Phrase-level Temporal Relationship Mining for Temporal Sentence Localization

Minghang Zheng[1], Sizhe Li[1], Qingchao Chen[2], Yuxin Peng[1,3], Yang Liu[1,4*]

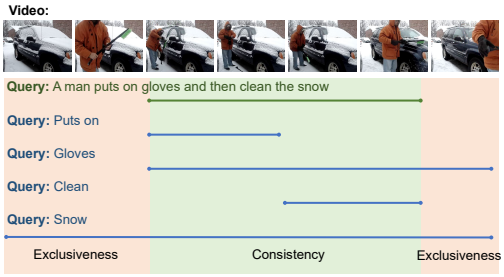
[1]Wangxuan Institute of Computer Technology, Peking University, Beijing, China
[2]National Institute of Health Data Science, Peking University, Beijing, China
[3]Peng Cheng Laboratory, Shenzhen, China
[4] National Key Laboratory of General Artificial Intelligence, BIGAI, Beijing, China
{minghang, lisizhe, qingchao.chen, pengyuxin, yangliu}@pku.edu.cn

Abstract

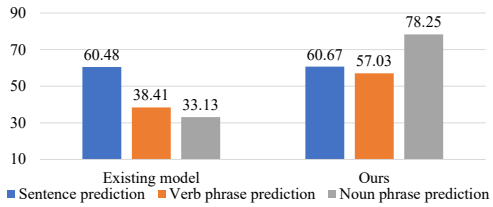
In this paper, we address the problem of video temporal sentence localization, which aims to localize a target moment from videos according to a given language query. We observe that existing models suffer from a sheer performance drop when dealing with simple phrases contained in the sentence. It reveals the limitation that existing models only capture the annotation bias of the datasets but lack sufficient understanding of the semantic phrases in the query. To address this problem, we propose a phrase-level Temporal Relationship Mining (TRM) framework employing the temporal relationship relevant to the phrase and the whole sentence to have a better understanding of each semantic entity in the sentence. Specifically, we use phrase-level predictions to refine the sentence-level prediction, and use Multiple Instance Learning to improve the quality of phrase-level predictions. We also exploit the consistency and exclusiveness constraints of phrase-level and sentence-level predictions to regularize the training process, thus alleviating the ambiguity of each phrase prediction. The proposed approach sheds light on how machines can understand detailed phrases in a sentence and their compositions in their generality rather than learning the annotation biases. Experiments on the ActivityNet Captions and Charades-STA datasets show the effectiveness of our method on both phrase and sentence temporal localization and enable better model interpretability and generalization when dealing with unseen compositions of seen concepts. Code can be found at https://github.com/minghangz/TRM.

Introduction

Video temporal sentence localization has become an important research problem due to its potential for a wide range of practical applications, requiring intelligent systems to identify the start and end timestamps of segments (i.e., moments) with respect to any given language queries in an untrimmed video. Using free-form natural language as queries allows users to freely search for interesting content without being restricted to pre-defined classes, which makes sentence localization have greater application potential. The model is expected to understand the visual and language concepts and their compositions to achieve robust performance.



(a) Sentence and phrase level prediction.



(b) Performance on Charades-STA (IoU@0.3).

Figure 1: (a) The sentence-level (in green) and phrase-level (in blue) prediction. We make two assumptions about the relationship between phrases and sentences: 1) Consistency: for each phrase, the phrase-level prediction should overlap the sentence ground truth (in green); 2) Exclusivity: for each video clip that does not intersect with sentence ground truth (in red), at least one phrase's prediction does not overlap it. (b) shows the evaluation results of the existing model (Wang et al. 2021b) and our method on the Charades-STA (R@1, IoU=0.3) when using sentences or phrases as queries.

Fully supervised approaches have made steady progress in the last decades when the queries are complete sentences. However, human-generated queries 'in the wild' vary a lot in terms of specificity, we expect models to deal with both complete sentences (the query marked in green in Fig. 1(a)) and short phrases (the query marked in blue in Fig. 1(a)) to be competent for real-world applications. However, we empirically observe that even the most recent open-source models learned by using sentence annotations lack the capability to deal with the phrase-level query, as shown in Fig. 1(b). We evaluate the existing method (Wang et al. 2021b) on the

*Corresponding author

Charades-STA dataset, and observe a sheer drop in prediction accuracy: IoU@0.3 is dropped by 22.07% and 27.35% when dealing with simpler verb queries and noun queries.

Usually, a word or group of words forms a syntactic constituent with a single grammatical function (ie. verb, subject, or object), representing a more straightforward semantic meaning than sentences (no need to understand their compositions). The typical failure in much more straightforward scenarios reveals the following problems. *First, existing models tend to capture the annotation bias in the benchmark but lack sufficient understanding of the intrinsic relationship between simple visual and language concepts.* Consequently, existing models may easily fail when the unrealistic assumption of the in-distribution test setting does not hold, i.e., incapable of generalizing to novel combinations of visual entities and text, which is also revealed by (Otani et al. 2020; Yuan et al. 2021; Li et al. 2022a). *Second, the models' interpretability and robustness are questioned* since they fail to deal with simple (atomic) concepts, even though they achieve decent results in sentence-level prediction tasks. This may hinder the application of these methods in real scenarios.

