Weakly Supervised Video Moment Localization with Contrastive Negative Sample Mining

Minghang Zheng[1], **Yanjie Huang**[2*], **Qingchao Chen**[3], **Yang Liu**[1,4†]

[1]Wangxuan Institute of Computer Technology, Peking University
[2]School of Integrated Circuits and Electronics, Beijing Institute of Technology
[3]National Institute of Health Data Science, Peking University
[4]Beijing Institute for General Artificial Intelligence
{minghang, qingchao.chen, yangliu}@pku.edu.cn

Abstract

Video moment localization aims at localizing the video segments which are most related to the given free-form natural language query. The weakly supervised setting, where only video level description is available during training, is getting more and more attention due to its lower annotation cost. Prior weakly supervised methods mainly use sliding windows to generate temporal proposals, which are independent of video content and low quality, and train the model to distinguish matched video-query pairs and unmatched ones collected from different videos, while neglecting what the model needs is to distinguish the unaligned segments within the video. In this work, we propose a novel weakly supervised solution by introducing Contrastive Negative sample Mining (CNM). Specifically, we use a learnable Gaussian mask to generate positive samples, highlighting the video frames most related to the query, and consider other frames of the video and the whole video as easy and hard negative samples respectively. We then train our network with the Intra-Video Contrastive loss to make our positive and negative samples more discriminative. Our method has two advantages: (1) Our proposal generation process with a learnable Gaussian mask is more efficient and makes our positive sample higher quality. (2) The more difficult intra-video negative samples enable our model to distinguish highly confusing scenes. Experiments on two datasets show the effectiveness of our method. Code can be found at https://github.com/minghangz/cnm.

Introduction

Video moment localization is an important yet challenging task with potential applications in video surveillance (Collins et al. 2000), robot manipulation (Kemp, Edsinger, and Torres-Jara 2007), etc. The goal is to localize temporally a video segment (i.e., start and end time) that best corresponds to a query sentence from untrimmed videos. Fully supervised video moment localization has witnessed remarkable progress recently (Zhao et al. 2021; Wang et al. 2021a; Zhou et al. 2021). However, annotating the ground truth temporal boundary for each query sentence is labor-intensive and time-consuming, which undermines fully supervised approaches in real-world large-scale scenarios. Therefore the weakly supervised setting, where only video-level descriptions are available during training, is more practical and draws increasing attention from the community.

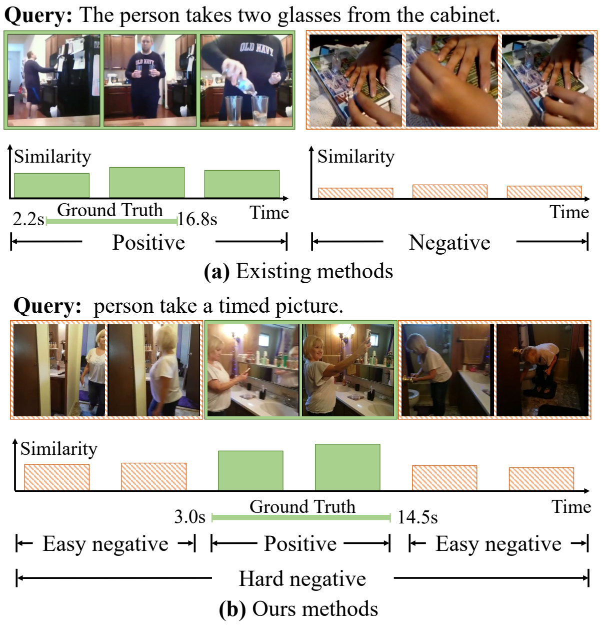


Figure 1: (a) Existing methods focus on distinguishing matched and unmatched video-query pairs (collected from different videos), while neglecting the matching level of different segments within one video. (b) We attach importance to mine positive and negative samples within the same video. We predict a learnable positive sample and consider the segments outside the positive sample as the easy negative sample, and the whole video as the hard negative sample.

However, existing weakly supervised solutions (Mithun, Paul, and Roy-Chowdhury 2019; Gao et al. 2019; Lin et al. 2020; Ma et al. 2020; Huang et al. 2021) have two limitations: *Firstly, given a specific query, they mainly focus on distinguishing different videos by calculating the semantic*

[*]This work was done during an internship at Wangxuan Institute of Computer Technology, Peking University

[†]Corresponding author

consistency between the video and the query while neglecting the matching level of different video segments within one video. Specifically, as shown in Figure 1(a), most existing multi-instance learning (MIL) based solutions learn the visual-text alignment in the video level by maximizing the matching scores of the paired sentences and videos while suppressing that of the unpaired ones. Reconstruction-based solutions (Lin et al. 2020; Song et al. 2020) solve this task through joint learning with the reconstruction mechanism, assuming that the video segment that best matches the query should best reconstruct the entire query. However, for all of them, given a specific query, all the negative samples are collected from the other videos, which is not optimal as it neglects the fact that the mismatched segments contained in the same video (as shown in Figure 1(b)) is often much harder to distinguish in the activity temporal localization setup due to similar background and video style. *Secondly, the existing proposal generation procedure is independent of the video and query sentence, which is less informative and inefficient.* The existing dominant proposal generation procedures are mainly based on sliding windows, which can not be dynamically adapted for different videos. They pool the frame features within the proposal, ignoring the inherent temporal structure of an event (i.e., the beginning, climax, and ending), and may derive semantically irrelevant visual-text relationships which are less generalizable. Moreover, to keep a high recall rate, a large number of dense proposals are required for long videos, which leads to an increase in computational complexity.

