STVGBert: A Visual-linguistic Transformer based Framework for Spatio-temporal Video Grounding

Rui Su
Platform & Content Group, Tencent
rayruisu@tencent.com

Qian Yu
College of Software, Beihang University
qianyu@buaa.edu.cn

Dong Xu*
School of Electrical and Information Engineering, The University of Sydney
dong.xu@sydney.edu.au

Abstract

Spatio-temporal video grounding (STVG) aims to localize a spatio-temporal tube of a target object in an untrimmed video based on a query sentence. In this work, we propose a one-stage visual-linguistic transformer based framework called STVGBert for the STVG task, which can simultaneously localize the target object in both spatial and temporal domains. Specifically, without resorting to pre-generated object proposals, our STVGBert directly takes a video and a query sentence as the input, and then produces the cross-modal features by using the newly introduced cross-modal feature learning module ST-ViLBert. Based on the cross-modal features, our method then generates bounding boxes and predicts the starting and ending frames to produce the predicted object tube. To the best of our knowledge, our STVGBert is the first one-stage method, which can handle the STVG task without relying on any pre-trained object detectors. Comprehensive experiments demonstrate our newly proposed framework outperforms the state-of-the-art multi-stage methods on two benchmark datasets VidSTG and HC-STVG.

1. Introduction

Vision and language play important roles for human to understand the world. In recent years, with remarkable progress of deep neural networks, various vision-language tasks (e.g., image captioning [15, 22], dense video caption [37, 36] and visual grounding [10, 35]) have attracted increasing attention from researchers.

Spatial-Temporal Video Grounding (STVG), which was introduced in the recent work [39], is a new and challenging vision-language task. Given an untrimmed video and a textual description of an object, the STVG task aims to produce a spatio-temporal tube (*i.e.*, a sequence of bounding boxes [21, 20]) for the target object described by the given text description. Different from the existing grounding tasks in images, both spatial and temporal localizations are required in the STVG task. Besides, how to effectively align visual and textual information through cross-modal feature learning in both spatial and temporal domains is also a key issue for accurately localizing the target object, especially in the challenging scenarios where different persons often perform similar actions within one scene.

Spatial localization in images/videos is a related visual grounding task, and spatial localization results have been improved in recent works [13, 35, 14, 27, 30, 31, 12, 29, 2]. In most existing works, a pre-trained object detector is often required to pre-generate object proposals. However, these approaches suffer from the following limitations: (1) The localization performance heavily relies on the quality of the pre-generated object proposals. (2) It is difficult for a pre-trained object detector to be well generalized to any new datasets with unseen classes. (3) Additional training data and computational cost are required for pre-training the object detectors. Although the recent works [34, 10, 16, 33] have attempted to remove the pre-generation process in the image grounding task, such efforts have not been made for the video grounding task.

For the STVG task, we are required to conduct localization both spatially and temporally. Intuitively, we can solve this task by using a two-stage approach, in which the temporal visual grounding methods [5, 1] are first used to localize the starting and ending frames of the target objects, and spatial localization is then performed by using the spatial visual grounding approaches [3, 17, 29] on the temporally trimmed videos. However, by handling the two sub-tasks separately, the pipeline becomes more complicated as each sub-task is handled by an independent network. Moreover,

*Dong Xu is the corresponding author.

we can also learn better representations by solving both spatial localization and temporal localization in an end-to-end optimized network. Therefore, it is desirable to propose a unified one-stage framework for the STVG task.

Motivated by the above observations, in this work, we propose a one-stage visual-linguistic transformer based framework STVGBert for the STVG task, which can directly generate spatio-temporal object tubes from the input videos and query descriptions without relying on any pre-trained object detectors. Specifically, our method first takes a pair of video clip and textual query as the input to produce the cross-modal features. The cross-modal features are then used to produce the bounding box for each frame as well as predict the starting and ending frames, which are then used to generate the spatio-temporal tubes for the target object.

Considering that promising results have been achieved by using transformers for various tasks [41, 40, 28], our STVGBert also builds on a visual-linguistic transformer. Specifically, the key component is a cross-modal feature learning module called ST-ViLBert. Different from the relevant work ViLBERT [15] which only encodes temporal information, our newly proposed ST-ViLBert also preserves spatial information in the visual input feature. As a result, our STVGBert can effectively learn the cross-modal representation based on both spatial and temporal visual information and produce the spatio-temporal object tubes *without* requiring any pre-trained object detectors. We evaluate our proposed framework STVGBert on two benchmark datasets, VidSTG [39] and HC-STVG [25], and the experiments demonstrate that our framework outperforms all state-of-the-art methods.

Our contributions can be summarized as follows:

(1) We propose a new one-stage visual-linguistic transformer based framework STVGBert for the spatio-temporal video grounding task. To the best of our knowledge, this is the first end-to-end optimized STVG framework that does not require any pre-trained object detectors.

(2) We introduce a new cross-modal feature learning module, ST-ViLBert, to model spatio-temporal information and align cross-modal representation at the same time.

(3) Comprehensive experiments conducted on two benchmark datasets, VidSTG and HC-STVG, demonstrate the effectiveness of our framework for the STVG task. Our one-stage scheme outperforms all multi-stage state-of-the-art methods by a significant margin.

