Weakly Supervised Temporal Sentence Grounding with Uncertainty-Guided Self-training

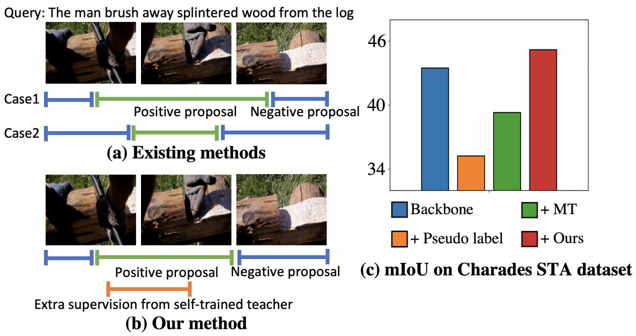
Yifei Huang[1,2*], Lijin Yang[1*], Yoichi Sato[1]
[1]The University of Tokyo, [2]Shanghai Artificial Intelligence Laboratory
{hyf, yang-lj, ysato}@iis.u-tokyo.ac.jp

Abstract

The task of weakly supervised temporal sentence grounding aims at finding the corresponding temporal moments of a language description in the video, given video-language correspondence only at video-level. Most existing works select mismatched video-language pairs as negative samples and train the model to generate better positive proposals that are distinct from the negative ones. However, due to the complex temporal structure of videos, proposals distinct from the negative ones may correspond to several video segments but not necessarily the correct ground truth. To alleviate this problem, we propose an uncertainty-guided self-training technique to provide extra self-supervision signal to guide the weakly-supervised learning. The self-training process is based on teacher-student mutual learning with weak-strong augmentation, which enables the teacher network to generate relatively more reliable outputs compared to the student network, so that the student network can learn from the teacher's output. Since directly applying existing self-training methods in this task easily causes error accumulation, we specifically design two techniques in our self-training method: (1) we construct a Bayesian teacher network, leveraging its uncertainty as a weight to suppress the noisy teacher supervisory signals; (2) we leverage the cycle consistency brought by temporal data augmentation to perform mutual learning between the two networks. Experiments demonstrate our method's superiority on Charades-STA and ActivityNet Captions datasets. We also show in the experiment that our self-training method can be applied to improve the performance of multiple backbone methods.

1. Introduction

One of the most important directions in video understanding is to temporally localize the start and end timestamp of a given sentence description. Also known as temporal sentence grounding, this task has a wide range of potential applications ranging from video summarization [45,66], video action segmentation [23,28,59], to Human-computer interaction systems [8,22,30,52,63]. While most existing works deal with this task in a supervised manner, manually annotating temporal labels of the starting and ending timestamps of each sentence is extremely laborious, which harms the scalability and viability of this task in real-world applications. To escalate practicability, recent research attention has been drawn towards weakly supervised temporal sentence grounding, where video-language correspondence is given as annotation only at video-level for model training.



(a) Existing methods

(b) Our method

(c) mIoU on Charades STA dataset

Figure 1. (a) Existing methods [70, 71] find it hard to distinguish the two cases since they learn positive proposals purely based on negative proposals. (b) Our method provides extra supervision signals for learning positive proposals. (c) Performance of the backbone network [71], backbone network trained with existing self-training methods pseudo labeling [29], Mean Teacher (MT) [50], and backbone network trained with our method. Directly applying self-training methods for semi-supervised learning negatively influences the performance, while our self-training method can improve the backbone performance.

Previous weakly supervised temporal sentence grounding works [16,21,36,38] mainly adopt the multiple instance learning (MIL) method. They generate mismatched video-language pairs as negative samples and train the model to distinguish the positive/negative samples, in order to learn a cross-modal latent space for a language feature to highlight a certain time period of the video. Some methods find negative samples by selecting sentences that describe another

[*] equal contribution. Author order is determined by a coin toss.

video [38, 62], but these negative samples are often easy to distinguish and thus cannot provide strong supervision signals. Recent works [68, 70, 71] select negative samples by sampling video segments within the same video, allowing the model to distinguish more confusing video segments.

One major limitation of these methods is that they learn the models completely depending on negative samples, since the objectives of these methods are to generate positive proposals that are distinct from the negative ones, where the distance is usually measured by a certain metric such as the ability to reconstruct the query using only the video segment inside the proposal [32, 60, 70, 71]. However, due to the complex temporal structure of videos that often contain multiple events, being distinct from the negative proposals does not always guarantee the quality of the positive proposals. For example in Figure 1(a), it is hard for existing methods like [70, 71] to distinguish the two cases since in both cases the positive proposals can better reconstruct the query sentence than the negative proposals.

However, in the absence of strong supervision, it is not straightforward to positively guide the process of temporal sentence grounding. Our solution is to leverage self-training to produce extra supervision signals (Figure 1(b)). As for self-training, one may consider to directly apply existing techniques originally designed for semi-supervised learning such as pseudo label [29] or Mean Teacher with weak-strong augmentation [50]. However, as shown in Figure 1(c), our preliminary experiment suggests that the teacher's supervision tends to be noisy and would degrade performance due to error accumulation. This is mainly because unlike semi-supervised learning [72], no strong supervision is used for initializing the teacher network.

