Boosting Weakly-Supervised Temporal Action Localization with Text Information

Guozhang Li[1],De Cheng[1],Xinpeng Ding[2], Nannan Wang[1]*, Xiaoyu Wang[3], Xinbo Gao[1,4]
[1]Xidian University,[2]The Hong Kong University of Science and Technology,
[3]The Chinese University of Hong Kong (Shenzhen) [4]Chongqing University of Posts and Telecommunications
liguozhang@stu.xidian.edu.cn, dcheng@xidian.edu.cn, xdingaf@connect.ust.hk
nnwang@xidian.edu.cn,fanghuaxue@gmail.com, gaoxb@cqupt.edu.cn

Abstract

Due to the lack of temporal annotation, current Weakly-supervised Temporal Action Localization (WTAL) methods are generally stuck into over-complete or incomplete localization. In this paper, we aim to leverage the text information to boost WTAL from two aspects, i.e., ***(a)*** *the discriminative objective to enlarge the inter-class difference, thus reducing the over-complete;* ***(b)*** *the generative objective to enhance the intra-class integrity, thus finding more complete temporal boundaries. For the discriminative objective, we propose a Text-Segment Mining (TSM) mechanism, which constructs a text description based on the action class label, and regards the text as the query to mine all class-related segments. Without the temporal annotation of actions, TSM compares the text query with the entire videos across the dataset to mine the best matching segments while ignoring irrelevant ones. Due to the shared sub-actions in different categories of videos, merely applying TSM is too strict to neglect the semantic-related segments, which results in incomplete localization. We further introduce a generative objective named Video-text Language Completion (VLC), which focuses on all semantic-related segments from videos to complete the text sentence. We achieve the state-of-the-art performance on THUMOS14 and ActivityNet1.3. Surprisingly, we also find our proposed method can be seamlessly applied to existing methods, and improve their performances with a clear margin. The code is available at https://github.com/lgzlIlllI/Boosting-WTAL.*

1. Introduction

Temporal action localization attempts to temporally localize the action instances of interest in untrimmed videos. Although current fully-supervised temporal action localization methods [5, 26, 42, 51] have achieved remarkable progress, time-consuming and labor-intensive frame-level annotations are required. To alleviate the annotation cost, weakly-supervised temporal action localization (WTAL) methods [15,22,32,35] have gained more attention recently, which only requires efficient video-level annotations.

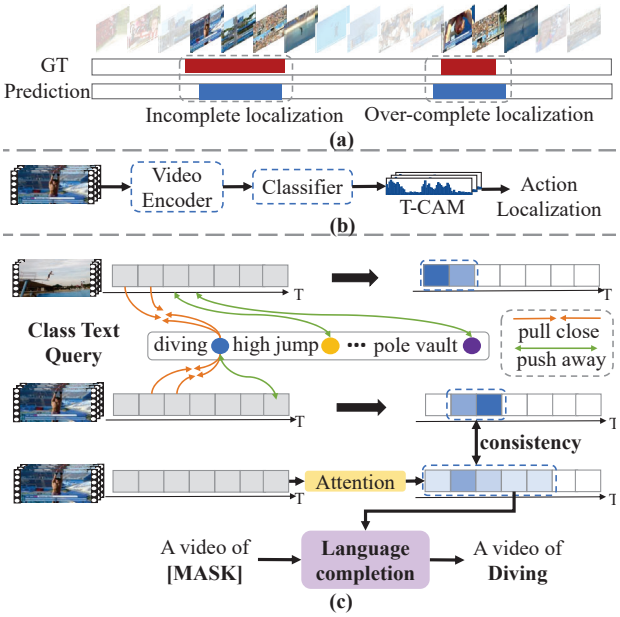

Figure 1. Comparison of our proposed framework with current WTAL methods. (a) Common failures in existing WTAL methods. (b) Existing WTAL model's pipeline. (c) The proposed framework with text-segment mining and video-text language completion, where the depth of color represents the degree of correlation between segments and texts.

With only video-level supervision, existing WTAL methods [15, 22, 35, 45] generally utilize video information to train a classifier, which is used to generate a sequence of class logits or predictions named temporal class activation map (T-CAM). While significant improvement has been achieved, current methods still suffer from two prob-

*Corresponding author

lems, *i.e.*, incomplete and over-complete localization. As shown in Fig. 1 (a), some sub-action with low discriminability may be ignored, while some background segments that contribute to classification can be misclassified as action, causing incomplete and over-complete localization.

Differently from current methods that only utilize the video information, in this paper, we aim to explore the text information to improve WTAL from two aspects: **(a)** the discriminative objective to enlarge the inter-class difference, thus reducing the over-complete; **(b)** the generative objective to enhance the intra-class integrity, thus finding more complete temporal boundaries. For the discriminative objective, we propose a Text-Segment Mining (TSM) mechanism, where the action label texts can be used as queries to mine all related segments in videos. Specifically, we first use the prompt templates to incorporate the class label information into the text query. Without temporal annotations, TSM requires to compare the text query with all segments of the different videos across the dataset, as shown in Fig. 1 (c). During the comparison, the segments that is best matching to the text query would be mined, while other irrelevant segments would be ignored, which is similar to 'matched filter' [43, 50]. In this way, the segments and text queries with the same class from all videos would be pulled close while pushing away others, hence enhancing the inter-class difference.

For different categories of videos, there are some shared sub-actions, *e.g.*, sub-action "Approach" exists in both "High Jump" and "Long Jump" videos. Merely using TSM is too strict to neglect the semantic-related segments, which results in incomplete localization, *e.g.*, neglecting "Approach" segments. To overcome this problem, we further introduce a generative objective named Video-text Language Completion (VLC) which focuses on all semantic-related segments to complete the text sentence. First, we construct a description sentence for the action label of the video and mask the key action words in the sentence, as shown in Fig. 2. Then an attention mechanism is design to collect semantic related segments as completely as possible to predict masked action text via the language reconstructor, which enhances the intra-class integrity. Combining TSM and VLC by a self-supervised constraint, our method achieves the new state-of-the-art on two popular benchmarks, *i.e.*, THUMOS14 [17] and ActivityNet1.3 [1]. Furthermore, we also find our proposed method can be applied into existing methods, and improve their performances with a clear margin.