To address the above limitations, we introduce a novel weakly supervised method for activity temporal localization by dynamically generating informative proposals and mining hard negative samples within the same video for training. We call it Contrastive Negative sample Mining (CNM). Firstly, to enable deeper coupling between the proposal generation procedure and the video-level supervision, we propose to generate a learnable Gaussian mask for each video, highlighting a video segment most relevant to the query, which serves as the positive sample. It is worth noting that our Gaussian mask can represent the temporal structure of an event and can be learned end-to-end. Secondly, to mine the negative samples within the video, we treat the video segments that are not highlighted by Gaussian mask (marked by shaded orange segments in Figure 1(b)) as easy negative samples. The whole video also serves as a hard negative sample as it often contains a lot of redundant information. We propose an Intra-Video Contrastive (IVC) Loss to ensure that the similarity with the query is sorted from large to small as positive, hard negative, and easy negative samples. By incorporating all of them into the training pipeline, we can learn a temporal-sensitive visual embedding for temporal localization and boost the performance.

To sum up, the main contributions of our work are:

- We propose to generate a Gaussian mask as a proposal, which can represent the temporal structure of an event and can be learned by the network.
- In contrast to collecting negative samples from different videos, we propose to mine the hard and easy negatives within the same video and train the system with Intra-Video Contrastive loss. Training with such a negative mining scheme makes our network capable of distinguishing highly confusing scenes.
- Experiments on the ActivityNet Captions (Caba Heilbron et al. 2015) and Charades-STA (Gao et al. 2017) datasets demonstrate the effectiveness of our method in weakly supervised video moment localization.

Related Work

Fully supervised video moment localization. In the fully supervised setting, the annotations of precise start and end timestamps for each video and query pair are available during training. The method mentioned in (Gao et al. 2017) uses a fully-connected layer to join the sentence and video feature together. The 2D Temporal Adjacent Networks (2D-TAN) (Zhang et al. 2020a) takes advantage of the feature of frames nearby. In addition, Boundary Proposal Net (BP-Net) (Xiao et al. 2021) fuses the generated segment-level feature and the query feature through multi-model fusion. In the work of (Rodriguez-Opazo et al. 2020), it constructs the Spatio-Temporal Graph, finding the relationship between object and human nodes. Multi-stage Aggregated Transformer Network (MSA) (Zhang et al. 2021) tires to utilize the feature of not only the start and end timestamps, but also the middle of the frame. The method proposed in (Zhou et al. 2021) uses the K-means algorithm to inference. Dual Path Interaction Network (DPIN) (Wang et al. 2020) and Structured Multi-Level Interaction Network (SMIN) (Wang et al. 2021a) build structured multi-level interaction module to optimize the use of the logic relationship between the query and the segment. However, such fully supervised methods need a huge amount of time and labor for annotation, limiting their scalability and practicability.

Weakly supervised video moment localization. Compared with the supervised setting, only video and query pairs are given in the weakly supervised setting. Firstly, some methods like the weakly supervised Semantic Completion Network (SCN) (Lin et al. 2020) introduce the reconstruction mechanism, asserting that a video segment paired with the query could reconstruct the sentence better. However, those reconstruction-based methods ignore the information from unmatched videos and queries for contrastive learning. Further, other works (Yang et al. 2021; Huang et al. 2021; Mithun, Paul, and Roy-Chowdhury 2019) utilize the Multi-Instance Learning (MIL) method, consider non-aligned video-query pairs from other videos as negative samples, and train the model to distinguish them from aligned ones through specially designed loss functions. However, to those MIL-based methods, their negative samples are not difficult enough for the model to distinguish, making the model unable to effectively distinguish highly confusing scenes within the video, because the contents of different videos are visually distinct. In our method, we not only use the reconstruction mechanism but also use negative samples for contrastive learning, and the SCN method functions as our baseline. We collect the easy negative sample outside the positive sample within the same video, and consider the whole video as the hard negative sample, thus adding difficulty during training, which enables our network to distinguish highly confusing scenes.

Secondly, the methods mentioned in (Mithun, Paul, and Roy-Chowdhury 2019; Chen et al. 2020; Huang et al. 2021) all use sliding windows to generate proposals. However, the proposals generated by these methods are not related to the content of the video. During training, these models would generate a large number of redundant proposals and use Non-Maximum Suppression (NMS) (Neubeck and Van Gool 2006) for post-processing, which involves heavy computational cost. In our method, we introduce the learnable Gaussian masks to help us generate positive samples, saving the labor of generating a large number of proposals through the sliding windows.

Approach

The overall framework of CNM is illustrated in Fig. 2(a). It consists of a mask generator and a mask conditioned reconstructor. In the mask generator, we fuse the multi-modal information of video and language to predict a Gaussian mask, highlighting a video segment most semantically relevant to the query, which serves as the positive sample. The Gaussian mask can be viewed as a high-quality content-based temporal proposal that can be learned end-to-end. To enable the model to distinguish highly confusing scenes, we mine negative samples within the same video. We treat the video segments that are not highlighted by the Gaussian mask as easy negative samples. Since the entire video contains a lot of redundant information, we also treat it as a hard negative sample. In the mask conditioned reconstructor, we use the reconstruction performance as a measurement of semantic similarity of the query, assuming that the segment that perfectly matches the query can better reconstruct the entire query. To make our reconstructor differentiable to the mask, we design the mask conditioned attention in Fig. 2(b), which collects contextual information within the video segment highlighted by the mask and uses the fused multi-modal information to reconstruct the query. Our mask conditioned attention weights the attention map by the values in the Gaussian mask, which prevents the leakage of frame features outside the mask. Finally, we optimize our mask conditioned reconstructor with the reconstruction loss $\mathcal{L}_{rec}$ for a better reconstruction, and optimize the mask generator with the Intra-Video Contrastive loss $\mathcal{L}_{IVC}$ by requiring the reconstruction results of the positive sample, hard negative sample, and easy negative sample are from good to bad.

