ACSNet: Action-Context Separation Network for Weakly Supervised Temporal Action Localization

Ziyi Liu¹, Le Wang¹*, Qilin Zhang², Wei Tang³, Junsong Yuan⁴, Nanning Zheng¹, Gang Hua⁵

¹Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University ²HERE Technologies
³University of Illinois at Chicago ⁴The State University of New York at Buffalo ⁵Wormpex AI Research
liuziyi@stu.xjtu.edu.cn, {lewang, nnzheng}@mail.xjtu.edu.cn,
tangw@uic.edu, jsyuan@buffalo.edu, {samqzhang, ganghua}@gmail.com,

Abstract

The object of Weakly-supervised Temporal Action Localization (WS-TAL) is to localize all action instances in an untrimmed video with only video-level supervision. Due to the lack of frame-level annotations during training, current WS-TAL methods rely on attention mechanisms to localize the foreground snippets or frames that contribute to the video-level classification task. This strategy frequently confuses context with the actual action, in the localization result. Separating action and context is a core problem for precise WS-TAL, but it is very challenging and has been largely ignored in the literature. In this paper, we introduce an Action-Context Separation Network (ACSNet) that explicitly takes into account context for accurate action localization. It consists of two branches (i.e., the Foreground-Background branch and the Action-Context branch). The Foreground-Background branch first distinguishes foreground from background within the entire video while the Action-Context branch further separates the foreground as action and context. We associate video snippets with two latent components (i.e., a positive component and a negative component), and their different combinations can effectively characterize foreground, action and context. Furthermore, we introduce extended labels with auxiliary context categories to facilitate the learning of action-context separation. Experiments on THUMOS14 and ActivityNet v1.2/v1.3 datasets demonstrate the ACSNet outperforms existing state-of-the-art WS-TAL methods by a large margin.

1 Introduction

Temporal Action Localization (TAL) aims to localize temporal starts and ends of specific action categories in a video. It serves as a fundamental tool for several practical applications such as action retrieval, intelligent surveillance and video summarization (Lee, Ghosh, and Grauman 2012; Vishwakarma and Agrawal 2013; Asadiaghbolaghi et al. 2017; Kang and Wildes 2016; Yao, Lei, and Zhong 2019). Although fully supervised TAL methods have recently achieved remarkable progress (Buch et al. 2017; Xu, Das, and Saenko 2017; Gao et al. 2017; Chao et al. 2018; Lin et al. 2018, 2019; Zeng et al. 2019), manually annotating the precise temporal boundaries of action instances in untrimmed videos is time-consuming and challenging. This limitation motivates the weakly supervised setting where only video-level categorical labels are provided for model training. Compared with temporal boundary annotations, video-level categorical labels are easier to collect, and they help avoid the localization bias introduced by human annotators.

Existing weakly-supervised temporal action localization (WS-TAL) methods (Wang et al. 2017; Nguyen et al. 2018; Paul, Roy, and Roy-Chowdhury 2018; Nguyen, Ramanan, and Fowlkes 2019) leverage attention mechanisms to categorize snippets or sampled frames into foreground and background based on their contribution to the video-level classification task, i.e., to find the blue dashed line in Figure 1. Then temporal action localization is reformulated as select-
The goal of this paper is to address the action-context separation (ACS) problem in the weakly-supervised setting so as to achieve more precise action localization. We first introduce auxiliary context categories for each action class during training. As shown in Figure 2, each video-level category is divided into two sub-categories, respectively corresponding to the actual action and its context. Prior methods exploit foreground attention to achieve foreground-background separation. However, this simple idea is not applicable to action-context separation due to two difficult issues. (1) Lack of explicit action-context constraints: The sum-to-one constraint (Nguyen, Ramanan, and Fowlkes 2019) of the foreground and background attention scores does not apply to action-context separation. (2) Lack of explicit supervision: Both action and context can contribute to action classification, so the only available video-level categorical labels cannot provide direct supervision for them.

To address these two difficult issues, we introduce the Action-Context Separation Network (ACSNet). As illustrated in Figure 3, it consists of two branches, i.e., the Foreground-Background branch (FB branch) and the Action-Context branch (AC branch). The FB branch divides an untrimmed video into foreground and background based on whether a snippet supports the video-level classification. This is achieved via snippet-level categorical predictions (SCPs) and snippet-level attention predictions (SAPs), e.g., foreground attention in Figure 2. Subsequently, the AC branch further divides the obtained foreground into action and context by associating each video snippet with two latent components, i.e., a positive component and a negative component. Different combinations of these two components respectively characterize the foreground, action, and context. This enables effective action-context separation with only video-level supervision. Finally, the output of AC branch facilitates the TAL by providing (1) temporal action proposals with more accurate boundaries and (2) more reliable proposal confidence scores.

