# Towards Robust Temporal Activity Localization Learning with Noisy Labels

**Daizong Liu[1], Xiaoye Qu[2], Xiang Fang[3], Jianfeng Dong[4], Pan Zhou[2†], Guoshun Nan[5], Keke Tang[6], Wanlong Fang[2] and Yu Cheng[7]**

[1]Wangxuan Institute of Computer Technology, Peking University
[2]School of Cyber Science and Engineering, Huazhong University of Science of Technology
[3]Interdisciplinary Graduate Programme, Nanyang Technological University
[4]College of Computer Science and Technology, Zhejiang Gongshang University
[5]School of Cyberspace Security, Beijing University of Posts and Telecommunications
[6]Cyberspace Institute of Advanced Technology, Guangzhou University
[7]Department of Computer Science and Engineering, The Chinese University of Hong Kong
dzliu@hust.edu.cn, xiaoye@hust.edu.cn, xfang9508@gmail.com, dongjf24@gmail.com, panzhou@hust.edu.cn,
nanguo2021@bupt.edu.cn, tangbohutbh@gmail.com, wanlongfang@gmail.com, chengyu@cse.cuhk.edu.hk

## Abstract

This paper addresses the task of temporal activity localization (TAL). Although recent works have made significant progress in TAL research, almost all of them implicitly assume that the dense frame-level correspondences in each video-query pair are correctly annotated. However, in reality, such an assumption is extremely expensive and even impossible to satisfy due to subjective labeling. To alleviate this issue, in this paper, we explore a new TAL setting termed Noisy Temporal activity localization (NTAL), where a TAL model should be robust to the mixed training data with noisy moment boundaries. Inspired by the memorization effect of neural networks, we propose a novel method called **C**o-**T**eaching **R**egularizer (CTR) for NTAL. Specifically, we first learn a Gaussian Mixture Model to divide the mixed training data into preliminary clean and noisy subsets. Subsequently, we refine the labels of the two subsets by an adaptive prediction function so that their true positive and false positive samples could be identified. To avoid single model being prone to its mistakes learned by the mixed data, we adopt a co-teaching paradigm, which utilizes two models sharing the same framework to teach each other for robust learning. A curriculum strategy is further introduced to gradually learn the moment confidence from easy to hard. Experiments on three datasets demonstrate that our CTR is significantly more robust to the noisy training data compared to the existing methods.

**Keywords:** Temporal activity localization, Noisy label

## 1. Introduction

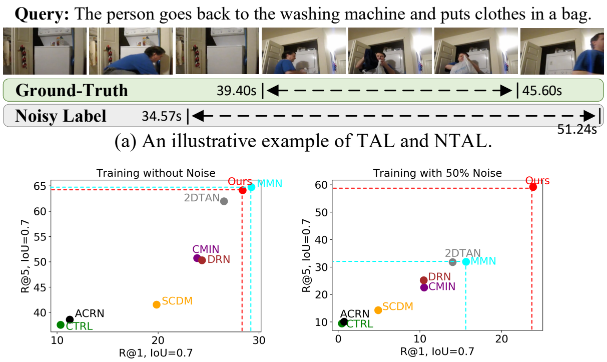
Temporal activity localization (TAL) aims at retrieving the start and end timestamps of the target moment in an untrimmed video semantically according to a sentence query. Figure 1 (a) shows an illustrative example of this task. It requires cooperation from both computer vision and natural language processing for the precisely semantic alignment, and has a wide range of applications such as video summarization (Chu et al., 2015; Jiang and Mu, 2022; Liu et al., 2023a,d,c,b, 2022a,c,e; Liu and Hu, 2022a,b; Liu et al., 2022b, 2023e, 2020b,a, 2021b,c,a, 2022d, 2024) and video question answering (Gao et al., 2019; Le et al., 2020; Fang et al., 2022a, 2021a, 2022b, 2020, 2021b, 2023b; Fang and Hu, 2020; Fang et al., 2023a, 2024).

Prior TAL works either exploit *propose-and-rank* frameworks (Anne Hendricks et al., 2017; Gao et al., 2017; Chen et al., 2018; Zhang et al., 2019b; Yuan et al., 2019a; Zhang et al., 2020b) which first generate multiple moment candidates and then utilize multimodal matching strategy to retrieve the most relevant candidate for a query, or follow the *boundary-regression* frameworks (Chen et al., 2020; Yuan et al., 2019b; Zeng et al., 2020; Zhang et al., 2020a; Nan et al., 2021; Zhang et al., 2021) to directly predict two probabilities (start/end) at each frame instead of relying on the moment candidates. Although the above two types of methods have achieved promising results, almost all of them depend on an implicit data assumption, *i.e.*, the moment boundary labels in training data are correctly annotated. However, in practical scenarios, it is extremely expensive and time-consuming to annotate or collect such dense labeled data. Actually, due to the subjectivity of different annotators, mixed web data, or the adversarial attackers, it is inevitable to collect some biased moment boundaries. As shown in Figure 1 (b), once such noisy samples are poisoned into the clean data, it will remarkably degrade both the robustness and performance of TAL methods. To the best of our knowledge, such an important noisy label problem has not been explored in TAL task yet.

In this paper, we propose a novel TAL subtask

[†]Corresponding author.

*LREC-COLING 2024*, pages 16630–16642
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0



(a) An illustrative example of TAL and NTAL.

(b) Importance of studying localization with noisy labels.

Figure 1: (a) Illustrative example of the TAL task. (b) Performance comparison on ActivityNet Caption, where the noisy samples will directly affect the performance and robustness of the existing TAL models.

termed **N**oisy Temporal activity localization (NTAL). Different from the standard TAL task, the training data in NTAL contains partially noisy samples, which corresponds to biased moment boundaries. Since NTAL problem has seldom been investigated, there are three main issues that need to be concerned about: 1) There are no extra annotations to distinguish the clean and noisy video-query samples. Therefore, it is hard to directly train a robust model in a fully-clean set. 2) Noisy samples also provide additional knowledge during the training. How to rectify their labels for assisting the model learning is worth investigating. 3) Utilizing a single model to distinguish samples and noisy labels might not be robust enough, since it may prone to specific mistakes during the training process.