Our contributions are summarized as three-folds: **(a)** To best of our knowledge, we are the first to leverage text information to boost WTAL. We also prove that our method can be easy to extend to existing state-of-the-art approaches and improve their performance. **(b)** To leverage the text information, we devise two objective: the discriminative objective to enlarge the inter-class difference, thus reducing the over-complete; and the generative objective to enhance the intra-class integrity, thus finding more complete temporal boundaries. **(c)** Extensive experiments illustrate our method outperforms current methods on two public datasets, and comprehensive ablation studies reveal the effectiveness of the proposed objectives.

2. Related Work

Weakly Supervised Temporal Action Localization. Weakly-supervised temporal action localization requires video-level labels only. Due to the lack of precise boundary labels, most advanced WTAL methods [15, 16, 29, 35] fall into a localization-by-classification pipeline to tackle the WTAL task. Erasing-based methods [29, 40, 48, 54] carefully design adversarial erase strategies, which find many less discriminative regions by erasing the most discriminant regions. Metric learning-based methods [12, 30, 33, 35] employ center loss or triple loss to decrease intra-class variations while increasing inter-class difference. In addition, background segments suppression-based methods [16, 22, 23] aim to separate action segments from background segments by setting additional background class to learn background suppression weights. Some pseudo-label-based methods [15, 28, 49] utilize video information to generate pseudo-labels to improve the quality of T-CAMs. Besides, lee *et al*. [21] used audio within the video as an auxiliary information. Existing methods can employ one or more of the above strategies to improve T-CAM quality and improve localization performance. Despite the success of these methods, however, the above strategies only make use of video information, and the semantic information encapsulated in category labels of the text form is not fully explored. In this paper, we design a novel framework consisting of two objectives, *i.e.,* text-segment mining and video-text language completion, to leverage action label text information to boost WTAL.

Self-Supervised Learning. Self-supervised learning leverages unlabeled data to make the model learn intrinsic information from data. Currently, several methods have been proposed to exploit the self-supervised learning strategy to learn better representation when lacking full annotation data. For example, Gong *et.al* [8] proposed self-supervised equivariant transform consistency constraint to realize self-supervised action localization. TSCN [49] and UGCT [47] utilize RGB and optical flow video features for cross-supervision to improve the performance of WTAL. Su *et.al* [41] utilizes temporal multi-resolution information to generate pseudo labels for better representation learning. VLC model tends to focus on all video segments related to action text to achieve text integrity, which can be used to alleviate excessive attention to important segments in the TSM model. In this paper, we utilize the label text informa-

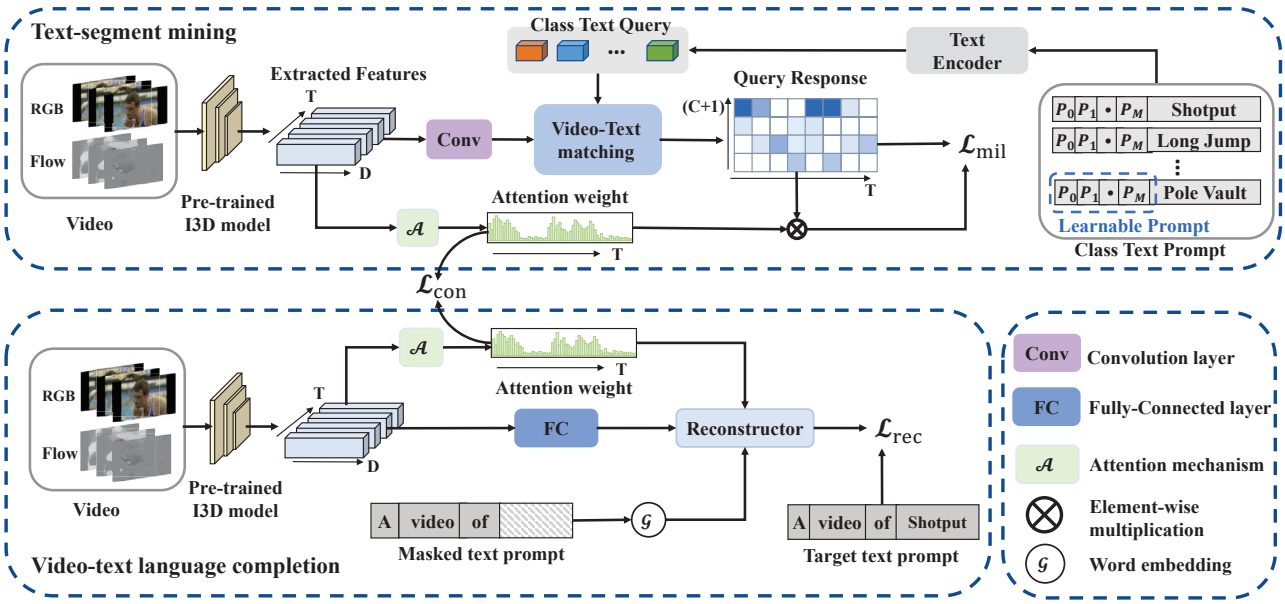

Figure 2. Illustration of the proposed framework. In this work, the text-segment mining objective uses the action label texts as a query to mine semantically related segments in the video to achieve action localization. In addition, the language completion objective aims to focus on the areas related to the action label texts in the video as comprehensively as possible to complete the masked keywords, and alleviate the localization errors caused by the excessive attention of the matching strategy to the most relevant segments in a self-supervised manner.

tion to construct a VLC model and design a self-supervised constraint between the TSM and the VLC model to achieve more complete localization results.

