Embracing Uncertainty: Decoupling and De-bias for Robust Temporal Grounding

Hao Zhou[1], Chongyang Zhang[1,2] *, Yan Luo[1], Yanjun Chen[1], Chuanping Hu[1,3]
[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China
[3]Zhengzhou University, Zhengzhou, China
{zhouhao_0039,sunny_zhang,luoyan_bb,erinchen}@sjtu.edu.cn, cphu@vip.sina.com

Abstract

*Temporal grounding aims to localize temporal boundaries within untrimmed videos by language queries, but it faces the challenge of two types of inevitable human uncertainties: query uncertainty and label uncertainty. The two uncertainties stem from human subjectivity, leading to limited generalization ability of temporal grounding. In this work, we propose a novel DeNet (**De**coupling and **De**-bias) to embrace human uncertainty: Decoupling — We explicitly disentangle each query into a relation feature and a modified feature. The relation feature, which is mainly based on skeleton-like words (including nouns and verbs), aims to extract basic and consistent information in the presence of query uncertainty. Meanwhile, modified feature assigned with style-like words (including adjectives, adverbs, etc) represents the subjective information, and thus brings personalized predictions; De-bias — We propose a de-bias mechanism to generate diverse predictions, aim to alleviate the bias caused by single-style annotations in the presence of label uncertainty. Moreover, we put forward new multi-label metrics to diversify the performance evaluation. Extensive experiments show that our approach is more effective and robust than state-of-the-arts on Charades-STA and ActivityNet Captions datasets.*

1. Introduction

As the increasing demand for video understanding, many related works have drawn increasing attention, *e.g*. action recognition [31, 37, 23] and temporal action detection [50, 20]. These tasks rely on trimmed videos or predefined action categories, yet most videos are untrimmed and associated with open-world language descriptions in real scenarios. Temporal grounding task aims to localize corresponding temporal boundaries in an untrimmed video by a language query. Thus, models need to understand both fine-grained video content and complex language queries. Recently, this task has also shown its potential in a wide range of applications, *e.g*. video captioning [26, 41, 5], video object segmentation [7, 13] and video question answering [19, 14, 35].

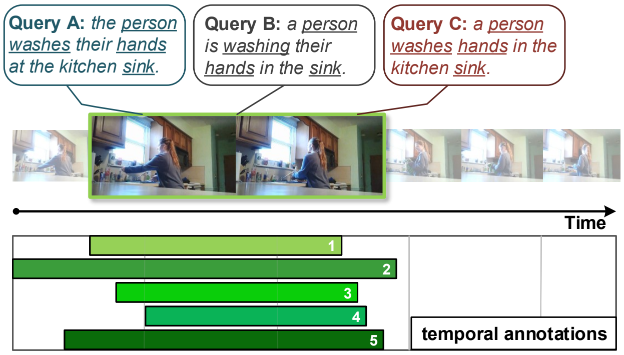

Query A: the person washes their hands at the kitchen sink.
Query B: a person is washing their hands in the sink.
Query C: a person washes hands in the kitchen sink.
Time
1
2
3
4
5
temporal annotations

Figure 1. Example of temporal grounding task with two types of uncertainties. Query uncertainty: For one same event, there are different language expressions. Label uncertainty: Given one same query and video, different annotators may provide a variety of temporal boundaries.

We observe there lies inherent uncertainty in temporal grounding task and classify it into two types: 1) One is query uncertainty stemming from different expressions for one same event. As shown in Figure 1, three queries are attached to the same moment. Previous approaches usually leverage LSTM-based [45, 47] networks to encode entire language as a deterministic vector. However, the variety of expressions makes it challenging to extract discriminative semantic features, sometimes leading to quite different predictions for the same event. 2) The other is label uncertainty representing subjective boundaries for one same event. As shown in Figure 1, for the same query A and video, temporal boundaries annotated by different people exist disagree-

*This is the corresponding author.

ment. Due to the expensive cost of multiple-labeling, most of previous models [24, 28] are optimized using single-style annotations (which means each sample is labeled by one annotator), whereas the inherent uncertainty of event localization [29] is ignored. As a result, models may learn single-style prediction bias from training datasets, leading to limited generalization performances.

Considering the fact that uncertainty can cover a broad range of human perspectives, it should be embraced to promote robust temporal grounding. Furthermore, we argue single-annotation, single prediction is not reasonable in the presence of uncertainty, and diversity of predictions is an effective way to alleviate the bias caused by single-style annotations. Therefore, the key challenge is how to obtain diverse predictions. Inspired by linguistic knowledge, we find consistent discriminative information lies in a skeleton-like relation phrase (including *nouns* and *verbs*), and query uncertainty mainly exists in a style-like modified phrase (including *adjectives*, *adverbs*, etc). On one hand, the relation phrase is beneficial to robust temporal grounding. On the other hand, the modified phrase may be largely associated with human preferences and brings personalized differences. Based on this intuition, our main idea is to leverage various expressions stemming from query uncertainty to obtain a diverse yet plausible prediction set that fits label uncertainty.

In this paper, we propose one novel DeNet (**De**coupling and **De**-bias) to embrace the two types of uncertainties in the temporal grounding task. First of all, a decoupling method is introduced to disentangle each query into a relation feature and a modified feature using Parts-of-Speech (PoS). While discriminative and consistent information is obtained from the relation feature, personalized information can be also reserved in the modified feature. Then, a de-bias mechanism is proposed to generate diverse predictions, which includes sampling operation, multiple choice learning (MCL) [10], clustering, etc. Specifically, we encode the modified feature as a Gaussian distribution and adopt a sampling operation in the latent space to obtain multiple query representations. To tackle the dilemma between multiple predictions and single-style annotations, we introduce a min-loss from MCL to optimize DeNet to generate diverse predictions. In the inference stage, multiple predictions are clustered into one diverse yet plausible prediction set. Moreover, we devise multi-label metrics to meet for multiple testing annotations situations. Finally, DeNet is evaluated on two popular datasets Charades-STA [6] and ActivityNet Captions [2, 16] in terms of standard metrics and new multi-label metrics. To sum up, the main contributions of our work are as follows:

(1) We first attempt to embrace two types of human uncertainties: query uncertainty and label uncertainty, in one unified network DeNet to model robust temporal grounding.