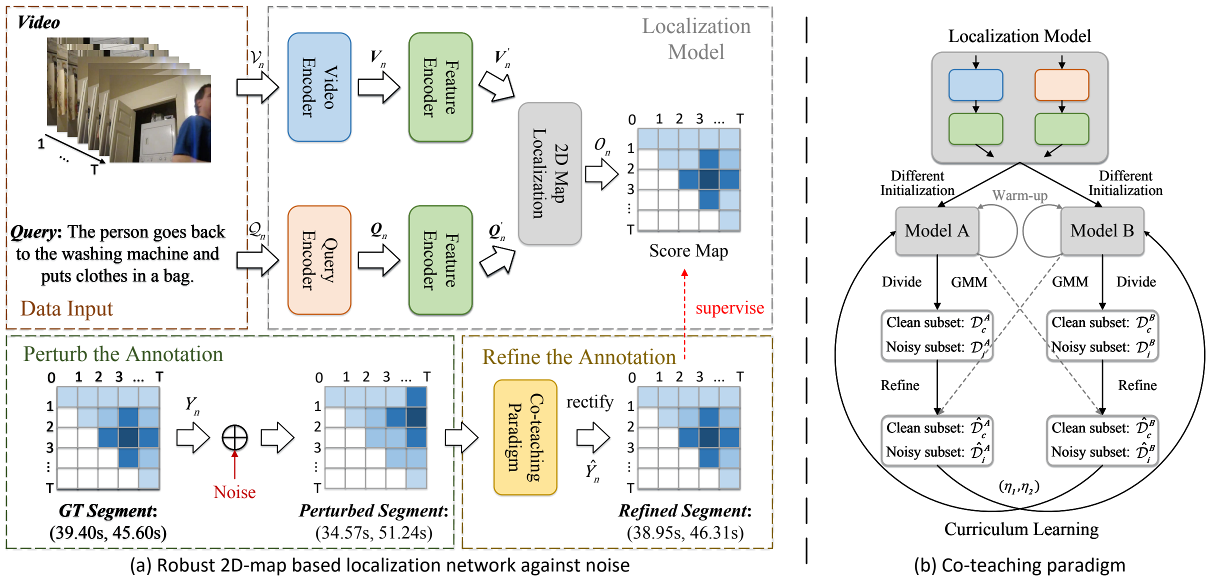
To tackle the above issues, we propose a novel framework, named Co-Teaching Regularizer (CTR). Our method is based on the memorization effect of DNNs observed in (Arpit et al., 2017; Xia et al., 2020), *i.e.*, DNNs tend to learn the simple patterns before fitting noisy samples. Specifically, we first introduce a Gaussian Mixture Model (GMM) to divide the video-query data into two data partitions, *i.e.*, clean and noisy subsets, based on their loss difference of a warm-up TAL model. Then, we develop an adaptive prediction function for label rectifying so that the false positives and the true positives could be identified from the clean and the noisy subsets, respectively. To avoid single model being prone to its mistakes learned by the mixed data, we adopt a co-teaching paradigm to utilize two models sharing the same framework to teach each other for more robust learning. In addition, we further employ a dynamic curriculum strategy to learn the rectified samples from easy-to-hard to ease the model optimization for better predicting the moment confidence. Experimental results show that our CTR is more robust to the noisy training data compared to the existing methods.

Our contributions are summarized as follows:

- To achieve robust temporal activity localization learning, we propose a more practical noisy TAL setting and make the first attempt to investigate it.
- We propose a novel Co-Teaching Regularizer (CTR) model to address corresponding issues in the NTAL task. Specifically, we develop a co-teaching paradigm to collaboratively divide the clean/noisy subsets and introduce a curriculum learning strategy to gradually learn the rectified samples from easy to hard.
- Experiments on three datasets demonstrate that our method significantly surpasses existing methods in the noisy setting. It is worth noticing that our method still achieves competitive performance on the fully-clean dataset.

## 2. Related Work

**Temporal activity localization.** Temporal activity localization (TAL) is a new task introduced recently (Gao et al., 2017; Anne Hendricks et al., 2017). Most previous algorithms (Anne Hendricks et al., 2017; Gao et al., 2017; Chen et al., 2018; Zhang et al., 2019b; Yuan et al., 2019a; Zhang et al., 2020b; Liu et al., 2021b) have been proposed within the *propose-and-rank* framework, which first generates moment candidates and then utilizes multimodal matching to retrieve the most relevant candidate for a query. Some of them (Anne Hendricks et al., 2017; Gao et al., 2017) take multiple sliding windows as candidates. To improve the quality of the candidates, (Zhang et al., 2019b; Yuan et al., 2019a) pre-cut the video on each frame by multiple pre-defined temporal scales, and directly integrate sentence information with fine-grained video clip for scoring. For instance, Xu *et al.* (Xu et al., 2019) introduce a multi-level model to integrate visual and textual features earlier and further re-generate queries as an auxiliary task. Although these methods achieve great performance, they are severely limited by the heavy computation on proposal matching/ranking, and sensitive to the quality of pre-defined proposals. Recently, many methods (Chen et al., 2020; Yuan et al., 2019b; Zeng et al., 2020; Zhang et al., 2020a; Nan et al., 2021; Zhang et al., 2021) propose to utilize the *boundary-regression* framework. Specifically, they directly predict two probabilities at each frame by leveraging cross-modal interactions between video and query, which indicate whether this frame is a start/end frame of the ground truth video moment.



(a) Robust 2D-map based localization network against noise

(b) Co-teaching paradigm

Figure 2: Overview of our proposed CTR framework. (a) In our NTAL setting, the clean annotation is perturbed by the random noise. To achieve robust training, we first utilize a co-teaching paradigm to rectify the noisy labels, and then introduce a curriculum learning strategy to iteratively supervise the model prediction. (b) The co-teaching paradigm first warms up two individual networks and employs the GMM model on them to divide the training data into clean and noisy subsets. After that, both two models will co-rectify the labels of each other for training.

Although the above methods have achieved promising results in recent years, all of them heavily rely on the correctly aligned multi-modal datasets. Therefore, it is highly expected to develop method which is robust against potential noisy correspondence, which has not been studied as far as we know. To this end, we make the first attempt to reveal the noisy label problem in TAL task and propose to eliminate the negative impact from the noisy samples.

**Learning with noisy labels.** Most existing methods for training DNNs with noisy labels seek to correct the loss function in the single-modal classification task (Song et al., 2022; Liu and Tao, 2015), which can be categorized in two types. The first type treats all samples equally and correct loss either explicitly or implicitly through relabeling the noisy samples. For relabeling methods, the noisy samples are modeled with directed graphical models (Xiao et al., 2015), Conditional Random Fields (Vahdat, 2017) or knowledge graph (Li et al., 2017). However, they require access to a small set of clean samples. Recently, (Tanaka et al., 2018; Yi and Wu, 2019) propose iterative methods which relabel samples using network predictions. The second type of correction focuses on reweighting training samples or separating clean and noisy samples, which results in correcting the loss function (Thulasidasan et al., 2019; Konstantinov and Lampert, 2019). A common method is to consider samples with smaller loss as clean ones (Shen and Sanghavi, 2019). Jiang *et al.* (Jiang et al., 2018) train a mentor network to guide a student network by assigning weights to samples. Arazo *et al.* (Arazo et al., 2019) calculate sample weights by modeling per-sample loss with a mixture model. Unlike the above noisy label studies, this paper focuses on a more challenging noisy label problem which considers mismatched multi-modal data pairs.

Note that, it is impossible to directly adopt previous noise label learning methods to solve our multi-modal noisy correspondence problem due to the following two reasons: First, most of the noisy label learning methods propose to use the model's prediction for label rectifying in the scenario of classification, while it is intractable to directly predict the aligned moment boundaries of given multi-modal pairs in TAL models. Second, even if we can rectify the noisy moment boundaries, the refined moment proposals are imbalanced and still hard to learn. To this end, we propose a novel co-teaching paradigm with curriculum learning strategy to address above issues.

