Mono3DVG: 3D Visual Grounding in Monocular Images

Yang Zhan[1], **Yuan Yuan**[1*], **Zhitong Xiong**[2*]

[1]School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China
[2]Technical University of Munich (TUM), Munich, Germany
{zhanyangnwpu, y.yuan1.ieee, xiongzhitong}@gmail.com

Abstract

We introduce a novel task of 3D visual grounding in monocular RGB images using language descriptions with both appearance and geometry information. Specifically, we build a large-scale dataset, Mono3DRefer, which contains 3D object targets with their corresponding geometric text descriptions, generated by ChatGPT and refined manually. To foster this task, we propose Mono3DVG-TR, an end-to-end transformer-based network, which takes advantage of both the appearance and geometry information in text embeddings for multi-modal learning and 3D object localization. Depth predictor is designed to explicitly learn geometry features. The dual text-guided adapter is proposed to refine multiscale visual and geometry features of the referred object. Based on depth-text-visual stacking attention, the decoder fuses object-level geometric cues and visual appearance into a learnable query. Comprehensive benchmarks and some insightful analyses are provided for Mono3DVG. Extensive comparisons and ablation studies show that our method significantly outperforms all baselines. The dataset and code will be released.

Introduction

For intelligent systems and robots, understanding objects based on language expressions in real 3D scenes is an important capability for human-machine interaction. Visual grounding (Deng et al. 2021; Yang et al. 2022; Zhan, Xiong, and Yuan 2023) has made significant progress in 2D scenes, but these approaches cannot obtain the true 3D extent of the objects. Therefore, recent researches (Chen, Chang, and Nießner 2020; Achlioptas et al. 2020) utilize RGB-D sensors for 3D scanning and build indoor point cloud scenes for 3D visual grounding. The latest work (Lin et al. 2023) focuses on outdoor service robots and utilizes LiDAR and an industrial camera to capture point clouds and RGB images as multimodal visual inputs. However, the practical application of these works is limited due to the expensive cost and device limitations of RGB-D scans and LiDAR scans.

Monocular 3D object detection (Huang et al. 2022a; Brazil et al. 2023) can obtain the 3D coordinates of all objects in the scene and only requires RGB images. While this approach has broad applications, it overlooks the semantic understanding of the 3D space and its objects, making it unable to accomplish specific object localization based on human instructions. To carry out more effective human-machine interaction on devices equipped with cameras, such as drones, surveillance systems, intelligent vehicles, and robots, it is necessary to perform visual grounding using natural language in monocular RGB images.

In this work, we introduce a task of 3D object localization through language descriptions with geometry information directly in a single RGB image, termed Mono3DVG (see Fig. 1). Specifically, we build a large-scale dataset, Mono3DRefer, which provides 41,140 natural language expressions of 8,228 objects. Mono3DRefer's descriptions contain both appearance and geometry information, generated by ChatGPT and refined manually. Geometry information can provide more precise instructions and identify invisible objects. Even if the appearance of an object is the primary visual perception for humans, they tend to use geometry information to distinguish objects.

To perform inference based on the language with appearance and geometry information, we propose a novel end-to-end transformer-based approach, namely Mono3DVG-TR, which consists of a multi-modal feature encoder, a dual text-guided adapter, a grounding decoder, and a grounding head. First, we adopt transformer and CNN to extract textual and multi-scale visual features. Depth predictor is designed to explicitly learn geometry features. Second, to refine multi-scale visual and geometry features of the referred object, we propose the dual text-guided adapter to perform text-guided feature learning based on pixel-wise attention. Finally, a learnable query first aggregates the initial geometric features, then enhances text-related geometric features by text embedding and finally collects appearance features from multiscale visual features. The depth-text-visual stacking attention fuses object-level geometric cues and visual appearance into the query, fully realizing text-guided decoding.

Our contributions can be summarized as follows:

- We introduce a novel task of 3D visual grounding in monocular RGB images using descriptions with appearance and geometry information, termed Mono3DVG.
- We contribute a large-scale dataset, which contains 41,140 expressions generated by ChatGPT and refined manually based on the KITTI, named Mono3DRefer.
- We propose an end-to-end transformer-based network,

*Corresponding Authors.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

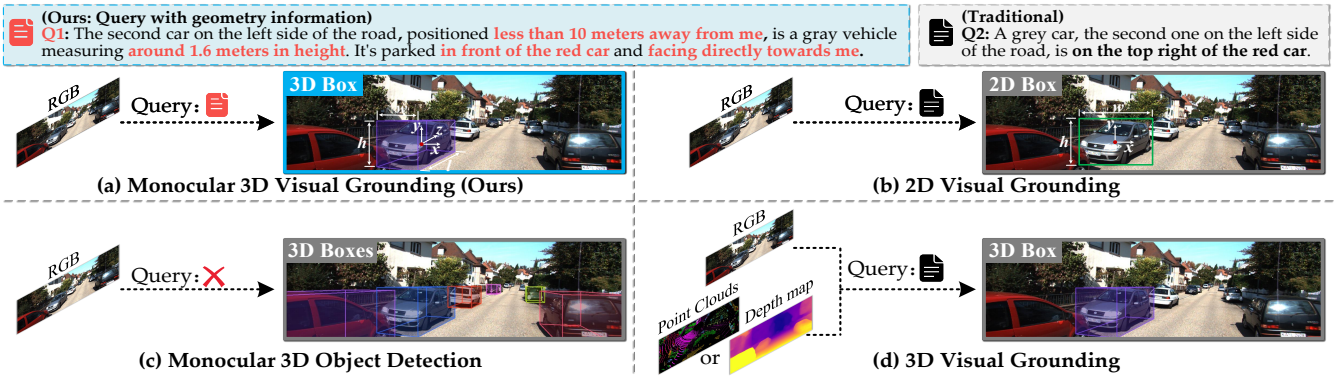

Figure 1: Introduction for 3D visual grounding in monocular images (Mono3DVG). (a) Mono3DVG aims to localize the true 3D extent of referred objects in an image using language descriptions with geometry information. (b) The counterpart 2D task does not capture the 3D extent of the referred object. (c) Localizing specific objects is not feasible for monocular 3D object detection. (d) 3D visual grounding requires laser radars or RGB-D sensors, which greatly limits its application scenarios.