The contribution of this paper is summarized below.

1. Prior WS-TAL approaches take it for granted that the foreground localized via the classification attention is equivalent to the actual action instance, and thus they unavoidably include the co-occurring context in the localization result. We address this challenge via a novel action-context separation network (ACSNet), which not only distinguishes foreground from background but also separates action and context within the foreground to achieve more precise action localization.

2. The proposed ACSNet features a novel Action-Context branch. It can individually characterize foreground, action, and context using different combinations of two latent components, i.e., the positive component and the negative component.

3. We propose novel extended labels with auxiliary context categories. By explicitly decoupling the actual action and its context, this new representation facilitates effective learning of action-context separation.

4. Extensive experimental results indicate the proposed ACSNet can effectively perform action-context separation. It significantly outperforms state-of-the-art methods on three benchmarks, and it is even comparable to recent fully-supervised methods.

## 2 Related Work of WS-TAL

Different from action recognition which is essentially a classification task (Feichtenhofer, Pinz, and Zisserman 2016; Simonyan and Zisserman 2014; Wang et al. 2016; Ji et al. 2013; Sun et al. 2015b; Tran et al. 2015; Feichtenhofer et al. 2019), TAL requires finer-grained predictions with temporal boundaries of the target action instances. WS-TAL methods address it without temporal annotations, which is first introduced in (Sun et al. 2015a). To distinguish action instances from background, the attention mechanism is widely adopted for foreground-background separation. UntrimmedNet (Wang et al. 2017) formulates the attention mechanism as a soft selection module to localize target action, and the final localization is achieved by thresholding the snippets’ action scores. STPN (Nguyen et al. 2018) proposes a sparsity loss based on the soft selection module of UntrimmedNet, which can facilitate the selection of action instances. Nguyen et al. (Nguyen, Ramanan, and Fowlkes 2019) characterize background by an additional background loss and
introduce other losses to guide the attention. For better evaluation of temporal action proposals, W-TALC (Paul, Roy, and Roy-Chowdhury 2018) proposes a co-activity loss to enforce the feature similarity among localized instances. AutoLoc (Shou et al. 2018) uses an “outer-inner-contrastive loss” to predict and regress temporal boundaries. Liu et al. (Liu, Jiang, and Wang 2019) exploit a multi-branch neural network to discover distinctive action parts and fuse them to ensure completeness. CleanNet (Liu et al. 2019b) designs a “contrast score” by leveraging temporal contrast in SCPs to predict and regress temporal boundaries. Liu et al. (Shou et al. 2018) uses an “outer-inner-contrastive loss” to predict and regress temporal boundaries.

However, driven by the video-level classification labels, the existing attention mechanism is merely able to capture the difference between foreground and background for classification, instead of action and non-action for localization. The proposed ACSNet manages to distinguish action instances from their surrounding context, and we extend labels by introducing auxiliary context categories to make the framework trainable.

3 Action-Context Separation Network

In this section, we introduce the extended video-level labels with auxiliary context categories (Section 3.1) and the proposed Action-Context Separation Network (ACSNets). As illustrated in Figure 3, the ACSNet consists of two branches, i.e., Foreground-Background branch (FB branch) and Action-Context branch (AC branch). After feature extraction from the given video (Section 3.2), FB branch distinguishes the foreground from background (Section 3.2). The obtained foreground contains both action and context. Subsequently, AC branch localizes the actual temporal action instances by performing action-context separation within the foreground (Section 3.3). To guide the training of ACS, additional losses are introduced (Section 3.4).

3.1 Extending Video-Level Labels

Suppose we are given a video $V$ with a video-level categorical label $y = [y(0), y(1), ..., y(N)]$, where $y(n) = 1$ if $V$ contains the $n$-th action category. $N$ is the total number of action categories, $y(0)$ represents the background category. To guide the division of foreground into action and context, we extend $y$ with auxiliary context categories as

$$\tilde{y} = [y_a(1), ..., y_a(N), y_c(1), ..., y_c(N)], \tilde{y} \in \mathbb{R}^{2N},$$

where $y_a(n)$ and $y_c(n)$ denote the $n$-th action category and its corresponding context, respectively. As shown in Figure 3, $y \in \mathbb{R}^{N+1}$ is used in FB branch and $\tilde{y} \in \mathbb{R}^{2N}$ is used in AC branch.

