Correspondence Matters for Video Referring Expression Comprehension

Meng Cao
mengcao@pku.edu.cn
SECE, Peking University

Ji Jiang
jiangji@stu.pku.edu.cn
SECE, Peking University

Long Chen∗
zjuchenlong@gmail.com
Columbia University

Yuexian Zou∗
zouyx@pku.edu.cn
SECE, Peking University & Peng Cheng Laboratory

ABSTRACT
We investigate the problem of video Referring Expression Comprehension (REC), which aims to localize the referent objects described in the sentence to visual regions in the video frames. Despite the recent progress, existing methods suffer from two problems: 1) inconsistent localization results across video frames; 2) confusion between the referent and contextual objects. To this end, we propose a novel Dual Correspondence Network (dubbed as DCNet) which explicitly enhances the dense associations in both the inter-frame and cross-modal manners. Firstly, we aim to build the inter-frame correlations for all existing instances within the frames. Specifically, we compute the inter-frame patch-wise cosine similarity to estimate the dense alignment and then perform the inter-frame contrastive learning to map them close in feature space. Secondly, we propose to build the fine-grained patch-word alignment to associate each patch with certain words. Due to the lack of this kind of detailed annotations, we also predict the patch-word correspondence through the cosine similarity. Extensive experiments demonstrate that our DCNet achieves state-of-the-art performance on both video and image REC benchmarks. Furthermore, we conduct comprehensive ablation studies and thorough analyses to explore the optimal model designs. Notably, our inter-frame and cross-modal contrastive losses are plug-and-play functions and are applicable to any video REC architectures. For example, by building on top of Co-grounding [44], we boost the performance by 1.48% absolute improvement on Accu.@0.5 for VID-Sentence dataset. Our codes are available at https://github.com/mengcaopku/DCNet.

KEYWORDS
Video Referring Expression Comprehension; Inter-Frame Contrastive Learning; Cross-Modal Contrastive Learning

1 INTRODUCTION
Referring Expression Comprehension (REC) [19, 20, 60, 61] is a challenging task which aims to localize the region of the image described by the natural language (i.e., referent). It has been extensively studied in both industry and academia, due to its wide applications in
Sparse Alignment
Dense Alignment

Figure 2: Left: Sparse Alignment is conducted only between the annotated bounding box region features; Right: Dense Alignment is conducted for all the patch features.

downstream tasks, e.g., visual question answering [1, 9, 11], captioning [7, 10], navigation [6], autonomous driving [25], e.t.c. Recently, several works [12, 47, 68, 69] are developed to conduct REC in the video domain. Compared to image REC, video REC is more arduous due to the complex temporal structures.

Although great improvement has been achieved by current methods, it is worth noting that current methods have two overlooked drawbacks: 1) Inconsistent localization results across video frames. Compared to image REC, some frames in video REC may suffer from the deteriorated visual quality, e.g., motion blur, light change, object occlusion, and out of focus, leading to unstable and interrupted localization results. Thus, it is hard and error-prone to identify the referent by only utilizing information within one isolated image. A state-of-the-art video REC method [44] also points out this issue and partially alleviates it by introducing the frame-wise attention mechanism. However, such attention fails to build the explicit correspondence and still leads to the sub-optimal performance. For example in Figure. 1a, the referent "elephant" is obscured in frame #2 and the method [44] mistakes another elephant as the object-of-interest. 2) Confusion between the referent and contextual objects. Recall that our goal is to ground the object related to the global sentence meaning. It is common for the sentence to contain multiple objects including the referent and contextual objects. Current methods tend to confuse these two types of objects since they do not introduce detailed alignment between all the mentioned object instances and object areas. As the example shown in Figure. 1b, the given query sentence ("A rabbit is running around a deer.") contains two instance, (i.e., "rabbit", "deer"). The state-of-the-art method [44] fails to identify the referred instance "rabbit" and falsely localize the contextual object "deer".

Based on these observations, we propose a Dual Correspondence Network (dubbed as DCNet) for more accurate video referring expression comprehension. Specifically, an inter-frame contrastive loss is designed to enhance patch-wise\(^1\) correspondence within frames and a cross-modal contrastive loss is built to model fine-grained word-patch alignment.

Firstly, we contend that **instance feature consistency** is crucial to generate more stable and consistent localization results. Concretely, each instance within the frame should be aware of its potential spatial localization in the contextual frames. Let’s revisit the case in Figure. 1a. Although being obscured, the feature response for the "elephant" in frame #2 should be consistent with that in frame #1 and frame #3. Therefore, we propose an inter-frame contrastive loss to enhance instance correspondence between frames. To achieve this, a straightforward way is to consider the annotated ground-truth areas of each frame as the positive samples (cf. Figure. 2 left). This method, however, can not build the dense correspondence for each instance. Besides, relying on annotations makes it only applicable to the supervised training scenario. To alleviate this, we take a rather simple manner to estimate the dense correspondence by computing the inter-frame patch-wise cosine similarity. Then the patch pairs with top similarity scores are selected as the positive samples, where their embedding features are mapped close.