Dataset	Publication	Expression Num.	Object Num.	Scene Num.	Range	Exp. Length	Vocab	Scene	Target
SUN-Spot	*ICCVW'2019*	7,990	3,245	1,948	–	14.04	2,690	Indoor	furni.
REVERIE	*CVPR'2020*	21,702	4,140	90	–	18.00	1,600	Indoor	furni.
ScanRefrer	*ECCV'2020*	51,583	11,046	704	10m	20.27	4,197	Indoor	furni.
Sr3d	*ECCV'2020*	83,572*	8,863	1,273	10m	–	196	Indoor	furni.
Nr3d	*ECCV'2020*	41,503	5,879	642	10m	11.40	6,951	Indoor	furni.
SUNRefer	*CVPR'2021*	38,495	7,699	7,699	–	16.30	5,279	Indoor	furni.
STRefer	*arXiv'2023*	5,458	3,581	662	30m	–	–	Outdoor	human
LifeRefer	*arXiv'2023*	25,380	11,864	3,172	30m	–	–	In/Outdoor	human
Mono3DRefer	–	**41,140**	**8,228**	**2,025**	**102m**	**53.24**	**5,271**	**Outdoor**	**human, vehicle**

Table 1: Statistic comparison of visual grounding datasets in the 3D scene, where 'num.' denotes number, 'exp.' indicates expression, and 'furni.' means furniture. '*' represents the unique text data automatically generated and the largest amount.

Mono3DVG-TR, which fully aggregates the appearance and geometry features in multi-modal embedding.

- We provide sufficient benchmarks based on two-stage and one-stage methods. Extensive experiments show that our method significantly outperforms all baselines.

Related Work

2D Visual Grounding

The earlier two-stage approaches (Zhang, Niu, and Chang 2018; Hu et al. 2017; Yu et al. 2018a; Liu et al. 2019b; Yu et al. 2018b; Chen, Kovvuri, and Nevatia 2017) adopt a pre-trained detector to generate region proposals and extract visual features. It obtains the optimal proposal by calculating scores with vision-language features and sorting. Additionally, NMTree (Liu et al. 2019a) and RvG-Tree (Hong et al. 2022) utilize tree networks by parsing the expression. To capture objects' relation, graph neural network is adopted by Yang, Li, and Yu (2019); Wang et al. (2019); Yang, Li, and Yu (2020). Recently, the one-stage pipeline has been widely used due to its low computational cost. Many works (Chen et al. 2018; Sadhu, Chen, and Nevatia 2019; Yang et al. 2019, 2020; Huang et al. 2021; Liao et al. 2022) use visual and text encoders to extract visual and textual features, and then fuse the multi-modal features to regress box coordinates. They do not depend on the quality of pre-generated proposals. Du et al. (2022) and Deng et al. (2021) first design the end-to-end transformer-based network, which has achieved superior results in terms of both speed and performance. (Li and Sigal 2021; Sun et al. 2022) propose the multi-task framework to further improve the performance. (Yang et al. 2022; Ye et al. 2022) focus on adjusting visual features by multi-modal features. Mauceri, Palmer, and Heckman (2019) present dataset for 2D visual grounding in RGB-D images. Qi et al. (2020) study 2D visual grounding for language-guided navigation in indoor scenes. However, these works cannot obtain the true 3D coordinates of the object in the real world, which greatly limits the application.

Monocular 3D Object Detection

The methods can be summarized into anchor-based, keypoint-based, and pseudo-depth based methods. The anchor-based method requires preset 3D anchors and regresses a relative offset. M3D-RPN (Brazil and Liu 2019) is an end-to-end network that only requires training a 3D region proposal network. Kinematic3D (Brazil et al. 2020) improves M3D-RPN by utilizing 3D kinematics to extract scene dynamics. Furthermore, some researchers predict key points and then estimate the size and location of 3D bounding boxes, such as SMOKE (Liu, Wu, and Tóth 2020),

Dataset	Language Context		Visual Context		Label	Task
	Form	Cost	Form	Cost		
SUN-Spot	manual	★★★★★	RGB-D	★★★☆☆	2D bbox	2D Visual Grounding in RGBD
REVERIE	manual	★★★★★	pc	★★★★☆	2D bbox	Localise Remote Object
ScanRefrer	manual	★★★★★	pc	★★★★☆	3D bbox	3D Visual Grounding
Sr3d	templated	★☆☆☆☆	pc	★★★★☆	3D bbox	3D Visual Grounding
Nr3d	manual	★★★★★	pc	★★★★☆	3D bbox	3D Visual Grounding
SUNRefer	manual	★★★★★	RGB-D	★★★☆☆	3D bbox	3D Visual Grounding in RGBD
STRefer	manual	★★★★★	pc & RGB	★★★★★	3D bbox	3D Visual Grounding in the Wild
LifeRefer	manual	★★★★★	pc & RGB	★★★★★	3D bbox	3D Visual Grounding in the Wild
Mono3DRefer	**ChatGPT+manual**	★★☆☆☆	**RGB**	★☆☆☆☆	**2D/3D bbox**	3D Visual Grounding in RGB

Table 2: The form, cost, and label of the datasets collected in Table 1 and the corresponding tasks. 'pc' denotes point cloud and 'bbox' means bounding box.