3.2 Baseline Modules

This section introduces the baseline modules used in ACSNet, including feature extraction and FB branch based on the attention mechanism. While they are not our main contribution, we introduce them for completeness. Similar modules have been explored and adopted by existing methods (Nguyen et al. 2018; Paul, Roy, and Roy-Chowdhury 2018; Nguyen, Ramanan, and Fowlkes 2019; Lee, Uh, and Byun 2020).

Feature Extraction

The input of the feature extraction module is the given video $V = \{s_t\}_{t=1}^T$, which is divided into $T$ non-overlapping snippets. The outputs are the corresponding features of each snippet. For each snippet $s_t$, the corresponding $D$-dimensional features are extracted from two streams, i.e., the spatial stream (RGB) and the temporal stream (optical flow), denoted as $F^{rgb}(t) \in \mathbb{R}^D$ and $F^{flow}(t) \in \mathbb{R}^D$, respectively. Afterwards, the video $V$ is represented as $F^{rgb} \in \mathbb{R}^{D \times T}$ and $F^{flow} \in \mathbb{R}^{D \times T}$.

For notational simplicity, we use superscript “$s$” to indicate the notations used in both streams in the rest of the paper. The notations of the spatial/temporal stream can be obtained by substituting the superscript “$s$” with “rgb/flow”. For example, $F^s$ can represent either $F^{rgb}$ or $F^{flow}$.

Foreground-Background Branch

The goal of the FB branch is to divide the entire video into two parts, i.e., foreground and background, which can be trained by the video-level categorical label $y = [y(0), y(1), ..., y(N)]$.

The inputs of FB branch are the features $F^s \in \mathbb{R}^{D \times T}$, and the outputs are the snippet-level attention predictions (SAPs, $\varphi \in \mathbb{R}^{1 \times T}$) and the snippet-level classification predictions (SCPs, $\Psi \in \mathbb{R}^{(N+1) \times T}$). Accordingly, FB branch consists of two sub-modules, i.e., attention module ($m_{fg}^s$) and Foreground-Background classification module ($m^s$). The SAPs and SCPs of each stream are obtained by

$$\varphi^s = m_{fg}^s(F^s), \quad \varphi^s \in \mathbb{R}^{1 \times T},$$

$$\Psi^s = m^s(F^s), \quad \Psi^s \in \mathbb{R}^{(N+1) \times T}.$$  

(2)  

Subsequently, the outputs of two streams are weighted to get the final SAPs and SCPs as

$$\varphi = \alpha \varphi^{rgb} + (1 - \alpha) \varphi^{flow},$$

$$\Psi = \alpha \Psi^{rgb} + (1 - \alpha) \Psi^{flow},$$

(4)  

where $\alpha = 0.5$ by default in our experiments. We implement $m_{fg}^s$ with a fully-connected (FC) layer followed by a sigmoid activation function. And $m^s$ is implemented with an FC layer.

To train $m_{fg}^s$ and $m^s$ with only video-level label, video-level prediction is needed. Therefore, we calculate the video-level foreground feature as

$$f_{fg}^s = \frac{1}{T} \sum_{t=1}^{T} \varphi^s(t) F^s(t), \quad f_{fg}^s \in \mathbb{R}^D.$$  

Similarly, the video-level background feature is obtained by

$$f_{bg}^s = \frac{1}{T} \sum_{t=1}^{T} (1 - \varphi^s(t)) F^s(t), \quad f_{bg}^s \in \mathbb{R}^D.$$  

(7)

After obtaining $f_{fg}^s$ and $f_{bg}^s$, we feed them into $m^s$ to obtain the video-level prediction, i.e., the foreground prediction ($p_{fg}^s \in \mathbb{R}^{N+1}$) and background prediction ($p_{bg}^s \in \mathbb{R}^{N+1}$), defined as

$$p_{fg}^s = m^s(f_{fg}^s), \quad p_{bg}^s = m^s(f_{bg}^s).$$  

(8)

Given video-level predictions in Eq. (8), the FB branch can be trained via regular cross-entropy loss. For $p_{fg}^s$, its label is $y$, where $y(n) = 1$ if $V$ contains the $n$-th action category, as shown in Figure 3. While for $p_{bg}^s$, assuming that all videos contain background snippets, its label is always $y(0) = 1$ and $y(n) = 0$, $n = 1, 2, ..., N$. 


