Denoised Dual-level Contrastive Network for Weakly-supervised Temporal Sentence Grounding

Yaru Zhang[1,2], Xiao-Yu Zhang[1,2](✉), and Haichao Shi[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
{zhangyaru, zhangxiaoyu, shihaichao}@iie.ac.cn

Abstract. The task of temporal sentence grounding aims to localize the target moment corresponding to a given natural language query. Due to the large burden of labeling the temporal boundaries, weakly-supervised methods have drawn increasing attention. Most of the weakly-supervised methods heavily rely on aligning the visual and textual modalities, ignoring modeling the confusing snippets within a video and non-discriminative snippets across different videos. Moreover, the error-prone caused by the sparsity of video-level labels is not well explored, which brings noisy activations and is not robust to real-world applications. In this paper, we present a novel Denoised Dual-level Contrastive Network, namely DDCNet, to overcome the above limitations. Particularly, DDCNet is equipped with a dual-level contrastive loss to explicitly address the incomplete predictions by simultaneously minimizing the intra-video and inter-video loss. Moreover, a ranking weight strategy is presented to select high-quality positive and negative pairs during training. Afterward, an effective pseudo-label denoised process is introduced to alleviate the noisy activations caused by the video-level annotations, thereby leading to more accurate predictions. Comprehensive experiments are conducted on two widely used benchmarks, i.e., Charades-STA and ActivityNet Captions, manifesting the superiority of our method in comparison to existing weakly-supervised methods.

Keywords: Temporal sentence grounding · Weakly-supervised learning · Contrastive learning · Video denoising.

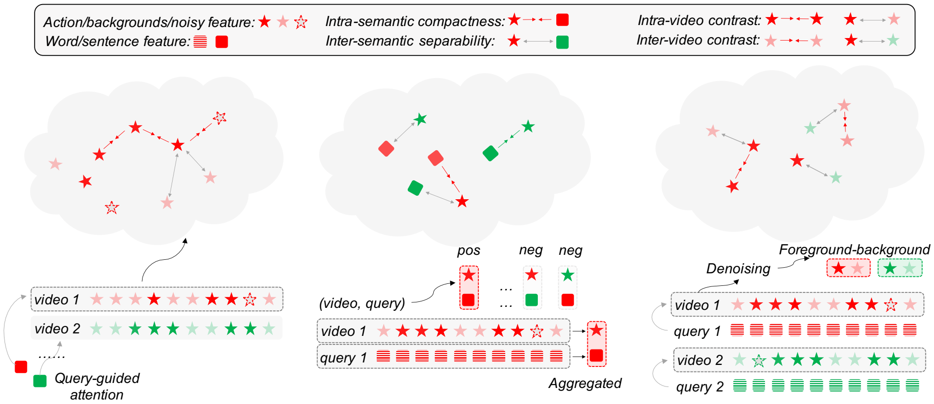
1 Introduction

Temporal sentence grounding aims to localize the temporal boundaries of the target moment that semantically corresponds to the given language query. As a fundamental vision-language problem, temporal sentence grounding has attracted extensive attention due to its broad applications, including surveillance [10], video summarization [25], and so forth. With the rapid development of deep learning technologies, fully-supervised methods [6] have made tremendous achievements in recent years, where precise temporal boundaries of each query are required for the model training. However, such manually eye-watching annotations are laborious and time-consuming, leading to expensive annotation costs. In addition, labeling temporal boundaries corresponding with the specific query is usually subjective and ambiguous, which narrows its scalability and practicability potential in real-world scenarios. Therefore, the weakly-supervised learning schemes, where only video-level natural language queries are needed, have rapidly attracted much more research interest due to the low annotation cost and time efficiency.

To identify the target moment that best matches the given query, it becomes crucial to improve the snippet-wise feature discrimination ability of various video snippets. Generally, the snippet-wise feature embedding space is expected to satisfy two properties: 1) the most relevant video snippet with the given query should be distinguished from the other snippets within a video, *i.e., intra-video separability*; 2) video snippets and queries with similar semantics should be closer than those of different semantics, *i.e., intra-semantic compactness & inter-semantic separability*. This has raised several studies exploring contrastive learning [21, 48, 49, 5, 23] to foster feature discrimination. Fig. 1 shows different contrastive learning schemes and their distinction with our proposed method. As illustrated in Fig. 1 (a), their focus is mostly on intra-video separability. After performing query-guided attention, snippet-wise target moments are pushed away from their backgrounds within a video. They unfortunately fail to capture the inter-semantic separability and discard the useful "global" contrast across different videos. In Fig. 1 (b), another type of effort strived to consider the matched and mismatched video-language pairs and engage them in the feature contrastive training process. Due to the uncertainty of sampling quality, this method would inevitably give rise to suboptimal performance of sentence grounding.

Due to the lack of frame-level temporal boundaries, snippet-wise pseudo-labels are often used to provide fine-grained supervision. For example, WSLLN [15] designs a parallel network with an extra surrogate module to generate pseudo labels, which will further encourage competition among candidate proposals and foster the feature discrimination. This also led to several pioneer studies exploring self-supervised learning [23], temporal adjacent network [37], pseudo-query generation [27] to refine the predictions. In spite of promising performance, the paradigm is easy to generate noisy activations, i.e., false positives and false negatives in the learned feature space. Most of the existing weakly-supervised temporal sentence grounding methods rely heavily on the pseudo-label strategy to provide refined supervision but do not explicitly handle the label noise.

To address the aforementioned problems, we propose a novel weakly-supervised method namely Denoised Dual-level Contrastive Network (DDCNet), by incorporating the label denoising process into video-language alignment under the constraints of intra-video and inter-video contrastive losses. DDCNet is reconstruction-based, dual-level contrast, and noise-label robust. As illustrated in Fig. 1 (c), for each video-language pair, we force the network to disentangle the query-related snippets (foreground) and other irrelative snippets (background) within a video. To attain the discriminative proposals in each video, we devise a margin ranking based intra-video contrastive loss to distinguish the foregrounds from backgrounds from easy to hard. Then all foreground and background representations among different videos are collected to engage in inter-video contrastive learning. As the natural language description is diverse and subjective, only foregrounds with similar semantic information or backgrounds with similar contexts can make a positive effect. In this case, we propose a similarity-based rank weighting module to reduce the impact of dissimilar positive pairs, and enhance the positive impact of similar pairs. Furthermore, to mitigate the negative influence of noisy



(a) Snippet-wise Contrast Within a Video (b) Video-query Pair Contrast (c) Our Denoised Dual-level Contrast

Fig. 1. Different contrastive learning schemes. (a) Exploiting *intra-video separability* to separate the target moments away from their backgrounds within a video. **(b)** Exploiting *intra-semantic compactness & inter-semantic separability* within minibatch to contrast the matched and mismatched video-language pairs. **(c)** Our denoised dual-level contrastive algorithm with 1) *intra-video contrast*, 2) *inter-video contrast* and 3) pseudo-label denoising module. The stars stand for the video features, and squares stand for the query features. Different colors indicate different videos or queries (better viewed in color).

pseudo-labels that are omnipresent in weakly-supervised methods, a determinant-based denoised loss is designed to generate reliable pseudo-labels and suppress the noisy activations. The denoised loss, on the other hand, is capable of encouraging a more robust joint feature space by enhancing the mutual information (MI) between query-related activations and pseudo-foreground within a video. To this end, the uncertainty of predictions is reduced, leading to more accurate predictions.

