Semi-supervised Video Paragraph Grounding with Contrastive Encoder

Xun Jiang[1], Xing Xu[1]*, Jingran Zhang[1], Fumin Shen[1], Zuo Cao[2], Heng Tao Shen[1,3]
[1]Center for Future Media & School of Computer Science and Engineering
University of Electronic Science and Technology of China [2]MeiTuan [3]Peng Cheng Lab

Abstract

Video events grounding aims at retrieving the most relevant moments from an untrimmed video in terms of a given natural language query. Most previous works focus on Video Sentence Grounding (VSG), which localizes the moment with a sentence query. Recently, researchers extended this task to Video Paragraph Grounding (VPG) by retrieving multiple events with a paragraph. However, we find the existing VPG methods may not perform well on context modeling and highly rely on video-paragraph annotations. To tackle this problem, we propose a novel VPG method termed Semi-supervised Video-Paragraph TRansformer (SVPTR), which can more effectively exploit contextual information in paragraphs and significantly reduce the dependency on annotated data. Our SVPTR method consists of two key components: (1) a base model VPTR that learns the video-paragraph alignment with contrastive encoders and tackles the lack of sentence-level contextual interactions and (2) a semi-supervised learning framework with multimodal feature perturbations that reduces the requirements of annotated training data. We evaluate our model on three widely-used video grounding datasets, i.e., ActivityNet-Caption, Charades-CD-OOD, and TACoS. The experimental results show that our SVPTR method establishes the new state-of-the-art performance on all datasets. Even under the conditions of fewer annotations, it can also achieve competitive results compared with recent VPG methods.

1. Introduction

Localizing events in a given untrimmed video is one of the challenging video understanding tasks, which is first proposed by [1, 7]. Following their works, a list of promising methods [16, 20, 45–47, 50] has been proposed. However, most existing methods focus on *Video Sentence Grounding* (VSG), addressing this problem in "*single-multi*" approaches (as shown in Fig. 1(a)), they ground a moment from a video that consists of several different events according to an individual sentence query. Contrastively, as illustrated in Fig. 1(b), *Video Paragraph Grounding* (VPG), which is recently proposed by [2], addresses the video events grounding task in the "*multi-multi*" manner. Specifically, in the VPG task, given a paragraph describing multiple events instead of a single sentence, it is expected to localize all of the related moments in an untrimmed video. Since a paragraph consisting of multiple sentences in time order contains more temporal information compared with the single sentence input, it is more informative for retrieving moments in videos.

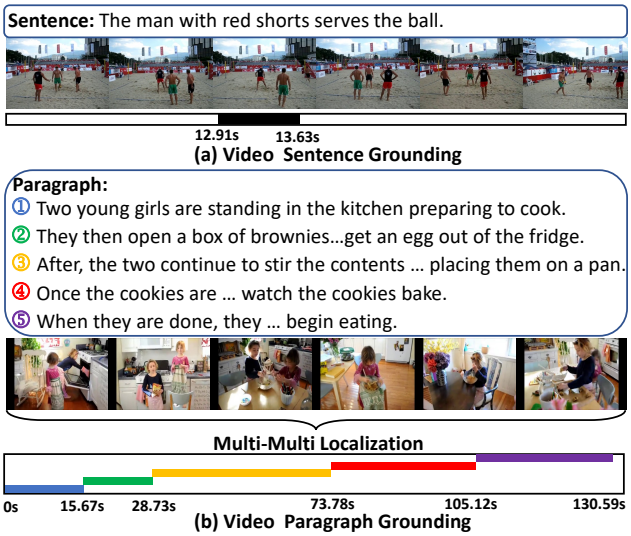

Figure 1. An illustrative example of VSG and VPG: (a) VSG aims at retrieving a particular moment with a single sentence. (b) VPG receives a paragraph consisting of multiple sentences as a query and localizes multiple events in the untrimmed video.

The previous VPG methods [2, 6, 50] first generate proposals for each sentence, then learn temporal order and semantic relations among these proposals to select desired candidates. Nevertheless, these methods exist three problems. Firstly, they rely on the temporal information of paragraphs but hardly exploit the contextual information well from the perspective of text modality. For example, shown in Fig. 1(b), all the sentences in the paragraph are describ-

*Corresponding author.

ing two girls cooking in the kitchen and each of the sentences is related to others contextually around the cooking topic. Recently, the informal work [30] tried to tackle VPG with Transformer [34], which proved the global contexts worked in this task. However, these methods including [30] still fall into the second defect: with the paragraph input, they only focus on the proposal-sentence matching but ignore the video-paragraph matching, which may lead to misalignment on cross-modal fusion. Lastly, compared with moment-sentence annotation, the video-paragraph annotated data are more expensive and hard to generate. All these VPG methods are required to be trained with temporal labeled data, which brings heavy costs to this task. Although there are also some weakly-supervised video grounding methods [5, 19, 43], most of them are "single-multi" methods and the performance is much worse than fully-supervised methods.

To tackle these problems, we first propose a novel base model termed *Video-Paragraph TRansformer* (VPTR), which introduces contrastive learning and semi-supervised learning into VPG. We further extend it to the semi-supervised version, the *Semi-supervised Video-Paragraph TRansformer* (SVPTR), to reduce the dependency on temporal annotations. Specifically, as the general framework of our proposed SVPTR method shown in Fig. 2, to explore the contextual information hidden in paragraphs, we extract hierarchical text features and design a sentence-based query mechanism in the decoder. The individual sentence queries interact with particular words and other sentences with such designs thus we can extract more contextual information. Moreover, to avoid misalignment between proposal moments and sentences, contrastive learning is introduced into the multimodal encoder to guide the cross-modal fusion at the video-paragraph level. As is shown in Fig. 2, the contrastive encoder separately encodes the two modalities and projects them into a common space via self-supervised learning. Finally, we develop an advanced semi-supervised learning VPG method SVPTR that is based on the teacher-student framework, which effectively reduces the consumption of video-paragraph temporal annotations.

