Weakly Supervised Temporal Sentence Grounding with Gaussian-based Contrastive Proposal Learning

Minghang Zheng[1] Yanjie Huang[1] Qingchao Chen[2] Yuxin Peng[1] Yang Liu[1,3*]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]National Institute of Health Data Science, Peking University
[3]Beijing Institute for General Artificial Intelligence
{minghang, qingchao.chen, pengyuxin, yangliu}@pku.edu.cn

Abstract

Temporal sentence grounding aims to detect the most salient moment corresponding to the natural language query from untrimmed videos. As labeling the temporal boundaries is labor-intensive and subjective, the weakly-supervised methods have recently received increasing attention. Most of the existing weakly-supervised methods generate the proposals by sliding windows, which are content-independent and of low quality. Moreover, they train their model to distinguish positive visual-language pairs from negative ones randomly collected from other videos, ignoring the highly confusing video segments within the same video. In this paper, we propose Contrastive Proposal Learning(CPL) to overcome the above limitations. Specifically, we use multiple learnable Gaussian functions to generate both positive and negative proposals within the same video that can characterize the multiple events in a long video. Then, we propose a controllable easy to hard negative proposal mining strategy to collect negative samples within the same video, which can ease the model optimization and enables CPL to distinguish highly confusing scenes. The experiments show that our method achieves state-of-the-art performance on Charades-STA and ActivityNet Captions datasets. The code and models are available at https://github.com/minghangz/cpl.

1. Introduction

Temporal sentence grounding aims at localizing the start and end time of the moment described by a given free-form natural language query in untrimmed videos. Automatic temporal sentence grounding enables us to efficiently find the video moment of interest rather than going through the whole video, which has broad application potential in video surveillance [6], video summarization [20], *etc*. Fully supervised temporal sentence grounding has witnessed tremendous achievements recently, however, it needs laborious manual annotations of temporal boundaries for every query thus limiting its scalability and practicability in real-world applications. Therefore, the weakly supervised learning schemes, where only the video and natural language query are required during training, have gained more attention due to their low annotation cost and reasonable efficiency.

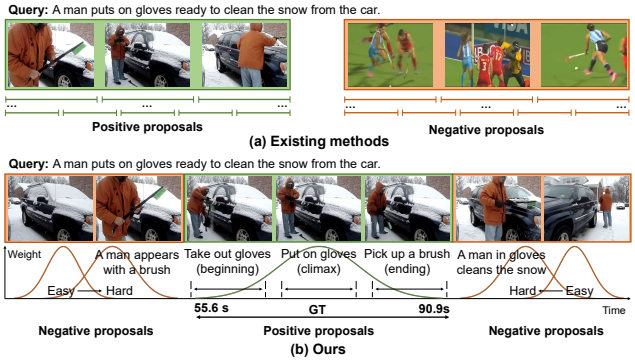

Figure 1. (a) Existing methods generate proposals by sliding window and focus on distinguishing negative proposals from other videos. (b) We use multiple learnable Gaussian functions to generate both positive and negative proposals to characterize the temporal structure of events. Our negative proposals are in the same video and collected from easy to hard.

Existing weakly supervised solutions employ either the multiple instance learning (MIL) based or reconstruction-based paradigms. Specifically, MIL-based methods [11, 12, 19, 21] normally define matched and mismatched video-language pairs as positive and negative samples, and learn the latent cross-modal semantic space by aligning the video-level visual-textual relationships. Reconstruction-based method [18, 24, 39] solves the task through joint learning with the reconstruction loss, assuming that the proposals

*Corresponding author

that match the text should best reconstruct the entire query. However, both paradigms have the following limitations:

Firstly, most *existing methods generate the same proposals for all samples via sliding window (shown in Fig. 1(a)), regardless of their contents and difficulty, which is inefficient and of low quality.* CNM [39] proposes to use single learnable Gaussian mask as the positive proposal which can characterize the inherent temporal structure of an event. However, an untrimmed long video usually comprises several events and these events often contain similar characters and backgrounds. This makes the model easy to optimize on some sub-optimal solutions when only predicting one positive proposal, resulting in a reduction in the recall rate. CNM [39] directly uses one minus the positive Gaussian mask as negative mask, which is unrealistic to describe the temporal structure of negative events and is easily distinguished by the model. Secondly, most *existing methods heavily depend on the quality of randomly selected negative samples (other unpaired videos), as shown in Fig. 1(a), which are often easy to distinguish and cannot provide strong supervision signals.* However, what the model needs for temporal sentence grounding is to distinguish the highly confusing video segments within the same video (*e.g.* a man in gloves and a man puts on gloves as shown in Fig. 1(b)). However, directly using the video segments outside the positive proposals as negative proposals will harm the model training during early training stage due to some misidentified negative proposals.

To address the above limitations, we introduce a novel weakly supervised method namely Contrastive Proposal Learning(CPL), by generating multiple content-dependent proposals and mining negative samples from easy to hard within the same video. On the one hand, *to characterize the multiple events in a long video, we use multiple learnable Gaussian functions to generate both positive (green curve in Fig. 1(b)) and negative (orange curves in Fig. 1(b)) proposals*, where the negative proposals should be out of the positive ones and do not cover the corresponding positive proposals [1]. Moreover, to distinguish the positive from negative proposals in each video, we introduce the entire video as a reference point, as it contains both the ground truth and a large amount of redundant information. We require that the semantic alignment between the positive proposal and the query is expected to be higher than that of the entire video, while the semantic alignment of negative proposals should be lower. On the other hand, in contrast to learning from randomly selected negative samples from unpaired videos, we *propose a controllable easy-to-hard negative proposal mining strategy.* We collect negative proposals within the same video and enforce the negative proposals further away from the positive ones at the early training stages while proposals closer to the positive are learned at later stages. Because we observe that the negative proposals closer to the positive proposal are often harder to distinguish than those further away, mainly due to the smooth transition of events (similar background and semantics as shown in Fig. 1(b)). This dynamic curriculum strategy to mine samples can gradually reduce the ambiguity and thus facilitate to learn reliable intra-video samples and ease the model optimization.