Our contributions are summarized as follows:

- We introduce a novel denoised dual-level contrastive network, named DDCNet, for the problem of weakly-supervised temporal sentence grounding. To the best of our knowledge, we are the first to explicitly address the pseudo-label noises that are omnipresent in weakly-supervised methods.
- We present a dual-level contrastive loss to enhance the discriminability and completeness of the target moment. By disentangling each untrimmed video into query-related foreground and irrelevant backgrounds, our proposed method achieves intra-video separability, intra-semantic compactness and inter-semantic separability simultaneously.
- We design a pseudo-label noise removal process to guarantee the robustness of temporal sentence grounding. In contrast to ignoring noisy activations in feature interaction, our method reduces the negative impact and achieves refined predictions.
- Comprehensive experiments are performed on Charades-STA and ActivityNet Captions datasets, which demonstrate the effectiveness of our DDCNet when compared with existing weakly-supervised methods.

2 Related Work

2.1 Weakly-supervised Temporal Sentence Grounding

Weakly-supervised temporal sentence grounding is becoming more attractive due to its practical effects in reducing the burden of collecting frame-level annotations. Past efforts can be categorized as either multi-instance learning (MIL) based methods [26, 15, 38, 37, 24, 47, 39] or reconstruction-based methods [12, 21, 48, 49, 32]. The MIL-based methods treat an untrimmed video as a bag of instances with video-level query annotations, and typically learn to predict temporal boundaries with a triplet loss. Among them, TGA [26] first presents a text-guided attention to optimize the video-text alignment space. WSLLN [15] jointly learns the cross-modal alignment and discriminative proposal selection. Follow-up works expand the MIL-based framework by designing sophisticated cross-modal modules [38, 37, 33], proposing proposal selection strategies [24, 47], or building effective objective functions [11, 39, 43]. However, these MIL-based methods heavily rely on the quality of randomly-selected negative pairs, and cannot provide enough strong supervision signal. In contrast, reconstruction-based methods aim to select moments that can reconstruct the given language query, and use the intermediate results for predicting temporal boundaries. Based on this concept, SCN [21], where masked words and predicted moments are fed to reconstruct the origin query, assuming localized moments should be able to accomplish those important words. Besides, CNM [48] and CPL [49] recently introduce a learnable Gaussian mask to generate high-quality positive and negative proposals, which highly improves the grounding performances due to the superiority of content-related proposal generation. Inspired by such advances, our approach takes a further step to explore the denoised contrastive learning from a large number of weakly annotated videos, which fosters the discrimination and robustness from both intra-video and inter-video aspects.

2.2 Contrastive Representation Learning

Contrastive learning presents a remarkable performance due to its great potential capability for un-/self-supervised representation learning [31, 40, 34]. These approaches seek to learn such an embedding feature space in which similar (or positive) sample pairs should be pulled together while dissimilar (or negative) ones are pushed apart. Some approaches even achieve favorable performance without engaging negative pairs [3, 16, 8]. Following the success of contrastive representation learning, some recent efforts are making an attempt to adapt such a paradigm into the video domain. For instance, VideoMoCo [28] employs the image-based MoCo method for video representation, which largely improves the temporal representation capability for video-related tasks. In video grounding task, AsyNCE [11] proposes to reduce the impact of the false positives by leveraging flexible AsyNCE loss, encouraging effective communication between cross-modal interaction for weakly-supervised grounding. To improve the training efficiency, CCL [47] develops a counterfactual contrastive framework, which verifies the effectiveness and robustness of vision-language grounding. However, these methods are mostly based on NCE loss and its variants, while other types of loss formulation have not been well explored. Different from these approaches, in this paper we

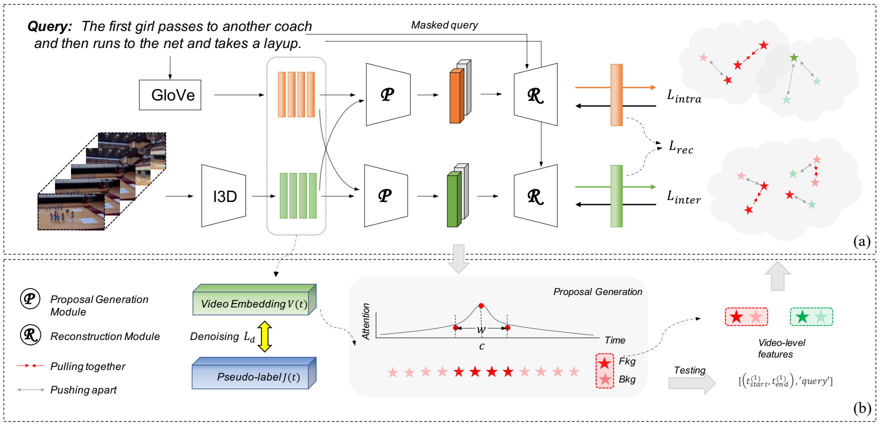


Fig. 2. The overall framework of our DDCNet. The focus of our method is to jointly enhance the discriminability and completeness of latent moment embeddings and addresses the pseudo-label noisy omnipresent in weakly-supervised learning. The upper stream (a) presents our method trained with dual-level contrastive loss, which consists of an intra-video and inter-video contrastive loss to optimize the proposal generation module $\mathcal{P}$. In the bottom stream (b), we propose a denoised algorithm aiming to reduce the impact of noisy activations in temporal sentence grounding. Besides query reconstruction loss $\mathcal{L}_{rec}$ in $\mathcal{R}$, the network is trained jointly using loss terms $\mathcal{L}_{intra}, \mathcal{L}_{inter}$ and $\mathcal{L}_d$.

perform contrastive learning on weakly-supervised temporal sentence grounding with both intra-video and inter-video contrast, and achieve compelling results both quantitatively and qualitatively.

3 The Proposed Method

3.1 Problem Formulation

Given a pair of untrimmed video and associated language query (V_i, Q_i), where V_i and Q_i separately represents a video and the corresponding language query. The goal of weakly-supervised temporal sentence grounding is to ascertain a moment τ that temporally matches the query Q_i with only the video-level annotation given. More specifically, we denote the input video as a frame sequence with l_V snippets, termed as $V_i = \{v_t\}_{t=1}^{l_V}$, where v_t represents the video snippet at timestamp t. Similarly, the associated language query can be represented as $Q_i = \{w_j\}_{j=1}^{l_S}$, where w_j and l_S represent the j-th single word in the language and the number of total words. Under this primary notation, our model is to learn a mapping function for predicting the moment boundary with parameter Θ, which can be formulated as follows:

$$f_\Theta : (V_i, Q_i) \to (\tau_s, \tau_e), \tau_s < \tau_e,$$

where τ_s and τ_e indicate the indices of the start and end timestamp of the predicted boundary, respectively.

