Leaky Gated Cross-Attention for Weakly Supervised Multi-Modal Temporal Action Localization

Jun-Tae Lee Sungrack Yun Mihir Jain
Qualcomm AI Research*
{juntlee,sungrack,mijain}@qti.qualcomm.com

Abstract

As multiple modalities sometimes have a weak complementary relationship, multi-modal fusion is not always beneficial for weakly supervised action localization. Hence, to attain the adaptive multi-modal fusion, we propose a leaky gated cross-attention mechanism. In our work, we take the multi-stage cross-attention as the baseline fusion module to obtain multi-modal features. Then, for the stages of each modality, we design gates to decide the dependency on the other modality. For each input frame, if two modalities have a strong complementary relationship, the gate selects the cross-attended feature, otherwise the non-attended feature. Also, the proposed gate allows the non-selected feature to escape through it with a small intensity, we call it leaky gate. This leaky feature makes effective regularization of the selected major feature. Therefore, our leaky gating makes cross-attention more adaptable and robust even when the modalities have a weak complementary relationship. The proposed leaky gated cross-attention provides a modality fusion module that is generally compatible with various temporal action localization methods. To show its effectiveness, we do extensive experimental analysis and apply the proposed method to boost the performance of the state-of-the-art methods on two benchmark datasets (ActivityNet1.2 and THUMOS14).

1. Introduction

Temporal action localization has been an essential task in computer vision with its great importance for diverse applications in video understanding. In a weakly-supervised setting, when only video-level labels are available, most of the expensive and time-consuming frame-level annotation is avoided. Hence, a plenty of works [9, 13, 14, 19, 22–24, 26, 27, 34, 41] have studied the temporal action localization task with weak supervision.

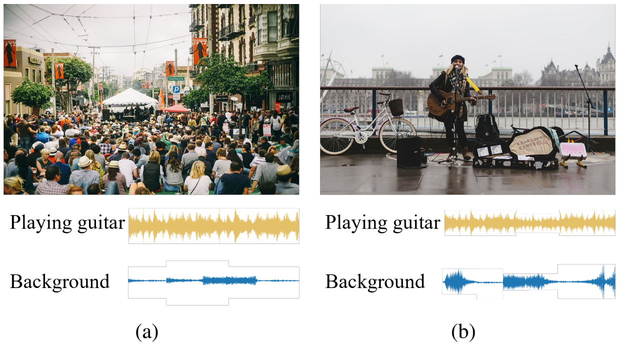


Figure 1. When one of the modalities is weak or dominated by the background noise/clutter, it can contaminate the fused signal. (a) Audio is clear but visual clutter affects the fusion. (b) Background noise in audio signal deteriorates the fused representation compared to only visual representation. Here, conditional to the input sequence, gating the impact of modalities provides robustness to these practical scenarios.

To address the challenging scenario, most of the existing methods depend on fusion of RGB and optical flow features, but the fusing strategies are rather simple like concatenation. These simple strategies are often insufficient to capture and fuse complementary information of different modalities. More recently, learning models from audio-visual inputs is also explored [13,31,36,37] with a dedicated fusion module. Nevertheless, there has been less effort to collaboratively combine different modalities for this task.

Recently, Lee *et al.* [13] have developed the cross-attention mechanism, and repeatedly applied it, in order to enhance both audio and visual features from their complementary relationship. This results in a much improved fusion by modeling the dependency of one modality on the other. But what if one of the modalities is noisy or restrained? Modality that is weak or dominated by the background noise/clutter can potentially contaminate the fused representation. This is illustrated in Figure 1(a) when audio is clear but visual clutter affects the fusion, and Figure 1(b) when background noise in audio deteriorates the fused representation compared to only visual representation. Similar

[*] Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

inconsistencies can sometimes occur in RGB and optical flow fusion as well. For instance, appearance cues that are also present outside the duration of action are better not relied upon. Therefore, instead of indiscriminately integrating information from multiple modalities, it is important to decide how or when to integrate the multi-modal information for accurate temporal action localization.

To attain the adaptive multi-modal fusion, we propose a leaky gated cross-attention mechanism. In our work, we take the multi-stage cross-attention [13] as the baseline fusion module to obtain multi-modal features. For each modality, we design multiple gates over the stages to decide the dependency on the other modality. For each input frame, if two modalities have strong complementary relationship, the gate selects the cross-attended feature, otherwise the non-attended feature. Also, the proposed gate allows the non-selected feature to leak with a small intensity, hence the name *leaky gate*. This leaky feature provides effective regularization by acting as noise signal to the selected feature. We experimentally illustrate this. Our leaky gating provides necessary flexibility to cross-attention to learn to limit the impact of a modality when it is weak or unhelpful. Thus making the leaky gated cross-attention more robust when the modalities have weak complementary relationship.

The proposed leaky gated cross-attention provides a modality fusion module with a key role of collaboratively and adaptively fusing two modalities. Moreover, such fusion module is generally compatible with various temporal action localization methods. To show its effectiveness, we apply leaky gated cross-attention with four recent methods [13, 14, 19, 22] and boost the performance to beyond state-of-the-art on two benchmark datasets (ActivityNet1.2[1] and THUMOS14[2]). In summary, we make the following contributions:

- We propose a leaky gated cross-attention that adaptively selects the better one between the cross-attended and non-attended features.
- By marginally leaking the non-selected feature, the leaky gating gives the regularization effect on the selected feature.
- Detailed experimental analysis is done for each component of the approach on two temporal action localization datasets, where we achieve state-of-the-art performance.