Our contributions are summarized as follows: (1) We propose to use multiple Gaussian functions to generate both positive and negative proposals from the same video. By introducing the entire video as a reference point, our proposal generation is content-dependent and efficient. (2) We propose a controllable Easy to Hard Negative sample mining strategy to collect negative proposals within the video and ease the model optimization. This enables our network to distinguish highly confusing scenes. (3) Experiments on Charades-STA [10] and ActivityNet Captions datasets [2, 15] demonstrate our method significantly outperforms existing weakly supervised methods.

2. Related Work

Fully Supervised Temporal Sentence Grounding. In the fully supervised setting, the precise start and end timestamps annotations for each video and query pair are required in training [10, 22, 30, 31, 38, 40]. Specifically, TALL [10] makes the first attempt to integrate the query and video feature to predict the start and end timestamps directly. The Structured Multi-Level Interaction Network (SMIN) [31] constructs a structured multi-level interaction module to optimize the use of the logic relationship between query and video segment. However, the fully supervised methods need laborious manual annotation of temporal boundaries thus limiting its scalability and practicability. Moreover, as studied in [22], the temporal boundary annotation is sometimes subjective and, possibly, not consistent across different annotators. And such issues are not adequately considered by many of the existing approaches.

Weakly Supervised Temporal Sentence Grounding. Different from the fully supervised setting, the start and end timestamps are inaccessible during training for weakly supervised temporal sentence grounding.

Firstly, the methods proposed in [5, 12, 18, 21, 26, 32] use sliding windows to generate proposals. The proposals are content-independent and heavily rely on prior knowledge to the length distribution of ground truth for the specific datasets and bring much extra computational cost for pre-processing. [13] proposes to use a learnable network to generate gate-shape masks as proposals for the action localisation task. However, the gate-shape mask assumes that all frames in the proposal are equally important, which is not

[1] We plot a single positive proposal in Fig. 1(b) as an example, practically, we generate multiple positive proposals via Gaussian functions to improve the recall rate.

optimal for sentence localization, as the events described in free-form natural language queries have more complex temporal structure. CNM [39] proposes to use single learnable Gaussian mask as the positive proposal and use one minus the Gaussian mask as negative proposal. However, an untrimmed long video usually comprises several events with similar characters and backgrounds, which makes predicting only single proposal easy to optimize on a sub-optimal solution. In addition, such method does not reflect the temporal structure of negative proposals, which is easy for the model to distinguish them from positive ones. In our method, we utilize multiple Gaussian functions to generate both positive and negative content-dependent proposals with learnable parameters efficiently, and design a diversity loss to require these Gaussian functions to focus on different events in the video. These proposals will jointly participate in the inference process, from which the most relevant proposal is selected.

Secondly, some methods like the weakly supervised Semantic Completion Network (SCN) [18] assume that a video segment paired with the query could reconstruct the sentence better. However, they do not consider the information contained in unpaired videos and queries, which could be used for contrastive learning. In our method, we also use the reconstruction mechanism to measure the semantic alignment, but utilize the negative proposals to perform contrastive learning. Further, other works [12, 21, 35] train the model to distinguish aligned video-query pairs and non-aligned ones collected from other videos. However, their video-query samples are often easy to distinguish, ignoring what the model really needs to distinguish is the highly confusing segments in the same video. Moreover, RTBPN [36] and CNM [39] consider the confrontment of the intra and inter samples made up of video-query pairs. However, if a mistake is made in the early stage of training, as training continues, the correct video segments would be suppressed, which will harm the training of the model. In our algorithm, we collect negative proposals outside the positive proposal, and introduce the whole video as a reference. Our negative proposals are more difficult to distinguish, enabling our network to distinguish highly confusing scenes. Additionally, in the early stage of training, our negative proposals are far from the positive ones, which reduces the negative impact of introducing negative proposals when the accuracy of positive proposal is not high in the early training stage.

Curriculum learning. The curriculum learning method emulates the learning behavior of humans [1, 25]. It has many applications (*e.g.* image classification [27], object detection [16], *et al.*). As far as we know, we make the first attempt to explore a curriculum-style easy to hard negative proposals mining strategy and verify its effectiveness in temporal sentence grounding.

3. Proposed Method

3.1. Overall Framework

CPL comprises the proposal generation module and the mask conditioned reconstruction module (In Fig. 2).

For the proposal generation module, we use Gaussian masks to represent both positive and negative proposals within the same video. The frame features within each proposal will be aggregated based on the weight in the Gaussian curve, which characterizes the inherent temporal structure of events. To distinguish positive and negative proposals, we introduce the entire video as a reference, requiring the semantic similarity between the positive proposals and the query should be higher than that of the reference and the semantic similarity of the negative proposals should be lower. To ease the model optimization, we mine the negative proposals in the same video, collecting them from easy to hard (gradually close to the positive proposal but never overlap, as shown in Fig. 3). Note that we generate multiple positive proposals and try to make them diverse via a diversity loss $\mathcal{L}_{div}$ to improve the recall rate.

