Iterative Proposal Refinement for Weakly-Supervised Video Grounding

Meng Cao[1*], Fangyun Wei[2], Can Xu[3], Xiubo Geng[3], Long Chen[4],
Can Zhang[1], Yuexian Zou[1], Tao Shen[3], Daxin Jiang[3†]

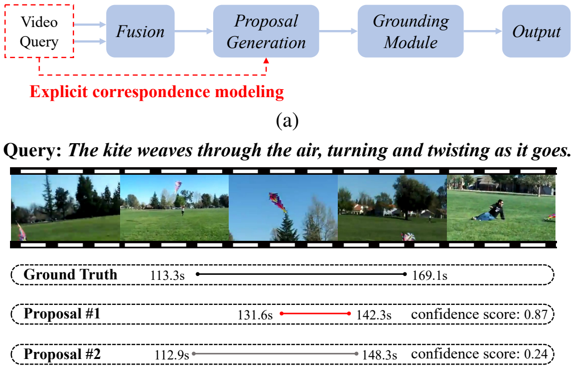
[1]School of Electronic and Computer Engineering, Peking University [2]Microsoft Research Asia
[3]Microsoft [4]The Hong Kong University of Science and Technology

Abstract

Weakly-Supervised Video Grounding (WSVG) aims to localize events of interest in untrimmed videos with only video-level annotations. To date, most of the state-of-the-art WSVG methods follow a two-stage pipeline, i.e., *firstly generating potential temporal proposals and then grounding with these proposal candidates. Despite the recent progress, existing proposal generation methods suffer from two drawbacks: 1) lack of explicit correspondence modeling; and 2) partial coverage of complex events. To this end, we propose a novel* ***I****te****R****ative pr****O****posal refi****N****ement network (dubbed as IRON) to gradually distill the prior knowledge into each proposal and encourage proposals with more complete coverage. Specifically, we set up two lightweight distillation branches to uncover the cross-modal correspondence on both the semantic and conceptual levels. Then, an iterative Label Propagation (LP) strategy is devised to prevent the network from focusing excessively on the most discriminative events instead of the whole sentence content. Precisely, during each iteration, the proposal with the minimal distillation loss and its adjacent ones are regarded as the positive samples, which refines proposal confidence scores in a cascaded manner. Extensive experiments and ablation studies on two challenging WSVG datasets have attested to the effectiveness of our IRON. The code will be available at https://github.com/mengcaopku/IRON.*

1. Introduction

Weakly-Supervised Video Grounding (WSVG) [21, 37, 41, 72, 73] aims to localize the moment of interest from an untrimmed video according to a query sentence without frame-wise annotations. It has drawn increasing attention in both industry and academia due to its wide applications, *e.g.*, video retrieval [13, 19], video question answering [1], human-computer interaction [49], *etc*. Currently, the overwhelming majority of state-of-the-art WSVG methods follow a two-stage pipeline, *i.e.*, they firstly generate potential proposals and then use these proposals to conduct grounding via multi-instance learning (MIL) [21,27,40,41] or query reconstruction [37, 40, 50, 72]. This paradigm commonly relies on densely-placed proposals to achieve high recall and ensure as much coverage as possible, which causes severe computation redundancy. Recent works [72, 73] reduce the number of required proposals by predicting Gaussian masks to highlight query-relevant segments. However, a such constraint is too rigorous and lacks flexibility. Thus, in this paper, we work toward designing sparse and reliable proposals without any distribution assumptions.



(a)

(b)

Figure 1. (a) The conventional WSVG pipeline (*i.e.*, baseline) lacks **explicit correspondence modeling**. (b) **Partial coverage of complex events**. Proposal #1 with the high confidence score (*i.e.*, 0.87) tends to be of short duration. The more reasonable proposal #2 has the lower confidence score.

Despite of the dominated performance achieved, it is worth noting that current proposal generation methods suffer from two inherent drawbacks: 1) **Lack of explicit correspondence modeling**: A simple pipeline[1] for the conventional proposal-based WSVG methods is illustrated in Fig-

[*] Work done during the internship at Microsoft.
[†]Corresponding author: Daxin Jiang (djiang@microsoft.com).

[1] We call this pipeline as *baseline* and refer to the appendix for details.

ure 1a. As shown, under the weakly-supervised scenario, there exist no explicit regression supervisions (*e.g.*, temporal boundary annotations) for the proposal generation procedure. Accordingly, the proposal coordinate update is solely based on the outputs of the grounding module. This leads to a *chicken and egg situation*, *i.e.*, the succeeding grounding module requires plausibly reliable proposals to achieve accurate localization results while the proposal distributions rely on decent grounding results to update. 2) **Partial coverage of complex events.** Compared to the atomic action instances in Temporal Action Localization (TAL) [8,47,71], the query sentences in WSVG are much more complex, *e.g.*, containing multiple events. Empirically, it is easy to excessively concentrate on the most discriminative parts instead of the whole picture [37]. For example, the case in Figure 1b aims to ground the complete process of the kite, *i.e.*, `weaving`, `turning`, and `twisting`. However, the top-ranking proposal #1 only covers the `weaving` process and overlooks the other parts. In contrast, a more accurate proposal #2 has much lower confidence scores. In Figure 2a, we compute the length distribution of the proposals with highest confidence score (Charades-STA [48] test set), which are always obviously shorter than their ground truths.

To alleviate these aforementioned problems, we propose a novel **I**te**R**ative pr**O**posal refi**N**ement network for WSVG (dubbed as **IRON**), which distills the prior knowledge into the proposal generation in a cascaded manner.

For correspondence modeling, we contend that it should be conducted from two aspects: 1) **Semantic-level**: The overall semantics of the proposals should match the query sentence. Specifically, we respectively feed the proposal frames and the query sentence into the visual and language encoders of the pre-trained video-language (VL) model [56] to estimate their semantic similarity. Due to the powerful transfer ability of pre-trained VL models [28, 39, 44, 56], we use the estimated similarity as the semantic distillation target. Then a lightweight semantic distillation branch is leveraged to optimize towards this target, referred to as the *semantic distillation loss*. 2) **Conceptual-level**: The proposal ought to be sensitive to the *linguistic salient concepts* including *object* words (*e.g.*, `kite`), *attribute* words (*e.g.*, `white`) and *relationship* words (*e.g.*, `through`). This is similar to the human way of reasoning, *i.e.*, tending to focus on the most prominent objects when assessing given videos. Here we define the *concepts* as the high-frequency words (*i.e.*, verbs, adjectives, and nouns) in the dataset corpus. Then, one multi-hot label is generated for each query sentence according to whether it hits the corresponding concept. We introduce a concept classification branch to estimate proposal-wise concept predictions supervised by the multi-hot label, yielding the *conceptual distillation loss*.