2. Related Work

2.1. Vision-language Modelling

Transformer-based neural network has been widely explored for various vision-language tasks [24, 9, 15, 22], such as visual question answering, image captioning, and image-text retrieval. For example, the work in [24] proposed to use two single-modal transformers together with a cross-modal transformer to learn the cross-modal representations for the visual question answering task. In [9], Li *et al*., pretrained a single cross-stream transformer for the sentence-image alignment task and used the pretrained model to tackle the image-text retrieval task. In the works ViLBERT [15] and VL-BERT [22], the authors trained general transformer-based neural networks based on large visual-linguistic datasets, which can benefit several downstream tasks. Apart from these works designed for the image-based visual-linguistic tasks, Sun et al. [23] proposed VideoBERT for the video captioning task by modeling temporal variations across multiple video frames. However, this work does not model spatial information within each frame, so it cannot be applied for the spatio-temporal video grounding task discussed in this work.

The aforementioned transformer-based neural networks take the features extracted from either Region of Interests (RoIs) in images or video frames as the input features, so spatial information in the feature space cannot be preserved when transforming them to visual tokens. In this work, we propose an improved version of ViLBERT [15] to better model spatio-temporal information in videos and learn better cross-modal representations.

2.2. Visual Grounding in Images/Videos

Visual grounding in images/videos aims to localize the object of interest in an image/video based on a query sentence. In most existing methods [13, 35, 14, 27, 30, 31, 12, 29, 2, 39], a pre-trained object detector is often required to pre-generate object proposals. The proposal that best matches the given input description is then selected as the final result. The work MattNet [35] used a modular network to explore the attributes and object relationships. For the visual grounding tasks in images, some recent works [34, 10, 16, 33] proposed new one-stage grounding frameworks without using the pre-trained object detectors. For example, Liao et al. [10] used the anchor-free object detection method [42] to localize the target objects based on the cross-modal representations. Yang *et al*. [33] used sub-queries to generate text-conditional visual features to improve the performance of one-stage grounding method.

For the video grounding task, Zhang et al. [39] proposed a new method (referred to as STGRN) that does not rely on the pre-generated tube proposals. Unfortunately, this work [39] still requires a pre-trained object detector to first generate object proposals since the output bounding boxes are retrieved from these candidate bounding boxes. Similar as in our proposed framework, the recent work STGVT [25] also adopted a visual-linguistic transformer to learn cross-modal representations for the spatio-temporal

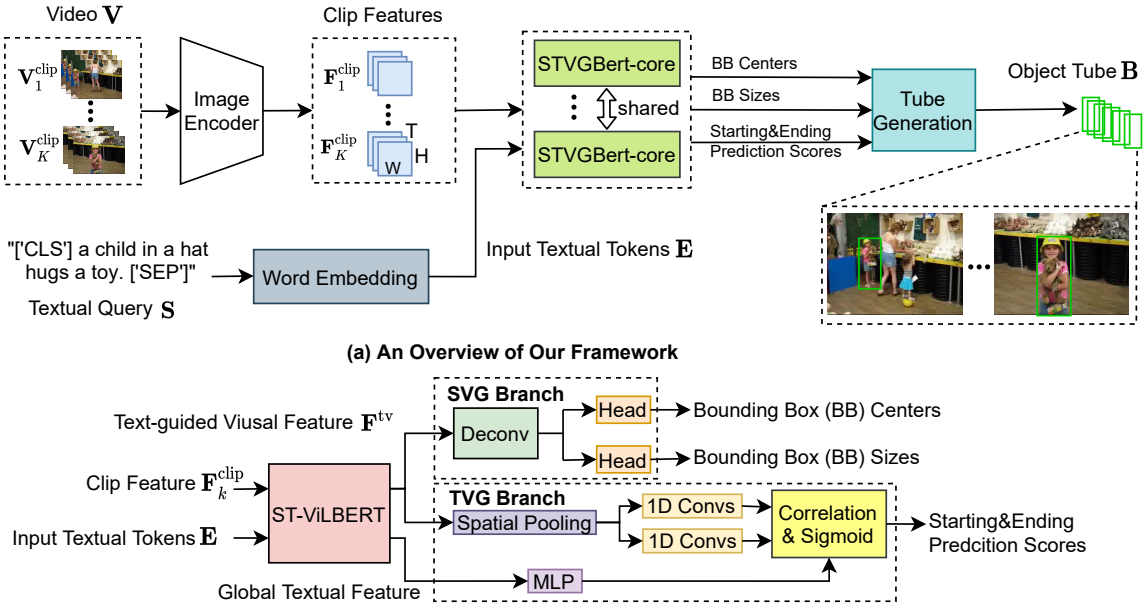


Figure 1. (a) The overview of our spatio-temporal video grounding framework STVGBert. Our one-stage framework consists of the STVGBert-core module and the Tube Generation module, which takes video and textual query pairs as the input to generate spatio-temporal object tubes containing the object of interest. In (b), the STVGBert-core module consists of two branches, the Spatial Visual Grounding (SVG) Branch and the Temporal Visual Grounding (TVG) Branch, which simultaneously generates bounding boxes and predicts the probability of each frame being the starting/ending frame.

video grounding task. But this work [25] also needs to first generate the tube proposals as in most existing methods [2, 29]. Additionally, in the works [18, 8, 38, 32], the pre-trained object detectors are also required to generate object proposals for object relationship modelling.