Motivated by the above observations, we attempt to take phrase-level prediction into consideration of temporal localization models' designation. To avoid the high annotation cost and subjective annotation bias of fine-grained phrases, we propose phrase-level Temporal Relationship Mining (TRM) framework to improve the phrase temporal localization using sentence-level supervision only. The two key ideas underpinning this framework are as follows. First, inspired by the successful application of Multiple Instance Learning (MIL) to weakly supervised temporal sentence localization, we train the model to discriminate between matched and unmatched video-phrase pairs without phrase-level annotations. Second, in order to consider the constraints of sentence-level annotations on phrase-level predictions, we exploit the temporal localization relationship relevant to the phrase and the whole sentence and follow the two design principles -*consistency* and *exclusiveness*. Specifically, *consistency* requires every phrase-level prediction should share a period with the annotated sentence-level ground truth. As shown in Fig 1(a), all predictions of the phrases "puts on", "gloves", "clean" and "snow" should overlap with the sentence ground truth annotation (in green). *Exclusiveness* requires that every period not intersect the sentence ground truth (as shown in red boxes in Fig 1(a)) is at least excluded from one phrase-level prediction (not intersect at least one phrase prediction). Combining the above two key ideas, the performance of our model on phrase level prediction has been significantly improved (18.62% improvement for verb phrases and 45.12% improvement for noun phrases in Fig. 1(b)).

Our contributions are summarized as follows: **(1)** We highlight the importance of phrases in video temporal localization and exploit the temporal relationship relevant to phrases and the whole sentence. **(2)** We propose phrase-level Temporal Relationship Mining (TRM) framework to investigate phrase-level prediction using sentence-level supervision only, which proposes the consistency and exclusiveness constraints to regularize the training process. **(3)** Experiments on Charades-STA and ActivityNet Captions demonstrate our method's ability to improve phrase-level performance while performance in sentence-level settings remains stable, achieving better generalization performance.

Related Work

Temporal Sentence Localization

Since being proposed by TALL (Gao et al. 2017), the task has drawn wide attention. Most previous methods either generate candidate proposals and rank them using multi-modal features (Zhang et al. 2020b), or use multi-modal features to generate timestamp predictions directly (Zhang et al. 2020a). Recent works have started to consider fine-grained vision and language information. For example, for vision information, DORi (Rodriguez-Opazo et al. 2021) and MARN (Liu et al. 2022b) consider the features of objects within the video and improve models' performance. Correspondingly, for language features. LGI (Mun, Cho, and Han 2020) generates sub-query features to implicitly consider fine-grained text features and boost sentence localization performance. MMN (Wang et al. 2021b) trains the model to distinguish matched and unmatched video-sentence pairs collected from both intra-video and inter-video. MGSL-Net (Liu et al. 2022a), which uses memory to reinforce uncommon samples in the training process. EMB (Huang et al. 2022) constructs elastic boundaries to handle the uncertainties in temporal boundary. VISA (Li et al. 2022a) considers the distribution of different entities and conducts the Charades-CG and ActivityNet-CG dataset splits to test the compositional generalization, where the novel composition of seen phrases will appear in the test split. However, we found that existing approaches perform poorly when using simpler phrases as queries, suggesting that they do not really understand the intrinsic connection between vision and language. In this paper, we propose a unified framework dealing with both sentence and phrase queries simultaneously and improve the performance.

Multiple Instance Learning

Multiple Instance Learning (MIL) has been widely applied in computer vision, such as content-based image retrieval (Song et al. 2013), object localization, and segmentation (Xu et al. 2015), computer-aided diagnosis and detection (Xu et al. 2014), etc. Although (Huang et al. 2021; Yang et al. 2021; Huang et al. 2021; Zheng et al. 2022a,b) use MIL to solve the weakly supervised temporal sentence localization, where only videos and natural language queries are available during training, no previous work has tried to use it to solve the phrase-level video temporal localization problem.

Moreover, directly regarding phrase-level prediction as a weakly supervised task and introducing MIL ignores the constraint of sentence-level annotations on phrase-level predictions. Thus, we exploit the relationship between phrase-level predictions and sentence-level annotations and put forward the assumptions of consistency and exclusivity.

Phrase in Vision-Language Tasks

Phrase-level features can provide models with more fine-grained text representations and have wide applications in vision-language tasks, such as video grounding (Mun, Cho, and Han 2020; Rohrbach et al. 2016), video captioning (Ryu et al. 2021; Zhang and Peng 2019), etc. LGI (Mun, Cho, and Han 2020) first exploits sub-query features. However, it simply fuses them in an early stage to obtain fine-grained sentence features, neither locating the phrases directly nor considering the relationship between the localization results of phrases and sentence. This results of LGI still perform poorly when encountering phrases as queries, as shown in Tab. 2. For the first time, PLPNet (Li et al. 2022b) directly considers the problem of locating a phrase and improves the performance of phrase-level localization through contrastive learning. However, it has no extra constraints on phrase-level predictions and sentence-level predictions, dismissing the intrinsic connection between the video periods related to a sentence and its phrases. In this paper, we propose a unified framework to deal with both sentence and phrase queries simultaneously and improve the performance of both. We introduce constraints from the perspective of prediction results so that the TRM model can directly supervise the predicted phrase-level timestamps without extra phrase-level annotations. To our best knowledge, we are the first to investigate the temporal relationship between phrase-level prediction and sentence-level prediction explicitly. This setting is more in line with real-world application scenarios and enables the model to generalize to unseen combinations of seen phrases.