# 3. Proposed Method

## 3.1. Overview

**Problem formulation.** Given the multi-modal training data $\mathcal{D} = \{\mathcal{V}_n, \mathcal{Q}_n, \mathcal{Y}_n\}_{n=1}^{N}$, each untrimmed video $\mathcal{V}_n$ is represented as $\mathcal{V}_n = \{v_{n,t}\}_{t=1}^{T}$ clip-by-clip, where $v_{n,t}$ is the $t$-th clip of $n$-th video and $T$ is the number of total clips. Similarly, the sentence query $\mathcal{Q}_n$ with $M$ words is denoted as $\mathcal{Q}_n = \{q_{n,m}\}_{m=1}^{M}$ word-by-word. For each video-query pair $(\mathcal{V}_n, \mathcal{Q}_n)$, TAL aims to retrieve a specific moment $\mathcal{Y}_n = (\gamma_n^s, \gamma_n^e)$ starting at timestamp $\gamma_n^s$ and ending at timestamp $\gamma_n^e$ in video $\mathcal{V}_n$, which corresponds to the same semantic as query $\mathcal{Q}_n$. To simulate the practical labeling process of the NTAL setting, for noisy labeling samples, the target moment boundaries are randomly perturbed, *i.e.*, $(\gamma_n^s, \gamma_n^e)$ is practically disturbed by adding random offsets $(\delta_n^s, \delta_n^e)$ as $(\gamma_n^s + \delta_n^s, \gamma_n^e + \delta_n^e)$.

**Overall framework.** Our proposed CTR framework consists of four major steps: *Step 1: Model initialization.* Given the mixed training data, we first initialize two individual networks $A, B$ sharing the same architecture shown in Figure 2 (a) and warm them up. *Step 2: Clean/noisy subset identification.* Then, we utilize the GMM model to divide the clean and noisy subsets based on the computed losses of each network. *Step 3: Label refinement.* After that, we co-rectify the labels of two networks to recall the possible true positives from noisy subset and eliminate the negative impact of the possible false positives from clean subset. *Step 4: Co-updating with curriculum strategy.* At last, we introduce an adaptive curriculum strategy to re-train the two networks in a swapping way from easy to hard till convergence to get the optimum.

### 3.2. Preliminary

**Feature extraction.** Given an untrimmed video $\mathcal{V}_n$ and a sentence query $\mathcal{Q}_n$, we first encode them into feature vectors. To be specific, the video $\mathcal{V}_n$ is encoded with a pre-trained 3D convolutional network (Tran et al., 2015; Carreira and Zisserman, 2017), and represented as $\boldsymbol{V}_n = \{\boldsymbol{v}_{n,t}\}_{t=1}^{T} \in \mathbb{R}^{T \times d_v}$, where $\boldsymbol{v}_{n,t}$ denotes the $t$-th clip feature of $n$-th video and $d_v$ refers to its feature dimension. For the query $\mathcal{Q}_i$, each word is embedded using GloVe (Pennington et al., 2014) and represented as $\boldsymbol{Q}_n = \{\boldsymbol{q}_{n,m}\}_{m=1}^{M} \in \mathbb{R}^{T \times d_m}$, where $d_m$ is the word feature dimension. After that, we utilize two linear layers to project both video features $\boldsymbol{V}_n$ and query features $\boldsymbol{Q}_n$ to the same dimension $d$. As position encoding offers a flexible way to embed a sequence when the sequence order matters, we first incorporate a position embedding to every input of both video and query sequences. Then we refer to previous works (Yuan et al., 2019b; Zhang et al., 2020a, 2021) and use four convolutions layers, a multi-head attention layer, and a feed-forward layer to generate contextualized representations $\boldsymbol{V}'_n = \{\boldsymbol{v}'_{n,t}\}_{t=1}^{T} \in \mathbb{R}^{T \times d}$ and $\boldsymbol{Q}'_n = \{\boldsymbol{q}'_{n,m}\}_{m=1}^{M} \in \mathbb{R}^{T \times d}$.

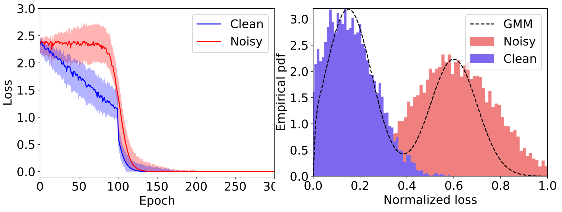


Figure 3: *Left:* Directly training our TAL model to fit the noisy labels, where the losses of clean and noisy samples are different at the beginning. *Right:* The probability density function (PDF) on the clean and noisy samples. We can find that GMM fits their distribution well and can be utilized to distinguish them.

**2D-Map based localization.** To localize the target video moment, we follow previous *propose-and-rank* framework to first define multiple moment proposals within the video and then choose the best matched one as the final prediction. Specifically, we enumerate all the possible consecutive video clips as proposals by constructing a 2D proposal feature map $\boldsymbol{F}_n = \{\boldsymbol{f}_{n,a,b}\}_{a=1,b=1}^{a=T,b=T} \in \mathbb{R}^{T \times T \times d}$ like (Zhang et al., 2020b), where each moment proposal $\boldsymbol{f}_{n,a,b}$ of $n$-th video is obtained by maxpooling its contained clips as $\boldsymbol{f}_{n,a,b} = maxpool(\boldsymbol{v}'_{n,a}, \boldsymbol{v}'_{n,a+1}, ..., \boldsymbol{v}'_{n,b})$, and $a, b$ represent the indexes of start and end video clips of the proposal. To interact each query and proposal pair, we first apply a LSTM layer on $\boldsymbol{Q}'_n$ to generate sentence-level feature and then employ a low-rank bilinear function (Kim et al., 2016) for cross-modal fusion as:

$$\boldsymbol{f}'_{n,a,b} = \boldsymbol{W}_1(\boldsymbol{W}_2 LSTM(\boldsymbol{Q}'_n) \odot \boldsymbol{W}_3 \boldsymbol{f}_{n,a,b}), \tag{1}$$

where $\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3 \in \mathbb{R}^{d \times d}$ are learnable embedding matrices and $\odot$ is the Hadamard product operator. After that, we utilize the temporal adjacent network (Zhang et al., 2020b) over the fused 2D feature map $\boldsymbol{F}'_n = \{\boldsymbol{f}'_{n,a,b}\}_{a=1,b=1}^{a=T,b=T} \in \mathbb{R}^{T \times T \times d}$ with a scaled function to generate the 2D score map $O_n = \{o_{n,a,b}\}_{a=1,b=1}^{a=T,b=T} \in \mathbb{R}^{T \times T}$. During the training, we employ a binary cross entropy loss to learn the model as:

$$\mathcal{L}_n = \sum_{a=1,b=1}^{a=T,b=T} y_{n,a,b} log(o_{n,a,b}) + (1 - y_{n,a,b}) log(1 - o_{n,a,b}), \tag{2}$$

$$\mathcal{L} = \frac{1}{N \times T \times T} \sum_{n=1}^{N} \mathcal{L}_n, \tag{3}$$

where $y_{n,a,b}$ is the ground-truth score.