In contrast to these works [39, 25, 18, 8, 38, 32], we introduce a new framework with a newly proposed cross-modal feature learning module to generate object tubes without requiring any pre-trained object detectors.

3. Methodology

In this section, we briefly introduce the overall framework of our proposed method in Section 3.1, and then we present how to encode the visual features and the textual features from the input videos and the textual query descriptions (Section 3.2), respectively, as well as introduce our newly proposed multi-modal representation learning module ST-ViLBERT (Section 3.3). We then describe the spatial and temporal localization process for generating spatio-temporal object tubes in Section 3.4. Finally, the training details are introduced in Section 3.5

3.1. Overview

We denote an untrimmed video $\mathbf{V}$ with $K * T$ frames as a set of non-overlapping video clips, namely we have $\mathbf{V} = \{\mathbf{V}_k^{\text{clip}}\}_{k=1}^{K}$, where $\mathbf{V}_k^{\text{clip}}$ indicates the k-th video clip consisting of T frames, and K is the total number of video clips in the untrimmed video. We also denote a textual description as $\mathbf{S} = \{s_n\}_{n=1}^{N}$, where s_n indicates the n-th word in the description $\mathbf{S}$, and N is the total number of words. The STVG task aims to output the spatio-temporal tube $\mathbf{B} = \{\mathbf{b}_t\}_{t=t_s}^{t_e}$ containing the object of interest (*i.e.*, the target object) between the t_s-th and the t_e-th frames, where $\mathbf{b}_t$ is a 4-d vector indicating the top-left and bottom-right spatial coordinates of the target bounding box in the t-th frame. t_s and t_e represent the temporal starting and ending frames of the object tube $\mathbf{B}$, respectively.

In this task, it is required to perform both spatial localization and temporal localization based on the query sentence. We propose a unified STVG framework STVGBert to simultaneously localize the target of interest in both spatial and temporal domains. Unlike the previous methods [29, 2], which first output a set of tube proposals by linking the pre-detected object bounding boxes, our STVGBert does not require any pre-trained object detectors. As shown in Fig. 1, our STVGBert extracts the visual features and the textual embedding from the video frames and the textual query, respectively, and then produces the text-guided visual feature by using our newly introduced cross-modal feature learning module, and finally spatially and temporally localizes the object of interest to generate an object tube, including

the bounding boxes $\mathbf{b}_t$ for the target object in each frame and the indexes of the starting and ending frames. In the following sections, we will describe each step in details.

3.2. Visual Feature and Textual Feature Encoding

We first use ResNet-101 [6] as the image encoder to extract the visual feature. The output from the 4th residual block is reshaped to the size of $HW \times C$ with H, W and C indicating the height, the width, and the number of channels of the feature map, respectively, which is then used as the extracted visual feature. For the k-th video clip, we stack the extracted visual features from each frame in this video clip to construct the clip feature $\mathbf{F}_k^{\text{clip}} \in R^{T \times HW \times C}$, which is then fed into our cross-modal feature learning module to produce the multi-modal visual feature.

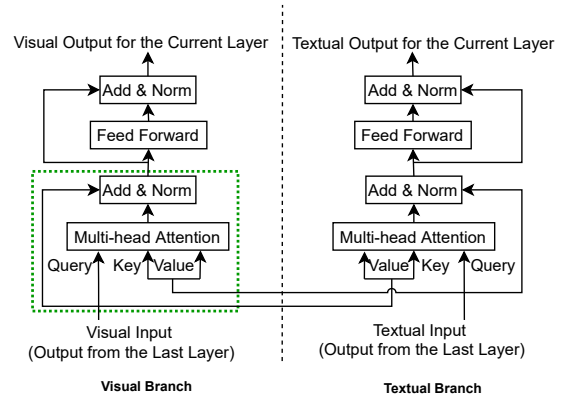
For the textual descriptions, we use a word embedding module to map each word in the description as a word vector, and each word vector is considered as one textual input token. Additionally, we add two special tokens, ['CLS'] and ['SEP'], before and after the textual input tokens of the description to construct the complete textual input tokens $\mathbf{E} = \{e_n\}_{n=1}^{N+2}$, where e_n is the n-th textual input token.

3.3. Multi-modal Feature Learning

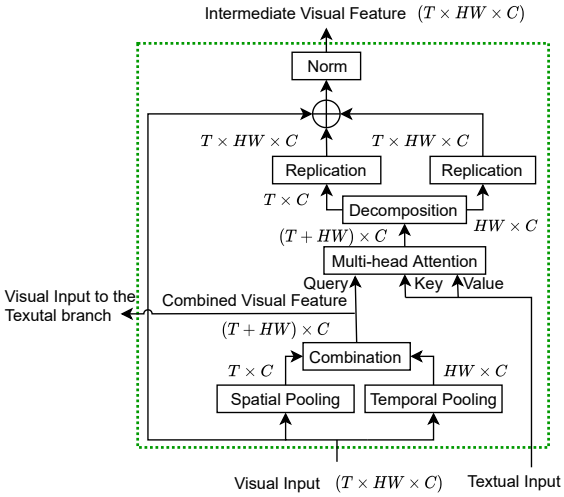
Given the visual input feature $\mathbf{F}_k^{\text{clip}}$ and the textual input tokens $\mathbf{E}$, we develop a new cross-modality modeling module called ST-ViLBERT, to learn the visual-linguistic representation. Following the structure of ViLBERT [15], our ST-ViLBERT module consists of a visual branch and a textual branch where both branches adopt the multi-layer transformer encoder [26] structure. As shown in Fig. 2(a), the visual branch interacts with the textual branch via a set of co-attention layers, which exchanges information between the key-value pairs to generate the text-guided visual feature (or vice versa). Please refer to the work in [15] for further details.