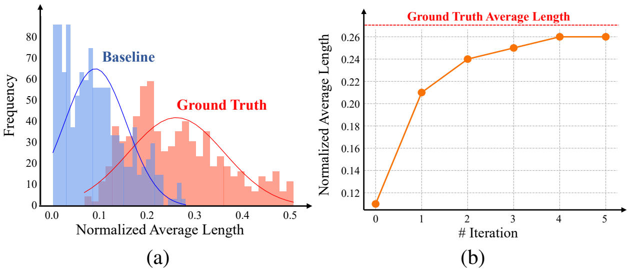


Figure 2. (a) The gaussian distributions of **normalized average length** of the **ground truth** and the proposals with highest confidence scores in **baseline**[1]. Results are calculated based on the test set of Charades-STA [48]. (b) The **normalized average length of proposals** with highest confidence scores *v.s.* **iteration numbers**.

To mitigate the partial coverage issue, we devise a Label Propagation (LP) algorithm, which aims to refine proposal-wise confidence scores in an iterative manner. Our motivation lies in that proposals with the minimal distillation loss (both semantic and conceptual distillation loss) can be regarded as the *biased indicator*, *i.e.*, these proposals always contain some salient events-of-interest and have short durations (*cf*. Sec. 4.5). Therefore, during each iteration in LP, we assign the positive pseudo label to the proposal with the minimal distillation loss and its *adjacent* ones[2]. Based on the generated pseudo label, the proposal confidence scores are rectified via a binary cross-entropy loss in the consequent stage. After the multi-step refinements, our IRON gradually converges to more complete intervals instead of parts (*cf*. Figure 2b).

In summary, we make three contributions in this paper:

- We propose to model explicit correspondence for each proposal at both semantic and conceptual levels, which distills in-depth knowledge from the well-trained VL model and the linguistic structure of the query sentence.
- To avoid biased and partial grounding results, a label propagation algorithm is crafted to refine proposal-wise confidence scores iteratively.
- Extensive experiments on both Charades-STA and ActivityNet Captions datasets have witnessed the state-of-the-art performance of our proposed IRON.

2. Related Work

Weakly-Supervised Video Grounding. Compared to the fully-supervised counterpart [5–7, 20, 35], WSVG [21, 37, 41, 72, 73] has gained intensive attention, which grounds the referent with only video-level annotations (*i.e.*, natural language queries). Currently, most of the WSVG methods can be classified into two major categories: multi-instance learning (MIL) based methods and reconstruction based methods. For MIL-based methods [21, 27, 40, 41], they learn the latent visual-textual alignment by attracting

[2] The *adjacent* proposals are defined as the proposals having high overlaps with the most confident one.

the matched video-language pairs while repelling the mismatched ones. On the contrary, reconstruction based methods [37, 40, 50, 72] rank proposals by a language reconstruction metric, asserting that the most matching proposals should best reconstruct the entire language query.

Since the current state-of-the-art methods [72, 73] rely on proposals for grounding, in this paper, we focus on generating high-quality proposals to fit the weakly supervised scenario. The most commonly used proposals are manually designed by employing multi-scale sliding windows [37, 41, 52]. Although straightforward, this methodology is computationally intensive and relies on heuristic rules. As an improvement, the recent works CNM [72] and CPL [73] use learnable Gaussian functions to generate both positive and negative proposals. To sum up, current proposal generation methods fail to provide explicit supervisions for proposal updates, which inevitably leads to suboptimum performance. In contrast, our IRON provides the supervision for the proposal generation stage at both the semantic and conceptual levels, facilitating the succeeding grounding module and generating more accurate results.

Knowledge Transfer of VL Pre-trained Models. As a breakthrough in the vision-language domain, large-scale VL pre-trained models (*e.g.*, CLIP [44], DeCLIP [36] and ALIGN [28]) have demonstrated great potential for learning transferable representations over diverse downstream tasks. CoOp [74] designs learnable prompts for textual inputs instead of handcrafted ones. CLIP-Adapter [22] conducts fine-tuning with a lightweight feature adapter and Tip-Adapter [68] proposes a non-parametric adapter via a key-value cache model. CLIP4clip [39] proposes to transfer the knowledge of CLIP to video retrieval in an end-to-end manner. ActionCLIP [57] formulates action recognition as a multi-modal learning framework to behave like video-text pre-training via prompt engineering. In this paper, our IRON distills the prior knowledge with two lightweight branches in a *multi-task* manner, which greatly improves the learning efficiency and prediction accuracy by using the knowledge in pre-trained models as an inductive bias [10].

Partial Coverage in Weakly-Supervised Learning. For the weakly-supervised learning, the localization tasks (*e.g.*, object detection [3, 45, 53] and TAL [33, 65, 67]) are always trained as the image/video classification task. Therefore, it tends to be easily responsive to trivial and sparse discriminative regions due to the inherent contradiction between the classifier and the detector. To tackle this, several TAL works [33, 38, 66] try to extend the discriminative regions by suppressing the dominant response or randomly hiding patches. However, such a heuristic multiple-run erasing model is unstable and not end-to-end trainable. In object detection, the mainstream methods [29, 34, 53, 62] mine high-quality bounding boxes by selecting top-scoring proposals from the preceding predictions. Such a strategy is solely based on the confidence score and cannot guarantee correctness, especially when encountering the open-set localization scenario like our targeted WSVG (*cf*. Sec. 4.5). In contrast, we progressively transfer the knowledge from the pre-trained VL models, which offers more reliable refinement labels due to their proven effectiveness on open-vocabulary detection [15, 23].

3. Approach

The schematic illustration of our IRON is illustrated in Figure 3. In Sec. 3.1, we present the preliminaries of IRON including feature extraction, proposal generation, semantic & conceptual distillation target generation. Then we detail the proposed iterative proposal refinement in Sec. 3.2. Finally, the grounding module is presented in Sec. 3.3.

3.1. Preliminary of IRON

Feature Extraction. Given an untrimmed video and a natural language query, we first feed them into the respective encoders to obtain the embedded features. Specifically, The encoded video feature is represented as $\boldsymbol{v} \in \mathbb{R}^{T\times C}$, where T is the number of sampled frames[3] and C is the feature dimension. The query embedding is represented as $\boldsymbol{q} \in \mathbb{R}^{S\times C}$, where S denotes the total word length.