### 3.3. Co-Teaching Learning Paradigm

**Preliminary dividing data by GMM.** When injecting the noisy labels into the clean dataset, some early empirical studies (Arpit et al., 2017; Xia et al., 2020) show that DNNs tend to first learn simple samples and then gradually fit the noisy samples as the noisy labels are somewhat hard samples. This so-called memorization effect of DNNs will lead to a relatively low loss for the clean samples and a higher loss for the noisy samples. To investigate the training procedure of NTAL models, we plot the loss results of clean and noisy samples in Figure 3 (*left*). It shows that noisy labels take longer to learn than clean labels. Therefore, one can infer from the loss value that a sample is more likely to be clean or noisy.

To avoid fitting noisy labels, we first divide the mixed data into two preliminary accurate data partitions, *i.e.*, "clean" and "noisy" subsets, based on the loss difference. Specifically, we follow the mixture models like (Ma and Leijon, 2011; Han et al., 2018; Yu et al., 2019) to utilize the difference of loss distribution between the clean and noisy samples to divide the training data. As shown in Figure 3 (*right*), the loss distribution of our TAL model can be well approximated by a Gaussian Mixture Model (GMM) (Permuter et al., 2006) due to its flexibility in the sharpness of distribution. Therefore, we fit the per-sample loss $\mathcal{L}_n$ in Eq.(2) of all training data by using a two-component (clean-noisy) GMM, and compute corresponding probability density function (pdf) of K components (K=2) as:

$$p(\mathcal{L}_n) = \sum_{k=1}^{K} \alpha_k p(\mathcal{L}_n|k), \tag{4}$$

where $\alpha_k$ and $p(\mathcal{L}_n|k)$ are the mixture coefficient and the probability density of the k-th component, respectively. Based on the memorization effect of DNNs, we treat the component with a smaller mean value (*i.e.*, smaller loss) as the clean set, and the other as the noisy set. We utilize the Expectation-Maximization algorithm (Moon, 1996) to optimize the GMM model, and compute the posterior probability as the clean probability $w_n$ of each sample as:

$$w_n = p(k|\mathcal{L}_n) = \frac{p(k)p(\mathcal{L}_n|k)}{p(\mathcal{L}_n)}, \tag{5}$$

where $k$ is the Gaussian component with the smaller mean. In this way, we can divide the training data into a clean subset and a noisy subset by setting a threshold $\tau_1$ on $\{w_n\}_{n=1}^{N}$.

Directly training a single model using the data divided by itself could lead to confirmation bias (*i.e.*, the model is prone to confirm its mistakes (Tarvainen and Valpola, 2017)), as noisy samples that are wrongly grouped into the clean subset would keep having lower loss due to the model overfitting to their labels. Therefore, we adopt a co-teaching paradigm to avoid such error accumulation. Being diverged offers the two networks distinct abilities to filter different types of error, making the model more robust to noise. Specifically, we individually train two networks $A$ and $B$ of the same architecture as in Sec.3.2 with different initializations and batch sequences. Following the observation in Figure 3 (a), these two networks are first trained on all training data to achieve initial convergence by Eq.(2)(3). Then, at each latter epoch, network $A$ or $B$ will model its per-sample loss distribution with a GMM and divide the dataset into clean and noisy subsets which are used for training each other.

**Refine the noisy labels.** For either of model $A$ and $B$, the mixed data $\mathcal{D}$ will be divided into the clean subset $\mathcal{D}_c^k = \{\mathcal{V}_n^{c,k}, \mathcal{Q}_n^{c,k}, \mathcal{Y}_n^{c,k}\}_{n=1}^{N_c}$ and the noisy subset $\mathcal{D}_i^k = \{\mathcal{V}_n^{i,k}, \mathcal{Q}_n^{i,k}, \mathcal{Y}_n^{i,k}\}_{n=1}^{N_i}$, where $k \in \{A, B\}$.

As for the clean subset $\mathcal{D}_c^k$, we tend to refine its labels to eliminate the negative impact of the possible false positives. Specifically, for each sample $\mathcal{V}_n^{c,k}, \mathcal{Q}_n^{c,k}, \mathcal{Y}_n^{c,k}$, we first transform the label $\mathcal{Y}_n^{c,k}$ into a 2D score map $Y_n^{c,k} = \{y_{n,a,b}^{c,k}\}_{a=1,b=1}^{a=T,b=T} \in \mathbb{R}^{T\times T}$, and then refine each sub-label $y_{n,a,b}$ by:

$$\widetilde{y}_{n,a,b}^{c,k} = w_n^k y_{n,a,b}^{c,k} + (1-w_n^k) o_{n,a,b}^{c,k}, \tag{6}$$

where $o_{n,a,b}^{c,k}$ is the network prediction. Considering that only few moment proposals in 2D map are most closed to the $\mathcal{Y}_n^{c,k}$, we further apply a sharpening function on the refined labels within each video to reduce their smoothness as:

$$\widehat{y}_{n,a,b}^{c,k} = \frac{(\widetilde{y}_{n,a,b}^{c,k})^{\frac{1}{\lambda}}}{\sum_{a=1,b=1}^{a=T,b=T}(\widetilde{y}_{n,a,b}^{c,k})^{\frac{1}{\lambda}}}, \tag{7}$$

where hyperparameter $\lambda$ is set to 0.5.

As for the noisy subset $\mathcal{D}_i^k$, we tend to refine its labels to recall the possible true positives. In particular, we discard the original labels (noise) and rectify the labels by averaging the predictions of both two models $A$ and $B$ as:

$$\widehat{y}_{n,a,b}^{i,k} = \frac{o_{n,a,b}^{i,A} + o_{n,a,b}^{i,B}}{2}. \tag{8}$$

Here, we do not apply the sharpen function since there is no true label for the noisy data.

**Co-updating the two models.** After obtaining the refined labels $\widehat{Y}_n^A, \widehat{Y}_n^B$ based on two models, we feed corresponding updated datasets $\widehat{\mathcal{D}}^A, \widehat{\mathcal{D}}^B$ to train the network $B$ and $A$ in a swapping way. Such co-updating process makes two models more robust to noise by teaching each other implicitly.

## 3.4. Curriculum Learning with Refined Labels

However, directly training each model ($A$ or $B$) with the refined labels of each video-query pair via Eq.(2) may suffer from two challenging issues:

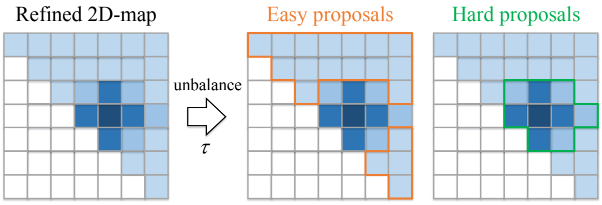


Figure 4: Illustration of the unbalanced easy and hard moment proposals.