Secondly, we believe that building patch-word fine-grained alignment helps distinguish the referent and the contextual objects. For example in Figure. 1b, the query sentence contains two keywords "rabbit" and "deer", which are semantically associated with certain patches in the video frame. Therefore, we propose to model such fine-grained cross-modal alignment. To establish this detailed regularization, we also resort to the cosine similarity computation manner to mine the desired correspondence. Specifically, given one patch feature within the frame, we compute its cosine similarity with each word feature and select those with the highest similarities as positive pairs.

Equipped with the two mentioned contrastive losses, our DCNet constructs both the inter-frame and cross-modal correspondence with the absence of explicit annotations. These two losses could serve as plug-and-play functions that can be easily inserted into existing video REC methods. As the example shown in Figure. 3, by building on top of the baseline model [44], our inter-frame and cross-modal contrastive losses lead to more consistent and accurate feature responses.

In summary, we make three contributions in this paper:

- We contend that the inter-frame and cross-modal alignments are crucial for accurate video REC. To this end, we propose a simple yet effective cosine similarity mining method to extract dense patch-wise and fine-grained patch-word correspondence.

---

\(^1\)We use "patch" to denote each spatial feature on the embedded feature maps.
Based on the extracted correspondence, we propose the inter-frame and cross-modality contrastive loss, which boost more consistent video feature representations and optimize fine-grained visual-text relations, respectively.

Experimental results show that our DCNet achieves state-of-the-art performance on both video and image REC benchmarks. Besides, our proposed cross-modal and inter-frame contrastive loss can be further seamlessly integrated into any video REC architectures and achieve promising performance gains.

2 RELATED WORK

Referring Expression Comprehension. REC aims to ground a language query to the corresponding region on an image. Most of the advances in REC can be broadly classified into two categories, i.e., two-stage methods and one-stage methods.

Two-stage REC methods [8, 18, 19, 31, 32, 35, 50, 51, 53, 59, 60, 65, 72] adopt the propose-and-rank pipeline, where region candidates are firstly generated using unsupervised methods [37, 49] or the pre-trained object detector [59, 65]. Then the similarity scores between the sentence and region candidates are computed, which can be discriminatively trained with the binary classification [49, 66] or maximum-margin ranking [35, 36, 50].

One-stage REC methods [21, 29, 34, 45, 54, 58, 69] get rid of the computation-intensive proposal generation process and aim to directly perform bounding box prediction based on the the language-attended feature maps. The pioneering work [56] fuses the language features into the YOLOv3 detector [39] to ground the referred instance. To take one step further, ReSC [55] proposes a recursive sub-query construction framework, which conducts multi-round image-query reasoning and reduces the referring ambiguity. To achieve real-time inference, RCCF [29] formulates the REC problem as a correlation filtering process [3, 17] by employing the peak value in the correlation heatmap as the center points of the target box. Recently, several works [13, 23, 71] rely on the Transformer-family structures [48] to perform the end-to-end grounding without post-processing procedures.

Video Referring Expression Comprehension. Besides the image REC, there are several works exploring the task in the video domain. Compared to video grounding [4, 5, 62–64], video REC requires both spatial and temporal localization results. According to the specific settings, they can be classified into three threads. 1) Several works [22, 43, 54, 58, 69] aim to localize all the mentioned objects in the corresponding regions for each frame of the video; 2) There are also works identifying spatial-temporal tubes within untrimmed videos from referring expressions [67, 68]; 3) The other works [12, 42, 44] localize only the referred object corresponding to the global query sentence semantics within the trimmed videos.

This paper targets the third setting since it is a more basic and fundamental problem. Chen et al. [12] take the two-stage manner by firstly extracting potential spatio-temporal tubes and then aligning these candidates to the sentence to find the best matching one. [44] proposes a one-stage framework built on YOLO-v3 [39], i.e., it fuses visual-text features and predicts bounding boxes densely at all spatial locations. [42] investigates the role of object relations across both time and space by modeling self-attention with relative position encoding. These methods, however, do not introduce detailed alignment between the mentioned object instances and object areas, making it hard to identify the referent. In this paper, we aim to build this comprehensive alignment for all densely placed instances and achieve more accurate grounding results.

3 APPROACH

This work aims to introduce a dual correspondence network for video REC. We firstly present the backbone of our DCNet including the video-language fusion process in Sec. 3.1. Then we elaborate the proposed inter-frame and cross-modality contrastive loss in Sec. 3.2 and Sec. 3.3, respectively. The final training objectives are presented in Sec. 3.4.