For the mask conditioned reconstruction module, we use the video frames in each proposal to reconstruct the original query by a transformer to measure the proposal-query alignment, assuming that the better-aligned proposals can reconstruct the query better. Since the negative proposals do not contain any frames in the ground truth, we should not require them to reconstruct the query. Therefore, the reconstruction loss $\mathcal{L}_{rec}$ consists only of the cross-entropy loss of the positive proposal and the reference. Finally, we introduce an intra-video contrastive loss $\mathcal{L}_{IVC}$ to ensure that the reconstruction results of positive proposals are better than that of the entire video, while the reconstruction results of negative proposals should be worse.

3.2. Multiple Positive Proposals Generation

In this module, we fuse the information of video and text to generate multiple positive proposals, which depend on the content of video and query. Following CNM [39], we use the Gaussian masks as the proposals, which can characterize the inherent temporal structure of events. Unlike CNM [39] which only generates single positive proposal, we predict multiple positive proposals at the same time and encourage these proposals to be different through a diversity loss. Since a long untrimmed video usually contains multiple events, our method can efficiently find the potential event of interest and improve the recall rate.

Feature Extraction. Given an untrimmed video and a natural language query, we firstly encode them into feature vectors. To be specific, each word of the query is embedded using GloVe [23] and the query is represented as $T = \{t_1, t_2, ..., t_M\} \in \mathbb{R}^{M \times D_T}$, where M is the number of words and D_T is the word feature dimension. The video is

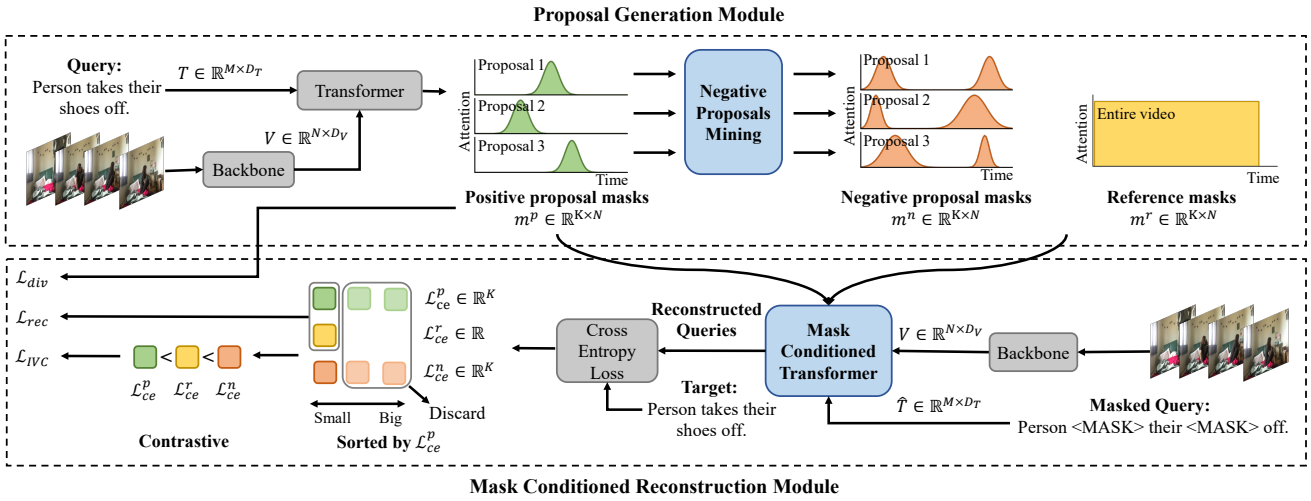

Figure 2. The framework of our method. The proposal generation module uses Gaussian masks to represent both positive and negative proposals within the same video. To distinguish the positive and negative proposals, we introduce the entire video for each sample as a reference. Note that we generate multiple positive proposals and try to make them diverse via $\mathcal{L}_{div}$ to improve the recall rate. The mask conditioned reconstruction module uses the frame features conditioned on the proposal masks to reconstruct the query, as a measurement of the similarity between the proposal and query. $\mathcal{L}_{rec}$ is used to optimize our networks for a better reconstruction, and $\mathcal{L}_{IVC}$ performs a contrast between proposals, requiring the semantic similarity between the proposals and the query from large to small are positives, reference, and negatives.

encoded with pre-trained 3D convolutional network [3, 28], and represented as $V = \{v_1, v_2, ..., v_N\} \in \mathbb{R}^{N \times D_V}$, where N is the number of extracted video features and D_V is the feature dimension.

Proposal Generation. We use transformer [29] to handle the multi-modal interaction of the video and text. Firstly, to pool the frame features and obtain the video representation, we append an additional learnable [CLASS] token v_{cls} [7] at the end of the video features: $\hat{V} = \{v_1, v_2, ..., v_N, v_{\text{cls}}\}$. We use transformer to conduct cross-modal interaction between the embedded texts T and video features $\hat{V}$ and obtain the hidden features $H = \{h_1, h_2, ..., h_N, h_{\text{cls}}\}$ that incorporates semantic and vision information: $H = \text{D}(\hat{V}, \text{E}(T)) \in \mathbb{R}^{N \times D_H}$, where $\text{E}(\cdot)$ is the transformer encoder, $\text{D}(\cdot)$ is the transformer decoder, and D_H is the dimension of the hidden features. More details about $\text{E}(\cdot)$ and $\text{D}(\cdot)$ will be provided in Sec. 4. To ensure the end-to-end training of our model, we adopt Gaussian functions as proposals. As h_{cls} combines all the frame and word features, we predict the center $c^p \in \mathbb{R}^K$ and width $w^p \in \mathbb{R}^K$ of our positive proposals through h_{cls} with a fully connected layer activated by Sigmoid function. To improve the recall rate, we will predict K Gaussian masks $m^{\text{p}} \in \mathbb{R}^{K \times N}$ as potential positive proposal candidates:

$$m^p_{ki} = \frac{1}{\sqrt{2\pi}(w^p_k/\sigma)} \exp(-\frac{(i/N - c^p_k)^2}{2(w^p_k/\sigma)^2}), \quad k = 1, ..., K; \quad i = 1, ..., N \tag{1}$$

where c^p_k, w^p_k are the center and width of the k-th positive proposal which are learnable, and σ is a hyperparameter which controls the width of the Gaussian curve.