(2) We develop a decoupling module in the language encoding, and one de-bias mechanism in the temporal regression. With the two designs, diverse yet plausible predictions can be obtained to fit human diversity in real scenarios.

(3) We devise new multi-label metrics to meet multiple annotations and verify the effectiveness and robustness of DeNet on both Charades-STA and ActivityNet Captions.

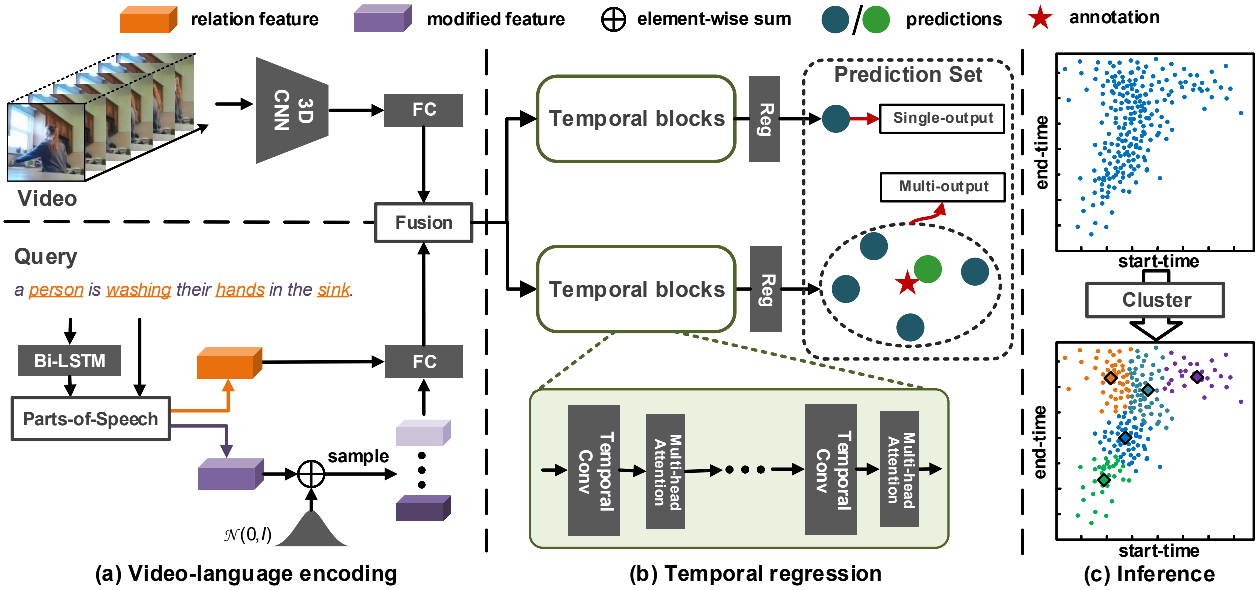
2. Related Work

Temporal grounding. As a challenging task in video understanding, temporal grounding needs to capture semantic information in both videos and language queries.

In the video encoding component, most previous approaches [6, 22, 1, 38, 47] follow a proposal-based framework, where untrimmed videos are clipped into multi-scale segments as proposal candidates. Gao *et al*. [6] and Liu *et al*. [22] adopt a sliding window to combine each central-clip feature and its context-clip features as one proposal candidate. Hendricks *et al*. [1] and Wang *et al*. [38] concatenate local features and global feature to better cover contexts. To further explore dependencies across multiple candidates, Zhang *et al*. [47] generate multi-scale segments and construct a 2D temporal adjacent map. However, too many proposals will burden models during the training process. Recently, some approaches [24, 45, 28, 11] adopt a proposal-free framework. For example, Zeng *et al*. [45] extract sequential clip-level features, then directly predict temporal boundaries in a subsequent network. In this paper, the proposed DeNet follows the proposal-free framework to reduce the training computation cost.

Language encoding also plays an important role in the temporal grounding task. Most approaches employ LSTM-based layers [45, 47, 24, 39] or GRU-based layers [28, 43] to encode entire language queries. Recently, some approaches [49, 48, 21] leverage syntactic dependency parser to capture underlying semantic structures. Besides, Mun *et al*. [24] and Yuan *et al*. [44] attempt to capture discriminative features from queries using an attention mechanism. These methods aim to obtain more subtle query representations, yet we follow a different motivation. On the one hand, we hope to obtain discriminative information from various expressions to achieve robust predictions. On the other hand, we attempt to reserve personalized differences to achieve diversified predictions. Thus, we adopt an explicit decoupling method to disentangle each query into the relation feature and the modified feature.

Multiple choice learning. In contrast to single-output learning, multiple choice learning (MCL) [10] is proposed to produce multiple outputs based on one min-loss. Given a training sample, MCL takes account of all hypotheses and only updates networks according to the best hypothesis. One accurate and diverse prediction set can be obtained in this way. Inspired by MCL, we consider diversity is an



(a) Video-language encoding (b) Temporal regression (c) Inference

Figure 2. An overview of our proposed model for the temporal grounding task. (a) In the video-language encoding component, we use a pretrained 3D CNN to extract the sequential video feature and disentangle the query into relation feature and modified feature by Parts-of-Speech. Then, a sampling operation is applied in the latent space to generate multiple query representations. (b) In the temporal regression component, two independent branches are set to generate multiple predictions. (c) In the inference stage, we adopt a clustering method to obtain a fixed-size prediction set.

effective way to model human uncertainty, and introduce the min-loss into temporal grounding to predict all possible temporal boundaries in the absence of multiple annotations. However, note that our proposed method is significantly different from traditional MCLs. Firstly, MCL focuses on ensemble learning, whereas we focus on temporal grounding. Then, most MCL approaches [18, 17, 32] produce the multi-output $\{f_i(x)\}_{i=1}^N$ based on multiple ”base classifiers”, whereas our method generates the multi-output $\{f(\hat{x}_i)\}_{i=1}^N$ via multiple features.