3.2 Visual-Text Feature Extraction

Before generating more expressive representations, we first embed the given video and its corresponding language query into a continuous high-dimension feature space. For each video, we employ a pretrained feature extractor (e.g., C3D [35] or I3D[4]) to extract video representations $\mathcal{V}$ and then apply the temporal pooling operation on them to reduce the feature dimension. Here the extracted video features can be represented as $\mathcal{V} = \{V_1, V_2, \ldots, V_{l_V}\} \in \mathbb{R}^{l_V \times d_V}$, where d_V stands for the video feature dimension. As for the language query, we employ the GloVe [29] model to obtain the query embedding with respect to each word. In this case, the query embedding can be naturally represented as $\mathcal{Q} = \{W_i, W_2, \ldots, W_{l_S}\} \in \mathbb{R}^{l_S \times d_Q}$, where d_Q denotes the query embedding dimension. It's worth noting that we didn't finetune the pretrained feature extractor on the given untrimmed video datasets in order to guarantee a fair comparison with existing proposed methods.

3.3 Gaussian-based Proposal Generation

Following the standard practice, we utilize a Gaussian-based mask generator $\mathcal{P}$ [48] to generate high-quality proposals with query-related semantics. Inspired by the recent success of Transformer [36], we first use the multi-head attention module to capture long-range semantic representations from the query, dubbed as $F = MHA(\mathcal{Q})$, and then arrive at fused hidden feature $\mathcal{H}$ that incorporate both video and query semantics, given by

$$\mathcal{H} = TransEncoder(\mathcal{V}, F, \mathcal{V}) \in \mathbb{R}^{l_V \times d_H}, \tag{1}$$

where $TransEncoder(\cdot)$ represents a Transformer-based encoder architecture, and d_H denotes the feature dimension. As $\mathcal{H}$ combines both semantic and vision information, we predict the center and width of our target proposal through a fully connected layer followed by a *Sigmoid* calculation, which can be denoted as G_c and G_w respectively. Afterwards, the Gaussian-based mask can be derived as

$$\varphi^p(i) = \exp\left(-\frac{\alpha(i/N - G_c)^2}{G_w^2}\right), i = 1, \ldots, N \tag{2}$$

where $\varphi^p(i)$ represents the probability of the i-th video snippet being the foreground proposal in the Gaussian mask, and α denotes a hyperparameter that controls the variance of the Gaussian curve.

To obtain more complete predictions, we encourage to produce K Gaussian masks through a multi-branch module. To avoid the branches lazily concentrating on the same video snippet, a diversity loss is imposed on them:

$$\mathcal{L}_{div} = \frac{1}{K} \sum_{k=1}^{K} \max\left(\left\|\varphi^p \varphi^{p\top} - \mathbf{I}\right\|_F^2 - b, 0\right), \tag{3}$$

where K is the hyperparameter, b is a balance vector that controls the extent of overlap between different masks. After that, we average K Gaussian masks to obtain the final proposal:

$$\varphi^{avg} = \frac{1}{K} \sum \varphi_k^p(i). \tag{4}$$

In this case, the average mask captures and combines different action parts, effectively encode the entire action.

3.4 Intra-video Contrastive Learning

Although we have obtained a series of content-based proposals based on the Gaussian generation module, there still exist a few highly-confusing snippets inside the video that puzzle the generator, thereby leading to inaccurate boundary prediction. To enable the generator more distinguishable, we study the intra-video contrastive representation learning with both easy and hard negative snippets. Intuitively, we regard $\varphi^e = (1 - \varphi^p) \in \mathbb{R}^N$ as the easy negative sample, which corresponds to those video snippets that mostly do not match the given query. As for the hard negative sample, we refer to the entire video as it contains overlapping snippets with both foregrounds and semantically related backgrounds, given by

$$\varphi^h = [1, 1, \ldots, 1] \in \mathbb{R}^N. \tag{5}$$

Training the generator with both easy and hard negative samples can help the model to locate more accurate predictions and prevent the model from outputting longer boundaries that include the ground truth.

As discussed earlier, our goal is to highlight the salient moment that best matches the language query. To measure the semantic relevance between the moment proposal and query, we introduce the semantic completion module $\mathcal{R}$ [21] to calculate reconstruction scores and regard them as feedback to refine previous proposals. Firstly, we mask 1/3 keywords of the original query and then attain the masked query embedding, dubbed as $\hat{\mathcal{Q}}$, which is subsequently fed into the $\mathcal{R}$ together with original video features and foreground Gaussian mask. The specific process can be formulated as follows:

$$\mathcal{W}^p = TransDecoder\left(\hat{\mathcal{Q}}, U, \varphi^p\right) \in \mathbb{R}^{l_S \times d_U}, \tag{6}$$

where $\hat{\mathcal{Q}}$ represents the masked query embedding, $U = TransEncoder\left(\mathcal{V}, \mathcal{V}, \varphi^p\right)$ aiming to attain visual representations with respect to φ^p, and $TransDecoder(\cdot)$ denotes the completion module that can be used to achieve the reconstructed feature with respect to each word.

To predict the masked words, we apply a single fully connected layer on $\mathcal{W}^p$ and output the probability distribution $\mathcal{P}^p$ of total reconstructed query. Finally, we use the cross-entropy loss to measure the similarity distance between the reconstructed query and the original query, which is given by

$$\mathcal{D}_{ce}^p = -\sum_{i=1}^{l_S - 1} \log \mathcal{P}^p\left(w_{i+1} \mid \mathcal{V}, \mathcal{Q}_{1:i}\right). \tag{7}$$

Similarly, we arrive at $\mathcal{D}_{ce}^e$ and $\mathcal{D}_{ce}^h$ by replacing φ^p with φ^e and φ^h. As only positive sample and hard negative sample contain video snippets corresponding to the query, the final reconstruction loss is defined as:

$$\mathcal{L}_{rec} = \mathcal{D}_{ce}^p + \mathcal{D}_{ce}^h. \tag{8}$$

To contrast positive and highly-confusing negative proposals, we utilize the ranking-motivated loss for intra-video contrastive learning, which can be formulated as:

$$\mathcal{L}_{intra} = [\mathcal{D}_{ce}^{p} + \lambda_1 - \mathcal{D}_{ce}^{e}]_{+} + \left[\mathcal{D}_{ce}^{p} + \lambda_2 - \mathcal{D}_{ce}^{h}\right]_{+}, \tag{9}$$

where $[\cdot]_+$ is the hinge function, λ_1 and λ_2 are hyperparameters with a constraint $\lambda_1 < \lambda_2$.

3.5 Inter-video Contrastive Learning

While the intra-video contrastive learning gives us good representations for distinguishing highly-confusing snippets and the real proposals, it fails to address the incompleteness issue of proposals, especially when encountering condition variations with respect to complex semantics (e.g. various scales, viewpoints, or illumination conditions). In this case, it is natural to resort to exploring the correspondence and knowledge transfer across different videos, extending intra-video contrastive learning into inter-video contrastive learning. In this section, we embrace the fact that learning from cross-video foreground-background contrast produces more reliable foreground proposals, and design an inter-video contrastive loss from two aspects: 1) the representation of different video snippets of the same semantics should be close and 2) other representations of opposite semantics should be pushed apart.