Vision-Language Models. Recently, a series of works on the interaction of vision and language has attracted increasing attention in the past few years, such as vision language pre-training [18, 38], video caption [44], video grounding [6, 31, 52], video question answering [24] and so on. However, how to make full use of the information encapsulated in action label texts in the WTAL task has not yet been explored. In this paper, we design a novel framework to explore leverage text information of action label to boost WTAL task. By combining discriminative objective TSM and the generative objective VLC, the proposed framework realizes the indirect use of text information to boost WTAL.

3. The Proposed Method

3.1. Overall Architecture

Problem formulation. In WTAL, we are provided with a set of N untrimmed videos defined as $\{V_j\}_{j=1}^{N}$, and all of them are annotated with their corresponding video-level action category labels $\{\mathbf{y}_j\}_{j=1}^{N}$. Generally, the label $\mathbf{y}_j$ is discretized into a binary vector indicating the presence/absence of each category of action in the video v_j. Each video V contains a set of segments: $V = \{v_t\}_{t=1}^{\mathcal{T}}$, where $\mathcal{T}$ is the number of segments in the video. Generally, $\mathcal{T}$ segments are fed into a pre-trained 3D CNN model [2] to extract both RGB features $\mathbf{X}_r \in \mathbb{R}^{\mathcal{T}\times 1024}$ and FLow video features $\mathbf{X}_f \in \mathbb{R}^{\mathcal{T}\times 1024}$. During inference, we predict a sequence of actions $\{c_i, s_i, e_i, conf_i\}$ for an input video, where c_i is the action category, s_i and e_i represent the start and end time, and $conf_i$ is confidence score.

Overview. The proposed overall framework is shown in Figure 2, which leverages text information of action labels to boost WTAL from two aspects, *i.e.*, Text-Segment Mining (TSM) and Video-text Language Completion (VLC). For the TSM in Section 3.2, the RGB and Flow video features $\mathbf{X}_r$ and $\mathbf{X}_f$ is fed into a video embedding module consisting of convolution layers to generate video features embedding at first. Second, we construct text descriptions for action labels via prompt template and generate text queries according to the description by the text encoder. Then in the video-text matching module, TSM compares the text queries with all segments of videos to generate query responses to mine semantically related video segments. In addition, we generate attention weights for each video segment to further suppress the response of the background segments to text queries. For the VLC in Section 3.3, the extracted video features $\mathbf{X}_r$ and $\mathbf{X}_f$ are fed into a fully connected layer to get video feature embedding at first. Later, we construct a description sentence for the action label of the video and mask the key action words of the sentence. Then, an attention mechanism is designed to collect semantically related segments to predict masked action words via the language reconstructor. Finally, in Section 3.4, we combine the TSM and VLC via imposing self-supervised constraints between attentions of them to obtain more accurate

and complete localization results.

3.2. Text-Segment Mining

In this section, we introduce the text-segment mining objective (TSM) to make full use of information encapsulated in action label texts. Specifically, the TSM consists of a video embedding module, a text embedding module and a video-text feature matching module.

Video embedding module. Similar to other WTAL models, the video embedding module is composed of two 1D convolutions followed by RelU and Dropout layers. We use a strategy similar to [11] to fuse RGB and Flow features to obtain video features $\mathbf{X} \in \mathbb{R}^{\mathcal{T}\times 2048}$ as the input of the video embedding module. Then, the corresponding video feature embedding $\mathbf{X}_e \in \mathbb{R}^{\mathcal{T}\times 2048}$ can be obtained by $\mathbf{X}_e = emb(\mathbf{X})$, where $emb(\cdot)$ represents the video embedding module. Besides, following previous works [11,16], an attention mechanism is utilized to generate attention weight $\mathbf{att}_m \in \mathbb{R}^{\mathcal{T}\times 1}$ for each video segment V_j,

$$\mathbf{att}_m = \sigma(\mathcal{A}(\mathbf{X})), \tag{1}$$

where $\mathcal{A}(\cdot)$ is the attention mechanism consisting of several convolution layers, and $\sigma(\cdot)$ means the sigmoid function.

Text embedding module. The text embedding module aims to use action label text to generate a series of queries for mining segments related to category text in the videos. We adopt category-specific learnable prompts for C category of action label texts, to form the input of the text embedding module L_q:

$$\mathbf{L}_q = [\mathbf{L}_s; \mathbf{L}_p; \mathbf{L}_e], \tag{2}$$

where $\mathbf{L}_s$ denotes the [START] token initialized randomly, $\mathbf{L}_p$ denotes the learnable textual contexts with the length N_p, and $\mathbf{L}_e$ denotes action label text features embed by GloVe [36]. Besides, the $C+1$-th additional background class embedding is initialized by zero.

Then a Transformer encoder $trans(\cdot)$ is used as the text embedding module to generate text queries. Specifically, the class text queries $\mathbf{X}_q$ can be obtained by $\mathbf{X}_q = trans(\mathbf{L}_q)$, where $\mathbf{X}_q \in \mathbb{R}^{(C+1)\times 2048}$.

Video-text feature matching. The video-text feature matching module is used to match semantic-related text query and video segment features.

To be specific, we conduct the inner product operation between the video embedding feature $\mathbf{X}_e$ and the text queries $\mathbf{X}_q$ to generate the segment-level video-text similarity matrix $\mathbf{S} \in \mathbb{R}^{\mathcal{T}\times(C+1)}$.

Besides, following the background suppression-based methods [11,16,22], we also apply attention weight $\mathbf{att}_m$ to suppress the response of the background segment to the action text. The background suppressed segment-level matching result $\bar{\mathbf{S}} \in \mathbb{R}^{\mathcal{T}\times C+1}$ can be obtained by $\bar{\mathbf{S}} = \mathbf{att}_m * \mathbf{S}$, where '*' means element-wise multiplication in this paper.