3. Proposed Method

3.1. Method overview

Given an untrimmed video $\mathbf{V}$ and an open-world language description $\mathbf{Q}$ as a query, temporal grounding aims to localize the start-end boundary $\mathbf{b}_{se}$ within $\mathbf{V}$. Specifically, the untrimmed video is represented as $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^T$, where $\mathbf{v}_i$ denotes the i-th video clip and T is the total number of video clips. The query is represented as $\mathbf{Q} = \{\mathbf{w}_i\}_{i=1}^S$, where $\mathbf{w}_i$ denotes the i-th word and S is the total number of words. In this work, models should output matched temporal times $\{\mathbf{b}_{se}\}^N = \{(t_s, t_e)\}^N$ corresponding to the query $\mathbf{Q}$, where N is the number of predictions.

As illustrated in Figure 2, DeNet contains two main components: video-language encoding and temporal regression. In the video-language encoding component, we adopt a decoupling method to disentangle each query into a relation feature and a modified feature using PoS, where the modified feature is encoded as a distribution. Then, the video-language feature is fed into the temporal regression component to predict multiple temporal boundaries. In the training stage, two independent branches are optimized by single-output loss and multi-output loss, respectively. In the inference stage, we cluster the collection of predictions into a fixed-size prediction set and evaluate them in both standard metrics and new multi-label metrics.

3.2. Video-language encoding

Video encoding. Firstly, an untrimmed video is represented as a collection of clips $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^T$, where each clip covers C frames ($C = 16$ in this work). Analogous to [47], we use a pretrained 3D CNN model to extract clip-level features, then sample fixed T_m clips from T clips so as to obtain a fixed-length video feature $\widetilde{\mathbf{V}} \in \mathbb{R}^{d_v \times T_m}$, where d_v is the dimension of the video feature. Furthermore, a zero-padding operation is applied if there are less than T_m clips in an untrimmed video. Finally, two extra Fully Connected layers are implemented to obtain a final video embedding $\mathbf{F}^V \in \mathbb{R}^{d_v \times T_m}$ as:

$$\mathbf{F}^V = \mathbf{W}_2 \mathrm{ReLU}(\mathbf{W}_1 \widetilde{\mathbf{V}}), \tag{1}$$

where $\mathbf{W}_1$, $\mathbf{W}_2 \in \mathbb{R}^{d_v \times d_v}$ are learnable parameters, the superscript V indicates the video modality.

Language encoding. For a language query $\mathbf{Q} = \{\mathbf{w}_i\}_{i=1}^{S}$ with S words, we take advantage of Glove [27] to map each word to a 300-dimensional vector, then set two Bi-LSTM layers to get word-level features $\{\mathbf{h}_i\}_{i=1}^{S} \in \mathbb{R}^{d_l \times S}$, where d_l is the feature dimension of each word. In our observation, query uncertainty mainly lies in the modified phrase and discriminative information are in the relation phrase. For example, "*a person is washing their hands in the sink*" can be broken down into relation phrase [*person, washing, hands, sink*] and modified phrase [*a, is, their, in, the*]. Here, the spaCy toolbox[1] is used to generate PoS tags that denote word types, like *verbs*, *adjectives*. Then, we average word-level features associated with the relation phrase to get a relation feature $\mathbf{f}_r^L$. Similarly, the remaining word-level features are selected and averaged as a modified feature $\mathbf{f}_m^L$.

Then, we concatenate the two types of features and set a Fully Connected layer to obtain a final query embedding as:

$$\mathbf{f}^L = \mathbf{W}_3[\mathbf{f}_r^L, \mathbf{f}_m^L] + \mathbf{b}_3, \tag{2}$$

where $\mathbf{W}_3 \in \mathbb{R}^{d_l \times 2d_l}$, $\mathbf{b}_3 \in \mathbb{R}^{d_l}$ are the learnable parameters, $[\cdot, \cdot]$ denotes concatenation and the superscript L indicates the language modality. Considering the fact that most variances stem from the modified phrase, we encode corresponding modified feature as a distribution instead of a deterministic vector. Here, we adopt the Gaussian distribution $\mathcal{N}(\mathbf{u}, \sigma^2)$ as in many existing works [40]. From a probabilistic perspective, it means that the feature is regarded as a random variable to model uncertainty [42]. $\mathbf{f}_m^L$ is set as the distribution center $\mathbf{u}$ and a collection of modified features are sampled from the Gaussian distribution $\mathcal{N}(\mathbf{f}_m^L, \sigma^2)$. A reparameterisation trick is used to obtain the modified feature $\hat{\mathbf{f}}_m^L = \mathbf{f}_m^L + \epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^\mathbf{2})$. Finally, a variant query embedding is formulated as:

$$\hat{\mathbf{f}}^L = \mathbf{W}_4[\mathbf{f}_r^L, \hat{\mathbf{f}}_m^L] + \mathbf{b}_4. \tag{3}$$

From another perspective, the distribution representation is equivalent to adding small perturbations in the modified feature. We provide two rationales illustrating its advantages. On the one hand, models will further focus on the relation feature and pay less attention to the modified feature. Thereby, the model is more robust. On the other hand, the sampling process can be viewed as query augmentation. Based on multiple query features, models can generate multiple personalized predictions.

Multimodal fusion. When both videos and language embeddings are obtained, we need to model the interaction of them. First of all, $\mathbf{f}^L$ and $\hat{\mathbf{f}}^L$ are replicated for T_m times to get sequential embeddings $\mathbf{F}^L$, $\hat{\mathbf{F}}^L \in \mathbb{R}^{d_l \times T_m}$, respectively. Then, multimodal features $\mathbf{F}^M, \hat{\mathbf{F}}^M \in \mathbb{R}^{d_m \times T_m}$ are produced by fusing video embedding and query embedding:

$$\mathbf{F}^M = ||\mathbf{F}^V \circ \mathbf{F}^L||_F, \tag{4}$$

$$\hat{\mathbf{F}}^M = ||\mathbf{F}^V \circ \hat{\mathbf{F}}^L||_F, \tag{5}$$

where $\circ$ denotes the Hadamard product and $||\cdot||_F$ is the Frobenius normalization (ℓ_2-norm). Note that d_m, d_v and d_l are consistent for dimension matching.