2. Related Work

Weakly-supervised action localization. Many attempts have been made to solve temporal action localization with weak supervision. Most of them focus on utilizing RGB and optical flow as input. In [34], Wang *et al.* learned attention weights on pre-cut video segments using a temporal softmax function and thresholded the attention weights to generate action proposals. This was extended by Nguyen *et al.* [23] by introducing a class-agnostic attention model with sparsity constraints. To reduce classifier's dependence on specific instances, Singh and Lee [28] developed a technique to randomly hide several frames during training in order to force the network respond to multiple relevant parts. Pual *et al.* [26] introduced a co-activity similarity loss to enforce the feature similarity for video pairs with a common class. Narayan *et al.* [22] devised three loss functions to ensure the separability of instances at local-level, to enhance discriminability of action categories at global-level, and to delineate adjacent action sequences. Nguyen *et al.* [24] proposed attention modules to describe both foreground and background frames. In [14], background frames are modeled as out-of-distribution samples, and then separated from foreground action frames by maximizing the entropy of action probability distribution from background frames. Jain *et al.* [9] segmented a video into interpretable fragments, called ActionBytes, and used it to generate action proposals. To distinguish action and near-action snippets, Shi *et al.* [27] developed the class-agnostic frame-wise probability conditioned on the attention using conditional variational auto-encoder, and Ma *et al.* [19] learned a separate class-agnostic model to predict if an instance includes any classes of target actions. Luo *et al.* [18] exploited an expectation-maximization on the multi-instance learning where the key instance is formulated by a hidden variable.

Recently, more works [15, 31, 37] have attempted to fuse audio and visual modalities to localize actions. These performed well on trimmed videos containing audio-visual events with strong audio cues, such as playing guitar, *etc.* Lee *et al.* [13] took it further to localize unconstrained actions in untrimmed video by proposing multi-stage cross-attention mechanism to collaboratively combine audio and visual features. In our work, we exploit their cross-attention as the baseline module fuse multiple modalities (both audio-visual and RGB-flow pairs) in a more adaptive manner.

Conditional computation via gating Conditional computation in neural networks aims to adaptively allocate the components of the model (e.g. layers, sub-networks, etc.) depending on inputs. The conditional computation-based methods can be grouped into two categories according to the objectives. The first focuses on reducing the computational cost by dynamically deciding the topology of networks. In [32, 33], the residual connection is gated in each residual block conditioned on the input. In [12, 30], the input sample can exit the network early via the intermediate classifiers. In [39], considering the spatial redundancy of input samples, sub-networks with hierarchical depths and input resolution are used together with early exit strategy.

[1] http://activity-net.org/download.html

[2] http://crcv.ucf.edu/THUMOS14/download.html

In [5], a cascade of gating modules are used to determine when to stop further processing of the video frames.

The second line of works focuses on the accurate inference. In [2, 35, 38], to obtain the descent unified representation of sequential input, the weight of each element is computed by gating functions. In [29], the gates locally filter out the less informative spatial regions before the spatio-temporal fusion. In [3], to relatively suppress a weak modality, the weight of a modality is computed by the complement of that of the other modality. In this field, the gating functions are mostly designed by soft or hard gates, where the latter is more efficient and often more effective too. In our work, we propose the leaky gate, which is a hard gate with a leakage to further exploit the less informative modality for regularization.

3. Method

In this section, we introduce the proposed leaky gated cross-attention mechanism for collaborative multi-modal fusion in weakly-supervised temporal action localization. Our framework for two-stage case is illustrated in Figure 2.

Problem setting: As in the most previous methods, we consider two input modalities $\mathcal{M}$ and $\mathcal{N}$ in this task. For an input video, L non-overlapping snippets are uniformly sampled. Then, for each modality, a pre-trained network extracts snippet-wise features by:

$$X_{\mathcal{M}} = (\mathbf{x}^l_{\mathcal{M}})^L_{l=1}, \quad X_{\mathcal{N}} = (\mathbf{x}^l_{\mathcal{N}})^L_{l=1} \tag{1}$$

where $\mathbf{x}^l_{\mathcal{M}} \in R^d_{\mathcal{M}}$ and $\mathbf{x}^l_{\mathcal{N}} \in R^d_{\mathcal{N}}$ denote the feature representation of snippet l for modalities $\mathcal{M}$ and $\mathcal{N}$, respectively. Thus, $X_{\mathcal{M}} \in \mathbb{R}^{d_{\mathcal{M}} \times L}$ and $X_{\mathcal{N}} \in \mathbb{R}^{d_{\mathcal{N}} \times L}$.

The weakly-supervised action localization network, $f : (X_{\mathcal{M}}, X_{\mathcal{N}}) \rightarrow \mathcal{Y}$, outputs a sequence of snippet-level class-activation scores often referred as class activation sequence. This can be divided into two modules: multi-modal fusion module $r(\cdot)$ and localization head $h(\cdot)$, so that $f(X_{\mathcal{M}}, X_{\mathcal{N}}) = h(r(X_{\mathcal{M}}, X_{\mathcal{N}}))$. The fusion module largely determines how effectively the two modalities are combined, while the localization head accommodates a combination of classification and localization losses. In testing phase, the class activation sequences are post-processed to temporally localize action instances, which is a *de facto* standard of this task.

In our work, we focus on developing an effective multi-modal fusion module $r(\cdot)$ which is generally applicable to diverse localization heads.

3.1. Multi-stage cross-attention

As preliminary, we briefly describe our baseline fusion module, multi-stage cross-attention mechanism [13].