Following previous works [12, 31, 34, 50], our method also apply the weak-strong augmentation technique, where the student network takes data with strong augmentation as input, while the teacher network gets as input weakly augmented data. Thus, compared to the student network, the teacher network can generate output less affected by heavy augmentation, providing supervisory guidance to the student network. To realize self-training in the weakly supervised temporal sentence grounding task, we specifically design the following two techniques: **(1)** As the teacher network itself is initially trained with only weak supervision and may generate erroneous supervision signals, we apply a Bayesian teacher network, enabling an uncertainty estimation of its output. The estimated uncertainty is used to weigh the teacher supervision signal thus reducing the chance of error accumulation. **(2)** To efficiently update both networks, we develop cyclic mutual learning, where the forward cycle forces the student network to output temporally consistent representations with the teacher, and the backward cycle encourages the teacher's output to be consistent with the average of multiple student outputs generated by inputs with different augmentations. This mutual-learning method allows the teacher to update more carefully than the student, preventing over-fitting to the low-quality supervision. On the other hand, a better teacher will provide reliable uncertainty measures for learning the student network. Our self-training technique can be applied to most existing methods and we observe performance improvement on multiple public datasets.

Our contributions can be summarized as follows: (1) We propose a novel method for temporal sentence grounding based on self-training. To the best of our knowledge, this is the *first attempt* to apply self-training to the weakly supervised temporal sentence grounding task. (2) To realize self-training for this task, we design a Bayesian teacher network to alleviate the negative effect of low-quality teacher supervision, and we use a mutual-learning strategy based on the consistency of the data augmentation to better update the teacher and student networks. (3) Our experiments on two standard datasets Charades-STA and ActivityNet Captions demonstrate that our method can effectively improve the performance of existing weakly supervised methods.

2. Related Work

Temporal sentence grounding with strong supervision. Many previous works focus on Temporal sentence grounding with strong supervision [1, 17, 19, 33, 43, 64]. With precise start and end timestamps annotations for each video and query pair, TALL [15] makes the first attempt to directly regress the start and end timestamp with video and language inputs. LGI [39] further used multi-granularity textual features and predict the timestamps considering local-global video-text interactions. However, these require manual annotation of temporal boundaries for each sentence, which is labor-consuming and subjective [42] (inconsistent among different annotators). This harms the potential of these approaches in real-world applications.

Weakly supervised temporal sentence grounding. To avoid laborious annotation and subjective annotation bias, methods for weakly supervised temporal sentence grounding do not use precise start and end timestamps, but only use video-level video-sentence correspondence during training [10, 49, 58, 69]. Without explicit temporal annotations, one group of methods [16, 21, 36, 38] adopt the multi-instance learning (MIL) technique. These methods construct negative video-language pairs by selecting sentences from other videos, and learning the video-level visual-text correspondence by maximizing the matching scores of the positive pairs while suppressing that of the negative pairs. Then the learned correspondence is used to find the optimal temporal regions that best match the given queries during inference. However, generating negative pairs either from other videos [38] or within the same video [70] can only encourage the models to output proposals that are distinct

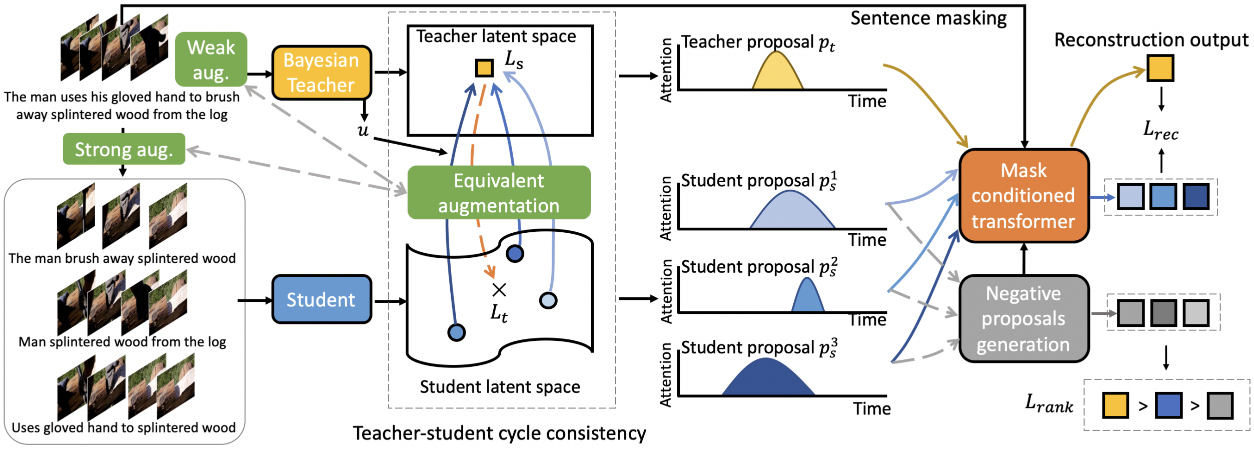


Figure 2. Overview of our proposed method. The teacher network takes as input weakly augmented data while the student network takes as input multiple strongly augmented data. Then teacher-student cycle consistency is used for mutual learning of the two networks, considering the uncertainty u into consideration. Gaussian masks are generated to represent the proposals, and we further use reconstruction loss L_{rec} and ranking loss L_{rank} to ensure high-quality proposals.

from the negative proposals. Since videos usually contain multiple complex temporal events, proposals distinct from the negative ones may just represent some other events but not correspond with the ground truth. In our method, we design a self-training method based on a teacher-student structure, where the teacher network can provide extra self-supervision signals to learn a better student network, and inversely the student network transfers learned knowledge to the teacher network by cycle consistency.

Another line of research aims to select the video segments which can best reconstruct the given query sentence [32, 48, 70]. The reconstruction result can also be used for contrastive learning [71]. In our method, we also leverage the reconstruction performance to guide the mutual learning process of the teacher-student method.