Proposal Generation. We follow [72, 73] to conduct the proposal generation by predicting upon the video-language fusion results. Firstly, a learnable `[CLS]` token is inserted at the end of the input video features. Then a vanilla Transformer [55] is used to conduct cross-modal interactions and the `[CLS]` token output $\boldsymbol{h}_{\text{cls}}$ comprehensively interacts with all the frame and word features. Based on $\boldsymbol{h}_{\text{cls}}$, we predict a set of N proposals $\boldsymbol{u} \in \mathbb{R}^{N\times 2}$ by applying a fully connected layer activated by `sigmoid` function. Then, the proposal feature $\boldsymbol{p} \in \mathbb{R}^{N\times C}$ is generated by pooling video feature $\boldsymbol{v}$ with the corresponding coordinates of $\boldsymbol{u}$.

Semantic Distillation Target. Vision-language pretraining [44, 51] has shown great potential in learning transferable representations in a common feature space. Therefore, we resort to them to evaluate the global semantic alignment between the visual proposal and textual query. Firstly, we crop the video frames according to the proposal coordinates. Then, as shown in the top-right part of Figure 3, the cropped videos and the raw query sentence are fed into the visual and language encoders of pre-trained VL model [56] to obtain the cross-modal similarity. We consider the similarity values $\hat{\boldsymbol{s}} \in \mathbb{R}^{N\times 1}$ activated by `sigmoid` function as the semantic distillation targets.

Conceptual Distillation Target. Textual concepts are widely used for image-grounded cross-modal representations [42, 64], which serves as a complement to high-level

[3] We slightly abuse "frame" here, and we refer to it as a video segment consisting of several consecutive frames.

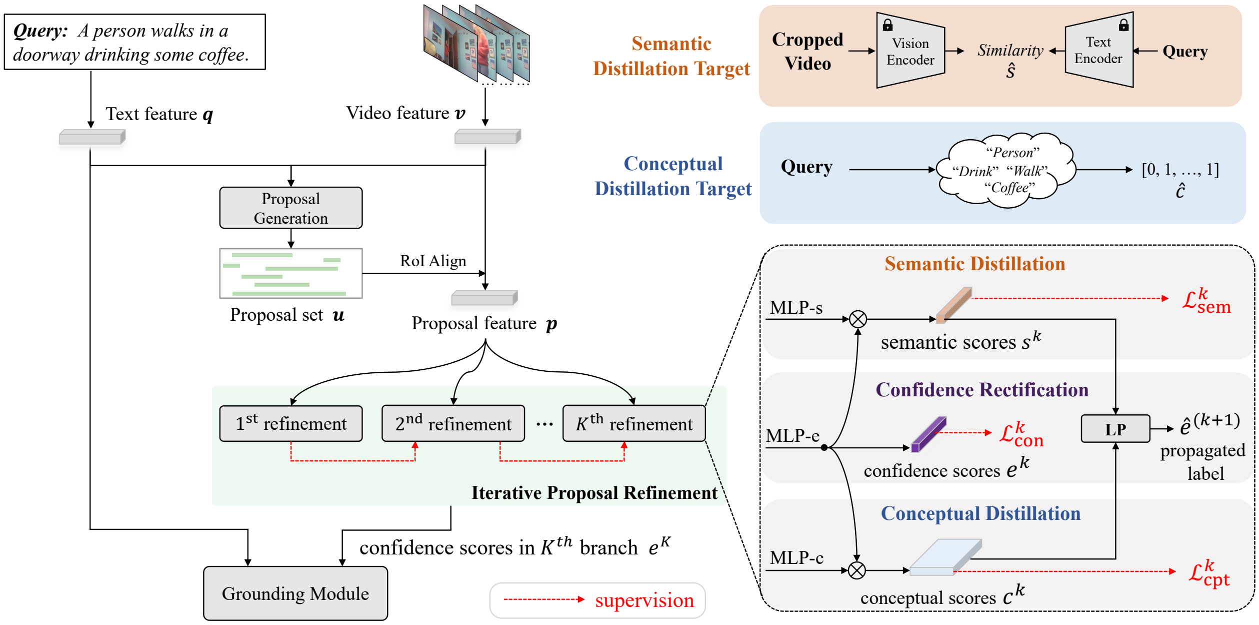

Query: A person walks in a doorway drinking some coffee.
Text feature $\boldsymbol{q}$
Video feature $\boldsymbol{v}$
Proposal Generation
RoI Align
Proposal set $\boldsymbol{u}$
Proposal feature $\boldsymbol{p}$
1st refinement
2nd refinement
···
Kth refinement
Iterative Proposal Refinement
confidence scores in K^{th} branch e^K
Grounding Module
supervision
Semantic Distillation Target
Cropped Video
Vision Encoder
Similarity $\hat{s}$
Text Encoder
Query
Conceptual Distillation Target
Query
"Person" "Drink" "Walk" "Coffee"
[0, 1, …, 1] $\hat{c}$
Semantic Distillation
MLP-s
semantic scores s^k
$\mathcal{L}_{\text{sem}}^k$
Confidence Rectification
MLP-e
confidence scores e^k
$\mathcal{L}_{\text{con}}^k$
LP
$\hat{e}^{(k+1)}$ propagated label
Conceptual Distillation
MLP-c
conceptual scores c^k
$\mathcal{L}_{\text{cpt}}^k$

Figure 3. **An overview of IRON.** The proposal generation module firstly integrates the text feature $\boldsymbol{q}$ and the video feature $\boldsymbol{v}$ with a vanilla Transformer. Then, a set of proposals $\boldsymbol{u}$ is predicted and the corresponding proposal features $\boldsymbol{p}$ are generated. **Semantic distillation targets** $\hat{\boldsymbol{s}} \in \mathbb{R}^{N\times 1}$ are estimated by computing the similarity scores between the cropped videos and the query sequences via the pre-trained VL model [56]. **Conceptual distillation targets** $\hat{\boldsymbol{c}} \in \mathbb{R}^{M\times 1}$ are customized according to whether the input query hits the pre-defined high-frequency words. In k^{th} iteration, proposal-wise semantic scores $\boldsymbol{s}^k \in \mathbb{R}^{N\times 1}$, conceptual scores $\boldsymbol{c}^k \in \mathbb{R}^{N\times M}$, and confidence scores $\boldsymbol{e}^k \in \mathbb{R}^{N\times 1}$ are predicted via three parallel MLPs. The **label propagation** (LP) strategy generates the pseudo confidence score labels $\hat{\boldsymbol{e}}^{k+1}$ for $(k+1)^{th}$ branch based on the current results. The confidence scores from the last refinement branch $\boldsymbol{e}^K$ are fed into the **grounding module**, which can be implemented with either MIL or query reconstruction (*cf*. Figure 4).

semantic information. We customize the conceptual distillation target for each proposal following [17]. Specifically, in Figure 3, the *concept corpus* is defined to be the most frequent M words (*e.g.*, nouns, verbs, and adjectives) among all the query sentences within the dataset. For each input query, we define its conceptual distillation target $\hat{\boldsymbol{c}} \in \mathbb{R}^{M\times 1}$ as a multi-hot vector, *i.e.*, $\hat{\boldsymbol{c}}_m = 1$ if the m^{th} concept word exists in the query sentence.