### 3.3 Action-Context Branch

The attention mechanism trained by \( y \) will be distracted by context because both action and context can support video-level classification. To avoid such distraction, after distinguishing the foreground from background, we further separate action and context within the foreground to locate the actual action instances in this section.

The inputs of the AC branch are features from two streams (\( F^s \) obtained in Section 3.2) and SAPs (\( \varphi \) obtained in Section 3.2). The AC branch consists of three sub-modules, i.e., latent components generation, latent components combination, and action-context separation.

**Latent Components Generation.** We introduce the concept of positive component (\( L^+ \) \( \in \mathbb{R}^{1 \times T} \)) and negative component (\( L^- \) \( \in \mathbb{R}^{1 \times T} \)) to characterize foreground, action, and context. Assuming the foreground is represented by two latent components, we define the one corresponding to the actual action as positive component, while the other one as negative component. They are obtained similarly as the SAPs in Eq.(2), by feeding features into positive module (\( m^+ \)) and negative module (\( m^- \))

\[
L^+ = m^+_s(F^s), \quad L^- = m^-_s(F^s). \tag{9}
\]

\( m^+_s \) and \( m^-_s \) share the same architecture (parameters are not shared), with two temporal convolution (Conv1d) layers followed by a ReLU and a sigmoid activation function for the first and the second layer, respectively.

**Latent Components Combination.** Given \( L^+ \) and \( L^- \), we use the combination of them to construct the snippet-level foreground attention (\( \hat{\varphi}_g \in \mathbb{R}^{1 \times T} \)), action attention (\( \hat{\varphi}_a \in \mathbb{R}^{1 \times T} \)), and context attention (\( \hat{\varphi}_c \in \mathbb{R}^{1 \times T} \)). Specifically, for each stream, we have

\[
\hat{\varphi}_g = \sigma(L^+ + L^-), \tag{10}
\]

\[
\hat{\varphi}_a = \sigma(L^+), \tag{11}
\]

\[
\hat{\varphi}_c = \sigma(L^- - L^+), \tag{12}
\]

where \( \sigma(\cdot) \) denotes the sigmoid function. Subsequently, the outputs from two streams are fused by weighted average similar to Eq.(4),

\[
\hat{\varphi}_z = \alpha \hat{\varphi}^g_z + (1 - \alpha) \hat{\varphi}^a_z \hat{\varphi}^c_z, \quad \hat{\varphi}_z \in \{fg, a, c\}, \tag{13}
\]

where \( \hat{\varphi}_z \in \mathbb{R}^{1 \times T} \). For notational simplicity, we use subscript “\( z \)” to denote either “fg”, “a” or “c” if necessary. By substituting the subscript “\( z \)” with “fg/alc”, \( \hat{\varphi}^g_\| \hat{\varphi}_a \| \hat{\varphi}_c \) are obtained following Eq.(13).

Instead of directly imposing simple constraints like foreground and background following (Nguyen, Ramanan, and Fowlkes 2019), i.e., \( \hat{\varphi}^c = 1 - \hat{\varphi}^a \), we adopt the combinations of \( L^+ \) and \( L^- \) to characterize \( \hat{\varphi}^a \) and \( \hat{\varphi}^c \) individually. We compared different approaches to obtain \( \hat{\varphi}^c \) in supplementary material.

**Action-Context Separation.** After obtaining \( \hat{\varphi}^g_\| \hat{\varphi}_a \| \hat{\varphi}_c \), we can start the action-context separation by leveraging label with auxiliary context categories (i.e., \( \mathcal{Y} \in \mathbb{R}^{2N} \) introduced in Section 3.1). First of all, we select all temporal indices corresponding to foreground snippets as

\[
I = \{t \mid \varphi(t) > 0.5\}, \quad |I| = T', \tag{14}
\]

where \( |\cdot| \) denotes the cardinality (number of elements). Subsequently, the video-level feature representations of foreground, action and context are obtained as

\[
\hat{f}_z = \frac{1}{T'} \sum_{t \in I} \hat{\varphi}_z(t) F(t), \quad z \in \{fg, a, c\}, \tag{15}
\]

where \( \hat{f}_z \in \mathbb{R}^{2D \times 1} \) and \( F(t) = \langle F^{rgb}(t), F^{flow}(t) \rangle \) (\( F(t) \in \mathbb{R}^{2D \times 1} \)) is the concatenated feature from both streams and \( \langle \cdot \rangle \) means concatenation. By substituting the subscript “\( z \)” with “fg/alc”, \( \hat{f}_{fg}, \hat{f}_a \) and \( \hat{f}_c \) are calculated following Eq.(15).

