Uncertainty-Based Spatial-Temporal Attention for Online Action Detection

Hongji Guo[1], Zhou Ren[2], Yi Wu[2], Gang Hua[2], and Qiang Ji[1]

[1] Rensselaer Polytechnic Institute, Troy NY 12180, USA
{guoh11, jiq}@rpi.edu
[2] Wormpex AI Research, Bellevue WA 98004, USA
{renzhou200622, ywu.china, ganghua}@gmail.com

Abstract Online action detection aims at detecting the ongoing action in a streaming video. In this paper, we proposed an uncertainty-based spatial-temporal attention for online action detection. By explicitly modeling the distribution of model parameters, we extend the baseline models in a probabilistic manner. Then we quantify the predictive uncertainty and use it to generate spatial-temporal attention that focus on large mutual information regions and frames. For inference, we introduce a two-stream framework that combines the baseline model and the probabilistic model based on the input uncertainty. We validate the effectiveness of our method on three benchmark datasets: THUMOS-14, TVSeries, and HDD. Furthermore, we demonstrate that our method generalizes better under different views and occlusions, and is more robust when training with small-scale data.

Keywords: Online action detection, Spatial-temporal attention, Uncertainty modeling, Generalization, Robustness

1 Introduction

Traditional offline action detection [52,49,33] takes the entire sequence as the input to temporally localize the actions. Differently, online action detection (OAD) aims at detecting the ongoing action in a streaming video with only the previous and current frames. An illustration is shown in Figure 1. Online action detection has many practical applications since many real world tasks do not provide future observations and require real-time responses such as autonomous driving [19], anomaly detection [37], sports analysis [38]. Online action detection is very challenging due to the following reasons: (1) the beginnings of actions are unknown; (2) the observations of actions are incomplete; (3) background and irrelevant actions in the video may cause problems to the detection of the ongoing action; (4) there is a large within-class variability and the distribution of training data is imbalanced; and (5) the training data is limited in many situations.

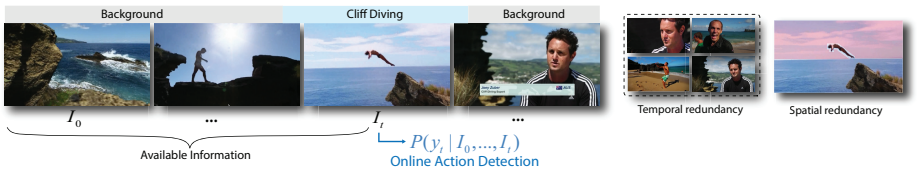


Figure 1: **An illustration of online action detection, temporal and spatial redundancy.** Online action detection aims at detecting the ongoing action without seeing the future. The available information includes all the frames up to the current.

Online action detection relies on existing observations, so features selection is crucial for the task. However, both temporal and spatial domains contain redundant information as illustrated in Figure 1, which may degenerate the performance since the prediction is made based on the irrelevant features. This problem can be alleviated by the attention mechanism, which automatically assigns weights to features according to their importance. In this way, the model performs better since the input features are more relevant and discriminative. For instance, Transformer [40] is a kind of attention model that captures the pair-wise dependency by scaled dot-product. However, when the amount of training data is limited or the model is applied to datasets with distribution shifts, traditional deterministic attention modeling methods become less robust and generalize poorly. Also, the purely learning-based attention methods are less interpretable in terms of attention generalization since the attentions are predicted to minimize the final loss function without considering the underlying dynamics.

To address these issues, we propose an uncertainty-based spatial-temporal attention for online action detection. Specifically, the model is extended in a probabilistic manner and the predictive uncertainty is quantified to compute the attention weights. The attention associated with a certain frame or region is based on its potential to reduce the prediction uncertainty of the ongoing action. In this way, the frames and regions with higher mutual information are assigned larger attention weights and the model can benefit from these discriminative features. When training data is insufficient, the model should be aware of that thanks to the quantified epistemic uncertainty in the probabilistic formulation. On the other hand, the generated attention is based on the input, thus the model can also generalize better when dealing with datasets with different distributions.

In general, our main contributions are summarized as: (1) we proposed an uncertainty-based spatial-temporal attention based on predictive uncertainty for online action detection; (2) the proposed attention mechanism can discriminate high mutual information frames and regions for better prediction of the ongoing action; (3) our proposed attention is validated on three benchmark datasets with multiple baseline methods and it shows the performance improvement; (4) we demonstrate that our proposed attention generalizes better under different views and occlusions and is more robust with small-scale training data

2 Related Work

Online action detection. Online action detection [4] is an important emerging research topic on account of the requirements of many real-time applications. Here we review the online action detection methods and related works chronologically. Gao *et al.* [10] proposed an encoder-decoder network trained by reinforcement learning for online action anticipation. To distinguish ambiguous background, Shou *et al.* [32] designed a hard negative samples generation module and an adaptive sampling method is used to handle the scarcity of the important training frames around the action starts. Xu *et al.* [45] proposed temporal recurrent network (TRN) that aggregates the features from the past and the future under the LSTM framework. To specifically detect the action starts, StarNet [11] combines an action classification network and a localization network to boost the performance. To deal with background and irrelevant features from the past, Eun *et al.* [6,7] introduced information discrimination unit (IDU) and temporal filtering network (TFN) to accumulate information based on its relevance to the current action and filter out irrelevant features. Zhao *et al.* [51] proposed a learning-with-privileged framework for online action detection by combing a offline teacher model and an online student model. With only video-level annotations, Gao *et al.* [12] proposed WORD for weakly supervised online action detection by introducing a proposal generator and an action recognizer. Recently, Transformer [40] is utilized to model the pairwise dependencies among frames. Wang *et al.* introduced OadTR [42] based on the standard Transformer architecture. Further, Xu *et al.* [46] proposed long short-term Transformer (LSTR) to simultaneously capture long-range and short-term information. Besides video modality, skeleton-based online action detection was also explored in [24,26,25].

