \leq i \leq n, 1 \leq t \leq n$. This importance weighting mechanism decides which frame of the video to pay attention to. The final feature map $\boldsymbol{Y}_t$ that is input to the ConvLSTM is a weighted sum of the features from all of the frames:

$$\boldsymbol{Y}_t = \frac{1}{n}\sum_{i=1}^{n} w_{ti}\tilde{\boldsymbol{X}}_i, \tag{4}$$

where $\tilde{\boldsymbol{X}}_i$ denotes the i-th masked frame of each video.

We use the following initialization strategy [46] for the ConvLSTM cell state and hidden state for faster convergence:

$$\boldsymbol{C}_0 = g_c(\frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{X}}_i), \qquad \boldsymbol{H}_0 = g_h(\frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{X}}_i), \tag{5}$$

where g_c and g_h are two layer convolutional networks with batch normalization [11].

We calculate the average hidden states of ConvLSTM over time length n, $\overline{\boldsymbol{H}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i$ and send it to a fully connected classification layer for the final video action classification.

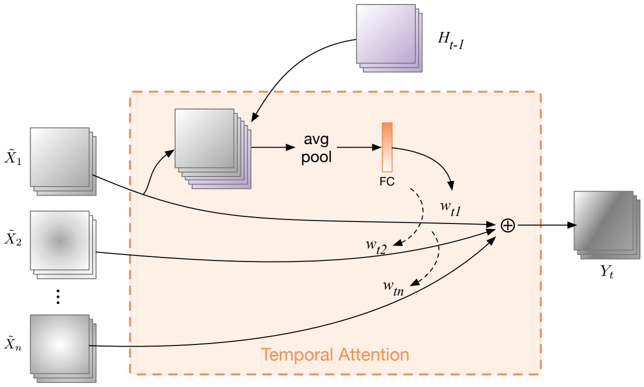

Figure 3. **Temporal attention component.** The temporal attention learns a temporal attention weight w_{ti} at each time step t. The final feature map $\boldsymbol{Y}_t$ at time t to the ConvLSTM is a weighted sum of the feature from all the previous masked frames. For details please refer to Sec. 3.3.

3.4. Loss function

Considering the spatial and temporal nature of our video action recognition, we would like to learn (1) a sensible attention mask for spatial attention, (2) reasonable importance weighting scores for different frames, and (3) improve the action recognition accuracy at the same time. Therefore, we define our loss function L as:

$$L = L_{\text{CE}} + \lambda_{\text{TV}} L_{\text{TV}} + \lambda_{\text{contrast}} L_{\text{contrast}} + \lambda_{\text{unimodal}} L_{\text{unimodal}}, \tag{6}$$

where L_{CE} is the cross entropy loss for classification, L_{TV} represents the total variation regularization [24]; L_{contrast} represents the mask and background contrast regularizer; and L_{unimodal} represents unimodality regularizer. λ_{TV}, $\lambda_{\text{contrast}}$ and $\lambda_{\text{unimodal}}$ are the weights for the corresponding regularizers.

The total variation regularization L_{TV} of the learnable attention mask encourages spatial smoothness of the mask and is defined as:

$$L_{\text{TV}} = \sum_{i=1}^{n}\left(\sum_{j,k}|\boldsymbol{M}_i^{j+1,k} - \boldsymbol{M}_i^{j,k}| + \sum_{j,k}|\boldsymbol{M}_i^{j,k+1} - \boldsymbol{M}_i^{j,k}|\right) \tag{7}$$

where $\boldsymbol{M}_i$ is the mask for the i-th frame, and $\boldsymbol{M}_i^{j,k}$ is the entry at the (j,k)-th spatial location of the mask. The contrast regularization L_{contrast} of learnable attention mask is used to suppress the irrelevant information and highlight important parts:

$$L_{\text{contrast}} = \sum_{i=1}^{n}\left(-\frac{1}{2}\boldsymbol{M}_i \odot \boldsymbol{B}_i + \frac{1}{2}\boldsymbol{M}_i \odot (1 - \boldsymbol{B}_i)\right) \tag{8}$$

where $\boldsymbol{B}_i = \mathbb{I}\{\boldsymbol{M}_i > 0.5\}$ represents the binarized mask, $\mathbb{I}$ is an indicator function applied element-wise.