The primary contributions in this work are as follows:

- We explore contextual information in the paragraph query with hierarchical text features and the sentence-based query mechanism. It effectively improves the precision of localizing events in untrimmed videos.
- We combine self-supervised learning to optimize the cross-modal fusion in video paragraph grounding. Particularly, we design a contrastive loss at the video-paragraph level without proposing moment candidates.
- We design a semi-supervised learning framework for VPG and achieve promising results with less annotated data. To the best of our knowledge, we are the first to explore the semi-supervised learning on video paragraph grounding.

To evaluate the proposed SVPTR method, we conduct extensive experiments on three widely-used datasets: ActivityNet-Caption [12], Charades-CD-OOD [44], and TACoS [27]. The comprehensive results demonstrate the superiority of our SVPTR method compared with a handful of state-of-the-art VPG approaches under both fully-supervised and semi-supervised settings.

2. Related Work

Video Sentence Grounding. Video Sentence Grounding (VSG) is first proposed by [1, 7], which determines the start and end time points of a single event by a query sentence. Early works in VSG [1, 7, 8, 45, 48, 50] adopt a two-stage model that first generates proposal candidates then models these video segments and sentence queries jointly. Meanwhile, a part of VSG methods [15, 20, 22, 47] follow the proposal-free model and treat the VSG as a regression task that predicts the timestamps directly. Recently, researchers [16, 39, 49, 51] are studying new frameworks to localize the target events. Notably, Zhang *et al*. [49] introduced Transformer [34] into VSG to improve the quality of cross-modal modeling, which proved its effectiveness for video grounding task. Additionally, there are also weakly-supervised methods [5, 19, 43] tackling the VSG problem from the perspective of overcoming the costs of temporal annotations. Since the lack of location information, these methods usually perform much worse than popular fully-supervised methods. Recently, an informal work [18] proposed a novel VSG framework that achieves a trade-off between annotation costs and performance with semi-supervised learning. Nevertheless, most of these methods above are limited in single event grounding and not suitable for localizing multiple events at the same time.

Video Paragraph Grounding. Different from VSG, Video Paragraph Grounding (VPG) treats video events grounding as a "multi-multi" problem. Specifically, the natural language query in VPG is a long paragraph, which describes multiple events in the untrimmed video. The VPG is first defined by Bao *et al*. [2] as the inverse problem of the Dense Video Caption [35]. In this initial work [2] for VPG, they first extended two previous methods [6, 50] to baselines, then designed a novel VPG model DepNet. Following that, a recent informal work [30] proposed a method that utilized Transformer [34] to tackle this task.Although these methods above have achieved promising improvement compared with conventional VSG methods, they ignore crucial contextual information hidden in paragraphs or the high costs of annotations caused by paragraph input, which motivates us to develop this work.

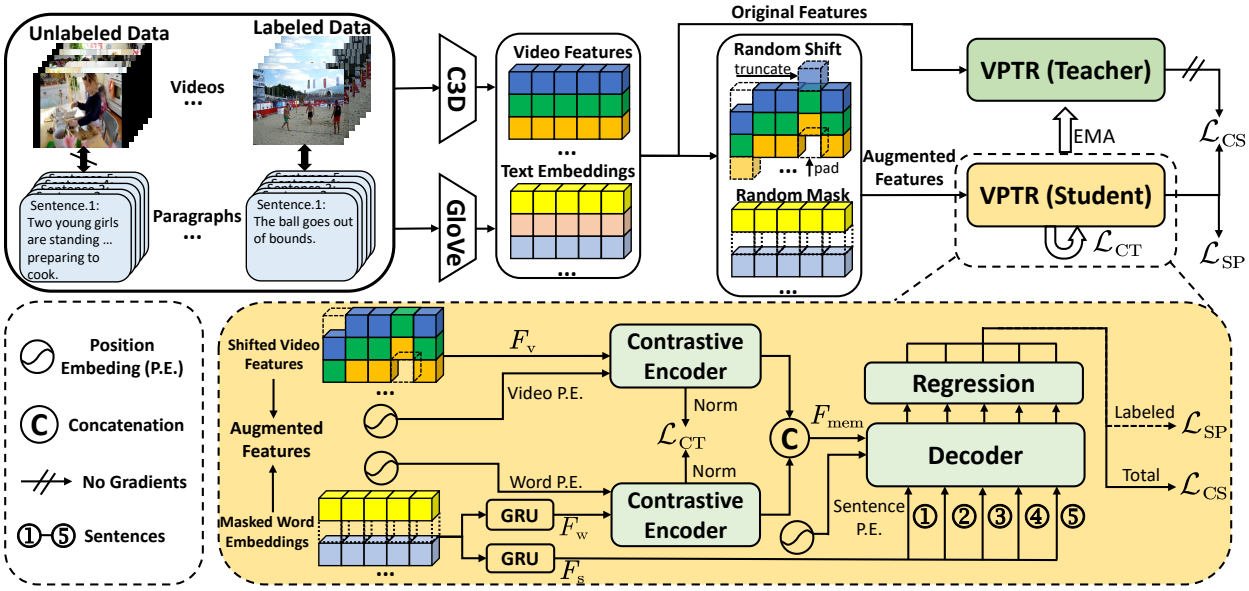

Figure 2. An overview of our SVPTR method. The first part on the top shows the semi-supervised framework of our method, which maintains a teacher model and a student model. Two models receive different inputs via multimodal feature perturbations. We update the teacher model with Exponential Moving Average (EMA) instead of backpropagation. The second part in the yellow box illustrates the details of the proposed base model VPTR.

Semi-supervised Learning. Semi-supervised Learning (SSL) is a class of dominant methods in machine learning to learn from limited labeled data and massive unlabeled data. In general, it can be roughly categorized into two types: pseudo label generation and consistency regularization. The former [14, 25, 40] usually make predictions from a model trained on labeled data to impute approximate labels for unlabeled data, while the latter [13, 31] encourage models to reduce the discrepancy between predictions made perturbed data points. Recently, some works [10, 18, 37] introduced semi-supervised learning into video understanding task. Ji *et al*. [10] designed two sequential perturbations based on the mean teacher framework. Wang *et al*. [37] combined self-supervised learning and SSL to reduce the reliance on labeled data for Temporal Action Proposal. Meanwhile, Luo *et al*. [18] recently explored the SSL in VSG, which is the most related to our work.