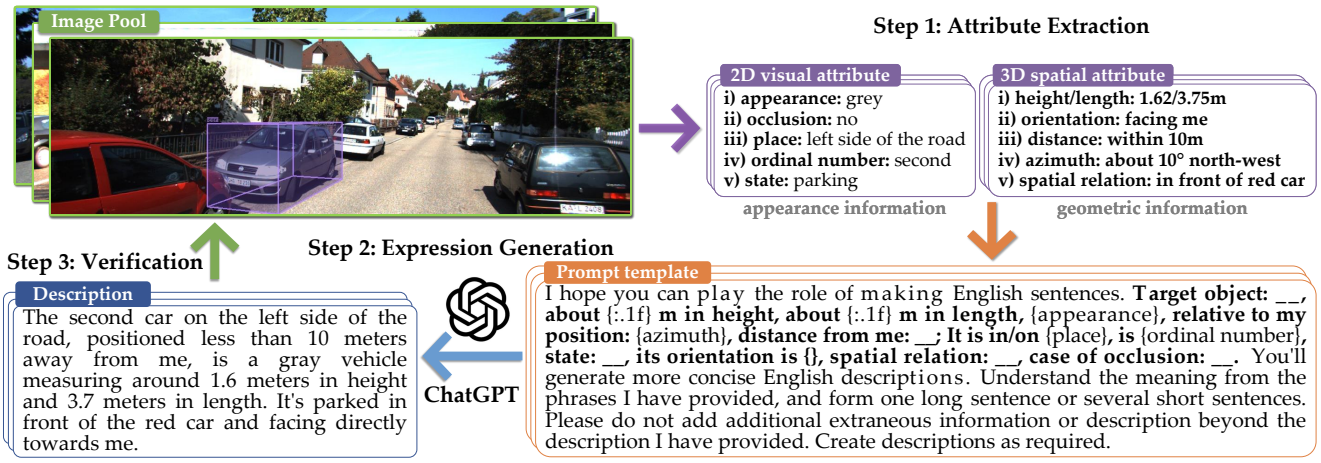

Figure 2: Our data collection pipeline: i) 2D visual attributes that provide appearance information and 3D spatial attributes that provide geometric information of the target are extracted. ii) fill in the prompt template we designed with attributes, and input the complete prompt into ChatGPT to get descriptions. iii) check whether the description can uniquely identify the object.

FCOS3D (Wang et al. 2021), MonoGRNet (Qin, Wang, and Lu 2019), and MonoFlex (Zhang, Lu, and Zhou 2021). However, due to the lack of depth information, pure monocular approaches have difficulty accurately localizing targets. Other works (Bao, Xu, and Chen 2020; Ding et al. 2020; Park et al. 2021; Chen, Dai, and Ding 2022) utilize extra depth estimators to supplement depth information. However, existing models only extract spatial relationships and depth information from visual content. Hence, we propose to explore the impact of language with geometry attributes on 3D object detection.

3D Visual Grounding

To handle this task, Scanrefer (Chen, Chang, and Nießner 2020) and Referit3D (Achlioptas et al. 2020) first create datasets. Similar to the counterpart 2D task, earlier works adopt the two-stage pipeline which uses a pre-trained detector to generate object proposals and extract features, such as PointNet++ (Qi et al. 2017). SAT (Yang et al. 2021) adopts 2D object semantics as extra input to assist training. InstanceRefer (Yuan et al. 2021) converts this task into an instance matching problem. To understand complex and diverse descriptions in point clouds directly, Feng et al. (2021) construct a language scene graph, a 3D proposal relation graph, and a 3D visual graph. 3DVG-Trans (Zhao et al. 2021), TransRefer3D (He et al. 2021), Multi-View Trans (Huang et al. 2022b), and LanguageRefer (Roh et al. 2022) all develop transformer-based architectures. D3Net (Chen et al. 2022) and 3DJCG (Cai et al. 2022) both develop a unified framework for dense captioning and visual grounding. Liu et al. (2021) present a novel task for 3D visual grounding in RGB-D images. The previous works are all in indoor environments and target furniture as the object. To promote the application, Lin et al. (2023) introduce the task in large-scale dynamic outdoor scenes based on online captured 2D images and 3D point clouds. However, capturing visual data through LiDAR or the industrial camera is expensive and not readily available for a wide range of applications. Our work focuses on the 3D visual grounding in a single image.

Mono3DRefer Dataset

As shown in Table 1 and Table 2, previous SUN-Spot (Mauceri, Palmer, and Heckman 2019) and REVERIE (Qi et al. 2020) only focus on 2D bounding boxes in the 3D