The unimodality regularizer L_{unimodal} encourages the temporal attention weights to be unimodal, biasing against spurious temporal weights. This stems from our observation that in most cases only one activity would be present in the considered frame window, with possible irrelevant information on either or both sides. Here we use the log concave distribution to encourage the unimodal pattern of temporal attention weights:

$$L_{\text{unimodal}} = \sum_{t=1}^{n} \sum_{i=2}^{n-1} \max\{0, w_{t,i-1} w_{t,i+1} - w_{t,i}^2\} \quad (9)$$

3.5. Log-concave regularization

A probability distribution is unimodal if it has a single peak or mode. The temporal attention weights are a univariate discrete distribution over the frames, indicating the importance of the frames for the task of classification. In the context of activity recognition, it is reasonable to assume that the frames that contain salient information should be consecutive, instead of scattered around. Therefore, we introduce a mathematical concept called the log-concave sequence and design a regularizer that encourages unimodality. We first give a formal definition of the unimodal sequence.

Definition 1. *A sequence* $\{a_i\}_{i=1}^n$ *is unimodal if for some integer m,*

$$\begin{cases} a_{i-1} \le a_i & \text{if } i \le m, \\ a_i \ge a_{i+1} & \text{if } i \ge m. \end{cases}$$

A univariate discrete distribution is unimodal, if its probability mass function forms a unimodal sequence. The log-concave sequence is defined as follows.

Definition 2. *A non-negative sequence* $\{a_i\}_{i=1}^n$ *is log-concave if* $a_i^2 \ge a_{i-1} a_{i+1}$.

This property gets its name from the fact that if $\{a_i\}_{i=1}^n$ is log-concave, then the sequence $\{\log a_k\}_{i=1}^n$ is concave. The connection between unimodality and log-concavity is given by the following proposition.

Proposition 1. *A log-concave sequence is unimodal.*

The proof of this proposition is given in the supplementary material. Given the definition of log-concavity, it is straightforward to design a regularization term that encourages log-concavity:

$$R = \sum_{i=2}^{n-1} \max\{0, a_{i-1} a_{i+1} - a_i^2\}. \quad (10)$$

By Proposition 1, this regularizer also encourages unimodality.

4. Experiments

In this section, we first conduct experiments to evaluate our proposed method on video action recognition tasks on three publicly available datasets. We then evaluate our spatial attention mechanism on the spatial localization task and our temporal attention mechanism on the temporal localization task respectively.

4.1. Video action recognition

We first conduct extensive studies on the widely used HMDB51 [16] and UCF101 [33] datasets. The purpose of these experiments is mainly to examine the effects of different sub-components. We then show that our method can be applied to the challenging large-scale Moments in Time dataset [21].

Datasets. HMDB51 dataset [16] contains 51 distinct action categories, each containing at least 101 clips for a total of 6,766 video clips extracted from a wide range of sources. These videos include general facial actions, general body movements, body movements with object interaction, and body movements for human interaction.

UCF101 dataset [33] is an action recognition dataset of realistic action videos, collected from YouTube, with 101 action categories.

Moments in Time Dataset [21] is a collection of one million short videos with one action label per video and 339 different action classes. As there could be more than one action taking place in a video, action recognition models may predict an action correctly yet be penalized because the ground truth does not include that action. Therefore, it is believed that top 5 accuracy measure will be more meaningful for this dataset.

Baselines. For HMDB51 and UCF101, we use three attention-based methods (Visual attention [26], VideoLSTM [18], and Attentional Pooling [6]) and our own method without attention (ResNet101-ImageNet) as baselines. For the Moments in Time dataset, as there are no attention-based methods results available, we use ResNet50-ImageNet [21], TSN-Spatial [40] and TRN-Multiscale [54] as our baselines.

Experimental setup. We use the same parameters for HMDB51 and UCF101: a single convolutional LSTM layer with 512 hidden-state dimensions, sequence length $n = 25$, $\lambda_{\text{TV}} = 10^{-5}$, $\lambda_{\text{contrast}} = 10^{-4}$, and $\lambda_{\text{unimodal}} = 1$. For the Moments in Time dataset, we use time sequence length $n = 15$. The dataset pre-processing and data augmentation are the same as the ResNet ImageNet experiment [8]. For more details on the experimental setup, please refer to the supplementary material.