1) The confidence scores of some query-related moment proposals are rectified from small to high. It is hard to directly fit its distribution by a single learning step. 2) The numbers of query-relevant and -irrelevant moments are imbalance. Optimizing the average scores of all moments may fail to capture and learn the sharp rectified moment scores. To this end, we propose an adaptive curriculum learning strategy like (Bengio et al., 2009) by considering two aspects: 1) For 2D score map, most moment proposals are query-irrelevant with lower confidence scores, while a few proposals are query-related but are ambiguous to determine which is the best (especially the adjacent proposals). Therefore, the irrelevant proposals can serve as the easy instances and the relevant proposals can serve as the hard instances for gradually network learning. 2) The imbalance problem of query-irrelevant and -relevant proposals can be addressed by applying balanced weights. Specifically, we utilize a threshold $\tau_2$ to distinguish both query-irrelevant and -relevant moment proposals, and then apply corresponding balanced weights with easy-hard controllers $\eta_1, \eta_2$ to loss function as:

$$\mathcal{L}'_n = \sum_{a=1,b=1}^{a=T,b=T} \eta_1 \frac{N_{total}}{N_{O_n \geq \tau_2}} \widehat{y}_{n,a,b} log(o_{n,a,b}) + \eta_2 \frac{N_{total}}{N_{O_n < \tau_2}} (1 - \widehat{y}_{n,a,b}) log(1 - o_{n,a,b}), \tag{9}$$

where $N_{total}$ is the total moment number, $\eta_1, \eta_2$ are dynamically changed during the iterative learning steps via $\eta_1 = e^{(\widehat{y}_{n,a,b} log(o_{n,a,b}))/(1+0.2*step)}$ and $\eta_2 = e^{((1-\widehat{y}_{n,a,b}) log(1-o_{n,a,b}))/(1+0.2*step)}$. Specifically, lower evaluation loss in $\eta_1, \eta_2$ indicates a less satisfactory learning status (hard examples) of the current proposal, and will lead to a lower weight to be learned. In reverse, the easy samples will obtain higher weights at the beginning and be gradually weakened by the decay schedule, while the hard samples will gradually get increasing weights during the co-updating. The overall co-teaching strategy is illustrated in Algorithm 1.

**Algorithm 1** Co-teaching Strategy

```
Input: A mixed training data $\mathcal{D}$, two localization models $A$ and $B$ sharing the same 2D-Map based framework
Warm-up two models $A$ and $B$ using Eq.(2)
for epoch=1:num_epoch do:
    $(\mathcal{D}_c^A, \mathcal{D}_i^A) \leftarrow$ divide data by $GMM(\mathcal{D}, A)$
    $(\mathcal{D}_c^B, \mathcal{D}_i^B) \leftarrow$ divide data by $GMM(\mathcal{D}, B)$
    for $k = \{A, B\}$ do:
        $(\widehat{\mathcal{D}}_c^k, \widehat{\mathcal{D}}_i^k) \leftarrow$ refine labels by Eq.(6)(7)(8)
        for $(\eta_1, \eta_2)$ in 1:step do:
            train the other model on $(\widehat{\mathcal{D}}_c^k, \widehat{\mathcal{D}}_i^k)$ with easy-hard controller $(\eta_1, \eta_2)$ by Eq.(9)
Output: Well-trained models $A$ and $B$
```

# 4. Experiments

## 4.1. Datasets

**ActivityNet Caption.** Activity Caption (Krishna et al., 2017) contains 20000 videos with 100000 descriptions from YouTube (Caba Heilbron et al., 2015). Since the test split is withheld for competition, following public split (Gao et al., 2017), we use 37421, 17505, and 17031 sentence-video pairs for training, validation, and testing respectively.

**TACoS.** TACoS is collected by (Regneri et al., 2013) for video grounding and dense video captioning tasks. For fair comparisons, we follow the same split of the dataset as (Gao et al., 2017), which has 10146, 4589, and 4083 video-query pairs for training, validation, and testing respectively.

**Charades-STA.** It is built upon the Charades (Sigurdsson et al., 2016) dataset. Following previous work (Gao et al., 2017), we utilize 12408 video-query pairs for training and 3720 pairs for testing.

## 4.2. Implementation Details

We pre-extract vision feature using the C3D (Tran et al., 2015) model for ActivityNet Caption and TACoS, and I3D (Carreira and Zisserman, 2017) model for Charades-STA. We use GloVe (Pennington et al., 2014) to extract word embeddings for each word. We set the kernel size of the convolution layer of the encoder to 7 and the head size of multi-head attention to 8. The joint video-query embedding dimension $d$ is set to 512. The random offsets of each perturbed sample are constrained to be smaller than the segment length. To divide the data, we set the threshold $\tau_1$ to 0.5. To distinguish the query-relevant and query-irrelevant moment proposals, we set the threshold $\tau_2$ to 0.55. The iterative step of curriculum learning is set to 10. We warm-up two localization models for 50 epochs. We train the whole network using the Adam optimizer with learning rate set to 0.0004. At the inference