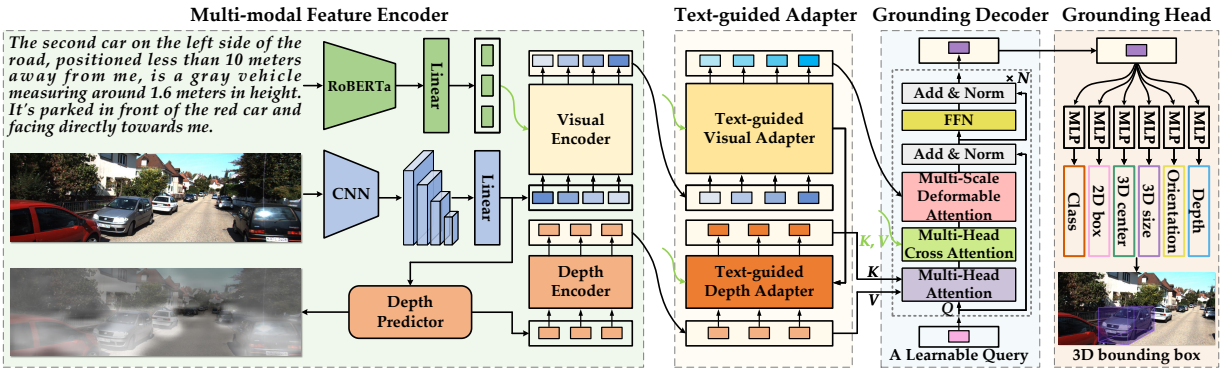

Figure 3: Overview of the proposed framework. The multi-modal feature encoder first extracts textual, multi-scale visual, and geometry features. The dual text-guided adapter refines visual and geometry features of referred objects based on pixel-wise attention. A learnable query fuses geometry cues and visual appearance of the object using depth-text-visual stacking attention in the grounding decoder. Finally, the grounding head adopts multiple MLPs to predict the 2D and 3D attributes of the target.

scene. Subsequently, ScanRefer (Chen, Chang, and Nießner 2020), Sr3d, Nr3d (Achlioptas et al. 2020), and SUNRefer (Liu et al. 2021) are built to investigate 3D visual grounding, but they are limited to indoor static scenes. Although STRefer and LifeRefer (Lin et al. 2023) focus on outdoor dynamic scenes, they require LiDAR and industrial cameras. To facilitate the broad application of 3D visual grounding, we employ both manually annotated and ChatGPT to annotate a large-scale dataset based on KITTI (Geiger, Lenz, and Urtasun 2012) for Mono3DVG.

Data Annotation

To cover all scenes and reduce inter-frame similarity, we performed scene clustering on the original KITTI dataset and sampled 2025 images from each category. The annotation pipeline of Fig. 2 consists of three stages. **Step 1: Attribute extraction.** The attributes of objects are divided into 2D visual attributes (appearance, occlusion, place, ordinal number, state) and 3D spatial attributes (height/length, orientation, distance, azimuth, spatial relationship). The color of appearance is preliminarily extracted by the HSV color recognition method. Occlusion and height/length are directly obtained from labels of the raw KITTI. Based on the 302 category results of scene clustering, unified rough annotations are performed for the scene place and state of objects in each category. Distance and azimuth are calculated by the coordinates of 3D boxes. Spatial relations include i) Horizontal Proximity, ii) Between, and iii) Allocentric such as far from, next to, between A and B, on the left, and in front. The judgment model is established based on 3D boxes and space geometry to preliminarily extract ordinal number, orientation, and spatial relation. Finally, to ensure correctness, we organize four people to verify and correct 2D and 3D attributes that provide appearance and geometric information. **Step 2: Expression generation.** We customize the prompt template for generating expressions for ChatGPT. Fill in the template with each attribute of objects and input the complete prompt into ChatGPT to obtain the descriptions. **Step 3: Verification.** To guarantee the correctness of descriptions, four persons from our team jointly verify the dataset.

Dataset Statistics

Table 1 summarizes the statistical information of the dataset. We sample 2025 frames of images from the original KITTI for Mono3DRefer, containing 41,140 expressions in total and a vocabulary of 5,271 words. In addition to the Sr3d generated through templates, Mono3DRefer has a similar number of expressions as the ScanRefer and Nr3d. For the range, 10m is the range of the whole scene pre-scanned by RGB-D sensors, 30m is the approximated perception radius with annotations for the LiDAR sensor, and 102m is the distance range of objects with annotations for our dataset. The average length of expressions generated by ChatGPT is 53.24 words involving visual appearance and geometry information. Table 2 shows that the Mono3DVG task has relatively low language data collection costs and the lowest visual data collection costs. We provide more detailed statistics and analyses in the supplementary materials.

Methodology

As shown in Fig. 3, we propose an end-to-end transformer-based framework, Mono3DVG-TR, which consists of four main modules: 1) the encoder; 2) the adapter; 3) the decoder; 4) the grounding head.

Multi-modal Feature Encoder

We leverage pre-trained RoBERTa-base (Liu et al. 2019c) and a linear layer to extract the textual embeddings $\boldsymbol{p}_t \in \mathbb{R}^{C\times N_t}$, where N_t is the length of the input sentence. For the image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, we utilize a CNN backbone (*i.e.*, ResNet-50 (He et al. 2016) and an additional convolutional layer) and a linear layer to obtain four level multi-scale visual features $\boldsymbol{f}_v \in \mathbb{R}^{C\times N_v}$, where $C = 256$ and $N_v = \frac{H}{8}\times\frac{W}{8} + \frac{H}{16}\times\frac{W}{16} + \frac{H}{32}\times\frac{W}{32} + \frac{H}{64}\times\frac{W}{64}$. Following Zhang et al. (2022), we use the lightweight depth predictor to get the geometry feature $\boldsymbol{f}_g \in \mathbb{R}^{C\times N_g}$, where $N_g = \frac{H}{16}\times\frac{W}{16}$. Then we design visual encoder and depth encoder to conduct global context inference and generate embeddings with long-term dependencies, denoted as $\boldsymbol{p}_v \in \mathbb{R}^{C\times N_v}$, $\boldsymbol{p}_g \in \mathbb{R}^{C\times N_g}$. The depth encoder is composed of one transformer

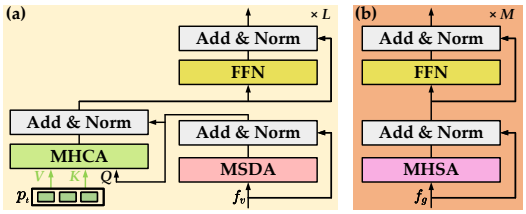

Figure 4: Detail of visual encoder (a) and depth encoder (b).