3.3. Temporal regression

When we obtain a collection of multimodal features, a temporal regression network is constructed to predict matched temporal boundaries. It is composed of two independent branches, where each branch contains a stack of temporal blocks and a regression layer. The single-output branch associated with $\mathbf{F}^M$ produces a top-1 prediction. The multi-output branch associated multiple $\hat{\mathbf{F}}^M$ produces multiple predictions covering possible annotations.

Each temporal block contains a Temporal Convolutional layer and a Multi-head Attention layer [34]. The Temporal Convolutional layer aims to capture temporal dependencies in the neighbor clips and the Multi-head Attention layer is to capture long-range temporal dependencies. For the n-th temporal block, its output $\mathbf{F}^{(n)} \in \mathbb{R}^{d_m \times T_m}$ can be formulated as:

$$\widetilde{\mathbf{F}}^{(n)} = \mathbf{F}^{(n-1)} + \mathrm{Conv}(\mathbf{F}^{(n-1)}), \tag{6}$$

$$\mathbf{F}^{(n)} = \widetilde{\mathbf{F}}^{(n)} + \mathrm{MultiheadAttention}(\widetilde{\mathbf{F}}^{(n)}), \tag{7}$$

where $\mathbf{F}^{(n-1)}$ is the output of previous temporal block. $\mathrm{Conv}(\cdot)$ represents a mapping function in the Temporal Convolutional layer that contains two 1D convolutional layers with batch normalization.

Following a stack of temporal blocks, an attention-guided regression layer is employed to output the start-end prediction $\mathbf{b}_{se}$. An auxiliary head is implemented here to predict the center-width $\mathbf{b}_{cw}$ to assist temporal grounding. Thus, the regression layer is formulated as:

$$\mathbf{a} = \mathrm{softmax}(\mathbf{W}_6 \mathrm{Tanh}(\mathbf{W}_5 \mathbf{F})), \tag{8}$$

$$\mathbf{b}_{cw} = (t_c, t_w) = \mathbf{Reg}_{cw}(\sum_{i=1}^{T_m} a_i \mathbf{F}_i), \tag{9}$$

$$\mathbf{b}_{se} = (t_s, t_e) = \mathbf{Reg}_{se}(\sum_{i=1}^{T_m} a_i \mathbf{F}_i), \tag{10}$$

where $\mathbf{a} \in \mathbb{R}^{T_m}$ is an attention coefficient and $\mathbf{Reg}_{cw}$, $\mathbf{Reg}_{se}$ are two independent Fully Connected layers. Note that all of predictions are normalized to [0,1].

3.4. Optimization and inference

Optimization. According to the definition of equation 6-10, we feed $\mathbf{F}^M$ and the collection of $\hat{\mathbf{F}}^M$ into the two branches of temporal regression network and obtain a single prediction $(\mathbf{b}_{se}, \mathbf{b}_{cw}, \mathbf{a})$ and multiple predictions $\{(\hat{\mathbf{b}}_{se}, \hat{\mathbf{b}}_{cw}, \hat{\mathbf{a}})\}^K$, respectively.

[1] https://spacy.io/

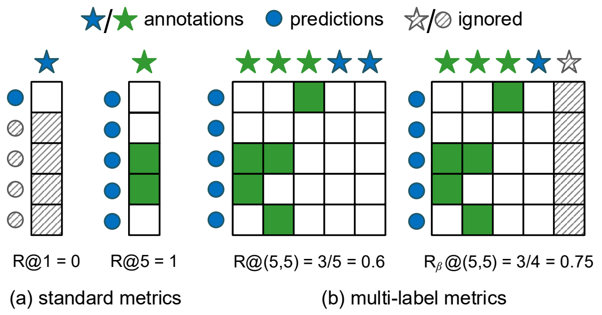


Figure 3. A example to illustrate differences between proposed multi-label metrics and standard metrics. The green star denotes the corresponding annotation that is matched with at least one prediction (with IoU larger than α) and otherwise the corresponding star is blue. The grey star is the low-quality annotation (with average IoU smaller than β).

The single-output branch is optimized with two kinds of loss functions, and one is a regression loss as follows:

$$\mathcal{L}_{reg}(\mathbf{b}_{se}, \mathbf{b}_{cw}) = L_1(\mathbf{b}_{se} - \mathbf{y}_{se}) + L_1(\mathbf{b}_{cw} - \mathbf{y}_{cw}), \quad (11)$$

where L_1 denotes L1 distances, and $\mathbf{y}_{se}$, $\mathbf{y}_{cw} \in [0, 1]$ denote the start-end and center-width groundtruth, respectively. The other one is an attention loss [44] that forces the model to focus on clips within groundtruth interval:

$$\mathcal{L}_{att}(\mathbf{a}) = -\frac{\sum_{i=1}^{T_m} m_i \text{log} a_i}{\sum_{i=1}^{T_m} m_i}, \quad (12)$$

where $m_i = 1$ if the i-th clip is within the groundtruth interval and otherwise $m_i = 0$.

For the multi-output branch, it's not reasonable to regress all of $\{(\hat{\mathbf{b}}_{se}, \hat{\mathbf{b}}_{cw}, \hat{\mathbf{a}})\}^K$ with one single annotation. To tackle the dilemma between multiple predictions and single-style annotations, we introduce a min-loss from MCL [10] to learn diverse predictions without extra annotations. It only computes a loss between the closest prediction to the existing annotations. Finally, all of the loss functions are jointly considered as follows:

$$\begin{aligned} \mathcal{L}_{all} &= \mathcal{L}_{single} + \lambda \mathcal{L}_{multi} \\ &= \mathcal{L}_{reg}(\mathbf{b}_{se}, \mathbf{b}_{cw}) + \mathcal{L}_{att}(\mathbf{a}) \\ &\quad + \lambda \min_{i \in [K]} [\mathcal{L}_{reg}(\hat{\mathbf{b}}_{se,i}, \hat{\mathbf{b}}_{cw,i}) + \mathcal{L}_{att}(\hat{\mathbf{a}}_i)], \end{aligned} \quad (13)$$

where λ is a trade-off parameter between two regression branches, and $[K]$ denotes the set $\{1, ..., K\}$.