The work ViLBERT [15] takes the visual features extracted from all pre-generated proposals within an image as the visual input to learn the visual-linguistic representation (see Fig. 2(a)). However, since these visual features are spatially pooled by using the average pooling operation, spatial information in the visual input feature space will be lost. While such information is important for predicting bounding boxes, this is not an issue for ViLBERT as it assumes the bounding boxes are already generated by using the pre-trained object detectors.

Our ST-ViLBERT module is designed for spatial localization without requiring any pre-generated bounding boxes, where the key is to preserve spatial information when performing cross-modal feature learning. Specifically, we introduce a Spatio-temporal Combination and Decomposition (STCD) module to replace the Multi-head Attention and Add & Norm modules for the visual branch in ViLBERT. As shown in Fig. 2(b), our STCD module respectively applies the *spatial* and *temporal* average pooling operations on the input visual feature (*i.e.*, the visual output from the last layer) to produce the initial temporal feature with the size of $T \times C$ and the initial spatial feature with the size of $HW \times C$, which are then concatenated to construct the combined visual feature with the size of $(T + HW) \times C$. We then pass the combined visual feature to the textual branch, which is used as key and value in the multi-head attention block of the textual branch. Additionally, the combined visual feature is also fed to the multi-head attention block of the visual branch together with the textual input (*i.e.*, the textual output from the last layer) to



(a) One Co-attention Layer in ViLBERT



(b) Spatio-temporal Combination and Decomposition Module

Figure 2. (a) The overview of one co-attention layer in ViLBERT [15]. The co-attention layer, consisting of a visual branch and a textual branch, generates the visual-linguistic representations by exchanging the key-value pairs for the multi-head attention blocks. (b) The structure of our Spatio-temporal Combination and Decomposition (STCD) module, which replaces the "Multi-head attention" and "Add & Norm" blocks in the visual branch of ViLBERT (marked in the green dotted box in (a)).

generate the initial text-guided visual feature with the size of $(T+HW)\times C$, which is then decomposed into the text-guided *temporal* feature with the size of $T\times C$ and the text-guided *spatial* feature with the size of $HW\times C$. These two features are then respectively replicated HW and T times to match the dimension of the input visual feature. The replicated features and the input visual feature are added up and normalized to generate the intermediate visual feature with the size of $T\times HW\times C$. The remaining part (including the textual branch) are the same as that in ViLBERT [15]. In our ST-ViLBERT, we take the output from the visual and textual branches in the last co-attention layer as the **t**ext-guided **v**isual feature $\mathbf{F}^{\text{tv}}\in R^{T\times HW\times C}$ and the visual-guided textual feature, respectively.

3.4. Spatial and Temporal Localization

In our framework, the cross-modal features from our ST-ViLBert module, including the text-guided visual feature $\mathbf{F}^{\text{tv}}$ and the visual-guided textual feature, are fed into two branches, SVG branch and TVG branch, for **s**patial **v**isual **g**rounding and **t**emporal **v**isual **g**rounding, respectively.

Spatial Localization As shown in Fig. 1(b), the SVG branch takes the text-guided visual feature $\mathbf{F}^{\text{tv}}$ to predict the bounding box $\mathbf{b}_t$ at each frame. We first reshape $\mathbf{F}^{\text{tv}}$ to the size of $T\times H\times W\times C$. Taking the feature from each individual frame (with the size of $H\times W\times C$) as the input of three deconvolution layers, we then upsample the spatial resolution by a factor of 8. Similar as in CenterNet [42], the upsampled feature is used as the input of two parallel detection heads, with each head consisting of a 3×3 convolution layer for feature extraction and a 1×1 convolution layer for dimension reduction. The first detection head outputs a heatmap $\mathbf{A}\in R^{8H\times 8W}$, where the value at each spatial location indicates the probability of this position being the bounding box center of the target object, and the second head regresses the size (*i.e.*, the height and width) of the target bounding box at each position. In the heatmap, we select the spatial location with the highest probability as the predicted bounding box center, and use the corresponding predicted height and width at this selected position to calculate the top-left and the bottom-right coordinates of the predicted bounding box.

