Local-Global Multi-Modal Distillation for Weakly-Supervised Temporal Video Grounding

Peijun Bao[1], Yong Xia[*2], Wenhan Yang[3], Boon Poh Ng[1], Meng Hwa Er[1], Alex C. Kot[1]

[1]Nanyang Technological University
[2]Northwestern Polytechnical University
[3]Peng Cheng Laboratory
peijun001@e.ntu.edu.sg, yxia@nwpu.edu.cn, yangwh@pcl.ac.cn, {ebpng, emher, eackot}@ntu.edu.sg

Abstract

This paper for the first time leverages multi-modal videos for weakly-supervised temporal video grounding. As labeling the video moment is labor-intensive and subjective, the weakly-supervised approaches have gained increasing attention in recent years. However, these approaches could inherently compromise performance due to inadequate supervision. Therefore, to tackle this challenge, we for the first time pay attention to exploiting complementary information extracted from multi-modal videos (*e.g.,* RGB frames, optical flows), where richer supervision is naturally introduced in the weakly-supervised context. Our motivation is that by integrating different modalities of the videos, the model is learned from synergic supervision and thereby can attain superior generalization capability. However, addressing multiple modalities would also inevitably introduce additional computational overhead, and might become inapplicable if a particular modality is inaccessible. To solve this issue, we adopt a novel route: building a multi-modal distillation algorithm to capitalize on the multi-modal knowledge as supervision for model training, while still being able to work with only the single modal input during inference. As such, we can utilize the benefits brought by the supplementary nature of multiple modalities, without undermining the applicability in practical scenarios. Specifically, we first propose a cross-modal mutual learning framework and train a sophisticated teacher model to learn collaboratively from the multi-modal videos. Then we identify two sorts of knowledge from the teacher model, *i.e.,* temporal boundaries and semantic activation map. And we devise a local-global distillation algorithm to transfer this knowledge to a student model of single-modal input at both local and global levels. Extensive experiments on large-scale datasets demonstrate that our method achieves state-of-the-art performance with/without multi-modal inputs.

Introduction

Given a natural language query and an untrimmed video, the task of temporal video grounding (Gao et al. 2017; Krishna et al. 2017) aims to temporally localize the video moment described by the language query. It is one of the most fundamental tasks in video understanding and has a wide range of real-world applications (Qi et al. 2021; Bao et al. 2023; Sreenu and Durai 2019; Zhu et al. 2021), such as video localization, video summarization, as well as video surveillance analysis. While achieving remarkable performance, the fully-supervised temporal video grounding (Liu et al. 2018; Zhang et al. 2019a,b, 2020a; Bao, Zheng, and Mu 2021) necessitates laborious manual annotations of temporal moment boundaries. Consequently, the weakly-supervised setting (illustrated in Fig 1) has recently received growing attention (Chen et al. 2020; Tan et al. 2021; Lin et al. 2020; Zheng et al. 2022a,b), where only paired videos and natural language queries are required during training. However, the grounding capability of the existing weakly-supervised methods is still unsatisfactory and lags behind the fully-supervised counterparts because the incomprehensive annotations do not provide sufficient supervisory signals.

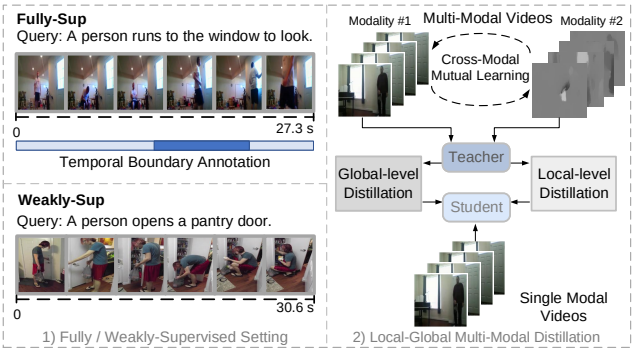

Figure 1: 1) Due to lacking temporal boundary annotations, weakly-supervised temporal video grounding faces ineffective supervision compared to fully-supervised scenarios. 2) We alleviate this issue by exploiting complementary multi-modal videos as an auxiliary supervisory signal. We propose a local-global multi-modal distillation algorithm that transfers the multi-modal knowledge from the teacher model to a single-modal student model at local and global levels.

Different from the prevailing works on weakly-supervised learning only consider RGB frames for video features (Gao et al. 2019; Chen et al. 2020; Lin et al. 2020; Tan et al. 2021; Zheng et al. 2022a,b), we pay attention to exploring the potential of using different modalities of the videos (*e.g.,* RGB frames, optical flow, audio), whose complementary informa-

*Corresponding author.

tion can naturally result in the improvement of the grounding accuracy. For instance, the features of RGB frames can capture useful appearances to align the objects and scenes between the sentence and video, while the explicit modeling of the motion is absent. Besides, they are also sensitive to occlusions and lighting conditions. Comparatively, optical flow features can complement this with richer motion information, which facilitates action understanding and improved robustness to occlusions and lighting changes. Therefore, intuitively, it is beneficial to utilize the synergic cues from the multi-modalities of the videos instead of only tackling RGB frames. However, while integrating multiple modalities[†] can improve the generalization capability and robustness of the model, it also brings about potential negative impacts. First, the additionally introduced model parameters lead to increased computational costs. Second, the use of multiple modalities limits the practicability of the method, both for computational consideration (*e.g.,* the heavy computational burden of optical flows (Dosovitskiy et al. 2015; Lucas and Kanade 1981)) and from the perspective of data availability (*e.g.,* audio modality is often missed in surveillance videos).