3. Proposed Method

3.1. Problem Formulation

Given an untrimmed video V and a paragraph P consisting of K sentences, our goal is to localize timestamps of K events in the video which are most related to these sentences respectively. Specifically, we represent the untrimmed video V as $\mathrm{V} = \{\mathrm{v}_i\}_{i=1}^{l^{\mathrm{V}}}$, where l^{V} is the frame number of V. The paragraph P is presented as $\mathrm{P} = \{\mathrm{S}_i\}_{i=1}^{\mathrm{K}}$ where S_i denotes i-th sentence in the paragraph. Let t_{s} and t_{e} be the start time and end time of one target video segment respectively, the VPG task can be formulated as follows:

$$\mathrm{M}_{\mathrm{VPG}}(\mathrm{V}, \mathrm{P}) \rightarrow \{(t_{\mathrm{s}}, t_{\mathrm{e}})_i\}_{i=1}^{\mathrm{K}}, \quad t_{\mathrm{s}} < t_{\mathrm{e}}, \tag{1}$$

where $(t_{\mathrm{s}}, t_{\mathrm{e}})_i$ is the retrieved result for i-th sentence in P.

3.2. Video-Paragraph TRansformer

Depicted in Fig. 2, we first propose a base model VPTR, which tackles the problems of learning contextual information and aligning the two modalities.

Video Modality. Given an untrimmed video V with l^{V} frames, we divide them into a group of small clips without overlap and each clip contains the same constant number of frames. Afterward, we extract visual features with pre-trained C3D backbones [32], where l^F denotes the total length of video features. Let $m_v(\cdot)$ be the 3D CNN backbone and $\phi_v(\cdot)$ be the projecting layer with normalization, the video feature extraction can be expressed as $\mathbf{F}_v = \phi_v(m_v(V)) = \{\mathbf{f}_i\}_{i=1}^{l^F}$.

Text Modality. Most existing works in VSG [15, 16, 51] denote the query into a group of word-level features. However, in VPG, we receive multiple sentences from the paragraph query, where each sentence contains different semantics. To this end, we hierarchically extract the text features from paragraph input. Specifically, we first extract the word-level features from the whole paragraph to learn global semantics. Following that, we extract the sentence-level features from each sentence individually. Given a paragraph query $\mathrm{P} = \{\mathrm{S}_i\}_{i=1}^{l^{\mathrm{S}}} = \{\mathrm{W}_j\}_{j=1}^{l^{\mathrm{W}}}$, we use 2-layer bidirectional GRUs to obtain the word-level features

$\mathbf{F}_w = \{\mathbf{w}_j\}_{j=1}^{l^{\mathrm{W}}}$, where l^{S}, l^{W} denote the number of sentences and words in the paragraph respectively and w_j denote j-th word features. Moreover, we split the paragraph into sentences and each sentence is encoded by the GRUs individually. The sentence-level feature of i-th sentence in the paragraph is obtained by the concatenation of hidden states in both directions, which is denoted as $\mathbf{s}_i$. Hence, we obtain the sentence-level features of the whole paragraph $\mathbf{F}_s = \{\mathbf{s}_i\}_{i=1}^{l^{\mathrm{S}}}$. Formally, the language feature extraction can be represented as following:

$$\begin{cases} \mathbf{s}_i = \mathrm{BiGRU}\left(\mathbf{w}_k^i, h_{k-1}^i\right), \\ \mathbf{w}_j = \mathrm{BiGRU}\left(\mathbf{w}_j, h_{j-1}\right), \end{cases} \tag{2}$$

where $\mathbf{w}_k^i$ represents the k-th word in i-th sentence and the h is the hidden state of GRUs. Afterward, the word-level features $\mathbf{F}_w$ are processed by a projecting layer with normalization $\phi_w(\cdot)$, which is similar to video modality.

Contrastive Encoder. Inspired by the success of multimodal Transformer [33], the recent work [49] introduces it into VSG and achieves competitive results. However, we observed that the mixed input of encoders results in degeneration of learning intra-modality information. To maintain the intra-modality modeling and obtain inter-modality aligning for VPG, we design a contrastive encoder, learning the semantic consistency [42] via self-supervision. As illustrated in Fig. 3(a), our contrastive encoder projects the video features $\mathbf{F}_v$ and word-level text features $\mathbf{F}_w$ into a common subspace for semantic alignment [41]. The objective of our contrastive encoder is to pull the positive video and paragraph pairs together and push the negative pairs away. Specifically, as depicted in Fig. 2, we apply the transformer encoder [3] $\Phi(\cdot)$ and normalization layer Norm to acquire the transferred video feature $\tilde{\mathbf{F}}_v = \mathrm{Norm}(\Phi(F_v))$ and text features $\tilde{\mathbf{F}}_w = \mathrm{Norm}(\Phi(\mathbf{F}_w))$. Additionally, we construct a triplet tuple $\left(\tilde{\mathbf{F}}_v^+, \tilde{\mathbf{F}}_w^+, \tilde{\mathbf{F}}_w^-\right)$, where $\left(\tilde{\mathbf{F}}_v^+, \tilde{\mathbf{F}}_w^+\right)$ is a positive pair and $\left(\tilde{\mathbf{F}}_v^+, \tilde{\mathbf{F}}_w^-\right)$ a negative pair, for our contrastive learning. Let $\mathcal{T}_{\mathrm{v}}$ and $\mathcal{T}_{\mathrm{w}}$ represent triplet tuples $\left(\tilde{\mathbf{F}}_v^+, \tilde{\mathbf{F}}_w^+, \tilde{\mathbf{F}}_w^-\right)$ and $\left(\tilde{\mathbf{F}}_w^+, \tilde{\mathbf{F}}_v^+, \tilde{\mathbf{F}}_v^-\right)$ respectively, the contrastive loss can be expressed as:

$$\mathcal{L}_{CT} = \sum_{(\tilde{\mathbf{F}}_v^+, \tilde{\mathbf{F}}_w^+)} \left\{ \sum_{\tilde{\mathbf{F}}_w^-} \mathcal{L}_T(\mathcal{T}_{\mathrm{v}}) + \sum_{\tilde{\mathbf{F}}_v^-} \mathcal{L}_T(\mathcal{T}_{\mathrm{w}}) \right\}, \tag{3}$$

where $\mathcal{L}_T(\cdot)$ [19] ensures the positive pair's similarity score is better than the negative pair's by at least a margin. Meanwhile, we concatenate the encoded features of video and paragraph together, which are used as the memory of decoder. Let $[;]$ denote the concatenating operation, we obtain the multimodal memory $\mathbf{F}_{\mathrm{mem}} = [\mathbf{F}_v; \mathbf{F}_w]$.