Spatial-temporal attention. In video-based action recognition and detection, spatial-temporal attention modeling aims at learning the discriminative feature representation of actions from the input. In particular, attention modeling in online action detection is crucial since the input contains a lot of redundant or irrelevant information, which may cause the degeneracy of performance. In this part, we review the recent spatial-temporal attention modeling methods. In general, attention mechanisms in video understanding can be divided into spatial attention [28,18] and temporal attention [5,35,8,48], which model the discriminative regions and frames respectively. And the joint modeling yields the spatial-temporal attention [5,35,8,48]. The Transformer [40] is a fixed attention mechanism achieved by the scaled dot-product. Wang *et al.* proposed Non-Local (NL) network [43] by modeling the dependencies between human and objects across frames. To utilize the information of interaction between actors and context, ACAR-Net [29] models the relation between the actors and the context for spatial-temporal action localization. Recently, Zhao *et al.* introduced Tubelet Transformer (TubeR) [50] that can learn tubelet-queries and capture the dynamic spatial-temporal nature of videos through tubelet-attention. To model the spatial-temporal attention for action detection by utilizing the self-attention, Dai *et al.* introduced MS-TCT [3] with a multi-scale feature mixer module un-

der the Transformer framework to capture global and local temporal relations at multiple temporal resolutions. Existing attention modeling methods either generate the attention by a fixed mechanism such as scaled dot-product or let the neural network handle the input. These may work well when training data is sufficient but may not be robust under less training data or when generalized to different datasets. Furthermore, the attention generation process is less interpretable. When the task becomes challenging such as online action detection, these purely learning-based attention methods may confront problems. The trained model may perform poorly when generalizing to different datasets and become less robust when training data is limited. In this work, we aim at addressing these issues by generating the attention based on predictive uncertainty.

Uncertainty modeling. In machine learning systems, uncertainty quantification is crucial for better understanding the prediction the and improving the model [20]. Under the probabilistic setting, the prediction uncertainty of the model can be quantified using various approaches such as deep ensembles [23], dropout [9] and prior network [27]. Recently, uncertainty modeling has been applied to many computer vision tasks. By explicitly modeling the epistemic and aleatoric uncertainty, the predictions can be better interpreted and further be used to guide the model for specific tasks. Subedar *et al.* [36] quantified the uncertainty in a Bayesian framework and use it for the fusion of audio data and visual data. Want *et al.* [44] computed the data uncertainty to guide the semi-supervised object detection. Specifically, the image uncertainty guides the easy data selection and the region uncertainty guides RoI re-weighting. Yang *et al.* proposed UGTR [47] to perform the weakly-supervised action detection. Arnab *et al.* proposed a probabilistic variant of Multiple Instance Learning where the uncertainty of each prediction is estimated. Guo *et al.* proposed UGPT [13] for complex action recognition by utilizing the model uncertainty. In this paper, we quantify the predictive uncertainty for online action detection.

3 Method

In this section, we first formulate the problem of online action detection and spatial-temporal attention in Sec. 3.1. Then we introduce our uncertainty quantification method in Sec. 3.2 and how to compute the spatial-temporal attention in Sec. 3.3. The mechanism of the proposed uncertainty-based attention with respect to mutual information is discussed in Sec. 3.4. Finally, we introduce our two-stream inference framework in Sec. 3.5.

3.1 Problem setup

Online action detection (OAD) aims at identifying the ongoing action in a streaming video without seeing the future frames. Mathematically, denote an untrimmed video as $\boldsymbol{V} = [I_1, I_2, ..., I_T]$, where T is the video length and I_t represents the frame at time t. The available frames at time t is $\boldsymbol{V_t} = \{I_{t'}\}_{t'=1}^{t}$.

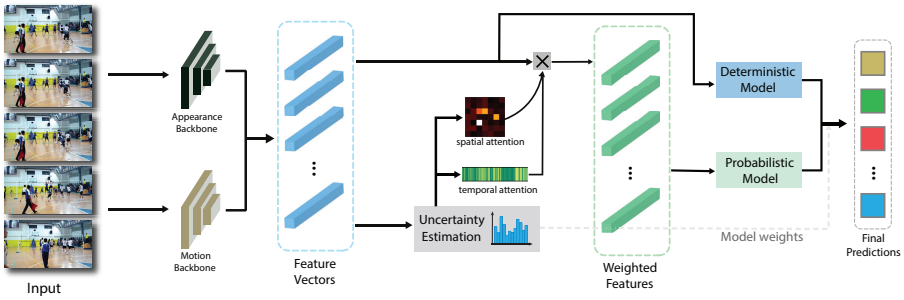


Figure 2: **Overall framework**. Firstly, feature vectors are constructed by concatenating the appearance features and motion features. Then we estimate the uncertainty based on the input and use the quantified uncertainty to generate the spatial-temporal attention. Finally, the prediction is made by dynamically combining both the deterministic model and probabilistic model, whose inputs are original features and attention-weighted features respectively.

Then online action detection can be formulated as a classification problem of frame t given $\boldsymbol{V_t}$:

$$y_t^* = \underset{c}{\operatorname{argmax}}\, P(\hat{y}_t = c|\boldsymbol{V}_t) \tag{1}$$

where $\hat{y}_t$ is the prediction of frame t. Class c belongs to an action set $\mathcal{Y} = \{0, 1, ..., C\}$, where 0 represents background class and C is the number of action classes.