Qualitative results. We visualize the spatial attention and temporal attention results in Fig. 4. The spatial attention can

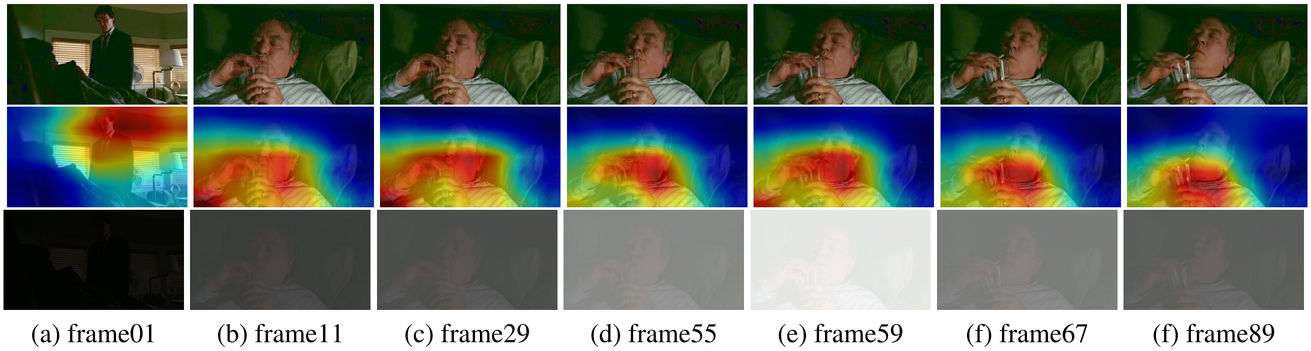

Figure 4. **Examples of spatial temporal attention.** (Best viewed in color.) A frame sequence from a video of *Drink* action in HMDB51. The original images are shown in the top row, spatial attention is shown as heatmap (red means important) in the middle row, and temporal attention score is shown as the gray image (the brighter the frame is, the more crucial the frame is) in the bottom row. It shows that spatial attention can focus on important areas while temporal attention can attend to crucial frames. The temporal attention also shows a unimodal distribution for the entire action from starting to drink to completing the action.

correctly focus on the important spatial areas of the image, and the temporal attention shows a unimodal distribution for the entire action from starting frame to final frame. More results are shown in the supplementary material.

Quantitative results. We show the top-1 video action classification accuracy for HMDB51 and UCF101 datasets in Table 2. Our proposed model outperforms previous attention based models [6, 18, 26] and ResNet101-ImageNet with the same input. The ablation experiments demonstrate that all the sub-components of the proposed method contribute to improving the final performance. The results on the Moments in Time dataset are reported in Table 3. Our method achieves the best accuracy comparing to other methods using just RGB modality input. Furthermore, our spatial-temporal attention mechanism is an easy plug-in model which could be based on different architectures. As shown in Table 4, our spatial-temporal attention mechanism can boost performance for different base networks. It also indicates that base networks are also crucial for the final performance as higher-capacity base networks can provide better image features as input to our model.

4.2. Weakly supervised action localization

Due to the existence of spatial and temporal attention mechanisms, our model can not only classify the action of the video, but also give a better interpretability of the results, *i.e.* telling which region and frames contribute more to the prediction. In other words, our proposed model can also localize the most discriminant region and frames at the same time. To verify this, we conduct the spatial localization and temporal localization experiments.

4.2.1 Spatial action localization

The spatial localization is designed to detect the most important region in the image. With the explicit attention mask prediction in our spatial attention module, the region with higher value in the mask is considered naturally more important.

Dataset. UCF101-24 is a subset of 24 classes out of 101 classes of UCF101 [33] that comes with spatio-temporal localization annotation, in the form of bounding box annotations for humans, with THUMOS2013 and THUMOS2014 challenges [13].

Baselines and error metric. We use supervised methods Fast action proposal [48] and Learning to track [42]

Model	HMDB51	UCF101
Visual attention [26]	41.31	84.96
VideoLSTM [18]	43.30	79.60
Attentional Pooling [6]	50.80	–
ResNet101-ImageNet	50.04	83.30
Ours	**53.07**	**87.11**
Ablation Experiments		
Ours w/o spatial attention	51.98	85.78
Ours w/o temporal attention	52.25	85.86
Ours w/o L_{TV}	52.01	85.89
Ours w/o L_{contrast}	52.10	85.98
Ours w/o L_{unimodal}	52.05	86.10

Table 2. **Top-1 accuracy (%) on HMDB51 and UCF101 dataset.**

Model	**Top-1** (%)	**Top-5** (%)
ResNet50-ImageNet [21]	26.98	51.74
TSN-Spatial [40]	24.11	49.10
TRN-Multiscale [54]	27.20	53.05
Ours	**27.86**	**53.52**

Table 3. **Results on Moments in Time dataset with RGB modality.** ResNet50-ImageNet and TRN-Multiscale spatial results reported here are based on authors' publicly released trained model.