Self-training in weakly supervised learning. Self-training is originally proposed in semi-supervised learning and has been adopted in other scenarios such as domain adaptation [5, 31]. Many methods also use self-training to improve the model performance for weakly supervised tasks, for example, text classification [37], semantic segmentation [35, 46, 55] or object/action detection [6,9,24,53,57,61,65,67]. To the best of our knowledge, we make the first attempt to explore the use of self-training on weakly supervised temporal sentence grounding.

Bayesian deep learning. To provide posterior uncertainty estimates, there has been a long presence of Bayesian inference in machine learning [2]. Since Bayesian inference on neural networks is difficult, early works explored a variety of methods such as Markov Chain Monte Carlo (MCMC) [40] or variational inference [20]. Bayesian deep learning has thus been applied to various tasks such as unsupervised domain adaptation [5] and time series forecasting [25]. In this work, we utilize a Bayesian network to acquire uncertainty estimation for self-training.

3. Proposed Method

Problem formulation and overview. We first demonstrate the problem formulation before going into details of our proposed method. Given a set of N videos $\{v_1, \cdots, v_N\}$ and their corresponding query sentences $\{q_1, \cdots, q_N\}$ that describe each video, our goal is to ground each sentence to a specific temporal segment in video with start and end timestamps.

Figure 2 shows the overview of our method. Our self-training method consists of a Bayesian network as the teacher network and an identical network as the student network. We can apply the network architecture of most existing weakly supervised methods as the teacher/student. In the following part of this section, we showcase the backbone with the state-of-the-art method CPL [71] which outputs Gaussian attention proposals. Following the weak-strong augmentation [34, 41], the teacher network takes a weakly augmented video-language pair as input, whereas the student network takes a strongly augmented video-language pair as input. Both networks are first initialized with the training approach of the backbone. We then perform self-training and update both networks using uncertainty estimated by the teacher network (Section 3.1) and temporal augmentation cycle-consistency (Section 3.2).

3.1. Uncertainty estimation via Bayesian teacher

Since the teacher network itself is not learned by strong supervision, it may generate low-quality supervision signals even given weak augmentation. In fact, our preliminary experiment in Figure 1(c) shows that directly applying the output of the teacher network as supervision can even downgrade the overall performance. Thus, it is essential to suppress the influence of low-quality outputs. Inspired by the success of Bayesian deep learning [26], we propose to use Bayesian inference on the teacher network to get an uncertainty estimation.

Since all parameters are considered as random variables in a Bayesian network, obtaining the posterior distribution is often intractable. Recent works use variational inference as an approximation [3]: given an input I, the predictive distribution of output O is acquired by D-time repeated stochastic forward passes with network parameters sampled from an approximating variational distribution $q(w)$:

$$\begin{aligned} p(O|I) &= \int p(O|I,w)q(w)dw \\ &\approx \frac{1}{D}\sum_{i=1}^{D} p(O|I,w_i), \quad w_i \sim q(w), \end{aligned} \tag{1}$$

where $p(O|I,w_i)$ is one forward pass with model parameters w_i. In practice, we use the trick in [14] to perform Bayesian inference without changing model structure and parameter by sampling model parameters with dropout. Using a temporal sentence grounding model $F(v,q,w)$ with weights w which outputs a temporal segment proposal p given one video v and a sentence query q as input, the uncertainty estimation u can be computed as:

$$\bar{p} = \frac{1}{D}\sum_{i=1}^{D} F(v,q,w_i), \quad w_i \sim dropout(w) \tag{2}$$

$$u = \frac{1}{D}\sum_{i=1}^{D} p_i^2 - \bar{p}^2. \tag{3}$$

This uncertainty estimation is then used in the teacher-student mutual learning stage to alleviate the negative influence of the low-quality supervision signals from the teacher network.

As a proof of concept, we visualize the correlation between the outputs' uncertainty and their mean Intersection over Union (mIoU) with the ground truth segment in v that corresponds to the given query sentence q. For better visualization, we transform uncertainty to confidence by $1-u$, and visualize it with the mIoU score on the test set of the Charades-STA dataset in Figure 3. From the figure, we can see that samples with low IoU scores tend to also have low confidence scores. This study shows that we can leverage uncertainty measurement u to represent the quality of the teacher network's output.

3.2. Mutual learning with temporal augmentation cycle consistency

To effectively encourage the student to learn knowledge from teacher, and allow the learned knowledge to transfer back to the teacher, we design teacher-student mutual learning with temporal augmentation cycle consistency. We feed the student network a set of temporally augmented videos, including temporal scaling, shifting, and masking. We also add augmentation to the sentence queries by decomposition using semantic role labeling [18] (details in Section 3.2). The teacher network is given only weakly augmented videos and sentence queries as input. Details of the mutual learning are as follows:

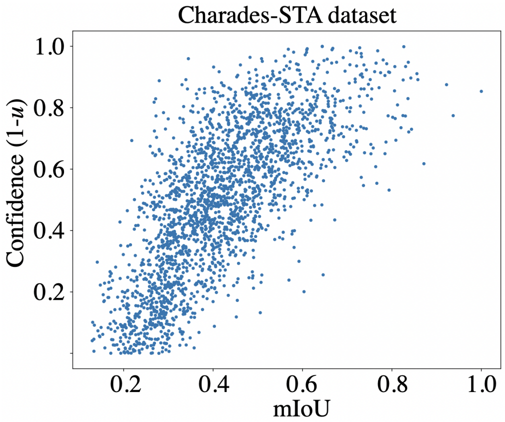


Figure 3. Model confidence (computed by $1-u$) and mIoU on the Charades-STA dataset are highly correlated, indicating that we can leverage the uncertainty estimation u to represent the quality of the network output.