Spatial-temporal attention (STA). For video-based action detection, spatial-temporal attention modeling aims at discriminating salient regions in certain frames that contain useful information for the tasks. By assigning higher weights to these regions, the model can take advantage of more discriminative features and improve the performance. Also, the spatial-temporal attention can make the model more interpretable by visualizing the attention weights. For online action detection, denote the extracted features of available frames at time t as $\boldsymbol{F_t} = \{f_{t'}\}_{t'=1}^{t}$, where $f_{t'}$ is the feature of frame t'. The spatial-temporal attention generates attention-weighted features $\boldsymbol{F'_t} = \{f'_{t'}\}_{t'=1}^{t}$, where

$$f'_{t'} = a_{t'} \times (b_{t'} \odot f_{t'}) \tag{2}$$

where $a_{t'}$ is the temporal attention weight and $b_{t'}$ is the spatial attention weight. $a_{t'}$ and $b_{t'}$ measure the temporal and spatial importance respectively. By applying the spatial-temporal attention, the impact of redundant or irrelevant information should be alleviated since they are assigned smaller weights.

3.2 Uncertainty quantification

In this part, we introduce the concepts of uncertainties and their quantification methods. Assume the problem is in classification setting. Denote the trained

model parameters as Θ^*. Given a test sample $\boldsymbol{X'}$, the output is the conditional probability distribution $P(y'|\boldsymbol{X'}, \Theta^*)$ over a set of action classes. From the output, we can quantify the total predictive uncertainty. We measure it by the entropy of the output distribution:

$$\mathcal{H}[y'|\boldsymbol{X'}, \Theta^*] = -\sum_{y' \in \mathcal{Y}} P(y'|\boldsymbol{X'}, \Theta^*) log P(y'|\boldsymbol{X'}, \Theta^*) \tag{3}$$

where $\mathcal{H}$ represents the entropy, y' is the predicted label, and $\mathcal{Y}$ is the action class set.

There are two sources of predictive uncertainty. One is the model parameters when the model is inadequately learned due to the the insufficient data. We refer this kind of uncertainty as **epistemic uncertainty** [20]. Increasing the data can reduce epistemic uncertainty. On the other hand, uncertainty also comes from data. When the input data is noisy, the uncertainty is large. We refer this kind of uncertainty as **aleatoric uncertainty** [20]. It cannot be reduced by increasing data. These two kinds of uncertainties add up to the total predictive uncertainty.

In a probabilistic model, denote all the model parameters as $\Theta = \{\Theta_d, \Theta_p\}$, where Θ_d and Θ_p represent deterministic and probabilistic model parameters respectively. The epistemic uncertainty is quantified as the mutual information between the prediction and the probabilistic model parameters [34]:

$$\mathcal{U}_E = \mathcal{I}\left[y', \Theta_p | \boldsymbol{X'}, \Theta_d^*\right] \tag{4}$$

where $\mathcal{I}$ represents the mutual information.

The aleatoric uncertainty measures the inherent noise in the observation. It is quantified as the expectation of the predictive uncertainty:

$$\mathcal{U}_A = \mathbb{E}_{\mathrm{p}(\Theta_p|\boldsymbol{X'}, \Theta_d^*)}\left[\mathcal{H}[y'|\boldsymbol{X'}, \Theta_p]\right] \tag{5}$$

The total uncertainty can be rewritten as the sum of epistemic uncertainty and aleatoric uncertainty as below:

$$\mathcal{H}\left[y'|\boldsymbol{X'}, \Theta^*\right] = \mathcal{I}\left[y', \Theta_p|\boldsymbol{X'}, \Theta_d^*\right] + \mathbb{E}_{\mathrm{p}(\Theta_p|\boldsymbol{X'}, \Theta_d^*)}\left[\mathcal{H}[y'|\boldsymbol{X'}, \Theta_p]\right] \tag{6}$$

Directly computing the uncertainty is infeasible since it is intractable to integrate over the true distribution. Thus, we generate samples and approximate the uncertainty by the sample average. After obtaining K output samples by repeating the probabilistic forward process, the epistemic uncertainty can be estimated as:

$$\mathcal{I}(y', \Theta_p|\boldsymbol{X'}, \Theta_d^*) \approx \mathcal{H}[\frac{1}{K}\sum_{k=1}^{K} P(y'|\boldsymbol{X'}, \Theta_d^*, \Theta_p^k)] - \frac{1}{K}\sum_{k=1}^{K} \mathcal{H}[y'|\boldsymbol{X'}, \Theta_d^*, \Theta_p^k] \tag{7}$$

Notice that we rearrange Eq. (6) and the first term on the right is the entropy of the average of the sample predictions, which is the estimation of total uncertainty. The second term is the average of the entropy of the sample predictions,

which is the estimation of the aleatoric uncertainty.

Motivation: For online action detection, most existing methods [45,12,10] directly take the raw features as the input without attention modeling, the drawbacks related to background and irrelevant actions are ignored. Recently, methods including IDN [6] and Transformer [42,46] model the temporal dependencies to alleviate the impacts of background and irrelevant actions. The euclidean distance between features and the scaled dot-product are used as measures for dependencies respectively. In this paper, we consider this problem from the information theory perspective: we assume the regions or frames that have large mutual information with respect to the ongoing action are more relevant and important to the current detection, which is analyzed in Sec. 3.4. Based on the assumption, the proposed uncertainty-based attention can identify high mutual information regions and frames. Also, the probabilistic extension of the framework can improve the robustness and generalization of the model.