Mask Generator

To generate high-quality and content-based proposals, our mask generator fuses information from two modalities of vision and language and predicts a Gaussian mask as our positive sample. Different from previous works, which use sliding windows to generate proposals, our Gaussian mask is learnable and can characterize the inherent temporal structure of events (begining, climax, and ending). To enable our model to distinguish highly confusing scenes, we mine negative samples within the same video: the video frames outside the Gaussian mask are regarded as easy negative samples, and the whole video, which contains a lot of irrelevant redundant information, is regarded as a hard negative sample.

Feature Extraction. We first encode the videos and queries into feature vectors. Specifically, each word of the query is embedded using GloVe (Pennington, Socher, and Manning 2014) and the query is represented as $W = \{w_1, w_2, ..., w_M\} \in \mathbb{R}^{M \times D_W}$, where M is the number of words and D_W is the word feature dimension. The video is sampled as images at a fixed frame rate, and each image is independently encoded by Pre-trained vision backbone networks. The video is represented as $V = \{v_1, v_2, ..., v_N\} \in \mathbb{R}^{N \times D_V}$, where N is the number of video frames and D_V is the video feature dimension. During training, the word embedding and vision backbone networks are frozen.

Mask Generation. As an event usually includes a beginning, a climax and an ending, we propose to use the Gaussian mask as the proposal to characterize the inherent temporal structure of events. Because transformer (Vaswani et al. 2017) has achieved great success in sequence analysis, we use it to handle the multi-modal interaction of video sequence and text sequence, and obtain the fused features $H = \{h_1, h_2, ..., h_N\}$ that incorporate semantic and vision information:

$$H = \mathrm{D}(V, \mathrm{E}(W)) \in \mathbb{R}^{N \times D_H} \tag{1}$$

where $\mathrm{E}(\cdot)$ is the transformer encoder, $\mathrm{D}(\cdot)$ is the transformer decoder, D_H is the hidden feature dimension. Since h_N combines all the frame and word features, we predict our Gaussian center c and width w through h_N:

$$c = \mathrm{Sigmoid}(\mathrm{FC}(h_N)) \in \mathbb{R} \tag{2}$$

$$w = \mathrm{Sigmoid}(\mathrm{FC}(h_N)) \in \mathbb{R} \tag{3}$$

where $\mathrm{FC}(\cdot)$ denotes a single layer fully connected network. The video segment with center c and width w is our positive sample, and the corresponding positive Gaussian mask m^p is formulated as:

$$m_i^p = \exp(-\frac{\alpha(i/N - c)^2}{w^2}), i = 1, ..., N \tag{4}$$

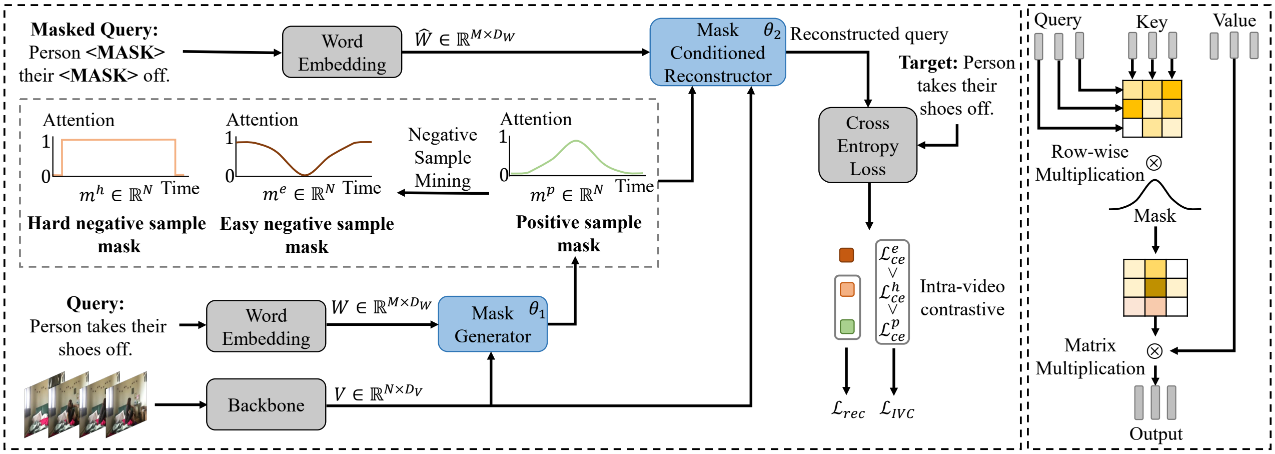
where m_i^p is the weight of the i-th video frame in the Gaussian mask, and α is a hyperparameter that controls the variance of the Gaussian function.

Negtive Sample Mining. To enable our model to distinguish highly confusing scenes, we mine negative samples within the same video. Those negative samples inside the video are what the model needs to distinguish during inference. Compared with other methods that simply use other unmatched videos as negative samples, our negative samples can provide richer information for the model.