Finally, similar to current approaches [30, 35], We also use top-k multi-instance learning to calculate matching loss. Specifically, we calculate the average value of top-k similarity in the temporal dimension corresponding to a specific category of text query as the video-level video-text similarity.

For the j-th action category, video-level similarity $\mathbf{v}_j$ and $\bar{\mathbf{v}}_j$ are generated from $\mathbf{S}$ and $\bar{\mathbf{S}}$, respectively:

$$\mathbf{v}_j = \max_{\substack{l\subset\{1,\dots,\mathcal{T}\}\\|l|=k}} \frac{1}{k}\sum_{i\in l}\mathbf{S}_i(j), \quad \bar{\mathbf{v}}_j = \max_{\substack{l\subset\{1,\dots,\mathcal{T}\}\\|l|=k}} \frac{1}{k}\sum_{i\in l}\bar{\mathbf{S}}_i(j), \tag{3}$$

where l is a set containing the index of the top-k segments with the highest similarity to the j-th text query, and k is the number of selected segments. Then, We apply softmax to $\mathbf{v}_j$ and $\bar{\mathbf{v}}_j$ to generate video-level similarity score $\mathbf{p}_j$ and $\bar{\mathbf{p}}_j$.

We encourage the positive score of video-text category matching to approach 1, while the negative score to approach zero to train the TSM objective,

$$\mathcal{L}_{mil} = -(\sum_{j=1}^{C+1}\mathbf{y}_j\log(\mathbf{p}_j) + \sum_{j=1}^{C+1}\hat{\mathbf{y}}_j\log(\hat{\mathbf{p}}_j)), \tag{4}$$

where $\mathbf{y}_j$ and $\hat{\mathbf{y}}_j$ are labels for video-text matching. In addition, the additional $C+1$-th background class is 0 in $\hat{\mathbf{y}}_j$ and 1 in $\mathbf{y}_j$.

Besides, in this work, follow [11, 16], we also adopt co-activity loss [30, 35], normalization loss [22, 23] and guide loss [11, 16] to train the TSM model. Since they are not the main contributions of this work, we do not elaborate on them in this paper.

3.3. Video-Text Language Completion

The Video-text Language Completion (VLC) objective aims to complete the masked keywords in the video description, by focusing on the text-related video segments related as comprehensively as possible. The proposed VLC also contains a video embedding module and a text embedding module. Besides, a transformer reconstructor is used for multi-modal interaction and completion of the original text description.

Video embedding module. Given the original video feature $\mathbf{X} \in \mathbb{R}^{\mathcal{T}\times 2048}$ as described in sec.3.2, we can obtain the corresponding video feature embedding $\mathbf{X}_v \in \mathbb{R}^{\mathcal{T}\times 512}$ through for a full connection layer the VLC module.

To mine positive areas of text-semantic-related video, the proposed completion model specially designs an attention mechanism with the same structure as Sec. 3.2. The attention weight for VLC $\mathbf{att}_r \in \mathbb{R}^{\mathcal{T}\times 1}$ can be obtained by:

$$\mathbf{att}_r = \sigma(\mathcal{A}(\mathbf{X})), \tag{5}$$

where $\mathcal{A}(\cdot)$ is the attention mechanism composed of several convolution layers, and $\sigma(\cdot)$ represents the sigmoid function.

Text embedding module. The datasets of the WTAL task only provide action videos and their action labels but does not contain any sentences describing the corresponding videos. Hence, we first use the prompt template "a video of [CLS]" and the action label texts to construct a description sentence for the video. Then, we mask the key action words of the description sentence, and embed the masked sentence with GloVe [36] and a fully connected layer to get sentence feature embedding $\hat{\mathbf{X}}_s \in \mathbb{R}^{M\times 512}$, where M is the length of the sentence.

Transformer reconstructor. In the video-text language completion model, a transformer reconstructor is used to complete the masked description sentence. Firstly, following [27], we randomly mask 1/3 of the words in the sentence as the alternative description sentence, which could result in a high probability to mask the action label texts. Then, the encoder of the transformer is used to get the foreground video feature $\mathbf{F} \in \mathbb{R}^{\mathcal{T}\times 512}$ by:

$$\mathbf{F} = E(\mathbf{X}_v, \mathbf{att}_r) = \delta(\frac{\mathbf{X}_v\mathbf{W}_q(\mathbf{X}_v\mathbf{W}_k)^T}{\sqrt{D_h}} * \mathbf{att}_r)\mathbf{X}_v\mathbf{W}_v, \tag{6}$$

where $E(\cdot,\cdot)$ denotes the transformer encoder, $\delta(\cdot)$ denotes the softmax function, $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{512\times 512}$ are learnable parameters, and $D_h = 512$ is the feature dimension of $\mathbf{X}_v$.

The decoder of the transformer is used to obtain multi-modal representation $\mathbf{H} \in \mathbb{R}^{M\times 512}$ to reconstruct the masked sentence:

$$\begin{aligned}\mathbf{H} &= D(\hat{\mathbf{X}}_s, \mathbf{F}, \mathbf{att}_r) \\ &= \delta(\frac{\hat{\mathbf{X}}_s\mathbf{W}_{qd}(\mathbf{F}\mathbf{W}_{kd})^T}{\sqrt{D_h}} * \mathbf{att}_r)\mathbf{F}\mathbf{W}_{vd},\end{aligned} \tag{7}$$

where $D(\cdot,\cdot,\cdot)$ denotes the transformer decoder, and $W_{qd}, W_{kd}, W_{vd} \in \mathbb{R}^{512\times 512}$ are learnable parameters.

Finally, the probability distribution $\mathbf{P} \in \mathbb{R}^{M\times N_v}$ of the i-th word w_i on the vocabulary can be obtained by:

$$\mathbf{P}(w_i|\mathbf{X}_v, \hat{\mathbf{X}}_{s[0:i-1]}) = \delta(FC(\mathbf{H})), \tag{8}$$

where FC$(\cdot)$ denotes the fully connected layer, $\delta(\cdot)$ denotes the softmax function, and N_v is the vocabulary size.