3.3 Uncertainty-based spatial-temporal attention

Spatial attention. Given the input $\boldsymbol{V}_t = \{I_1, ..., I_t\}$ at time t, features extracted by the backbone are denoted as $\boldsymbol{F}_t = \{f_t, ..., f_t\}$, where $f_{t'} \in \mathbb{R}^{h \times w}$. Spatial attention aims at identifying relevant discriminative regions within each frame. Specifically, a spatial attention mask $b_{t'} \in \mathbb{R}^{h \times w}$ is generated for the feature of each frame and is applied by the Hadamard product [15] as $f^s_{t'} = b_{t'} \odot f_{t'}$.

To model the spatial attention mask by uncertainty. We model each element of the mask with a Gaussian distribution:

$$b^s_{t'}(i,j) \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}),\ i = 1, ...h; j = 1, ...w \tag{8}$$

During the training, the reparameterization trick [22] is adopted to perform the forward process: $b^s_{t'}(i,j) = \mu_{ij} + \epsilon\sigma_{ij}$, where $\epsilon \sim \mathcal{N}(0,1)$. In this way, the spatial attention mask can capture the randomness within the spatial domain of the input feature and further be utilized to quantify the predictive uncertainty. Noticed that except for these spatial attention masks, other model parameters are deterministic, which are fixed during the inference.

For inference, we estimate the predictive uncertainty to apply the attention. Based on the learned distribution of the spatial attention mask, we sample from it to generate multiple output from the input. Specifically, feature $f_{t'}$ goes through the spatial-attention mask for K times and generate K masks denoted as $\{b^s_{t'1}, ..., b^s_{t'K}\}$. From these samples, we estimate the uncertainty of each pixel (i,j) in the feature map. At time t', the epistemic uncertainty of each element can be estimated as:

$$\mathcal{I}_{t'}(i,j) = \mathcal{H}[\frac{1}{K}\sum_{k=1}^{K} P(y|x, b^s_{t'k}(i,j))] - \frac{1}{K}\sum_{k=1}^{K}\mathcal{H}[y|x, b^s_{t'k}(i,j)] \tag{9}$$

With the estimated uncertainty, we generate the spatial attention mask by a normalization:

$$b_{t'}(i,j) = 1 + \mathcal{U}_{t'}(i,j) \Big/ \sum_{i,j}\mathcal{U}_{t'}(i,j) \tag{10}$$

The equation above indicates that regions with high uncertainty are assigned large weights. Later we show these regions also have high mutual information. These attention weights are applied to the feature by $f^s_{t'} = b_{t'} \odot f_{t'}$.

Temporal attention. Temporal attention aims at identifying important frames from the input sequence. Specifically, a temporal attention weight $\alpha_{t'}$ is generated and applied to the input sequence by multiplication with the corresponding frame: $f'_{t'} = a_{t'} \times f_{t'}$.

For online action detection, the prediction of each frame is made by a fully connected classifier with a softmax layer in the end. Similar as spatiatl attention, we model the paramters in these fully connected layers in a probabilistic manner [39]. In this way, all the aggregated features are considered and evaluated for the current detection. Similarly, we assume the parameters Θ follows Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. The probabilistic parameters are computed by $\Theta = \boldsymbol{\mu} + \boldsymbol{\epsilon}\Sigma$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. By learning the mean $\boldsymbol{\mu}$ and covariance matrix Σ, we can estimate the distribution of Θ.

With probabilistic model parameters, we generate K samples from the input. Then the epistemic uncertainty corresponding to frame t' is estimated as:

$$\mathcal{I}_{t'} = \mathcal{H}[\frac{1}{K}\sum_{k=1}^{K} p(y^k_{t'})] - \frac{1}{K}\sum_{k=1}^{K} \mathcal{H}[p(y^k_{t'})] \tag{11}$$

To discriminate important frames, we make the temporal attention weight positively correlated to the epistemic uncertainty. We show this mechanism works with mutual information in Sec. 3.4. The temporal attention is computed by normalizing the weights of all considered frames as below:

$$a_{t'} = 1 + \mathcal{U}_{t'} \Big/ \sum_{t''=t-\tau}^{t} \mathcal{U}_{t''} \tag{12}$$

where τ is the number of past frame we consider. Then the generated weights are multiplied to the corresponding input. By modeling the temporal attention, the frames used to make the online prediction are evaluated and optimized based on their importance. So the discriminative information from the past are better utilized with less redundancy. Different from the other attention modeling methods, our proposed attention is based on the prediction uncertainty, which is more interpretable in terms of the generation process.

Spatial-temporal attention. To jointly model the spatial and temporal attention, we combine them to formulate a unified spatial-temporal attention using Eq. (2). The samples generated from the same input are used to estimate the predictive uncertainty for both spatial and temporal attention simultaneously. In this way, the spatial and temporal attention can be generated with the same estimated uncertainty, which is more computationally efficient. The training procedure is summarized as Algorithm 1.