Method

Overview

Fig. 2 illustrates the overall architecture of our proposed TRM framework. We first extract video representation and generate a 2D Temporal Map (Zhang et al. 2020b). Meanwhile, the query encoder generates phrases and extracts text features for both phrases and sentences. To represent the similarity between the text and each video proposal, we generate score maps using the 2D temporal map and the text feature for sentences and all phrases. Due to the lack of phrase-level annotation, we explored the *consistency* and *exclusiveness* relationship between phrases and sentences as the loss function to regularize the training process and improve the accuracy of phrase score maps. Since the phrase-level score maps can provide more fine-grained information for the sentence, we use them to refine the sentence score map with a weighted sum option as well, and the weight of each phrase represent its importance. Finally, we optimize the refined sentence score map with an IoU regression loss and a contrastive learning loss.

Model Architecture

Video Encoder The video encoder aims at extracting video features and generating a 2D temporal map for similarity learning. We extract features from the input video and encode them as a 2D temporal adjacent feature map following MMN (Wang et al. 2021b). For an input video, we first split it into small video clips, each containing equal frames. Then we extract the clip-level visual feature with a pre-trained CNN model. We can obtain N clip-level features $\{f_i^V\}_{i=1}^N \in \mathbb{R}^{N\times d}$, where N is the number of clips and d is the feature dimension. Then, we build up the 2D proposal feature map $F^V \in \mathbb{R}^{N\times N\times d}$ following MMN (Wang et al. 2021b), where proposal $F_{i,j}^V$ represents the video candidate starting from the i-th clip and ending with the j-th clip.

Query Encoder The query encoder aims to generate fine-grained phrases for a sentence and extract both sentence and phrase-level text features. More specifically, given a query sentence S, we first parse N_p phrases $[p_1, p_2, ..., p_{N_p}]$ using pre-trained SRLBERT(Shi and Lin 2019). SRLBERT assigns semantic role labels to each word in the sentence, while we only keep the semantic roles with more than 1000 occurrences in the training set as phrases. Then, we use a pre-trained DistilBERT (Sanh et al. 2019) model following MMN (Wang et al. 2021b) to extract the features of sentences and phrases at the same time. Phrases provide fine-grained information to the sentence, and the sentence provides global information to phrases. Therefore, we further interact sentence and phrase features through a single-layer transformer encoder (Vaswani et al. 2017). The final sentence feature and phrase features are represented as $f^S \in \mathbb{R}^d$ and $f^P \in \mathbb{R}^{N_p\times d}$ respectively.

Similarity Learning Module To learn the semantic relevance of each sentence and phrase with each temporal proposal, we generate score maps for both sentence and phrases according to the similarity of text and video features. In order to improve the quality of phrase score maps, we propose two assumptions of consistency and exclusivity to constrain the phrase score maps. Since phrases provide finer-grained semantic information for sentences, we use the phrase score maps to refine the sentence score map so that it can summarize the attentional information for each phrase. We use a weighted sum option over the phrase score maps and leverage phrase weights to describe the importance of different phrases. Finally, we optimize the refined sentence score map with an IoU regression loss and a contrastive learning loss.

Score Map Generation. For the sentence, we perform 1×1 convolution operation on visual feature map F and perform a linear projection on text features f^S respectively to project the features of two modalities into the same dimension d^H. The final representations of sentence features $f_{iou}^S \in \mathbb{R}^{d^H}$ and visual features $F_{iou}^V \in \mathbb{R}^{N\times N\times d^H}$ are:

$$f_{iou}^S = \text{FC}_{iou}(f^S), F_{iou}^V = \text{Conv}_{iou}(F^V) \tag{1}$$

where $\text{FC}(\cdot)$ is a fully connected network and $\text{Conv}(\cdot)$ is an 1×1 convolution. Then we regard the cosine similarity of f_{iou}^S and F_{iou}^V as sentence-level score map: $S^s = F_{iou}^{VT} f_{iou}^S \in \mathbb{R}^{N\times N}$, in which $S_{i,j}^s$ represents the similarity score between the sentence and the proposal from the i-th video clip to the j-th video clip.

Temporal Relation Mining. In previous works (Wang et al. 2021b; Zhang et al. 2020b), the sentence score map is directly used to predict the timestamps. However, it dismisses the fine-grained phrases inside the query, and has poor performance when the query is a single phrase. To solve