Temporal Localization In addition to bounding box prediction by using the SVG branch, our TVG branch predicts the positions of the starting frame and the ending frame based on both text-guided visual feature and the visual-guided textual feature. As shown in Fig. 1(b), we apply spatial average pooling on the text-guided visual feature $\mathbf{F}^{\text{tv}}$ to produce the **g**lobal **t**ext-guided **v**isual feature $\mathbf{F}^{\text{gtv}}\in R^{T\times C}$ for each input video clip with T frames. The global text-guided visual feature is then fed into two parallel temporal convolution blocks to produce the starting and ending visual features, where each temporal convolution block consists of three 1D convolutional layers with the kernel size of 3. The size of both starting and ending visual features is $T\times C$. We also feed the global textual feature (*i.e.*, the visual-guided textual feature corresponding to the token ['CLS']) into a MLP layer to produce the intermediate textual feature in the C-dimension common feature space. By using the correlation operation, we can then compute the initial starting (*resp.*, ending) prediction score for each frame between the corresponding starting (*resp.*, ending) visual feature and the intermediate textual feature in the common feature space. After applying the Sigmoid activation function, we produce the final starting (*resp.*, ending) prediction score p^s (*resp.*, p^e) for each frame in one video clip, which indicates the probability of each frame being the starting (*resp.*, the ending) frame for the target tube.

Tube Generation After producing bounding boxes and the starting and ending prediction scores for all frames from all video clips, we then combine them across the temporal domain to construct an initial object tube, namely, a sequence of starting scores and a sequence of ending scores for all $K*T$ frames in the whole video. After that, the temporal positions of the starting and ending frames (t_s, t_e) can be determined by selecting the frames with the largest starting score and the largest ending score, respectively. Moreover, the bounding boxes from the frames before t_s and after t_e are removed. Finally, the temporal boundaries (t_s, t_e) and the predicted bounding boxes $\mathbf{b}_t$ form the tube prediction result $\mathbf{B}=\{\mathbf{b}_t\}_{t=t_s}^{t_e}$.

3.5. Loss Functions

We use a combination of three focal losses and a L1 loss to train our STVGBert. At the training stage, we randomly sample a set of video clips with T consecutive frames, and then we select the video clips as the training samples which contain at least one frame having a ground-truth bounding box. Next, we take one video clip with T consecutive frames as an example for better explanation. Specifically, for the i-th frame in each training video clip, we denote the center, width, and height of the ground-truth bounding box as $(\hat{x}_i, \hat{y}_i)$, $\hat{w}_i$, $\hat{h}_i$, respectively. Additionally, we also denote the indexes of the ground-truth starting and ending frames as $\hat{t}_s$ and $\hat{t}_e$. Based on the ground-truth bounding box, we follow [42] to generate a center heatmap $\hat{\mathbf{A}}$ for the i-th frame by using the Gaussian kernel $\hat{a}_i^{x,y}=\exp(-\frac{(x-\hat{x}_i)^2+(y-\hat{y}_i)^2}{2\sigma_i^2})$, where $\hat{a}_i^{x,y}$ is the value of $\hat{\mathbf{A}}_i\in R^{8H\times 8W}$ at the spatial location (x, y), and σ_i is the bandwidth parameter, which is adaptively determined based on the object size [7]. Similarly, we can generate two 1D temporal heatmaps $\hat{\mathbf{p}}_s$ and $\hat{\mathbf{p}}_e\in R^T$ for the starting and ending positions, respectively. For training our STVGBert, the objective function L_{total} for each training video clip is

defined as follows:

$$
\begin{aligned}
L_{\text{total}} =& \lambda_1 \frac{1}{T} \sum_{i=1}^{T} L_{\text{size}}(w_i^{\hat{x}_i,\hat{y}_i}, h_i^{\hat{x}_i,\hat{y}_i}, \hat{w}_i, \hat{h}_i) \\
&+ \lambda_2 L_s(\mathbf{p}_s, \hat{\mathbf{p}}_s) + \lambda_3 L_e(\mathbf{p}_e, \hat{\mathbf{p}}_e) \\
&+ \lambda_4 \frac{1}{T} \sum_{i=1}^{T} L_c(\mathbf{A}_i, \hat{\mathbf{A}}_i),
\end{aligned} \tag{1}
$$

where $\mathbf{A}_i \in R^{8H \times 8W}$ is the predicted heatmap, $\mathbf{p}_s$ and $\mathbf{p}_e$ are the predicted starting and ending score sequences, $w_i^{\hat{x}_i,\hat{y}_i}$ and $h_i^{\hat{x}_i,\hat{y}_i}$ are the predicted width and height of a bounding box centered at the location $(\hat{x}_i,\hat{y}_i)$ for the i-th frame. L_c, L_s and L_e are the focal loss [11] for predicting the bounding box center and the temporal positions of starting and ending frames, respectively; L_{size} is a L1 loss for regressing the size of the bounding box. We empirically set the loss weights as $\lambda_1 = 0.1$ and $\lambda_2 = \lambda_3 = \lambda_4 = 1$.

4. Experiment

4.1. Experiment Setup

Datasets We evaluate our proposed framework on the VidSTG [39] dataset and the HC-STVG [25] dataset.

-VidSTG. This dataset consists of 99,943 sentence descriptions with 44,808 declarative sentences and 55,135 interrogative sentences describing 79 types of objects appearing in the untrimmed videos. Following [39], we divide the sentence descriptions into the training set, the validation set, and the testing set with 36,202 (*resp.*, 44,482), 3,996 (*resp.*, 4,960), and 4,610 (*resp.*, 5,693) declarative (*resp.*, interrogative) sentences. The described objects in the untrimmed videos are annotated with the spatio-temporal tubes.