Inference. We only focus on the single prediction $\mathbf{b}_{se}$ and the collection of predictions $\{\hat{\mathbf{b}}_{se}\}^K$ in the inference stage, where K depends on the number of query embeddings $\hat{\mathbf{F}}^M$ sampled in the latent space. Previous approaches adopt NMS to reduce predictions, yet this method faces two issues: 1) Since the collection of predictions is dense, predictions are mistakenly suppressed easily. 2) Confidence scores are necessary to rank predictions. Most approaches build up an extra branch to predict the confidence scores or IoU scores, whereas performances are limited. To address above two issues, we leverage K-Means to cluster $\{\hat{\mathbf{b}}_{se}\}^K$ into a fixed-size prediction set $\{\hat{\mathbf{b}}_{se}\}^N$ without NMS, where N is a pre-defined constant. If necessary, we can rank $\{\hat{\mathbf{b}}_{se}\}^N$ using the distance from single-style prediction $\mathbf{b}_{se}$.

3.5. New evaluation metrics

The standard evaluation metric is "R@N, IoU=α". It is defined as the percentage of at least one of the top-N predictions having IoU larger than α. This metric only focuses on whether the single groundtruth is localized successfully. Due to the label uncertainty, different people localize various moment boundaries for the same query. That is to say, there are multiple acceptable labels for each query. Thus, we consider the prediction set should be evaluated with multi-labels instead of a single label.

Recently, Otani *et al.* [25] provide 5 annotations for each testing sample on two public datasets. We propose two multi-label metrics to meet for multi-label situations. The first metric is "R@(N, G), IoU=α" that evaluates performances with N predictions and G annotations for each query. It is defined as the percentage of annotations that match at least one prediction (with IoU larger than α) in top-N predictions. This metric is equivalent to the standard metric "R@N, IoU=α" if G is set as 1. The second metric is "R$_\beta$@(N, G), IoU=α", where low-quality annotations (with average IoU among annotations smaller than β) are ignored. Intuitively, when one annotation has a small average IoU, it tends to be low-quality. Thus, "R$_\beta$@(N, G), IoU=α" is equivalent to "R@(N, G), IoU=α" if β is set as 0. When there is only one testing sample, Figure 3 illustrates the results in different metrics. The standard metrics only compute the matched percentage of single annotation (*e.g.* R@1 = 0 and R@5 = 1), our multi-label metrics considers whether multiple annotations are matched (*e.g.* R@(5,5) = 0.6 and R$_\beta$@(5,5) = 0.75). We note that some methods [1, 12] consider multiple annotations based on standard metrics that use their aggregator over three out of the four human annotators. Similar to our proposed "R$_\beta$@(N, G), IoU=α", they ignore part of multi-labels when evaluating. However, instead of discarding one of four labels that has the lowest evaluation score, we evaluate the disagreements among labels and filter out low-quality labels adaptively.

4. Experiments

4.1. Datasets

Charades-STA. This dataset contains 9,848 videos built on the Charades dataset [30]. Gao *et al.* [6] provide single temporal annotation for each language query as Charades-STA, where 12,408 samples are split into the training set

Method	Feature	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
CTRL [6]	C3D	23.63	8.89	58.92	29.52
SMRL [38]	C3D	24.36	11.17	61.25	32.08
MAC [8]	C3D	30.48	12.20	64.84	35.13
MLVI [39]	C3D	35.60	15.80	79.40	45.40
CBP [36]	C3D	36.80	18.87	70.94	50.19
SAP [4]	VGG	27.42	13.36	66.37	38.15
MAN [46]	VGG	41.24	20.54	83.21	51.85
2D-TAN [47]	VGG	42.80	23.25	80.54	54.14
EXCL [9]	I3D	44.10	22.40	-	-
TMLGA [9]	I3D	52.02	33.74	-	-
DRN [45]	I3D	53.09	31.75	89.06	60.05
SCDM [43]	I3D	54.44	33.43	74.43	58.08
LGI [24]	I3D	59.46	35.48	-	-
DeNet(ours)	I3D	**59.70**	**38.52**	**91.24**	**66.83**

Table 1. Comparison with state-of-the-art methods on Charades-STA using standard metrics; bold font indicates best results, underlined second-best.

and 3,720 samples are into the testing set. Recently, Otani *et al*. [25] extend 5 temporal annotations for each query (1,000 queries totally) in the testing set.

ActivityNet Captions. This dataset [2] contains 19,209 videos, which was originally proposed by [16] for dense video captioning task. As the largest dataset in temporal grounding task, it contains 10,024, 4,926, and 5,044 samples for the training set, val_1 set, and val_2 set. Due to the lack of of the testing set, we follow a popular split method [39] that combines the two validation sets as the testing set. Besides, Otani *et al*. [25] extend 5 temporal annotations for each query (1,288 queries totally) in the validation sets.

4.2. Implementation details

In the video encoding, we use pretrained 3D CNN networks to extract clip-level features, where each clip contains 16 consecutive frames. Following previous works, we adopt I3D features [3] for Charades-STA and C3D features [33] for ActivityNet Captions. The max video length T_m is set as 128. In the language encoding, we draw 5 samples $\hat{\mathbf{F}}^M$ from the latent space and set the standard deviation σ as an identical matrix *I* during the training procedure. For dimension matching, dimension of video embedding d_v, dimension of query embedding d_l and dimension of multimodal feature d_m are all set as 512. In the inference procedure, we set deviation σ as 2*I* to enlarge the personalized differences from modified feature, and cluster about 200 results into fixed 5 predictions using K-means. The trade-off parameter λ in Equation 13 is set as 0.02. In all experiments, we use Adam [15] and batch size of 32 for optimization.