Given n video sequences, we first compute a set of Gaussian mask corresponding to their queries based on the generation module $\mathcal{P}$. Then, φ^p and $(1 - \varphi^p)$ are multiplied by $\mathcal{V}$ to disentangle each video into a foreground f and a background b representation. To this end, we can collect n *negative* foreground-background pair $\{(f_i, b_j)\}_{i=1}^{n}$ for the total video set. In this case, the negative contrastive loss is designed as:

$$\mathcal{L}_{inter}^{Neg} = -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\log\left(1 - \varDelta(f_i, b_j)\right), \tag{10}$$

where $\varDelta(i, j)$ is the cosine similarity between f_i and b_j. The $\mathcal{L}_{inter}^{Neg}$ considers the semantic contrasts both within a single video $(i = j)$ and cross different videos $(i \neq j)$.

To boost the discrimination of activated foregrounds and suppress the co-occurring backgrounds, we consider the other two *positive pairs* (f_i, f_j), (b_i, b_j) from different videos and intend to pull these positive pairs together in the feature space. However, only *positive pairs* with similar semantics that really benefit the model training while those with large distances will degrade the training process. To cope with it, we design a distance-based rank weighting strategy to automatically learn the effect of different positive pairs. It can reduce the impact of those dissimilar pairs to some extent for better contrastive learning. Formally, the positive contrastive loss $\mathcal{L}_{inter}^{Pos}$ is defined as a combination of $\mathcal{L}_{f}^{Pos}$ and $\mathcal{L}_{b}^{Pos}$, which is represented as:

$$\mathcal{L}_{f}^{Pos} = -\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\nVdash_{[i \neq j]}\left(w_{i,j}^{f} \cdot \log\left(\varDelta(f_i, f_j)\right)\right) \tag{11}$$

$$\mathcal{L}_b^{Pos} = -\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \nVdash_{[i \neq j]} \left(w_{i,j}^{b} \cdot \log \left(\varDelta(b_i, b_j) \right) \right), \tag{12}$$

where $\nVdash$ is an indicator function that equals 1 if $(i \neq j)$. To this end, the overall inter-video objective loss can be formulated as a combination of $\mathcal{L}_{inter}^{Neg}$ and $\mathcal{L}_{inter}^{Pos}$, which is defined as:

$$\mathcal{L}_{inter} = \mathcal{L}_{inter}^{Pos} + \mathcal{L}_{inter}^{Neg}. \tag{13}$$

When contrastive loss $\mathcal{L}_{inter}$ is applied, our proposed network will enhance more complete proposal predictions and simultaneously suppress the co-occurring query-related backgrounds in the training process.

3.6 Pseudo-label Noise Removal

The aforementioned two loss functions, however, only ensure φ^p to be more discriminative and completely cover the target moment, without considering the noisy activation in multi-modal interaction process. To enhance the robustness of foreground snippets and further refine the predictions, we propose to denoise the snippet-wise pseudo-labels by capturing the mutual information between the temporal activation and their corresponding pseudo-labels. Unlike directly reducing the impact of noisy features, the pseudo-label denoising process can serve as initial fine-grained annotations and be more applicable to existing weakly-supervised methods.

Based on the above observations, we first generate snippet-wise pseudo-label $\mathcal{J}$ to refine foreground and background regions, then build a denoised loss $\mathcal{L}_d$ to improve the robustness of foreground activation with respect to the noisy activation. Intuitively, we calculate snippet-wise pseudo-labels by computing the similarity of each video snippet and the corresponding query. The specific process is formulated as:

$$\mathcal{J}(t) = \frac{1}{2} \left(1 + \varDelta \left(\mathcal{V}(t), \mathcal{Q}_{ref} \right) \right), \quad t \in [1, l_V], \tag{14}$$

where $\mathcal{V}(t)$ is the video feature corresponding to the timestamp t, $\mathcal{Q}_{ref}$ denotes the mean of query features over m iterations.

Let $t_f = \{t : \mathcal{J}(t) > 0.5\}$ and $t_b = \{t : \mathcal{J}(t) < 0.5\}$ represent the time slots for selecting the foreground and background snippets with respect to $\mathcal{J}(t)$, we can estimate the snippet-wise label for both foreground and background snippets.

After generating snippet-wise pseudo-labels, we need to reduce the impact of label noise caused by the absence of ground truth, thereby improving the accuracy of the predicted moments. The denoised loss $\mathcal{L}_d$ is designed to assign a confidence score to each snippet that estimates the probability of its pseudo-label being a trustworthy true label, which exploits the mutual information between query-related activation and corresponding labels. Concretely, our denoised loss is inspired by the Determinant based Mutual Information (DMI) [42], which is proposed for multi-class classification tasks and robust to a variety of noise patterns. The original DMI is first defined to compute the determinant of a joint distribution matrix, i.e., $Determin(\mathcal{Z}, \mathcal{Y}) = |\det(\mathcal{C})|$. Here, $\mathcal{Z}$ and $\mathcal{Y}$ denote the predicted probabilities and the ground-truth labels. $\mathcal{C} = 1/n\mathcal{Z}\mathcal{Y}$ is the joint distribution over $\mathcal{Z}$ and $\mathcal{Y}$. Therefore, the denoised loss function is defined as:

$$\mathcal{L}_d = -\mathcal{E}[\log(Determin(\mathcal{C}))], \tag{15}$$

where $\mathcal{E}$ represents the *Expectation* function. Taking the set of snippet-wise pseudo-labels into account, we construct a prediction matrix that considers the set of pseudo-foreground/background temporal locations. Therefore, the final prediction matrix and pseudo-label matrix are given by

$$\hat{\mathcal{Z}} = \begin{bmatrix} \mathcal{J}_f & \mathcal{J}_b \\ 1-\mathcal{J}_f & 1-\mathcal{J}_b \end{bmatrix}, \quad \hat{\mathcal{Y}} = 1/z \begin{bmatrix} 1_{n_f} & 0_{n_f} \\ 0_{n_b} & 1_{n_b} \end{bmatrix}, \tag{16}$$

where $z = n_f + n_b$, $n_f = |t_f|$ and $n_b = |t_b|$ represent the width of constructed pseudo-foreground/background snippets. To avoid an explicit computation cost that caused by a large number of video snippets, we use an approximate formulation [17] to replace the original loss function. Finally, the denoised loss is defined as:

$$\mathcal{L}_d = -\mathcal{E}[\log(Determin(\hat{\mathcal{Z}}\hat{\mathcal{Y}}))] = \mathbb{E}\left[\log\left(\mathit{\Gamma}\right)\right], \tag{17}$$

where $\mathit{\Gamma}$ is the condition number of $\hat{\mathcal{Z}}\hat{\mathcal{Y}}$.

3.7 Training and Inference

In this section, we elaborate on the details of network training and the inference.