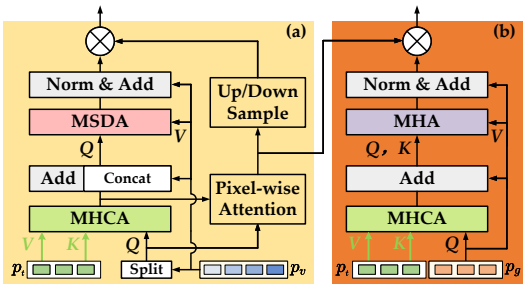

Figure 5: Detail of text-guided visual and depth adapter.

encoder layer to encode geometry embeddings. In Fig. 4(a), visual encoder replaces multi-head self-attention (MHSA) with multi-scale deformable attention (MSDA) to avoid excessive attention computation on multi-scale visual features. Moreover, we insert an additional multi-head cross-attention (MHCA) layer between MSDA layer and feed-forward network (FFN), providing textual cues for visual embeddings.

Dual Text-guided Adapter

To exploit the appearance and geometry information in text, the dual adapter is proposed. In Fig. 5(b), the depth adapter takes the geometry embedding $\boldsymbol{p}_g$ as the query for MHCA and takes the text embedding $\boldsymbol{p}_t$ as the key and value. Then, a multi-head attention (MHA) layer is used to apply implicit text-guided self-attention to the geometry features. Original geometry embedding $\boldsymbol{p}_g$ as the value. The refined geometry feature is denoted as $\boldsymbol{p}_g^{''}$. Visual adapter requires splitting and concatenating multi-scale visual embeddings $\boldsymbol{p}_v$ before and after MHCA which uses $\boldsymbol{p}_v^{\frac{1}{16}}$ with the size of $\frac{H}{16} \times \frac{W}{16}$ as the query. Then, MSDA is used instead of MHA, and the refined visual feature is denoted as $\boldsymbol{p}_v^{''}$.

Then, we linearly project $\boldsymbol{p}_v^{\frac{1}{16}}$ and the output of MHCA in the visual adapter to obtain the original visual feature map $\boldsymbol{F}_{orig} \in \mathbb{R}^{C \times \frac{H}{16} \times \frac{W}{16}}$ and the text-related $\boldsymbol{F}_{text} \in \mathbb{R}^{C \times \frac{H}{16} \times \frac{W}{16}}$, respectively. To explore the alignment relationship and fine-grained correlation between vision and language, we compute the attention score $\boldsymbol{s}_{ij} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$ for each region (i, j) in the feature map as follows:

$$\boldsymbol{F}_{orig} = \|\boldsymbol{F}_{orig}\|_2, \ \boldsymbol{F}_{text} = \|\boldsymbol{F}_{text}\|_2, \tag{1}$$

$$\boldsymbol{a}_{ij}^c = \boldsymbol{F}_{orig}^c(i,j) \odot \boldsymbol{F}_{text}^c(i,j), c = 1, 2, \ldots, C \tag{2}$$

$$\boldsymbol{s}_{ij} = \sum_{c=1}^{C} \boldsymbol{a}_{ij}^c. \tag{3}$$

where, $\|\cdot\|_2$ and $\odot$ indicate l_2-norm and element-wise product respectively. Then, we further model the semantic similarity $\boldsymbol{S}^{\frac{1}{16}}$ with the size of $\frac{H}{16} \times \frac{W}{16}$ between each pixel feature and the text feature using the Gaussian function:

$$\boldsymbol{S}^{\frac{1}{16}} = \alpha \cdot \exp(-\frac{(1 - \boldsymbol{s}_{ij})^2}{2\sigma^2}), \tag{4}$$

where, α and σ are a scaling factor and standard deviation, respectively, and both are learnable parameters. We upsample $\boldsymbol{S}^{\frac{1}{16}}$ using bilinear interpolation and downsample $\boldsymbol{S}^{\frac{1}{16}}$ using max pooling. Then we concatenate the flattened score maps to obtain the multi-scale attention score $S \in \mathbb{R}^{N_v}$:

$$S = \text{Concat}[\text{Up}(\boldsymbol{S}^{\frac{1}{16}}), \boldsymbol{S}^{\frac{1}{16}}, \text{Down}(\boldsymbol{S}^{\frac{1}{16}}), \text{Down}(\boldsymbol{S}^{\frac{1}{16}})]. \tag{5}$$

Based on pixel-wise attention scores, the visual and geometry features are focused on the regions relevant to the textual description. We use the features $p_v^{''}$ and $p_g^{''}$ and scores ($S^{\frac{1}{16}} \in \mathbb{R}^{N_d}$ is flattened) to perform element-wise multiplication, resulting in adapted features of the referred object:

$$\tilde{\boldsymbol{p}}_v = \boldsymbol{p}_v^{''} \cdot S, \ \tilde{\boldsymbol{p}}_g = \boldsymbol{p}_g^{''} \cdot S^{\frac{1}{16}}. \tag{6}$$

Grounding Decoder

As shown in Fig. 3, the n-th decoder layer consists of a block composed of MHA, MHCA, and MSDA, and an FFN. The learnable query $\boldsymbol{p}_q \in \mathbb{R}^{C \times 1}$ first aggregates the initial geometric information, then enhances text-related geometric features by text embedding, and finally collects appearance features from multi-scale visual features. This depth-text-visual stacking attention adaptively fuses object-level geometric cues and visual appearance into the query.