Model initialization. The first step is to initialize the teacher and student models. Initialization is a critical step for all self-training methods since we rely on the teacher to generate reliable supervision signals to optimize the student network. In our self-training method, it is possible to apply most existing weakly supervised temporal sentence grounding methods to initialize the model. We denote the initialization loss of the model as L_{init}. Note that the teacher and student models are initialized with the same parameters.

Data augmentation. We use the weak-strong augmentation strategy to allow the teacher network provide supervision signal to the student network. For the data augmentation, we apply random temporal shifting, random temporal scaling, and random temporal masking on the input video. Specifically, we first randomly scale the temporal length of each video with a ratio of $l\%$, and crop the scaled video to its original size with a temporal shifting ratio $s\%$ (*i.e.*, the start timestamp of cropping is at $s\%$ of the scaled length). After this, we randomly choose $m\%$ of the timestamps and replace the feature on the corresponding timestamp with zero vectors. As for the sentence query, we randomly drop words at 50% probability while keeping the sentence containing all words in at least one semantic structure decomposed by semantic role labeling (SRL) [18]. We repeat this augmentation k times, thus each pair of video-sentence input (v_t, q_t) is augmented into k video-subsentence pairs $\{(v_s^1, q_s^1), \cdots (v_s^k, q_s^k)\}$.

Teacher-student cycle consistency. We perform the teacher-student mutual learning leveraging the teacher-student cycle consistency. Denote the proposal output of the teacher network as p_t, we first apply an equivalent transformation $\mathcal{T}^k$ to echo the student's data augmentation (scaling and shifting) of the video. Since p_t represents a temporal segment (a start and an end timestamp) of the original video, this equivalent transformation is straightforward. Because of our augmentation strategy, the cycle consistency lies in that all the k student outputs p_s^k should be close to the teacher's output p_t, and inversely, the average of student outputs $Avg(p_s^k)$ should cycle back to the original teacher output. This cycle consistency enables both student learning and teacher learning. The student output can be directly supervised by the transformed teacher output:

$$L_s = \frac{\sigma(\lambda_1 u)}{k} \sum_{i=1}^{k} \left| \mathcal{T}^k(p_t) - p_s^k \right|, \tag{4}$$

where σ denotes the sigmoid function, u is the teacher's uncertainty, λ_1 is a hyper-parameter that controls the scale of the uncertainty measure, and p_s is the temporal proposal of the student.

The teacher's learning can be expressed as:

$$L_t = \sigma(\lambda_2 u) \left| \mathcal{T}^k(p_t) - \frac{1}{k} \sum_{i=1}^{k} (p_s^k) \right|, \tag{5}$$

where λ_2 is another hyper-parameter.

Enhancing self-training with reconstruction. To enhance the self-training, we additionally apply a triplet ranking loss based on masked reconstruction, as shown in the rightmost part of Figure 2. Different from previous methods that rank only between a proposal and its negative component sampled within or from other videos, we rank the reconstruction result based on the teacher's proposal, student's proposal, and a negative proposal taken by negative proposal mining from [71]. To be specific, we randomly choose one of the sentences from either the teacher input or the student input, and then randomly replace 30% of the words in the sentence query with a mask token, and predict the next word using the prefix of the sentence and the visual features within each proposal by a mask conditioned transformer [70]. Please refer to [71] and [70] for details of negative proposal mining and mask conditioned transformer. Denote the cross-entropy loss of the reconstruction by teacher proposal, student proposals, and the negative proposal as $L_{ce}(p_t), L_{ce}(p_s), L_{ce}(p_n)$, respectively, our ranking target is:

$$\begin{aligned} L_{rank} = &max(L_{ce}(p_t, q_t) - L_{ce}(p_s, q_t) + m_1, 0) \\ &+ max(L_{ce}(p_s, q_t) - L_{ce}(p_n, q_t) + m_2, 0). \end{aligned} \tag{6}$$

To learn the reconstruction, we apply cross-entropy loss using the student and teacher's proposals, without the negative proposals as [71]:

$$L_{rec} = L_{ce}(p_t, q_t) + L_{ce}(p_s, q_t) \tag{7}$$

Updating teacher and student. In our method, the student and teacher are updated asynchronously. After initialization, we first fix the teacher network and learn the student by L_{init}, L_s, L_{rec} and L_{rank}, and then fix the student network and train the teacher by L_{init}, L_t, L_{rec} and L_{rank}. Details of training can be found in Section 4.1. We use the result of the teacher network as the final output in inference.

4. Experiments

Our experiments are performed on two publicly available datasets Charades-STA [15] and ActivityNet Captions [27], following the common practice of previous works. Charades-STA is a subset of the Charades dataset [47] with sentence annotations and temporal timestamp annotations. It contains 12,408/3720 video-query pairs in the training/testing set. We report our results on the test split. ActivityNet Captions is a subset of the ActivityNet dataset [4] which contains a number of 37,417/17,505/17,031 annotated video-sentence pairs in the train/val_1/val_2 split. Following the majority of the previous works, we also report our results on the val_2 split.

As for the evaluation metric, we adopt the "IoU@n" metric under recall rate of top-1 prediction. A predicted proposal is considered correct if its Intersection over Union with the ground-truth proposal is greater than the predefined IoU threshold n. We choose $n = \{0.3, 0.5, 0.7\}$ for the Charades-STA dataset and $n = \{0.1, 0.3, 0.5\}$ on the ActivityNet Captions dataset.

4.1. Implementation detail

As for data pre-processing, we follow [71] to use C3D [51] feature for ActivityNet Captions and I3D [7] feature for Charades-STA. The features are extracted by first downsampling each video at a rate of 8. Pre-trained GloVe word2vec [44] are used to extract word embeddings. We follow [70] to set the maximum sentence length as 20, the maximum video length as 200, and the vocabulary size for the Charades-STA and ActivityNet Captions datasets as 1,111 and 8,000, respectively.