Algorithm 1 Training

Input: $\mathcal{D} = \{\boldsymbol{X_n} \in \mathbb{R}^{T_n \times d}, \boldsymbol{y_n} \in \mathbb{R}^{T_n}\}_{n=1}^{N}$: training data
Output: Θ: model parameters

for $\{\boldsymbol{X_n}, \boldsymbol{y_n}\}$ in $\mathcal{D}$ **do**
  Generate K samples
  Compute uncertainty by Eq. (7)
  Compute temporal attention $\boldsymbol{a}$ and spatial attention $\boldsymbol{b}$
  Generate attention-weighted feature by Eq. (2)
  Optimize Θ with weighted features
end for
Return Θ

Algorithm 2 Inference

Input: $\mathcal{D}' = \{\boldsymbol{X'_n}\}$: testing data
Output: $\{\boldsymbol{y'}\}$: predicted labels

for $\boldsymbol{X'_n}$ in $\mathcal{D}'$ **do**
  Generate K samples
  Compute uncertainty by Eq. (7)
  Compute temporal attention $\boldsymbol{a}$ and spatial attention $\boldsymbol{b}$
  Generate attention-weighted feature by Eq. (2)
end for
Compute w_u and w_b by Eq. (16)
Make prediction $\{\boldsymbol{y'}\}$ by Eq. (15)
Return $\{\boldsymbol{y'}\}$

3.4 Mechanism of uncertainty-based attention

In this part, we relate the predictive uncertainty with mutual information to demonstrate the underlying mechanism of our proposed attention mechanism. For online action detection, the relevant frames in the past should have high mutual information with the current frame, which can guide the generation of attention.

Uncertainty and mutual information. For a past time t', denote $F_{-t'} = \{F_{t-T}, ..., F_{t'-1}, F_{t'+1}, ..., F_t\}$. Then the mutual information between its feature and the current action can be written as:

$$\begin{aligned}\mathcal{I}[y_t; F_{t'}|F, \Theta_d] =& \mathcal{H}\big[E_{p(\Theta_p|F_{-t'})}[P(y_t|F_{-t'}, \Theta_p, \Theta_d)]\big] \\ &- \mathcal{H}\big[E_{P(\Theta_p|F)}[P(y_t|F, \Theta_p, \Theta_d)]\big]\end{aligned} \tag{13}$$

Combining Eq. 6 and Eq. 13 yields:

$$\begin{aligned}\mathcal{I}[y_t; F_{t'}|F_{-t'}] =& \mathcal{H}\big[E_{p(\Theta_p|F_{-t'})}[P(y_t|F_{-t'}, \Theta_p, \Theta_d)]\big] - \mathcal{I}[y_t; \Theta_p|F, \Theta_d] \\ &- E_{(P(\Theta_p|F)}\big[\mathcal{H}[y_t|F, \Theta_p, \Theta_d]\big]\end{aligned} \tag{14}$$

The mutual information on the left is what we desired. It is negatively correlated to the predictive uncertainty of the current action. So the features that lead to lower predictive uncertainty have higher mutual information with the ongoing action. In another words, the information that leads to low predictive uncertainty is treated higher weight.

Analysis and insights. By explicitly modeling the distributions of model parameters, our probabilistic architecture can well capture the stochasticity of the data and model. On the other hand, deterministic methods [5,35,8] directly generate the attention from the input feature. The network for attention generation needs to be trained well with enough data. Thus, the uncertainty-based model should have better generalization ability and more robustness than the deterministic methods. To demonstrate our propositions, we perform the generalization experiments and insufficient data experiments.

3.5 Two-stream framework

To leverage both the probabilistic model and deterministic model, we combine the baseline model with the uncertainty-based model dynamically based on the input uncertainty. The final prediction model is formulated as:

$$P(y|\boldsymbol{X'},\Theta^*) = w_u(\boldsymbol{X'})p_u(y|\boldsymbol{X'},\Theta^*) + w_b(\boldsymbol{X'})p_b(y|\boldsymbol{X'},\Theta^*) \tag{15}$$

where w_u and w_b are the weights of uncertainty-based model and baseline model respectively. They are computed based on the predictive uncertainty as below:

$$w_u(\boldsymbol{X'}) = \sigma\big(w\frac{\mathcal{U}_{max} - \mathcal{U}(\boldsymbol{X'})}{\mathcal{U}_{max} - \mathcal{U}_{min}} + b\big),\ w_b(\boldsymbol{X'}) \quad = 1 - w_b(\boldsymbol{X'}) \tag{16}$$

where $\mathcal{U}_{max}$ and $\mathcal{U}_{min}$ are the maximum and minimum uncertainty respectively. w and b are learnable parameters. The inference procedure is summarized as Algorithm 2.

4 Experiments

In this section, we first introduce benchmark datasets and evaluation metrics of online action detection in Sec. 4.1 and Sec. 4.2 respectively. Implementation details are provided in Sec. 4.3. The main experimental results on baseline methods are discussed in Sec. 4.4. Some qualitative results are presented in Sec. 4.5. The ablation studies are shown in Sec. 4.6.

4.1 Datasets

THUMOS-14 [16]. THUMOS-14 is a dataset of videos for temporal action localization. Following the settings in existing works [10,45], we use the 200 videos in the validation set for training and 213 videos in the test set for evaluation. There are totally 20 sports action classes as well as background in these videos. Each video contains 15.8 actions on average and the background frames occupy 71% of the video.

TVSeries [4]. TVSeries contains 27 episodes untrimmed videos from six TV series. There are totally 16 hours videos and 30 daily action classes such as 'eat', 'smoke'. It is a challenging dataset due to the diversity of actions, moving cameras, and heavy occlusion. This dataset provide metadata such as viewpoints and occlusions, which are used for our generalization experiments.

HDD [31]. HDD is a dataset for driving scene understanding. It includes 104 hours of real human driving in the San Francisco Bay Area collected by an instrumented vehicle. There are totally 11 goal-oriented driving actions such as passing, right turn. Following the settings in [31], we use 100 sessions for training and 37 sessions for testing.