3.1 Backbone of DCNet

The schematic illustration of our DCNet is illustrated in Figure 4. Given one video-query pair, we firstly feed the video and language modalities to their respective encoders \( f_p \) and \( f_q \) to obtain embedded features. Specifically, during each iteration, we randomly sample \( T \) video frames and each two sampled frames are with \( F \) frame temporal distance. The encoded video feature is represented as \( v = \{v^i\}_{i=1}^T \). The \( i \)-th frame feature is denoted as \( v^i = \{v^i_p\}_{p=1}^P \), where \( v^i_p \in \mathbb{R}^D \) is the \( p \)-th patch feature of the \( i \)-th frame\(^2\). \( p \) is the total patch number and \( D \) is the feature dimension. The text embedding is represented as \( q = \{q^s\}_{s=1}^S \), where \( q^s \in \mathbb{R}^D \) is the \( s \)-th word embedding and \( S \) is the total word length.

Then the two modality features are fused via a video-language fusion module. Without loss of generality, we utilize the attention mechanism to aggregate the word features for each patch feature following [12, 68]. Specifically, given \( i \)-th frame patch feature with the index \( p \in [1, P] \), i.e., \( v^i_p \), we compute its attentive scores with each query word \( q^s \), \( s \in [1, S] \) and apply a Softmax to normalize\(^3\):

\[
\epsilon^{i}_{p,s} = \text{Softmax} \left( W^p \tanh(W^v v^i_p + W^q q^s) \right),
\]

where \( W, W^v, W^q \in \mathbb{R}^{D \times D} \) are learnable parameters. \( \epsilon^{i}_{p,s} \) is the computed attention weight for patch feature \( v^i_p \) with regard to word feature \( q^s \). Accordingly, the cross-modal fused features are computed as follows:

\[
f^i_p = \sum_{s=1}^{S} \epsilon^{i}_{p,s} q^s,
\]

where \( f^i_p \) is the textual-aware video feature for \( p \)-th patch at \( i \)-th frame. Thus, the cross-modality fused feature of \( i \)-th frame is denoted as \( f^i = \{f^i_p\}_{p=1}^P \). Through this, we characterize the matching behaviors between each patch and the query word features, leading to well-fused cross-modal features.

3.2 Inter-frame Contrastive Learning

We propose to build the patch-wise correspondence and conduct the inter-frame contrastive learning. We contend that introducing such alignment helps incorporate more contextual information and calibrate the frame features to be more temporally consistent. The most

\(^2\)The \( i \)-th frame feature \( v^i_p \) is a 2-D feature and the \( p \)-th patch is indexed by the order of firstly-row-and-then-column.

\(^3\)For conciseness, we omit all the bias terms of linear transformations in this paper.
Figure 4: An overview of DCNet. \( f_0(\cdot) \) and \( f_q(\cdot) \) are video and language encoders, respectively. Inter-frame correspondence mines the dense patch-wise coherence while cross-modal correspondence finds the detailed patch-word alignment. Note that for clarity, we only present the correspondence discovery for one selected anchor patch \( v_p \) (marked by orange square). Then the discovered inter-frame positive patches \( m_p \) are marked by green while the cross-modal positive words \( w_p \) are marked by blue. Based on this, inter-frame and cross-modal contrastive loss \( L_{\text{inter}}, L_{\text{cross}} \) are applied, i.e., \( v_p \leftrightarrow m_p, v_p \leftrightarrow w_p \). The video-language fusion module fuses the cross-modal features via the attention mechanism. Based on it, \( L_{\text{cls}}, L_{\text{reg}} \) conduct the classification and regression loss for each spatial position of the features. Note that we omit the video-language fusion as well as the classification and regression loss for frame feature \( v_i \) for clarity.

As shown in Figure 4, we take two encoded image frame features \( v^i \) and \( v^j \), \( i, j \in [1, T] \) as an example to illustrate the details. For each patch within \( v^i \), e.g., \( v_{p}^i, p \in [1, P] \), we firstly compute its cosine similarity with all patch feature in \( v^j \). The items with top \( K_{\text{inter}} \) similarity scores are selected and pooled as the positive samples.

\[
m_p^i = \text{avgpool}\left(\arg\max_{q \in [1, P]} (v^i_p \cdot v^j_q)\right),
\]

where \( m_p^i \) is the final positive sample for \( v_p^i \), and \( v_q^j \) is the \( q \)th patch of \( v^j \). \((u \cdot v) = u^T v / \|u\| \|v\| \) represents the cosine similarity between \( \ell_2 \) normalized \( u \) and \( v \). The selected positive number is defined as 

\[
K_{\text{inter}} = \max(1, \lfloor P / R_{\text{inter}} \rfloor),
\]

where \( P \) is the total patch number and 1/\( R_{\text{inter}} \) is the sampling ratio.