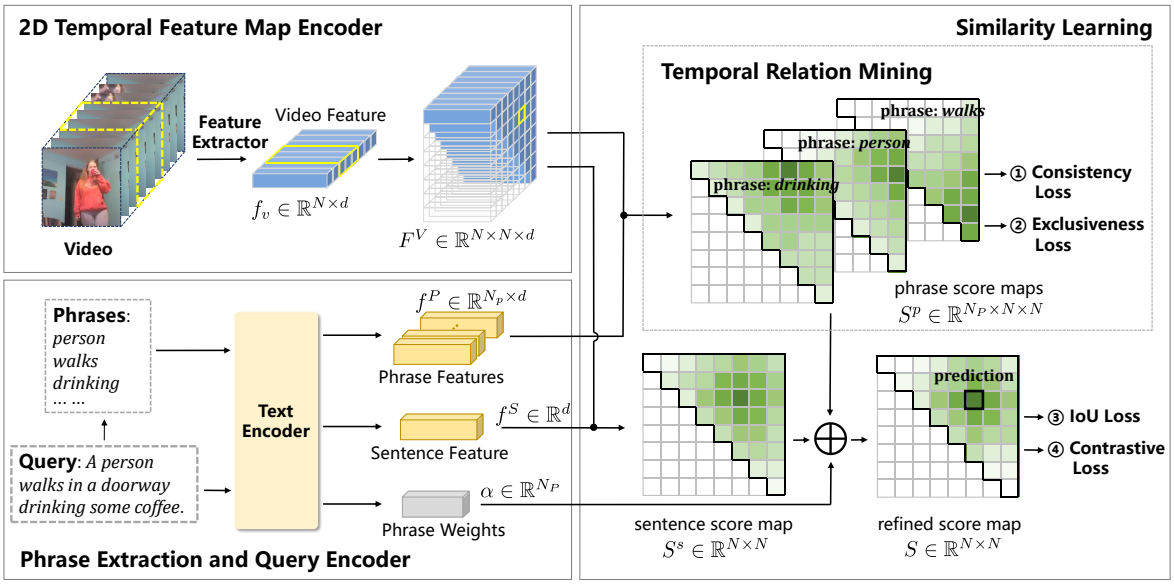


Figure 2: Our proposed TRM model framework focuses on the temporal relationship between a sentence and its phrases. Our model consists of three modules: a video encoder extracts video features and generates a 2D temporal map; a query encoder extracts both sentence-level and phrase-level features and a similarity learning module to mine the temporal relationship of phrases and sentences based on our two constraints(consistency and exclusiveness) and leverage sentence-level contrastive learning. We apply the phrase-level constraint loss considering the intrinsic relationship between sentences and phrases.

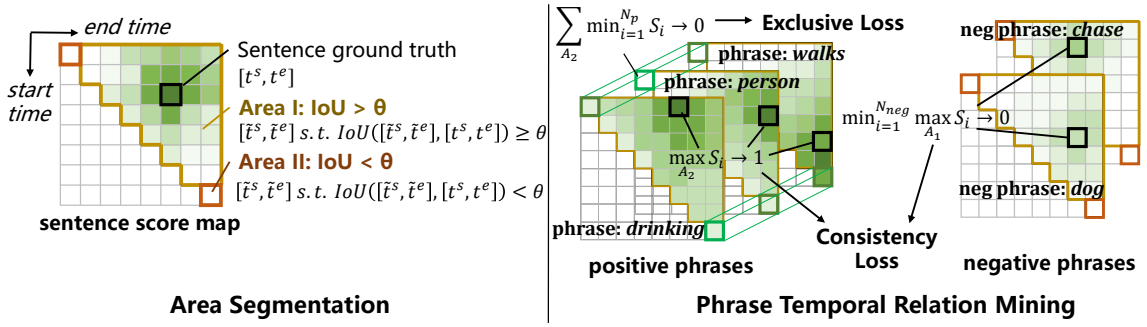


Figure 3: The specific process of proposal segmentation and implementation of our consistency and exclusiveness principles.

this problem, we build phrase score maps and mine the temporal relationship between the phrases and the sentence. Due to the lack of phrase-level annotation data, we impose constraints between the phrase score maps for training purposes. We have the following two hypotheses considering the relationship between phrases and sentences:

1. *consistency*: For paired sentences and videos, every phrase-level prediction should share a period with the annotated sentence-level ground truth. For unpaired sentences and videos, at least one phrase-level prediction does not share a period with the annotated ground truth.
2. *exclusiveness*: Each frame outside the ground truth is not contained in at least one phrase-level prediction result.

In detail, we first obtain the text feature $f^P_{i,iou} \in \mathbb{R}^{d^H}$ for the i-th phrase through Eq (1). Then we regard the cosine similarity as moments' estimation score map S^p of each phrase: $S^p_i = F^{VT}_{iou} f^P_{i,iou} \in \mathbb{R}^{N\times N}$. Inspired by Multiple Instance Learning, we also randomly sample unmatched phrases in a batch and compute their score map $\hat{S}^p$. Based on the degree of intersection with the sentence ground truth, we divide all proposals into two subsets. As shown in the left half of Fig. 3, all the proposals in Area I have an IoU with the ground-truth moment large than a certain threshold θ, while the opposite is true for all proposals in Area II.

Our consistency loss ensures that each phrase-level prediction should be located in Aera I, which is illustrated in Fig.3. That is: for each phrase score map, the max score (marked by black) in Area I should be 1. Our consistency loss also requires that for a negative sentence, there should be at least one phrase that mismatches any proposal in Area I, which is represented in Fig.3 as $\min_{i=1}^{N_{neg}} \max_{A_1} S_i \to 0$. The

consistency loss can be described as follows:

$$\mathcal{L}_{con} = \max_{i=1}^{N_p}(L_f(\max_{(s,t)\in A_1} S_i^p[s,t],1)) + \min_{i=1}^{N_p}(L_f(\max_{(s,t)\in A_1} \hat{S}_i^p[s,t],0)) \quad (2)$$

where $\mathcal{L}_f$ is the focal loss (Lin et al. 2017) to balance the positive and negative samples, A_1 represents Area I, and A_2 represents Area II.

Our exclusiveness loss requires that each proposal in Area II should mismatch at least one phrase of the query sentence. That is: as shown in Fig. 3, at least one of the phrase's scores should be 0 (i.e. the minimum score marked by green should be 0) for all the proposals in Area II. The exclusiveness loss can be described as follows:

$$\mathcal{L}_{ex} = \frac{1}{|A_2|}\sum_{(s,t)\in A_2} L_f(\min_{i=1}^{N_p}(S_i^p[s,t]),0) \quad (3)$$

Sentence Score Map Refinement.. Since the phrase-level score maps can provide more fine-grained information for the sentence, we use them to refine the original sentence score map $S^s \in R^{N\times N}$. We gain the final sentence score map $S \in R^{N\times N}$ by aggregating the score maps of the sentence and all of its phrases, which is shown as follows:

$$\alpha = \text{softmax}(\text{MLP}_{\text{satt}}([p_1, p_2, ..., p_{N_p}])) \quad (4)$$

$$S = S^s + \sum \alpha_i S_i^p \in R^{N\times N} \quad (5)$$

where $\alpha \in \mathbb{R}^{N_p}$is the phrase weights that describe the importance of different phrases, MLP_{satt} denotes a multilayer perception with a output layer of 1-dimension.