In order to make the K proposals as different as possible, we apply a diversity loss $\mathcal{L}_{div}$ introduced in [17] to m^p:

$$\mathcal{L}_{div} = ||m^p m^{p\top} - \lambda I||^2_F \tag{2}$$

where $||\cdot||_F$ denotes Frobenius norm of a matrix, and $\lambda \in [0, 1]$ is a hyperparameter which controls the extent of overlap between proposals. The loss encourages proposals to have less overlap, prevents them from converging to the same center and width, and improves the recall rate.

3.3. Negative Proposal Mining

Unlike CNM [39] which directly uses one minus the Gaussian mask of positive sample as negative samples, we point out that negative proposals should have the same temporal structure of events as positive proposals, but are not so semantically relevant to the query. Thus we also use Gaussian functions to represent the them. Inspired by curriculum learning, we collect negative proposals from easy to hard to ease the optimization. We observe when the negative proposals are close to the positive ones, they are more confusing due to the similar background and semantics. Thus, we enforce the negative proposals further away from the positive ones at the early training stages while proposals closer to the positive are learned at later stages. Finally we use the mask conditioned reconstruction module to measure the

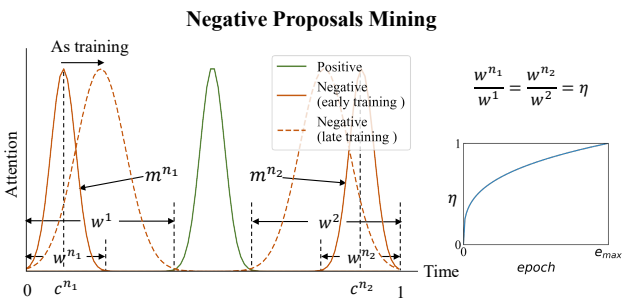

Figure 3. We mine the difficulty of negative proposals, assuming that those negative proposals close to the positive proposal are more difficult to distinguish. We learn them from easy to hard to ease the model optimization.

semantic similarity between proposals and query, and contrast between the positive and negative proposals to train our model.

Negative Proposal Mining. As shown in Fig. 3, we generate two negative proposals (before and after the positive proposal). We fix one end of the two negative proposals at the boundary of the video, and represent the distance η between the negative proposals and the positive proposal by the ratio of negative proposal's width w^{n_1}, w^{n_2} to the width of the corresponding side of the positive proposal w^1, w^2:

$$w^1 = c^p - \frac{w^p}{2}, w^2 = 1 - c^p - \frac{w^p}{2} \tag{3}$$

$$\frac{w^{n_1}}{w^1} = \frac{w^{n_2}}{w^2} = \eta \tag{4}$$

As the training progresses, η will gradually increase to make the negative proposals approach the positive proposal:

$$\eta = (\frac{e}{e_{max}})^{0.5} \in [0, 1] \tag{5}$$

where e is the current training epoch and e_{max} is the total number of training epochs. Since one end of the negative proposal is fixed, the center c^{n_1}, c^{n_2} of negative proposals can be calculated by: $c^{n_1} = \frac{w^{n_1}}{2}, c^{n_2} = 1 - \frac{w^{n_2}}{2}$. Then, similar to Eq.(1), we can get the Gaussian mask of the negative proposals m^{n_1}, m^{n_2} respectively.

To help the model distinguish the positive and negative proposals, we introduce the entire video as a per sample reference m^r:

$$m^r = [1, 1, ..., 1] \in \mathbb{R}^N \tag{6}$$

Since the entire video contains the ground truth segment as well as lots of redundant information, the semantic similarity of the proposals and the query should satisfy:

$$\mathrm{R}(m^p, Q) > \mathrm{R}(m^r, Q) > \mathrm{R}(m^n, Q) \tag{7}$$

where $\mathrm{R}(\cdot)$ is a function that measures the semantic similarity between the query Q and the proposal specified by the mask m.

Mask Conditioned Semantic Completion. To contrast between the positive and negative proposals, we use the mask conditioned reconstruction completion module inspired by SCN [18] and CNM [39] to measure the semantic relevance between the proposal and the query, assuming that the most relevant proposal should be able to best reconstruct the query using only the visual features in the proposal.

We randomly mask 1/3 of the words in the original query with a specific symbol, and require the model to predict the next word given a prefix of the query and visual features within the proposal. We embed the masked query using GloVe [23] and reconstruct the original query based on the visual features within each positive and negative proposal using the mask conditioned transformer proposed in CNM [39]. The mask conditioned transformer will multiply the mask by the attention map before aggregating contextual information to prevent the leakage of vision features outside the proposal and keep the entire module differentiable to the proposal. Finally, we use the cross-entropy loss to measure the similarity of the reconstructed query and the original query. We denote the cross-entropy loss of the positive proposals, negative proposals, and the reference as $\mathcal{L}^p_{ce}, \mathcal{L}^{n1}_{ce}, \mathcal{L}^{n2}_{ce} \in \mathbb{R}^K$, and $\mathcal{L}^r_{ce} \in \mathbb{R}$ respectively.