To this end, we develop a novel technical route to exploit multi-modal data more effectively and flexibly: 1) *training* the model with the *multi-modal complementary input*; 2) *inference* using only *single-modal data*. As such, the method successfully leads to an improved modeling capacity while maintaining practicality. As illustrated in Fig 1, our idea is to first train a sophisticated teacher model to collaboratively learn from the multi-modal videos. Subsequently, this teacher model is treated as a pseudo annotator to provide a student model with the ground truth of temporal boundaries, as well as the underlying semantic structure between the video and language. Because the student model only digests single-modal videos as input, it maintains the computational cost and eliminates additional multi-modal videos during inference. To the best of our knowledge, this is the first attempt of distilling the multi-modal knowledge to alleviate the challenge of weak supervision in the literature of temporal video grounding. Compared with the conventional knowledge distillation in fully / semi-supervised setting (Hinton, Vinyals, and Dean 2015; Tarvainen and Valpola 2017; Qiao et al. 2018), the case in our situation is more difficult, as the insufficient supervisory signals from incomplete annotations in our weakly-supervised context inherently pose a challenge.

Specifically, 1) we first devise a cross-modal mutual learning framework to train the teacher model under the scenarios of inputting multi-modal videos. The supplemental cues from different modality sources are leveraged to explicitly compensate for the errors of each single modality. 2) We then identify two sorts of knowledge from the teacher model, *i.e.,* temporal boundaries and semantic activation map. And we propose a multi-modal distillation algorithm to transfer this knowledge to a student model of single-modal input. At the local level, the semantic activation maps which denote the underlying similarity of video snippets and language are enforced to be consistent between the teacher and student model. At the global level, the predictions of temporal boundaries from the teacher model are regarded as pseudo labels to train the student model. In this way, the student model can exploit the extra knowledge from the multi-modal videos to handle the issue of weak supervisory signals, while keeping the single-modal videos as the input. 3) In addition, we propose a local-global contrastive learning algorithm for a single-modal baseline, where local and global levels of contrastive learning are devised to align the semantics of language and videos. This single-modal baseline model can still outperform state-of-the-art weakly-supervised methods even without touching any multi-modal videos during training or inference.

Our contributions are summarized as follows: 1) To the best of our knowledge, we are the first to make use of multi-modal videos to mitigate the inadequate supervision problem in weakly-supervised temporal video grounding. A multi-modal distillation algorithm is proposed to transfer knowledge to a single-modal student model at both local and global levels. 2) As a byproduct, we also for the first time explore the weakly-supervised temporal video grounding with the input of multi-modal videos. A mutual learning algorithm is crafted to collaboratively learn from different modality sources and compensate each other for reduced grounding errors. 3) We design a novel single-modal baseline with local-global contrastive learning, avoiding the utilization of multi-modal videos in either training or inference. 4) Extensive experiments on two large-scale datasets show that our methods achieve state-of-the-art results, regardless of whether employing multi-modal inputs.

Related Works

Fully-supervised temporal video grounding. The task of temporal video grounding is first introduced by Gao et al. (2017) which aims to determine the start and end time points of moment given by a query sentence. Liu et al. (2018) advise applying attention mechanism to highlight the crucial part of visual features. An event propagation network is developed in (Bao, Zheng, and Mu 2021) to localize video moments that are semantically related and temporally coordinated. While obtaining promising performance (Mun, Cho, and Han 2020; Wang, Ma, and Jiang 2020; Bao and Mu 2022; Zhang et al. 2019a, 2020a), these fully-supervised methods rely on the labor-intensive annotations of the temporal boundaries.

Weakly-supervised temporal video grounding. Existing works (Gao et al. 2019; Zheng et al. 2022a,b; Bao et al. 2024; Chen et al. 2020; Lin et al. 2020; Tan et al. 2021) on weakly-supervised temporal video grounding take the RGB frames of the video as the input. Early works (Mithun, Paul, and Roy-Chowdhury 2019; Tan et al. 2021) use joint visual-semantic embeddings and text-guided attention to avoid laborious temporal boundary annotations. Recently, Zheng et al. (2022a) design contrastive proposal learning to distinguish the positive video segments from the highly confusing ones within the same video. Different from existing works only considering RGB frames, we innovate to capitalize on synergic multi-modal videos as assistive training guidance to handle the dilemma of incomprehensive annotations.

[†]For clarity, single- / multi-modality in this paper refers specifically to the videos, although temporal video grounding itself is a multi-modal task in the definition.