Training. The total loss of our DCCNet comprises four parts: the reconstruction loss $\mathcal{L}_{rec}$ is in charge of optimizing the semantic completion module, which guarantees the network to predict the reconstructed query that is conditioned on the given mask; the dual-level loss $\mathcal{L}_{dual}$ is used to ensure the video feature more distinguishable and distinct from highly confusing backgrounds within and without a video; the diversity loss $\mathcal{L}_{div}$ is used to encourage the K proposals as different as possible (if added); the denoised loss $\mathcal{L}_d$ is adopted to reduce the noisy activation caused by the absence of frame-level annotations.

To encourage the candidate predictions to best reconstruct the given query, we optimize the whole framework by alternately executing the following two steps:

1. Update reconstructor parameter by $\mathcal{L}_{rec} + \mathcal{L}_d$ while freezing the mask generator:

$$\alpha_1^* = \arg\min_{\alpha_1} L_{rec}\left(\alpha_1, \alpha_2\right) + L_d\left(\alpha_1, \alpha_2\right). \tag{18}$$

2. Update the mask generator with optimal α_1^* by minimizing $\mathcal{L}_{dual} + \mathcal{L}_{div}$:

$$\alpha_2^* = \arg\min_{\alpha_2} L_{dual}\left(\alpha_1^*, \alpha_2\right) + L_{div}\left(\alpha_1^*, \alpha_2\right). \tag{19}$$

where $\mathcal{L}_{dual} = \mathcal{L}_{intra} + \mathcal{L}_{inter}$, α_1 and α_2 are the parameters of the reconstructor and mask generator, respectively.

Inference. During inference, we can obtain the temporal boundary $\tau = (\tau_s, \tau_e)$ of predicted Gaussian mask through Eqn. 2. The predicted start and end timestamps are calculated as follows:

$$\begin{aligned} \tau_s &= \max(G_c - G_w/2, 0) * T_v \\ \tau_e &= \min(G_c + G_w/2, 1) * T_v, \end{aligned} \tag{20}$$

where T_v represents the duration of the target video to be locate. Since we do not use multi-scale sliding windows to generate proposal candidates, it's noteworthy that we do not have to perform complex post-processing operations like Non-Maximum Suppression (NMS).

4 Experiments

4.1 Datasets

To validate the effectiveness of our proposed DDCNet, we perform experiments for weakly-supervised temporal sentence grounding on two prevailing and challenging datasets: Charades-STA [14] and ActivityNet Captions [20].

Charades-STA. The Charades-STA dataset is originally constructed from Charades [30] dataset which contains $9,848$ untrimmed videos about human daily indoor activities. Based on the Charades dataset, Gao *et al.* [14] develops a semi-automatic method to annotate each video with a moment-sentence pair. Concretely, the dataset consists of $12,408$ moment-sentence pairs for training and $3,720$ pairs for testing. The average duration, moment length, and query length are 29.8 seconds, 8.09 seconds and 7.22 words, respectively.

ActivityNet Captions. The ActivityNet Captions dataset is a large-scale dataset for temporal sentence grounding. It originally stems from ActivityNet dataset [2] for human activity understanding task, which comprises $14,926$ untrimmed videos and $71,953$ moment-sentence annotations. Following the standard experimental setting, we utilize val_1 as the validation set and val_2 as the testing set, which consists of $37,417$ pairs of video moments and descriptions for training, $17,505$ and $17,031$ pairs for validation and testing, respectively. Each video has an average of 4.82 temporal moments with their language descriptions. And the moment length and query length are about 37.14 seconds and 14.41 words on average.

4.2 Evaluation Metric

To evaluate the performance of our proposed method, we employ the commonly used $\langle \mathrm{R@}n, \mathrm{IoU@}m \rangle$ as our evaluation metric. Concretely, this metric is defined to compute the percentage of language queries whose predicted moments have at least one correct prediction in the top-n results. Specifically, a predicted moment is correct only if its IoU (i.e., Inter-section over Union) is larger than m in contrast with the ground truth. In our experimental setting, we report results for $n \in \{1, 5\}$ with $m \in \{0.3, 0.5, 0.7\}$ on Charades-STA, and $m \in \{0.1, 0.3, 0.5\}$ for ActivityNet Captions datasets.

4.3 Implementation Details

Data Preprocessing. For Charades-STA dataset, we utilize the publicly available I3D[4] network to extract visual features. For ActicvityNet Captions dataset, we employ the C3D [35] model pre-trained on Sport1M [18] dataset to obtain $4,096$ dimension features, which is subsequently reduced to 500 dimensions with the PCA algorithm. For a fair comparison, the feature extractor is not finetuned on both datasets. The input video is downsampled every 8 frame and the maximum length of frames is set to 200. For the sentence query, we adopt NLTK [22] to split each sentence into several words and employ the pre-trained GloVe [29] word2vec model to initialize the word embeddings. The maximum length of words is set to 20. And the vocabulary size is set to $1,111$ and $8,000$ for Charades-STA and ActivityNet Captions, respectively.

Table 1. Performance comparison between the proposed model and the state-of-the-arts on Charades-STA dataset.

Method	Rank@1, IoU=			Rank@5, IoU=		
	0.3	0.5	0.7	0.3	0.5	0.7
CTRL [4]	-	23.63	8.89	-	58.92	29.52
QSPN [41]	54.70	35.60	15.80	95.60	71.80	38.87
MAN [1]	-	46.53	22.72	-	86.23	33.09
2D-TAN [45]	-	39.81	23.25	-	79.33	52.15
TGA [26]	32.14	19.94	8.84	86.58	65.52	33.51
WSRA [13]	50.13	31.20	11.01	86.75	70.50	39.02
WSTAN [37]	43.39	29.35	12.28	93.04	76.13	41.53
VLANet [24]	45.24	31.83	14.17	95.70	82.85	33.09
SCN [21]	42.96	23.58	9.97	95.56	71.80	38.87
MARN [32]	48.55	31.94	14.81	90.70	70.00	37.40
CNM [48]	60.39	35.43	15.45	-	-	-
WSTG [5]	43.31	31.02	16.53	95.54	77.53	41.91
RTBPN [46]	60.04	32.36	13.24	**97.48**	71.85	41.18
DDCNet (Ours)	**63.96**	37.14	16.05	-	-	-
DDCNet*(Ours)	63.71	**46.58**	**20.68**	97.12	**84.45**	**50.03**

Model Setting. To improve the training stability, we utilize the multi-head mechanism proposed in [36] for the mask generator and semantic completion module. Specifically, the encoder and decoder are both equipped with 3 layers and 4 multi-attention heads. And the dimension of the hidden state is set to 256. In the training phase, we employ Adam [19] as our optimizer without weight decay. The learning rate is set to $4e^{-4}$ for Charades-STA and ActivityNet Captions. Besides, the hyperparameters λ_1 and λ_2 are set to 0.1 and 0.15, respectively. K is set to 5. Following the standard practice, we mask $1/3$ important words in each sentence by replacing them with a special token when reconstructing the origin query, in which nouns and verbs are more likely to be selected to be the keywords. Moreover, the maximum width of the predicted moments is limited to shorter than 0.45 as the inherent property on the Charades-STA dataset.