Although we predicted K positive proposals, only one video segment corresponds to the query. Therefore, we only keep the k^*-th positive proposal with the smallest loss $\mathcal{L}^p_{ce}$, because it is most semantically related to the query:

$$k^* = \arg\min_k(\mathcal{L}^p_{ce}[k]) \tag{8}$$

As only the positive proposal and the reference contain the segment related to the query, only $\mathcal{L}^p_{ce}[k^*]$ and $\mathcal{L}^r_{ce}$ will participate in the optimization of our reconstruction network. The final reconstruction loss $\mathcal{L}_{rec}$ is formulated as:

$$\mathcal{L}_{rec} = \mathcal{L}^p_{ce}[k^*] + \mathcal{L}^r_{ce} \tag{9}$$

As shown in (7), the semantic similarity between the positive proposal, negative proposal, and the reference should satisfy a certain relationship. Following CNM [39], we use the Intra-video Contrastive loss $\mathcal{L}_{IVC}$ to contrast between positive and negative proposals:

$$\begin{aligned}\mathcal{L}_{IVC} =& \max(\mathcal{L}^p_{ce}[k^*] - \mathcal{L}^r_{ce} + \beta_1, 0) + \\ & \max(\mathcal{L}^p_{ce}[k^*] - \mathcal{L}^{n_1}_{ce}[k^*] + \beta_2, 0) + \\ & \max(\mathcal{L}^p_{ce}[k^*] - \mathcal{L}^{n_2}_{ce}[k^*] + \beta_2, 0)\end{aligned} \tag{10}$$

where β_1 and β_2 are hyperparameters satisfying $\beta_1 < \beta_2$. $\mathcal{L}_{IVC}$ requires the loss $\mathcal{L}_{ce}$ of the positive proposal should be at least β_1 smaller than that of the reference, and at least β_2 smaller than that of the negative proposals.

3.4. Model Training and Inference

In this section, we describe the loss function we optimize to train our network and the model inference process.

Training. Our network includes three parts of loss: the reconstruction loss $\mathcal{L}_{rec}$ in Eq.(9) is used to help the model to reconstruct query through the video features within our proposals, serving as a measurement of the alignment between the proposal and query; the Intra-video Contrastive loss $\mathcal{L}_{IVC}$ in Eq.(10) is used to train the model to generate the most semantically relevant positive proposals for the query; the diversity loss $\mathcal{L}_{div}$ in Eq.(2) is used to encourage the model to produce multiple different positive proposals. Finally, we compute a multi-task loss to train our network in an end-to-end manner, denoted by:

$$\mathcal{L} = \mathcal{L}_{rec} + \alpha_1 \mathcal{L}_{IVC} + \alpha_2 \mathcal{L}_{div} \tag{11}$$

where α_1, α_2 are hyperparameters to balance the losses.

Inference. Firstly, we can obtain the center c^p and width w^p in Eq.(1) of our predicted K positive proposals. To select the top-1 prediction from our K proposals, we design two selection strategies: loss-based strategy and vote-based strategy.

For the loss-based strategy, the cross-entropy loss $\mathcal{L}^p_{ce}$ of reconstructed query serves as the measurement of the reliability of each proposal. Thus, we select the positive proposal with the smallest loss as our final prediction. For the vote-based strategy, every positive proposal will participate in the selection. Inspired by the ensemble learning [41], we use the K positive proposals to vote with each other to decide which one is our final top-1 prediction. Specifically, for each positive proposal, we calculate the IoU with the remaining $K-1$ positive proposals and the sum of IoUs is the number of votes it obtained. And finally we choose the one with the highest number of votes as the final prediction.

Finally, for the selected k^*-th positive proposal, our predicted start st and end en timestamps are:

$$\begin{aligned} st &= \max(c^p_{k^*} - w^p_{k^*}/2, 0) * \text{Duration} \\ en &= \min(c^p_{k^*} + w^p_{k^*}/2, 1) * \text{Duration} \end{aligned} \tag{12}$$

To obtain the top-k predictions, we sort all the positive proposals by $\mathcal{L}^p_{ce}$ from small to large, and output the start and end timestamps by Eq.(12) for the top-k positive proposals.

4. Experiments

4.1. Datasets

In order to evaluate the effectiveness of our method, we perform experiments on two publicly available datasets: Charades-STA [10], and ActivityNet Captions [2, 15].

Charades-STA. Charades-STA dataset contains 5338/1334 videos and 12,408/3720 video-query pairs for training/testing. We report our results on the test split.

Table 1. Evaluation Results on the Charades-STA dataset ($n \in \{1, 5\}$ and $m \in \{0.3, 0.5, 0.7\}$). The numbers in bold are the best result, and the underlined ones are the second best result. Our CPL uses the loss-based strategy during inference and CPL* uses the vote-based strategy during inference.