| Noise Ratio | Method | ActivityNet Caption | | | | TACoS | | | | Charades-STA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1, IoU=0.5 | R@1, IoU=0.7 | R@5, IoU=0.5 | R@5, IoU=0.7 | R@1, IoU=0.3 | R@1, IoU=0.5 | R@5, IoU=0.3 | R@5, IoU=0.5 | R@1, IoU=0.5 | R@1, IoU=0.7 | R@5, IoU=0.5 | R@5, IoU=0.7 |
| 0% | SCDM (Yuan et al., 2019a) | 36.75 | 19.86 | 64.99 | 41.53 | 26.11 | 21.17 | 40.16 | 32.18 | **54.44** | 33.43 | 74.43 | 58.08 |
| | VSLNet (Zhang et al., 2020a) | 43.22 | 26.16 | - | - | 29.61 | 24.27 | - | - | 54.19 | **35.22** | - | - |
| | CMIN (Zhang et al., 2019b) | 43.40 | 23.88 | 67.95 | 50.73 | 24.64 | 18.05 | 38.46 | 27.02 | - | - | - | - |
| | 2DTAN (Zhang et al., 2020b) | 44.51 | 26.54 | 77.13 | 61.96 | 37.29 | 25.32 | 57.81 | 45.04 | 39.81 | 23.25 | 79.33 | 51.15 |
| | DRN (Zeng et al., 2020) | 45.45 | 24.36 | 77.97 | 50.30 | - | 23.17 | - | 33.36 | 53.09 | 31.75 | 89.06 | 60.05 |
| | MMN (Wang et al., 2022) | **48.59** | **29.26** | 79.50 | **64.76** | 39.24 | 26.17 | **62.03** | **47.39** | 47.31 | 27.28 | 83.74 | 58.41 |
| | **CTR** | 46.74 | 28.39 | **79.62** | 64.15 | **39.97** | **27.86** | 60.73 | 47.28 | 45.04 | 27.91 | **89.50** | **58.77** |
| 20% | SCDM (Yuan et al., 2019a) | 23.95 | 11.09 | 52.51 | 32.04 | 16.47 | 13.05 | 29.73 | 25.34 | 44.12 | 26.23 | 71.92 | 47.57 |
| | VSLNet (Zhang et al., 2020a) | 31.17 | 17.72 | - | - | 18.94 | 14.83 | - | - | 43.84 | 26.66 | - | - |
| | CMIN (Zhang et al., 2019b) | 33.56 | 16.35 | 56.48 | 40.39 | 15.33 | 10.26 | 28.19 | 18.65 | - | - | - | - |
| | 2DTAN (Zhang et al., 2020b) | 35.24 | 19.07 | 66.94 | 53.21 | 26.06 | 18.48 | 45.96 | 36.81 | 31.16 | 19.32 | 69.01 | 40.43 |
| | DRN (Zeng et al., 2020) | 33.31 | 14.49 | 64.37 | 40.86 | - | 17.33 | - | 25.98 | 42.58 | 23.74 | 75.76 | 46.28 |
| | MMN (Wang et al., 2022) | 36.83 | 21.44 | 64.75 | 52.72 | 28.80 | 18.62 | 49.53 | 37.15 | 36.39 | 21.05 | 71.11 | 45.64 |
| | **CTR** | **45.10** | **26.57** | **78.29** | **62.45** | **38.64** | **26.39** | **59.38** | **45.72** | **44.60** | **27.03** | **88.71** | **56.95** |
| 50% | SCDM (Yuan et al., 2019a) | 12.27 | 4.90 | 22.31 | 14.28 | 12.04 | 9.88 | 16.19 | 13.56 | 29.25 | 12.57 | 30.73 | 20.62 |
| | VSLNet (Zhang et al., 2020a) | 19.14 | 10.38 | - | - | 12.27 | 10.52 | - | - | 28.64 | 13.16 | - | - |
| | CMIN (Zhang et al., 2019b) | 21.85 | 10.52 | 26.76 | 22.44 | 12.59 | 8.71 | 15.45 | 9.30 | - | - | - | - |
| | 2DTAN (Zhang et al., 2020b) | 24.36 | 14.01 | 38.26 | 31.80 | 23.92 | 14.35 | 30.41 | 23.28 | 16.26 | 8.94 | 27.85 | 14.39 |
| | DRN (Zeng et al., 2020) | 22.03 | 10.47 | 35.72 | 25.19 | - | 12.67 | - | 15.88 | 22.47 | 11.51 | 30.73 | 18.99 |
| | MMN (Wang et al., 2022) | 25.58 | 15.65 | 36.94 | 31.93 | 26.06 | 15.11 | 35.74 | 24.52 | 18.72 | 10.09 | 29.38 | 18.60 |
| | **CTR** | **40.92** | **23.86** | **74.37** | **59.17** | **34.29** | **22.93** | **55.44** | **41.96** | **41.18** | **23.51** | **84.64** | **53.27** |

Table 1: Performance comparison on ActivityNet Caption, TACoS, and Charades-STA datasets.

| Model | Co-Teaching Paradigm | | | Curriculum Learning | | R@1, IoU=0.5 | R@1, IoU=0.7 | R@5, IoU=0.5 | R@5, IoU=0.7 |
|---|---|---|---|---|---|---|---|---|---|
| | Divide data | Refine label | Warm-up | Balanced weights | Controllers | | | | |
| Backbone | × | × | × | × | × | 23.82 | 15.25 | 39.14 | 32.37 |
| ① | ✓ | × | ✓ | × | × | 31.47 | 18.71 | 58.93 | 46.59 |
| ② | × | ✓ | ✓ | × | × | 25.63 | 15.19 | 42.38 | 33.43 |
| ③ | ✓ | ✓ | × | × | × | 3.94 | 0.85 | 16.46 | 11.71 |
| ④ | ✓ | ✓ | ✓ | × | × | 35.88 | 21.02 | 65.60 | 51.92 |
| ⑤ | ✓ | ✓ | ✓ | ✓ | × | 38.96 | 22.74 | 71.23 | 56.25 |
| ⑥ | ✓ | ✓ | ✓ | ✓ | ✓ | **40.92** | **23.86** | **74.37** | **59.17** |

Table 2: Main ablation study on the ActivityNet Caption dataset with 50% noise ratio.

| Noise Level | Method | R@1, IoU=0.5 | R@1, IoU=0.7 | R@5, IoU=0.5 | R@5, IoU=0.7 |
|---|---|---|---|---|---|
| 0.0 | 2DTAN (Zhang et al., 2020b) | 44.51 | 26.54 | 77.13 | 61.96 |
| | DRN (Zeng et al., 2020) | 45.45 | 24.36 | 77.97 | 50.30 |
| | MMN (Wang et al., 2022) | **48.59** | **29.26** | 79.50 | **64.76** |
| | **CTR** | 46.74 | 28.39 | **79.62** | 64.15 |
| 0.2 | 2DTAN (Zhang et al., 2020b) | 32.88 | 17.31 | 63.46 | 51.32 |
| | DRN (Zeng et al., 2020) | 31.06 | 12.38 | 63.19 | 38.60 |
| | MMN (Wang et al., 2022) | 34.35 | 19.84 | 61.57 | 50.51 |
| | **CTR** | **43.69** | **25.80** | **76.83** | **61.07** |
| 0.5 | 2DTAN (Zhang et al., 2020b) | 19.21 | 10.15 | 32.98 | 27.35 |
| | DRN (Zeng et al., 2020) | 16.85 | 7.46 | 30.04 | 20.73 |
| | MMN (Wang et al., 2022) | 20.17 | 11.39 | 31.82 | 27.06 |
| | **CTR** | **38.33** | **21.81** | **71.20** | **56.73** |

Table 3: Performance comparison on the ActivityNet dataset with different noise level.

stage, we average the similarities predicted by two networks for the localization evaluation.

## 4.3. Comparisons to the State-of-The-Arts

As shown in Table 1, we report the results with three different noise ratios (denoting the percentages of video-query pairs containing noise), *i.e.*, 0%, 20%, and 50%. When the noise rate is 0%, we directly refer to the results reported in the corresponding papers. For the noisy cases, we re-train the compared models with our noisy setting. From this table, we can find that our method is very competitive in the 0% noise. Moreover, when increasing the noise ratio in the training data, all compared methods are very sensitive to the noisy labels and degenerate the performance a lot. Instead, our CTR is more robust to the noisy labels and remarkably outperforms all the baselines by a large margin on all three datasets under different noise settings.

We also report the results with three different noise level (denoting the amount of noise in each video-query pair), *i.e.*, 0.0, 0.2, and 0.5. As shown in Table 3, the experiments are conducted on ActivityNet, where we denote the noise level on each sample as the $1 - IoU(gt, noise\_label)$ ( *i.e.*, the larger the noise level, the lower IoU score with GT). It shows that our CTR is much more robust.

## 4.4. Ablation Study

We perform ablation studies of our CTR on the ActivityNet Caption dataset with 50% noise.