In a single stage, to encode inter-modal information and also preserve the exclusive and meaningful intra-modal characteristics, features are separately learned for each modality under constraints from the other modality. To this end, the inter-modal relevance is measured by the cross-correlation which is computed using a learnable matrix W:

$$\Lambda = X^T_{\mathcal{M}} W X_{\mathcal{N}} \tag{2}$$

In the cross-correlation matrix, a high correlation coefficient means that the pair of the corresponding features in different modalities are highly relevant. Based on this, the cross attention weights $A_{\mathcal{M}}$ and $A_{\mathcal{N}}$ are generated by column-wise soft-max of Λ and Λ^T, respectively. Then, for each modality, the attention weights are used to re-weight the snippet features. Formally, the attention-weighted features $\tilde{X}_{\mathcal{M}}$ and $\tilde{X}_{\mathcal{N}}$ are represented by:

$$\tilde{X}_{\mathcal{M}} = X_{\mathcal{M}} A_{\mathcal{M}} \quad \text{and} \quad \tilde{X}_{\mathcal{N}} = X_{\mathcal{N}} A_{\mathcal{N}} \tag{3}$$

When multiple stages are used, the cross-attention is repeatedly applied. To prevent the over-suppression of original modality-specific characteristics, the dense skip-connection [8] is exploited. Then, at stage t, the attended features for two modalities are obtained by:

$$X^{(t)}_{\text{att},\mathcal{M}} = \tanh\Big(\sum_{i=0}^{t-1} X^{(i)}_{\text{att},\mathcal{M}} + \tilde{X}^{(t)}_{\mathcal{M}} \Big) \tag{4}$$

and

$$X^{(t)}_{\text{att},\mathcal{N}} = \tanh\Big(\sum_{i=0}^{t-1} X^{(i)}_{\text{att},\mathcal{N}} + \tilde{X}^{(t)}_{\mathcal{N}} \Big) \tag{5}$$

where $X^{(0)}_{\text{att},\mathcal{M}}$ and $X^{(0)}_{\text{att},\mathcal{N}}$ are $X_{\mathcal{M}}$ and $X_{\mathcal{N}}$, respectively. $\tanh(\cdot)$ denotes the hyperbolic tangent activation function.

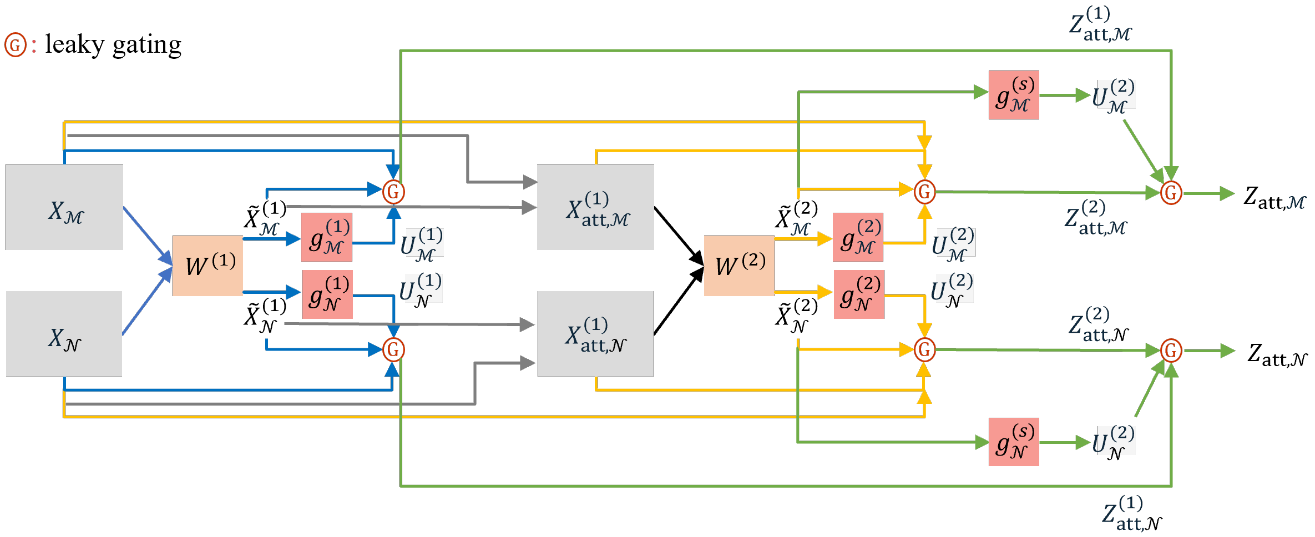
At the last stage t_{e}, the final attended features are concatenated to yield multi-modal features as:

$$X_{\text{att}} = [\, X^{(t_{\text{e}})}_{\text{att},\mathcal{M}};\, X^{(t_{\text{e}})}_{\text{att},\mathcal{N}} \,] \tag{6}$$

3.2. Leaky gating for multi-stage cross-attention

Considering that the two modalities may be incompatible for fusion in several frames, it is not effective to treat them uniformly over every stage. To that end, we need a mechanism to control the impact of each modality via skip-connections and over stages. To adaptively decide how/when to fuse two modalities, we develop a gating controller *leaky gate* for cross-attention. Following [13], we set t_{e} as two. Then, we can consider two kinds of gating: dense skip-connection and stage gates.

Gate design: We design an efficient gating layer with a single fully-connected (fc) layer. To implement gating, we activate the output of fc layer using soft-max with small temperature [7, 21]. Specifically, the temperature is set as 0.1, experimentally. With the small temperature, the gate output is almost 1 on the selected *major* path, and close to 0 in other *leakage* paths. Hence, the unselected features leak



G: leaky gating
$Z^{(1)}_{\text{att},\mathcal{M}}$
$X_\mathcal{M}$
$X_\mathcal{N}$
$W^{(1)}$
$\tilde{X}^{(1)}_\mathcal{M}$
$\tilde{X}^{(1)}_\mathcal{N}$
$g^{(1)}_\mathcal{M}$
$g^{(1)}_\mathcal{N}$
$U^{(1)}_\mathcal{M}$
$U^{(1)}_\mathcal{N}$
$X^{(1)}_{\text{att},\mathcal{M}}$
$X^{(1)}_{\text{att},\mathcal{N}}$
$W^{(2)}$
$\tilde{X}^{(2)}_\mathcal{M}$
$\tilde{X}^{(2)}_\mathcal{N}$
$g^{(2)}_\mathcal{M}$
$g^{(2)}_\mathcal{N}$
$U^{(2)}_\mathcal{M}$
$U^{(2)}_\mathcal{N}$
$g^{(s)}_\mathcal{M}$
$g^{(s)}_\mathcal{N}$
$U^{(2)}_\mathcal{M}$
$U^{(2)}_\mathcal{N}$
$Z^{(2)}_{\text{att},\mathcal{M}}$
$Z^{(2)}_{\text{att},\mathcal{N}}$
$Z_{\text{att},\mathcal{M}}$
$Z_{\text{att},\mathcal{N}}$
$Z^{(1)}_{\text{att},\mathcal{N}}$

Figure 2. Illustration of the proposed leaky gated cross-attention exemplified on the two-stage case. Colorized arrows denote skip-connection gating on stage-1 (blue) and stage-2 (yellow), and stage gating (green).

through these leakage paths, with a very small intensity. We call such feature *leaky feature*. Acting as a noise signal, it can provide a regularization effect [25]. We empirically observe that when the attended feature is not selected, it provides regularization in training through the leakage path.