Method	R@1			R@5		
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
TGA [21]	32.14	19.94	8.84	86.58	65.52	33.51
CTF [5]	39.8	27.3	12.9	-	-	-
SCN [18]	42.96	23.58	9.97	95.56	71.80	38.87
WSTAN [32]	43.39	29.35	12.28	93.04	76.13	41.53
BAR [34]	44.97	27.04	12.23	-	-	-
VLANet [19]	45.24	31.83	14.17	95.70	82.85	33.09
LoGAN [26]	48.04	31.74	13.71	89.01	72.17	37.58
MARN [24]	48.55	31.94	14.81	90.70	70.00	37.40
WSRA [9]	50.13	31.20	11.01	86.75	70.50	39.02
CCL [37]	-	33.21	15.68	-	73.50	41.87
CRM [12]	53.66	34.76	16.37	-	-	-
VCA [33]	58.58	38.13	19.57	**98.08**	78.75	37.75
LCNet [35]	59.60	39.19	18.87	94.78	80.56	45.24
RTBPN [36]	60.04	32.36	13.24	97.48	71.85	41.18
CNM [39]	60.39	35.43	15.45	-	-	-
CPL (ours)	**66.40**	**49.24**	22.39	96.99	**84.71**	**52.37**
CPL* (ours)	65.99	49.05	**22.61**	96.99	**84.71**	**52.37**

ActivityNet Captions. ActivityNet Captions contains 10,009/4917/5044 videos and 37,417/17,505/17,031 video-query pairs for training/validation/testing. We report our results on the val_2 split.

4.2. Evaluation Metric

In order to test our method, similar to what have done in the previous work [12, 18], we choose the computation result of 'R@n, IoU=m' as our evaluation metric, where m is the predefined temporal Intersection over Union (IoU) threshold, and n refers to the recall rate of top-n predictions. In particular, this metric means that the percentage of predicted moments that have the IoU value larger than m in our top n predictions. We report results for R@1 and R@5 on both Charades-STA and ActivityNet Captions datasets.

4.3. Implementation Details

Data Preprocessing. We downsample each video every 8 frames and pre-extract it's vision feature using the C3D [28] model for ActivityNet Captions and I3D [3] model for Charades-STA. We use the pre-trained GloVe [23] word2vec for each word token to extract word embeddings. We set the maximum description length to 20, and the vocabulary size is 8000.

Model Settings. For the transformer and the mask conditioned transformer, there are 3 layers with 4 attention heads for both the encoder and decoder. The dimension of their hidden state is 256. For the number of positive proposals, we set K to 8 for Charades-STA and 5 for ActivityNet Captions. For the hyperparameters, we set $\sigma = 9$, $\lambda = 0.15$, $\beta_1 = 0.1$, $\beta_2 = 0.15$, $\alpha_1 = 1$ for both datasets. We find

Table 2. Evaluation Results on the ActivityNet Captions dataset ($n \in \{1, 5\}$ and $m \in \{0.1, 0.3, 0.5\}$). The numbers in bold are the best result, and the underlined ones are the second best result. Our CPL uses the loss-based strategy during inference and CPL* uses the vote-based strategy during inference.

Method	R@1			R@5		
	IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.1	IoU=0.3	IoU=0.5
WS-DEC [8]	62.71	41.98	23.34	-	-	-
VCA [33]	67.96	50.45	31.00	92.14	71.79	53.83
EC-SL [4]	68.48	44.29	24.16	-	-	-
MARN [24]	-	47.01	29.95	-	72.02	57.49
SCN [18]	71.48	47.23	29.22	90.88	71.56	55.69
BAR [34]	-	49.03	30.73	-	-	-
RTBPN [36]	73.73	49.77	29.63	93.89	79.89	60.56
CTF [5]	74.2	44.3	23.6	-	-	-
WSLLN [11]	75.4	42.8	22.7	-	-	-
LCNet [35]	78.58	48.49	26.33	<u>93.95</u>	**82.51**	62.66
CCL [37]	-	50.12	31.07	-	77.36	61.29
WSTAN [32]	79.78	52.45	30.01	93.15	79.38	<u>63.42</u>
CRM[2] [12]	<u>81.61</u>	55.26	<u>32.19</u>	-	-	-
CNM [39]	78.13	<u>55.68</u>	**33.33**	-	-	-
CPL, $\alpha_2 = 0.1$	79.86	53.67	31.24	87.24	63.05	43.13
CPL*, $\alpha_2 = 0.1$	**82.55**	**55.73**	31.37	87.24	63.05	43.13
CPL, $\alpha_2 = 1$	71.23	50.07	30.14	**94.28**	<u>81.32</u>	**65.79**

our model is sensitive to α_2. We set α_2 to 1 on Charades-STA, and to 0.1 or 1 on ActivityNet Captions (Details are in Sec.4.4). For the model training, we use Adam [14] optimizer with learning rate set to 0.0004.

4.4. Comparisons to the State-Of-The-Art

Tab. 1 and Tab. 2 compare the overall performance of CPL with previous works, where CPL uses the loss-based inference strategy and CPL* uses the vote-based inference strategy. We can draw the following conclusions: (1) On Charades-STA dataset, compared with previous methods, our CPL achieves 10.05% absolute gain on 'R@1, IoU=0.5'. Our loss-based strategy and vote-based strategy have similar performance on Charades-STA dataset. (2) On ActivityNet Captions dataset, for R@1 we outperform all existing methods with vote-based strategy. We find that the accuracy of reconstruction on ActivityNet Captions is relatively lower, and choosing the final prediction by voting is more reliable. For R@5, by carefully selecting $\alpha_2 = 1$ to encourage more diversity among multiple proposals, our model performs best. (3) In practical applications, α_2 can be set flexibly under different application requirements. The main reason is that the queries containing a complex relationship of multiple events are more difficult to reconstruct, resulting in less reliable measurement when choosing the best one from those positive proposals with a large diversity.

[2]Directly comparing CRM with others (including our CPL) is not fair. CRM requires a paragraph description annotation (multiple events described sequentially) per video in training, which is not always available.