For data augmentation, we use different parameters for different datasets. For student input, on Charades-STA, when generating each augmented data, l is a random number in [100, 150], s is randomly chosen from [-25, 25], and m is set to 10. We use $k = 2$ for Charades-STA since the sentences are typically short. On the ActivityNet Captions dataset, l is fixed as 100 and s is selected randomly from [-50, 50]. We set $m = 30$ and $k = 4$. On both datasets, when the augmentation causes an index out-of-range error, we repeat the feature on the nearest timestamp. We use the parameters l, s to perform the equivalent transformation $\mathcal{T}$. For the teacher input, we only apply random frame feature masking at 10%. All data augmentation is done only in the training stage, we use the original video-sentence pair as input to the teacher network to get results on the test sets.

As for the training, we first initialize the model with L_{init} for 15 epochs for the Charades-STA dataset and 30 epochs for the ActivityNet Captions dataset. After initialization, we repeat the following step 15 times: (1) fix the teacher network and train the student network for 3 epochs, using; (2) fix the student network and train the teacher network for 1 epoch. We use Adam optimizer with learning rate set to 0.0004 for training both networks, the learning rate is decayed with an inverse square root scheduler. We set $\lambda_1 = 1, \lambda_2 = 2$ for model training on both datasets. We give L_s and L_t a weight of 10 while giving other losses a weight of 1 during training.

Our method does not introduce additional parameters to the backbone network. When applying on networks that generate multiple proposals, we only use the top-1 proposal to compute reconstruction and ranking losses. When applying to backbones that do not contain reconstruction-based loss, we simply discard the L_{rec} and L_{rank} during training.

4.2. Results and comparisons

The top block of Table 1 and 2 shows the performance of previous state-of-the-art weakly supervised temporal sentence grounding methods. Compared with our method in the last row of each table, we observe best performance is achieved on both datasets with our method.

In the bottom block of each table, we list the performance of the backbone method CPL trained with original data but is directly inferenced with augmented data (**CPL (aug)**), CPL both trained and inferenced with augmented data (**CPL + aug**). Also, we show the backbone method CPL applied with the standard teacher-student self-training method MT [50] (**CPL + MT**) with weak-strong augmentation. Note that in Table 2, we show both the results reported in the original CPL paper (**CPL (ori.)**) and the results of our replication (**CPL (rep.)**).

We note that, while the backbone network CPL [71] performs worse when it is directly inferenced on augmented data, simply training with data augmentation already results in a good performance on both datasets. This implies the success of our self-training method, since if the backbone network performs consistently on strongly augmented data, no extra knowledge can be learned from the self-training. Compared to the backbone method CPL [71], our method can consistently increase its performance on all of the metrics. This is proof that the student network learned useful knowledge from the positive guidance provided by the teacher and subsequently transferred the knowledge back to the teacher, thanks to the teacher-student cycle consistency. Importantly, we observe larger improvement in IoU at higher thresholds (IoU@0.5 and 0.7 on Charades-STA, IoU@0.3 and 0.5 on ActivityNet Captions). This is because the backbone CPL judges each proposal using reconstruction error, thus tending to produce long proposals (see Section 4.4) to ensure a good reconstruction. The largest performance gap exists on the ActivityNet Captions dataset at IoU@0.3 and 0.5. We believe this is because our sentence augmentation technique makes the sentences shorter thus reconstruction task becomes easier to accomplish, which addressed the limitation stated in [71], *i.e.*, worse performance on long sentences.

Method	IoU@0.3	IoU@0.5	IoU@0.7	mIoU
TGA [38]	32.14	19.94	8.84	-
SCN [32]	42.96	23.58	9.97	-
WSTAN [54]	43.39	29.35	12.28	-
VLANet [36]	45.24	31.83	14.17	-
MARN [48]	48.55	31.94	14.81	-
CRM [21]	53.66	34.76	16.37	-
VCA [56]	58.58	38.13	19.57	38.49
LCNet [62]	59.60	39.19	18.87	38.94
RTBPN [68]	60.04	32.26	13.24	-
CNM [70]	60.04	35.15	14.95	38.11
CPL [71]	66.40	49.24	22.39	43.48
CPL (aug)	56.46	38.47	17.64	36.78
CPL + aug	67.35	50.09	23.75	44.39
CPL + MT [50]	65.17	32.55	11.40	39.31
CPL + Ours	**69.16**	**52.18**	**23.94**	**45.20**

Table 1. IoU@{ 0.3, 0.5, 0.7} and mIoU results on the Charades-STA dataset test split. The bold numbers represent the top-1 result.

Comparing the performance of the backbone and our method with the standard self-training method MT [50], we can see that MT slightly degrades the backbone performance, while our method can increase the backbone performance. This is expected since (1) MT does not use uncertainty-guided training, resulting in the accumulation of errors, and (2) MT updates the teacher network via Exponential Moving Average (EMA), however, the student network does not take the whole sentence query as input like the teacher network, thus directly updating model weights to the teacher network performs unfavorably in our setting.

4.3. Ablation study

We conduct ablation studies to show the effectiveness of each component in our method, as well as the influence of different augmentation techniques.