The final VLC loss function can be formulated as:

$$\mathcal{L}_{rec} = -\sum_{i=1}^{M} \log \mathbf{P}(w_i|\mathbf{X}_v, \hat{\mathbf{X}}_{txt[0:i-1]})). \tag{9}$$

To further improve the mined positive areas of text-semantic-related video, we also adopt a contrastive loss [53] in the completion model. Specifically, positive areas mined by attention weight $\mathbf{att}_r$ should be more compatible with the sentence than the entire video, and those negative areas mined by $1 - \mathbf{att}_r$. Therefore, following Eq. 6-9, we can obtain the completion loss $\mathcal{L}^e_{rec}$ and $\mathcal{L}^n_{rec}$, where the attention weight $\mathbf{att}_r$ used in the transformer is replaced with 1 and $1 - \mathbf{att}_r$, respectively.

Finally, the contrastive loss $\mathcal{L}_c$ can be formulated as:

$$\mathcal{L}_c = \max(\mathcal{L}_{rec} - \mathcal{L}^e_{rec} + \gamma_1, 0) + \max(\mathcal{L}_{rec} - \mathcal{L}^n_{rec} + \gamma_2, 0), \tag{10}$$

where γ_1 and γ_2 are hyper-parameters.

3.4. Self-Supervised Consistency Constraint

The matching strategy used in TSM tends to focus on the video segments that better match the text, while excluding other text-unrelated segments as they could lead to localization error. On the other hand, the VLC tends to focus on all video clips that are related to action text to achieve description completion. Hence, we impose self-supervised constraints between attentions of these two objectives, *i.e.,* the discriminative objective TSM and the generative objective VLC, to alleviate the excessive attention paid to the most semantic-related segments by TSM. The consistency constraint loss L_{con} can be obtained by:

$$\mathcal{L}_{con} = MSE(\mathbf{att}_m, \psi(\mathbf{att}_r)) + MSE(\mathbf{att}_r, \psi(\mathbf{att}_m)), \tag{11}$$

where $\psi(\cdot)$ represents a function that truncates the gradient of the input, and MSE$(\cdot,\cdot)$ denotes the Mean Squared Error loss.

The consistency constraint loss can encourage $\mathbf{att}_m$ trained by TSM and $\mathbf{att}_r$ trained by VLC to focus on the same action area within the video. In this way, the localization errors which are caused by the excessive attention of the matching strategy on the most relevant segments can be alleviated. Besides, the information of the action label text can be transmitted from the video-text language completion model to the WTAL model through the attention mechanism indirectly.

3.5. Model Training and Inference

Optimizing Process. Considering all the aforementioned objectives, our final objective function of the whole framework arrives at:

$$\mathcal{L} = \mathcal{L}_{mil} + \alpha\mathcal{L}_{rec} + \beta\mathcal{L}_c + \lambda\mathcal{L}_{con}, \tag{12}$$

where α, β, λ are the hyper-parameters to balance these four loss terms.

Model Inference. In the test stage, we follow the process of [11, 16]. Firstly, we select those classes with video-level category scores above a threshold for generating proposals. Then for the selected action classes, we obtain the class-agnostic action proposals by thresholding the attention weights and selecting the continuous components of

Table 1. Experimental results of different methods in THUMOS14 dataset.

Method	mAP@IoU(%)					Avg
	0.3	0.4	0.5	0.6	0.7	0.3:0.7
BasNet(2020) [22]	44.6	36.0	27.0	18.6	10.4	25.2
RPN(2020) [12]	48.2	37.2	27.9	16.7	8.1	27.6
TSCN(2020) [49]	47.8	37.7	28.7	19.4	10.2	28.8
HamNet(2021) [16]	50.3	41.1	31.0	20.7	11.1	30.8
UGCT(2021) [47]	55.5	46.5	35.9	23.8	11.4	34.6
CO2Net(2021) [11]	54.5	45.7	38.3	26.4	13.4	35.6
FACNet(2021) [13]	52.6	44.3	33.4	22.5	12.7	33.1
FTCL(2022) [7]	55.2	45.2	35.6	23.7	12.2	33.4
ASMLoc(2022) [9]	**57.1**	46.8	36.6	25.2	13.1	34.4
DCC(2022) [25]	55.9	45.9	35.7	24.3	13.7	35.1
RSKP(2022) [15]	55.8	47.5	38.2	25.4	12.5	35.9
Ours	56.2	**47.8**	**39.3**	**27.5**	**15.2**	**37.2**

Table 2. Experimental results of different methods in ActivityNet1.3 dataset.

Method	mAP@IoU(%)			Avg
	0.5	0.75	0.95	0.5:0.95
BasNet(2020) [22]	34.5	22.5	4.9	22.2
TSCN(2020) [49]	25.3	21.4	5.3	21.7
UGCT(2021) [47]	39.1	22.4	5.8	23.8
FACNet(2021) [13]	37.6	24.2	6.0	24.0
ACMNet(2021)	40.1	24.2	6.2	24.6
FTCL(2022) [7]	40.0	24.3	**6.4**	24.8
ASMLoc(2022) [9]	41.0	24.9	6.2	25.1
DCC(2022) [25]	38.8	24.2	5.7	24.3
RSKP(2022) [15]	40.6	24.5	5.9	25.0
Ours	**41.8**	**26.0**	6.0	**26.0**

the remaining segments. The obtained i-th candidate action proposal can be denoted as $\{c_i, s_i, e_i, conf_i\}$. For the confidence score $conf_i$, we follow the AutoLoc [39] to calculate the outer-inner score of each action proposal through $\bar{\mathbf{S}}$. Finally, we remove the overlapping proposals using soft non-maximum suppression.

4. Experiments

4.1. Datasets

THUMOS14. THUMOS14 [17] dataset contains 200 validation videos and 213 test videos. There are a totally of 20 categories in the dataset, with an average of 15.5 actions per video. Following the same setting as [14, 33–35], we adopt 200 validation videos for training and 213 test videos for testing.