Based on the established patch-wise correspondence, we perform the inter-frame contrastive learning following the cross-modal InfoNCE [15] loss.

\[
L_{\text{inter}}^{ij} = \frac{1}{P} \sum_{p=1}^{P} - \log \frac{\exp\left(\frac{v_p^i \cdot m_p^j}{\tau}\right)}{\sum_{q=1}^{P} \exp\left(\frac{v_p^i \cdot v_q^j}{\tau}\right)},
\]

where \( \tau \) is the temperature parameter.

To aggregate \( L_{\text{inter}}^{ij} \) among all sampled \( T \) frames, we propose two potential ways to obtain the final inter-frame contrastive loss \( L_{\text{inter}} \). As shown in Figure 5, the adjacent strategy only conducts the inter-frame contrastive loss between the nearest frames while the fully-connected strategy aggregates frame-wise associations from all the other frames.

3.3 Cross-modal Contrastive Learning

Although REC only requires to ground the object corresponding to the overall semantic meanings, we contend that conducting the fine-grained alignment for each object is crucial to differentiate the referent and the contextual objects, leading to more accurate localization results. Thus we propose to build the alignment for
We first introduce the experimental settings in Sec. 4.1. The comparison on the feature similarity scores. Given the $p^{th}$ patch of the $i^{th}$ frame $v_p^i$ and the $s$-th word feature $q^s$, we compute the similarity scores between them and selected the top Kcross responsive items as the positive samples:

$$w_p^i = \text{avgpool} \left( \arg \max_{s \in [1, S]} \left( v_p^i \cdot q^s \right) \right).$$  

(5)

where $w_p^i$ is the pooled positive cross-modal features for the image patch $v_p^i$, $K_{cross} = \max(1, \lfloor S/R_{cross} \rfloor)$, where $S$ is the total word length and $1/R_{cross}$ is the sampling ratio.

Then we build the cross-modal contrastive learning based on the mined patch-word correspondence as follows:

$$L_{cross} = \frac{1}{p} \sum_{p=1}^{P} -\log \frac{\exp \left( \frac{v_p^i \cdot w_p^i}{\tau} \right)}{\sum_{i'=1}^{S} \exp \left( \frac{v_p^i \cdot q^{i'}}{\tau} \right)}.$$  

(6)

Thus the final cross-modal contrastive loss is represented as $L_{cross} = \frac{1}{T} \sum_{i=1}^{T} L_i$.

3.4 Training Objectives

Besides the aforementioned inter-frame and cross-modal contrastive loss, we also apply a localization loss $L_{loc}$ to regress the bounding box towards the ground-truth and a classification loss $L_{cls}$ to select the best matching prediction term [12, 44]. We apply these two losses based on the cross-modal features illustrated in Sec. 3.1. For each spatial position of the cross-modal features, we generate bounding box predictions centered at the current location, together with a confidence score to indicate the probability to be the final grounding output. Specifically, the regression and classification losses are implemented as the MSE (mean square error) loss and the cross-entropy loss, respectively. During inference, the bounding box with the highest confidence is selected as the final prediction. Please refer to [39, 56] for more details.

Integrating the above constraints, final loss function is as follows.

$$L = \lambda_{loc} L_{loc} + \lambda_{cls} L_{cls} + \lambda_{inter} L_{inter} + \lambda_{cross} L_{cross}.$$  

(7)

where $\lambda_{loc}$, $\lambda_{cls}$, $\lambda_{inter}$, and $\lambda_{cross}$ are balance factors.

4 EXPERIMENTS

We first introduce the experimental settings in Sec. 4.1. The comparisons with the state-of-the-art methods are discussed in Sec. 4.2. Then, we present the model architecture agnostic analysis in Sec. 4.3. The extensive exploratory studies on cross-modal correspondence and inter-frame correspondence learning are provided in Sec. 4.4 and Sec. 4.5, respectively. Note that all ablation studies are conducted on the VID-Sentence dataset. Further, we provide the visualization results in Sec. 4.6.

4.1 Experimental Settings

Datasets. We evaluated the performance of REC on two challenging video REC datasets and one image REC dataset. 1) VID-Sentence (VID) [12]: It is re-labeled by Chen et al. [12] based on the ImageNet video object detection dataset [41]. It includes 7,654 trimmed videos with annotated language captions. The size of the vocabulary is 1,823 and the average length of the descriptions is 13.2. Following [12, 44], we adopt 6,582 instance pairs for training and 536 pairs for testing. 2) Lingual OTB99 (LiOTB) [28]: This dataset is built based on the popular object tracking dataset OTB100 [33] which contains 100 videos. We take the same strategy as in [28, 44], splitting the dataset into 51 and 48 instances for training and testing, respectively. 3) RefCOCO [60]: It consists of 142,210 referring expressions for 50,000 objects in 19,994 images. The average length of each expression is 3.5 words. The dataset is split into train, validation, testA, and testB sets. The testA and testB sets contain images with multiple people and objects, respectively.