Comparison on self-training methods. To confirm the effect of each proposed component of our method, we apply standard self-training methods to see their usefulness in the weakly supervised temporal sentence grounding task. We also remove different components of our method to see the effect of each component. We specifically compare with the

Method	IoU@0.1	IoU@0.3	IoU@0.5	mIoU
WS-DEC [13]	62.71	41.98	23.34	28.23
VCA [56]	67.96	50.45	31.00	33.15
MARN [48]	-	47.01	29.95	-
SCN [32]	71.48	47.23	29.22	-
RTBPN [68]	73.73	49.77	29.63	-
CTF [11]	74.2	44.3	23.6	32.2
WSLLN [16]	75.4	42.8	22.7	32.2
LCNet [62]	78.58	48.49	26.33	34.29
WSTAN [54]	79.78	52.45	30.01	-
CRM [21]	81.61	55.26	32.19	-
CNM [70]	79.74	54.61	30.26	36.59
CPL (rep.) [71]	81.14	53.99	29.38	35.55
CPL (ori.)	79.86	53.67	31.24	-
CPL (aug)	78.52	51.32	28.69	34.53
CPL + aug	**82.53**	54.90	30.19	36.87
CPL + MT [50]	79.50	52.20	28.03	34.92
CPL + Ours	82.10	**58.07**	**36.91**	**41.02**

Table 2. IoU@{0.1, 0.3, 0.5} and mIoU results on the ActivityNet Captions dataset val_2 split. The bold numbers represent the top-1 result. CPL (ori) denotes the results reported in [71], while CPL(rep) is our replicated result.

Method	IoU@0.3	IoU@0.5	IoU@0.7	mIoU
Backbone alone	66.40	49.24	22.39	43.48
Pseudo label [29]	61.59	44.93	19.85	40.24
Pseudo label + u	64.85	48.29	22.70	42.75
Mean Teacher [50]	65.17	32.55	11.40	39.31
Mean Teacher + u	66.02	39.68	14.44	40.43
w.o. u	68.62	48.89	21.88	42.88
w.o. L_{rank}, L_{rec}	68.24	48.80	22.07	43.39
Ours full	**69.16**	**52.18**	**23.94**	**45.20**

Table 3. Ablation study on the Charades-STA dataset.

following baselines on the Charades-STA dataset: **Pseudo-label** [29] is one straightforward technique for self-training, where the model iteratively refines its prediction based on the previous prediction. Since pseudo labeling often generates low-quality outputs causing error accumulation, we add another comparison where Bayesian inference is applied and the uncertainty u is utilized to weigh the pseudo labels. We denote this setting as **Pseudo label + u**. Also, we use **Mean Teacher** as a representative of standard self-training with weak-strong augmentation in which the teacher network is updated by exponential moving average (EMA), *i.e.*, L_t is not used for teacher update. To better see the effect of uncertainty guidance in self-training, we equip Mean Teacher with a Bayesian teacher network and denote this baseline as **Mean Teacher + u**. We further remove one of the other components, *i.e.*, the Bayesian inference of teacher (**w.o. u**) and the reconstruction loss (**w.o. L_{rank}, L_{rec}**), to indicate the effectiveness of each ingredient.

Results can be seen from Table 3. Not surprisingly, directly using the pseudo label technique downgrades the backbone performance due to the error accumulation. Mean Teacher's EMA-based teacher update is not suitable to our method, due to the difference in input between the teacher and student networks. Adding uncertainty by Bayesian inference to Pseudo label and Mean Teacher can mitigate the error accumulation to some extent. Our method cannot get optimal performance without the Bayesian inference of the teacher network, which proves our assumption that the student can learn better from the high-quality supervision signals of the teacher. Similarly with other weakly supervised temporal sentence grounding approaches [70, 71], we can also observe that reconstruction loss contributes to the final result. Our full method with all the uncertainty measurement, cycle consistency, and reconstruction loss performs the best, indicating that the design of all these components is of great importance for our network.

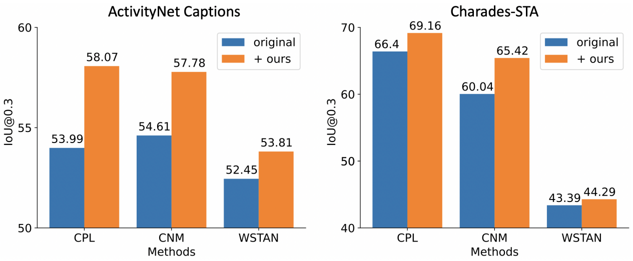


Figure 4. Results on the ActivityNet Captions (left) and Charades-STA (right) datasets when our method is applied on different backbone networks.

Augmentation			Charades-STA		ActivityNet Captions	
V	M	D	IoU@0.3	IoU@0.5	IoU@0.3	IoU@0.5
			66.40	49.24	53.99	29.38
✓			68.02	51.49	54.40	31.13
	✓		68.59	51.39	54.83	30.15
		✓	66.44	49.73	56.96	34.06
✓	✓	✓	69.16	52.18	58.07	36.91

Table 4. Results of our method when using different augmentation techniques. V: video temporal scaling and shifting; M: video temporal masking; D: decomposition of sentence queries with SRL.

Changing backbones. As explained before, our self-training method can work with multiple backbones and improve their performance. Here we show the performance of three backbone networks trained with our method. In Table 4, we demonstrate the IoU@0.3 performance of CPL [71], CNM [70] and WSTAN [54] before and after applying our method on two datasets. Our method can bring positive improvement to all the three backbones, indicating the generalizability of our proposed method.