Skip-connection gating: Dense skip-connection is helpful to preserve the information of previous stages. To selectively exploit the skip-connections, we add a skip-connection gate at the end of every stage of each modality.

Firstly, for the modality $\mathcal{M}$ on the stage-1, the two-way leaky gate $g^{(1)}_\mathcal{M}$ takes the attention-weighted snippet features $\tilde{X}^{(1)}_\mathcal{M}$ as input, and yields the gating matrix $U^{(1)}_\mathcal{M} \in \mathbb{R}^{2\times L}$. Then, repeating each row of $U^{(1)}_\mathcal{M}$ to a $d_\mathcal{M} \times L$-sized matrix, we obtain the gated feature of the stage-1 by:

$$Z^{(1)}_{\text{att},\mathcal{M}} = ReLU(X_\mathcal{M} \otimes U^{(1)}_{\mathcal{M},0} + \tilde{X}^{(1)}_\mathcal{M} \otimes U^{(1)}_{\mathcal{M},1}) \tag{7}$$

where $U^{(1)}_{\mathcal{M},0}$ and $U^{(1)}_{\mathcal{M},1}$ are the matrices obtained by row repetition, and $\otimes$ denotes the element-wise multiplication.

Next, on the stage-2, we develop the ternary gate $g^{(2)}_\mathcal{M}$ which takes the attention-weighted snippet features $\tilde{X}^{(2)}_\mathcal{M}$ as input. Then, similarly to the stage-1, row-wisely reshaping the gating matrix $U^{(2)}_\mathcal{M} \in \mathbb{R}^{3\times L}$ to $d_\mathcal{M} \times L$-sized matrices, skip-connection gating is performed by:

$$\begin{aligned}\tilde{Z}^{(2)}_{\text{att},\mathcal{M}} &= (X_\mathcal{M} + X^{(1)}_{\text{att},\mathcal{M}} + \tilde{X}^{(2)}_\mathcal{M}) \otimes U^{(2)}_{\mathcal{M},0} \; + \\ &(X_\mathcal{M} + \tilde{X}^{(2)}_\mathcal{M}) \otimes U^{(2)}_{\mathcal{M},1} + (X^{(1)}_{\text{att},\mathcal{M}} + \tilde{X}^{(2)}_\mathcal{M}) \otimes U^{(2)}_{\mathcal{M},2}\end{aligned} \tag{8}$$

and

$$Z^{(2)}_{\text{att},\mathcal{M}} = ReLU(\tilde{Z}^{(2)}_{\text{att},\mathcal{M}}) \tag{9}$$

where $U^{(2)}_{\mathcal{M},0}$, $U^{(2)}_{\mathcal{M},1}$ and $U^{(2)}_{\mathcal{M},2}$ are the matrices obtained by repeating the first, second, and last rows of $U^{(2)}_\mathcal{M}$, respectively. The first term in Eq.8 sums the stage-2 feature with the attended stage-1 feature and the initial unattended feature. Selecting only this term is same as multi-stage cross-attention of [13], ignoring the use of ReLU activation. Second term do not consider the attended feature of stage-1, and third term ignores the unattended features. We tried other variations and found the combination of these three terms empirically best. Hence, the leaky gated cross-attention has multiple options (plus leakage through gates) for the adaptive selection dependent on the input sample. Akin to the modality $\mathcal{M}$, the skip-connection gating is conducted on both stages-1 and stage-2 of the modality $\mathcal{N}$.

Stage gating: To adaptively select stages as well, we devise the stage gating. In the stage gating, taking the last attention-weighted features $\tilde{X}^{(2)}_\mathcal{M}$ as input, the gate $g^{(s)}_\mathcal{M}$ computes the gating matrix $U^{(s)}_\mathcal{M} \in \mathbb{R}^{2\times L}$. Then, the final feature is one of the two skip-connection gated features given by:

$$Z_{\text{att},\mathcal{M}} = Z^{(1)}_{\text{att},\mathcal{M}} \otimes U^{(s)}_{\mathcal{M},0} + Z^{(2)}_{\text{att},\mathcal{M}} \otimes U^{(s)}_{\mathcal{M},1} \tag{10}$$

where $U^{(s)}_{\mathcal{M},0}$ and $U^{(s)}_{\mathcal{M},1}$ are the $d_\mathcal{M} \times L$-sized expanded matrices of rows of $U^{(s)}_\mathcal{M}$, respectively. Similarly, we also perform the stage gating for the modality $\mathcal{N}$. Finally, the multi-modal feature is obtained by the concatenation of stage gated features, which is represented as:

$$r(Z_\mathcal{M}, Z_\mathcal{N}) = [Z_{\text{att},\mathcal{M}}; Z_{\text{att},\mathcal{N}}] \tag{11}$$

From the proposed leaky gating technique, the cross-attention mechanism automatically selects attention-level and strengthens important features for multi-modal temporal action localization.

Table 1. Analysis of the proposed leaky gated cross-attention applied to CAAV [13] on the THUMOS14 dataset. Avg mAP@[0.1:0.1:0.7] scores (%) are reported with the computational costs.