Grounding Head

Our grounding head employs multiple MLPs for 2D and 3D attribute prediction. The output of the decoder, *i.e.*, the learnable query, is denoted by $\tilde{\boldsymbol{p}}_q \in \mathbb{R}^{C \times 1}$. Then, $\tilde{\boldsymbol{p}}_q$ is separately fed into a linear layer for predicting the object category, a 3-layer MLP for the 2D box size (l, r, t, b) and projected 3D box center (x_{3D}, y_{3D}), a 2-layer MLP for the 3D box size (h_{3D}, w_{3D}, l_{3D}), a 2-layer MLP for the 3D box orientation θ, and a 2-layer MLP for the depth d_{reg}. (l, r, t, b) represents the distances between the four sides of the 2D box and the projected 3D center point (x_{3D}, y_{3D}). Similar to (Zhang et al. 2022), the final predicted depth d_{pred} is computed.

Loss Function

We group the category, 2D box size, and projected 3D center as 2D attributes, and the 3D box size, orientation, and depth as 3D attributes. The loss for 2D is formulated as:

$$\mathcal{L}_{2D} = \lambda_1 \mathcal{L}_{class} + \lambda_2 \mathcal{L}_{lrtb} + \lambda_3 \mathcal{L}_{GIoU} + \lambda_4 \mathcal{L}_{xy3D}, \tag{7}$$

where, $\lambda_{1\sim4}$ is set to $(2, 5, 2, 10)$ following (Zhang et al. 2022). $\mathcal{L}_{class}$ is Focal loss (Lin et al. 2017) for predicting nine categories. $\mathcal{L}_{lrtb}$ and $\mathcal{L}_{xy3D}$ adopt the L1 loss. $\mathcal{L}_{GIoU}$ is the GIoU loss (Rezatofighi et al. 2019) that constrains the 2D bounding boxes. The loss for 3D is defined as:

$$\mathcal{L}_{3D} = \mathcal{L}_{size3D} + \mathcal{L}_{orien} + \mathcal{L}_{depth}. \tag{8}$$

Method	Type	Unique		Multiple		Overall		Time cost
		Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	(ms)
CatRand	Two-Stage	100	100	24.47	24.43	38.69	38.67	0
Cube R-CNN + Rand	Two-Stage	32.76	14.61	13.36	7.21	17.02	8.60	153
Cube R-CNN + Best	Two-Stage	35.29	16.67	60.52	32.99	55.77	29.92	153
ZSGNet + backproj	One-Stage	9.02	0.29	16.56	2.23	15.14	1.87	31
FAOA + backproj	One-Stage	11.96	2.06	13.79	2.12	13.44	2.11	144
ReSC + backproj	One-Stage	11.96	0.49	23.69	3.94	21.48	3.29	97
TransVG + backproj	Tran.-based	15.78	4.02	21.84	4.16	20.70	4.14	80
Mono3DVG-TR (Ours)	Tran.-based	**57.65**	**33.04**	**65.92**	**46.85**	**64.36**	**44.25**	110

Table 3: Comparison with baselines. The underline means performance exceeding our bolded results.

We use the 3D IoU oriented loss (Ma et al. 2021), MultiBin loss (Chen et al. 2020), and Laplacian aleatoric uncertainty loss (Chen et al. 2020) as $\mathcal{L}_{size3D}$, $\mathcal{L}_{orien}$, and $\mathcal{L}_{depth}$ to optimize the predicted 3D size, orientation, and depth. Following (Zhang et al. 2022), we use Focal loss to supervise the prediction of the depth map, denoted as $\mathcal{L}_{dmap}$. Finally, our overall loss is formulated as:

$$\mathcal{L}_{overall} = \mathcal{L}_{2D} + \mathcal{L}_{3D} + \mathcal{L}_{dmap}. \tag{9}$$

Experiments

Implementation Details. We split our dataset into 29,990, 5,735, and 5,415 expressions for train/val/test sets respectively. We train 60 epochs with a batch size of 10 by AdamW with 10^{-4} learning rate and 10^{-4} weight decay on one GTX 3090 24-GiB GPU. The learning rate decays by a factor of 10 after 40 epochs. The dropout ratio is set to 0.1.

Evaluation metric. Similar to (Chen, Chang, and Nießner 2020; Liu et al. 2021; Lin et al. 2023), we use the accuracy with 3D IoU threshold (Acc@0.25 and Acc@0.5) as our metrics, where the threshold includes 0.25 and 0.5.

Baselines. To explore the difficulty and enable fair comparisons, we design several baselines and validate these methods using a unified standard. **Two-stage:** 1) CatRand randomly selects a ground truth box that matches the object category as the prediction result. This baseline measures the difficulty of our task and dataset. 2) (Cube R-CNN (Brazil et al. 2023) + Rand) randomly selects a bounding box that matches the object category as the prediction result from predicted object proposals of Cube R-CNN, the best monocular 3D object detector. 3) (Cube R-CNN (Brazil et al. 2023) + Best) selects a bounding box that best matches the ground truth box from predicted object proposals. This baseline provides the upper bound on how well the two-stage approaches work for our task. **One-stage:** 2DVG backproj baselines adapt the results of 2D visual grounding to 3D by using back-projection. We select three SOTA one-stage methods, *i.e.*, ZSGNet (Sadhu, Chen, and Nevatia 2019), FAOA (Yang et al. 2019) , ReSC (Yang et al. 2020), and the transformer-based TransVG (Deng et al. 2021).