Table 3. The ablation study of our different losses.

Loss Used			R@1			
$\mathcal{L}_{rec}$	$\mathcal{L}_{IVC}$	$\mathcal{L}_{div}$	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
✓	✗	✗	54.24	20.49	6.74	33.99
✓	✓	✗	60.39	32.08	12.95	37.98
✓	✓	✓	**66.40**	**49.24**	**22.39**	**43.48**

Table 4. The ablation study of positive and negative proposal generation processes.

Positive Proposal	Negative Proposal	R@1			
		IoU=0.3	IoU=0.5	IoU=0.7	mIoU
Fixed	None	55.07	28.97	10.13	34.06
Learnable	None	61.68	45.47	20.14	40.43
Learnable	Other video	65.64	47.56	21.37	42.34
Learnable	Intra-video	**66.40**	**49.24**	**22.39**	**43.48**

Table 5. The ablation study of our training strategy that only the positive proposal with the smallest L_{ce}^{p} participates in the optimization.

Strategy	R@1			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
All (equally)	54.84	36.16	18.49	36.28
All (weighted)	54.94	37.49	17.48	36.44
One w/ smallest $\mathcal{L}_{ce}^{p}$	**66.40**	**49.24**	**22.39**	**43.48**

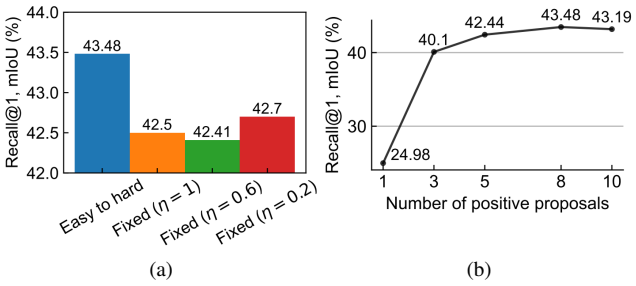

Figure 4. Fig. 4a shows the effectiveness of learning negative proposals from easy to hard. Fig. 4b shows the ablation study of different number of positive proposals.

4.5. Ablation Study

To prove the effectiveness of different components of our CPL, we perform ablation studies on the Charades-STA with our loss-based strategy.

Effectiveness of different losses. As Tab. 3 shows, we evaluate the effectiveness of $\mathcal{L}_{rec}$, $\mathcal{L}_{IVC}$, and $\mathcal{L}_{div}$. The model with all these three losses performs best, indicating that the design of all of them is significant for our network. We also find that the diversity loss can significantly improve the performance of the model, which proves that generating

multiple different positive proposals is beneficial to more accurately locating the event of interest in long videos.

Effectiveness of Proposal Generation. (1) As the first two rows in the Tab. 4 show, we evaluate the effectiveness of our learnable positive proposals. The 'Fixed' means that we use sliding windows and policy gradient algorithm used by SCN [18] to select positive proposals, and the 'None' means no negative proposals are used. We can see that our learnable proposals are of high quality. (2) As the last two rows in the Tab. 4 show, we evaluate the effect of our negative proposals in the same video. It reveals the fact that mining negative proposals within the video plays an important role in improving the performance of our method. (3) Moreover, as shown in the Fig. 4a, we evaluate the effectiveness of introducing negative proposals from easy to hard. When negative proposals are always far away from positive proposals (*i.e.* $\eta = 0.2$), they are easy to distinguish and provide little information. When they are always close to the positive proposal (*i.e.* $\eta = 1$), they may bring wrong information especially at the early training stage where the accuracy of positive proposals is low. Dynamically adjusting η and learning from easy to hard can effectively balance these two situations.

Effect of multiple Positive Proposals. (1) As Fig. 4b shows, we evaluate the effectiveness of the number of positive proposals. We can see that the more the number of positive proposals, the higher the mIoU in general. But when the number of positive proposals is greater than 8, the gain is very small. Continuing to increase the number of positive proposals will increase the computational cost. (2) Tab. 5 shows the effectiveness of our training strategy that only the positive proposal with the smallest $\mathcal{L}_{ce}^{p}$ participates in optimization. The 'equally' means all the positive proposals participate equally, and 'weighted' means positive proposals with smaller $\mathcal{L}_{ce}^{p}$ will participate more. We can see that it's helpful to only optimize the positive proposal with the smallest $\mathcal{L}_{ce}^{p}$ (encourage one-to-one correspondence between positive proposals and query).

4.6. Qualitative Results

Fig. 5 shows some qualitative examples. P1 and P2 (in blue) are our positive proposals with the lowest top-2 loss $\mathcal{L}_{ce}^{p}$. (1) As shown in Fig. 5(a) and (b), our method can achieve better results than SCN, proving that our negative proposals and references can provide more information. (2) As shown in Fig. 5(a), (b) and (c), comparing P1 and P2, the higher the ranking, and the more relevant the semantic meaning with the ground truth's. (3) Fig. 5(c) shows that the performance of CPL is relatively poor when the query contains complex relationships of multiple events. This may be caused by the low reconstruction accuracy in this situation.