Evaluation Metrics. Following [27, 44, 57], we adopt “Accu.@a” as the metrics. Specifically, we compute the intersection over union (IoU) for the predicted bounding box and the ground truth per frame. The prediction for one video is considered as positive if the IoU of all frames exceeds the threshold $a$. We set $a$ to 0.4, 0.5, 0.6. Besides, following [44, 57], we additionally adopt “success” and “precision” as the evaluation metrics. The success score is the same as the AUC (area under curve) metric while the precision score represents the ratio of frames where the predicted bounding box falls within a threshold of 20 pixels around the target.

Implementation Details. We chose Darknet-53 [39] pre-trained on MSCOCO [30] and BERT [14] pre-trained on English Wikipedia and Toronto Book Corpus as the visual and language encoder, respectively. The long edge of the input image was resized to the size of 256 and then the whole image was padded to 256 × 256. The extracted image feature was with the multi-scale resolution, i.e., 8 × 8, 16 × 16 and 32 × 32. Random crop, affine transformation, horizontal flip, and color jitter were used for image data augmentation. The network was optimized with RMSProp [46] for 10 epochs in video datasets and 100 epochs in image dataset. The initial learning rate was set to 1e-4 and decayed with a polynomial schedule. Experiments were conducted on 4 V100 GPUs and the batch sizes for training VID, LiOTB, and RefCOCO were set to 32, 8, and 32, respectively. We set the balancing weight $\lambda_{loc} = 5$, $\lambda_{cls} = 1$, $\lambda_{inter} = 1$, and $\lambda_{cross} = 1$. During each iteration, the sampled frame number $T$ was set to 4, and the frame distance $F$ was set to 3. The temperature factor $\tau$ used in contrastive learning was set to 0.07 following [38, 52]. The positive sample sampling rate $1/R_{inter}$ and $1/R_{cross}$ are set to 1/8 and 1/3, respectively. Note that for the image dataset RefCOCO, we omitted the inter-frame contrastive loss $L_{inter}$.

4.2 Comparisons with State-of-the-Arts

Results on Video REC datasets. The comparison results on VID-Sentence (VID) and Lingual OTB99 (LiOTB) datasets are summarized in Table. 1 and Table. 2, respectively. Specifically, we provide three sets of comparison methods: 1) Per-frame grounding: We apply widely adopted image REC methods including Yang et. al. [56], DVSA [24], and GrounderK [40] for frame-wise grounding. 2) Per-frame tracking: We firstly use [56] to obtain the tracking template on the first/middle/last/random frame and employ tracker [26] to track the given template. 3) Other SOTA video REC methods: WSTG [12], LSAN [28], and Co-grounding [44].

---

4971
Table 1: Comparison (%) with state-of-the-art methods on VID-Sentence dataset.

| Method                                | Accu. | Succ. | Prec. |
|---------------------------------------|-------|-------|-------|
|                                       | 0.4   | 0.5   | 0.6   |
| Yang et al. (w/ BERT) [56]            | -     | 52.39 | 42.7  |
| Yang et al. (w/ LSTM) [56]            | -     | 54.78 | 45.1  |
| DVSA [24]-Avg                         | 36.2  | 29.7  | 23.5  |
| DVSA [24]-NetVLAD [2]                 | 31.2  | 24.8  | 18.5  |
| DVSA [24]-LSTM                        | 38.2  | 31.2  | 23.5  |
| GrounDeR [40]-Avg                     | 36.7  | 31.9  | 25.0  |
| GrounDeR [40]-NetVLAD [2]             | 26.1  | 22.2  | 15.1  |
| GrounDeR [40]-LSTM                    | 36.8  | 31.2  | 27.1  |
| First-frame tracking [26]             | -     | 36.97 | 33.4  |
| Middle-frame tracking [26]            | -     | 44.90 | 38.4  |
| Last-frame tracking [26]              | -     | 36.26 | 32.8  |
| Random-frame tracking [26]            | -     | 40.20 | 35.6  |
| WSTSTG [12]                           | 44.60 | 38.20 | 28.90 |
| Co-grounding [44]                     | 63.35 | 60.25 | 53.89 |
| DCNet (Ours)                          | 65.54 | 61.69 | 55.11 |

Table 2: Comparison (%) with state-of-the-art methods on LiOTB dataset.