To analyze the importance of other information besides the category, we report metrics of these baselines on 'unique' and 'multiple' subsets in Table 3. The 'unique' subset means cases where there is one object that matches the category, while the 'multiple' subset contains multiple confused objects with the same category. To analyze the task difficulty, we report metrics at varying levels of depth d as near: $0 < d \leq 15$m, medium: 15m $< d \leq$ 35m, far: 35m $< d \leq \infty$ in Table 4. Considering that occlusion or truncation of the objects adds challenge to the task, we also show metrics at varying levels of difficulty as easy: no occlusion and truncation < 0.15, moderate: no/partial occlusion and truncation < 0.3, hard: others. For more convincing results, we show the average of 5 evaluations with different random seeds for CatRand and Cube R-CNN Rand.

Quantitative Analysis and Task Difficulty

In Table 3, CatRand achieves 100% accuracy on the 'unique' subset but only 24% on the 'multiple'. Cube R-CNN Rand also performs better on the 'unique' subset compared to the 'multiple'. If there is only one car in an image, inputting the "car" is sufficient. However, if there are multiple cars, additional information beyond the category is necessary. The significant gap between Cube R-CNN Best and CatRand on the 'unique' subset indicates tremendous research potential in monocular 3D object detection. Overall, while our result is close to the CatRand, there is still room for improvement.

In Table 4, CatRand performs much better on the 'far' subset compared to 'near' and 'medium'. Our method and other baselines show a decreasing performance as the depth increases. The 'far' subset contains fewer ambiguous objects, so CatRand's random selection of ground truth can achieve better results. Other methods rely on predicted bounding boxes. Generally, objects that are farther away from the camera are more challenging to accurately predict their depth and 3D extent. Cube R-CNN Best exhibits excellent results on Acc@0.25. The accuracy gap between CatRand and our method on the 'far' subset indicates that accurately predicting the depth of target objects based on a single image and natural language is a challenge in our task.

For 'easy-moderate-hard' subsets, Cube R-CNN Best has suboptimal results on Acc@0.25, but a lower Acc@0.5, indicating that the best object detector has the ability to detect occluded or truncated objects, but the accuracy needs to be improved. Our method fully fuses visual and textual features to accurately detect occluded and truncated objects, achieving better results than CatRand.

Our method outperforms all 2DVG backproj baselines by a significant margin in Tables 3-4. It is inefficient to obtain accurate 3D bounding boxes from 2D localization results by back projection. The methods of 2DVG can only predict the extent of the object in the 2D plane and lack the ability to

Method	Type	Near / Easy		Medium / Moderate		Far / Hard	
		Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
CatRand	Two-Stage	31.16/47.29	31.05/47.26	35.49/33.92	35.49/33.92	52.11/30.83	52.11/30.74
Cube R-CNN + Rand	Two-Stage	17.40/21.12	11.45/11.41	18.01/17.85	8.15/8.01	14.91/10.56	6.38/5.18
Cube R-CNN + Best	Two-Stage	67.76/59.66	41.45/33.05	60.69/60.56	30.35/33.45	34.72/46.25	17.01/22.52
ZSGNet + backproj	One-Stage	24.87/21.33	0.59/3.35	16.74/13.87	3.71/0.63	2.15/7.57	0.07/0.84
FAOA + backproj	One-Stage	18.03/17.51	0.53/3.43	15.64/12.18	3.95/1.34	4.86/8.83	0.62/0.90
ReSC + backproj	One-Stage	33.68/27.90	0.59/5.71	24.03/19.23	6.15/1.97	4.24/14.41	1.25/1.02
TransVG + backproj	Tran.-based	29.34/28.88	0.86/6.95	25.05/16.41	8.02/2.75	4.17/12.91	0.97/1.38
Mono3DVG-TR (Ours)	Tran.-based	**64.74/72.36**	**53.49/51.80**	**75.44/69.23**	**55.48/48.66**	**45.07/49.01**	**15.35/29.91**

Table 4: Results for 'near'-'medium'-'far' subsets and 'easy'-'moderate'-'hard' subsets. The underline means performance exceeding our bolded results.

Figure 6: Qualitative results from baseline methods and our Mono3DVG-TR. Blue, green, and red boxes denote the ground truth, prediction with IoU higher than 0.5, and prediction with IoU lower than 0.5, respectively.

estimate depth, resulting in inaccurate 3D localization.

Qualitative Analysis

Fig. 6 displays the 3D localization results of Cube R-CNN Best, ReSC backproj, TransVG backproj, and our proposed method. Although the approximate range of objects can be obtained, Cube R-CNN Best fails to provide precise bounding boxes. ReSC backproj and TransVG backproj depend on the accuracy of 2D boxes and are unable to estimate depth, thus unable to provide accurate 3D bounding boxes. Our method includes text-RGB and text-depth two branches to make full use of the appearance and geometry information for multi-modal fusion, but there are also some failures. We provide more detailed analyses in the supplementary.

Ablation Studies

We conduct detailed ablation studies to validate the effectiveness of our proposed network and report the Acc@0.25 and Acc@0.5 overall on the Mono3DRefer test set. In Table 5, we report results of a comprehensive ablation experiment on the main components. The first row shows the results by directly using visual and geometry features of the CNN backbone and depth predictor to decode. The second row shows a significant improvement with the addition of the encoder. In the third row, we only utilize the text-guided visual adapter. After adding the complete adapter, the results can be improved by approximately 4%-5%. We provide more detailed analyses of ablation studies in the supplementary.

Grouning Decoder	Encoder V.	Encoder D.	Adapter V.	Adapter D.	Acc0.25	Acc@0.5
✓					47.31	24.38
✓	✓	✓			60.21	38.52
✓	✓	✓	✓		61.98	40.12
✓	✓	✓	✓	✓	**64.36**	**44.25**

Table 5: The ablation studies of the proposed components of our approach. 'V.' and 'D.' denote visual and depth.