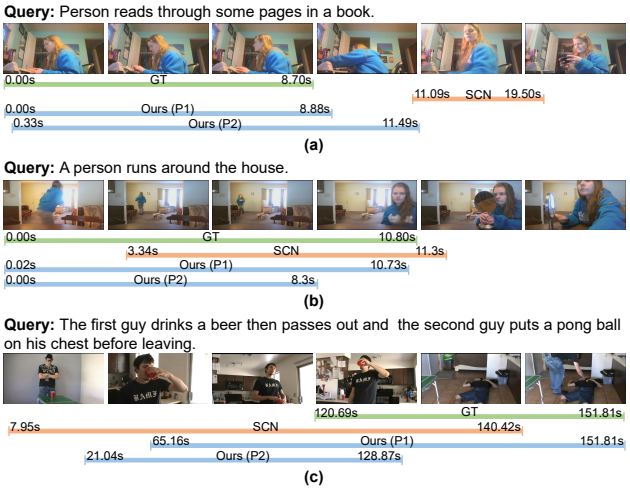

Figure 5. Qualitative examples of our top-2 predictions. Fig. 5 (a) and (b) are from the Charades-STA dataset, and Fig. 5 (c) is from the ActivityNet Captions dataset respectively.

5. Conclusion

In this work, we propose a novel weakly supervised video moment localization method, called Contrastive Proposal Learning(CPL). Our CPL generates several learnable Gaussian masks as proposals, which are effective and of high quality. We propose a novel method to mine the negative proposals within the same video, and introduce the entire video as the reference, which enables the network to distinguish highly confusing scenes. Inspired by the curriculum learning, the difficulty of the negative proposals increases as the training continues, benefiting the optimization. Experiments on the Charades-STA and ActivityNet Captions datasets show the outstanding performance of CPL. Extensive ablation studies also verify the effectiveness of the components in CPL.

Limitation Discussion: In this work, we focus on exploring how to learn high-quality proposals through the contrast of positive and negative proposals within the video. However, we find that when the query describes several events with complex relationship (with specified chronological order), our method may fail. How to better explore and represent the complex relationship between different events can be studied in future work.

6. Acknowledgements

This work is supported by the grants from the National Natural Science Foundation of China (61925201, 62132001, U21B2025) and Zhejiang Lab (NO. 2022NB0AB05).

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery. 3

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 2, 6

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 4, 6

[4] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8435, 2021. 7

[5] 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. 2, 6, 7

[6] Robert T Collins, Alan J Lipton, Takeo Kanade, Hironobu Fujiyoshi, David Duggins, Yanghai Tsin, David Tolliver, Nobuyoshi Enomoto, Osamu Hasegawa, Peter Burt, et al. A system for video surveillance and monitoring. *VSAM final report*, 2000(1-68):1, 2000. 1

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[8] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. *arXiv preprint arXiv:1812.03849*, 2018. 7

[9] Zhiyuan Fang, Shu Kong, Zhe Wang, Charless Fowlkes, and Yezhou Yang. Weak supervision and referring attention for temporal-textual association learning, 2020. 6

[10] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query, 2017. 2, 6

[11] Mingfei Gao, Larry S Davis, Richard Socher, and Caiming Xiong. Wslln: Weakly supervised natural language localization networks. *arXiv preprint arXiv:1909.00239*, 2019. 1, 7

[12] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7199–7208, 2021. 1, 2, 3, 6, 7

[13] Chen Ju, Peisen Zhao, Siheng Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Divide and conquer for single-frame temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13455–13464, 2021. 2

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[15] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 6

[16] Siyang Li, Xiangxin Zhu, Qin Huang, Hao Xu, and C. C. Jay Kuo. Multiple instance curriculum learning for weakly supervised object detection, 2017. 3

[17] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017. 4

[18] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network, 2020. 1, 2, 3, 5, 6, 7, 8

[19] 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 *European Conference on Computer Vision*, pages 156–171. Springer, 2020. 1, 6

[20] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proceedings of the Tenth ACM International Conference on Multimedia*, MULTIMEDIA '02, page 533–542, New York, NY, USA, 2002. Association for Computing Machinery. 1

[21] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2019. 1, 2, 3, 6

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

[23] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3, 5, 6

[24] 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. 1, 6, 7

[25] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey, 2021. 3

[26] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2083–2092, 2021. 2, 6

[27] Ye Tang, Yu Bin Yang, and Yang Gao. Self-paced dictionary learning for image classification. In *Proceedings of the 20th ACM international conference on Multimedia*, 2012. 3

[28] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 4, 6

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 4

[30] Hao Wang, Zheng-Jun Zha, Xuejin Chen, Zhiwei Xiong, and Jiebo Luo. Dual path interaction network for video moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 4116–4124, New York, NY, USA, 2020. Association for Computing Machinery. 2

[31] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. Structured multi-level interaction network for video moment localization via language query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7026–7035, June 2021. 2

[32] Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li. Weakly supervised temporal adjacent network for language grounding. *IEEE Transactions on Multimedia*, 2021. 2, 6, 7

[33] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metze, and Balakrishnan Prabhakaran, editors, *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1459–1468. ACM, 2021. 6, 7

[34] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos, 2020. 6, 7

[35] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE Transactions on Image Processing*, 30:3252–3262, 2021. 3, 6, 7

[36] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4098–4106, 2020. 3, 6, 7

[37] Zhu Zhang, Zhou Zhao, Zhijie Lin, jieming zhu, and Xiuqiang He. Counterfactual contrastive learning for weakly-supervised vision-language grounding. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18123–18134. Curran Associates, Inc., 2020. 6, 7

[38] Yang Zhao, Zhou Zhao, Zhu Zhang, and Zhijie Lin. Cascaded prediction network via segment tree for temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4197–4206, June 2021. 2

[39] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 1, 2, 3, 4, 5, 6, 7

[40] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. Embracing uncertainty: Decoupling and debias for robust temporal grounding, 2021. 2

[41] Zhi-Hua Zhou. Ensemble learning. In *Machine Learning*, pages 181–210. Springer, 2021. 6