To supervise the sentence score map, we apply the binary cross entropy loss to regress the IoU score of each proposal. Following (Zhang et al. 2020b), we adopt a scaled IoU value y_i as the supervision scale, but not a hard binary score. Then the binary cross entropy loss can be expressed as

$$\mathcal{L}_{iou} = -\frac{1}{C}\sum_{i=1}^{C}(y_i \log S_i + (1-y_i)\log(1-S_i)), \quad (6)$$

where C is the number of proposals.

Sentence-level Contrastive Learning. Following MMN (Wang et al. 2021b), we also use contrastive learning to provide more supervised signals to the model. We collect positive and negative sentence-video pairs within and between videos, and use noise contrastive estimation (Oord, Li, and Vinyals 2018) to estimate two conditional distributions $p(s|v)$ and $p(v|s)$. The former represents the probability that a sentence s matches the video v when giving v, and the latter represents the probability that a video v matches the sentence s when giving s. We adopt the contrastive loss to help capture better information between modalities as follows:

$$\mathcal{L}_{cont} = -(\sum_{s\in\mathbb{S}} \log p(v_s|s) + \sum_{v\in\mathbb{V}} \log p(s_v|v)) \quad (7)$$

where $\mathbb{S}, \mathbb{V}$ are the sets of training sentences and video in a batch, v_s is the video that matches the sentence s, and s_v is the sentence that matches the video v.

Training and Inference

Training The total loss of our model is as follows.

$$\mathcal{L} = \mathcal{L}_{iou} + \mathcal{L}_{cont} + \mathcal{L}_{con} + \mathcal{L}_{ex} \quad (8)$$

Given the lack of phrase-level annotations, we can still optimize the understanding of phrases during training with the constraints between the whole sentence and phrases.

Inference At the inference time, when given a sentence query, we can obtain the refined score maps S through Eq(5) to make predictions. When given a single phrase query, we can treat it as a sentence (as the text encoders for phrase and sentence are shared). In this case, the score maps of the sentence and phrase are the same and both can be used to output phrase predictions.

Experiments

Dataset

Charades-STA Charades-STA (Gao et al. 2017) originates from Charades (Sigurdsson et al. 2016) dataset, containing indoor videos with sentence queries and corresponding annotations. There are 12,408 and 3,720 video-query pairs for training and testing respectively. Our sentence-level results are reported on the test split.

ActivityNet Captions ActivityNet Captions (Krishna et al. 2017) contains 20K videos, with 37,417/17,505/17,031 video-query pairs in the train /val_1/val_2 split. We adopt standard splits and report the sentence-level results on the val_2 split.

Experiment Settings

Evaluation Metric. Following (Gao et al. 2017), we adopt the "R@1, IoU $= m$" and mIoU (the mean average IoU) metrics to evaluate the model's performance. Specifically, this metric evaluates the percentage of predicted moments that have the temporal Intersection over Union (IoU) larger than the threshold m, and m is set to $\{0.3, 0.5, 0.7\}$.

Evaluation for phrase. When evaluating the performance of phrases, we use a single phrase rather than a complete sentence as the query, in which case the score map of the sentence and phrase is the same and both can be used to output predictions. Due to the lack of phrase-level annotations, we adopt the action annotation used for the Temporal Action Localization task and use the action names as the query phrases. Although we only tested with verbs, our model can handle arbitrary phrases. To prove this, we also use the object annotations on the Charades-STA dataset provided by (Yuan et al. 2017). We collect the common noun phrases in the sentences, and get the time of the first appearance and the last disappearance of the object in the object annotation as the noised noun phrase ground truth timestamps. We report the evaluation results of our model when using noun phrases as queries in the ablation section. It is worth noting that we only use the phrase-level annotations for evaluating the model's performance on phrases, and avoid using them in the training process. So our experiment setting is fair compared with others.

Method	feature	sentence prediction				phrase prediction			
		IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
SAP (Chen and Jiang 2019)	VGG	—	27.42	13.36	—				
MAN (Zhang et al. 2019)		—	41.24	20.54	—				
LGI (Mun, Cho, and Han 2020)		57.20	40.70	20.13	38.75				
FVMR (Gao and Xu 2021)		—	42.36	24.14	—				
DRN (Zeng et al. 2020)		—	42.90	23.68	—				
SSCS (Ding et al. 2021)		—	43.15	25.54	—				
CBLN (Liu et al. 2021)		—	43.67	24.44	—				
CPN (Zhao et al. 2021)		**64.41**	46.08	25.06	**43.90**				
2D-TAN (Zhang et al. 2020b)		57.31	42.8	23.25	—	45.15	23.22	10.14	—
MMN (Wang et al. 2021b)		60.48	47.45	27.15	—	38.41	22.19	10.1	—
PLPNet (Li et al. 2022b)		57.82	41.88	20.56	39.12	46.24	22.94	7.69	28.46
TRM (ours)	VGG	60.67	**47.77**	**28.01**	42.77	**57.03**	**33.69**	**11.86**	**35.82**

Table 1: Sentence-level and Phrase-level prediction accuracy on Charades-STA.