Base network	ResNet50	ResNet101	ResNet152
w/o attention	47.5	50.0	50.4
w attention	**49.8**	**53.1**	**54.4**

Table 4. **Top-1 accuracy (%) on HMDB51 of our spatio-temporal attention mechanism with different base networks.** It shows our spatial-temporal attention mechanism can boost performance for different base networks. It also indicates base networks plays a vital role in the final performance as high-capacity base model can provide more accurate image features.

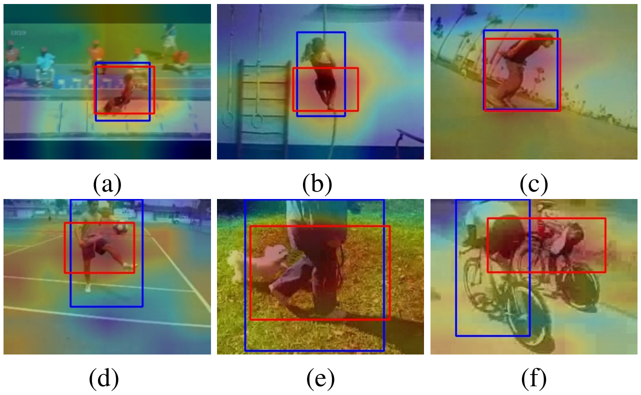
(a) (b) (c) (d) (e) (f)

Figure 5. **Examples of spatial attention for action localization.** (Best viewed in color.) Blue bounding boxes represent ground truth while the red ones are predictions from our learned spatial attention. (a) long jump, (b) rope climbing, (c) skate boarding, (d) soccer juggling, (e) walking with dog, (f) biking.

as our baselines as currently there are no quantitative results available for weakly supervised setting. Fast action proposal [48] is formulated as a maximum set coverage problem and solved by a greedy-based method, and can capture spatial-temporal video tube of high potential to localize human action. Learning to track [42] first detects frame-level proposals and scores them with a combination of static and motion CNN features. It then tracks high-scoring proposals throughout the video using a tracking-by-detection approach. Note that these supervised spatial localization methods require bounding box annotations, while our method only needs classification labels.

We use Intersection Over Union (IoU) of the predicted and ground truth bounding boxes as our error metric.

Experimental setup. For training, we only use the classification labels without spatial bounding box annotations. For evaluation, we threshold the produced saliency mask at 0.5 and the tightest bounding box that contains the thresholded saliency map is set as the predicted localization box for each frame. The predicted localization boxes are compared with the ground truth bounding boxes at different IoU levels.

Qualitative results. We show some qualitative results in Fig. 5. Our spatial attention can attend to important action areas. The ground truth bounding boxes include all the human actions, while our attention could attend to crucial parts of an action such as in Fig. 5(d) and (e). Furthermore, our attention mechanism is able to attend to areas with multiple human actions. For instance, in Fig. 5(f) the ground truth only includes one person bicycling, but our attention can include both people bicycling. More qualitative results including failure cases are included in the supplementary material.

Quantitative results. Table 5 shows the quantitative results for UCF101-24 spatial localization results. Our attention mechanism works better compared with the baseline methods when the IoU threshold is lower mainly because our model only focuses on important spatial areas rather than the entire human action annotated by bounding boxes. Compared with the baseline methods trained with ground truth bounding boxes, we only use the action classification labels, no ground truth bounding boxes are used.

4.2.2 Temporal action localization

The temporal localization is designed to find the start and end frame of the action in the video. Intuitively, the weight of each frame predicted by the temporal attention module in our model indicates the importance of each frame.

Dataset. The action detection task of THUMOS14 [13] consists of 20 classes of sports activities, and contains 2765 trimmed videos for training, and 200 and 213 untrimmed videos for validation and test, respectively. Following the standard practice [47, 53], we use the validation set as training and evaluate on the testing set. To avoid the training ambiguity, we remove the videos with multiple labels. We extract RGB frames from the raw videos at 10 fps.

Baselines and error metric. We compare our method with a reinforcement learning based method REINFORCE [47] and a weakly supervised method UntrimmedNets [39] for temporal action localization. [47] formulates the model as a recurrent neural network-based agent that interacts with a video over time, and use REINFORCE to learn the agent's decision policy. UntrimmedNets [39] couple both the classification module and the selection module to learn the action models and reason about the temporal duration of action instances.