Discussion on different augmentation techniques. In our method, we use multiple augmentation techniques during the teacher-student mutual learning. We show the effect of each augmentation in Table 4. Here V stands for video temporal scaling and shifting, M denotes video temporal mask-

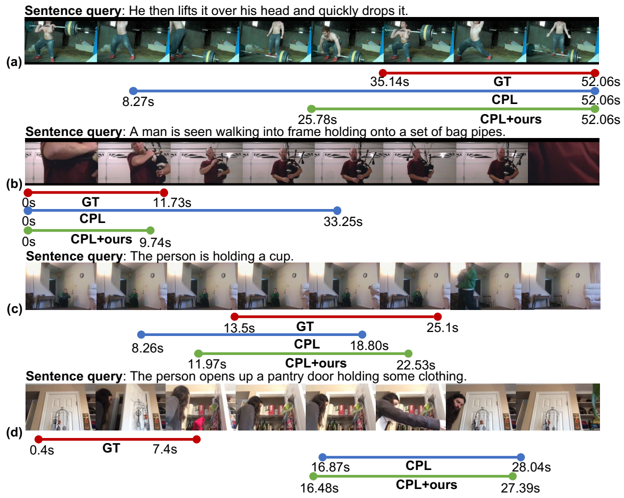

Sentence query: He then lifts it over his head and quickly drops it.
(a)
35.14s
GT
52.06s
8.27s
CPL
52.06s
25.78s
CPL+ours
52.06s
Sentence query: A man is seen walking into frame holding onto a set of bag pipes.
(b)
0s
GT
11.73s
0s
CPL
33.25s
0s
CPL+ours
9.74s
Sentence query: The person is holding a cup.
(c)
13.5s
GT
25.1s
8.26s
CPL
18.80s
11.97s
CPL+ours
22.53s
Sentence query: The person opens up a pantry door holding some clothing.
(d)
0.4s
GT
7.4s
16.87s
CPL
28.04s
16.48s
CPL+ours
27.39s

Figure 5. Qualitative examples of the ground truth (GT), the backbone network (CPL), and the backbone method with our mutual learning (CPL+ours). Examples (a, b) are from the ActivityNet Captions dataset, and (c, d) are from the Charades-STA dataset.

ing, and D denotes the decomposition of sentence queries with SRL. To better show the performance gap, we show the original backbone in the first row of Table 4 with gray background. We can see from the table that different augmentation techniques have different influences on each of the datasets: the augmentation on videos V and M are more effective on the Charades-STA dataset, while the augmentation of sentence decomposition works better on the ActivityNet Captions dataset. We think this is mainly because of the difference in the length of sentences. In the Charades-STA dataset, the sentences are mostly short with an average of 6.2 words per sentence, while in the ActivityNet Captions dataset the average number of words per sentence is 13.5. Thus, the effect of decomposition by semantic role labeling is more significant in the sentences of the ActivityNet Captions dataset.

4.4. Qualitative results

We show several qualitative examples in Figure 5. From this figure we can obtain several interesting observations: (1) As shown in Figure 5(a) (b) and (c), our method can achieve better results than the backbone CPL, proving that our self-training technique can positively provide extra guidance to the network. (2) As shown in Figure 5(a) (b) and (c), the backbone method CPL tends to output longer proposals, while our method can effectively reduce the length of the proposals to a reasonable range. This is mainly because CPL purely relies on reconstruction results as an indicator of the quality of each proposal, thus the proposals tend to be long in order to guarantee a successful reconstruction. (3) Fig. 5(d) shows that when the performance of CPL is too off, it is also hard for our method to refine this result. This reveals one limitation of our work, *i.e.*, relies on the performance of the backbone network.

Method	Recall@1	Recall@5		
	IoU@0.3	IoU@0.3	IoU@0.5	IoU@0.7
VLANet [36]	45.24	95.70	82.85	33.09
VCA [56]	58.58	**98.08**	78.75	37.75
LCNet [62]	59.60	94.78	80.56	45.24
RTBPN [68]	60.04	97.48	71.85	41.48
CPL [71]	66.40	96.99	84.71	52.37
CPL + aug	67.35	97.37	**85.40**	**52.74**
CPL + Ours	**69.16**	96.96	84.86	52.58

Table 5. IoU of different methods at Recall@5 on the Charades-STA dataset. Recall@1, IoU@0.3 is shown for reference.

4.5. Limitation and future work

As discussed in the previous section, although our method can be applied to multiple backbone networks and improve their performance, one limitation is that the performance relies on the backbone network. When the backbone network does not generate reliable results, our method can only marginally improve the overall performance.

Most methods generate multiple proposals for each sentence. However, we found that while our method increases the IoU performance of the top-1 proposal, our method can only marginally improve the results under recall rate of top-5 prediction (Recall@5). In Table 5 we show the Recall@5 performance on the Charades-STA dataset. We can see that the backbone network CPL directly trained with data augmentation can achieve consistently higher performance on Recall@5 compared to our method. We think the reason is that although our self-training method can generate more accurate top-1 proposals, it simultaneously harms the diversity of the output proposals. We leave the goal of increasing Recall@5 for future work.

5. Conclusion

In this work, we propose the first self-training-based method for weakly supervised temporal sentence grounding. Our self-training framework includes a pair of mutually learned teacher and student networks. We give weak-strong augmentation to the teacher-student networks and learn the networks by teacher-student cycle consistency loss and reconstruction-based losses. Experiments on public datasets demonstrate the outstanding performance of our method. We also show in ablation studies the effectiveness of the components in our method and our method's capability of working with different backbone networks.