ActivityNet. ActivityNet [1] dataset offers a larger benchmark for temporal action localization. There are 10,024 training videos, 4,926 validation videos, and 5,044 testing videos with 200 action categories. Following the experimental setting in [13–15, 47], we adopt all the training videos to train our model and evaluate our proposed method in all the testing videos.

4.2. Implementation Details

Evaluation Metrics. We evaluate the proposed method for action localization using mean Average Precision (mAP). The prediction proposal is considered as correct if its action category is predicted correctly and overlaps significantly with the ground truth segment (based on the IoU threshold). We adopt the official evaluation code of ActivityNet to evaluate our method [1].

Feature Extractor. Following previous work [4,30,33,35], the optical flow maps are generated by using the TV-L1 algorithm [46], and we use I3D network [2] pre-trained on the Kinetics dataset [19] to extract both RGB and optical flow features without fine-turning.

Training Settings. we use Adam [20] with a learning rate of 0.0005 and weight decay of 0.001 to optimize our model for about 5,000 iterations on THUMOS14. For ActivityNet1.3, the learning rate is 0.00003 to optimize our model for about 50,000 iterations. For the hyper-parameters in L_c, we set γ_1 as 0.1 and γ_2 as 0.2. Besides, for the hyper-parameter α, β, λ, we set it as 1.0,1.0,1.5 on THUMOS14 and 1.0,1.0, 0.25 on ActivityNet1.3, respectively. Our model is implemented by PyTorch 1.8 and trained under Ubuntu 18.04 platform. Hyper-parameter sensitivity analysis can be found in the supplementary materials.

4.3. Comparison with the State-of-the-Arts

We compare the proposed method with state-of-the-art weakly-supervised methods in this section. The results are shown in Table 1 and Table 2. For THUMOS14 datasets, the proposed framework evidently outperforms current state-of-the-art WTAL approaches, especially in high IoU experimental settings. On the important criterion: average mAP (0.3:0.7), we surpass the state-of-the-art method [15] by 1.3%, even surpassing some fully-supervised methods. For the larger ActivityNet1.3 dataset, our method still obtains 1.0%mAP improvement over existing the state-of-the-art weakly-supervised methods [15] on average.

4.4. Ablation Study

Effectiveness of each component. The proposed framework mainly contains three ingredients: (1) the text-segment mining (TSM) module to replace the existing WTAL model that only uses video information; (2) Additional video-text language complements (VLC) are used to constrain WTAL models in a self-supervised manner, denoted as $\mathcal{L}_{rec} + \mathcal{L}_{con}$; (3) contrastive loss in the video-text language complements model, denoted as $\mathcal{L}_c$.

To verify the effectiveness of each component in the proposed framework, we conduct a comprehensive ablation

study to analyze different components in Table 3. Specifically, we implement four variants of the proposed method as follows: (1) "Baseline": Using a convolution layer as a classifier instead of video-text matching in TSM, and only using video information to train the WTAL model; (2) "Baseline + $\mathcal{L}_{rec} + \mathcal{L}_{con}$": Additional video text language complements (VLC) is used to constrain baseline WTAL models in a self-supervised manner; (3) "Baseline + $\mathcal{L}_{rec} + \mathcal{L}_{con} + \mathcal{L}_c$": Using contrastive loss in the video-text language complements model; (4) "TSM+ $\mathcal{L}_{rec} + \mathcal{L}_{con} + \mathcal{L}_c$" : The final framework, replacing the baseline WTAL model with the proposed TSM on the basis of (3);

Table 3. Effectiveness of each component on THUMOS14 datasets.

Method	mAP@IoU(%)			Avg
	0.3	0.5	0.7	
Baseline	54.5	36.5	13.0	34.9
Baseline + $\mathcal{L}_{rec} + \mathcal{L}_{con}$	55.0	37.8	14.0	35.9
Baseline + $\mathcal{L}_{rec} + \mathcal{L}_{con} + L_c$	55.7	38.3	13.8	36.3
TSM + $\mathcal{L}_{rec} + \mathcal{L}_{con} + L_c$	56.2	39.3	15.2	37.2

By comparing the performance of methods "TSM + $\mathcal{L}_{rec} + \mathcal{L}_{con} + \mathcal{L}_c$" and "Baseline + $\mathcal{L}_{rec} + \mathcal{L}_{con} + \mathcal{L}_c$", we can conclude that the text-segment mining is better than generally WTAL model using only convolution classifier without action label text information, which brings about 0.9% performance improvement on THUMOS14 dataset. When we ablate the measurement of learning loss $\mathcal{L}_c$ and the additional video-text language completion model $\mathcal{L}_{rec} + \mathcal{L}_{con}$ step-by-step, the performance under all the experiment settings could be gradually decreased. To be specific, by comparing the methods "Baseline + $\mathcal{L}_{rec} + \mathcal{L}_{con}$" with "Baseline", we can conclude that the proposed video-text language model can constrain the WTAL model by self-supervision and indirectly transfer text information to it, which brings about 1.0% mAP performance improvement on THUMOS14 dataset. Besides, comparing the methods "Baseline + $\mathcal{L}_{rec} + \mathcal{L}_{con} + \mathcal{L}_c$" with "Baseline + $\mathcal{L}_{rec} + \mathcal{L}_{con}$", we can also verify the effectiveness of the contrastive loss in the VLC.

Futhermore, we evaluated the frame-level classification results on THUMOS14. Compared with the baseline, after using the TSM model, the false positive rate (FPR) dropped from 26.0% to 23.8%, and after using the VLC model, the false negative rate (FNR) decreased from 28.0% to 26.9%. This shows that TSM can effectively alleviate the problem that the background segment is misclassified as a groundtruth action, thus effectively alleviating the over-complete problem while VLC can effectively alleviate the problem that the groundtruth action segment is misclassified as background, thus effectively alleviating the incomplete problem.