| Method                                | Accu.@0.5 | Succ. | Prec. |
|---------------------------------------|-----------|-------|-------|
|                                       |           |       |       |
| Yang et al. (w/ BERT) [56]            | 49.13     | 35.8  | 46.8  |
| Yang et al. (w/ LSTM) [56]            | 49.16     | 33.3  | 41.4  |
| First-Frame Tracking [26]             | 50.93     | 39.1  | 48.2  |
| Middle-Frame Tracking [26]            | 43.08     | 35.6  | 42.1  |
| Last-Frame Tracking [26]              | 44.17     | 32.7  | 39.1  |
| Random-Frame Tracking [26]            | 25.16     | 28.8  | 32.9  |
| LSAN [28]                             |           | 25.9  |       |
| Co-grounding [44]                     | 52.26     | 39.2  | 50.0  |
| DCNet (Ours)                          | 53.21     | 41.0  | 51.7  |

Table 3: Comparison (%) with state-of-the-art methods on RefCOCO.

| Method                                | Vis.Enc. | val | testA | testB |
|---------------------------------------|----------|-----|-------|-------|
|                                       |          |     |       |       |
| Two Stage:                            |          |     |       |       |
| VC [65]                               | VGG16    | -   | 73.33 | 67.44 |
| ParalAttn [72]                        | VGG16    | -   | 75.31 | 65.52 |
| MAttN† [59]                           | RN101    | 76.40 | 80.43 | 69.28 |
| CM-Att-Erase [32]                     | RN101    | 78.35 | 83.14 | 71.32 |
| DGA [33]                              | VGG16    | -   | 78.42 | 65.53 |
| RvG-Tree [18]                         | RN101    | 75.06 | 78.61 | 69.85 |
| NMTTree† [31]                         | RN101    | 76.41 | 81.21 | 70.09 |
| Ref-NMS (CM-A-E) [8]                  | RN101    | 80.70 | 84.00 | 76.04 |
|                                       |          |     |       |       |
| One Stage:                            |          |     |       |       |
| RealGIN [70]                          | DN53     | 77.25 | 78.70 | 72.10 |
| FAOA [56]                             | DN53     | 71.15 | 74.88 | 66.32 |
| RCCF [29]                             | DLA34    | -   | 81.06 | 71.85 |
| MCN [34]                              | DN53     | 80.08 | 82.29 | 74.98 |
| ReSC† [55]                            | DN53     | 77.63 | 80.45 | 72.30 |
| LBYL [21]                             | DN53     | 79.67 | 82.91 | 74.15 |
| TransVG [13]                          | RN101    | 81.02 | 82.72 | 78.35 |
| Trar† [71]                            | DN53     | -   | 81.40 | **78.60** |
| Co-grounding [44]                     | DN53     | 77.65 | 80.75 | 73.37 |
| DCNet (Ours)                          | DN53     | **81.12** | **83.12** | 77.53 |

4.3 Architecture Agnostic

Loss Component Ablations. We ablate the loss components of our DCNet to see the difference. As shown in Table 5, both the inter-frame and cross-modal contrastive losses \( L_{inter} \) and \( L_{cross} \) are crucial. For example, when removing \( L_{inter} \) (mode #2 in Table 5), Accu.@0.4 drops by 4.55% compared to the full version (mode #1).

Extending to existing methods. Our inter-frame and cross-modal contrastive losses \( L_{inter} \) and \( L_{cross} \) are annotation-free and can be seamlessly inserted into existing video REC methods. As shown in Table 6, we select two representative methods including the one-stage method [44] and the two-stage method [12] as baselines. The results in Table 6 show that both \( L_{inter} \) and \( L_{cross} \) benefit the two baseline models. For example, on top of [44], \( L_{inter} \) and \( L_{cross} \) lead to 0.78% and 1.57% performance boost on Accu.@0.4, respectively. These results demonstrate that our inter-frame and cross-modal contrastive losses are general and compatible with various video REC methods.

4.4 Ablations on Inter-frame Correspondence

Sparse v.s. Dense Correspondence. We design the inter-frame contrastive learning to build the dense patch-wise correspondence...
Table 4: (a) Comparisons of different alignment manners in inter-frame contrastive learning. Dense: our proposed patch-wise correspondence; Sparse: alignment conducted between annotated bounding box features. Refer to Figure 2 for schematic illustrations. (b) Comparisons of different aggregation manners. adjacent: aggregating from the nearest frames; fc (fully-connected): aggregating from all the other frames. Refer to Figure 5 for schematic illustrations. (c) Ablations of the positive sample ratio 1/R_inter. (d) Ablations of cross-modal correspondence discovery strategies. (f) Ablations of the positive sample ratio 1/R_cross.