Acknowledgement This work is supported by JSPS KAKENHI Grant Number JP22KF0119 and JP20H04205, JST AIP Acceleration Research Grant Number JPMJCR20U1, as well as JST SPRING, Grant Number JPMJSP2108.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017. 2

[2] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Number 4. 2006. 3

[3] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017. 4

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 5

[5] Minjie Cai, Feng Lu, and Yoichi Sato. Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In *CVPR*, pages 14392–14401, 2020. 3

[6] Tianyue Cao, Lianyu Du, Xiaoyun Zhang, Siheng Chen, Ya Zhang, and Yan-Feng Wang. Cat: Weakly supervised object detection with category transfer. In *ICCV*, pages 3070–3079, 2021. 3

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 5

[8] Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama. Commonsense knowledge aware concept selection for diverse and informative visual storytelling. In *AAAI*, pages 999–1008, 2021. 1

[9] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *ECCV*, pages 192–208, 2022. 3

[10] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *CVPR*, pages 8425–8435, 2021. 2

[11] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308*, 2020. 7

[12] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021. 2

[13] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *NeurIPS*, 2018. 7

[14] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016. 4

[15] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017. 2, 5

[16] Mingfei Gao, Larry Davis, Richard Socher, and Caiming Xiong. WSLLN:weakly supervised natural language localization networks. In *EMNLP-IJCNLP*, 2019. 1, 2, 7

[17] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *WACV*, pages 245–253. IEEE, 2019. 2

[18] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288, 2002. 4

[19] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*, pages 8393–8400, 2019. 2

[20] GE Hinton and Drew van Camp. Keeping neural networks simple by minimising the description length of weights. In *Proceedings of COLT-93*, pages 5–13, 1993. 3

[21] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *ICCV*, pages 7199–7208, 2021. 1, 2, 6, 7

[22] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *ECCV*, pages 754–769, 2018. 1

[23] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *CVPR*, pages 14024–14034, 2020. 1

[24] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *CVPR*, pages 1377–1385, 2017. 3

[25] Xue-Bo Jin, Wen-Tao Gong, Jian-Lei Kong, Yu-Ting Bai, and Ting-Li Su. A variational bayesian deep network with data self-screening layer for massive time-series data forecasting. *Entropy*, 24(3):335, 2022. 3

[26] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *NeurIPS*, 30, 2017. 4

[27] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 5

[28] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, pages 156–165, 2017. 1

[29] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML*, 2013. 1, 2, 7

[30] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019. 1

[31] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *CVPR*, pages 7581–7590, 2022. 2, 3

[32] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, 2020. 2, 3, 6, 7

[33] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*, pages 11235–11244, 2021. 2

[34] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *ICLR*, 2021. 2, 3

[35] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In *CVPR*, pages 1726–1736, 2021. 3

[36] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *ECCV*, pages 156–171, 2020. 1, 2, 6, 8

[37] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised hierarchical text classification. In *AAAI*, pages 6826–6833, 2019. 3

[38] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, pages 11592–11601, 2019. 1, 2, 6

[39] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, pages 10810–10819, 2020. 2

[40] Radford M Neal. *Bayesian learning for neural networks*, volume 118. 2012. 3

[41] Takehiko Ohkawa, Yu-Jhe Li, Qichen Fu, Rosuke Furuta, Kris M Kitani, and Yoichi Sato. Domain adaptive hand keypoint and pixel localization in the wild. *ECCV*, 2022. 3

[42] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. *BMVC*, 2020. 2

[43] Sudipta Paul, Niluthpol Chowdhury Mithun, and Amit K Roy-Chowdhury. Text-based temporal localization of novel events. In *ECCV*, pages 567–587, 2022. 2

[44] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 5

[45] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *ECCV*, pages 347–363, 2018. 1

[46] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *ICCV*, pages 5208–5217, 2019. 3

[47] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526, 2016. 5

[48] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*, 2020. 3, 6, 7

[49] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *WACV*, pages 2083–2092, 2021. 2

[50] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 2017. 1, 2, 6, 7

[51] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 5

[52] Alessandro Vinciarelli, Anna Esposito, Elisabeth André, Francesca Bonin, Mohamed Chetouani, Jeffrey F Cohn, Marco Cristani, Ferdinand Fuhrmann, Elmer Gilmartin, Zakia Hammal, et al. Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive Computation*, 7:397–413, 2015. 1

[53] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *CVPR*, pages 14090–14100, 2021. 3

[54] Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li. Weakly supervised temporal adjacent network for language grounding. *IEEE TMM*, 24:3276–3286, 2022. 6, 7

[55] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, pages 12275–12284, 2020. 3

[56] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *ACM MM*, pages 1459–1468, 2021. 6, 7, 8

[57] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE TPAMI*, 39(11):2314–2320, 2016. 3

[58] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In *ACM MM*, page 1283–1291, 2020. 2

[59] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14722–14732, 2022. 1

[60] Lijin Yang, Quan Kong, Hsuan-Kung Yang, Wadim Kehl, Yoichi Sato, and Norimasa Kobori. Deco : Decomposition and reconstruction for compositional temporal grounding via coarse-to-fine contrastive ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[61] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In *CVPR*, pages 53–63, 2021. 3

[62] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE TIP*, 30:3252–3262, 2021. 2, 6, 7, 8

[63] Thorsten O Zander, Matti Gaertner, Christian Kothe, and Roman Vilimek. Combining eye gaze input with a brain–computer interface for touchless human–computer inter-

action. *Intl. Journal of Human–Computer Interaction*, 27(1):38–51, 2010. 1

[64] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, pages 1247–1257, 2019. 2

[65] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE TPAMI*, 44(9):5866–5885, 2021. 3

[66] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV*, pages 766–782, 2016. 1

[67] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *ECCV*, pages 597–613, 2018. 3

[68] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *ACM MM*, pages 4098–4106, 2020. 2, 6, 7, 8

[69] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *NeurIPS*, pages 18123–18134, 2020. 2

[70] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *AAAI*, page 3, 2022. 1, 2, 3, 5, 6, 7

[71] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *CVPR*, pages 15555–15564, 2022. 1, 2, 3, 5, 6, 7, 8

[72] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009. 2