3.2. Iterative Proposal Refinement

Our iterative proposal refinement consists of K cascaded steps to distill prior knowledge and rectify the confidence scores. For clarity, we take the k^{th} $(1 \leq k \leq K)$ stage as an example to illustrate the stage-wise training.

Semantic & Conceptual Distillation. We aim to generate the proposal-wise semantic & conceptual scores and optimize towards the obtained distillation targets. As shown in the right part of Figure 3, we apply three branches of Multilayer Perceptrons (MLPs) to generate the corresponding predictions, *i.e.*, semantic scores $\boldsymbol{s}^k \in \mathbb{R}^{N\times 1}$, conceptual scores $\boldsymbol{c}^k \in \mathbb{R}^{N\times M}$, and proposal-wise confidence scores $\boldsymbol{e}^k \in \mathbb{R}^{N\times 1}$ as follows[4].

$$\begin{aligned} \boldsymbol{e}^k &= \mathrm{Sigmoid}\left(\boldsymbol{p}\cdot \mathbf{W}_{\mathrm{e}}^k\right) \\ \boldsymbol{s}^k &= \mathrm{Sigmoid}\left(\boldsymbol{p}\cdot \mathbf{W}_{\mathrm{s}}^k\right)\cdot \boldsymbol{e}^k, \\ \boldsymbol{c}^k &= \mathrm{Sigmoid}\left(\boldsymbol{p}\cdot \mathbf{W}_{\mathrm{c}}^k\right)\cdot \boldsymbol{e}^k, \end{aligned} \tag{1}$$

where $\mathbf{W}_{\mathrm{s}}^k, \mathbf{W}_{\mathrm{e}}^k \in \mathbb{R}^{C\times 1}$ and $\mathbf{W}_{\mathrm{c}}^k \in \mathbb{R}^{C\times M}$ are learnable parameters in the k^{th} iteration.

For the n^{th} proposal in the k^{th} iteration, the semantic loss $\mathcal{L}_{\text{sem}}^{k,n}$ and conceptual loss $\mathcal{L}_{\text{cpt}}^{k,n}$ are implemented in the ℓ_1 and binary cross-entropy form, respectively.

$$\mathcal{L}_{\text{sem}}^{k,n} = \left|\boldsymbol{s}_n^k - \hat{\boldsymbol{s}}_n\right|, \tag{2}$$

$$\mathcal{L}_{\text{cpt}}^{k,n} = -\textstyle\sum_{m=1}^{M} \hat{\boldsymbol{c}}_m \log \boldsymbol{c}_{n,m}^k, \tag{3}$$

where $\boldsymbol{s}_n^k$ is the semantic scores of n^{th} proposal. $\boldsymbol{c}_{n,m}^k$ denotes the the m^{th} concept score in the n^{th} proposal. $\hat{\boldsymbol{s}}_n$ and $\hat{\boldsymbol{c}}_m$ are corresponding labels (*cf*. Sec. 3.1).

Label Propagation. To alleviate the partial coverage of the learned proposals, we propose a label propagation (LP) strategy to refine the proposal-wise confidence score in a multi-stage architecture, which generates the supervision label for the next stage based on the current outputs.

[4] We omit all the bias terms of linear transformations for conciseness.

Algorithm 1 **Label Propagation**

```
Inputs: Semantic and conceptual scores in $k^{th}$ iteration $\boldsymbol{s}^k$, $\boldsymbol{c}^k$; Proposal-wise confidence score in $k^{th}$ iteration $\boldsymbol{e}^k$.
Outputs: Pseudo label of the confidence scores in $(k+1)^{th}$ iteration $\hat{\boldsymbol{e}}^{k+1}$.
Hyper-parameters: The IoU threshold $\beta$; The number of proposals $N$.
 1: function LP($\boldsymbol{s}^k$, $\boldsymbol{c}^k$, $\boldsymbol{e}^k$)
 2:     for $k = 1$ to $K-1$ do
 3:         $i^k \leftarrow \arg\min_{n\in[1,N]} \mathcal{L}_{\text{sem}}^{k,n} + \mathcal{L}_{\text{cpt}}^{k,n}$        ▷ cf. Eq. (4)
 4:         $\hat{\boldsymbol{e}}_{i^k}^{k+1} \leftarrow 1$
 5:         for $n = 1$ to $N$ do
 6:             $I_n \leftarrow$ calc_IoU $(\boldsymbol{u}_n, \boldsymbol{u}_{i^k})$        ▷ IoU with $\boldsymbol{u}_{i^k}$
 7:             if $I_n > \beta$ then        ▷ Trigger propagation
 8:                 $\hat{\boldsymbol{e}}_n^{k+1} \leftarrow 1$
 9:             else
10:                 $\hat{\boldsymbol{e}}_n^{k+1} \leftarrow 0$
11:             end if
12:         end for
13:     end for
14:     return $\hat{\boldsymbol{e}}^{k+1}$
15: end function
```

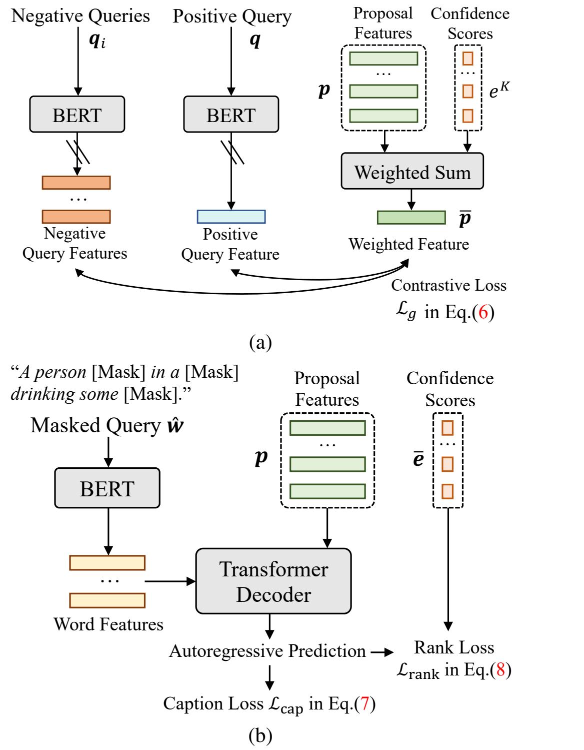


Figure 4. **The architectures of the grounding module** based on (a) MIL or (b) query reconstruction.