4.4 Comparisons with State-of-the-Art Methods

We compare our proposed DDCNet with existing state-of-the-art approaches in recent years, including both fully-supervised and weakly-supervised methods.

Results on Charades-STA. We compare our DDCNet with the state-of-the-art fully-supervised and weakly-supervised methods on the Charades-STA testing set. The best results are highlighted in **bold** and the second best results are underlined in tables. As shown in Table 1, our method achieves impressive performance on almost all metrics except a slightly worse one, which verifies the effectiveness of our proposed

Table 2. Performance comparison between the proposed model and the state-of-the-arts on ActivityNet Captions dataset.

Method	Rank@1, IoU=			Rank@5, IoU=		
	0.1	0.3	0.5	0.1	0.3	0.5
TGN [6]	-	43.81	27.93	-	54.56	44.20
CTRL [4]	49.10	28.70	14.00	-	58.92	29.52
ABLR [44]	73.30	55.67	36.79	-	-	-
2D-TAN [45]	-	59.45	44.51	-	85.53	77.13
WS-DEC [12]	62.71	41.98	23.34	-	-	-
VCA [39]	67.96	50.45	31.00	92.14	71.79	53.83
EC-SL [7]	68.48	44.29	24.16	-	-	-
MARN [32]	-	47.01	29.95	-	72.02	57.49
SCN [21]	71.48	47.23	29.22	90.88	71.56	55.69
CTF [9]	74.2	44.3	23.6	-	-	-
WSLLN [15]	75.4	42.8	22.7	-	-	-
CCL [47]	-	50.02	31.07	-	77.36	**61.29**
CNM [48]	78.13	55.68	**33.33**	-	-	-
DDCNet (Ours)	79.36	56.53	31.81	-	-	-
DDCNet*(Ours)	**79.51**	**57.57**	32.29	**92.65**	**77.96**	60.54

DDCNet. Specifically, it can be seen that our approach achieves 63.96% on "Rank@1, IoU=0.3" and 37.14% on "Rank@1, IoU=0.5", bringing the compelling result by a large margin. Notably, we can see that a variant version of our approach (DDCNet*) with a multiple proposal generation scheme (Eqn. 3 and 4) outperforms other previous weakly-supervised methods at most of the IoU thresholds, demonstrating the superiority of denoised contrastive learning criteria without the precise frame-level annotations. It can be noticed that our method also attains competitive results even compared with some fully-supervised counterparts (in upper parts of the tables).

Results on ActivityNet Captions. As shown in Table 2, we also give a thorough study of the ActivityNet Captions dataset and report the corresponding results. Similarly, we compare the overall performance with both fully-supervised and weakly-supervised methods, where DDCNet* indicates an advanced version of our method with multiple proposals generation. As can be seen, our method shows significant improvements over existing weakly-supervised methods while maintaining competitive results with other fully-supervised methods. Specifically, we observe that our DDCNet attains the highest performance except for the "Rank@1, IoU=0.5" metric. This may stem from the intrinsic characteristics of this dataset. Since the query characteristics in ActivityNet Captions are diverse and complicated, there is a high probability to make the training models confused and ineffective. Compared with other recently proposed methods, however, like CCL [47] and CNM [48], our DDCNet still outperforms the MIL-based and reconstruction-based methods to a large extent. This suggests that our

Table 3. Ablation studies of the proposed model on Charads-STA dataset.

ID	$\mathcal{L}_{rec}$	$\mathcal{L}_{intra}$	$\mathcal{L}_{inter}$	$\mathcal{L}_{d}$	$\mathcal{L}_{div}$	IoU=0.1	IoU=0.3	IoU=0.5	mIoU
1	✓					76.85	53.74	29.20	53.26
2	✓	✓				76.79	60.54	35.59	57.64
3	✓	✓	✓			78.75	61.43	37.21	59.13
4	✓			✓		74.35	58.58	36.42	56.45
5	✓	✓	✓	✓		79.89	**63.96**	37.14	60.33
6	✓	✓	✓	✓	✓	**79.94**	63.71	**46.58**	**63.41**

method is robust and applicable to a large-scale dataset of various query semantics. By extending the framework to multiple proposal generation, our method nearly achieves consistent improvements among different IoU metrics beyond all doubt. Besides, our method achieves favorable performance even in contrast with existing fully-supervised methods, which reduces the performance gap by a large margin and benefits the practicability to real-world applications.

4.5 Ablation Study and Analysis

To investigate the effectiveness of our proposed DDCNet for weakly-supervised temporal sentence grounding, we conduct extensive ablation studies on both datasets. The results are summarized in Table. 3$\sim$ Table. 5.

Q1: How does the proposed multi-task loss help? To evaluate the effectiveness of our carefully-designed multi-task loss, we conduct ablation studies with respect to different losses, i.e., $\mathcal{L}_{intra}$, $\mathcal{L}_{inter}$, $\mathcal{L}_{d}$ and $\mathcal{L}_{div}$. The results are summarized in Table 3. As we can see, introducing the intra-video contrastive learning loss $\mathcal{L}_{intra}$ improves the Rank@1 mIoU from 53.26% to 57.64%, demonstrating that snippet-wise variances within the same video are essential for capturing discriminative representations. Furthermore, our method, which adds $\mathcal{L}_{inter}$ to perform inter-video contrastive learning, boosts the Rank@1 mIoU to 59.13%. This suggests $\mathcal{L}_{inter}$ effectively guides the network to produce more complete predictions by exploring cross-video relations. In addition, we also find that adding the denoised loss $\mathcal{L}_{d}$ achieves an absolute 1.2% improvement. And the modified version DDCNet* with $\mathcal{L}_{div}$ achieves the best average performance on Charades-STA dataset. This shows that each component provides an indispensable contribution to the learning model.

Q2: Is it necessary to consider both HP and HN terms in $\mathcal{L}_{inter}$ loss? While we have validated that our inter-video loss helps the training model achieve better performance, it should also be considered whether both HP and HN terms are essential components. To explore this, we conduct experiments that use two variants of the $\mathcal{L}_{inter}$ loss, each of which contains one kind of the loss term in Eqn. 13, i.e., $\mathcal{L}_{inter}^{Pos}$ and $\mathcal{L}_{inter}^{Neg}$, respectively. We summarize the corresponding results in Table 4. As we can see, the performance drops largely when either type of sub-loss is removed, demonstrating that both loss terms contribute to the improved prediction. Compared with the baseline, our

Table 4. Ablation studies of on $\mathcal{L}_{inter}$ terms on Charads-STA dataset.

Setting	Loss	IoU=0.5 ($\triangle$)
DDCNet (Ours)	$\mathcal{L}_{rec} + \mathcal{L}_{inter}$	**37.14%**
baseline	$\mathcal{L}_{rec}$	29.20% (-7.49%)
DDCNet w/o HN trm.	$\mathcal{L}_{rec} + \mathcal{L}_{inter}^{Pos}$	36.11% (-1.03%)
DDCNet w/o HP trm.	$\mathcal{L}_{rec} + \mathcal{L}_{inter}^{Neg}$	36.38% (-0.76%)

Table 5. The effectiveness of training strategy.