Intersection Over Union (IoU) of the predicted and ground truth temporal labels are used as our error metric.

Experimental setup. We use the same hyperparameters for THUMOS14 as HMDB51, UCF101 and UCF101-24. For training, we only use the classification labels without temporal annotation labels. For evaluation, we threshold the normalized temporal attention importance weight at 0.5. These predicted temporal localization frames are then

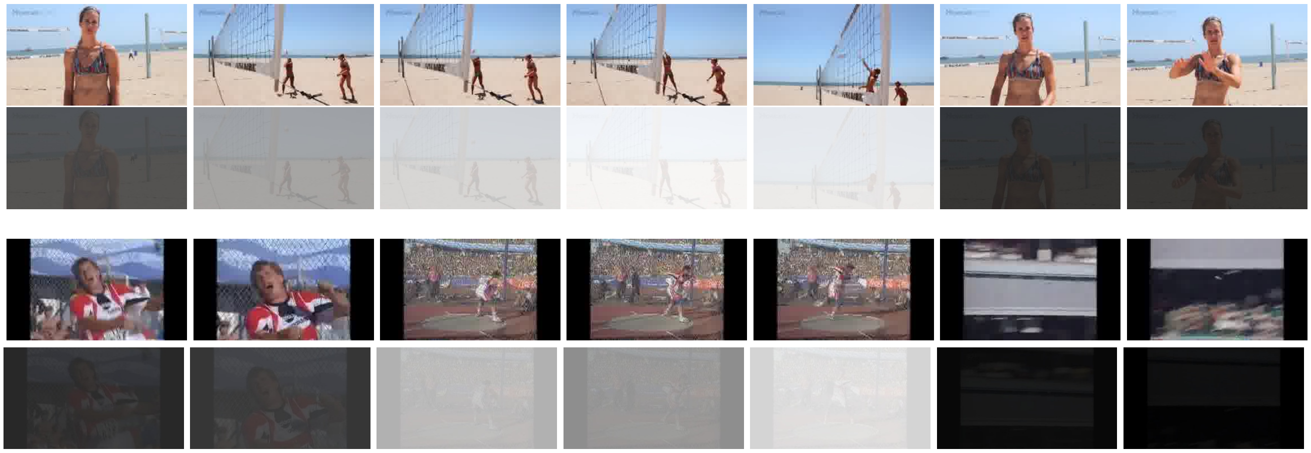

Figure 6. **Examples of temporal attention from THUMOS14.** The upper two rows show original images of the *Volleyball* action, and the corresponding images overlaid with temporal attention weights. The lower two rows show the *Throw Discus* action. Our temporal attention module can automatically highlight important frames and avoid irrelevant frames corresponding to non-action poses or background.

compared with the ground truth annotation at different IoU thresholds.

Qualitative results. We first visualize some examples of learned attention weights on the test data of THUMOS14 in Fig. 6. We see that our temporal attention module is able to automatically highlight important frames and to avoid irrelevant frames corresponding to background or non-action human poses. More qualitative results are included in the supplementary material.

Quantitative results. With our spatial temporal attention mechanism, the video action classification accuracy for the THUMOS'14 20 classes improved from 74.45% to 78.33%: a 3.88% increase. Besides improving the classification accuracy, we show our temporal attention mechanism is able to highlight discriminative frames quantitatively in Table 6. Compared with reinforcement learning based method [47] and weakly supervised method [39], our method achieves the best accuracy in terms of different levels of IoU thresholds.

5. Conclusion

In this paper we have introduced an interpretable spatial-temporal attention mechanism for video action recognition. Also, we propose a set of regularizers that ensure our attention mechanism attends to coherent regions in space and time, further improving the performance. Besides boosting video action recognition accuracy, our easy-plug-in spatio-temporal mechanism can increase the model interpretability. Furthermore, we qualitatively and quantitatively demonstrate that our spatio-temporal attention is able to localize discriminative regions and important frames, despite being trained in a purely weakly-supervised manner with only classification labels. Experimental results demonstrate the efficacy of our approach, showing superior or comparable accuracy with the state-of-the-art methods. For future work, it will be promising to integrate our spatio-temporal attention mechanism into 3D ConvNets [2, 7, 37] for improving action recognition performance and increasing model interpretability.