4.2 Evaluation metrics

mean Average Precision (mAP): following existing works [4,45,6,42], we use mAP as the evaluation metric for THUMOS-14 and HDD dataset. It is computed by taking the mean of the average precision of each action class over all frames.
mean calibrated Average Precision (mcAP) [4] is used as the evaluation metric for TVSeries dataset. As mAP is sensitive to the ratio of positive frames versus negative background frames, it is difficult to compare two classes with different positive vs. negative ratio. To address this issue, the mcAP is used. The calibrated precision is defined as:

$$cPrec = \frac{TP}{TP + \frac{FP}{\omega}} = \frac{\omega \times TP}{\omega \times TP + FP} \tag{17}$$

where ω is the ratio between negative frames and positive frames. Then the calibrated average precision (cAP) is computed similarly as mAP:

$$cAP = \frac{\sum_k cPrec(k) \times \mathbb{1}(k)}{P} \tag{18}$$

where P is the total number of positive frames and $\mathbb{1}(k)$ is an indicator function that is equal to 1 if frame k is a true positive. The mcAP is the mean of calibrated average precision of all action classes.

4.3 Implementation details

Feature extraction. We use TSN [41] for feature extraction. The video frames are extracted at 24 fps and the chunk size is set to 6. We adopt a two-stream architecture with ResNet-200 [14] for appearance features and BN-Inception [17] for motion features. Specifically, the network pretrained on ActivityNet [1] outputs 3072-dimensions features. The appearance features have 2048 dimensions and the motion features have 1024 dimension. And the network pretrained on Kinetics [2] generates 4096-dimensions features. Both appearance features and motion features have 2048 dimensions.
Settings. We implemented our proposed uncertainty-based spatial-temporal attention in PyTorch [30]. The training is conducted by the Adam optimizer [21]. For TRN, the learning rate is set to 5×10^{-5} with a weight decay rate of 5×10^{-5}. The batch size is set to 12. The number of epochs is set to 25. For OadTR, we set the learning rate to 10^{-4} with a weight decay rate of 10^{-4}. The batch size is set to 128.

4.4 Main experimental results

To demonstrate the effectiveness of our proposed uncertainty-based spatial-temporal attention, we apply it on three baseline methods: TRN [45], OadTR [42], and LSTR [46]. Experimental results on THUMOS-14 are shown in Tab. 1. Results on TVSeries and HDD are shown in Tab. 2 and Tab. 3 respectively.

Table 1: Experimental results on THUMOS-14 with ActivityNet features and Kinetics features in terms of mAP (%)

Method	ActivityNet	Kinetics
TRN [45]	47.2	62.1
TRN + Spatial	48.3	62.5
TRN + Temporal	50.1	62.8
TRN + Spatial-Temporal	51.3	63.1
OadTR [42]	58.3	65.2
OadTR + Spatial	58.9	66.9
OadTR + Temporal	59.4	66.4
OadTR + Spatial-Temporal	60.7	67.5
LSTR [46]	65.3	69.5
LSTR + Spatial	65.7	69.8
LSTR + Temporal	65.9	69.9
LSTR + Spatial-Temporal	66.0	69.9

Table 2: Experimental results on TVSeries

Method	ActivityNet	Kinetics
TRN [45]	83.7	86.2
TRN + STA	85.2	86.9
OadTR [42]	85.4	87.2
OadTR + STA	86.6	87.7
LSTR [46]	88.1	89.1
LSTR + STA	88.3	89.3

Table 3: Experimental results on HDD

Method	mAP (%)
TRN [45]	29.2
TRN + STA	29.6
OadTR [42]	29.8
OadTR + STA	30.1

From the results, our proposed uncertainty-based spatial-temporal attention improves the performance of all baseline methods on three datasets. Both spatial and temporal attention improve the online action detection, especially for the RNN-based method TRN. For Transformer-based method, LSTR, the performance gain with STA is not as significant as TRN with STA. This is because the self-attention mechanisms have already used in the baseline approach which may reduce the benefits of our proposed attention. The performance of different portions of videos on TVSeries is shown in Tab. 4. The methods with STA outperforms baseline methods at every stage of action instances.

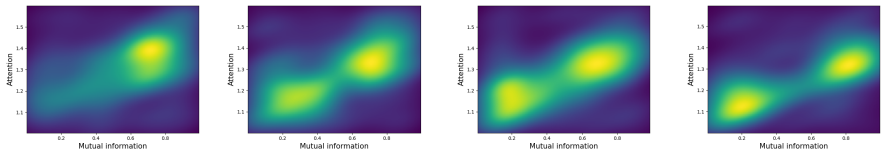
Table 4: Experimental results on TVSeries of different portions of videos in terms of mcAP (%). Each portion is only used to compute mcAP after detecting the current actions on all frames in an online manner.

Method	Portion of video									
	0%-10%	10%-20%	20%-30%	30%-40%	40%-50%	50%-60%	60%-70%	70%-80%	80%-90%	90%-100%
TRN [45]	78.8	79.6	80.4	81.0	81.6	81.9	82.3	82.7	82.9	83.3
TRN+STA	81.4	80.2	80.6	81.3	83.7	85.8	84.9	83.	83.5	83.7
OadTR [42]	79.5	83.9	86.4	85.4	86.4	87.9	87.3	87.3	85.9	84.6
OadTR + STA	79.9	84.4	87.2	85.7	86.5	88.4	88.0	88.2	87.4	85.1
LSTR [46]	83.6	85.0	86.3	87.0	87.8	88.5	88.6	88.9	89.0	88.9
LSTR + STA	83.7	85.2	87.2	87.1	88.3	88.7	88.6	89.2	89.5	89.0

4.5 Qualitative results

Attention and mutual information. To verified the mechanism in Sec. 3.4, we plot the distribution of attention and mutual information in Fig. 3. The distributions are obtained on THUMOS-14 dataset. When computing the mutual information, we select top-k high-probability actions to reduce the impact of

low-probability actions. From the visualization, the attention is approximately positively related to the mutual information, which is as expected.