**Main ablation.** To investigate the effectiveness of both co-teaching and curriculum learning modules

| Module | Change | R@1, IoU=0.5 | R@1, IoU=0.7 | R@5, IoU=0.5 | R@5, IoU=0.7 |
|---|---|---|---|---|---|
| Divide data | w/. GMM | **40.92** | **23.86** | **74.37** | **59.17** |
| | w/. BMM | 35.37 | 20.65 | 68.24 | 55.91 |
| | $\tau_1 = 0.4$ | 37.75 | 21.44 | 70.18 | 56.83 |
| | $\tau_1 = 0.5$ | **40.92** | **23.86** | **74.37** | **59.17** |
| | $\tau_1 = 0.6$ | 39.03 | 22.61 | 71.94 | 57.50 |
| Refine clean subset | w/. sharpen | **40.92** | **23.86** | **74.37** | **59.17** |
| | w/o. sharpen | 39.01 | 22.34 | 72.11 | 57.62 |
| Refine noisy subset | w/. sharpen | 38.87 | 22.45 | 71.79 | 57.14 |
| | w/o. sharpen | **40.92** | **23.86** | **74.37** | **59.17** |

Table 4: The ablation study of the co-teaching paradigm on the ActivityNet Caption dataset with 50% noise ratio.

| Module | Change | R@1, IoU=0.5 | R@1, IoU=0.7 | R@5, IoU=0.5 | R@5, IoU=0.7 |
|---|---|---|---|---|---|
| Distinguish proposals | $\tau_2 = 0.45$ | 36.37 | 21.09 | 68.75 | 54.96 |
| | $\tau_2 = 0.50$ | 38.84 | 22.51 | 71.82 | 57.54 |
| | $\tau_2 = 0.55$ | **40.92** | 23.86 | **74.37** | **59.17** |
| | $\tau_2 = 0.60$ | 39.65 | **24.08** | 74.14 | 59.13 |
| Iterative learning | step = 5 | 39.16 | 22.35 | 72.25 | 57.84 |
| | step = 10 | 40.92 | **23.86** | 74.37 | **59.17** |
| | step = 15 | **41.03** | 23.79 | **74.42** | 58.99 |
| | step = 20 | 40.85 | 23.47 | 74.16 | 58.64 |

Table 5: The ablation study of the curriculum learning on the ActivityNet Caption dataset with 50% noise ratio.

in this paper, we conduct the main ablation study as shown in Table 2. Here, we remove the above two modules of CTR to build the baseline. This table shows that the baseline model achieves poor performance (similar to existing methods in Table1) on the noise setting. By adding the whole co-teaching paradigm module on the baseline, model ④ brings significant improvement of 12.06%, 5.77%, 26.46% and 19.55% on all metrics. It demonstrates that co-teaching paradigm helps to distinguish and rectify the noisy labels in the mixed training data. Model ①, ② and ③ also illustrate the contributions of different progress steps (*i.e.*, divide data, refine label, warm-up) in the co-teaching module. From model ⑤ and ⑥, we can find that the balanced weights and easy-to-hard controllers of curriculum learning module also contribute a lot to the final performance. Overall, Table 2 demonstrates that our designed modules are the keys to improve the model robustness to the noisy labels.

**Investigation on the co-teaching paradigm.** As shown in Table 4, we investigate different settings for each component in the co-teaching paradigm. As for dividing the clean and noisy data subsets, we find that GMM performs better than BMM. This is because BMM tends to produce undesirable flat distributions and may fail when the label noise is asymmetric, while GMM is flexible in the sharpness of distribution. The model achieves the best performance with GMM when the dividing threshold $\tau_1$ is set to 0.5. As for refining the labels of clean data, utilizing sharpen function helps to reduce the smoothness among the moment proposals for better optimization. Instead, sharpen function does not work on the noisy labels. We think this reason is that sharpen function may break the correspondence among the proposals and guide wrong focus on the negative proposals.

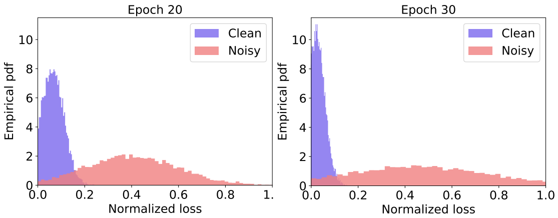


Figure 5: *Left:* The probability density function (PDF) on the clean and noisy sample when we re-train the model with 20 epochs. *Right:* The PDF when we re-train the model with 30 epochs.

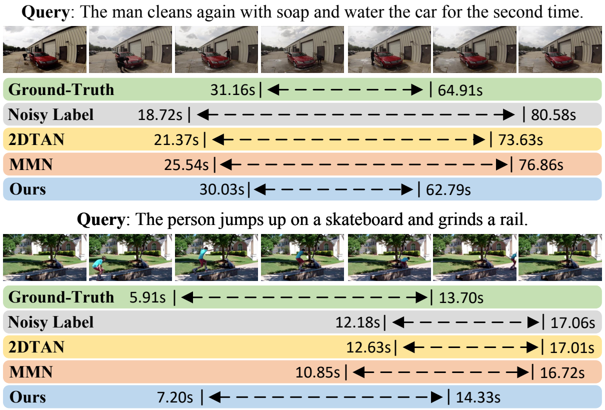

Query: The man cleans again with soap and water the car for the second time.
Ground-Truth 31.16s 64.91s
Noisy Label 18.72s 80.58s
2DTAN 21.37s 73.63s
MMN 25.54s 76.86s
Ours 30.03s 62.79s
Query: The person jumps up on a skateboard and grinds a rail.
Ground-Truth 5.91s 13.70s
Noisy Label 12.18s 17.06s
2DTAN 12.63s 17.01s
MMN 10.85s 16.72s
Ours 7.20s 14.33s

Figure 6: Qualitative examples on the noisy samples of the mixed training data. Our method is more robust.

**Investigation on the curriculum learning.** As shown in Table 5, we also investigate different settings of each component in the curriculum learning module. As for distinguishing the query-relevant and query-irrelevant moment proposals, the model achieves the best results when the threshold $\tau_2$ is set to 0.55. As for the iterative optimization strategy in curriculum learning, the model with either 10 steps or 15 steps achieves great performance. To balance the overall effectiveness and efficiency, we choose step=10 in our all experiments.

## 4.5. Qualitative Results

We first investigate the influence of the label refinement in our method. As shown in Figure 5, we carry out experiments by visualizing the loss distribution of different epochs. It shows that losses of the rectified soft labels of most clean pairs gradually become smaller while those of most noisy labels gradually become larger. Besides, the densities of the former gradually become larger while those of most noisy labels gradually become smaller. It demonstrates that our model enforces the similarity of true positives larger than that of the negatives during training, thus eliminating the negative impact of the noisy labels and then rectifying them for assisting the model learning. As shown in Figure 6, we further show the localization results on the noisy samples of the mixed training data. It illustrates that previous methods are not robust and tend to fit the noisy labels during the model training. Instead, our framework is able to recognize and rectify the noisy labels for better learning.