Setting	Rank@1			
	IoU=0.1	IoU=0.3	IoU=0.5	mIoU
$\min_{\alpha_1,\alpha_2} (\mathcal{L}_{recon} + \mathcal{L}_{gen})$	67.38	54.28	23.05	48.57
$\min_{\alpha_1} \mathcal{L}_{recon} + \min_{\alpha_2} \mathcal{L}_{gen}$	**79.89**	**63.96**	**37.14**	**60.33**

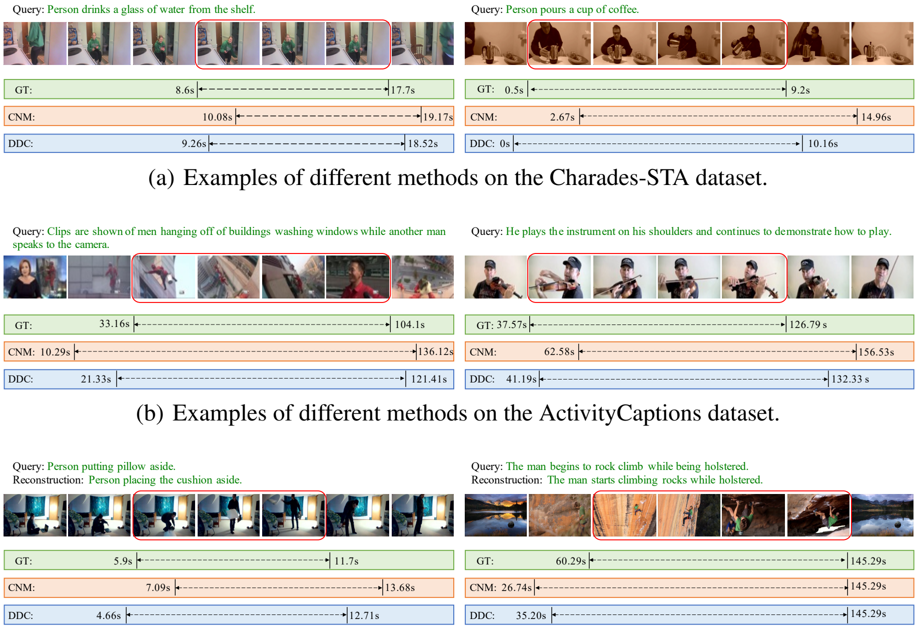
DDCNet is beneficial to make representations of similar snippets closer and helps to transfer informative knowledge. Overall, the above analyses strongly justify the significance of the two items in our proposed $\mathcal{L}_{inter}$ loss.

Q3: How does different training strategy effect the performance? As Table 5 shows, we conduct experiments to study how different training strategies influence the performance on the Charades-STA dataset. The first row indicates our DDCNet is trained by optimizing the mask generator and reconstruction module separately, where the weight of the generator is frozen when optimizing the query reconstruction module, and vice versa. In contrast, the second row demonstrates the entire model is optimized with $\mathcal{L}_{recon}$ and $\mathcal{L}_{gen}$ jointly. As we could see, the DDCNet performs better results when $\mathcal{L}_{recon}$ and $\mathcal{L}_{gen}$ work separately, with a consistent improvement in terms of Rank@1 metric at all IoU thresholds. This is because the iterative training manner can avoid a trivial solution that the reconstruction module always gives the predicted negative samples low scores at early training, thereby contributing to superior results.

4.6 Qualitative Results

Intuitively, we provide qualitative results from Charades-STA and ActivityNet Captions to further demonstrate the superiority of our DDCNet. As shown in Fig. 3, each video is presented with a human-annotated query description, along with the ground truth and predictions with different methods.

Specifically, Fig. 3 (a) displays two typical examples of the detected moments on the Charades-STA dataset. Compared with the ground truth, we can easily find that our method is capable of detecting more precise boundaries than CNM, especially when dealing with the easy-confusing backgrounds. In Fig. 3 (b), we visualize two qualitative examples on the large-scale ActivityNet Captions dataset. The first example demonstrates a set of consecutive scenes, where several firemen are washing windows while


Query: Person drinks a glass of water from the shelf.
GT: 8.6s 17.7s
CNM: 10.08s 19.17s
DDC: 9.26s 18.52s
Query: Person pours a cup of coffee.
GT: 0.5s 9.2s
CNM: 2.67s 14.96s
DDC: 0s 10.16s
Query: Clips are shown of men hanging off of buildings washing windows while another man speaks to the camera.
GT: 33.16s 104.1s
CNM: 10.29s 136.12s
DDC: 21.33s 121.41s
Query: He plays the instrument on his shoulders and continues to demonstrate how to play.
GT: 37.57s 126.79 s
CNM: 62.58s 156.53s
DDC: 41.19s 132.33 s
Query: Person putting pillow aside.
Reconstruction: Person placing the cushion aside.
GT: 5.9s 11.7s
CNM: 7.09s 13.68s
DDC: 4.66s 12.71s
Query: The man begins to rock climb while being holstered.
Reconstruction: The man starts climbing rocks while holstered.
GT: 60.29s 145.29s
CNM: 26.74s 145.29s
DDC: 35.20s 145.29s

(a) Examples of different methods on the Charades-STA dataset.

(b) Examples of different methods on the ActivityCaptions dataset.

(c) Examples about query reconstruction and moment prediction on the Charades-STA and ActivityCaptions datasets.

Fig. 3. Qualitative visualization on both two datasets, i.e., Charades-STA and ActivityCaptions. The horizontal axis denotes the timestamps.

another man is speaking to the camera. As we can see, even though the backgrounds are diverse and the language description is complicated, our model successfully localizes the entire salient moment and suppress the false positive predictions. The second example demonstrates a "*play the instrument*" action observed with highly-confusing backgrounds, leading to inaccurate predictions with the CNM model. Our DDCNet, however, still performs well in this case except for a few failures in the end. In addition, we simultaneously present the reconstruction and prediction results to better reveal the rationale behind our DDCNet. We show two examples from both Charades-STA and ActivityNet Captions in Fig. 3 (c). As expected, we observe that our DDCNet achieves higher IoU results between the predicted moment and the ground truth, and the reconstructed query is also close to the original one. This demonstrates that our denoised dual-level contrastive framework captures more fine-grained semantic information inside and outside the video to reconstruct the query, leading to more complete and robust predictions.