Method	feature	sentence prediction				phrase prediction			
		IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
DORi (Rodriguez-Opazo et al. 2021)	C3D	57.89	41.49	26.41	42.78				
BPNet (Xiao et al. 2021)		58.98	42.07	24.69	42.11				
VSLNet (Zhang et al. 2020a)		63.16	43.22	26.16	43.19				
DeNet (Zhou et al. 2021)		61.93	43.79	—	—				
CPN (Zhao et al. 2021)		62.81	45.10	28.10	45.70				
DRN (Zeng et al. 2020)		—	45.45	24.36	—				
SeqPAN (Zhang et al. 2021a)		61.65	45.50	28.37	45.11				
FIAN (Qu et al. 2020)		64.10	47.90	29.81	—				
CBLN (Liu et al. 2021)		66.34	48.12	27.60	—				
SMIN (Wang et al. 2021a)		—	48.46	30.34	—				
MGSL-Net (Liu et al. 2022a)		—	**51.87**	**31.42**	—				
LGI(Mun, Cho, and Han 2020)		58.48	41.65	24.1	41.48	35.39	21.07	9.76	25.14
2D-TAN(Zhang et al. 2020b)		59.45	44.51	27.38	—	51.71	42.19	32.22	—
MIGCN(Zhang et al. 2021b)		60.03	44.94	27.85	43.59	42.25	33.75	16.37	30.9
RaNet(Gao et al. 2021)		60.96	45.59	28.67	44.82	47.44	37.51	27.58	38.45
MMN(Wang et al. 2021b)		65.05	48.59	29.26	—	51.91	42.27	32.88	—
PLPNet (Li et al. 2022b)		56.92	39.20	20.91	39.53	50.10	38.12	25.24	37.96
TRM (ours)	C3D	**66.41**	50.44	31.18	**47.68**	**52.46**	**42.84**	**33.68**	**43.29**

Table 2: Sentence-level and phrase-level prediction accuracy on ActivityNet Captions.

Implementation Details. For the 2D temporal feature map encoder, we use exactly the same settings with 2D-TAN (Zhang et al. 2020b) and MMN (Wang et al. 2021b) for fair comparisons. We use the VGG (Simonyan and Zisserman 2014) features for the Charades-STA dataset and C3D features (Tran et al. 2015) for the ActivityNet Captions dataset, and the number of sampled clips N is 16 for Charades-STA and 64 for ActivityNet Captions. For the text encoder, we use the HuggingFace (Wolf et al. 2019) implementation of DistilBERT (Sanh et al. 2019) with pre-trained model following MMN (Wang et al. 2021b). We use AdamW (Loshchilov and Hutter 2017) optimizer with learning rate 1×10^{-4} and batch size 12 for Charades, learning rate 1×10^{-4} and batch size 20 for ActivityNet Captions. The learning rate of DistilBERT is 1/10 of our main model.

Comparison with Other Methods

This part compares state-of-the-art models and TRM's ability to deal with sentence-level and phrase-level prediction. On both Charades-STA and ActivityNet Captions datasets, we use sentences and verb phrases (obtained from action labels used for the temporal action localization task) as queries respectively. We reproduce some of the open-source methods to test the performance of phrase-level localization. For fair comparison, all methods use C3D (Tran et al. 2015) features on ActivityNet Captions and VGG (Simonyan and Zisserman 2014) features on Charades-STA.

As shown in Tab. 1, TRM achieves comparable results when using completed sentences as queries and achieves an absolute advantage when using verb phrases as queries. All the existing methods we reproduced have a sheer drop when using phrases as queries. This reveals that existing models lack sufficient understanding of the intrinsic relationship between simple visual and language concepts. As shown in Tab. 2, on ActivityNet Captions, our sentence prediction is 1.92% higher than baseline MMN (IoU=0.7) and achieves comparable results with MGSL-Net.

As shown in Tab. 3, we test the compositional generalization of our method on ActivityNet-CG (Li et al. 2022a) dataset. VISA (Li et al. 2022a) re-splits the ActivityNet datasets and constructs the ActivityNet-CG datasets. The test-trivial split has the same distribution as the training set, the novel-composition split includes unseen compositions of seen phrases, and the novel-word split includes unseen

	Method	Test-Trivial			Novel-Composition			Novel-Word		
		IoU=0.5	IoU=0.7	mIoU	IoU=0.5	IoU=0.7	mIoU	IoU=0.5	IoU=0.7	mIoU
Weakly-supervised	WSLL (Duan et al. 2018)	11.03	4.14	15.07	2.89	0.76	7.65	3.09	1.13	7.10
RL-based	TSP-PRL (Wu et al. 2020)	34.27	18.80	37.05	14.74	1.43	12.61	18.05	3.15	14.34
Proposal-free	LGI (Mun, Cho, and Han 2020)	43.56	23.29	41.37	23.21	9.02	27.86	23.10	9.03	26.95
	VLSNet (Zhang et al. 2020a)	39.27	23.12	42.51	20.21	9.18	29.07	21.68	9.94	29.58
	VISA (Li et al. 2022a)	47.13	29.64	44.02	31.51	16.73	**35.85**	30.14	15.90	35.13
Proposal-based	TMN (Liu et al. 2018)	16.82	7.01	17.13	8.74	4.39	10.08	9.93	5.12	11.38
	2D-TAN (Zhang et al. 2020b)	44.50	26.03	42.12	22.80	9.95	28.49	23.86	10.37	28.88
	TRM (Ours)	**55.22**	**35.06**	**51.85**	**33.80**	**16.86**	35.80	**35.49**	**17.68**	**37.50**

Table 3: Compositional generalization results on ActivityNet-CG dataset.