Method	Avg mAP	MFLOPs
single-stage baseline	32.8	1.39
two-stage baseline	35.6	1.65
single-stage baseline + leaky gate	34.1	1.40
two-stage baseline + leaky gate	37.5	1.69

Table 2. Analysis of the proposed leaky gated cross-attention applied to 3C-Net [22] on the THUMOS14 dataset. Avg mAP@[0.1:0.1:0.7] scores (%) are reported with the computational costs.

Method	Avg mAP	MFLOPs
Baseline (w/o cross-attention)	33.1	4.44
+ single-stage naive cross-attention	33.1	3.40
+ single-stage leaky gated cross-attention	34.4	3.41
+ two-stage naive cross-attention	32.4	4.45
+ two-stage leaky gated cross-attention	32.9	4.50

4. Experiments

In this section, we provide experimental analysis and comparative evaluation to show the effectiveness of the proposed method. To this end, we exploit four baselines: 3C-Net [22], W-TAL [14], ASL [19], and CAAV [13]. The proposed leaky gated cross-attention is added before their localization heads.

4.1. Datasets and evaluation method

Datasets: We evaluate our approach on ActivityNet1.2 and THUMOS14 datasets.

THUMOS14 dataset consists of videos temporally annotated for 20 classes. We follow the convention to train on the validation set of 200 videos and evaluate on the test set of 212 videos. A video includes 15.5 instances on average, often with less than a second background between actions.

ActivityNet1.2 dataset contains 4,819 train and 2,383 validation videos, which in the literature is used for evaluation. It is temporally annotated has 100 action classes, with on an average 1.5 action instances per video. The average length of the videos in this dataset is 115 seconds.

Evaluation metric: Following the standard evaluation protocol, we generate the action segments (start and end time) from snippet-wise prediction, and then measure mean average precision (mAP) at different intersection over union (IoU) thresholds. Average of these mAPs are reported as *Avg mAP*.

4.2. Implementation details

To obtain input multi-modal features, we follow the standard of the baselines. We use the I3D network [1] to extract the visual features. The I3D network is pre-trained on Kinetics-400 [11], and the features consist of two components: RGB and optical flow. For the audio features, as in [13], we use the VGG-like network [6], which is pre-trained on AudioSet [4].

For the baselines, we use the authors' source codes with their default setting, such as hyper-parameters, optimizer, learning rate, etc. Then, to apply the proposed leaky gated cross-attention, we add the leaky gate cross-attention on top of the earlier layers of RGB-flow-based baselines [14, 19, 22]. Especially, in the 3C-Net [22] baseline, we replace the 2nd modality-specific fc layers with the proposed leaky gated cross-attention, rather than simply attaching it. In the audio-visual baseline [13], the naive multi-stage cross-attention is already included in the model. Hence, we add the leaky gates on top of the cross-attention for every stage of each modality. Commonly, the resulting features of the leaky gated cross-attention are activated by ReLU, while the cross attended features are activated by tanh as in [13].

4.3. Analysis on leaky gated cross-attention

Impact of leaky gating on cross-attention: We first analyze the impact of applying leaky gating over 1-stage and 2-stage cross-attention for both audio-visual and RGB-flow fusion on THUMOS14 dataset. Varying the number of stages, we apply the leaky gated cross-attention to the audio-visual-based CAAV [13] (in Table 1) and to the RGB-flow-based 3C-Net [22] (in Table 2). The computational costs for various cases are also reported in these tables.

For CAAV, only cross-attention results (without leaky gating) are the baselines taken from [13]. We apply our leaky gating to them and improve the performance of both single stage and two stage baselines by 1.3% and 1.9% in Avg mAP, respectively. Further, the additional computational cost of adding leaky gating is negligible: 0.01 MFLOPs for 1-stage and 0.04 MFLOPs for 2-stage. Also, compared to single-stage baseline, the two-stage requires 0.29 more MFLOPs to obtain 1.9% Avg mAP improvement. Therefore, with the leaky gated cross-attention, impressive gains are obtained for action localization with practically no loss in efficiency.

For 3C-Net, the result without cross-attention is taken from [22]. We add four different fusion modules to the baseline, which are single-stage and two-stage cross-attention models with or without leaky gates. The naive (not gated) single-stage cross-attention does not show an improvement, but with the proposed leaky gated cross-attention, the Avg mAP increases by 1.3% to 34.4%. The performance is degraded with additional second stage due to over-fitting while fusing RGB and Flow (discussed later). However,

Table 3. Impact of applying the proposed leaky gated cross-attention to the four recent action localization methods on the THUMOS14 dataset. The mAPs (%) at different IoU thresholds and Avg mAP across the IoU thresholds are reported. Results improve in all four cases for all the IoU thresholds.

Method	Fusion	mAP@IoU								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	Avg	
3C-Net [22]	RGB-flow	57.1	51.4	41.7	33.7	26.0	15.1	6.4	33.1	
CAAV [13]	Audio-visual	61.6	57.2	45.7	36.3	26.2	15.6	6.4	35.6	
ASL [19]	RGB-flow	67.3	61.2	51.5	41.3	30.6	19.7	11.1	40.4	
W-TAL [14]	RGB-flow	65.5	59.9	51.3	42.4	32.7	20.8	10.4	40.4	
With leaky gated cross-attention										ΔAvg to baseline
Ours (on 3C-Net)	RGB-flow	59.2	53.6	43.6	33.8	26.6	16.1	7.8	34.4	+0.3
Ours (on CAAV)	Audio-visual	63.1	58.3	48.4	38.4	30.1	17.1	7.1	37.5	+1.9
Ours (on ASL)	RGB-flow	68.0	61.3	52.6	42.8	31.4	20.1	11.1	41.0	+0.6
Ours (on W-TAL)	RGB-flow	67.5	61.6	53.4	42.4	34.1	22.3	10.4	41.7	+1.3

Table 4. Impact of applying the proposed leaky gated cross-attention to the two recent action localization methods on the ActivityNet1.2 dataset. The mAPs (%) at different IoU thresholds and Avg mAP across the IoU thresholds are reported. Results improve for CAAV method while remaining similar for 3C-Net for the reasons mentioned in the text.