-HC-STVG. This dataset consists of 5,660 video-description pairs and all videos are untrimmed. This dataset is human-centric since all videos are captured in multi-person scenes and the descriptions contain rich expressions related to human attributes and actions. This dataset is divided into the training set and the testing set with 4,500 and 1,160 video-sentence pairs, respectively. All target persons are annotated with spatio-temporal tubes.

Implementation details We use the ResNet-101 [6] network pretrained on ImageNet [4] as our image encoder to extract the visual features from the RGB frames in the input videos. Our whole framework, including the ResNet-101, is end-to-end optimized. For the ST-ViLBERT module in our STVGBert, we employ the ViLBERT model pretained on the Conceptual Caption dataset [19] for initialization. Following [39], we sample the input videos at the frame rate of 5fps. The batch size and the initial learning rate are set to be 6 and 0.00001, respectively. After training our model for 50 epochs, we decrease the learning rate by a factor of 10 and then train our model for another 10 epochs. The temporal length of each video clip T is set as 20. Our method is implemented by using PyTorch on the machine with a single V100 GPU.

Evaluation metrics We follow [39] to use m_vIoU and vIoU@R as our evaluation criteria. The vIoU is calculated as vIoU $= \frac{1}{|S_U|} \sum_{t \in S_I} r_t$, where r_t is the IoU between the detected bounding box and the ground-truth bounding box at frame t, the set S_I contains the intersected frames between the detected tubes and the ground-truth tubes (*i.e.*, the intersection set between the frames from both tubes), and S_U is the union of two sets of frames from the detected tubes and the ground-truth tubes. The m_vIoU score is defined as the average vIoU score over all testing videos, and vIoU@R refers to the ratio of the testing videos with vIoU $>$ R over all the testing videos.

Baseline Methods We compare our method with the existing STVG methods proposed in [39] and [25].

-STGRN [39] is the state-of-the-art method on the VidSTC dataset. Although this method does not need to pre-generate the tube proposals, it still requires the pre-trained detector to produce the bounding box proposals in each frame, which are then used to build the spatial relation graph and the temporal dynamic graph. And the final bounding boxes are *selected* from these proposals. Therefore, its performance is highly dependent on the quality of the pre-generated proposals.

-STGVT [25] is the state-of-the-art method on the HC-STVG dataset. Similar to our proposed ST-ViLBert, it also adopts a visual-linguistic transformer module to learn the cross-modal representations. However, STGVT relies on a pre-trained object detector and a linking algorithm to generate the tube proposals, while our framework does not require any pre-generated tubes.

In addition, given the recent research progress in spatial visual grounding and temporal visual grounding, six baseline methods can be constructed by combining the methods from these two tasks. Specifically, we follow [39] and [25] to first respectively use the temporal visual grounding methods TALL [5] and L-Net [1] to predict the temporal positions of the starting and ending frames of the target objects, which are then used to temporally trim the input videos. Based on the temporally trimmed videos generated by either TALL or L-Net, a frame-level visual grounding method GroundeR [17] and two tube-level video grounding approaches STPR [29] and WSSTG [3] are employed to generate the bounding boxes for the target objects, which are then used to produce the final spatio-temporal object tubes. These six baseline methods are referred to as **GroundeR + TALL**, **STPR + TALL**, **WSSTG + TALL**, **GroundeR + L-Net**, **STPR + L-Net** and **WSSTG + L-Net**, respectively.

Table 1. Results of different methods on the VidSTG dataset. "*" indicates the results are quoted from the work in [39]

Methods	Declarative Sentence			Interrogative Sentence		
	m_vIoU(%)	vIoU@0.3(%)	vIoU@0.5(%)	m_vIoU(%)	vIoU@0.3(%)	vIoU@0.5(%)
GroundeR [17] + TALL [5]*	9.78	11.04	4.09	9.32	11.39	3.24
STPR [29] + TALL [5]*	10.40	12.38	4.27	9.98	11.74	4.36
WSSTG [3] + TALL [5]*	11.36	14.63	5.91	10.65	13.90	5.32
GroundeR [17] + L-Net [1]*	11.89	15.32	5.45	11.05	14.28	5.11
STPR [29] + L-Net [1]*	12.93	16.27	5.68	11.94	14.73	5.27
WSSTG [3] + L-Net [1]*	14.45	18.00	7.89	13.36	17.39	7.06
STGRN [39]*	19.75	25.77	14.60	18.32	21.10	12.83
STVGBert (Ours)	**23.97**	**30.91**	**18.39**	**22.51**	**25.97**	**15.95**

Table 2. Results of different methods on the HC-STVG dataset. "*" indicates the results are quoted from the work in [25]

Methods	m_vIoU	vIoU@0.3	vIoU@0.5
STGVT [25]*	18.15	26.81	9.48
STVGBert (Ours)	**20.42**	**29.37**	**11.31**

4.2. Comparison with the State-of-the-art Methods

We compare our proposed framework with the state-of-the-art methods on both VidSTG and HC-STVG datasets. The results on these two datasets are shown in Table 1 and Table 2. From the results, we have the following observations. 1) Our proposed method outperforms the state-of-the-art methods by a large margin on both datasets in terms of all evaluation metrics. 2) On the VidSTG dataset, for the first six baseline methods in Table 1, we first perform temporal visual grounding by using either TALL or L-Net, and then perform spatial visual grounding to produce the final results. In contrast, our method can simultaneously generate bounding boxes and temporal boundaries to form spatio-temporal object tubes, and our method significantly outperforms these two-stage baseline approches, which demonstrates the effectiveness of our proposed one-stage approach STVGBert. 3) In Table 2, both STGVT and our STVGBert apply the visual-linguistic transformer to learn cross-modal representations, but our method outperforms STGVT by a noticeable margin. Additionally, STGVT requires the pre-trained object detectors to generate a set of proposals, while our method STVGBert, which includes the improved transformer module ST-ViLBert, can directly deal with the input video clips.