Sentence-based Decoder. In previous work [2], multiple sentences brought more temporal information, however, the contextual information among these sentences was mined insufficiently. To overcome this defect, our decoder receives sentence-level features from the paragraph query rather than learnable embeddings. As depicted in Fig. 3(b), this mechanism conducts the sentence-level context modeling, turning each sentence gets related temporally and contextually. Concretely, given the sentence-level features $\mathbf{F}_s = \{\mathbf{s}_i\}_{i=1}^{l^{\mathrm{S}}}$, we first use the position embedding layer to encode temporal information, and then the interactions among sentences are conducted by the self-attention layers. Following that, each sentence query generates a feature for describing events from the multimodal memory via cross-attention layers. With the hierarchical text features, the sentence-level features can also interact with particular words, allowing the decoder to learn more contextual information. Finally, a parallel regression layer is employed to compute the timestamps of each sentence-wise feature. The procedures of sentence-based decoder are formulated as follows:

$$\mathbf{T} = \mathrm{MLP}\left(\Psi\left(\mathbf{F}_{mem}, \mathbf{F}_s\right)\right), \tag{4}$$

where the $\Psi(\cdot)$ represents the decoders of transformer which conduct the position embedding on sentence level. $\mathbf{T}$ is the localizing results consisting of l^{S} valid timestamps.

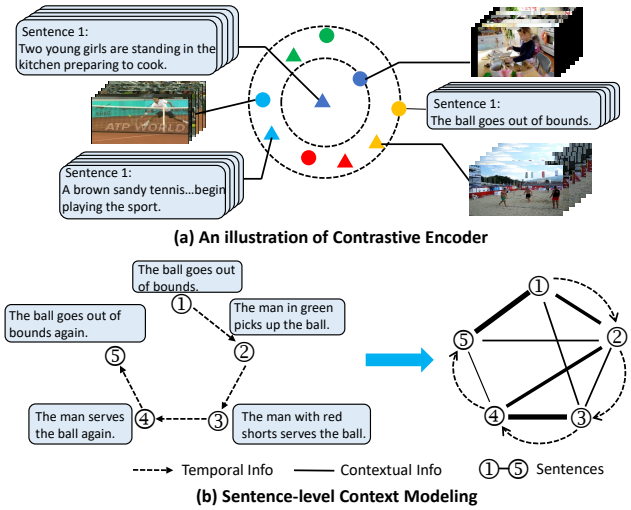

Figure 3. A detailed illustration for VPTR. (a) The contrastive encoder projects multimodal pairs with similar semantics closer. The circles and triangles represent text features and video features respectively. (b) The sentence-based decoder maintains sentence-level temporal and contextual information. The thickness of lines represents the strength of contextual connection.

3.3. Semi-supervised VPTR

Based on the teacher-student framework, we extend our VPTR to Semi-supervised VPTR (SVPTR), which consists of two base models and a multimodal feature perturbation module. With such a semi-supervised learning pipeline, we significantly reduce the consumption of annotated data.

Feature Perturbations. In previous works on semi-supervised learning [31, 37], stochastic perturbations have been found effective for improving the robustness of models. Moreover, the perturbations can also be regarded as data augmentation that helps self-supervised learning. Illustrated in Fig. 2, we conduct the feature perturbations both on two modalities. For video modality, we follow [37] and employ the random temporal shift as our video perturbation module. Concretely, we randomly choose μ channels first, then $\mu/2$ feature channels are moved forward and the other $\mu/2$ channels are moved backward. Different from the conventional temporal shift, the random selection brings more diversity into the perturbation, which augments the video modality features for semi-supervised learning and self-supervised learning. As for text modality, inspired by BERT [4], we randomly mask a part of words from the whole paragraph. Furthermore, we also apply dropout strategy with probability hyperparameter λ on both modalities.

Mean Teacher Framework. Mean Teacher [31] is a semi-supervised learning method based on consistency regularization. As illustrated in Fig. 2, two base models are maintained: a student VPTR model Γ and the teacher VPTR model Γ'. The student model Γ learns from the annotated data, and the Contrastive Encoder also allows us to train it with self-supervised learning. The teacher model Γ' is a duplicate of the student model, whose weights are updated with a sequence of student models during training via the Exponential Moving Average (EMA) strategy. Specifically, the Mean Teacher Framework can be formulated as follow:

$$\Gamma'_t = \tau\Gamma'_{t-1} + (1-\tau)\,\Gamma_t, \tag{5}$$

where t, τ denote the number of training iterations and smoothing coefficient respectively.

The input of each iteration consists of labeled and unlabeled data. During the training, the student VPTR receives input from the feature perturbation module and predicts the timestamps, while the teacher VPTR is fed by the original data to get predictions. Using the predictions from the two models, we compute a consistency loss, which induces our model to learn from labeled and unlabeled data jointly:

$$\mathcal{L}_{\text{CS}} = \frac{1}{N}\sum_{i=1}^{N} \|\Gamma\left(X\right)_i - \Gamma'\left(X'\right)_i\|^2, \tag{6}$$

where X and X' represent agumented features and original features respectively, and N denotes the total number of events in each input.