Afterwards, they are fed into the action-context classification module \( \hat{m} \) to get the video-level action-context prediction as

\[
\hat{p}_z = \hat{m}(\hat{f}_z), \quad \hat{p}_z \in \mathbb{R}^{2N}, \quad z \in \{fg, a, c\}. \tag{16}
\]
Different from the video-level prediction from FB branch (i.e., \( \hat{p}_c \in \mathbb{R}^{N+1} \) in Eq.(8)), \( \hat{p}_c \in \mathbb{R}^{2N} \) provides predictions on both action and context categories. Specifically, if the video contains the \( n \)-th category, the label for \( \hat{p}_g \) is \( \hat{y} = [y_a(1), \ldots, y_a(N), y_c(1), \ldots, y_c(N)] \), where \( y_a(n) = y_c(n) = 0.5 \). While for \( \hat{p}_a \) and \( \hat{p}_c \), the labels are \( (y_a(n) = 1, y_c(n) = 0) \) and \( (y_a(n) = 0, y_c(n) = 1) \), respectively, as shown in Figure 3. After obtaining \( \hat{p}_a \) and the corresponding labels, the AC branch is also trained via regular cross-entropy loss.

Applying \( \hat{m} \) to each snippet, the snippet-level action-context predictions are obtained as
\[
\hat{\Psi}' = \hat{m}(\hat{F}), \hat{\Psi}' \in \mathbb{R}^{2N \times T},
\]
where \( \hat{F} \in \mathbb{R}^{2D \times T} \) is the concatenated feature. \( \hat{\Psi}' \) is leveraged to promote the action and suppress the context, by defining an “action-context offset (\( \hat{\Psi} \in \mathbb{R}^{N \times T} \))” as
\[
\hat{\Psi}'(n,t) = \begin{cases} \hat{\Psi}'(n,t) - \hat{\Psi}'(2n,t) & \text{if } t \in \mathbb{I}, \\ 0 & \text{otherwise}, \end{cases}
\]
where \( \hat{\Psi}'(n,t) \) (or \( \hat{\Psi}'(2n,t) \)) is the prediction of the \( n \)-th action (or corresponding context) of the \( t \)-th snippet. Intuitively, \( \hat{\Psi}'(n,t) \) means “offsets” for the \( n \)-th class of the \( t \)-th snippet, compared the prediction of action (\( \hat{\Psi}'(n,t) \)) with context (\( \hat{\Psi}'(2n,t) \)).

In summary, the AC branch outputs snippet-level action score (\( \hat{\varphi}_a \in \mathbb{R}^{1 \times T} \)) and the action-context offset (\( \hat{\Psi} \in \mathbb{R}^{N \times T} \)) for the subsequent localization task.

### 3.4 Additional Losses

In addition to the regular cross-entropy losses, more constraints are required to train the ACSNet successfully, since there are neither temporal annotations nor action/context annotations available. In this section, we introduce two additional losses to provide extra guidance for ACSNet training, i.e., \( L_g \) and \( L_{mse} \).

For guidance loss \( L_g \), due to the lack of ground truth labeled action or context categories, confusion between action and context (e.g., \( \hat{\varphi}_a \) and \( \hat{\varphi}_c \), \( L_a^+ \) and \( L_c^+ \)) will occur due to symmetry. Therefore, additional guidance should be introduced to distinguish action from context, which is achieved by minimizing \( L_g \). Specifically, the differences between two streams are leveraged. We adopt weighted binary logistic regression loss function \( L_r \) to guide \( \hat{\varphi}_a \) and \( \hat{\varphi}_c \), where \( L_r(\hat{p}, \hat{q}) \) is denoted as
\[
L_r(\hat{p}, \hat{q}) = - \sum_{i=1}^{l^+} \frac{q_i \cdot \log(p_i)}{l^+} + \frac{(1-q_i) \cdot \log(1-p_i)}{l^-},
\]
where \( \hat{p}, \hat{q} \in \mathbb{R}^{1 \times T} \) and \( q \) is a binary vector indicating positive and negative samples (snippets). \( \hat{p} \) is the prediction to be regressed. \( l^+ = \sum q_i \) and \( l^- = \sum (1-q_i) \). For action attention \( \hat{\varphi}_a \), positive time index set (\( P_a \)) and negative time index set (\( N_a \)) are defined as
\[
P_a = \{ t \mid \hat{\varphi}_a^{rgb}(t) > \theta_h & \& \hat{\varphi}_a^{flow}(t) > \theta_h \},
\]
\[
N_a = \{ t \mid \hat{\varphi}_a^{rgb}(t) < \theta_l & \& \hat{\varphi}_a^{flow}(t) < \theta_l \},
\]
where \( \theta_h \) and \( \theta_l \) indicate high and low thresholds, respectively. Intuitively, the snippets with high/low thresholds on both streams are regarded as positive/negative samples for action snippets. For context attention \( \hat{\varphi}_c \), we assume context contains scenes (excluding action instances), so that the corresponding positive/negative snippet index sets are defined as
\[
P_c = \{ t \mid \hat{\varphi}_c^{rgb}(t) > \theta_h & \& \hat{\varphi}_c^{flow}(t) < \theta_l \},
\]
\[
N_c = P_a \cup N_a.
\]
Subsequently, the guidance loss \( L_g \) is calculated as
\[
L_g = L_r(\hat{\varphi}_a', [1(|P_a|), 0(|N_a|)]) + L_r(\hat{\varphi}_c', [1(|P_c|), 0(|N_c|)]),
\]
where \( G(\cdot) \) is a Gaussian smoothing function. Finally, the AC branch is trained by minimizing the total loss \( L \), calculated as
\[
L = L_{ch} + \lambda(L_{mse} + L_g),
\]
where \( L_{ch} \) is the sum of cross-entropy losses mentioned in Section 3.3. \( \lambda \) is the balancing weight set as 1.