The pseudo code of LP is shown in Algorithm 1. Firstly, we identify the proposal with the minimum distillation loss in the current iteration as follow.

$$i^k = \mathop{\arg\min}_{n\in[1,N]} \mathcal{L}_{\text{sem}}^{k,n} + \mathcal{L}_{\text{cpt}}^{k,n}. \tag{4}$$

where $\mathcal{L}_{\text{sem}}^{k,n}$ and $\mathcal{L}_{\text{cpt}}^{k,n}$ are as defined in Eq. (2) and Eq. (3), respectively. We regard the selected proposal $\boldsymbol{u}_{i^k}$ as the positive sample in the $(k+1)^{th}$ iteration, due to its potential high quality (*cf*. Sec. 4.5), *i.e*., $\hat{\boldsymbol{e}}_{i^k}^{k+1} = 1$.

To force a more complete coverage, we contend that highly spatially overlapped instances should have the same label. Therefore, we inspect the other proposals and filter out those whose IoU with the proposal $\boldsymbol{u}_{i^k}$ is larger than threshold β. We also mark these surrounding proposals to be positive samples (*cf*. line 8 in Algorithm 1). Through this online labelling process, we not only *exploit* the most trustworthy proposal, but also *explore* its adjacent[2] ones to prevent the local optimum.

Confidence Rectification. Once obtaining the propagated pseudo label $\hat{\boldsymbol{e}}_{i^k}^{k+1}$, we compute the confidence rectification loss in the binary cross-entropy form.

$$\mathcal{L}_{\text{con}}^{k,n} = -\hat{\boldsymbol{e}}_n^k \log \boldsymbol{e}_n^k, \tag{5}$$

where $\boldsymbol{e}_n^k$ is the confidence score for the n^{th} proposal in k^{th} iteration ($2 \leq k \leq K$). Note that, we omit the confidence loss in the first stage ($k = 1$) since no supervisions are available. Finally, we use the confidence scores learnt in the K^{th} branch, *i.e*., $\boldsymbol{e}^K$, for the downstream grounding.

3.3. Grounding Module

Our IRON is flexible and can be compatible with both MIL-based and reconstruction-based grounding modules. Without loss of generality, we experiment with the vanilla versions without bells and whistles.

For the MIL-based method (Figure 4a), we firstly compute the video-level representations as the weighted sum of each proposal, *i.e*., $\overline{\boldsymbol{p}} = \sum_{n=1}^{N} \boldsymbol{e}_n^K \boldsymbol{p}_n$, where $\overline{\boldsymbol{p}} \in \mathbb{R}^{C\times 1}$ denotes the attended video feature representations. Then the matched video-query pairs are selected as the positive samples, where their embedding features are mapped close following the cross-modal InfoNCE loss [24].

$$\mathcal{L}_{\text{g}} = -\log \frac{\exp\left(\overline{\boldsymbol{p}}\cdot\boldsymbol{q}/\tau\right)}{\sum_{i=1}^{B} \exp\left(\overline{\boldsymbol{p}}\cdot\boldsymbol{q}_i/\tau\right)}, \tag{6}$$

where $\boldsymbol{q}_i \in \mathbb{R}^{C\times 1}$ is the query features extracted from the i^{th} sentences within the batch. B is the batch size and τ is the temperature parameter.

The reconstruction-based method is shown in Figure 4b. Specifically, the original query sentence $\boldsymbol{w} = \{\boldsymbol{w}_s\}_{s=1}^{S}$ is randomly blocked out with a specific `[MASK]` symbol by 1/3 of the total words S, yielding $\hat{\boldsymbol{w}}$. We reconstruct the query sentence in an auto-regressive manner following [72, 73], *i.e*., conditioning each word prediction on the previously generated outputs and the proposal features. Finally, the standard caption loss is employed to measure the reconstruction quality in the cross-entropy form.

$$\mathcal{L}_{\text{cap}}^{n} = -\sum_{s=1}^{S-1} \log p\left(\boldsymbol{w}_{s+1} \mid \hat{\boldsymbol{w}}_{1:s}, \boldsymbol{p}_{n}\right). \tag{7}$$

where $\mathcal{L}_{\text{cap}}^{n}$ is the caption loss for the n^{th} proposal. Besides, following [37], a rank loss is provided to correct the confidence score as follows.

$$\mathcal{L}_{\text{rank}}^{n} = -R_n \log\left(\frac{\exp \boldsymbol{e}_n^K}{\sum_{i=1}^{N} \exp \boldsymbol{e}_i^K}\right), \tag{8}$$

where R_n is the reward for n^{th} proposal to encourage proposals with lower reconstruction loss. Specifically, we sort all the proposals in ascending order according to the reconstruction loss $\mathcal{L}_{\text{cap}}^{n}$. Then rewards for the sorted proposals are reduced from one to zero in steps of $1/(N-1)$. Therefore, the grounding loss for reconstruction-based methods is computed as follows.

$$\mathcal{L}_{\text{g}} = \sum_{n=1}^{N} \left(\mathcal{L}_{\text{cap}}^{n} + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}}^{n}\right), \tag{9}$$

where λ_{rank} is the balance hyper-parameter.

Our final loss function is expressed as follows by integrating all the above constraints.

$$\mathcal{L} = \sum_{n=1}^{N}\sum_{k=1}^{K}(\mathcal{L}_{\text{sem}}^{k,n} + \mathcal{L}_{\text{cpt}}^{k,n}) + \lambda_{\text{con}} \sum_{n=1}^{N}\sum_{k=2}^{K} \mathcal{L}_{\text{con}}^{k,n} + \lambda_g \mathcal{L}_g, \tag{10}$$

where λ_{con} and λ_g are balance factors.

During inference, the well learned proposal confidence sores $\boldsymbol{e}^K$ are used to select the best proposals and the corresponding coordinates $\boldsymbol{u}$ are taken as the grounding result.

4. Experiments

4.1. Experimental Settings

Datasets. We validated the performance of our proposed IRON on two benchmarked datasets. 1) **ActivityNet Captions** [32]: It is built upon ActivityNet v1.3 dataset [4], which covers 19,290 untrimmed videos of complex human activities. The videos in this dataset last for 2 minutes on average while the timing length of the annotated temporal segments varies largely, ranging from several seconds to over 3 minutes. Since the test split is withheld for competition, we follow the public split [20, 72, 73] which uses 37,421 segment-query pairs for training, 17,505 pairs for validation, and 17,031 pairs for testing. 2) **Charades-STA** [20]: It is re-labeled by Gao *et al.* [20] based on the Charades dataset [48]. The videos in Charades-STA focus on indoor activities and the average video length is around 30 seconds. Following the official splits, 12,408 segment-query pairs are used for training, and 3,720 pairs for testing.