3.4. Objective Function

The loss function of the proposed SVPTR method consists of three parts: supervised loss, contrastive loss and consistency loss. We represent the overall loss as follows:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{SP}} + \beta\mathcal{L}_{\text{CT}} + \gamma\mathcal{L}_{\text{CS}}. \tag{7}$$

The supervised loss guides our model to learn location information from the annotated data. Concretely, for total N events, we denote the supervised loss as follows:

$$\mathcal{L}_{\text{SP}} = \frac{1}{N}\sum_{i=1}^{N}\left(\|T_i - \hat{T}_i\|_1 + \mathcal{L}_{\text{iou}}\left(T_i, \hat{T}_i\right) + \mathcal{L}_{\text{attn}}\left(a_i, \hat{a}_i\right)\right), \tag{8}$$

where the T_i, $\hat{T}_i$ represent the predictions and ground truth respectively. $\mathcal{L}_{\text{iou}}$ is temporal IoU loss which based GIoU loss [28]. $\mathcal{L}_{\text{attn}}$ is the attention guided loss for cross-attention layers in decoder referring to [20], where a_i and $\hat{a}_i$ represent the attention weights of valid video features and the fine-grained one-hot ground truth in i-th event. α, β, γ are hyperparameters for balancing different part of losses.

4. Experiments

4.1. Experimental Settings

Following the previous VPG methods [2, 50], we evaluate our SVPTR method on three benchmark datasets:

ActivityNet-Caption (Activity). [12] It is the largest dataset in video grounding task, which contains around 20k open domain videos. On average, each video contains 3.65 queries, and each query has an average of 13.48 words. Since the original test set is not released, we follow the previous work [2] and split the dataset into the training, val_1, val_2 of 10009/37421, 4917/17505, and 4885/17031 video/sentence respectively, where the val_2 is used for test.

Charades-CD-OOD (Charades). [7, 44] The dataset contains 6672 indoor daily life videos and is first released by Gao *et al*. [7] namely Charades-STA. To better evaluate the effectiveness of existing VSG methods, Yuan *et al*. [44] re-organizes the original dataset and splits it into train, val, and test_ood with 4563/11071, 333/859, and 1442/3375 video/sentence respectively, in which the training and testing data are designed to have different distributions.

TACoS. [27, 29] It is based on MPII Cooking Composite Activities video corpus [29] and enriched by Regneri *et al*. [27] with natural language descriptions and temporal annotations. All the videos are in kitchen room and the videos are much longer than the other two datasets. A standard split consists of 75/10146, 27/4589, and 25/4083 video/sentence pairs for training, validation, and testing.

Implementation Details. Following the previous work [2, 7, 50], we use pre-trained C3D [32] model without fine-tuning to extract the video features, and employ GLoVe embeddings [24] to receive the text vector representations. For Activity, the video features are preprocessed by PCA [12]. All the video features are sampled evenly to a fixed length L first. As for the videos which are shorter than L, we apply the zero padding and avoid the invalid padding features with padding masks. We set the number of encoder and decoder layers to 2 for all datasets. The smoothing coeffi-

cient for EMA in Mean Teacher is set to 0.999. In feature perturbations, μ and λ are set to 64 and 0.2 respectively. We train the model with Adam optimizer [11], which has a fixed learning rate 4×10^{-5}. The weight decay factors are set to 1×10^{-5} for three datasets. Moreover, we follow the metrics adopted in most video grounding works, which is denoted as *Recall@k, IoU=m*, where k is the number of generated candidates and m is the threshold. In our methods, k is set to 1 since the SVPTR predicts the timestamps directly, while m is set to $\{0.3, 0.5, 0.7\}$, $\{0.3, 0.5, 0.7\}$, and $\{0.1, 0.3, 0.5\}$ for Activity, Charades, and TACoS respectively. We also adopt the *mIoU* metric, which shows the average effect of our model. More detailed implementations are reported in the supplementary materials.

4.2. Overall Comparison Results

We compare our proposed SVPTR with the existing state-of-the-art VPG methods DepNet [2], and two natural extension methods Beam Search and 3D-TPN, which are reported by Bao *et al*. [2]. Moreover, to show the superiority of exploring the contextual information hidden in paragraph, we also compare our model with recent VSG methods, including CTRL [7], ACRN [17], WSSL [5] ,ABLR [46], [38], 2D-TAN [50], DRN [47], CBP [36], LGI [20], CPNet [15], BPNet [39], CBLN [16], DeNet [51], MATN [49], I^{2}N [21]. Furthermore, Luo *et al*. [18] introduces semi-supervised learning into VSG recently, we compare their informal work on Charades under the same settings. Note that the experiments on **DepNet* are implemented by us based on the open source project [2]. The results of compared VSG methods on Charades refer to [44].

Table 1. Comparisons with state-of-the-arts on the Activity.

Method	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
LGI [20] (CVPR'20)	58.53	41.51	23.07	41.13
DRN [47] (CVPR'20)	-	45.45	24.36	-
CPNet [15] (AAAI'21)	-	40.56	21.63	40.65
BPNet [39] (AAAI'21)	58.98	42.07	24.69	42.11
CBLN [16] (CVPR'21)	66.34	48.12	27.6	-
DeNet [51] (CVPR'21)	61.93	43.79	-	-
MATN [49] (CVPR'21)	-	48.02	31.78	-
Beam Search [6]	62.53	46.43	27.12	-
3D-TPN [50] (AAAI'20)	67.56	51.49	30.92	-
DepNet [2] (AAAI'21)	72.81	55.91	33.46	-
SVPTR (Ours)	**78.07**	**61.70**	**38.36**	**55.91**

Comparison with Fully-supervised Learning. For fair comparisons, we first train our method with 100% labeled data and compare the test results with recent SOTA fully-supervised methods. The experimental results on Activity, Charades, and TACoS are reported in Table 1, Table 2, and Table 3 respectively. Based on these results, we list following observations: (1) On all datasets, our proposed SVPTR outperforms recent state-of-the-art VPG methods on most metrics under the same labeled proportion. Notably, shown in Table 1, with 100% labeled data, our method brings at least 4.9% improvement on all the metrics compared with DepNet (2) Compared with VSG methods, our SVPTR method also shows obvious advantages. The reason is that our SVPTR method receives a paragraph as input rather than an individual sentence. By the well-designed sentence-based decoder, our SVPTR method effectively mines the contextual information among the sentences and learns more temporal features from the two modalities jointly. (3) Compared with the previous works, our SVPTR method reveals superiority on *mIoU* metric. It demonstrates our method also has more stable performance and localizes events more precisely.