Method			Sentence prediction			Verb phrase prediction			Noun phrase prediction		
Phrase	Consistency	Exclusiveness	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
✗	✗	✗	60.48	47.45	27.15	38.41	22.19	10.01	33.13	8.17	3.15
✓	✗	✗	59.84	46.65	26.99	41.13	22.63	10.60	35.41	7.36	2.68
✓	✓	✗	60.22	46.56	27.31	56.69	30.85	10.85	71.12	51.67	8.57
✓	✗	✓	60.13	45.89	27.80	38.90	22.11	10.46	36.88	8.63	3.01
✓	✓	✓	**60.67**	**47.77**	**28.01**	**57.03**	**33.69**	**11.86**	**78.25**	**57.10**	**10.17**

Table 4: Ablation studies on the influence of phrase and score map and the implementation of our hypotheses.

words. We achieve the best performance on all the splits, which proves that learning phrase prediction helps generalize to novel phrase compositions and novel words.

Ablation Studies

In this section, we conduct ablative experiments on the Charades-STA dataset to analyze the necessity of phrase-level information and phrase-level constraints.

As shown in Tab. 4, comparing the first and second rows, we find that simply introducing fine-grained phrase features without considering the relationship between phrase and sentence-level predictions has limited performance improvement for phrase prediction. From the third row, we see that consistency loss can greatly improve the performance of phrase prediction. From the fourth row, it can be seen that training with only exclusiveness loss has a negative impact on the model. This is because only the exclusivity loss is incomplete because the all-zero scores map of phrases is a set of trivial solutions. From the fifth row, we can see that the consistency loss and exclusiveness loss together can further improve the performance of both sentences and phrases. The results show that exploiting the consistency and exclusiveness constraints of phrase-level predictions and sentence-level predictions can regularize the training process, thus alleviating the ambiguity of each phrase localization.

Qualitative Results

In Fig. 4, we visualize an example on Charades-STA Dataset. As we see, our prediction for the sentence matches the ground truth (in green) well. Also, TRM understands that the entire sentence consists of three phrases: 'drinking', 'some coffee', and 'walks'. All the predictions satisfy our constraints of consistency and exclusiveness. This shows TRM understands the intrinsic relationship between simple visual and language concepts.

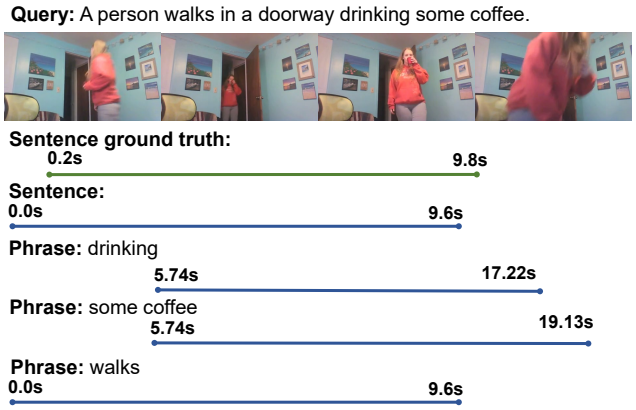


Figure 4: Qualitative results on Charades-STA.

Conclusion

In this work, we propose the phrase-level Temporal Relationship Mining (TRM) framework considering both phrase and sentence queries, making the first attempt to mine the phrase-proposal relation in the temporal localization task. We develop a method to constrain phrase-level prediction in training, tackling the lack of phrase-level annotation. We propose the consistency and exclusiveness constraints of phrase-level and sentence-level predictions to regularize the training process, thus alleviating the ambiguity of each phrase prediction. Experimental results on Charades-STA and ActivityNet Captions indicate that our model surpasses other models in phrase-level prediction while sentence-level results remain stable, demonstrating our model's competence, interpretability, and generalization performance.

Acknowledgement

This work was supported by the grants from the National Natural Science Foundation of China (61925201, 62132001, U21B2025, 62201014), Zhejiang Lab (NO. 2022NB0AB05), the National Key R&D Program of China (2021YFF0901502) and CAAI-Huawei MindSpore Open Fund.

References

Chen, S.; and Jiang, Y.-G. 2019. Semantic Proposal for Activity Localization in Videos via Sentence Query. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 8199–8206.

Ding, X.; Wang, N.; Zhang, S.; Cheng, D.; Li, X.; Huang, Z.; Tang, M.; and Gao, X. 2021. Support-Set Based Cross-Supervision for Video Grounding. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11553–11562.

Duan, X.; Huang, W.; Gan, C.; Wang, J.; Zhu, W.; and Huang, J. 2018. Weakly supervised dense event captioning in videos. *Advances in Neural Information Processing Systems*, 31.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, 5267–5275.

Gao, J.; Sun, X.; Xu, M.; Zhou, X.; and Ghanem, B. 2021. Relation-aware Video Reading Comprehension for Temporal Language Grounding. *ArXiv*, abs/2110.05717.

Gao, J.; and Xu, C. 2021. Fast Video Moment Retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1503–1512.

Huang, J.; Jin, H.; Gong, S.; and Liu, Y. 2022. Video Activity Localisation with Uncertainties in Temporal Boundary. In *European Conference on Computer Vision*, 724–740. Springer.

Huang, J.; Liu, Y.; Gong, S.; and Jin, H. 2021. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7199–7208.

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

Li, J.; Xie, J.; Qian, L.; Zhu, L.; Tang, S.; Wu, F.; Yang, Y.; Zhuang, Y.; and Wang, X. E. 2022a. Compositional Temporal Grounding with Structured Variational Cross-Graph Correspondence Learning. *CoRR*, abs/2203.13049.