Furthermore, we also visualize the frame-by-word attention to understand the cross-modal interaction process. As a fundamental component in temporal sentence grounding, this type of visualization also helps our model explain how frame-by-word attention works when reconstructing the original sentence. As shown in Fig. 4, the correlation of the pair of frame and word representations is displayed, where the darker color repre-

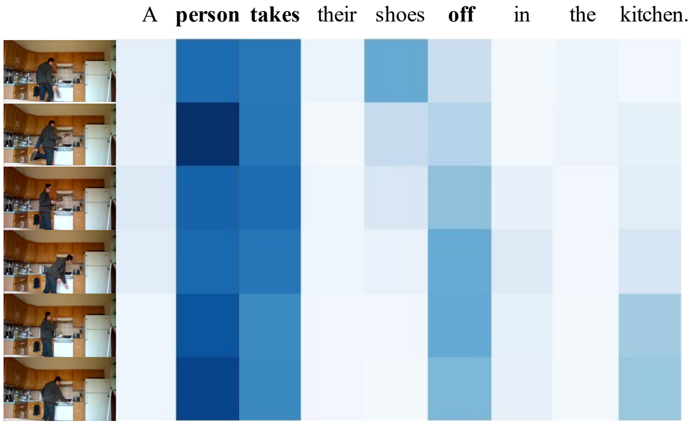


Fig. 4. Visualization results of the frame-by-word attention. The darker the color is, the larger the related attention value is.

sents a higher correlation. The typical case depicts that our DDCNet tends to seek more semantically related words in the sentence while ignoring other irrelevant words with subtle information. For instance, the 4-th positional frame is focused on the semantic correlated words “person”, “takes off” and neglects the remote irrelative words like “their” and “the”. This suggests our DDCNet is able to capture the semantic connections between visual and text representations, thereby leading to more accurate predictions.

5 Conclusion

In this paper, we propose a Denoised Dual-level Contrastive Network, DDCNet, for weakly-supervised temporal sentence grounding. Our method aims to encourage the completeness and robustness of the predicted moment. Specifically, we present a dual-level contrastive learning strategy to enable the completeness and robustness of the predicted moments. Then a ranking weight strategy based on the feature similarity is devised to guide the selection of positive and negative proposals. Furthermore, we introduce an effective pseudo-label denoised process to alleviate the false activations, which can ease the model training and enables DDCNet to predict more accurate localizations. The experiments are conducted on two publicly available datasets, namely Charades-STA and ActivityNet Captions, demonstrating the effectiveness and superiority of our DDCNet when compared with existing weakly-supervised methods.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) (Grant 62376265).

References

1. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE international conference on computer vision. pp. 5803–5812 (2017)
2. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
3. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV). pp. 132–149 (2018)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
5. Chen, J., Luo, W., Zhang, W., Ma, L.: Explore inter-contrast between videos via composition for weakly supervised temporal sentence grounding **36**(01), 267–275 (2022)
6. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.S.: Temporally grounding natural sentence in video. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp. 162–171 (2018)
7. Chen, S., Jiang, Y.G.: 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. pp. 8425–8435 (2021)
8. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
9. Chen, Z., Ma, L., Luo, W., Tang, P., Wong, K.Y.K.: Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. arXiv preprint arXiv:2001.09308 (2020)
10. Collins, R.T., Lipton, A.J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P., et al.: A system for video surveillance and monitoring. VSAM final report **2000**(1-68), 1 (2000)
11. Da, C., Zhang, Y., Zheng, Y., Pan, P., Xu, Y., Pan, C.: Asynce: Disentangling false-positives for weakly-supervised video grounding. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1129–1137 (2021)
12. Duan, X., Huang, W., Gan, C., Wang, J., Zhu, W., Huang, J.: Weakly supervised dense event captioning in videos. Advances in Neural Information Processing Systems **31** (2018)
13. Fang, Z., Kong, S., Wang, Z., Fowlkes, C., Yang, Y.: Weak supervision and referring attention for temporal-textual association learning. arXiv preprint arXiv:2006.11747 (2020)
14. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
15. Gao, M., Davis, L.S., Socher, R., Xiong, C.: Wslln: Weakly supervised natural language localization networks. arXiv preprint arXiv:1909.00239 (2019)
16. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
17. Islam, A., Radke, R.: Weakly supervised temporal action localization using deep metric learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 547–556 (2020)

18. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017)
21. Lin, Z., Zhao, Z., Zhang, Z., Wang, Q., Liu, H.: Weakly-supervised video moment retrieval via semantic completion network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11539–11546 (2020)
22. Loper, E., Bird, S.: Nltk: The natural language toolkit. arXiv preprint cs/0205028 (2002)
23. Luo, F., Chen, S., Chen, J., Wu, Z., Jiang, Y.G.: Self-supervised learning for semi-supervised temporal language grounding. arXiv preprint arXiv:2109.11475 (2021)
24. Ma, M., Yoon, S., Kim, J., Lee, Y., Kang, S., Yoo, C.D.: Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In: European conference on computer vision. pp. 156–171. Springer (2020)
25. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: Proceedings of the tenth ACM international conference on Multimedia. pp. 533–542 (2002)
26. Mithun, N.C., Paul, S., Roy-Chowdhury, A.K.: Weakly supervised video moment retrieval from text queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11592–11601 (2019)
27. Nam, J., Ahn, D., Kang, D., Ha, S.J., Choi, J.: Zero-shot natural language video localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1470–1479 (2021)
28. Pan, T., Song, Y., Yang, T., Jiang, W., Liu, W.: Videomoco: Contrastive video representation learning with temporally adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11205–11214 (2021)
29. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
30. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision. pp. 510–526. Springer (2016)
31. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems **29** (2016)
32. Song, Y., Wang, J., Ma, L., Yu, Z., Yu, J.: Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. arXiv preprint arXiv:2003.07048 (2020)
33. Tan, R., Xu, H., Saenko, K., Plummer, B.A.: 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. pp. 2083–2092 (2021)
34. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: European conference on computer vision. pp. 776–794. Springer (2020)
35. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

37. Wang, Y., Deng, J., Zhou, W., Li, H.: Weakly supervised temporal adjacent network for language grounding. IEEE Transactions on Multimedia (2021)
38. Wang, Y., Zhou, W., Li, H.: Fine-grained semantic alignment network for weakly supervised temporal language grounding. arXiv preprint arXiv:2210.11933 (2022)
39. Wang, Z., Chen, J., Jiang, Y.G.: Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1459–1468 (2021)
40. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
41. Xu, H., He, K., Plummer, B.A., Sigal, L., Sclaroff, S., Saenko, K.: Multilevel language and vision integration for text-to-clip retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9062–9069 (2019)
42. Xu, Y., Cao, P., Kong, Y., Wang, Y.: L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. Advances in neural information processing systems **32** (2019)
43. Yang, W., Zhang, T., Zhang, Y., Wu, F.: Local correspondence network for weakly supervised temporal sentence grounding. IEEE Transactions on Image Processing **30**, 3252–3262 (2021)
44. Yuan, Y., Mei, T., Zhu, W.: To find where you talk: Temporal sentence localization in video with attention based location regression. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9159–9166 (2019)
45. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12870–12877 (2020)
46. Zhang, Z., Lin, Z., Zhao, Z., Zhu, J., He, X.: Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4098–4106 (2020)
47. Zhang, Z., Zhao, Z., Lin, Z., He, X., et al.: Counterfactual contrastive learning for weakly-supervised vision-language grounding. Advances in Neural Information Processing Systems **33**, 18123–18134 (2020)
48. Zheng, M., Huang, Y., Chen, Q., Liu, Y.: Weakly supervised video moment localization with contrastive negative sample mining. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 1, p. 3 (2022)
49. Zheng, M., Huang, Y., Chen, Q., Peng, Y., Liu, Y.: Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15555–15564 (2022)