| Mode   | Accu.@0.4 | @0.5  | @0.6  |
|--------|-----------|-------|-------|
| Dense  | 65.54     | 61.69 | 55.11 |
| Sparse | 64.15     | 61.07 | 54.12 |

| 1/R_inter | Accu.@0.4 | @0.5  | @0.6  |
|-----------|-----------|-------|-------|
| 1/12      | 64.81     | 61.02 | 54.68 |
| 1/8       | 65.54     | 61.69 | 55.11 |
| 1/4       | 61.99     | 60.64 | 53.89 |

| Mode    | Accu.@0.4 | @0.5  | @0.6  |
|---------|-----------|-------|-------|
| adjacent| 65.62     | 61.83 | 55.26 |
| fc      |           |       |       |

| 1/R_cross | Accu.@0.4 | @0.5  | @0.6  |
|-----------|-----------|-------|-------|
| 1/5       | 62.72     | 59.04 | 52.11 |
| 1/3       | 65.54     | 61.69 | 55.11 |
| 1/2       | 63.28     | 60.36 | 53.94 |

Table 5: Ablations of loss components.

| Mode | L_inter | L_cross | Accu.@0.4 | @0.5  | @0.6  |
|------|---------|---------|-----------|-------|-------|
| #1   | ✓       | ✓       | 65.54     | 61.69 | 55.11 |
| #2   | ✓       | ✗       | 60.99     | 59.64 | 52.89 |
| #3   | ✗       | ✓       | 63.13     | 60.09 | 53.05 |
| #4   | ✗       | ✗       | 57.86     | 54.78 | 49.35 |

Table 6: Architecture Agnostic analysis. We insert the proposed inter-frame and cross-modal contrastive losses L_inter, L_cross to existing video REC methods [12, 44].

| Method     | Accu.@0.4 | @0.5  | @0.6  |
|------------|-----------|-------|-------|
| Co-grounding [44] | 63.35 | 60.25 | 53.89 |
| + L_inter   | 64.13     | 61.24 | 54.09 |
| + L_cross   | 64.92     | 61.24 | 54.21 |
| + L_cross + L_inter | 65.58 | 61.73 | 55.41 |
| WSSTG [12]  | 44.60     | 38.20 | 28.90 |
| + L_inter   | 46.81     | 39.92 | 30.64 |
| + L_cross   | 47.14     | 40.41 | 31.57 |
| + L_cross + L_inter | 48.66 | 42.38 | 32.83 |

between frames. As illustrated in Figure 2, there exists a simple manner to conduct inter-frame alignment between annotated ground-truth, which is called sparse alignment here. We conduct comparison experiments for both the dense and sparse alignment manners.

As shown in Table 4a, our proposed dense alignment outperforms the sparse one in all three metrics. Besides, our dense alignment requires no supervised annotations while the sparse alignment needs ground-truth annotations.

Multi-frame Aggregation Manner. As shown in Figure 5, we design two potential ways to aggregate the total T frames in the inter-frame contrastive learning. Specifically, the adjacent strategy only considers the nearest frames while the fully-connected strategy aggregates frame-wise associations from all the other frames.

The results in Table 4b show that the fully-connected aggregation manner slightly outperforms the adjacent aggregation one. For example on Accu.@0.4, the former achieves 0.08% boost compared to the latter one. Considering the computational costs of the adjacent aggregation are dramatically lower than that of the fully-connected one (O(T) v.s. O(T^2)), we adopt the adjacent aggregation manner.

Ablations of the Sample Distance F. We conduct the ablation studies on the sampled frame distance F to determine the optimum choice. From the results in Table 4c, we can observe that the optimum performance is achieved when setting F = 3. Too large or too small F value will both lead to the performance degradation.

Ablations of the Positive Sample Ratio 1/R_inter. We conduct ablative experiments on the positive sample ratio 1/R_inter in the inter-frame contrastive learning. The results in Table 4d show that the performance saturates at 1/R_inter = 1/8. This may be because too few patch-wise pairs are unable to establish accurate correspondences while setting too many positive samples makes it hard to find the most informative pairs.

4.5 Ablations on Cross-modal Correspondence

Cross-modal contrastive loss aims to capture the fine-grained patch-word alignment. In this section, we offer more potential correspondence discovery solutions to find the most suitable one.

Cross-modal Correspondence Discovery. In Sec. 3.3, we propose to find the cross-modal correspondence by selecting the most similar K_cross words for each patch region. Here we refer to it patch-topk. We provide two comparison discovery strategies: 1) word-topk: we select top-K_cross most responsive patches for each word; 2) random: we randomly select K_cross for each patch feature.

As shown in Table 4e, the random selection strategy performs the worst with Accu.@0.4 only reaching 61.35%. Besides, our patch-topk discovery strategy outperforms the word-topk manner. For example...
**Query:** An orange car is drifting on the road.