Evaluation Metrics. Following the previous works [26, 72], we adopt "Rn@m" as the metric, which is defined as the percentage of the language queries achieving at least one hitting (with IoU larger than m) in the top-n retrieved segments. We set $n \in \{1, 5\}$ for both datasets, $m \in \{0.3, 0.5, 0.7\}$ for Charades-STA, and $m \in \{0.1, 0.3, 0.5\}$ for ActivityNet Captions.

Implementation Details. For the video encoder, we chose C3D [54] pre-trained on sport1M [30] for ActivityNet Captions and I3D [9] pre-trained on Kinetics [9] for Charades-STA. Concretely, we took the output feature of `fc6` layer in C3D and the last average pooling result of I3D. As for the language encoder, we chose DistilBERT [46] pre-trained on English Wikipedia and Toronto Book Corpus for its light-weighted model capacity[5]. We set the maximum length of captions to 20 and the vocabulary size for ActivityNet Captions and Charades-STA was 8,000 and 1,111, respectively. For the fusion transformer, we set the hidden dimension to 256, the attention head number to 4, and the layer number to 3. We set the proposal number N to 8 for each video in both datasets. For the iterative refinement, we set the refinement number $K = 4$ and the IoU threshold $\beta = 0.6$. We took OA-Trans [56] as the frozen video-text pre-training model for semantic distillation target extraction. The size of the conceptual set M was set to 30. The temperature factor in Eq. (4) τ was set to 0.07 following [44, 61]. The balancing weights λ_{con} and λ_g were 5, 2, respectively. For the reconstruction-based grounding method, λ_{rank} was set to 0.1. All the models were trained for 50 epochs with a batch size of 32. Adam [31] was used as the optimizer, with a learning rate of 4×10^{-4}, linear decay of learning rate, and gradient clipping of 1.0.

4.2. Comparisons with State-of-the-Arts

The comparison results on Charades-STA and ActivityNet Captions datasets are summarized in Table 1 and Table 3, respectively. We can conclude with the following findings. **1)** Either grounding with MIL or reconstruction, our IRON outperforms previous state-of-the-art methods by a remarkable margin. For example on Charades-STA, when using the reconstruction strategy for grounding (*i.e.*, IRON in Table 1), our method surpasses the previous best performing method CPL [73] by 4.31% on R1@0.3. **2)** The reconstruction-based IRON outperforms the MIL-based one on both datasets. Besides, this gap is more obvious on ActivityNet Captions dataset. For example, on R1@0.3, the performance gap on Charades-STA and ActivityNet Captions datasets is 1.28% and 2.14%, respectively. This may be because the description annotations in ActivityNet Captions are more complex, which provides more sufficient textual contexts for the query reconstruction.

[5] We report the experimental results using Glove [43] as the language encoder in the appendix.

Table 1. **Comparisons (%) with state-of-the-art methods on Charades-STA dataset.** IRON* uses MIL for grounding and IRON follows the reconstruction strategy.

Method	R1@0.3	@0.5	@0.7	R5@0.3	@0.5	@0.7
CTF [12]	39.80	27.30	12.90	-	-	-
WSRA [18]	50.13	31.20	11.01	86.75	70.50	39.02
TGA [41]	32.14	19.94	8.84	86.58	65.52	33.51
SCN [37]	42.96	23.58	9.97	95.56	71.80	38.87
WSTAN [58]	43.39	29.35	12.28	93.04	76.13	41.53
BAR [60]	44.97	27.04	12.23	-	-	-
VLANet [40]	45.24	31.83	14.17	95.70	82.85	33.09
LoGAN [52]	48.04	31.74	13.71	89.01	72.17	37.58
MARN [50]	48.55	31.94	14.81	90.70	70.00	37.40
CCL [70]	-	33.21	15.68	-	73.50	41.87
CRM [27]	53.66	34.76	16.37	-	-	-
VCA [59]	58.58	38.13	19.57	98.08	78.75	37.75
LCNet [63]	59.60	39.19	18.87	94.78	80.56	45.24
RTBPN [69]	60.04	32.36	13.24	97.48	71.85	41.18
CNM [72]	60.39	35.43	15.45	-	-	-
CPL [73]	66.40	49.24	22.39	96.99	84.71	52.37
IRON* (Ours)	69.43	50.90	24.32	97.43	85.92	54.06
IRON (Ours)	**70.71**	**51.84**	**25.01**	**98.96**	**86.80**	**54.99**

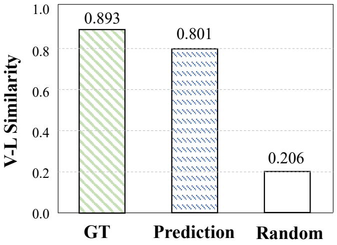


Figure 5. **Sanity check**: we calculate the similarity between the query sentence and ground truth segments, IRON predictions, and random segments through the pre-trained VL model [56], respectively.

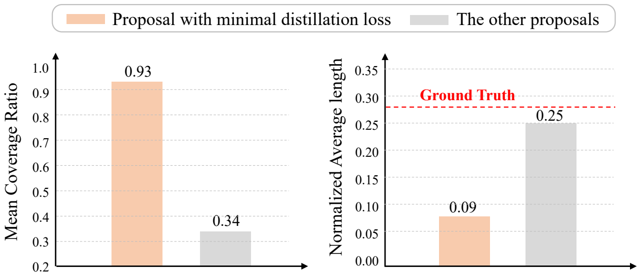


Figure 6. The mean values of **coverage ratio** (left) and **normalized average length** (right) of the proposals with minimal distillation loss and the other proposals, respectively.

4.3. Scalability Analysis[6]

Loss Component Ablations. We ablate the proposed semantic & conceptual distillation losses and the confidence rectification loss. As shown in Table 2a, $\mathcal{L}_{\text{sem}}$, $\mathcal{L}_{\text{cpt}}$, and $\mathcal{L}_{\text{con}}$ are all crucial to the overall performance. For example, when removing $\mathcal{L}_{\text{sem}}$ (mode #4), R1@0.3 drops by 8.67% compared to the full version (mode #1).