Method	Fusion	mAP@IoU											
		0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg	
3C-Net [22]	RGB-flow	36.5	33.3	30.0	27.0	23.8	20.5	17.2	13.6	9.4	4.2	21.6	
CAAV [13]	Audio-visual	44.8	42.1	37.8	34.2	30.8	26.7	22.5	15.9	4.0	1.0	26.0	
With leaky gated cross-attention													ΔAvg to baseline
Ours (on 3C-Net)	RGB-flow	36.5	33.1	30.0	27.2	23.9	20.5	17.2	13.2	8.9	4.4	21.5	-0.1
Ours (on CAAV)	Audio-visual	44.1	41.1	39.0	34.8	30.5	26.5	22.3	16.8	10.3	1.2	26.6	+0.6

the leaky gating reduces the degradation by 0.5% in Avg mAP, increasing the robustness to the number of stages. On computational front, again there is practically no loss of efficiency. In fact, the single-stage models need fewer FLOPs than baseline as in 3C-Net we replace two parallel modality-specific fc layers with the cross-attention or leaky gated cross-attention. The two-stage models need slightly higher number of FLOPs.

Efficacy of leaky gated cross-attention: Now, we show the effectiveness of the proposed leaky gated cross-attention by applying it to four different baseline methods: 3C-Net [22], W-TAL [14], ASL [19], and CAAV [13]. We use model with two stages for audio-visual CAAV, while only one stage is used for the other three methods for RGB-flow fusion. This is in accordance with the observations in last paragraph (Tables 1 and 2). For THUMOS14 dataset, we report the mAP scores for different thresholds [0.1:0.1:0.7] and the average (Avg) of the mAPs in Table 3. We see that, for every baseline, the proposed method improves the mAP scores on most thresholds. In terms of Avg mAP, the proposed method boosts the baselines by at least 0.3%. And the maximal gain is 1.9% (on CAAV).

In Table 4, we also show the mAP scores for the thresholds of [0.5:0.55:0.95] and the Avg mAPs on ActivityNet1.2 dataset. Since the hyper-parameters details of ASL [19] and W-TAL [14] are not available for this dataset, we could not reproduce the mAP scores they report. Hence, we only apply the leaky gated cross-attention to the CAAV and 3C-Net baselines. As demonstrated in Table 4, we improve the CAAV baseline by 0.6%, but slightly degrade the 3C-Net by 0.1% in Avg mAP. Note that, for both baselines, the proposed method is more effective on the THUMOS14 dataset rather than the ActivityNet1.2 dataset, especially for RGB-flow fusion. We analyze this in the following paragraph.

Relative significance of modalities: Now we analyze the impact of relative significance of one modality with respect to the other. This is to understand the different behaviour of RGB-flow fusion compared to the audio-visual fusion that we observed in Tables 1, 2, 3 and 4. In the upper part of Table 5, audio-visual fusion is analyzed on THUMOS14 and ActivityNet1.2 datasets. Here, the gain over each individual modality is significant for the cross-attentional fusion of CAAV, which is further improved by our leaky gating. The RGB-Flow fusion is analyzed in the lower part of Table 5. Here, both modalities are strong, and as a result relatively simpler late-fusion of 3C-Net already extracts much advantage of fusion. The gain of applying leaky gating to the 3C-Net method is small for THUMOS14, and diminishes for ActivtiyNet1.2 as the two modalities become even more comparable. Also both be-

Table 5. Analysis on relative significance of modalities for audio-visual fusion (with CAAV [13]) and RGB-flow fusion (with 3C-Net [22]). Avg mAP is computed on IoU thresholds [0.1:0.1:0.7] for THUMOS14 and [0.5:0.05:0.95] for ActivityNet1.2.

Method	Modality	THUMOS14	ActivityNet1.2
CAAV [13]	Audio-only	1.0	7.8
	Visual-only	32.4	22.1
	Audio-Visual	35.6	26.0
Ours (on CAAV)	Audio-Visual	37.5	26.6
3C-Net [22]	RGB-only	21.5	17.9
	Flow-only	30.3	14.7
	RGB-Flow	33.1	21.6
Ours (on 3C-Net)	RGB-Flow	34.4	21.5

longing to the visual domain, RGB and flow are not heterogeneous like audio and visual modalities. From this, we can infer that: (a) Leaky gated cross-attention is more beneficial when one modality is weaker than the other, as greater control over modality contribution through gating makes fusion more robust, and (b) Leaky gated cross-attention is more critical when the two modalities are heterogeneous. The more comparable two modalities become, the less we need the sophisticated fusion module, i.e., fewer stages are enough. Therefore, we use two stages for audio-visual fusion and only one stage for RGB-flow fusion.

Selection rate for paths: To see how adaptively the gates choose among available paths, we measure the selection rate (ζ) which computes how often a particular path is chosen as the major path for all snippets. To this end, our two-stage leaky gated model on top of CAAV is analyzed in Table 6. In the first stage, visual branch always selects unattended features, while for audio modality occasionally (1.6%) attended features are also selected. In the second stage, audio branch solely utilizes the combination of second stage features with the other two: stage-1 attended features and unattended features. While the visual branch selects all three combinations, majorly it focuses on the two combinations with unattended features. As far as stage gating is concerned, selected path always goes up to the second stage. Clearly being mutually heterogeneous, audio and visual modalities show different selection patterns.

We also analyze for RGB-Flow fusion with a single-stage leaky gated model on top of 3C-Net. Here, always the unattended features are chosen for both the modalities. Note that, the baseline also uses non-attended features only, but its performance is lower than the single-stage model with the leaky gated cross-attention in Table 2. This is because through the leakage path, attended features pass with the small intensity, which may be a noise signal, gives regularization effect to the major non-attended feature. This is further addressed in the following paragraph.