Firstly, we regard the frame suppressed by m^p as an easy negative sample m^e, expressed in the form of the mask as:

$$m^e = 1 - m^p \in \mathbb{R}^N \tag{5}$$

The easy negative sample is composed of frames in the video that are not related to the query, but those frames may be confusing because they have similar background and semantics to the positive sample. Training the model to distinguish


Masked Query:
Person <MASK> their <MASK> off.
Word Embedding
$\hat{W} \in \mathbb{R}^{M \times D_W}$
Mask Conditioned Reconstructor θ_2
Reconstructed query
Target: Person takes their shoes off.
Cross Entropy Loss
Attention
1
0
$m^h \in \mathbb{R}^N$ Time
Hard negative sample mask
$m^e \in \mathbb{R}^N$ Time
Easy negative sample mask
Negative Sample Mining
$m^p \in \mathbb{R}^N$ Time
Positive sample mask
$\mathcal{L}_{ce}^{e}$ < $\mathcal{L}_{ce}^{h}$ < $\mathcal{L}_{ce}^{p}$
Intra-video contrastive
Query:
Person takes their shoes off.
$W \in \mathbb{R}^{M \times D_W}$
Mask Generator θ_1
Backbone
$V \in \mathbb{R}^{N \times D_V}$
$\mathcal{L}_{rec}$
$\mathcal{L}_{IVC}$
Query
Key
Value
Row-wise Multiplication
Mask
Matrix Multiplication
Output

(a) Overall Framework

(b) Mask Conditioned Attention

Figure 2: The framework of our method in weakly supervised video moment localization. In Fig. 2(a), the mask conditioned generator fuses the information from the video and query and predicts a Gaussian mask, which highlights our positive sample. The video segments not highlighted by the Gaussian mask are considered as easy negative samples, and the whole video serves as the hard negative sample. The mask conditioned reconstructor uses the reconstruction results as a measurement for the semantic relevance between the query and positive and negative samples. We optimize our mask conditioned reconstructor with the reconstruction loss $\mathcal{L}_{rec}$ for a better reconstruction, and optimize the mask generator with the Intra-Video Contrastive loss $\mathcal{L}_{IVC}$ by requiring the reconstruction results of the positive sample, hard negative sample, and easy negative sample are from good to bad. In Fig. 2(b), to keep our reconstructor differentiable to the mask, we introduce the mask conditioned attention, which weights the attention map by the mask and collects contextual information within the frames highlighted by the mask.

the easy negative sample from the positive sample can improve the performance in highly confusing scenes.

Secondly, in most cases, the entire video can also be regarded as a negative example, because it contains a lot of irrelevant redundant information that has nothing to do with the query. Thus, we regard the entire video as a hard negative sample m^h, denoted as:

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

The hard negative sample is composed of the positive sample as well as a lot of irrelevant video frames and is more difficult for the model to distinguish. Training the model to distinguish the hard negative sample from the positive sample can help the model to locate more accurately and prevent the model from outputting longer predictions that include the ground truth.

Since the hard negative sample contains the positive sample as well as a lot of irrelevant redundant information and the easy negative sample does not contain any correct segment, the semantic relativity of the three samples and the query should satisfy:

$$\mathrm{R}(m^p, W) > \mathrm{R}(m^h, W) > \mathrm{R}(m^e, W) \tag{7}$$

where $\mathrm{R}(\cdot)$ is a function to evaluate the relevance of the query W and the video segment represented by the mask m, which will be discussed in the following section.

Mask Conditioned Reconstructor

Inspired by SCN, our mask conditioned reconstructor reconstructs the original query conditioned on arbitrary sample masks, the results of which serve as a measurement of the semantic similarity between positive/negative samples and the query. To keep our reconstructor differentiable to the mask, we introduce the mask conditioned attention, which weights the attention map by the mask and collects contextual information within the frames highlighted by the mask. Our reconstructor uses the standard transformer structure and replaces the traditional attention with our mask conditioned attention. To optimize our generated mask end-to-end, we design the Intra-Video Contrastive loss $\mathcal{L}_{IVC}$ which requires the reconstruction results of the positive sample, hard negative sample, and easy negative sample are from good to bad. To optimize our mask conditioned reconstructor, we use the reconstruction loss $\mathcal{L}_{rec}$ to minimize the cross-entropy loss of the reconstructed query and the original query.

Mask Conditioned Attention. Our reconstructor uses the standard transformer structure to perform multi-modal interaction. To keep the reconstructor differentiable to the mask, we introduce the mask conditioned attention in Fig. 2(b). We replace the conventional attention mechanism (Vaswani et al. 2017) in the transformer with our mask conditioned attention and keep other components unchanged. Our mask conditioned reconstructor includes an encoder $E_m(\cdot)$ and a decoder $D_m(\cdot)$, which can handle arbitrary mask as input and limit the attention to the frames highlighted by the mask.

The encoder $E_m(\cdot)$ takes the mask $m \in \mathbb{R}^N$ and vision features $V \in \mathbb{R}^{N \times D_V}$ as inputs, and exchange information within the frame features highlighted by the mask. We first project V to the attention queries $Q_a \in \mathbb{R}^{N \times D_H}$, keys $K_a \in \mathbb{R}^{N \times D_H}$ and values $V_a \in \mathbb{R}^{N \times D_H}$ with a fully connected layer. Then we calculate the similarity between Q_a and K_a and obtain the attention map $A = \frac{Q_a K_a}{\sqrt{D_H}} \in \mathbb{R}^{N \times N}$. To limit the attention to the frames highlighted by the mask m, we multiply m on each row of A. After a row-wise Softmax operation, the attention map is multiplied by V_a to output the aggregated context information:

$$E_m(V, m) = \text{Softmax}(A \otimes m)V_a \in \mathbb{R}^{N \times D_H} \tag{8}$$

where $\otimes$ means that m will be multiplied to each row of A, and the Softmax operation is applied in each row. The decoder $D_m(\cdot)$ takes the mask m, the query features W, and the outputs of $E_m(\cdot)$ as inputs, and collects contextual information for each word feature within the frame features highlighted by the mask. The calculation of $D_m(\cdot)$ is similar to that of $E_m(\cdot)$ except that the attention queries Q_a are projected from the query W, and the keys K_a and values V_a are projected from the outputs of $E_m(\cdot)$.