(a) All classes (b) Top-10 classes (c) Top-8 classes (d) Top-4 classes

Figure 3: Distributions of attention and mutual information. The attention is approximately positively related with the mutual information.

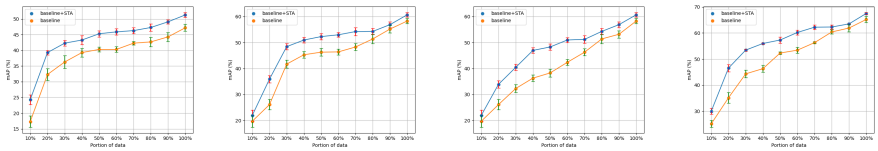
Temporal and spatial attention. Visualization of temporal and spatial attention are shown in Fig. 4. From the visualization, the action in the temporal and spatial domain are assigned with higher attention weights, which is as expected.

Figure 4: Visualization of attention. The temporal attention is shown on the left and the spatial attention is shown on the right.

4.6 Ablation studies

Training with small-scale data. The amount of training data is limited in many real situations, where the data-hungry methods may not work well. To demonstrate the robustness of our proposed uncertainty-based attention, we reduce the amount of training data from 100% to 10% and compare with the deterministic baseline methods. The experimental results on two baseline methods with two features are plotted in Fig. 5. For both baseline methods with ActivityNet features and Kinetics features, our uncertainty-based attention perform better, which shows that our method is more robust. The performance gaps are obvious when the amount of training data is between 20% and 70%. When the training data is extremely limited (10%), the uncertainty estimation failed and lead to marginal improvement on the baseline methods.

Generalization. Based on the meta annotations of TVSeries, we perform two kinds of generalization experiments. First, we divide the dataset into two parts



(a) TRN-ActivityNet (b) TRN-Kinetics (c) OadTR-ActivityNet (d) OadTR-Kinetics

Figure 5: Experiment results of training with small-scale data on THUMOS-14 with ActivityNet and Kinetics features. The uncertainty-based attention perform more robust than the standard attention on both baselines.

Table 5: Generalization experimental results.

Method	CV (%)	Occ. (%)
TRN [45]	65.8	85.2
TRN + STA	69.5	88.6
OadTR [42]	66.2	87.7
OadTR + STA	67.3	89.5

Table 6: Comparison of model complexity and computation cost.

Method	# of Paras	FLOPs	Per-frame Speed	Memory Cost
TRN [45]	357.8M	1.4G	0.0104 s	6479 MB
TRN + STA	379.2M	3.1G	0.0201 s	9662 MB
OadTR [42]	74.7M	2.5G	0.0069 s	1787 MB
OadTR + STA	78.5M	5.9G	0.0094 s	2835 MB

based on different viewpoints. We select frontal viewpoint frames and special viewpoint frames as the training set, and side viewpoint frames as testing set. Second, we divide the dataset based on occlusion conditions in the frames. The frames without occlusion are selected for training and the occluded ones are used for testing. The experimental results are shown in Table 5. For both cases, our proposed uncertainty-based attention outperformed the baseline methods, which demonstrates the generalization ability of our method.

Computation efficiency and model complexity. We made a comparison with baseline methods in Tab. 6. Compared with the baseline methods, our proposed uncertainty-based attention increase the computation cost of baseline methods since we need to sample from the parameter distribution and perform K times forward process. Compared with baseline ensembles, our method perform better with less computation cost and model complexity. We also make a comparison of inference speed and memory cost. The computation complexity increase linearly and our method can still achieve real-time responses.

5 Conclusion and Future Work

In this paper, we proposed uncertainty-based spatial-temporal attention for online action detection. By modeling the predictive uncertainty, the proposed attention mechanism improves the model with more discriminative features. Under the probabilistic setting, the generalization and robustness of the model are also improved. The proposed method is validated on three benchmark datasets to show the effectiveness, generalization, and robustness.

Future work may include the evaluation of different uncertainty quantification methods and improving the computation efficiency.

Acknowledgement: This project is supported in part by a gift from Wormpex AI Research to Rensselaer Polytechnic Institute.