Method	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5
MLVI [39]	45.30	27.70	75.70	59.20
TMLGA [9]	51.28	33.04	-	-
CBP [36]	54.30	35.76	77.63	65.89
ABLR [44]	55.68	36.79	-	-
2D-TAN [47]	56.92	42.08	82.64	73.01
DRN [45]	-	**43.95**	-	**74.87**
LGI [24]	58.52	41.51	-	-
DeNet(ours)	**61.93**	43.79	**86.02**	74.13

Table 2. Comparison with state-of-the-art methods on ActivityNet Captions (combination of two val_sets) using standard metrics; bold font indicates best results, underlined second-best.

4.3. Comparison with state-of-the-arts

First of all, we compare our model DeNet with other state-of-the-art methods using standard metrics on two datasets, which contains CTRL [6], SMRL [38], MAC [8], MLVI [39], CBP [36], SAP [4], MAN [46], 2D-TAN [47], EXCL [9], TMLGA [9], DRN [45], SCDM [43], LGI [24] and ABLR [44]. Table 1 and Table 2 report the results on Charades-STA and ActivityNet Captions, respectively. For a fair comparison, all of the performances listed in Table 2 are based on the combination of two validation sets on ActivityNet Captions. In the standard metrics, our method DeNet achieves competitive performances on both datasets, especially on the Charades-STA dataset. For example, DeNet obtains 3.04% gains in "R@1,IoU=0.7" and 6.78% gains in "R@5,IoU=0.7".

Then, to better evaluate performances of multiple predictions, we compare our model DeNet with some related methods (including 2D-TAN [47], DRN [45] and SCDM [43])[2] using "R@(N, G), IoU=α" and "R$_\beta$@(N, G), IoU=α". In this work, we take account of at most 5 predictions (N = 5) and 5 temporal annotations (G = 5). To reserve an average of 3 annotations for each query, β is set to 0.5 on Charades-STA, and 0.4 on ActivityNet Captions. Figure 4 illustrates the results. In contrast to performances in standard metrics, proposal-based methods (*i.e.* 2D-TAN and SCDM) outperform the proposal-free method (*i.e.* DRN) in new multi-label metrics. It means proposal-based methods tend to better cover multiple-styles annotations, yet most proposal-free models are biased to single-style annotations. We consider it is because most proposal-free models tend to produce dense predictions. However, our proposal-free-based DeNet still outperforms the above methods on both datasets, *e.g.* 1.75% gains on ActivityNet Captions in terms of R@(5,5). It validates our method has an advantage in matching the multi-styles annotations.

[2] We test 2D-TAN and DRN using pretrained official models and SCDM using third-party implementation [25].

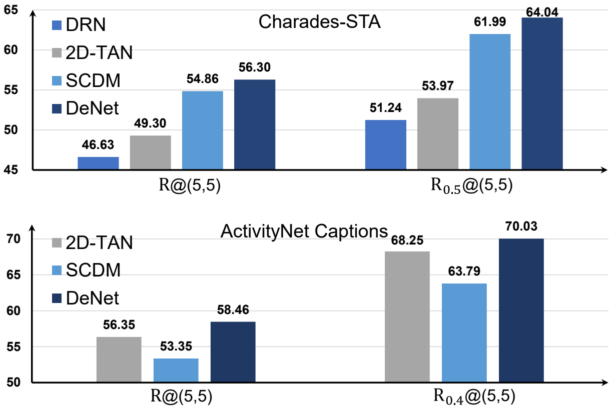


Figure 4. Performances on Charades-STA (top) and ActivityNet Captions (bottom) using multi-label metrics(IoU = 0.5), and at most 5 predictions and 5 annotations are taken into consideration. Best viewed in color.

Method	DRN [45]	2D-TAN [47]	SCDM [43]	DeNet
D_{var}	0.338	0.365	0.286	**0.223**

Table 3. Comparison of robustness for query uncertainty on Charades-STA. The lower value represents more consistent predictions for two siamese queries.

4.4. Ablation studies

Robustness for query uncertainty. We conduct experiments to evaluate the robustness for query uncertainty. Specifically, we explored whether predictions of models can be consistent when using different queries in the same temporal moment. A subset is selected from the Charades-STA testing set, where each temporal moment contains two queries. If a moment contains more queries in the original testing set, we randomly select two queries. Finally, the subset is composed of 848 testing samples (corresponding to 1696 queries). Then, we use $D_{var} = 1 - \text{IoU}$ to compute the average distance between top-1 predictions of two queries. The lower value of D_{var} represents more consistent predictions for the two corresponding queries. Table 3 shows a comparison between DeNet with some methods. Our DeNet outperforms them by 6.3%, which validates the robustness of our model for query uncertainty.

Robustness for label uncertainty. To evaluate the robustness for label uncertainty, we add perturbations in the temporal boundaries to enlarge the label uncertainty. During the training procedure, we take annotations t_s, t_e and generate new annotations $\hat{t}_s = t_s + \epsilon_s(t_e - t_s)$, $\hat{t}_e = t_e + \epsilon_e(t_e - t_s), \epsilon_s, \epsilon_e \in\sim U(-0.5, 0.5)$, where $U(-0.5, 0.5)$ is uniform distribution. We train our DeNet and 2D-TAN using new annotations, then still evaluate them using original annotations. For a fair comparison, 2D-TAN adopts the same I3D feature with DeNet. Here, we mainly investigate the impact of label uncertainty on multiple predictions. Figure 5 shows different results in the multi-label metrics, where ”original” adopts the previous annotations, and ”noise” adopts the new annotations. Compared to 2D-TAN, DeNet only drops slightly using noise annotations, *e.g.* 0.35% vs 5.54% in ”$\text{R}_{0.5}$@(5,5), IoU = 0.5”. It also means that our method can mitigate the reliance on precise annotations in real scenarios.