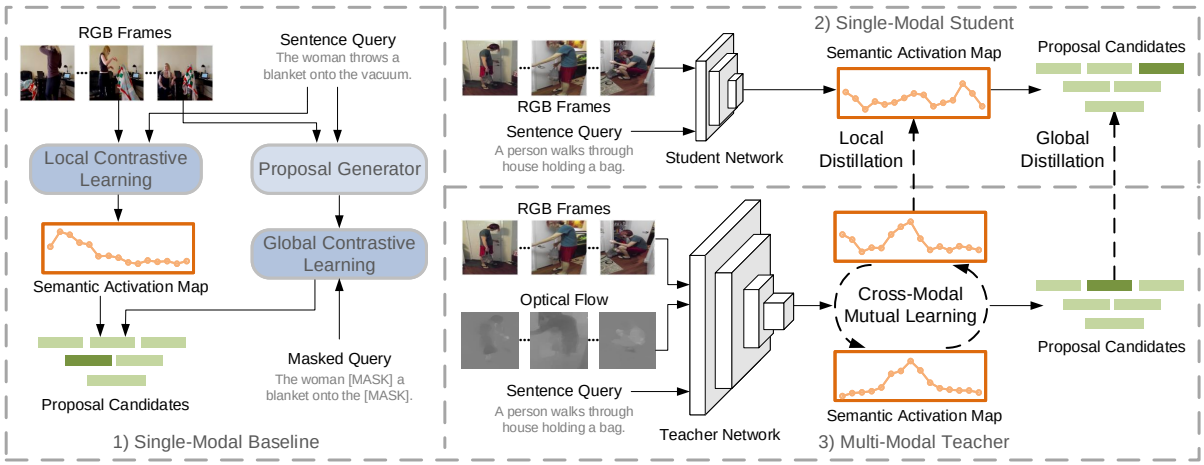

Figure 2: Overview of Local-Global Multi-Modal Distillation (MMDist). It comprises 1) a single-modal baseline using local-global contrastive learning, 2) a single-modal student model with a multi-modal distillation algorithm at local and global level, and 3) a multi-modal teacher model via cross-modal mutual learning. The proposal candidates that are in dark green represent the ones predicted as positive.

Multi-modal temporal video grounding. The only work in temporal video grounding that employs multi-modal videos is (Chen, Tsai, and Yang 2021). Their motivation is concentrated at the *feature level*: using multi-modal videos to augment the feature representation in the *fully-supervised* setting. We highlight that our motivation and formulation in the *weakly-supervised* context are distinct from theirs. Our goal to use multi-modal lies in *supervision level* i.e. addressing the deficient supervision problem by taking multi-modal as auxiliary supervision. This particular problem is unique to our weakly-supervised scenario and has not appeared in the fully-supervised counterpart. Besides, our formulation diverges from (Chen, Tsai, and Yang 2021) in that we only take multi-modal videos as extra supervision and do not require the multi-modal input during inference.

Knowledge distillation. Knowledge distillation is originally innovated in (Hinton, Vinyals, and Dean 2015) to transfer the knowledge acquired by a large, complex model to a smaller, more efficient model. In recent years, knowledge distillation is further applied in domain adaptation (Chen et al. 2019), zero-shot learning (Nayak et al. 2019), and multi-modal learning (Gupta, Hoffman, and Malik 2015; Wang et al. 2020). The most related works to ours are (Yu, Liu, and Chan 2021; Garcia, Morerio, and Murino 2018), which transfer knowledge of skeleton (Yu, Liu, and Chan 2021) or depth frames (Garcia, Morerio, and Murino 2018) to a student network of the RGB modality respectively. In contrast to them, we focus on the temporal grounding, and the identified local and global semantic knowledge to transfer is specific to our task.

Local-Global Multi-Modal Distillation

Method Overview

The proposed method Local-Global Multi-Modal Distillation (MMDist) explores leveraging multi-modal videos for weakly-supervised temporal video grounding (TVG). Our goal is not only to enhance the model with multi-modal input but further to take the multi-modal videos as auxiliary supervisory guidance for training the single-modal model, with the anticipation that it can mitigate the issue of deficient supervision. As illustrated in Fig 2, our methods consist of three parts: a single-modal baseline, a multi-modal teacher model, and a single-modal student model.

1) The single-modal baseline takes only single-modal videos as input. We propose local and global contrastive learning to align the semantic content of videos and sentences, simultaneously considering both local and global viewpoints.

2) The multi-modal teacher model collaboratively learns from multiple modality sources in videos. We devise cross-modal mutual learning to enforce the consistency of the semantic activation maps across different modalities. For each modality of the video, we first compute the semantic activation map between video snippets and query sentence respectively. Then the discrepancies arising from one modality are compensated for through the integration of the other, leading to enhanced overall performance and error mitigation.

3) The single-modal student model has the same network architecture design as the baseline model, but it receives additional supervision from the teacher model during training. More specifically, the multi-modal teacher model predicts more accurate temporal boundaries, whose ground truth is unknown in a weakly-supervised learning context. Also, the teacher model provides a better estimation of semantic activation maps, unveiling the intrinsic semantic relationship between language and videos. To this end, we design global-level and local-level distillation algorithms, which encourage the student model to mimic the predictions of temporal boundaries and semantic activation maps respectively. The student model is then trained with the supervisory signals from the multi-modal videos while still keeping the single-modal videos as the input during the inference stage.

Here we highlight our innovations. 1) We design local-global contrastive learning for the single-modal baseline.

Note that this baseline can beat state-of-the-art methods, without touching multi-modal videos during both training and inference. 2) Our student model is the first in the literature to capitalize on multi-modal videos to handle the insufficient supervision obstacle. And we craft a multi-modal distillation algorithm to distill multi-modal knowledge at both local and global scopes. 3) A novel cross-modal mutual learning framework is proposed for the teacher model to mutually compensate for the errors introduced by any single modality.

Contrastive Learning at Local and Global Level

The single-modal baseline aims to localize the temporal moment described in the sentence by using the single-modal video input in both training and testing. Previous approaches either solely emphasize semantic alignment between the overall proposal and language (Lin et al. 2020; Zheng et al. 2022a,b) *i.e.,* on a global scale, or specifically tackle the local similarity among the video snippets and sentence (Tan et al. 2021; Chen et al. 2020). However, local and global alignment can capture the underlying semantic structure and relationships between the sentence and video from different perspectives. Both of them serve to facilitate multi-modal knowledge transfer in the subsequent stages, thus establishing a foundational framework for the ensuing processes of local and global distillation. To this end, we propose to apply contrastive learning to simultaneously cater to both local and global scopes, formulating local-global contrastive learning.