References

1. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
3. Dai, R., Das, S., Kahatapitiya, K., Ryoo, M.S., Bremond, F.: Ms-tct: Multi-scale temporal convtransformer for action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20041–20051 (2022)
4. De Geest, R., Gavves, E., Ghodrati, A., Li, Z., Snoek, C., Tuytelaars, T.: Online action detection. In: European Conference on Computer Vision. pp. 269–284. Springer (2016)
5. Du, W., Wang, Y., Qiao, Y.: Recurrent spatial-temporal attention network for action recognition in videos. IEEE Transactions on Image Processing **27**(3), 1347–1360 (2017)
6. Eun, H., Moon, J., Park, J., Jung, C., Kim, C.: Learning to discriminate information for online action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 809–818 (2020)
7. Eun, H., Moon, J., Park, J., Jung, C., Kim, C.: Temporal filtering networks for online action detection. Pattern Recognition **111**, 107695 (2021)
8. Fu, Y., Wang, X., Wei, Y., Huang, T.: Sta: Spatial-temporal attention for large-scale video-based person re-identification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8287–8294 (2019)
9. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
10. Gao, J., Yang, Z., Nevatia, R.: Red: Reinforced encoder-decoder networks for action anticipation. arXiv preprint arXiv:1707.04818 (2017)
11. Gao, M., Xu, M., Davis, L.S., Socher, R., Xiong, C.: Startnet: Online detection of action start in untrimmed videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5542–5551 (2019)
12. Gao, M., Zhou, Y., Xu, R., Socher, R., Xiong, C.: Woad: Weakly supervised online action detection in untrimmed videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1915–1923 (2021)
13. Guo, H., Wang, H., Ji, Q.: Uncertainty-guided probabilistic transformer for complex action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20052–20061 (2022)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Horn, R.A.: The hadamard product. In: Proc. Symp. Appl. Math. vol. 40, pp. 87–169 (1990)
16. Idrees, H., Zamir, A.R., Jiang, Y.G., Gorban, A., Laptev, I., Sukthankar, R., Shah, M.: The thumos challenge on action recognition for videos "in the wild". Computer Vision and Image Understanding **155**, 1–23 (2017)
17. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. PMLR (2015)

18. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28**, 2017–2025 (2015)
19. Janai, J., Güney, F., Behl, A., Geiger, A., et al.: Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends® in Computer Graphics and Vision **12**(1–3), 1–308 (2020)
20. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
22. Kingma, D.P., Salimans, T., Welling, M.: Variational dropout and the local reparameterization trick. Advances in neural information processing systems **28**, 2575–2583 (2015)
23. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
24. Li, Y., Lan, C., Xing, J., Zeng, W., Yuan, C., Liu, J.: Online human action detection using joint classification-regression recurrent neural networks. In: European conference on computer vision. pp. 203–220. Springer (2016)
25. Liu, J., Li, Y., Song, S., Xing, J., Lan, C., Zeng, W.: Multi-modality multi-task recurrent neural network for online action detection. IEEE Transactions on Circuits and Systems for Video Technology **29**(9), 2667–2682 (2018)
26. Liu, J., Shahroudy, A., Wang, G., Duan, L.Y., Kot, A.C.: Ssnet: scale selection network for online 3d action prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8349–8358 (2018)
27. Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. Advances in neural information processing systems **31** (2018)
28. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: Advances in neural information processing systems. pp. 2204–2212 (2014)
29. Pan, J., Chen, S., Shou, M.Z., Liu, Y., Shao, J., Li, H.: Actor-context-actor relation network for spatio-temporal action localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 464–474 (2021)
30. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
31. Ramanishka, V., Chen, Y.T., Misu, T., Saenko, K.: Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7699–7707 (2018)
32. Shou, Z., Pan, J., Chan, J., Miyazawa, K., Mansour, H., Vetro, A., Nieto, X.G., Chang, S.F.: Online action detection in untrimmed, streaming videos-modeling and evaluation. In: ECCV. vol. 1, p. 5 (2018)
33. Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1049–1058 (2016)
34. Smith, L., Gal, Y.: Understanding measures of uncertainty for adversarial example detection. arXiv preprint arXiv:1803.08533 (2018)
35. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017)
36. Subedar, M., Krishnan, R., Meyer, P.L., Tickoo, O., Huang, J.: Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In: Pro-

ceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6301–6310 (2019)
37. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6479–6488 (2018)
38. Thomas, G., Gade, R., Moeslund, T.B., Carr, P., Hilton, A.: Computer vision for sports: Current applications and research topics. Computer Vision and Image Understanding **159**, 3–18 (2017)
39. Tran, D., Dusenberry, M., van der Wilk, M., Hafner, D.: Bayesian layers: A module for neural network uncertainty. Advances in Neural Information Processing Systems **32**, 14660–14672 (2019)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
41. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L.V.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016)
42. Wang, X., Zhang, S., Qing, Z., Shao, Y., Zuo, Z., Gao, C., Sang, N.: Oadtr: Online action detection with transformers. arXiv preprint arXiv:2106.11149 (2021)
43. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
44. Wang, Z., Li, Y., Guo, Y., Fang, L., Wang, S.: Data-uncertainty guided multi-phase learning for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4568–4577 (2021)
45. Xu, M., Gao, M., Chen, Y.T., Davis, L.S., Crandall, D.J.: Temporal recurrent networks for online action detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5532–5541 (2019)
46. Xu, M., Xiong, Y., Chen, H., Li, X., Xia, W., Tu, Z., Soatto, S.: Long short-term transformer for online action detection. Advances in Neural Information Processing Systems **34** (2021)
47. Yang, F., Zhai, Q., Li, X., Huang, R., Luo, A., Cheng, H., Fan, D.P.: Uncertainty-guided transformer reasoning for camouflaged object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4146–4155 (2021)
48. Yang, J., Zheng, W.S., Yang, Q., Chen, Y.C., Tian, Q.: Spatial-temporal graph convolutional network for video-based person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3289–3299 (2020)
49. Zhao, C., Thabet, A.K., Ghanem, B.: Video self-stitching graph network for temporal action localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13658–13667 (2021)
50. Zhao, J., Zhang, Y., Li, X., Chen, H., Shuai, B., Xu, M., Liu, C., Kundu, K., Xiong, Y., Modolo, D., et al.: Tuber: Tubelet transformer for video action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13598–13607 (2022)
51. Zhao, P., Xie, L., Zhang, Y., Wang, Y., Tian, Q.: Privileged knowledge distillation for online action detection. arXiv preprint arXiv:2011.09158 (2020)
52. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2914–2923 (2017)