Methods	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
Fast action proposal* [48]	42.8%	–	–	–
Learning to track* [42]	54.3%	51.7%	**47.7%**	**37.8%**
Ours	**67.0%**	**58.2%**	40.2%	30.7%

Table 5. **Spatial action localization results on UCF101-24 dataset** measured by mAP at different IoU thresholds α. *The baseline methods are strongly supervised spatial localization methods.

Method	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
REINFORCE [47]	48.9%	44.0%	36.0%	26.4%	17.1%
UntrimmedNet [39]	44.4%	37.7%	28.2%	21.1%	13.7%
Ours	**70.0%**	**61.4%**	**48.6%**	**32.6%**	**17.9%**

Table 6. **Temporal action localization results on THUMOS'14 dataset** measured by mAP at different IoU thresholds α.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 1, 3

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2, 8

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[4] W. Du, Y. Wang, and Y. Qiao. Recurrent spatial-temporal attention network for action recognition in videos. *IEEE Transactions on Image Processing (ICIP)*, 2018. 2

[5] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 2

[6] R. Girdhar and D. Ramanan. Attentional pooling for action recognition. In *NIPS*, 2017. 1, 2, 5, 6

[7] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018. 2, 8

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5

[9] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 3

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 3

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[12] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr. Learn to pay attention. *ICLR*, 2018. 2

[13] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://crcv.ucf.edu/THUMOS14/`, 2014. 6, 7

[14] J. Kim and J. Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *CVPR*, 2017. 1, 2

[15] Y. Kong and Y. Fu. Human action recognition and prediction: A survey. *arXiv preprint arXiv:1806.11230*, 2018. 2

[16] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 5

[17] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu. Tell me where to look: Guided attention inference network. *CVPR*, 2018. 2

[18] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. Snoek. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding (CVIU)*, 2018. 1, 2, 5, 6

[19] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016. 1

[20] A. Mahendran and A. Vedaldi. Salient deconvolutional networks. In *ECCV*, 2016. 2

[21] M. Monfort, B. Zhou, S. A. Bargal, A. Andonian, T. Yan, K. Ramakrishnan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick, et al. Moments in time dataset: one million videos for event understanding. *arXiv preprint arXiv:1801.03150*, 2018. 5, 6

[22] R. Ramprasaath, D. Abhishek, V. Ramakrishna, C. Michael, P. Devi, and B. Dhruv. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CVPR*, 2016. 2

[23] R. A. Rensink. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42, 2000. 1

[24] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 1992. 4

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 2

[26] S. Sharma, R. Kiros, and R. Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015. 1, 2, 5, 6

[27] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015. 4

[28] Z. Shou, D. Wang, and S. Chang. Action temporal localization in untrimmed videos via multi-stage cnns. In *CVPR*, 2016. 3

[29] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[30] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 2

[31] G. Singh, S. Saha, M. Sapienza, P. H. Torr, and F. Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *ICCV*, 2017. 3

[32] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, 2017. 2, 3

[33] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5, 6

[34] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 2

[35] A. Torabi and L. Sigal. Action classification and highlighting in videos. *arXiv preprint arXiv:1708.09522*, 2017. 1, 2, 3

[36] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *CVPR*, 2015. 2

[37] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 2, 8

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017. 1

[39] L. Wang, Y. Xiong, D. Lin, and L. Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017. 3, 7, 8

[40] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 5, 6

[41] Y. Wang, S. Wang, J. Tang, N. O'Hare, Y. Chang, and B. Li. Hierarchical attention network for action recognition in videos. *arXiv preprint arXiv:1607.06416*, 2016. 2

[42] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. In *CVPR*, 2015. 3, 6, 7, 8

[43] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016. 1

[44] H. Xu, A. Das, and K. Saenko. R-C3D: region convolutional 3d network for temporal activity detection. In *ICCV*, 2017. 3

[45] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016. 1

[46] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 4

[47] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *CVPR*, 2016. 3, 7, 8

[48] G. Yu and J. Yuan. Fast action proposals for human action detection and search. In *CVPR*, 2015. 3, 6, 7, 8

[49] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 2

[50] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016. 2

[51] Q. Zhang, Y. Nian Wu, and S.-C. Zhu. Interpretable convolutional neural networks. In *CVPR*, 2018. 2

[52] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang. MD-Net: A semantically and visually interpretable medical image diagnosis network. In *CVPR*, 2017. 1, 2

[53] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017. 7

[54] B. Zhou, A. Andonian, and A. Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018. 5, 6

[55] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2