### 4 Localization

After the inference, FB branch outputs SAPs (\( \varphi \in \mathbb{R}^{1 \times T} \)), SCPs (\( \hat{\Psi} \in \mathbb{R}^{(N+1) \times T} \)) and AC branch outputs action score (\( \hat{\varphi}_a \in \mathbb{R}^{1 \times T} \)), action-context offset (\( \hat{\Psi} \in \mathbb{R}^{N \times T} \)). These outputs are leveraged for the TAL task. We first introduce the TAL baseline using only outputs of FB branch. Secondly, we present the contribution of AC branch to the TAL task.

#### 4.1 Localization Baseline

The localization baseline uses only outputs of FB branch. The temporal action proposals are generated by thresholding \( \varphi \) with 0.5. The evaluation (scoring) of temporal action proposals is based on \( \Psi \).

After obtaining a proposal \( P = [t_s, t_e] \), where \( t_s \) and \( t_e \) denote the starting and ending snippet indices, respectively, \( P \) is scored by leveraging the Outer-Inner-Contrastive loss (Shou et al. 2018) as
\[
s(P, v) = \frac{\text{mean}(v(t_s : t_e)) - \text{mean}((v(t_s : t_e) - v(t_e : t_e + \tau)))}{\text{mean}((v(t_s - \tau : t_s), v(t_e : t_e + \tau)))},
\]
where \( v \in \mathbb{R}^{1 \times T} \) is the sequence for scoring. \( \tau = (t_e - t_s)/4 \) denotes the inflation length and \( \text{mean}(\cdot) \) is the averaging function. Specifically, when locating the \( n \)-th action category based on \( \Psi \), we make \( v = v_l = \Psi(\cdot, n) \), which is the predictions of the \( n \)-th action category of all snippets. After obtaining proposals and their scores, the TAL results are collected.
### 4.2 Improving Localization by AC branch

The two critical steps of performing TAL are the generation and evaluation of proposals. The outputs of AC branch can improve both of them. For proposal generation, in addition to thresholding $\varphi$ ($P_1$ in Table 4), we also perform thresholding step on $\hat{\varphi}_n$ and $\hat{\Psi}$ ($P_2$ and $P_3$ in Table 4). Since $\hat{\varphi}_n$ and $\hat{\Psi}$ are both action-aware and less susceptible to the influence of context, the proposals obtained by thresholding them can provide more accurate action boundaries and less context noise.

For proposal evaluation, we can improve the quality of $\hat{\Psi}(n,:)$ by suppressing the context and promoting the action as

$$v_2 = \hat{\Psi}(n,:) + \hat{\Psi}(n,:).$$

By replacing $v$ with $v_2$ in Eq.(28), we can evaluate proposals more accurately by alleviating the influence of context.

In summary, the contribution of AC branch to the TAL is reflected in three aspects, i.e., using its outputs ($\hat{\varphi}_a$ and $\hat{\Psi}$) to improve proposal generation ($P_2$ and $P_3$), and using $\hat{\Psi}$ to improve proposal scoring ($v_2$). These three aspects are validated in Table 4.