Table 2. Comparisons with state-of-the-arts on the Charades.

Method	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
CTRL [7] (ICCV'17)	44.97	30.73	11.97	-
ACRN [17] (SIGIR'18)	44.69	30.03	11.89	-
ABLR [46] (AAAI'19)	44.62	31.57	11.38	-
TSP-PRL [38] (AAAI'20)	31.93	19.37	6.20	-
2D-TAN [50] (AAAI'20)	43.45	30.77	11.75	-
DRN [47] (CVPR'20)	40.45	30.43	**15.91**	-
STLG [18] (arXiv'21)	48.30	30.39	9.79	-
*DepNet [2] (AAAI'21)	45.61	27.59	10.69	29.30
SVPTR (Ours)	**55.14**	**32.44**	15.53	**36.01**

Table 3. Comparisons with state-of-the-arts on the TACoS.

Method	IoU=0.1	IoU=0.3	IoU=0.5	mIoU
DRN [47] (CVPR'20)	-	-	23.17	-
FIAN [26] (MM'20)	39.55	33.87	-	-
2D-TAN [50] (AAAI'20)	47.59	37.29	-	-
BPNet [39] (AAAI'21)	-	25.96	20.96	19.53
I^{2}N [21] (TIP'21)	-	31.47	29.25	-
CBLN [16] (CVPR'21)	49.16	38.98	27.65	-
CPNet [15] (AAAI'21)	-	42.61	**28.29**	28.69
Beam Search [6]	48.46	38.14	25.72	-
3D-TPN [50] (AAAI'20)	55.05	40.31	26.54	-
DepNet [2] (AAAI'21)	56.10	41.34	27.16	-
SVPTR (Ours)	**67.91**	**47.89**	28.22	**31.42**

Comparison with Semi-supervised Learning. We list the semi-supervised training results of our SVPTR method and compare them with several state-of-the-art methods in Table 4, where the $\{\rho_1, \rho_2, \rho_3\}$ follows the same settings above for the three datasets. "*FS*", "*SS*", "*WS*" represent fully-supervised learning, semi-supervised learning, and weakly-supervised learning methods, respectively. From the results, we can observe that: (1) Our SVPTR method successfully utilizes the unlabeled data and significantly improves the grounding performance. Specifically, under the same situation, our SVPTR method clearly outperforms DepNet on all the datasets. (2) Using much less labeled data only, our method achieves similar or higher even performance compared with fully-supervised methods. It proves the superiority of our SVPTR method that reduces the dependency on expensive video-paragraph annotations and utilizes the labeled data more effectively. (3) The proposed method

SVPTR finds a fair trade-off between temporal annotation costs and performance. Compared with the weakly-supervised methods, our method brings more than 20% improvement on Activity and 5% approximately on Charades under the metric of R1@IoU=0.5. (4) It reveals the superiority of semi-supervised learning strategy that our SVPTR method achieves better performance than the base model VPTR does. However, we also note that on the TACoS dataset, the advantages of SVPTR are weakened. One probable reason is the diversity of videos is poor in this dataset, which has a negative influence on training robust models.

Table 4. Comparisons with state-of-the-arts methods using fewer temporal annotations.

Datasets	Types	Methods	IoU=ρ_1	IoU=ρ_2	IoU=ρ_3	mIoU
Activity	FS	3D-TPN	67.56	51.49	30.92	-
		DepNet@100%	72.81	55.91	**33.46**	-
	WS	WSSL	41.98	23.34	-	28.23
	SS	*DepNet@10%	61.46	45.14	26.78	44.11
		VPTR@10%	72.80	53.14	29.07	50.08
		SVPTR@10%	**73.39**	**56.72**	32.78	**51.98**
Charades	FS	STLG@100%	48.30	**30.39**	9.79	-
		*DepNet@100%	45.61	27.59	10.69	29.30
	WS	WSSL	35.86	23.67	8.27	-
	SS	STLG@30%	46.15	29.43	9.38	-
		*DepNet@30%	43.03	25.07	10.14	28.09
		VPTR@30%	45.13	24.98	10.22	28.92
		SVPTR@30%	**50.31**	28.50	**12.27**	**32.13**
TACoS	FS	3D-TPN	55.05	40.31	26.54	-
		DepNet@100%	56.1	**41.34**	**27.16**	-
	SS	*DepNet@50%	40.27	26.95	16.54	18.68
		VPTR@50%	61.31	40.59	21.39	**26.59**
		SVPTR@50%	**63.06**	40.19	20.05	26.10

4.3. Further Analysis

Effectiveness of Structures. To learn how each module of SVPTR performs in grounding, we conduct the structure ablation studies on Activity and Charades datasets with 10% and 30% labeled data respectively. As depicted in Table 5, we study the effectiveness of following components: sentence-level query (S.Q.), multimodal encoding (M.E.), contrastive loss ($\mathcal{L}_{\text{CT}}$), consistency loss ($\mathcal{L}_{\text{CS}}$), and feature perturbations (F.P.). According to the results, we can observe that: (1) The multimodal encoding performs a significant role in VPG. Specifically, with multimodal encoding, the precision on most metrics gets improved on the two datasets. The reason is that it allows the fine-grained cross-modal interactions between word-level text features and clip-level video features. (2) The consistency loss $\mathcal{L}_{\text{CS}}$ is essential for the whole semi-supervised learning framework. Without $\mathcal{L}_{\text{CS}}$, the generalization of models on unlabeled data will be degraded greatly thus leading to worse performance under semi-supervised conditions. (3) Using contrastive loss with the feature perturbations boosts the final results. A probable reason is that video perturbations bring more diversity into multimodal features, which has been proved [9, 23] to be crucial for contrastive learning.