Global contrastive learning. Our global contrastive learning module is similar to CPL network (Zheng et al. 2022b), which encompasses a proposal generator, and a sentence reconstructor. We use the proposal generator to generate a series of proposal candidates. These proposal candidates are defined by the center and width as (c_k, w_k) where $k = 1 \dots K$ and K is the number of proposal candidates. Then a transformer encoder as in CPL extracts the visual feature for the k-th proposal as v_k and sentence feature as q, with each feature vector having a dimension of d. The details of network architecture are omitted here and can be referred to (Zheng et al. 2022b). Then we randomly mask M words $w_i^m (i = 1 \dots W)$ in the sentence and enforce the reconstructor to reconstruct the masked words based on the video proposals, where W represents the number of words in the sentence. The reconstruction error is formulated as

$$\mathcal{L}_{rec} = \sum_{i=1}^{W} \mathcal{L}_{ce}(w_i^m) \tag{1}$$

The proposal that semantically matches the sentence query is regarded as a positive proposal, while the whole video is considered as a negative one. The positive proposal is assumed to be with a lower reconstruction error of the masked words than the negative proposal. We can heuristically select the positive proposal as the one k^* with the minimum reconstruction error

$$k^* = \text{argmin}_{k=1\dots K} \mathcal{L}_{rec}[k] \tag{2}$$

The global contrastive learning objective $\mathcal{L}^{\mathcal{B}}_{global}$ is formulated as

$$\begin{aligned} \mathcal{L}^{\mathcal{B}}_{global} &= \mathcal{L}_{rec}[k^*] + \mathcal{L}^{full}_{rec} \\ &+ \max(0, \mathcal{L}_{rec}[k^*] - \mathcal{L}^{full}_{rec} + \xi^{full}) \end{aligned} \tag{3}$$

where the reconstruction losses between the positive proposal and full video are contrasted with a margin of ξ^{full} and $\mathcal{L}^{full}_{rec}$ denotes the reconstruction loss by the full video.

Local contrastive learning. Specifically, we first enhance the local information of video snippet features $V \in \mathbb{R}^{L\times d}$ by applying a sequence of convolutional layers accompanied by the ReLU activation function, formulating context-enhanced local features $\hat{V} \in \mathbb{R}^{L\times d}$. Here L indicates the video snippet number and d is the channel dimension of the video features. Then we compute the semantic activation map $m \in \mathbb{R}^{L\times 1}$ which represents the semantic similarity between the video snippet and sentence as

$$m_l = \frac{\hat{V}_l \cdot q}{||\hat{V}_l|| \cdot ||q||} \tag{4}$$

where m_l signifies the value of semantic activation map for l-th video snippet and q denotes the sentence feature. Because the video is untrimmed, the foreground features relevant to the query sentence are intertwined with unrelated background elements. For a more accurate estimation of similarity between the i-th video and j-th sentence l^{ij} in a training batch, we adaptively select the top L_T values of m_l^{ij} and take their average, formulated as

$$l_{ij} = \sum_{l=1}^{L_T} \frac{\tilde{m}_l^{ij}}{L_T} \tag{5}$$

where $\tilde{m}^{ij}$ is a rearranged version of m^{ij}, sorted in descending order. Local contrastive learning encourages the model to maximize the similarity between the positive video-sentence pairs while minimizing the mismatched negative pairs. To achieve this, we first compute the probability p_i that the i-th video matches to the i-th sentence

$$p_i = \frac{\exp(\frac{l_{ii}}{\tau})}{\sum_{j=1}^{N} \exp(\frac{l_{ij}}{\tau})} \tag{6}$$

where τ is the temperature hyperparameter and N denotes the batch size. Then we can define the loss function of local contrastive learning $\mathcal{L}^{\mathcal{B}}_{local}$ as

$$\mathcal{L}^{\mathcal{B}}_{local} = -\frac{1}{N} \sum_{j=1}^{N} \log p_j \tag{7}$$

Local-global contrastive learning. The final objective function for local-global contrastive learning is formulated as

$$\mathcal{L}^{\mathcal{B}} = \mathcal{L}^{\mathcal{B}}_{global} + \alpha \mathcal{L}^{\mathcal{B}}_{local} \tag{8}$$

which jointly trains local and global contrastive learning. Here α is a weight hyperparameter to balance $\mathcal{L}^{\mathcal{B}}_{global}$ and $\mathcal{L}^{\mathcal{B}}_{local}$. The final score for the proposal p with start and end point (p_s, p_e) is computed from the local / global contrastive learning branches as

$$s_p = \gamma \sum_{l=p_s}^{p_e} \frac{m_l^{ii}}{p_e - p_s + 1} - r_p \tag{9}$$

where γ is a weight hyperparameter, m^{ii} signifies the semantic activation map of i-th video snippet and its query sentence, and r_p indicates the reconstruction error for proposal p as defined in Eq. 1. The proposal with the maximum score from the candidates is selected as the final prediction.

Multi-Modal Distillation at Local and Global Level

Assumed that one could train a powerful multi-modal model for weakly-supervised temporal video grounding (detailed in the subsection of "cross-modal mutual learning"). Thanks to utilizing accessory information from different modalities, the multi-modal model enjoys better localization accuracy and generalization capability than the single-modal one. But it also suffers from greater computational complexity and relies on multiple input modalities which might not be available in real-world applications. To alleviate this obstacle, we regard the multi-modality model as a teacher model $\mathcal{T}$ and transfer its multi-modal knowledge to a single-modal student model $\mathcal{S}$. The superiority of such multi-modal distillation lies in its ability to train the student model using the supervision of multiple modalities, while maintaining computational efficiency and taking single-modal input. Such a distillation paradigm can effectively cope with the deficient supervision obstacle for the weakly-supervised setting. We identify two sorts of multi-modal knowledge that are specific to our task, *i.e.,* knowledge of temporal boundaries at the global level, and knowledge of semantic activation maps at the local level. And correspondingly, a multi-modal distillation algorithm constituted by global-level distillation and local-level distillation is crafted to transfer these two sorts of knowledge respectively.