4.3. Ablation Study

In this section, we take the VidSTG dataset as an example to conduct ablation study and investigate the contributions of different components in our proposed framework.

Effectiveness of the one-stage framework As introduced in Sec. 3.4, with two carefully-designed branches, our method can handle spatial and temporal visual grounding simultaneously. In this section, we first conduct the experiments to demonstrate the performance of each single branch. And then we compare our one-stage scheme with its two-stage counterpart to demonstrate the effectiveness of the proposed one-stage framework. Specifically, we introduce the following variants of our STVGBert method for ablation study. (1) **STVGBert w/o TVG branch**: in this alternative method, we remove the TVG branch from our STVGBert framework, which only produces the spatial visual grounding results; (2) **STVGBert w/o SVG branch**: in this variant, we remove the SVG branch from our STVGBert framework, which only produces the temporal visual grounding results; (3) **STVGBert-2Stage**: in this alternative method, we first use *STVGBert w/o SVG branch* to generate a temporally trimmed input video, and then use *STVGBert w/o TVG branch* to generate the bounding boxes based on the temporally trimmed video and finally produce the spatio-temporal object tubes.

To further investigate how one-stage architecture improves the performance, we conduct the experiments under two different settings. 1) **w/o Tem. GT (Default)**: In this setting, we compare the spatio-temporal visual grounding results of the three baseline methods *STVGBert w/o TVG branch*, *STVGBert w/o SVG branch*, and *STVGBert-2Stage*, as well as our STVGBert without using the ground-truth temporal annotation (*i.e.*, the ground-truth starting and ending frames are not available). For *STVGBert w/o TVG branch*, since it only produces the spatial visual grounding results, the generated bounding boxes from the whole untrimmed input videos are used as the spatio-temporal video grounding results. For *STVGBert w/o SVG branch*, it only predicts the temporal boundaries (*i.e.*, the starting and ending frames), so we use the whole image of each frame between the predicted starting and ending frames as the spatio-temporal video grounding results. 2) **w/ Tem. GT**: In this alternative setting, the ground-truth temporal annotation is available and we compare the spatio-temporal visual grounding results generated by *STVGBert w/o TVG branch+Tem. GT* and *STVGBert+Tem. GT*. Note that for *STVGBert+Tem. GT*, we simply ignore the predicted temporal grounding results produced by our framework STVGBert, and instead use the ground-truth temporal annotation. All the experimental results are reported in Table 3.

In Table 3, under the default setting, we have the fol-

Table 3. Results of our method and its variants on the VidSTG dataset. Note that for *STVGBert+Tem. GT*, we simply ignore the temporal grounding results produced by the corresponding method, and use the ground-truth information instead.

Setting	Methods	Declarative Sentence Grounding			Interrogative Sentence Grounding		
		m_vIoU	vIoU@0.3	vIoU@0.5	m_vIoU	vIoU@0.3	vIoU@0.5
w/o Tem. GT (Default)	STVGBert w/o TVG branch	18.94	24.52	13.85	17.54	20.28	11.79
	STVGBert w/o SVG branch	7.21	4.59	1.21	6.98	4.21	1.05
	STVGBert-2Stage	22.51	29.74	17.58	21.05	24.99	15.01
	STVGBert-Simple	21.27	28.52	16.74	19.87	23.59	13.95
	STVGBert	**23.97**	**30.91**	**18.39**	**22.51**	**25.97**	**15.95**
w/ Tem. GT	STVGBert w/o TVG branch + Tem. GT	45.69	63.92	50.75	42.95	56.29	46.78
	STVGBert-Simple + Tem. GT	41.75	59.94	44.51	39.12	51.54	40.21
	STVGBert + Tem. GT	**47.25**	**66.16**	**53.09**	**44.12**	**59.77**	**49.28**

lowing observations. First, the results of *STVGBert w/o SVG branch* are very poor as it dose not output the bounding boxes. Second, the spatio-temporal video grounding results of *STVGBert w/o TVG branch* and *STVGBert w/o SVG branch* can be further improved by combining these two methods as an alternative two-stage approach (*i.e.*, *STVBert-2Stage*). Third, we also observe that our one-stage STVGBert framework outperforms this alternative method *STVBert-2Stage* in terms of all evaluation metrics, which demonstrates the effectiveness of the proposed one-stage framework. In Table 3, under the alternative setting w/ Tem. GT, both *STVGBert w/o TVG branch+Tem. GT* and *STVGBert+Tem. GT* use the ground-truth temporal annotation (*i.e.*, Tem. GT) as the temporal visual grounding results. In this case, the only difference between these two methods is from the spatial grounding results. Therefore, the improvement of *STVGBert+Tem. GT* over *STVGBert w/o TVG branch+Tem. GT* indicates that our newly proposed framework STVGBert achieves better spatial visual grounding performance, *i.e.*, our method produces more accurate bounding boxes. Finally, all these results also indicate that it is beneficial to jointly optimize the objective function related to both spatial and temporal visual grounding in a one-stage framework as our scheme benefits from multi-task learning.