**Query:** A white fox is sitting and barking on the ground.

---

**Figure 6:** (a) Input image frames. Localization results of DCNet w/o and w/ the inter-frame contrastive loss \( L_{\text{inter}} \) are marked in green and red, respectively. Ground-truth annotations are marked in blue. (b) Confidence score maps w/o \( L_{\text{inter}} \). (c) Confidence score maps w/ \( L_{\text{inter}} \). The score maps of DCNet w/ \( L_{\text{inter}} \) are more stable across video frames than that of DCNet w/o \( L_{\text{inter}} \). Pay special attention to the #2 frame of the two examples.

**Figure 7:** Visualizations of the cross-modal alignment. Sub-query is the specific word taken from the full query. The response map is generated by computing the cosine similarity between the sub-query and the input image.

---

Maps for DCNet w/ and w/o \( L_{\text{inter}} \). We present two video examples and we select three consecutive frames for each video.

As shown in Figure. 6, DCNet w/o \( L_{\text{inter}} \) leads to unstable and inconsistent feature responses. For example in Figure. 6 (left), due to the motion blur in frame #2, DCNet w/o \( L_{\text{inter}} \) falsely localizes the interfering object in front (marked in green). In contrast, introducing \( L_{\text{inter}} \) helps rectify the localization results in frame #2, leading to more stable results across frames.

**Visualizations of Cross-Modal Alignment.** The cross-modal contrastive loss \( L_{\text{cross}} \) builds fine-grained patch-word alignment. To prove this, we select specific words in the language query (denoted as sub-query in Figure. 7) and compute its cosine similarity with all patches within the frame.

As shown in Figure. 7, the similarity distribution focuses more on the areas semantically corresponding to the sub-query. These results demonstrate that our cross-modal contrastive loss \( L_{\text{cross}} \) can effectively model the patch-word alignment.

---

**5 CONCLUSIONS**

In this paper, we summarized two common problems in video REC, i.e., inconsistent localization and confusion between the referent and contextual objects. To this end, we proposed DCNet which introduces the inter-frame patch-wise and cross-modal patch-word correspondence without the reliance on annotations. Specifically, we estimated the patch-wise and patch-word alignment by computing the cosine similarity in feature space. Then the most responsive pairs were selected as positive samples, where the corresponding inter-frame and cross-modal contrastive learning were applied. Extensive experimental results on both video and image REC datasets have demonstrated the effectiveness of our proposed DCNet.

**Acknowledgements.** This paper was partially supported by NSFC (No: 62176008) and Shenzhen Science & Technology Research Program (No:GXWD20201231165807007-20200814115301001).
Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised...

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4252–4261.

Meng Cao, Ji Jiang, Long Chen, & Yuexian Zou. 2022. Unsupervised pre-training for temporal action localization tasks. In Proceedings of the 28th ACM International Conference on Multimedia. 1939–1947.

Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. 2021. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16010–16019.

Can Zhang, Meng Cao, Dongming Yang, Ji Jiang, and Yuexian Zou. 2021. Synergic learning for noise-insensitive weakly-supervised temporal action localization. Image and Vision Computing 113 (2021), 104247.

Can Zhang, Tianyu Yang, Junwu Weng, Meng Cao, Jie Wang, and Yuexian Zou. 2022. Unsupervised pre-training for temporal action localization tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 14031–14041.

Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. 2018. Grounding referring expressions in images by variational context. In CVPR 4158–4166.

Yuting Zhang, Luyao Yuan, Yijie Guo, Zhuyuan He, J-An Huang, and Honglak Lee. 2017. Discriminative bimodal networks for visual localization and detection with natural language queries. In CVPR. 557–566.

Zhu Zhang, Zhou Zhao, Zhiye Lin, Baoxing Huai, and Nicholas Jing Yuan. 2020. Weakly-supervised temporal action localization with snippet contrastive learning. arXiv preprint arXiv:2008.06941 (2020).

Zhu Zhang, Zhou Zhao, Qiang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. 2020. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10668–10677.

Luowei Zhou, Nathan Louis, and Jason J Corso. 2018. Weakly-supervised video object grounding from text by loss weighting and object interaction. arXiv preprint arXiv:1805.02834 (2018).

Yi Yi Zhou, Rongrong Ji, Gen Luo, Xiaoshuai Sun, Jinrong Su, Xinghao Ding, Chai-Wen Lin, and Qi Tian. 2021. A real-time global inference network for one-stage referring expression comprehension. IEEE Transactions on Neural Networks and Learning Systems (2021).

Yi Yi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. 2021. TRAR: Routing the Attention Spans in Transformer for Visual Question Answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2074–2084.

Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. 2018. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4252–4261.