Mask Conditioned Semantic Completion. To measure the semantic relevance of positive and negative samples to the query, we use our mask conditioned reconstructor to reconstruct the query conditioned on the frames highlighted by the mask, assuming that the frames that perfectly match the query can better reconstruct the entire query.

Following SCN, we randomly replace 1/3 of the words in the original query with a specific symbol, where nouns, verbs, and adjectives have a higher probability of being replaced. We denote $\hat{W}$ as the masked query embedded using GloVe (Pennington, Socher, and Manning 2014). Then we use our mask conditioned attention to obtain the cross-modal semantic representation H^p conditioned on the positive sample mask m^p:

$$H^p = D_m(\hat{W}, E_m(V, m^p), m^p) \in \mathbb{R}^{M \times D_H} \tag{9}$$

Then, a single fully connected layer is applied to H^p and outputs the probability distribution P^p of the next word on the vocabulary conditioned on the positive mask:

$$P^p(\widetilde{w}_{i+1} | V, \hat{W}_{1:i}) = \text{Softmax}(\text{FC}(H^p)) \in \mathbb{R}^{M \times N_w} \tag{10}$$

where FC$(\cdot)$ is a fully connected layer, and N_w is the vocabulary size. Then we use the cross-entropy loss to calculate the difference between P^p and the real distribution:

$$\mathcal{L}^p_{ce} = -\sum_{i=1}^{M-1} \log P^p(w_{i+1} | V, \hat{W}_{1:i}) \tag{11}$$

Similarly, we can get $\mathcal{L}^e_{ce}$ and $\mathcal{L}^h_{ce}$ by replacing m^p with m^e and m^h respectively. Because only the positive sample and the entire video (hard negative sample) contain the segment related to the query, only they can reconstruct the query in principle. So only the $\mathcal{L}^p_{ce}$ and $\mathcal{L}^h_{ce}$ will participate in the optimization of the mask conditioned reconstructor. The final reconstruction loss $\mathcal{L}_{rec}$ is formulated as:

$$\mathcal{L}_{rec} = \mathcal{L}^p_{ce} + \mathcal{L}^h_{ce} \tag{12}$$

Intra-Video Contrastive. To optimize our mask generator, we train our model to distinguish the positive and negative samples. As shown in (7), the semantic similarity between the query and the positive, hard negative, and easy negative samples should satisfy a certain relationship. Similar to margin ranking loss (Balntas et al. 2016), our Intra-Video Contrastive loss L_{IVC} can be formulated as:

$$\begin{aligned} \mathcal{L}_{IVC} = &\max(\mathcal{L}^p_{ce} - \mathcal{L}^h_{ce} + \beta_1, 0) + \\ &\max(\mathcal{L}^p_{ce} - \mathcal{L}^e_{ce} + \beta_2, 0) \end{aligned} \tag{13}$$

where β_1 and β_2 are hyperparameters satisfying $\beta_1 < \beta_2$. L_{IVC} requires that the loss of the positive sample is at least β_1 smaller than the loss of the hard negative sample, and at least β_2 smaller than the loss of the easy negative sample.

Model Training and Inference

In this section, we describe the loss function we optimize to train our network and our inference process. Our network mainly includes two parts of loss: the reconstruction loss $\mathcal{L}_{rec}$ is used to optimize the mask conditioned reconstructor, which encourages the network to accurately predict the description related to the given mask; the Intra-Video Contrastive loss L_{IVC} is used to optimize the mask generator, which encourages the network to choose an appropriate Gaussian mask to make the positive and negative samples more distinguishable.

Training. To require the reconstructor to try its best to reconstruct the query from a video segment regardless of whether it is positive or negative, the IVC loss is only used to train the mask generator, and the reconstruction loss is only used to train the mask conditioned reconstructor. Specifically, we first update the reconstructor by $\mathcal{L}_{rec}$, while freezing the mask generator; Then we update the mask generator by $\mathcal{L}_{IVC}$ while freezing the reconstructor:

$$\begin{aligned} \hat{\theta}_1 &= \arg\min_{\theta_1} \mathcal{L}_{IVC}(V, W | \theta_1, \theta_2) \\ \hat{\theta}_2 &= \arg\min_{\theta_2} \mathcal{L}_{rec}(V, W | \theta_1, \theta_2) \end{aligned} \tag{14}$$

where θ_1 is the parameters of the mask generator, and θ_2 is the parameters of the mask conditioned reconstructor. This design can avoid a trivial solution where the reconstructor always gives low scores to the predicted negative samples, which will easily accumulate errors at early training.

Inference. The inference process of our model is very simple. Through Equation (2) and (3), we can obtain the center c and width w of our predicted Gaussian mask. The temporal boundary (s, e) can be obtained by:

$$\begin{aligned} s &= \max(c - w/2, 0) * N \\ e &= \min(c + w/2, 1) * N \end{aligned} \tag{15}$$

Since there is no need to use sliding windows to generate dense proposals, our model abandons complex post-processing operations such as Non-Maximum Suppression (NMS) (Neubeck and Van Gool 2006).

Experiments

Datasets

To test the effectiveness of our proposed method, we perform experiments on two publicly available datasets, ActivityNet Captions (Caba Heilbron et al. 2015; Krishna et al. 2017) and Charades-STA (Gao et al. 2017), respectively.