Comparisons with different prompts in text-segment mining model. We compare the effects of handcraft prompts "a video of [CLS]" and the learnable prompt on the text-segment mining models in Table 4. Compared with the handcraft prompt in the text-segment mining model, the learnable prompt achieves better performance. It is because, by making it learnable, textual contexts can achieve better transferability in downstream video-text matching tasks by directly optimizing the contexts using back-propagation.

Table 4. Comparisons with different prompts in classification model on THUMOS14 dataset.

Method	mAP@IoU(%)			Avg
	0.3	0.5	0.7	
handcraft prompt	55.1	38.1	14.1	36.1
learnable prompt	**56.2**	**39.3**	**15.2**	**37.2**

Comparisons with different types of consistency constraint loss. We also evaluate the effect of different types of consistency constraints. Specifically, we implement five variants of the constraints on the VLC and TSM model in different ways: (1) "w/o $\mathcal{L}_{con}$": The VLC model is not used, and only TSM is used as the baseline; (2) "Share": $\mathcal{L}_{con}$ is not used, but the VLC and TSM share the parameters of the attention module; (3) "KL": Using Kullback Leibler divergence [37] as loss function $\mathcal{L}_{con}$; (4)"MAE": Using Mean Absolute Error as loss $\mathcal{L}_{con}$; (5)"MSE": Using Mean Square Error as loss $\mathcal{L}_{con}$.

The result in Table 5 shows that using an additional video-text completion model to constrain the WTAL model can effectively improve localization performance, and using MSE as the consistency constraint loss is more suitable.

Table 5. Comparisons with types of consistency constraint loss on THUMOS14 dataset.

Method	w/o	Share	KL	MAE	MSE
Avg mAP	35.4	35.8	36.9	36.4	**37.2**

Comparisons with different types of language reconstructor in the video-text language completion model.

We compare the performance impact of using different prompt templates to generate action descriptions in the completion model.

To verify the effectiveness of additional video-text language completion model, we compare the effects of different types of language reconstructor on the localization result. Specifically, we compared three different reconstructors, Transformer, GRU [3] and LSTM [10] in Table 6. In addition, "w/o" represents only the TSM model used. As shown in Table 6, we can conclude that no matter which language reconstructor is used, the video-text language completion model could improve the performance of the proposed framework, by imposing self-supervised constraints on TSM. Besides, we can conclude that the Transformer

structure is more suitable to be used as a language reconstructor in our framework.

Table 6. Comparisons with different types of language reconstructors in the video-text language completion model on THUMOS14 dataset.

Method	w/o	GRU	LSTM	Transformer
Avg mAP	35.4	36.7	36.1	**37.2**

Comparison of action descriptions generated by different prompt templates in language completion model. We compared the performance influence of action descriptions generated by different prompt templates in the language completion model in Table 7. The results of all types of prompt templates can outperform existing state-of-the-art results, as shown in Table 1. These results indicate that it is necessary to use video-text language completion model to constrain WTAL model.

Table 7. Comparisons with different prompts in completion model on THUMOS14 dataset.

Prompt	mAP@IoU(%)			Avg
	0.3	0.5	0.7	
a [CLS]	55.9	38.7	14.6	36.8
a video of action [CLS]	55.1	38.5	15.0	36.6
a video of the [CLS]	**56.2**	**39.3**	**15.2**	**37.2**

Integrating our framework to existing methods. The proposed method can be easily extended to existing WTAL models and improve their performance. To verify the scalability of the proposed framework, we design three sets of experiments to extend the proposed framework to existing methods: (1) "+TSM": Using the proposed TSM to replace convolution classifier of existing WTAL model; (2) "+VLC": Additional VLC model are used to constrain WTAL models in a self-supervised manner; (3) "+TSM+VLC": extended all components of our framework to existing WTAL model. As shown in Table 8, we can clearly conclude that both of the proposed TSM and VLC can greatly improve the performance of two existing methods, verifying the effectiveness of leveraging action label text information to expand WTAL model.

Table 8. Integrating our framework to existing methods on THUMOS14 dataset.

Method	mAP@IoU(%)			Avg
	0.3	0.5	0.7	
BaSNet [22]	44.6	27.0	10.4	27.3
BaSNet + TSM	48.2	31.7	9.7	29.5
BaSNet + VLC	48.6	32.0	10.3	30.2
BaSNet + TSM + VLC	**49.0**	**32.5**	**10.7**	**30.6**
HAMNet [16]	50.3	31.0	11.1	30.8
HAMNet + TSM	51.8	34.7	11.8	32.7
HAMNet + VLC	51.5	36.0	12.8	33.6
HAMNet + TSM + VLC	**52.3**	**37.4**	**13.4**	**34.5**

4.5. Qualitative Analysis

We visualize some examples of the detected action instances in Figure 3. For each example, the top line represents the segment of the video, the following four lines in order are the ground truth of the action in the video, the localization results generated by the baseline model, the localization results generated by the text-segment mining, and the localization results generated by our final framework. As can be seen from this figure, introducing the text information in the category annotation into the WTAL model in both direct and indirect ways, helps to generate more accurate localization results and suppress the response of background fragments to a certain extent.

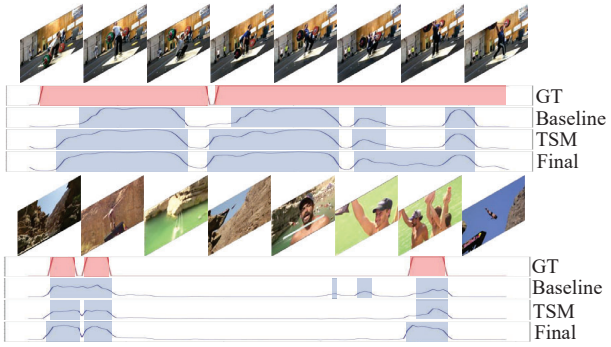

Figure 3. Two prediction examples on THUMOS14 dataset.