Effectiveness of the ST-ViLBERT Module In our proposed STVGBert framework, the key component is our ST-ViLBERT module. Different from ViLBERT [15], which does not model spatial information for the input visual features, the ST-ViLBERT module in our STVGBert framework can preserve both spatial and temporal information from the input visual features, so that our method can learn a better spatio-temporal representation. To evaluate the effectiveness of our ST-ViLBERT, we introduce an alternative method by replacing our ST-ViLBERT in our STVGBert framework with the existing scheme ViLBERT, which is referred to as **STVGBert-Simple**. Specifically, in *STVGBert-Simple*, for each input video clip, the spatial average pooling operation is applied on top of the clip features to produce the feature vector for each frame of the video clip. These feature vectors are then directly fed into ViLBERT as the visual input together with the textual input to generate the cross-modal representations. We then take the generated cross-modal representation as the residual feature and add it to the original clip features to produce the input features to the SVG branch. The experimental results are reported in Table 3 (*i.e.*, the Default setting). In addition to that, we also conduct the experiments for *STVGBert-Simple* under the alternative setting w/ Tem. GT, where the temporal grounding results generated by *STVGBert-Simple* are replaced by the ground-truth temporal annotation, and the method is referred to as **STVGBert-Simple+Tem. GT**.

As shown in Table 3, our *STVGBert* outperforms *STVGBert-Simple* under the default setting. Moreover, when the ground-truth temporal annotation is available, *STVGBert+Tem. GT* also performs much better than *STVGBert-Simple+Tem. GT*, i.e., the gains ranges from 5.0% to 9.0%, which indicates the effectiveness of our newly proposed ST-ViLBert module by additionally preserving spatial information.

5. Conclusion

In this work, we have proposed a new one-stage spatio-temporal video grounding framework STVGBert based on a visual-linguistic transformer to produce spatio-temporal object tubes for a given query sentence, which consists of a spatial visual grounding branch and a temporal visual grounding branch. Besides, we have introduced a new cross-modal feature learning method ST-ViLBERT within our STVGBert framework. With ST-ViLBert, our STVGBert framework can produce spatio-temporal object tubes without requiring any pre-trained object detector. Comprehensive experiments on two benchmark datasets VidSTG and HC-STVG demonstrate the effectiveness of our newly proposed framework for spatio-temporal video grounding.

Acknowledgement: This work is supported by the National Key Research and Development Project of China (No. 2018AAA0101900) and the National Natural Science Foundation of China (No. 62002012).

References

[1] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8175–8182, 2019. 1, 6, 7

[2] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1884–1894, 01 2019. 1, 2, 3

[3] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1884–1894, 2019. 1, 6, 7

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[5] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 1, 6, 7

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 6

[7] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018. 5

[8] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, 2020. 3

[9] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020. 2

[10] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020. 1, 2

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[12] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4673–4682, 2019. 1, 2

[13] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864, 2017. 1, 2

[14] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1950–1959, 2019. 1, 2

[15] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 1, 2, 4, 5, 8

[16] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10034–10043, 2020. 1, 2

[17] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016. 1, 6, 7

[18] Arka Sadhu, Kan Chen, and Ram Nevatia. Video object grounding using semantic roles in language description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10417–10427, 2020. 3

[19] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 6

[20] Rui Su, Wanli Ouyang, Luping Zhou, and Dong Xu. Improving action localization by progressive cross-stream cooperation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[21] Rui Su, Dong Xu, Luping Zhou, and Wanli Ouyang. Progressive cross-stream cooperation in spatial and temporal domain for action localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[22] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. 1, 2

[23] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019. 2

[24] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 2

[25] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 2, 3, 6, 7

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia

Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[27] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1960–1968, 2019. 1, 2

[28] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. StyleFormer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[29] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1453–1462, 2017. 1, 2, 3, 6, 7

[30] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4145–4154, 2019. 1, 2

[31] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4644–4653, 2019. 1, 2

[32] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1939–1947, 2020. 3

[33] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[34] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4683–4693, 2019. 1, 2

[35] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 1, 2

[36] Zhiwang Zhang, Dong Xu, Wanli Ouyang, and Chuanqi Tan. Show, tell and summarize: Dense video captioning using visual cue aided sentence summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):3130–3139, 2020. 1

[37] Zhiwang Zhang, Dong Xu, Wanli Ouyang, and Luping Zhou. Dense video captioning using graph-based sentence summarization. *IEEE Transactions on Multimedia*, 23:1799–1810, 2021. 1

[38] Zhu Zhang, Zhou Zhao, Zhijie Lin, Baoxing Huai, and Jing Yuan. Object-aware multi-branch relation networks for spatio-temporal video grounding. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 1069–1075, 7 2020. 3

[39] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020. 1, 2, 3, 6, 7

[40] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[41] Lichen Zhao, Jinyang Guo, Dong Xu, and Lu Sheng. Transformer3D-Det: Improving 3D object detection by vote refinement. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 2

[42] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 2, 5