Global-level distillation. In the weakly-supervised scenarios, only the video-sentence pairs are provided for training, and the ground truth temporal boundaries are not available. The multi-modal teacher model enjoys the advantage of accuracy and robustness in making global-level predictions of temporal boundaries. Therefore, we treat the predictions from the teacher model as pseudo-labels for the student model. Assume that the teacher model selects the $k^{\mathcal{T}}$-th proposal candidate as the prediction. In the design of the single-modal baseline, we heuristically choose the proposal candidate with minimum reconstruction loss as the potential ground truth proposal. However, such selection is often inaccurate due to the lack of sufficient training supervision. So for the student model, instead, we explicitly set the prediction from the teacher model *i.e.,* $k^{\mathcal{T}}$-th proposal candidate as the pseudo ground truth to train the student model. The global-level distillation loss $\mathcal{L}^{\mathcal{S}}_{global}$ is formulated as

$$\begin{aligned} \mathcal{L}^{\mathcal{S}}_{global} &= \mathcal{L}_{global}[k^{\mathcal{T}}] \\ k^{\mathcal{T}} &= \text{argmax}_k s^{\mathcal{T}}_k \end{aligned} \tag{10}$$

where $s^{\mathcal{T}}_k$ is the prediction score for the k-th proposal candidate evaluated by the teacher model, and $\mathcal{L}_{global}$ is the global contrastive learning loss function defined as in the single-modal baseline.

Local-level distillation. The semantic activation map $m \in \mathbb{R}^{L\times 1}$ is an intermediate output that estimates the similarity of the query sentence and each snippet of video at the local level. Unlike global-level knowledge of temporal boundaries, the local-level knowledge of the activation map provides a deeper understanding of the underlying data structure and relationships between the language and video. Therefore, mimicking the semantic activation map provides valuable guidance to transfer the multi-modal knowledge from the teacher model to the student model, resulting in improved generalization capabilities for the student model without inputs of multi-modal videos. To achieve this, we devise the local-level distillation loss $\mathcal{L}^{\mathcal{S}}_{local}$ as a consensus of semantic activation between the teacher and student model:

$$\mathcal{L}^{\mathcal{S}}_{local} = \varphi(m^{\mathcal{S}}, m^{\mathcal{T}}) \tag{11}$$

where φ is the distance function of the activation map such as L_1 or L_2 norm.

The final loss $\mathcal{L}^{\mathcal{S}}$ to train the single-modal student model $\mathcal{S}$ consists of both the distillation loss and the original loss for the baseline model, written as

$$\mathcal{L}^{\mathcal{S}} = \mathcal{L}^{\mathcal{B}} + \beta(\mathcal{L}^{\mathcal{S}}_{global} + \alpha\mathcal{L}^{\mathcal{S}}_{local}) \tag{12}$$

where β is a hyperparameter to balance the weight between the distillation and baseline losses.

Cross-Modal Mutual Learning

This subsection describes the cross-modal mutual learning algorithm for the teacher model of multi-modality. The teacher model $\mathcal{T}$ digests inputs of multi-modal video features, denoted as $V_1, V_2 \in \mathbb{R}^{L\times d}$. For the global contrastive module and proposal generator, the video features of different modalities are early fused by concatenation. For the local contrastive module, we first generate the semantic activation maps for the two modalities as $m_1, m_2 \in \mathbb{R}^{L\times 1}$ respectively. The final semantic activation map $m_{\mathcal{T}}$ of teacher model $\mathcal{T}$ is integrated as the average of the two modalities:

$$m_{\mathcal{T}} = \frac{m_1 + m_2}{2} \tag{13}$$

Note that different modalities contain complementary information and can thus compensate for the errors of each other. To enable collaborative learning from different modalities, we design a cross-modal mutual learning objective, where discrepancies arising from one modality can be compensated for through the integration of supplemental modalities. In more detail, for the semantic activation map of one modality, we regard the one from the other modality as a reference. Then we enforce the consistency of the semantic activation map and its reference, formulated as

$$\mathcal{L}_{mutual} = \varphi(m_1, \delta(m_2)) + \varphi(m_2, \delta(m_1)) \tag{14}$$

where φ represents the distance function of two vectors such as $L1$ or L_2 norm, and δ signifies gradient stopping operation.

Experiments

Datasets and Evaluation Metrics

We validate the performance of the proposed methods against the state-of-the-art approaches on two large-scale datasets: 1) **Charades-STA** (Gao et al. 2017) includes 9,848 videos of daily indoor activities. The average length of a sentence query is 8.6 words, and the average duration of the video is 29.8 seconds. It is originally designed for action recognition / localization (Sigurdsson et al. 2016), and later extended by Gao *et al.* (Gao et al. 2017) with language descriptions for temporal video grounding. 2) **ActivityNet Captions** (Krishna et al. 2017) consists of 19,290 untrimmed videos, whose