5. Conclusion

We introduce a new framework to leverage the text information to boost WTAL from two aspects, *i.e.* text-segment mining, and video-text language completion. With the help of text information, the proposed method can focus on the action-category-related areas in the video and improve the performance of WTAL tasks. Extensive experiments demonstrate that the proposed method achieves state-of-the-art performances on two popular datasets, and both of the proposed objectives can be directly extended to the existing WTAL methods to improve their performances.

Limitation. One major limitation in this work is that we must train the text-segment mining and video- text language completion models at the same time, resulting in the model size being twice as the original size. In the future, We will explore more efficient manners to make full use of text information in tags to boost WTAL.

Acknowledgments: This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103202; in part by the National Natural Science Foundation of China under Grants 62036007, U22A2096 and 62176198; in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15; in part by Open Research Projects of Zhejiang Lab under Grant 2021KG0AB01.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[2] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society, 2017.

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[4] Xinpeng Ding, Nannan Wang, Xinbo Gao, Jie Li, Xiaoyu Wang, and Tongliang Liu. Kfc: An efficient framework for semi-supervised temporal action localization. *IEEE Transactions on Image Processing*, 30:6869–6878, 2021.

[5] Xinpeng Ding, Nannan Wang, Jie Li, and Xinbo Gao. Cr-net: Centroid radiation network for temporal action localization. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 29–41. Springer, 2021.

[6] Xinpeng Ding, Nannan Wang, Shiwei Zhang, De Cheng, Xiaomeng Li, Ziyuan Huang, Mingqian Tang, and Xinbo Gao. Support-set based cross-supervision for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11573–11582, October 2021.

[7] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. *arXiv preprint arXiv:2203.16800*, 2022.

[8] Guoqiang Gong, Liangfeng Zheng, Wenhao Jiang, and Yadong Mu. Self-supervised video action localization with adversarial temporal transforms.

[9] Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. *arXiv preprint arXiv:2203.15187*, 2022.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. Cross-modal consensus network for weakly supervised temporal action localization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1591–1599, 2021.

[12] Linjiang Huang, Yan Huang, Wanli Ouyang, Liang Wang, et al. Relational prototypical network for weakly supervised temporal action localization. 2020.

[13] Linjiang Huang, Liang Wang, and Hongsheng Li. Foreground-action consistency network for weakly supervised temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8002–8011, 2021.

[14] Linjiang Huang, Liang Wang, and Hongsheng Li. Multi-modality self-distillation for weakly supervised temporal action localization. *IEEE Transactions on Image Processing*, 2022.

[15] Linjiang Huang, Liang Wang, and Hongsheng Li. Weakly supervised temporal action localization via representative snippet knowledge propagation. *arXiv preprint arXiv:2203.02925*, 2022.

[16] Ashraful Islam, Chengjiang Long, and Richard J. Radke. A hybrid attention mechanism for weakly-supervised temporal action localization, 2021.

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

[18] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*, 2021.

[19] Will Kay, J. Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. 05 2017.

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In *International Conference on Learning Representations*, 2021.

[22] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. *arXiv preprint arXiv:1911.09963*, 2019.

[23] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1854–1862, 2021.

[24] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019.

[25] Jingjing Li, Tianyu Yang, Wei Ji, Jue Wang, and Li Cheng. Exploring denoised cross-video contrast for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19914–19924, 2022.

[26] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[27] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11539–11546, 2020.

[28] Zhekun Luo, Devin Guillory, Baifeng Shi, Wei Ke, and Huijuan Xu. Weakly-supervised action localization with expectation-maximization multi-instance learning. 2020.

[29] Kyle Min and Jason J. Corso. Adversarial background-aware loss for weakly-supervised temporal activity localization, 2020.

[30] Kyle Min and Jason J. Corso. Adversarial background-aware loss for weakly-supervised temporal activity localization. 2020.

[31] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020.

[32] Sanath Narayan, Hisham Cholakkal, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. D2-net: Weakly-supervised action localization via discriminative embeddings and denoised activations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13608–13617, 2021.

[33] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 3c-net: Category count and center loss for weakly-supervised action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8679–8687, 2019.

[34] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6752–6761, 2018.

[35] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018.

[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[37] Daniel Polani. *Kullback-Leibler Divergence*. Encyclopedia of Systems Biology, 2013.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[39] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–171, 2018.

[40] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE international conference on computer vision (ICCV)*, pages 3544–3553. IEEE, 2017.

[41] Rui Su, Dong Xu, Luping Zhou, and Wanli Ouyang. Improving weakly supervised temporal action localization by exploiting multi-resolution information in temporal domain. *IEEE Transactions on Image Processing*, 30:6659–6672, 2021.

[42] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. *arXiv preprint arXiv:2102.01894*, 2021.

[43] George Turin. An introduction to matched filters. *IRE transactions on Information theory*, 6(3):311–329, 1960.

[44] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7622–7631, 2018.

[45] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4325–4334, 2017.

[46] Andreas Wedel, Thomas Pock, Christopher Zach, Horst Bischof, and Daniel Cremers. *An Improved Algorithm for TV-L1 Optical Flow*, pages 23–45. 07 2009.

[47] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 53–63, 2021.

[48] Runhao Zeng, Chuang Gan, Peihao Chen, Wenbing Huang, Qingyao Wu, and Mingkui Tan. Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization. *IEEE Transactions on Image Processing*, 28(12):5797–5808, 2019.

[49] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus networks for weakly-supervised temporal action localization. In *16th European Conference on Computer Vision (ECCV)*, August 2020.

[50] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4486–4496, June 2021.

[51] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. *arXiv preprint arXiv:2202.07925*, 2022.

[52] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877, 2020.

[53] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1, page 3, 2022.

[54] Jia-Xing Zhong, Nannan Li, Weijie Kong, Tao Zhang, Thomas H Li, and Ge Li. Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 35–44, 2018.