### 5 Experiments

In this section, we evaluate the proposed ACSNet via detailed ablation studies to explore the contribution brought by AC branch. Meanwhile, we compare our method with state-of-the-art WS-TAL methods and recent fully-supervised TAL methods on two standard benchmarks.

#### 5.1 Experimental Setting

**Evaluation Datasets.** THUMOS14 dataset (Jiang et al. 2014) provides temporal annotations for 20 action categories, including 200 untrimmed videos from validation set and 213 untrimmed videos from test set. On average, each video contains 15.4 action instances and 71.4% frames are non-action background. Following conventions, the validation and test sets are leveraged for training and testing, respectively. ActivityNet v1.2 & v1.3 (Fabian Caba Heilbron and Niebles 2015) provide temporal annotations for 100 / 200 action categories, including a training set with 4,819 / 10,024 untrimmed videos and a validation set with 2,383 / 4,926 untrimmed videos\(^1\).

**Evaluation metric.** Following the standard evaluation protocol, we evaluate the TAL performance using mean average precision (mAP) values at different levels of IoU thresholds. Specifically, the IoU threshold sets are [0.3 : 0.1 : 0.7] and [0.5 : 0.05 : 0.95] for THUMOS14 and ActivityNet, respectively. Both THUMOS14 and ActivityNet benchmarks provide standard evaluation implementations, which are directly exploited in our experiments for fair comparison.

#### 5.2 Comparisons with State-of-the-Art Methods

As presented in Table 1, the proposed ACSNet outperforms existing WS-TAL methods in terms of mAPs with all IoU threshold settings on THUMOS14 testing set with significant improvement. Also, the proposed ACSNet achieves state-of-the-art on ActivityNet v1.2 and v1.3, as presented in Table 2. However, such performance improvement is not as significant as that on THUMOS14, possibly due to ActivityNet v1.2/v1.3 only has 34.6%/35.7% non-action frames.

---

\(^1\)In our experiments, there are 4, 471/9, 937 and 2, 211/4, 575 videos accessible from YouTube in the training and validation set for ActivityNet v1.2 / v1.3, respectively.
Table 3: Classification and localization evaluation on different snippet sets on THUMOS14 test set. **Classification metric**: Average top1 classification accuracy ($A_1$), and proportion of groundtruth actions ($R_z$) defined in Eq.(30). **Localization metric**: Average mAP under the IoU thresholds from 0.3 to 0.7.

<table>
<thead>
<tr>
<th>Variants</th>
<th>$A_1$ (%)</th>
<th>$R_z$ (%)</th>
<th>mAP(%)@IoU</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{gt}$</td>
<td>91.4</td>
<td>62.4</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>$S_{fg}$</td>
<td>88.6</td>
<td>59.1</td>
<td>38.3</td>
<td>22.4</td>
</tr>
<tr>
<td>$S_c$</td>
<td>91.0</td>
<td>61.5</td>
<td>42.4</td>
<td>92.0</td>
</tr>
<tr>
<td>$S_c$</td>
<td>81.0</td>
<td>53.4</td>
<td>7.0</td>
<td>0.7</td>
</tr>
<tr>
<td>$S_{bg}$</td>
<td>26.7</td>
<td>15.1</td>
<td>0.1</td>
<td>0</td>
</tr>
</tbody>
</table>

per video on average, while THUMOS14 contains 71.4% on average. With lower non-action ratio, the improvement brought by context suppression could be less significant.

5.3 Ablation Study

**Is Context Really Useful for Classification?** We assume that the action-context confusion is caused by both action and context can support the classification, due to the high co-occurrence of them. To validate whether the context snippets estimated by AC branch meet our assumption or not, we collect the foreground/background and action/context snippets as follows. The $t$-th snippet belongs to foreground if $\varphi(t) > 0.5$ and otherwise it belongs to background. Among foreground snippets, if $\varphi_a(t) > 0.5$, the $t$-th snippet is assigned as action and otherwise as context. For reference, we also collect all ground truth snippets. Therefore, five snippet sets are collected, noted as $S_{fg}$, $S_{bg}$, $S_a$, $S_c$, and $S_{gt}$, respectively.