Method	Charades-STA			ActivityNet Captions		
	R@0.3	R@0.5	R@0.7	R@0.1	R@0.3	R@0.5
SCN (Lin et al. 2020)	42.96	23.58	9.97	71.48	47.23	29.22
BAR (Wu et al. 2020)	44.97	27.04	12.23	–	49.03	30.73
MARN (Song et al. 2020)	48.55	31.94	14.81	–	47.01	29.95
RTBPN (Zhang et al. 2020b)	60.04	32.36	13.24	73.73	49.77	29.63
CCL (Zhang et al. 2020c)	–	33.21	15.68	–	50.12	31.07
WSTAN (Wang et al. 2022)	43.39	29.35	12.28	79.78	52.45	30.01
LCNet (Yang et al. 2021)	59.60	39.19	18.87	78.58	48.49	26.33
VCA (Wang, Chen, and Jiang 2021)	58.58	38.13	19.57	67.96	50.45	31.00
CPL (Zheng et al. 2022b)	66.40	49.24	22.39	79.86	53.67	31.24
MMDist Teacher	**70.11**	**54.72**	**26.00**	82.89	58.53	**32.98**
MMDist Baseline	67.26	51.58	24.22	82.27	56.92	31.80
MMDist Student	68.90	53.29	25.27	**83.11**	**58.69**	32.52

Table 1: Comparisons with state-of-the-art methods on two large-scale datasets.

contents are diverse and open. The average duration of the video is 117.74 seconds and the average length of the description is 13.16 words. There are 2.4 annotated moments with a duration of 8.2 seconds in each video.

Following previous works (Gao et al. 2017; Lin et al. 2020; Zheng et al. 2022b,a), we adopt the evaluation metric "R@m" to evaluate the grounding accuracy of our method. Specifically, we calculate the Intersection over Union (IoU) between the predicted temporal moment and the ground truth. Then "R@m" is defined as the percentage of language queries having correct grounding results with its IoU being larger than m. As previous works, we report the results with $m = \{0.3, 0.5, 0.7\}$ on Charades-STA dataset, and $m = \{0.1, 0.3, 0.5\}$ on ActivityNet-Captions dataset.

Implementation Details

We consider the RGB frames and optical flows as the multi-modalities for the input videos. And I3D network (Carreira and Zisserman 2017) and C3D network (Tran et al. 2015) are used to extract RGB features for Charades-STA and ActivityNet-Captions respectively. TV-L1 algorithm (Zach, Pock, and Bischof 2007) and the I3D network are applied to compute the optical flow features. For the query sentence, we use the pre-trained GloVe word2vec (Pennington, Socher, and Manning 2014) to extract word features. We set the maximum description length to 20 on both datasets. The vocabulary size is 8000 on ActivityNet-Captions and 1111 on Charades-STA respectively. We mask 1/3 of words in the query sentence for reconstruction. The dimensions of the hidden state d for both language and visual features are set to be 256. The number of video snippets L is resampled to 200 on both datasets. We use the Adam optimizer (Kingma and Ba 2014) for the model training with a batch size of 32. For multi-modal distillation, we first train the teacher model with 15 epochs with a learning rate of 0.00035, and then distill it to the student model with another 15 epochs with a learning rate of 0.0005. The training of the single-modal baseline is independent of the teacher / student models, where the number of training epochs and learning rate for it are set to 15 and 0.0004 respectively. The hyperparameter of α, β, and γ is set to 4.5 0.9, and 3.0 respectively.

Variants	multi-modality	distillation
MMDist Teacher	✓	✗
MMDist Baseline	✗	✗
MMDist Student	✗	✓

Table 2: The difference between the variants of our models.

Performance Comparisons

Our methods have three variant models i.e., a multi-modal teacher model (MMDist Teacher), a single-modal baseline (MMDist Baseline), and a single-modal student model (MMDist Student). Table 2 presents their distinctions in the training and testing settings. While MMDist Teacher takes the multi-modal videos of RGB frames and optical flow as input, the other two models only consume the input of single-modal videos *i.e.,* RGB frames. The MMDist Student differs from the Baseline in that it exploits the distillation algorithm to learn from the teacher model. We verify the capability of the proposed methods on two widely-used datasets *i.e.,* Charades-STA and ActivityNet-Captions. Table 1 illustrates the performance comparison of our methods to previous methods of weakly-supervised TVG. All three proposed models beat the state-of-the-art methods by a clear margin. We denote the previous best method Gaussian-based Contrastive Proposal Learning (Zheng et al. 2022b) as CPL. The details of the comparison are concluded as follows.

1) MMDist Baseline is better than CPL. The MMDist Baseline model exclusively employs RGB frames from videos and refrains from incorporating any multi-modal videos throughout its training and testing phases. This ensures a fair comparison with state-of-the-art methods such as CPL. As indicated, MMDist Baseline consistently surpasses the previous best methods in all evaluation metrics. For instance, our proposed baseline achieves about 2 points higher than

Method	R@0.3	R@0.5	R@0.7
Baseline w/o l-cont	65.83	50.54	23.31
Baseline w/o g-cont	60.42	44.36	20.80
Baseline full	**67.26**	**51.58**	**24.22**

Table 3: Ablation study on the baseline.

Method	R@0.3	R@0.5	R@0.7
Student baseline	67.26	51.58	24.22
Student w/o l-dis	68.21	51.68	24.45
Student w/o g-dis	68.02	52.85	24.98
Student full	**68.90**	**53.29**	**25.27**

Table 4: Ablation study on our student model.

Method	R@0.3	R@0.5	R@0.7
Teacher baseline	67.92	52.31	24.64
Teacher w/o mutual	69.73	53.86	24.76
Teacher full	**70.11**	**54.72**	**26.00**

Table 5: Ablation study on the teacher model.