Li, S.; Li, C.; Zheng, M.; and Liu, Y. 2022b. Phrase-level Prediction for Video Temporal Localization. In *International Conference on Multimedia Retrieval (ICMR)*, 360–368.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, B.; Yeung, S.; Chou, E.; Huang, D.-A.; Fei-Fei, L.; and Niebles, J. C. 2018. Temporal modular networks for retrieving complex compositional activities in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 552–568.

Liu, D.; Qu, X.; Di, X.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022a. Memory-Guided Semantic Learning Network for Temporal Sentence Grounding. *arXiv preprint arXiv:2201.00454*.

Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Cheng, Y.; Wei, W.; Xu, Z.; and Xie, Y. 2021. Context-aware Biaffine Localizing Network for Temporal Sentence Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11235–11244.

Liu, D.; Qu, X.; Zhou, P.; and Liu, Y. 2022b. Exploring Motion and Appearance Information for Temporal Sentence Grounding. In *AAAI*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mun, J.; Cho, M.; and Han, B. 2020. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10810–10819.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Otani, M.; Nakashima, Y.; Rahtu, E.; and Heikkilä, J. 2020. Uncovering Hidden Challenges in Query-Based Video Moment Retrieval. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press.

Qu, X.; Tang, P.; Zou, Z.; Cheng, Y.; Dong, J.; Zhou, P.; and Xu, Z. 2020. Fine-grained iterative attention network for temporal language localization in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4280–4288.

Rodriguez-Opazo, C.; Marrese-Taylor, E.; Fernando, B.; Li, H.; and Gould, S. 2021. DORi: Discovering Object Relationships for Moment Localization of a Natural Language Query in a Video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1079–1088.

Rohrbach, A.; Rohrbach, M.; Hu, R.; Darrell, T.; and Schiele, B. 2016. Grounding of Textual Phrases in Images by Reconstruction. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, 817–834. Springer.

Ryu, H.; Kang, S.; Kang, H.; and Yoo, C. D. 2021. Semantic Grouping Network for Video Captioning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 2514–2522. AAAI Press.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shi, P.; and Lin, J. J. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *ArXiv*, abs/1904.05255.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. K. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. *ArXiv*, abs/1604.01753.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, X.; Jiao, L.; Yang, S.; Zhang, X.; and Shang, F. 2013. Sparse coding and classifier ensemble based multi-instance learning for image categorization. *Signal Processing*, 93(1): 1–11.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, H.; Zha, Z.-J.; Li, L.; Liu, D.; and Luo, J. 2021a. Structured Multi-Level Interaction Network for Video Moment Localization via Language Query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7026–7035.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2021b. Negative Sample Matters: A Renaissance of Metric Learning for Temporal Grounding. *CoRR*, abs/2109.04872.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wu, J.; Li, G.; Liu, S.; and Lin, L. 2020. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12386–12393.

Xiao, S.; Chen, L.; Zhang, S.; Ji, W.; Shao, J.; Ye, L.; and Xiao, J. 2021. Boundary Proposal Network for Two-Stage Natural Language Video Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2986–2994.

Xu, H.; Venugopalan, S.; Ramanishka, V.; Rohrbach, M.; and Saenko, K. 2015. A multi-scale multiple instance video description network. *arXiv preprint arXiv:1505.05914*.

Xu, Y.; Zhu, J.-Y.; Chang, E. I.-C.; Lai, M.; and Tu, Z. 2014. Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Analysis*, 18(3): 591–604.

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

Yuan, Y.; Lan, X.; Chen, L.; Liu, W.; Wang, X.; and Zhu, W. 2021. A Closer Look at Temporal Sentence Grounding in Videos: Datasets and Metrics. *CoRR*, abs/2101.09028.

Yuan, Y.; Liang, X.; Wang, X.; Yeung, D.-Y.; and Gupta, A. 2017. Temporal Dynamic Graph LSTM for Action-driven Video Object Detection. In *ICCV*.

Zeng, R.; Xu, H.; bing Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense Regression Network for Video Grounding. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10284–10293.

Zhang, D.; Dai, X.; Wang, X. E.; fang Wang, Y.; and Davis, L. S. 2019. MAN: Moment Alignment Network for Natural Language Moment Retrieval via Iterative Graph Adjustment. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1247–1257.

Zhang, H.; Sun, A.; Jing, W.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021a. Parallel Attention Network with Sequence Matching for Video Grounding. In *FINDINGS*.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2020a. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*.

Zhang, J.; and Peng, Y. 2019. Hierarchical Vision-Language Alignment for Video Captioning. In Kompatsiaris, I.; Huet, B.; Mezaris, V.; Gurrin, C.; Cheng, W.; and Vrochidis, S., eds., *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part I*, volume 11295 of *Lecture Notes in Computer Science*, 42–54. Springer.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020b. Learning 2D Temporal Adjacent Networks forMoment Localization with Natural Language. In *AAAI*.

Zhang, Z.; Han, X.; Song, X.; Yan, Y.; and Nie, L. 2021b. Multi-Modal Interaction Graph Convolutional Network for Temporal Language Localization in Videos. *IEEE Transactions on Image Processing*, 30: 8265–8277.

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

Zheng, M.; Huang, Y.; Chen, Q.; and Liu, Y. 2022a. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1, 3.

Zheng, M.; Huang, Y.; Chen, Q.; Peng, Y.; and Liu, Y. 2022b. Weakly Supervised Temporal Sentence Grounding With Gaussian-Based Contrastive Proposal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15555–15564.

Zhou, H.; Zhang, C.; Luo, Y.; Chen, Y.; and Hu, C. 2021. Embracing Uncertainty: Decoupling and De-bias for Robust Temporal Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8445–8454.