## 5. Conclusion

In this paper, we investigate a new but challenge problem of temporal activity localization (TAL), *i.e.*, learning TAL with noisy labels. To achieve this goal, we propose a novel Co-Teaching Regularizer (CTR) framework, to divide and rectify the noisy labels with two parallel models in a co-teaching manner. It can also tackle the problem of the partial data labelling by giving random boundaries to the unknown data. Experiments on three datasets demonstrate our effectiveness and robustness.

## 6. Bibliographical References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 233–242.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 41–48.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 162–171.

Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. 2020. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. 2015. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiang Fang and Yuchong Hu. 2020. Double self-weighted multi-view clustering via adaptive view fusion. *arXiv*.

Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Wu. 2021a. Animc: A soft approach for autoweighted noisy and incomplete multiview clustering. *TAI*.

Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Oliver Wu. 2020. V$^3$h: View variation and view heredity for incomplete multiview clustering. *TAI*.

Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Oliver Wu. 2021b. Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat. *TETCI*.

Xiang Fang, Daizong Liu, Wanlong Fang, Pan Zhou, Yu Cheng, Keke Tang, and Kai Zou. 2023a. Annotations are not all you need: A cross-modal knowledge transfer network for unsupervised temporal sentence grounding. In *Findings of EMNLP*.

Xiang Fang, Daizong Liu, Wanlong Fang, Pan Zhou, Zichuan Xu, Wenzheng Xu, Junyang Chen, and Renfu Li. 2024. Fewer steps, better performance: Efficient cross-modal clip trimming for video moment retrieval using language. *AAAI*.

Xiang Fang, Daizong Liu, Pan Zhou, and Yuchong Hu. 2022a. Multi-modal cross-domain alignment network for video moment retrieval. *TMM*.

Xiang Fang, Daizong Liu, Pan Zhou, and Guoshun Nan. 2023b. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *CVPR*, pages 2448–2460.

Xiang Fang, Daizong Liu, Pan Zhou, Zichuan Xu, and Ruixuan Li. 2022b. Hierarchical local-global transformer for temporal sentence grounding. *TMM*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5267–5275.

Lianli Gao, Pengpeng Zeng, Jingkuan Song, Yuan-Fang Li, Wu Liu, Tao Mei, and Heng Tao Shen. 2019. Structured two-stream attention network for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6391–6398.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in Neural Information Processing Systems (NIPS)*, 31.

Hao Jiang and Yadong Mu. 2022. Joint video summarization and moment localization by cross-task sample transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16388–16398.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

Nikola Konstantinov and Christoph Lampert. 2019. Robust learning from untrusted sources. In *International conference on machine learning*, pages 3488–3498. PMLR.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 706–715.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9972–9981.

Junnan Li, Richard Socher, and Steven CH Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. 2017. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918.

Daizong Liu, Xiang Fang, Wei Hu, and Pan Zhou. 2023a. Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding. *TMM*.

Daizong Liu, Xiang Fang, Xiaoye Qu, Jianfeng Dong, He Yan, Yang Yang, Pan Zhou, and Yu Cheng. 2024. Unsupervised domain adaptative temporal sentence localization with mutual information maximization. *AAAI*.

Daizong Liu, Xiang Fang, Pan Zhou, Xing Di, Weining Lu, and Yu Cheng. 2023b. Hypotheses tree building for one-shot temporal sentence localization. In *AAAI*.

Daizong Liu and Wei Hu. 2022a. Learning to focus on the foreground for temporal sentence grounding. In *COLING*.

Daizong Liu and Wei Hu. 2022b. Skimming, locating, then perusing: A human-like framework for natural language video localization. In *MM*.

Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. 2022a. Memory-guided semantic learning network for temporal sentence grounding. In *AAAI*.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, Guoshun Nan, Pan Zhou, Zichuan Xu, Lixing Chen, He Yan, and Yu Cheng. 2023c. Filling the information gap between video and query for language-driven moment retrieval. In *MM*.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. 2020a. Reasoning step-by-step: Temporal sentence localization in videos via deep rectification-modulation network. In *COLING*.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. 2021a. Adaptive proposal generation network for temporal sentence localization in videos. In *EMNLP*.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021b. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Zichuan Xu, Haozhao Wang, Xing Di, Weining Lu, and Yu Cheng. 2023d. Transform-equivariant consistency learning for temporal sentence grounding. *TOMM*.

Daizong Liu, Xiaoye Qu, and Wei Hu. 2022b. Reducing the vision and language bias for temporal sentence grounding. In *MM*.

Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020b. Jointly cross-and self-modal graph attention network for query-based moment localization. In *MM*.

Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. 2022c. Unsupervised temporal video grounding with deep semantic clustering. In *AAAI*.

Daizong Liu, Xiaoye Qu, and Pan Zhou. 2021c. Progressively guide to attend: An iterative alignment framework for temporal sentence grounding. In *EMNLP*.

Daizong Liu, Xiaoye Qu, Pan Zhou, and Yang Liu. 2022d. Exploring motion and appearance information for temporal sentence grounding. In *AAAI*.

Daizong Liu, Pan Zhou, Zichuan Xu, Haozhao Wang, and Ruixuan Li. 2022e. Few-shot temporal sentence grounding via memory-guided semantic learning. *TCSVT*.

Daizong Liu, Jiahao Zhu, Xiang Fang, Zeyu Xiong, Huan Wang, Renfu Li, and Pan Zhou. 2023e. Conditional video diffusion network for fine-grained temporal sentence grounding. *TMM*.

Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *Proceedings of the 41nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 15–24.

Tongliang Liu and Dacheng Tao. 2015. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461.

Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. 2019. DEBUG: A dense bottom-up grounding approach for natural language video localization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Zhanyu Ma and Arne Leijon. 2011. Bayesian estimation of beta mixture models with variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2160–2173.

Todd K Moon. 1996. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60.

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

Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. 2021. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Haim Permuter, Joseph Francos, and Ian Jermyn. 2006. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern recognition*, 39(4):695–706.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Yanyao Shen and Sujay Sanghavi. 2019. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pages 5739–5748. PMLR.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, pages 510–526.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5552–5560.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems (NIPS)*, 30.

Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. 2019. Combating label noise in deep learning using abstention. *arXiv preprint arXiv:1905.10964*.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497.

Arash Vahdat. 2017. Toward robustness against label noise in training deep discriminative neural networks. *Advances in neural information processing systems*, 30.

Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. 2022. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. 2020. Robust early-learning: Hindering the memorization of noisy labels. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699.

Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069.

Kun Yi and Jianxin Wu. 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7025.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning (ICML)*, pages 7164–7173.

Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019a. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *Advances in Neural Information Processing Systems (NIPS)*, pages 534–544.

Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019b. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166.

Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense regression network for video grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10287–10296.

Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019a. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1247–1257.

Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Parallel attention network with sequence matching for video grounding. *arXiv preprint arXiv:2105.08481*.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020a. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020b. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019b. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 655–664.