Effectiveness of Annotation Proportions. To study the quality of the semi-supervised learning framework, we evaluate our VPTR and SVPTR on different proportions of labeled data. For fair comparisons, we also train the recent method DepNet under the same conditions. Limited by the space, we report the experimental results on Activity. For more results on Charades, please refer to the supplementary materials. As illustrated in Fig. 4, we list following observations: (1) Our two models, VPTR and the complete model SVPTR, both outperform the previous method DepNet significantly, which reveals the robustness and effectiveness of our method. (2) The SVPTR achieves promising improvement compared with the base model VPTR, especially under the situation of less annotated data. It demonstrates the superiority of our SVPTR method that reduces the requirements of annotated data again. (3) The performance of our method can be improved with more labeled data. Specifically, with 100% labeled data, our SVPTR method achieves about 5% higher both on the two metrics compared with the results of 10% labeled data.

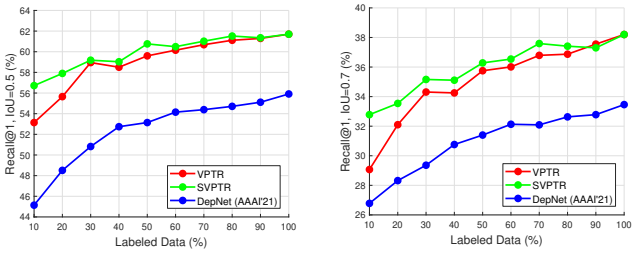

Figure 4. Analysis with respect to the proportion of annotations on Activity dataset.

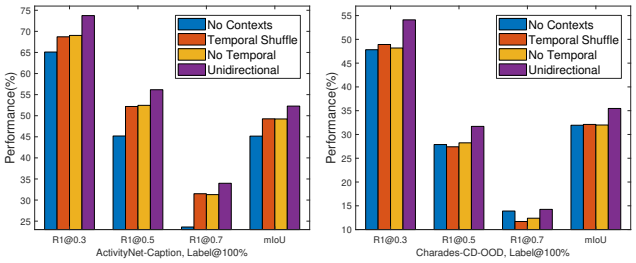

Figure 5. Analysis with respect to the sentence-level contexts.

Effectiveness of Contexts. We design four ablated models to study how the contextual information of paragraph influence the grounding quality, which is shown in Fig. 5: (1) "*No Contexts*": no sentence-level contexts are utilized and the model degenerates to a conventional VSG model. (2) "*Temporal Shuffle*": the order of sentences in a paragraph is shuffled randomly. (3) "*No Temporal*": we disable the sentence-level position embedding in a decoder. (4) "*Unidirectional*": the contexts are limited in the single direction.

Comparing these results, we can observe that our complete model achieves significant improvement on grounding precision, which proves again that contextual information is

Table 5. Ablation study on Activity with 10% labeled data and Charades with 30% labeled data.

S.Q.	M.E.	$\mathcal{L}_{CT}$	$\mathcal{L}_{CS}$	F.P.	Activity				Charades			
					IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
√					73.24	52.98	28.45	50.02	46.01	26.16	11.67	30.25
√	√				73.67	53.41	28.56	50.34	47.38	28.36	10.67	30.81
√	√	√			72.80	53.14	29.07	50.14	45.13	24.98	10.22	28.92
√	√		√		74.36	54.50	30.00	51.04	49.01	29.54	11.44	31.71
√	√		√	√	73.46	53.78	29.57	50.42	49.44	**29.87**	10.94	31.51
√	√	√	√		**74.69**	55.98	31.63	51.70	46.87	27.02	10.81	30.25
√	√	√	√	√	73.39	**56.72**	**32.78**	**51.98**	**50.31**	28.50	**12.27**	**32.14**

crucial for localizing events in untrimmed videos. However, we also note that "*No Contexts*" version performs better on Charades than it does on Activity. One essential reason is, compared with the latter, Charades contains a large number of videos consisting of sparse events. It thus leads to a deficiency of sentence-level contextual information, weakening the advantages of our SVPTR method.

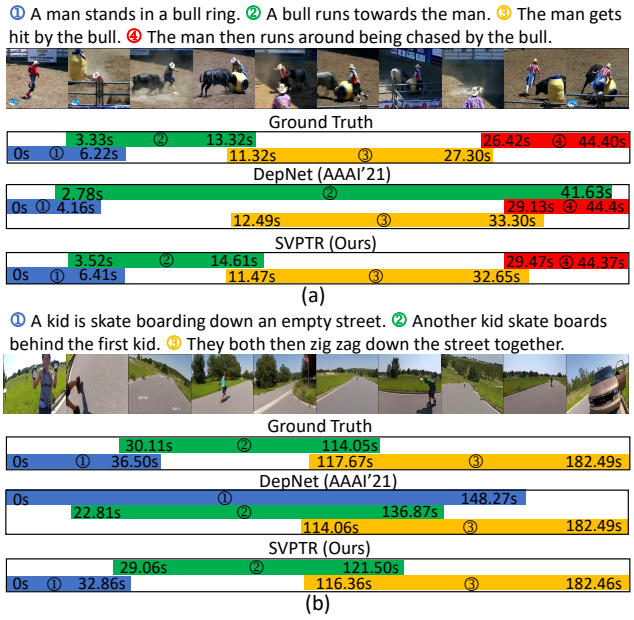

Figure 6. Visualization of two examples on Activity dataset.

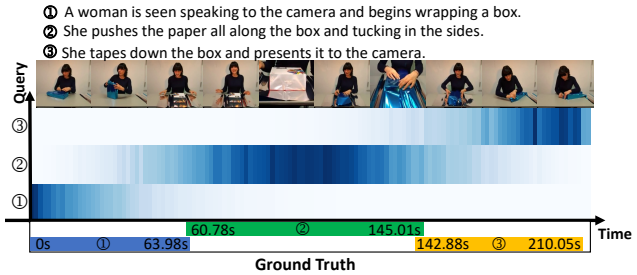

Figure 7. Visualization of cross-modal attention weights.