Inserting into existing methods. Our proposed iterative proposal refinement module could serve as plug-and-play and can be seamlessly inserted into existing WSVG methods. Here we select two representative methods, *i.e*., Gaussian proposal based CPL [73] and dense proposal based VLANet [40]. As shown in Table 2b, our refinement strategy benefits both methods remarkably. For example, on top of CPL [73], our refinement leads to a 3.74% improvement on R1@0.3. The consistent improvement demonstrates the flexibility and extensibility of our proposed strategy.

4.4. Ablations on Semantic & Concept Distillations[6]

Ablations on the semantic information source. We experiment with different pre-trained video-language models [2, 56] and the well-known CLIP model [44] with different backbones. For CLIP, the video-query similarity is computed as the mean of frame-query similarities. As shown in Table 2c, OA-Trans [56] achieves superior performance, which may be because it introduces detailed object-level information into pre-training. Either CLIP equipped with ResNet-50 [25] or ViT-B/16 [14] performs below expectations since it neglects to model the temporal information.

Correctness of the semantic distillation target. To validate the reasonableness of the semantic distillation target, we set up a sanity check. Specifically, for each input query, we calculate its similarity with the ground truth segments, IRON prediction results, and randomly sampled segments through VL pre-trained model [56], respectively. As shown in Figure 5, randomly sampled segments show the lowest similarity scores, which demonstrates that the pre-trained model can generate *distinguishable* similarity scores.

4.5. Ablations on Label Propagation[6]

Insights of proposals with minimal distillation loss. Our label propagation strategy is based on two assumptions of the proposal with minimum distillation loss: 1) *High hit rate*. They can always hit the region of interest. To demonstrate this, we devise a new metric *coverage ratio* to be the ratio of the length overlapping with the ground truth to the proposal total length. We compute the coverage ratio for the proposal with the minimal distillation loss and the mean values for other proposals, respectively. As shown in Figure 6 (left), proposals with minimal loss are almost completely covered by ground truth (with a coverage ratio reaching 0.93). 2) *Short duration*. We show the normalized average length of each proposal in Figure 6 (right). As shown, the proposal with the minimal loss bears an average length of 0.09, which is obviously shorter than the average ground truth length. Therefore, these two characteristics effectively support the rationality of our refinement strategy.

Ablations on label propagation strategy. In Sec 3.2, we

[6] All ablation studies are conducted on Charades-STA dataset with the reconstruction-based grounding module.

Table 2. (a) **Ablations of loss components**. (b) **Scalability analysis** by inserting the iterative proposal refinement module into existing WSVG methods [40, 73]. (c) **Ablations on the semantic information source**.

Mode	$\mathcal{L}_{\text{sem}}$	$\mathcal{L}_{\text{cpt}}$	$\mathcal{L}_{\text{con}}$	**R1@0.3**	**@0.5**	**@0.7**
#1	✓	✓	✓	**70.71**	**51.84**	**25.01**
#2	✓	✓	✗	65.83	46.96	20.38
#3	✓	✗	✓	66.24	47.13	20.90
#4	✗	✓	✓	62.04	40.33	17.20

(a)

Method	**R1@0.3**	**@0.5**	**@0.7**
CPL [73]	66.40	49.24	22.39
+ *refinement*	**70.14**	**50.55**	**24.61**
VLANet [40]	45.24	31.83	14.17
+ *refinement*	**49.90**	**33.70**	**15.18**

(b)

Method	**R1@0.3**	**@0.5**	**@0.7**
OA-Trans [56]	**70.71**	**51.84**	**25.01**
Frozen [2]	68.84	49.67	23.83
CLIP (RN-50) [44]	67.04	48.63	23.04
CLIP (ViT-B/16) [44]	68.56	49.35	23.61

(c)

Table 3. **Comparisons (%) with state-of-the-art methods on ActivityNet Captions dataset.** IRON* uses MIL for grounding and IRON follows the reconstruction strategy.

Method	**R1@0.1**	**@0.3**	**@0.5**	**R5@0.1**	**@0.3**	**@0.5**
CTF [12]	74.20	44.30	23.60	-	-	-
EC-SL [11]	68.48	44.29	24.16	-	-	-
WS-DEC [16]	62.71	41.98	23.34	-	-	-
MARN [50]	-	47.01	29.95	-	72.02	57.49
SCN [37]	71.48	47.23	29.22	90.88	71.56	55.69
VCA [59]	67.96	50.45	31.00	92.14	71.79	53.83
BAR [60]	-	49.03	30.73	-	-	-
RTBPN [69]	73.73	49.77	29.63	93.89	79.89	60.56
WSLLN [21]	75.40	42.80	22.70	-	-	-
LCNet [63]	78.58	48.49	26.33	93.95	82.51	62.66
CCL [70]	-	50.12	31.07	-	77.36	61.29
WSTAN [58]	79.78	52.45	30.01	93.15	79.38	63.42
CRM [27]	81.61	55.26	32.19	-	-	-
CNM [72]	78.13	55.68	33.33	-	-	-
CPL [73]	82.55	55.73	31.37	87.24	63.05	43.13
IRON* (Ours)	82.83	56.81	33.67	95.09	83.46	67.38
IRON (Ours)	**84.42**	**58.95**	**36.27**	**96.74**	**85.60**	**68.52**

Table 4. **Ablations** on the label propagation strategy (*cf*. Eq. (4)). Exp #1 mines the labels based on both semantic loss $\mathcal{L}_{\text{sem}}$ and conceptual loss $\mathcal{L}_{\text{cpt}}$. Exp #2 and Exp #3 only leverage either $\mathcal{L}_{\text{sem}}$ or $\mathcal{L}_{\text{cpt}}$. Exp #4 propagates labels by selecting proposals with maximum confidence score $\boldsymbol{e}$.

Exp	**propagation source**	**R1@0.3**	**@0.5**	**@0.7**
#1	$\min\,(\mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{cpt}})$	**70.71**	**51.84**	**25.01**
#2	$\min \mathcal{L}_{\text{sem}}$	68.65	50.10	24.26
#3	$\min \mathcal{L}_{\text{cpt}}$	69.14	51.21	24.35
#4	$\max \boldsymbol{e}$	67.12	50.01	23.79

Table 5. **Comparisons (%) with SOTA methods on Charades-STA dataset.** IRON follows the reconstruction strategy. For fair comparisons, we reproduce the results† of [72, 73] with different input features.