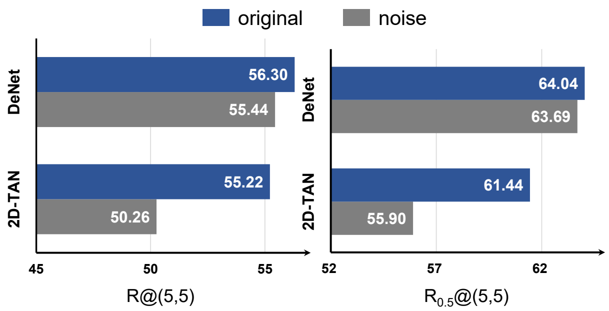


Figure 5. Performances of our DeNet and 2D-TAN with original annotations and noise annotations using multi-label metrics(IoU = 0.5). Best viewed in color.

Method	R@1	R@5	R@(5,5)	$\text{R}_{0.5}$@(5,5)
	IoU=0.5	IoU=0.5	IoU=0.5	IoU=0.5
DeNet w/o PoS	57.47	90.90	52.64	58.97
DeNet-Relation	58.12	**91.34**	54.32	61.76
DeNet-All	58.23	88.76	46.20	50.32
DeNet	**59.70**	91.24	**56.30**	**64.04**

Table 4. Ablation studies of language encoding on Charades-STA; bold font indicates best results.

Analysis on language encoding. In this subsection, we investigate the contribution of the language encoding under query uncertainty and set three variant implements. 1) ”DeNet w/o PoS” encodes entire language without PoS. 2) ”DeNet-Relation” encodes the relation feature as a Gaussian distribution rather than modified feature. 3) ”DeNet-All” encodes both relation feature and modified feature as Gaussian distributions. Table 4 shows the results.

Firstly, it’s more effective to disentangle language into two types of features (DeNet) than a single feature (DeNet w/o PoS). DeNet benefits from Parts-of-Speech parsing when extracting discriminative features. Secondly, for producing multiple predictions, it’s more beneficial to encode the modified feature as Gaussian distribution instead of the relation feature (DeNet-Relation). Thirdly, when both two types of features are encoded as distributions (DeNet-All), it will cause performance degradation.

Analysis on temporal regression. In this subsection, we investigate the contribution of our temporal regression under label uncertainty. Firstly, we set two variant imple-

Method	R@1 IoU=0.5	R@5 IoU=0.5	R@(5,5) IoU=0.5	$R_{0.5}$@(5,5) IoU=0.5
DeNet-Boundary	57.88	89.25	55.42	63.14
DeNet-Centerness	57.85	89.17	54.40	62.18
DeNet-Single	57.45	89.19	55.38	62.95
DeNet w/o min-loss	58.90	69.11	42.18	50.61
DeNet	**59.70**	**91.24**	**56.30**	**64.04**

Table 5. Ablation studies of temporal regression on Charades-STA; bold font indicates best results.

ments to validate the benefit of predicting the center-width as an auxiliary head. 1) ”DeNet-Boundary” only predicts the start-end boundary. 2) ”DeNet-Centerness” only predicts the center-width. As shown in Table 5, when supervised from two perspectives, our model DeNet can obtain gains in terms of all metrics.

Secondly, we set two variant implements to investigate settings of two independent branches. 1) ”DeNet-Single” represents that we only build a single-output branch. 2) ”DeNet w/o min-loss” replaces $\mathcal{L}_{multi}$ with $\mathcal{L}_{single}$ for multi-output branch. Table 5 summarizes different results. The original DeNet with two independent regression branches outperforms the model with only a single-output branch (DeNet-Single). For each sample, the single-output branch aims at matching the single-style annotations, yet the multi-output branch aims at matching potential multiple annotations. We consider the two different tasks may disturb each other once relied on one same branch. In terms of multiple predictions, performances will drop dramatically without min-loss (DeNet w/o min-loss), *e.g.* 22.13% drop in ”R@5, IoU = 0.5”. Thus, min-loss is necessary to learn multiple predictions for the multi-output branch.

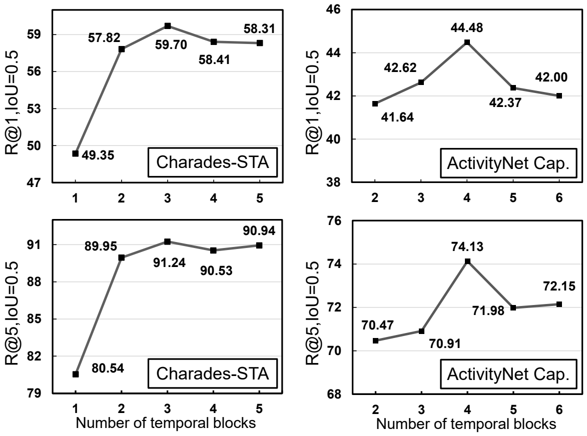


Figure 6. Effect of the number of stacked temporal blocks on Charades-STA and ActivityNet Captions.

Thirdly, we analyze the impact of the number of stacked temporal blocks. Each temporal block contains a Temporal Convolutional layer and a Multi-head Attention layer. Figure 6 shows results on the Charades-STA and ActivityNet Captions. We observe that our proposed method DeNet achieves best performances when the number of stacked temporal blocks reaches 3 for Charades-STA and 4 for ActivityNet Captions. We consider that fewer temporal blocks can not capture the long-range temporal dependencies, yet more temporal blocks may face over-fitting risk.

Qualitative results. Figure 7 illustrates multiple predictions generated by DeNet. We can find the temporal boundaries of different annotations exist disagreement for the same query. For the same query, the multiple predictions generated by DeNet can match each annotation as much as possible. For the same event, predictions of different queries (*i.e.* Query A and Query B) tend to be consistent.