Qualitative Analysis. To illustrate the localizing quality of the proposed SVPTR method, we visualize two localizing examples with different length of videos and compare them with the results of DepNet [2]. In Fig. 6(a), our proposed SVPTR method precisely localizes all the target moments, while the previous method DepNet fails to retrieve the second event. Moreover, as described in Fig. 6(b), the long video example is more challenging since the high demand on understanding the contexts. Focusing the proposal-sentence matching and ignoring crucial contexts in paragraph, the DepNet gives an absolutely incorrect result on the first event, which contains the whole video almost. Contrastively, our SVPTR method avoids this fatal defect effectively with the video-paragraph alignment and sentence-level context modeling. Additionally, to show how the proposed method SVPTR works, we visualize the cross-attention weights between sentence and video features from sentence-based decoders. As illustrated in Fig. 7, we can observe that each sentence gains more attention on the related moments and maintains the temporal order correctly.

5. Conclusion

In this work, we have introduced a novel Video Paragraph Grounding (VPG) framework dubbed *Semi-supervised Video-Paragraph TRansformer* (SVPTR) that learns contextual information from paragraphs and significantly reduces the dependency on annotated data. We evaluated our SVPTR method on three public datasets and conduct extensive experiments to prove its effectiveness and robustness. The results show that our proposed SVPTR model achieves competitive results with less annotated data. Moreover, with fully-supervised training, SVPTR outperforms the latest VPG methods. For future work, we will study furtherly the trade-off between performance and costs based on this work, and we believe it will inspire more research on video events grounding.

6. Acknowledgement

This work was supported in part by National Natural Science Foundation of China under Grants (No. 61976049 and 62072080); Sichuan Science and Technology Program, China (No. 2019ZDZX0008, 2020YFS0057) and Meituan.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *IEEE International Conference on Computer Vision*, pages 5803–5812, 2017. 1, 2

[2] Peijun Bao, Qian Zheng, and Yadong Mu. Dense events grounding in video. In *AAAI Conference on Artificial Intelligence*, pages 920–928, 2021. 1, 2, 4, 5, 6, 8

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, volume 12346, pages 213–229, 2020. 4

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019. 5

[5] Xuguang Duan, Wen-bing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In *Annual Conference on Neural Information Processing Systems*, pages 3063–3073, 2018. 2, 6

[6] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, 2017. 1, 2, 6

[7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *IEEE International Conference on Computer Vision*, pages 5267–5275, 2017. 1, 2, 5, 6

[8] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander G. Hauptmann. Excl: Extractive clip localization using natural language descriptions. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1984–1990, 2019. 2

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020. 7

[10] Jingwei Ji, Kaidi Cao, and Juan Carlos Niebles. Learning temporal action proposals with fewer labels. In *IEEE International Conference on Computer Vision*, pages 7073–7082, 2019. 3

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[12] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision*, pages 706–715, 2017. 2, 5

[13] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017. 3

[14] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013. 3

[15] Kun Li, Dan Guo, and Meng Wang. Proposal-free video grounding with contextual pyramid network. In *AAAI Conference on Artificial Intelligence*, pages 1902–1910, 2021. 2, 3, 6

[16] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11235–11244, June 2021. 1, 2, 3, 6

[17] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 15–24, 2018. 6

[18] Fan Luo, Shaoxiang Chen, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Self-supervised learning for semi-supervised temporal language grounding. *arXiv preprint arXiv:2109.11475*, 2021. 2, 3, 6

[19] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K. Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Computer Vision and Pattern Recognition*, pages 11592–11601, 2019. 2, 4

[20] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10807–10816, 2020. 1, 2, 5, 6

[21] Ke Ning, Lingxi Xie, Jianzhuang Liu, Fei Wu, and Qi Tian. Interaction-integrated network for natural language moment localization. *IEEE Transactions Image Process.*, 30:2538–2548, 2021. 6

[22] Cristian Rodriguez Opazo, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *IEEE Winter Conference on Applications of Computer Vision*, pages 2453–2462, 2020. 2

[23] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 7

[24] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. 5

[25] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V. Le. Meta pseudo labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021. 3

[26] Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Fine-grained iterative attention network for temporal language localization in

videos. In *ACM International Conference on Multimedia*, pages 4280–4288, 2020. 6

[27] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 2, 5

[28] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019. 5

[29] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119(3):346–373, 2016. 5

[30] Fengyuan Shi, Limin Wang, and Weilin Huang. End-to-end dense video grounding via parallel regression. *arXiv preprint arXiv:2109.11265*, 2021. 2

[31] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, pages 1195–1204, 2017. 3, 5

[32] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE international conference on computer vision*, pages 4489–4497, 2015. 3, 5

[33] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, 2019. 4

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Annual Conference on Neural Information Processing Systems*, pages 5998–6008, 2017. 2

[35] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7190–7198, 2018. 2

[36] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 12168–12175, 2020. 6

[37] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Changxin Gao, and Nong Sang. Self-supervised learning for semi-supervised temporal action proposal. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1905–1914, 2021. 3, 5

[38] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *AAAI Conference on Artificial Intelligence*, pages 12386–12393, 2020. 6

[39] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 2986–2994, 2021. 2, 6

[40] Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2020. 3

[41] Xing Xu, Kaiyi Lin, Lianli Gao, Huimin Lu, Heng Tao Shen, and Xuelong Li. Learning cross-modal common representations by private-shared subspaces separation. *IEEE Transactions on Cybernetics*, 2020. 4

[42] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. Cross-modal attention with semantic consistence for image–text matching. *IEEE transactions on neural networks and learning systems*, 31(12):5412–5425, 2020. 4

[43] 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. 2

[44] Yitian Yuan, Xiaohan Lan, Long Chen, Wei Liu, Xin Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Datasets and metrics. *arXiv preprint arXiv:2101.09028*, 2021. 2, 5, 6

[45] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 1, 2

[46] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166, 2019. 1, 6

[47] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10284–10293, 2020. 1, 2, 6

[48] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S. Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1247–1257, 2019. 2

[49] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. Multi-stage aggregated transformer network for temporal language localization in videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12669–12678, 2021. 2, 4, 6

[50] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI Conference on Artificial Intelligence*, pages 12870–12877, 2020. 1, 2, 5, 6

[51] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. Embracing uncertainty: Decoupling and debias for robust temporal grounding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8445–8454, 2021. 2, 3, 6