Regularization effect of leaky gate: Regularization by noise is a common technique to improve generalization performance of deep neural networks [25]. We verify the effectiveness of the leaky features as noise regularization in the proposed leaky gates in Table 7. We first compare our method with soft-gated cross-attention where all the features pass but are weighed differently. This yields 33.0% while 34.4% is obtained with leaky gating. The hard-gated cross-attention only let the selected feature pass through. To implement the hard gates, we employ the gumbel soft-max scheme [10, 20]. The leaky gate again yields higher Avg mAP of 34.4% compared to 33.9% by hard-gating.

Table 6. Selection rates of gates on the THUMOS14 dataset.

Gate	Path	CAAV [13]	3C-Net [22]
		two-stage ζ (Audio) / ζ (Visual)	single-stage ζ (RGB) / ζ (Flow)
Skip-connec. stage-1	X	98.4% / 100%	100% / 100%
	$\tilde{X}^{(1)}$	1.6% / 0%	0% / 0%
Skip-connec. stage-2	$X + X_{\text{att}}^{(1)} + \tilde{X}^{(2)}$	100% / 52.7%	-
	$X + \tilde{X}^{(2)}$	0% / 44.3%	-
	$X_{\text{att}}^{(1)} + \tilde{X}^{(2)}$	0% / 3.0%	-
Stage	$Z_{\text{att}}^{(1)}$	0% / 0%	-
	$Z_{\text{att}}^{(2)}$	100% / 100%	-

Table 7. Analysis on the regularization effect of the leaky gated cross-attention for 3C-Net [22] on the THUMOS14 dataset. Avg mAP@[0.1:0.1:0.7] scores (%) are reported.

Method	Avg mAP (%)
Baseline (w/o cross-attention)	33.1
Frozen cross-attention	32.4
Gaussian Noise (scale 1.0)	25.4
Gaussian Noise (scale 0.1)	33.8
Hard-gated cross-attention	33.9
Soft-gated cross-attention	33.0
Leaky gated cross-attention	34.4

We also compare the effectiveness of the leaky features with other regularization noises. First, after initializing the learnable matrices W using random Gaussian distribution, we freeze the learnable matrices during training. Then, the attention weighted feature is less informative and noisy. Secondly, rather than using cross-attention mechanism and leaky gates, we simply add the Gaussian random noise to the input features. In this case, we control the noise power with a scaling parameter (1.0 or 0.1) to verify for both high and small noise intensity. As shown in Table 7, the leaky feature of the proposed method improves the baseline by a margin larger than other noises. The gaussian noise with scale 1.0 shows performance even lower than the baseline. From this, we see that since the attention weighted feature conveys useful complementary information, its leakage can give more effective regularization compared to other noises.

Impact of various gates: In this section, we conduct an ablation study to verify the efficacy of each of the skip-connection and stage gating on the two-stage model of

Table 8. Ablation studies on the skip-connection and stage gating on the THUMOS14 dataset. Avg mAP@[0.1:0.1:0.7] scores (%) are reported.

Method	CAAV [13]
Baseline (with two-stage cross-attention)	35.6
+ skip-connection gate	37.0
+ stage gate	37.1
+ skip-connection & stage gates	37.5

Table 9. Comparison of our method with the state-of-the-art action localization methods on the THUMOS14 dataset. The average (Avg) mAP across the IoU thresholds [0.1:0.1:0.7] is reported.

Method	Supervision	Avg mAP (%)
Liu *et al.* [16]	Weak	32.4
3C-Net [22]	Weak	33.1
Nguyen *et al.* [22]	Weak	36.3
DGAM [27]	Weak	37.0
EM-MIL [18]	Weak	37.8
CAAV [13]	Weak	35.6
ASL [19]	Weak	40.4
W-TAL [14]	Weak	40.4
Ours (on 3C-Net)	Weak	34.4
Ours (on CAAV)	Weak	37.5
Ours (on ASL)	Weak	41.0
Ours (on W-TAL)	Weak	**41.7**

CAAV. First, we only use the skip-connection gates ($g_{\mathcal{M}}^{(2)}$ and $g_{\mathcal{N}}^{(2)}$) at the end of the stage-2. Next, to generate the stage-gating only model, we remove the all the skip-connection gating, and use the stage gates $g_{\mathcal{M}}^{(s)}$ and $g_{\mathcal{N}}^{(s)}$. Here, the features of each stage are obtained by (4) and (5). As demonstrated in Table 8, we observe that both skip-connection and stage gates are effective to boost the performance of the baseline. And we obtain the best performance by using the proposed two gates together.

4.4. Comparative evaluation

In this section, we compare the benefits of applying leaky gating with the current state-of-the-art methods on THUMOS14 and ActivityNet 1.2 datasets. Based on the earlier analysis, we apply the leaky gated cross-attention with two stages for the audio-visual method (CAAV) and with one stage for the RGB-flow methods (3C-Net, W-TAL, and ASL). In Table 9, the mAP scores for IoU thresholds [0.1:0.1:0.7] are reported on THUMOS14 dataset. Our method compares favourably to the other methods setting up new state-of-the-art on this dataset for weakly supervised temporal action localization. This shows the effectiveness of the leaky gates to collaboratively fuse two different modalities, be it audio-visual or RGB-flow.

Table 10. Comparison of our method with the state-of-the-art action localization methods on the ActivityNet1.2 dataset. The average (Avg) mAP across the IoU thresholds [0.5:0.05:0.95] is reported.