ActivityNet Captions. ActivityNet Captions dataset is released in (Krishna et al. 2017), which is made up of 19,290 videos with 37,417/17,505/17,031 moment of interests (MoIs) in the train/val_1/val_2 split. The length of the query, on average, is 14 words. The length of the MoIs and untrimmed videos are 36.2 and 117.6 seconds respectively. We adopt standard splits, and follow the common practice of the previous works SCN (Lin et al. 2020) and RTBPN (Zhang et al. 2020b), using the val_1 split for validation, val_2 split for testing.

Charades-STA. Charades-STA (Gao et al. 2017) dataset contains 12,408/3,720 video-query pairs. For training, 5,338 videos are available, while 1,334 videos can be used for testing. The query sentences are comprised of 7.2 words on average, while the average duration of target video moments and untrimmed videos are 8.1 and 30.6 seconds respectively, which is much shorter compared with those in the ActivityNet Captions dataset. We report our results on the test split.

Evaluation Metric

Following the previous work (Lin et al. 2020), we choose the result of 'IoU=m' as our evaluation metric. To be specified, this metric evaluates the percentage of predicted moments that have the temporal Intersection over Union (IoU) larger than the threshold m, and m is set to $\{0.1, 0.3, 0.5\}$ on the ActivityNet Captions dataset, and $\{0.3, 0.5, 0.7\}$ on the Charades-STA dataset respectively.

Implementation Details

Data Preprocessing. For each video, we pre-extract its visual features using the CLIP (Radford et al. 2021) for ActivityNet Captions and using I3D (Carreira and Zisserman 2017) for Charades-STA. We use the pre-trained GloVe (Pennington, Socher, and Manning 2014) word2vec for each word token to extract word embeddings. We set the maximum description length to 20, set the maximum number of frames to 200, and the vocabulary size for the ActivityNet Caption and Charades-STA is 8,000 and 1,111 respectively.

Model Settings. For the transformer in the mask generator and mask conditioned reconstructor, the dimension of their hidden state is 256, the number of attention heads is 4, and the number of layers is 3. During training, we use Adam (Ellouz et al. 1974) optimizer with the learning rate set to 0.0004. The hyperparameters β_1, β_2 are set to 0.1, 0.15 respectively for both datasets. α is set to 5 for ActivityNet captions and 5.5 for Charades-STA. Due to the shorter ground truth length on Charades-STA, we limit the maximum width of the prediction to 0.45 (multiplied by Eq. (3)).

Comparisons to the State-Of-The-Art

Tab. 1 and 2 compare CNM with previous works of weakly supervised video moment localization.

Table 1: Evaluation Results on the ActivityNet Captions Dataset ($m \in \{0.1, 0.3, 0.5\}$). The numbers in bold are the best result, and the numbers underlined are the second best.

Method	Recall		
	IoU=0.1	IoU=0.3	IoU=0.5
Random (Lin et al. 2020)	38.23	18.64	7.63
WS-DEC (Duan et al. 2018)	62.71	41.98	23.34
EC-SL (Chen and Jiang 2021)	68.48	44.29	24.16
MARN (Song et al. 2020)	-	47.01	29.95
SCN (Lin et al. 2020)	71.48	47.23	29.22
RTBPN (Zhang et al. 2020b)	73.73	49.77	29.63
WSTG (Chen et al. 2020)	74.2	44.3	23.6
WSLLN (Gao et al. 2019)	75.4	42.8	22.7
LCNet (Yang et al. 2021)	78.58	48.49	26.33
WSTAN (Wang et al. 2021b)	79.78	52.45	30.01
CRM [1] (Huang et al. 2021)	**81.61**	55.26	32.19
CNM (ours)	78.13	**55.68**	**33.33**

Table 2: Evaluation Results on the Charades-STA Dataset ($m \in \{0.3, 0.5, 0.7\}$). The numbers in bold are the best result, and the numbers underlined are the second best.

Method	Recall		
	IoU=0.3	IoU=0.5	IoU=0.7
Random (Lin et al. 2020)	20.12	8.61	3.39
TGA (Mithun, Paul, and Roy-Chowdhury 2019)	32.14	19.94	8.84
WSTG (Chen et al. 2020)	39.8	27.3	12.9
SCN (Lin et al. 2020)	42.96	23.58	9.97
WSTAN (Wang et al. 2021b)	43.39	29.35	12.28
VLANet (Ma et al. 2020)	45.24	31.83	14.17
LoGAN (Tan et al. 2021)	48.04	31.74	13.71
MARN (Song et al. 2020)	48.55	31.94	14.81
CRM [1] (Huang et al. 2021)	53.66	34.76	16.37
LCNet (Yang et al. 2021)	59.60	**39.19**	**18.87**
RTBPN (Zhang et al. 2020b)	60.04	32.36	13.24
CNM (ours)	**60.04**	35.15	14.95

We observe: (1) Our method surpasses the state-of-the-art methods for IoU=0.3 and IoU=0.5 on the ActivityNet Captions dataset. (2) On the ActivityNet Captions dataset with IoU=0.1, the result of our method is slightly lower than the CRM. However, CRM requires a paragraph description annotation (multiple events described sequentially) per video in training, while we do not. Paragraph-level descriptions provide additional event timing information but are not always available in practical applications. Thus, our approach addresses a more practical problem, i.e., uses less annotation information in training and achieves comparable performance to CRM. (3) On the Charades-STA dataset, our CNM achieves SOTA when IoU=0.3. However, there are still some gaps of CNM with SOTA when IoU=0.5 and 0.7. One possible reason is that influenced by reconstruction loss, our model tends to output longer predictions because it is more likely to contain the correct information for a better reconstruction.

[1] Directly comparing CRM with other methods (include our CNM) is unfair, because CRM requires an additional paragraph description annotation (multiple events described sequentially) per video in training, while we do not.