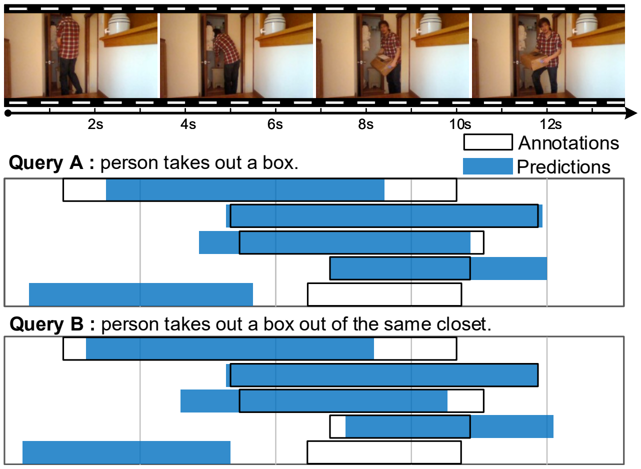

2s
4s
6s
8s
10s
12s
Annotations
Predictions
Query A : person takes out a box.
Query B : person takes out a box out of the same closet.

Figure 7. Qualitative results on Charades-STA dataset.

5. Conclusion

In this paper, we propose DeNet to embrace human uncertainty for temporal grounding. Firstly, DeNet adopts a decoupling method to decompose each query into relation feature and modified feature by PoS, where consistent query information and expression variance can be obtained respectively. Then, DeNet uses a de-bias mechanism to produce diverse yet plausible predictions, aims to mitigate the reliance on single-style annotations. Experiments on two datasets validate its effectiveness and robustness.

6. Acknowledgments

This work was partly funded by the National Key Research and Development Program (2017YFB1002401), NSFC(No.61971281), and STCSM(18DZ2270700).

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Int. Conf. Comput. Vis.*, pages 5803–5812, 2017. 2, 5

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 961–970, 2015. 2, 6

[3] João Carreira, Andrew Zisserman, and Quo Vadis. Action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4724–4733, 2018. 6

[4] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *AAAI*, volume 33, pages 8199–8206, 2019. 6

[5] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In *Adv. Neural Inform. Process. Syst.*, pages 3059–3069, 2018. 1

[6] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Int. Conf. Comput. Vis.*, pages 5267–5275, 2017. 2, 5, 6

[7] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees G. M. Snoek. Actor and action video segmentation from a sentence. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2018. 1

[8] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 245–253. IEEE, 2019. 6

[9] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. Excl: Extractive clip localization using natural language descriptions. *arXiv preprint arXiv:1904.02755*, 2019. 6

[10] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In *Adv. Neural Inform. Process. Syst.*, pages 1799–1807, 2012. 2, 5

[11] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*, volume 33, pages 8393–8400, 2019. 2

[12] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. *arXiv preprint arXiv:1809.01337*, 2018. 5

[13] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, pages 123–141. Springer, 2018. 1

[14] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Gaining extra supervision via multi-task learning for multi-modal video question answering. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 1

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

[16] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Int. Conf. Comput. Vis.*, pages 706–715, 2017. 2, 6

[17] Kimin Lee, Changho Hwang, Kyoung Soo Park, and Jinwoo Shin. Confident multiple choice learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2014–2023, 2017. 3

[18] Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. In *Adv. Neural Inform. Process. Syst.*, pages 2119–2127, 2016. 3

[19] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, 2018. 1

[20] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *ACM Int. Conf. Multimedia*, pages 988–996, 2017. 1

[21] Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Temporal modular networks for retrieving complex compositional activities in videos. In *Eur. Conf. Comput. Vis.*, pages 552–568, 2018. 2

[22] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 15–24, 2018. 2

[23] Weiwei Liu, Chongyang Zhang, Jiaying Zhang, and Zhonghao Wu. Global for coarse and part for fine: A hierarchical action recognition framework. In *IEEE Int. Conf. Image Process.*, pages 2630–2634. IEEE, 2018. 1

[24] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10810–10819, 2020. 2, 6

[25] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. In *Brit. Mach. Vis. Conf.*, 2020. 5, 6

[26] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4594–4602, 2016. 1

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 4

[28] Cristian Rodriguez, 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 *The IEEE Winter Conference on Applications of Computer Vision*, pages 2464–2473, 2020. 2

[29] Gunnar A Sigurdsson, Olga Russakovsky, and Abhinav Gupta. What actions are needed for understanding human actions in videos? In *Int. Conf. Comput. Vis.*, pages 2137–2146, 2017. 2

[30] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Eur. Conf. Comput. Vis.*, pages 510–526. Springer, 2016. 5

[31] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inform. Process. Syst.*, 1, 2014. 1

[32] Kai Tian, Yi Xu, Shuigeng Zhou, and J. Guan. Versatile multiple choice learning and its application to vision computing. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6342–6350, 2019. 3

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

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, pages 5998–6008, 2017. 4

[35] Anran Wang, Anh Tuan Luu, Chuan-Sheng Foo, Hongyuan Zhu, Yi Tay, and Vijay Chandrasekhar. Holistic multi-modal memory network for movie question answering. *IEEE Trans. Image Process.*, 29:489–499, 2019. 1

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

[37] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Eur. Conf. Comput. Vis.*, 2016. 1

[38] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 334–343, 2019. 2, 6

[39] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, volume 33, pages 9062–9069, 2019. 2, 6

[40] Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. Data augmentation for spoken language understanding via joint variational generation. In *AAAI*, volume 33, pages 7402–7409, 2019. 4

[41] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2016. 1

[42] Tianyuan Yu, Da Li, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. Robust person re-identification by modelling feature uncertainty. In *Int. Conf. Comput. Vis.*, pages 552–561, 2019. 4

[43] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *Adv. Neural Inform. Process. Syst.*, pages 536–546, 2019. 2, 6, 7

[44] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, volume 33, pages 9159–9166, 2019. 2, 5, 6

[45] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10287–10296, 2020. 1, 2, 6, 7

[46] 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 Conf. Comput. Vis. Pattern Recog.*, pages 1247–1257, 2019. 6

[47] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks formoment localization with natural language. In *AAAI*, 2020. 1, 2, 3, 6, 7

[48] Songyang Zhang, Jinsong Su, and Jiebo Luo. Exploiting temporal relationships in video moment localization with natural language. In *ACM Int. Conf. Multimedia*, pages 1230–1238, 2019. 2

[49] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 655–664, 2019. 2

[50] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Int. Conf. Comput. Vis.*, Oct 2017. 1