Method	Supervision	Avg mAP (%)
SSN [42]	Full	26.6
W-TALC [26]	Weak	18.0
3C-Net [22]	Weak	21.6
TSM [40]	Weak	17.1
CleanNet [17]	Weak	21.6
Liu *et al.* [16]	Weak	22.4
DGAM [27]	Weak	24.4
EM-MIL [18]	Weak	20.3
CAAV [13]	Weak	26.0
ASL [19]	Weak	25.8
W-TAL [14]	Weak	25.9
Ours (on 3C-Net)	Weak	21.5
Ours (on CAAV)	Weak	**26.6**

For the ActivityNet1.2 dataset, we report the mAP scores for IoU thresholds [0.5:0.05:0.95]. As demonstrated in Table 10, the proposed leaky gating applied to CAAV method compares favourably to all other weakly-supervised methods and improves the state-of-the-art performance. Furthermore, this performance is comparable to the fully-supervised method, SSN [42].

The better performance for audio-visual fusion and on THUMOS14 dataset compared to ActivityNet1.2, especially for RGB-flow fusion, is in accordance with the analyses done in Section 4.3 and Table 5. We observed that leaky gated cross-attentional fusion is critically important when modalities are heterogeneous or when one of the modalities is significantly weaker, as greater control over modalities through gating makes the fusion more robust.

5. Conclusion

We propose a leaky gated cross-attention for weakly-supervised temporal action localization. We build our approach on a the multi-stage cross-attention. For each modality, we add gates for the skip-connections and the stages to decide its dependency on the other modality. Thus, in addition to collaboratively fusing multiple modalities, our fusion module adaptively selects the better one between the cross-attended and non-attended features. This makes it robust even when the modalities have weak complementary relationship. Further, by letting the non-selected feature leak through with small intensity, the leaky gating provides regularization (by noise) for the selected feature. Finally, our fusion module is compatible with various temporal action localization methods. We demonstrate this by applying it to four recent methods. Each component of the proposed approach is analyzed and validated through extensive experiments. We report our results on two benchmark datasets (ActivityNet1.2 and THUMOS14) and improve the state-of-the-art on both of them.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*. 2017.

[2] Hui Chen, Zijia Lin, Guiguang Ding, Jianguang Lou, Yusen Zhang, and Borje Karlsson. Grn: Gated relation network to enhance convolutional neural network for named entity recognition. In *AAAI*. 2019.

[3] Zuyao Chen, Runmin Cong, Qianqian Xu, and Qingming Huang. Dpanet: Depth potentiality-aware gated attention network for rgb-d salient object detection. 2020.

[4] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*. 2017.

[5] Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibian. Frameexit: Conditional early exiting for efficient video recognition. In *CVPR*. 2021.

[6] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *ICASSP*. 2017.

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS*. 2015.

[8] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*. 2017.

[9] Mihir Jain, Amir Ghodrati, and Cees G. M. Snoek. ActionBytes: Learning from trimmed videos to localize actions. In *CVPR*. 2020.

[10] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*. 2017.

[11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. In *arXiv preprint arXiv:1705.06950*. 2017.

[12] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*. 2015.

[13] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In *ICLR*. 2021.

[14] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *AAAI*. 2021.

[15] Y. Lin, Y. Li, and Y. F. Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP*. 2019.

[16] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *CVPR*. 2019.

[17] Ziyi Liu, Le Wang, Qilin Zhang, Zhanning Gao, Zhenxing Niu, Nanning Zheng, and Gang Hua. Weakly supervised temporal action localization through contrast based evaluation networks. In *ICCV*. 2019.

[18] Zhekun Luo, Devin Guillory, Baifeng Shi, Wei Ke, Fang Wan, Trevor Darrell, and Huijuan Xu. Weakly-supervised action localization with expectation-maximization multi-instance learning. In *ECCV*. 2020.

[19] Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, and Guangwei Yu. Weakly supervised action selection learning in video. In *CVPR*. 2021.

[20] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*. 2017.

[21] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *NIPS*. 2019.

[22] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 3C-Net: Category count and center loss for weakly-supervised action localization. In *ICCV*. 2019.

[23] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *CVPR*. 2018.

[24] Phuc Xuan Nguyen, Deva Ramanan, and Charless C. Fowlkes. Weakly-supervised action localization with background modeling. In *ICCV*. 2019.

[25] Hyeonwoo Noh, Tackgeun You, Jonghwan Mun, and Bohyung Han. Regularizing deep neural networks by noise: Its interpretation and optimization. In *NeurIPS*. 2017.

[26] Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-TALC: Weakly-supervised temporal activity localization and classification. In *ECCV*. 2018.

[27] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative attention modeling. In *CVPR*. 2020.

[28] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*. 2017.

[29] Maitreya Suin and AN Rajagopalan. Gated spatio-temporal attention-guided video deblurring. In *CVPR*. 2021.

[30] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *ICPR*. 2016.

[31] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*. 2018.

[32] Andreas Veit and Serge Belongie. Convolutional networks with adaptive inference graphs. In *ECCV*. 2018.

[33] Thomas Verelst and Tinne Tuytelaars. Dynamic convolutions: Exploiting spatial sparsity for faster inference. In *CVPR*. 2020.

[34] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. UntrimmedNets for weakly supervised action recognition and detection. In *CVPR*. 2017.

[35] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. Gated self-matching networks for reading comprehension and question answering. In *ACL*. 2017.

[36] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *ICCV*. 2019.

[37] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *AAAI*. 2020.

[38] Lanqing Xue, Xiaopeng Li, and Nevin L Zhang. Not all attention is needed: Gated attention network for sequence data. In *AAAI*. 2020.

[39] Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai, and Gao Huang. Resolution adaptive networks for efficient inference. In *CVPR*. 2020.

[40] Tan Yu, Zhou Ren, Yuncheng Li, Enxu Yan, Ning Xu, and Junsong Yuan. Temporal structure mining for weakly supervised action detection. In *ICCV*. 2019.

[41] Yuan Yuan, Yueming Lyu, Xi Shen, Ivor W Tsang, and Dit-Yan Yeung. Marginalized average attentional network for weakly-supervised learning. In *ICLR*. 2019.

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