Table 3: The effectiveness of masks generator

Method	Recall			
	IoU=0.1	IoU=0.3	IoU=0.5	mIoU
Full Model	78.13	**55.68**	**33.33**	**37.14**
w/o. Mask	**79.35**	47.71	26.98	34.73

Table 4: The effect of intra-video negative samples mining.

Hard	Easy	Recall			
		IoU=0.1	IoU=0.3	IoU=0.5	mIoU
✓	✓	78.13	**55.68**	**33.33**	**37.14**
✗	✓	80.60	55.67	31.40	36.79
✓	✗	**80.99**	55.19	30.94	36.95
✗	✗	62.27	40.26	24.93	28.55

Ablation Study

We conduct ablation studies to investigate the sensitivity of the proposed CNM to different design choices. All the results are reported on the ActivityNet Caption dataset. The 'mIoU' in the tables below stands for the average IoU result.

Effect of Mask Generator. As Tab. 3 shows, we evaluate the effectiveness of our mask generator. We disable our mask generator during training. Instead, we use sliding windows and policy gradient method (Lin et al. 2020), just as the SCN does. We can see that the model with the mask generator performs better, especially when IoU=0.3 and IoU=0.5, revealing the fact that the design of the mask generator plays an important role in improving the performance of our method. It is because our Gaussian masks can be learned from end to end, and are more closely related to the content of the video. In addition, our mask generator allows us to abandon dense proposals, greatly improving the speed of inference. On one NVIDIA TITAN X, we can achieve the speed of 55.8ms per video on the ActivityNet Captions dataset, while the speed of SCN is 124ms per video.

Effect of Intra-Video Negative Samples Mining. As Tab. 4 shows, we evaluate the effectiveness of our negative samples. We compare the output results of the model without easy and hard negative samples. We can see that the model with both the easy and hard negative samples achieves the best result. This demonstrates that both easy and hard negative samples are essential for our training. It is because our negative samples are harder for the model to learn, providing stronger supervision signals.

Effect of Training Strategy. As Tab. 5 shows, we evaluate the effectiveness of our training strategy. In the first row, CNM uses $\mathcal{L}_{IVC}$ to optimize the mask generator, and uses $\mathcal{L}_{rec}$ to optimize the mask conditioned reconstructor respectively. In the second row, we show the results of using $\mathcal{L}_{rec}$ and $\mathcal{L}_{IVC}$ to jointly train the entire model. We can see that CNM is better in the three evaluation metrics, which proves that $\mathcal{L}_{rec}$ and $\mathcal{L}_{IVC}$ work best when used in the mask conditioned reconstructor and the mask generator separately. This design can avoid a trivial solution that the reconstructor always gives low scores to the predicted negative samples, which will easily accumulate errors at early training.

Table 5: The effectiveness of training strategy

Method	Recall			
	IoU=0.1	IoU=0.3	IoU=0.5	mIoU
$\min_{\theta_1} \mathcal{L}_{IVC} + \min_{\theta_2} \mathcal{L}_{rec}$	**78.13**	**55.68**	**33.33**	**37.14**
$\min_{\theta_1,\theta_2}(\mathcal{L}_{IVC} + \mathcal{L}_{rec})$	63.59	43.80	24.50	28.96

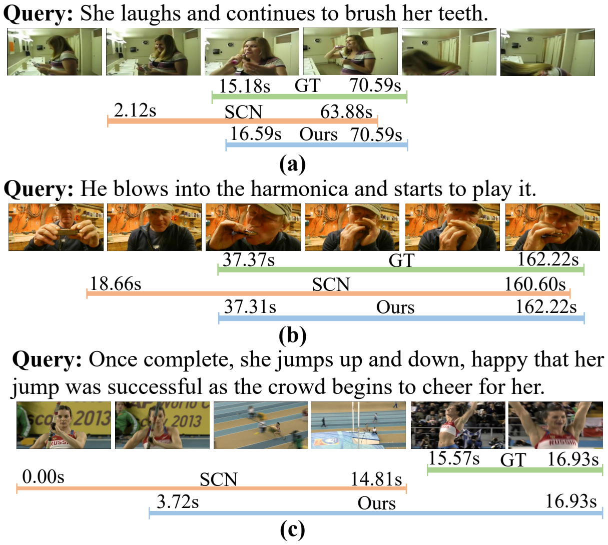

Query: She laughs and continues to brush her teeth.
15.18s
GT
70.59s
2.12s
SCN
63.88s
16.59s
Ours
70.59s
(a)
Query: He blows into the harmonica and starts to play it.
37.37s
GT
162.22s
18.66s
SCN
160.60s
37.31s
Ours
162.22s
(b)
Query: Once complete, she jumps up and down, happy that her jump was successful as the crowd begins to cheer for her.
15.57s
GT
16.93s
0.00s
SCN
14.81s
3.72s
Ours
16.93s
(c)

Figure 3: Qualitative examples on ActivityNet Captions.

Qualitative Results

Fig. 3 shows some qualitative examples from the ActivityNet Captions dataset. Each example provides the query, and the temporal boundaries of ground truth ('GT'), the SCN method, and ours respectively. It can be observed in Fig. 3(a) and (b) that the temporal boundaries of the prediction made by our method are more accurate than the SCN's, proving that our negative samples are beneficial for mask reconstruction during training. Fig. 3(c) demonstrates that our methods can still achieve better results when the SCN method goes wrong. In addition, Fig. 3(c) shows that our method may tend to output longer predictions because it is more likely to contain the correct information for a better reconstruction.