CPL in the metric of "R@0.5" on Charades-STA dataset, and 3.5 points higher in "R@0.3" on ActivityNet-Captions. This indicates that despite the absence of any utilization of multi-modal videos during training, our local-global contrastive learning baseline still exhibits better grounding ability.

2) MMDist Teacher surpasses CPL / MMDist Baseline. The incorporation of multi-modal inputs consistently leads to significant improvements across all evaluation metrics for the MMDist Teacher. The improvements can be attributed to the supplemental cues contained in multi-modal input sources, helpful in aligning the semantics of the video and the language query. MMDist Teacher demonstrates a 16.1% improvement in the metrics of R@0.7 on the Charades-STA and a 9.1% improvement in the metrics of R@0.3 on the ActivityNet-Captions, compared to CPL. Also, the grounding capability of MMDist Teacher is superior to MMDist Baseline with the extra information from multi-modal videos. This verifies the superiority of enhancing the model with multi-modal videos and the effectiveness of the proposed cross-modal mutual learning.

3) MMDist Student outperforms MMDist Baseline. Even without using multi-modal videos as input, MMDist Student outperforms MMDist Baseline significantly by the multi-modal knowledge distilled from the teacher model. And the grounding accuracy of MMDist Student also evidently surpasses CPL. We highlight that the MMDist Student model achieves almost similar accuracy to Teacher in ActivityNet-Captions where the student model is slightly better than the teacher in R@0.1 and R@0.3 (about 0.2 points), while slightly worse in R@0.5(about 0.4 points).

Ablation Studies

1) **The effectiveness of local-global contrastive learning**. We investigate the effectiveness of each proposed module on the model's performance and conduct ablation studies on the Charades-STA dataset. Table 3 explores the impact of local contrastive learning and global contrastive learning. When either local or global contrastive learning is removed, the model's performance declines significantly in each evaluation metric. This verifies the effectiveness of local-global contrastive learning and the necessity of learning the semantic alignment between the video and language at both local and global levels. Moreover, the following ablation study on multi-modal distillation further reveals that the semantic activation maps from the local contrastive learning play an important role in transferring the multi-modal knowledge from the teacher to the student model.

2) **The benefit of local-global multi-modal distillation**. This paper identifies two sorts of knowledge from the multi-modal teacher i.e., temporal boundaries at the global level and semantic activation maps at the local level. We correspondingly craft the multi-modal distillation algorithm composed of local and global level distillation to transfer them. Table 4 summarizes the ablation study on local and global distillation. On the one hand, when removing either of them from the full model, the performance decreases evidently. And especially when removing the local distillation part, the metric of "R@0.5" drops more than 1.5 points. On the other hand, both of them are still better than the "Student baseline". Here "Student baseline" denotes our single-modal baseline. This underscores that both local-level and global-level distillations are effective in leveraging the multi-modal training guidance.

3) **The efficacy of cross-modal mutual learning**. Here we study the effectiveness of cross-modal mutual learning in the teacher model. Table 5 presents the model accuracy after discarding the loss function of mutual learning. All three evaluation metrics values show about an evident drop. We also design a multi-modal baseline for the teacher model i.e., "Teacher baseline", which directly concatenates the multi-modal features at the input level. The model's localization accuracy surpasses our single-modal baseline with a clear margin, thanks to the supplementary information offered by the multi-modal videos. However, the teacher baseline model exhibits noticeable performance inferiority as it lacks the capability for collaborative learning from multi-modality.

Conclusion

This paper for the first time exploits multi-modal videos for weakly-supervised temporal video grounding. Firstly, we propose a cross-modal mutual learning framework to collaboratively train a teacher model with the input of multi-modal videos. Secondly, we devise local and global level distillation algorithms to transfer this knowledge from the teacher model to a single-modal student model. Moreover, we introduce a local-global contrastive learning framework as a baseline where the semantic contents of video and language are simultaneously aligned at both local and global scopes. Extensive experiments demonstrate the effectiveness of our methods on two widely-used datasets.

Acknowledgements

This work was carried out at the Rapid-Rich Object Search (ROSE) Lab, School of EEE, NTU, Singapore. The research is supported in part by the NTU-PKU Joint Research Institute (a collaboration between the Nanyang Technological University and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation).

References

Bao, P.; and Mu, Y. 2022. Learning Sample Importance for Cross-Scenario Video Temporal Grounding. In *ICMR*.

Bao, P.; Shao, Z.; Yang, W.; Ng, B. P.; Er, M. H.; and Kot, A. C. 2024. Omnipotent Distillation with LLMs for Weakly-Supervised Natural Language Video Localization: When Divergence Meets Consistency. In *AAAI*.

Bao, P.; Yang, W.; Ng, B. P.; Er, M. H.; and Kot, A. C. 2023. Cross-modal Label Contrastive Learning for Unsupervised Audio-Visual Event Localization. In *AAAI*.

Bao, P.; Zheng, Q.; and Mu, Y. 2021. Dense Events Grounding in Video. In *AAAI*.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*.

Chen, Y.-C.; Lin, Y.-Y.; Yang, M.-H.; and Huang, J.-B. 2019. CrDoCo: Pixel-Level Domain Transfer With Cross-Domain Consistency. In *CVPR*.

Chen, Y.-W.; Tsai, Y.-H.; and Yang, M.-H. 2021. End-to-end Multi-modal Video Temporal Grounding. In *NeurIPS*.

Chen, Z.; Ma, L.; Luo, W.; Tang, P.; and Wong, K.-Y. K. 2020. Look Closer to Ground Better: Weakly-Supervised Temporal Grounding of Sentence in Video. *ArXiv*.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Häusser, P.; Hazirbas, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; and Brox, T. 2015. FlowNet: Learning Optical Flow with Convolutional Networks. In *ICCV*.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *ICCV*.