Exp.	**Method**	**Feat**	**R1@0.5**	**@0.7**	**R5@0.5**	**@0.7**
1	CNM [72]	I3D	35.43	15.45	-	-
2	CPL [73]	I3D	49.24	22.39	84.71	52.37
3	CNM† [72]	OATrans	37.76	16.24	-	-
4	CPL† [73]	OATrans	50.13	22.84	85.22	53.01
5	CNM† [72]	I3D+OATrans	37.13	15.93	-	-
6	CPL† [73]	I3D+OATrans	50.02	22.14	84.98	52.73
7	IRON (Ours)	I3D+OATrans	**51.84**	**25.01**	**86.80**	**54.99**

design label propagation based on proposals with minimal semantic and conceptual distillation losses. Here we experiment other options, *i.e.*, based on proposals with only minimal semantic loss $\mathcal{L}_{\text{sem}}$ (#2 in Table 4), with only minimal conceptual loss $\mathcal{L}_{\text{cpt}}$ (#3) and with maximum confidence score $\boldsymbol{e}$ (#4). As shown in Table 4, propagation with only semantic or conceptual loss will degrade the performance. Besides, propagation with maximum confidence score is also sub-optimum, *e.g.*, 3.59% absolute drop on R1@0.3. This demonstrates that both semantic and conceptual losses are more reliable indicators.

Comparison Fairness Issue. For fair comparisons, we reproduce results of SOTA methods (CNM [72], CPL [73]) by using different features, *i.e.*, OATans [56] or channel-wise concatenation of I3D and OATans. As shown in Table 5, replacing I3D features with OATans features leads to performance gains (*e.g.*, exp.#1 *vs*. #3). However, simply concatenating OATans and I3D is slightly inferior to the OATans-based version (*e.g.*, exp.#3 *vs*. #5). Therefore, under the same setting, our proposed IRON outperforms the existing methods.

5. Conclusions

In this paper, we concentrate on designing appropriate and universal proposals for WSVG. To achieve this, an iterative proposal refinement network is proposed. Firstly, we distill the semantic and conceptual knowledge into each proposal from the well-trained video-language model and the linguistic structure of the query sentence, respectively. Besides, a label propagation strategy is put forward to avoid the biased localization results which only focus on the most discriminative action events. Extensive experimental results on WSVG datasets have illustrated the effectiveness of our proposed IRON.

Acknowledgements. This paper was partially supported by NSFC (No:62176008) and Shenzhen Science & Technology Research Program (No:GXWD20201231165807007-20200814115301001).

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 7, 8

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2846–2854, 2016. 3

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

[5] Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. On pursuit of designing multi-modal transformer for video grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9810–9823, 2021. 2

[6] Meng Cao, Ji Jiang, Long Chen, and Yuexian Zou. Correspondence matters for video referring expression comprehension. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4967–4976, 2022. 2

[7] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtp: Video-text pre-training for temporal localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 38–56. Springer, 2022. 2

[8] Meng Cao, Can Zhang, Long Chen, Mike Zheng Shou, and Yuexian Zou. Deep motion prior for weakly-supervised temporal action localization. *IEEE Transactions on Image Processing*, 31:5203–5213, 2022. 2

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 6

[10] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 3

[11] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8435, 2021. 8

[12] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308*, 2020. 7, 8

[13] Xuxin Cheng, Zhihong Zhu, Hongxiang Li, Yaowei Li, and Yuexian Zou. Ssvmr: Saliency-based self-training for video-music retrieval. *arXiv preprint arXiv:2302.09328*, 2023. 1

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 7

[15] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 3

[16] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. *Advances in Neural Information Processing Systems*, 31, 2018. 8

[17] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015. 4

[18] Zhiyuan Fang, Shu Kong, Zhe Wang, Charless Fowlkes, and Yezhou Yang. Weak supervision and referring attention for temporal-textual association learning. *arXiv preprint arXiv:2006.11747*, 2020. 7

[19] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer, 2020. 1

[20] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017. 2, 6

[21] Mingfei Gao, Larry Davis, Richard Socher, and Caiming Xiong. Wslln: Weakly supervised natural language localization networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1481–1487, 2019. 1, 2, 8

[22] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 3

[23] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 5

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[26] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4555–4564, 2016. 6

[27] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7199–7208, 2021. 1, 2, 7, 8

[28] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2, 3

[29] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1377–1385, 2017. 3

[30] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 6

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[32] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 6

[33] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3524–3533, 2017. 3

[34] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3512–3520, 2016. 3

[35] Hongxiang Li, Meng Cao, Xuxin Cheng, Zhihong Zhu, Yaowei Li, and Yuexian Zou. Generating templated caption for video grounding. *arXiv preprint arXiv:2301.05997*, 2023. 2

[36] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 3

[37] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11539–11546, 2020. 1, 2, 3, 6, 7, 8

[38] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1298–1307, 2019. 3

[39] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2, 3

[40] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *European conference on computer vision*, pages 156–171. Springer, 2020. 1, 2, 3, 7, 8

[41] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2019. 1, 2, 3, 7

[42] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6504–6512, 2017. 3

[43] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 6

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 6, 7, 8

[45] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10607, 2020. 3

[46] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 6

[47] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5734–5743, 2017. 2

[48] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526, 2016. 2, 6

[49] Joyeeta Singha, Amarjit Roy, and Rabul Hussain Laskar. Dynamic hand gesture recognition using vision-based approach for human–computer interaction. *Neural Computing and Applications*, 29(4):1129–1141, 2018. 1

[50] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*, 2020. 1, 3, 7, 8

[51] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 3

[52] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2083–2092, 2021. 3, 7

[53] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2843–2851, 2017. 3

[54] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 6

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[56] Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2022. 2, 3, 4, 6, 7, 8

[57] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 3

[58] Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li. Weakly supervised temporal adjacent network for language grounding. *IEEE Transactions on Multimedia*, 2021. 7, 8

[59] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1459–1468, 2021. 7, 8

[60] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1283–1291, 2020. 7, 8

[61] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 6

[62] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8372–8381, 2019. 3

[63] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE Transactions on Image Processing*, 30:3252–3262, 2021. 7, 8

[64] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE international conference on computer vision*, pages 4894–4902, 2017. 3

[65] Yuan Yuan, Yueming Lyu, Xi Shen, Ivor Tsang, and Dit-Yan Yeung. Marginalized average attentional network for weakly-supervised learning. In *ICLR 2019-Seventh International Conference on Learning Representations*, 2019. 3

[66] Runhao Zeng, Chuang Gan, Peihao Chen, Wenbing Huang, Qingyao Wu, and Mingkui Tan. Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization. *IEEE Transactions on Image Processing*, 28(12):5797–5808, 2019. 3

[67] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *CVPR*, pages 16010–16019, 2021. 3

[68] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 3

[69] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4098–4106, 2020. 7, 8

[70] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33:18123–18134, 2020. 7, 8

[71] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017. 2

[72] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1, page 3, 2022. 1, 2, 3, 5, 6, 7, 8

[73] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15555–15564, 2022. 1, 2, 3, 5, 6, 7, 8

[74] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3