Conclusion

In this work, we propose a novel weakly supervised video moment localization method, called Contrastive Negative sample Mining (CNM). Our CNM generates a learnable Gaussian mask as the positive sample, which ensures the balance between recall rate and efficiency. Our CNM also proposes a novel method to mine the hard and easy negative samples within the same video, which enables CNM to distinguish highly confusing scenes. Extensive experiments and ablation studies on ActivityNet Captions and Charades-STA dataset demonstrate the advantages of CNM.

Acknowledgement

This work is supported by Zhejiang Lab (NO. 2022NB0AB05), State Key Laboratory of Media Convergence Production Technology and Systems, National Engineering Laboratory for Big Data Analysis and Applications Technology. The authors would also like to thank Jiabo Huang and Sizhe Li for helpful suggestions.

References

Balntas, V.; Riba, E.; Ponsa, D.; and Mikolajczyk, K. 2016. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, 3.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 961–970.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, S.; and Jiang, Y.-G. 2021. 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*, 8425–8435.

Chen, Z.; Ma, L.; Luo, W.; Tang, P.; and Wong, K.-Y. K. 2020. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308*.

Collins, R. T.; Lipton, A. J.; Kanade, T.; Fujiyoshi, H.; Duggins, D.; Tsin, Y.; Tolliver, D.; Enomoto, N.; Hasegawa, O.; Burt, P.; et al. 2000. A system for video surveillance and monitoring. *VSAM final report*, 2000(1-68): 1.

Duan, X.; Huang, W.; Gan, C.; Wang, J.; Zhu, W.; and Huang, J. 2018. Weakly supervised dense event captioning in videos. *arXiv preprint arXiv:1812.03849*.

Ellouz, F.; Adam, A.; Ciorbaru, R.; and Lederer, E. 1974. Minimal structural requirements for adjuvant activity of bacterial peptidoglycan derivatives. *Biochemical and biophysical research communications*, 59(4): 1317–1325.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. TALL: Temporal Activity Localization via Language Query. arXiv:1705.02101.

Gao, M.; Davis, L. S.; Socher, R.; and Xiong, C. 2019. Wslln: Weakly supervised natural language localization networks. *arXiv preprint arXiv:1909.00239*.

Huang, J.; Liu, Y.; Gong, S.; and Jin, H. 2021. Cross-Sentence Temporal and Semantic Relations in Video Activity Localisation. *arXiv preprint arXiv:2107.11443*.

Kemp, C. C.; Edsinger, A.; and Torres-Jara, E. 2007. Challenges for robot manipulation in human environments [grand challenges of robotics]. *IEEE Robotics & Automation Magazine*, 14(1): 20–29.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Niebles, J. C. 2017. Dense-Captioning Events in Videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*.

Lin, Z.; Zhao, Z.; Zhang, Z.; Wang, Q.; and Liu, H. 2020. Weakly-Supervised Video Moment Retrieval via Semantic Completion Network. arXiv:1911.08199.

Ma, M.; Yoon, S.; Kim, J.; Lee, Y.; Kang, S.; and Yoo, C. D. 2020. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *European Conference on Computer Vision*, 156–171. Springer.

Mithun, N. C.; Paul, S.; and Roy-Chowdhury, A. K. 2019. Weakly Supervised Video Moment Retrieval From Text Queries. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 11592–11601. Computer Vision Foundation / IEEE.

Neubeck, A.; and Van Gool, L. 2006. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, 850–855. IEEE.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global Vectors for Word Representation. In *Conference on Empirical Methods in Natural Language Processing*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Rodriguez-Opazo, C.; Marrese-Taylor, E.; Fernando, B.; Li, H.; and Gould, S. 2020. DORi: Discovering Object Relationship for Moment Localization of a Natural-Language Query in Video. arXiv:2010.06260.

Song, Y.; Wang, J.; Ma, L.; Yu, Z.; and Yu, J. 2020. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*.

Tan, R.; Xu, H.; Saenko, K.; and Plummer, B. A. 2021. 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*, 2083–2092.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.

Wang, H.; Zha, Z.-J.; Chen, X.; Xiong, Z.; and Luo, J. 2020. Dual Path Interaction Network for Video Moment Localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 4116–4124. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Wang, H.; Zha, Z.-J.; Li, L.; Liu, D.; and Luo, J. 2021a. 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)*, 7026–7035.

Wang, Y.; Deng, J.; Zhou, W.; and Li, H. 2021b. Weakly Supervised Temporal Adjacent Network for Language Grounding. *IEEE Transactions on Multimedia*.

Xiao, S.; Chen, L.; Zhang, S.; Ji, W.; Shao, J.; Ye, L.; and Xiao, J. 2021. Boundary Proposal Network for Two-Stage Natural Language Video Localization. arXiv:2103.08109.

Yang, W.; Zhang, T.; Zhang, Y.; and Wu, F. 2021. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE Transactions on Image Processing*, 30: 3252–3262.

Zhang, M.; Yang, Y.; Chen, X.; Ji, Y.; Xu, X.; Li, J.; and Shen, H. T. 2021. Multi-Stage Aggregated Transformer Network for Temporal Language Localization in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12669–12678.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020a. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. arXiv:1912.03590.

Zhang, Z.; Lin, Z.; Zhao, Z.; Zhu, J.; and He, X. 2020b. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4098–4106.

Zhao, Y.; Zhao, Z.; Zhang, Z.; and Lin, Z. 2021. Cascaded Prediction Network via Segment Tree for Temporal Video Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4197–4206.

Zhou, H.; Zhang, C.; Luo, Y.; Chen, Y.; and Hu, C. 2021. Embracing Uncertainty: Decoupling and De-bias for Robust Temporal Grounding. arXiv:2103.16848.