Gao, M.; Davis, L. S.; Socher, R.; and Xiong, C. 2019. WSLLN:Weakly Supervised Natural Language Localization Networks. In *EMNLP*.

Garcia, N. C.; Morerio, P.; and Murino, V. 2018. Modality Distillation with Multiple Stream Networks for Action Recognition. In *ECCV*.

Gupta, S.; Hoffman, J.; and Malik, J. 2015. Cross Modal Distillation for Supervision Transfer. In *CVPR*.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *ArXiv*, abs/1503.02531.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. In *ICLR*.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *ICCV*.

Lin, Z.; Zhao, Z.; Zhang, Z.; Wang, Q.; and Liu, H. 2020. Weakly-Supervised Video Moment Retrieval via Semantic Completion Network. In *AAAI*.

Liu, M.; Wang, X.; Nie, L.; Tian, Q.; Chen, B.; and Chua, T.-S. 2018. Cross-modal moment localization in videos. In *ACM MM*.

Lucas, B. D.; and Kanade, T. 1981. An Iterative Image Registration Technique with an Application to Stereo Vision. In *IJCAI*.

Mithun, N. C.; Paul, S.; and Roy-Chowdhury, A. K. 2019. Weakly Supervised Video Moment Retrieval From Text Queries. In *CVPR*.

Mun, J.; Cho, M.; and Han, B. 2020. Local-Global Video-Text Interactions for Temporal Grounding. In *CVPR*.

Nayak, G. K.; Mopuri, K. R.; Shaj, V.; Babu, R. V.; and Chakraborty, A. 2019. Zero-Shot Knowledge Distillation in Deep Networks. In *ICCV*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*.

Qi, M.; Qin, J.; Yang, Y.; Wang, Y.; and Luo, J. 2021. Semantics-Aware Spatial-Temporal Binaries for Cross-Modal Video Retrieval. *TIP*.

Qiao, S.; Shen, W.; Zhang, Z.; Wang, B.; and Yuille, A. L. 2018. Deep Co-Training for Semi-Supervised Image Recognition. In *ECCV*.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. K. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *ECCV*.

Song, Y.; Wang, J.; Ma, L.; Yu, Z.; and Yu, J. 2020. Weakly-Supervised Multi-Level Attentional Reconstruction Network for Grounding Textual Queries in Videos. *ArXiv:2003.07048*.

Sreenu, G.; and Durai, M. A. S. 2019. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *Journal of Big Data*.

Tan, R.; Xu, H.; Saenko, K.; and Plummer, B. A. 2021. LoGAN: Latent Graph Co-Attention Network for Weakly-Supervised Video Moment Retrieval. In *WACV*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*.

Tran, D.; Bourdev, L. D.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV*.

Wang, J.; Ma, L.; and Jiang, W. 2020. Temporally Grounding Language Queries in Videos by Contextual Boundary-aware Prediction. In *AAAI*.

Wang, Q.; Zhan, L.; Thompson, P. M.; and Zhou, J. 2020. Multimodal Learning with Incomplete Modalities by Knowledge Distillation. In *KDD*.

Wang, Y.; Deng, J.; gang Zhou, W.; and Li, H. 2022. Weakly Supervised Temporal Adjacent Network for Language Grounding. *TMM*.

Wang, Z.; Chen, J.; and Jiang, Y.-G. 2021. Visual Co-Occurrence Alignment Learning for Weakly-Supervised Video Moment Retrieval. In *ACM MM*.

Wu, J.; Li, G.; Han, X.; and Lin, L. 2020. Reinforcement Learning for Weakly Supervised Temporal Grounding of Natural Language in Untrimmed Videos. In *ACM MM*.

Yang, W.; Zhang, T.; Zhang, Y.; and Wu, F. 2021. Local Correspondence Network for Weakly Supervised Temporal Sentence Grounding. *TIP*.

Yu, B. X. B.; Liu, Y.; and Chan, K. C. C. 2021. Multimodal Fusion via Teacher-Student Network for Indoor Action Recognition. In *AAAI*.

Zach, C.; Pock, T.; and Bischof, H. 2007. A duality based approach for realtime tv-l 1 optical flow. In *PR*.

Zhang, D.; Dai, X.; Wang, X.; Wang, Y.-F.; and Davis, L. S. 2019a. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020a. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *AAAI*.

Zhang, Z.; Lin, Z.; Zhao, Z.; and Xiao, Z. 2019b. Cross-Modal Interaction Networks for Query-Based Moment Retrieval in Videos. In *ACM SIGIR*.

Zhang, Z.; Lin, Z.; Zhao, Z.; Zhu, J.; and He, X. 2020b. Regularized Two-Branch Proposal Networks for Weakly-Supervised Moment Retrieval in Videos. In *ACM MM*.

Zhang, Z.; Zhao, Z.; Lin, Z.; Zhu, J.; and He, X. 2020c. Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding. In *NeurIPS*.

Zheng, M.; Huang, Y.; Chen, Q.; and Liu, Y. 2022a. Weakly Supervised Video Moment Localization with Contrastive Negative Sample Mining. In *AAAI*.

Zheng, M.; Huang, Y.; Chen, Q.; Peng, Y.; and Liu, Y. 2022b. Weakly Supervised Temporal Sentence Grounding with Gaussian-based Contrastive Proposal Learning. In *CVPR*.

Zhu, W.; Lu, J.; Li, J.; and Zhou, J. 2021. DSNet: A Flexible Detect-to-Summarize Network for Video Summarization. *TIP*.