Regarding the concatenated snippets as temporal proposals among each set, these snippet sets can be evaluated in both localization and classification tasks, as summarized in Table 3. For localization, we use the metrics introduced in Section 5.1 with $v = v_1 = \Psi(n, :)$ for proposal evaluation, since $\Psi(n, :)$ does not bias on either action or context. For classification, two metrics are adopted, i.e., the average top1 classification accuracy ($A_1$) and proportion of groundtruth actions defined as

$$R_z = \frac{\sum_{t \in S_{gt}} \Psi(n_{gt}, t) \sum_{n=1}^{N} \Psi(n, t), z \in \{fg, bg, a, c, gt\}}{\sum_{n=1}^{N} \sum_{t \in S_{gt}} \Psi(n, t)},$$

where $n_{gt}$ means the groundtruth category and $\Psi(n, t)$ is the $t$-th snippet’s classification prediction on the $n$-th class.

As presented in Table 3, context snippets $S_c$ contain more useful information compared with $S_{bg}$, indicated by the much better classification accuracy. However, in terms of localization task, both $S_c$ and $S_a$ perform poorly, which matches our assumption of context, i.e., snippets that can support classification but contain no actual actions.

**TAL Contribution of AC branch.** The contribution of the proposed AC branch towards the TAL task is reflected in three aspects as summarized in Section 4.2. To validate these three aspects, five ablated variants are evaluated in this section. For the convenience of the discussion, we define the following notations for experiment settings. For proposal generation settings, $P_1/P_3/P_4$ are defined as: Thresholding $\varphi_{\hat{a}}/\varphi_{\hat{c}}/\varphi(n, :)$ with $0.5/0.5/0$ to generate temporal action proposals for all/all/n-th action class. For proposal scoring settings, $S_1/S_2$ are defined as: Using $v_1/v_2$ as the $v$ in Eq.(28) for proposal evaluation. Therefore, the usage of $P_2$ reflects the contribution of $\varphi_a$ in aspects of proposal generation. The usage of $P_3$ and $S_2$ reflect the contribution of $\Psi$ in aspects of proposal generation and evaluation, respectively. The contribution of $P_2/P_3/S_2$ to TAL is evaluated individually below, as presented in Table 4.

With $P_1$ and $S_1$, the #0 variants are the baseline methods, which depend on FB branch and are non-related to the AC branch. Noted that baselines show super sensitivity towards hyper-parameter $\alpha$, we choose the best one ($\alpha = 0.4$) for comparison below. In contrast, all the other ablated variants are with simple average two-stream fusion ($\alpha = 0.5$). Comparison between baseline (#0) and #1 shows the contribution solely from $P_2$. Similarly, the contributions solely from $P_3$ and $S_2$ can be validated by the comparisons between #2 and #4, #1 and #2, respectively. Quantitatively, $P_2/P_3/S_2$ take up 33.3%/22.2%/44.5% of the performance gain upon baseline.

Besides, compared with #4 and #5, an obvious performance drop is observed, indicating the localization result from FB branch has been burden for the final localization. Without the proposals from FB branch, and with the help of $\varphi_a$ and $\Psi$ on proposal generation and evaluation, “#4” achieves the best localization performance.

### Table 4: Ablation studies of ACSNet on THUMOS14

<table>
<thead>
<tr>
<th>Variants</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>mAP(%)@IoU</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>#0($\alpha = 0.4$)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>31.4</td>
<td>17.0</td>
</tr>
<tr>
<td>#0($\alpha = 0.4$)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>31.4</td>
<td>17.0</td>
</tr>
<tr>
<td>#1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>25.4</td>
<td>22.4</td>
</tr>
<tr>
<td>#2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>19.5</td>
<td>19.5</td>
</tr>
<tr>
<td>#3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>13.6</td>
<td>13.6</td>
</tr>
<tr>
<td>#4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>9.2</td>
<td>9.2</td>
</tr>
<tr>
<td>#5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>4.8</td>
<td>4.8</td>
</tr>
</tbody>
</table>

6 Conclusions

We propose an ACSNet for weakly-supervised temporal action localization, which can separate action and context with only video-level categorical labels. This is achieved by characterizing foreground/action/context as combinations of positive and negative latent compositions. ACSNet significantly outperforms existing WS-TAL methods on three standard datasets, i.e., THUMOS14, ActivityNet v1.2 and v1.3. Moreover, ACSNet achieves competitive performance even compared with recent fully-supervised TAL methods. Experimental results validate the significance of action-context separation and the superiority of the proposed pipeline.
7 Acknowledgment

This work was supported partly by National Key R&D Program of China Grant 2018AAA0101400, NSFC Grants 61629301, 61773312, and 61976171, China Postdoctoral Science Foundation Grant 2019M653642, Young Elite Scientists Sponsorship Program by CAST Grant 2018QNRC001, and Natural Science Foundation of Shaanxi Grant 2020QJ-069.

References