Conclusion

We introduce the novel task of Mono3DVG, which localizes 3D objects in RGB images by descriptions. Notably, we contribute a large-scale dataset, Mono3DRefer, which is the first dataset that leverages the ChatGPT to generate descriptions. We also provide a series of benchmarks to facilitate future research. Finally, we hope that Mono3DVG can be widely applied since it does not require strict conditions such as RGB-D sensors, LiDARs, or industrial cameras.

Acknowledgments

This work was supported in part by grants from the National Science Fund for Distinguished Young Scholars (No.61825603), the National Key Research and Development Project (No.2020YFB2103900), and the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University (No.CX2023030).

References

Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; and Guibas, L. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, 422–440.

Bao, W.; Xu, B.; and Chen, Z. 2020. MonoFENet: Monocular 3D Object Detection With Feature Enhancement Networks. *IEEE Transactions on Image Processing*, 29: 2753–2765.

Brazil, G.; Kumar, A.; Straub, J.; Ravi, N.; Johnson, J.; and Gkioxari, G. 2023. Omni3D: A Large Benchmark and Model for 3D Object Detection in the Wild. In *CVPR*, 13154–13164.

Brazil, G.; and Liu, X. 2019. M3d-rpn: Monocular 3d region proposal network for object detection. In *ICCV*, 9287–9296.

Brazil, G.; Pons-Moll, G.; Liu, X.; and Schiele, B. 2020. Kinematic 3d object detection in monocular video. In *ECCV*, 135–152.

Cai, D.; Zhao, L.; Zhang, J.; Sheng, L.; and Xu, D. 2022. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *CVPR*, 16464–16473.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 202–221.

Chen, D. Z.; Wu, Q.; Nießner, M.; and Chang, A. X. 2022. D3Net: A Unified Speaker-Listener Architecture for 3D Dense Captioning and Visual Grounding. In *ECCV*, 487–505.

Chen, K.; Kovvuri, R.; and Nevatia, R. 2017. Query-guided regression network with context policy for phrase grounding. In *ICCV*, 824–832.

Chen, X.; Ma, L.; Chen, J.; Jie, Z.; Liu, W.; and Luo, J. 2018. Real-time referring expression comprehension by single-stage grounding network. *arXiv preprint arXiv:1812.03426*.

Chen, Y.; Tai, L.; Sun, K.; and Li, M. 2020. MonoPair: Monocular 3D Object Detection Using Pairwise Spatial Relationships. In *CVPR*, 12093–12102.

Chen, Y.-N.; Dai, H.; and Ding, Y. 2022. Pseudo-stereo for monocular 3d object detection in autonomous driving. In *CVPR*, 887–897.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. TransVG: End-to-End Visual Grounding With Transformers. In *ICCV*, 1769–1779.

Ding, M.; Huo, Y.; Yi, H.; Wang, Z.; Shi, J.; Lu, Z.; and Luo, P. 2020. Learning depth-guided convolutions for monocular 3d object detection. In *CVPR workshops*, 1000–1001.

Du, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2022. Visual Grounding with Transformers. In *2022 IEEE International Conference on Multimedia and Expo*, 1–6.

Feng, M.; Li, Z.; Li, Q.; Zhang, L.; Zhang, X.; Zhu, G.; Zhang, H.; Wang, Y.; and Mian, A. 2021. Free-form description guided 3d visual graph network for object grounding in point cloud. In *ICCV*, 3722–3731.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 3354–3361.

He, D.; Zhao, Y.; Luo, J.; Hui, T.; Huang, S.; Zhang, A.; and Liu, S. 2021. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2344–2352.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hong, R.; Liu, D.; Mo, X.; He, X.; and Zhang, H. 2022. Learning to Compose and Reason with Language Tree Structures for Visual Grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2): 684–696.

Hu, R.; Rohrbach, M.; Andreas, J.; Darrell, T.; and Saenko, K. 2017. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 1115–1124.

Huang, B.; Lian, D.; Luo, W.; and Gao, S. 2021. Look before you leap: Learning landmark features for one-stage visual grounding. In *CVPR*, 16888–16897.

Huang, K.-C.; Wu, T.-H.; Su, H.-T.; and Hsu, W. H. 2022a. MonoDTR: Monocular 3D Object Detection With Depth-Aware Transformer. In *CVPR*, 4012–4021.

Huang, S.; Chen, Y.; Jia, J.; and Wang, L. 2022b. Multi-view transformer for 3d visual grounding. In *CVPR*, 15524–15533.

Li, M.; and Sigal, L. 2021. Referring transformer: A one-step approach to multi-task visual grounding. In *Advances in Neural Information Processing Systems*, volume 34, 19652–19664.

Liao, Y.; Zhang, A.; Chen, Z.; Hui, T.; and Liu, S. 2022. Progressive Language-Customized Visual Feature Learning for One-Stage Visual Grounding. *IEEE Transactions on Image Processing*, 31: 4266–4277.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*, 2980–2988.

Lin, Z.; Peng, X.; Cong, P.; Hou, Y.; Zhu, X.; Yang, S.; and Ma, Y. 2023. WildRefer: 3D Object Localization in Large-scale Dynamic Scenes with Multi-modal Visual Data and Natural Language. *arXiv preprint arXiv:2304.05645*.

Liu, D.; Zhang, H.; Wu, F.; and Zha, Z.-J. 2019a. Learning to assemble neural module tree networks for visual grounding. In *ICCV*, 4673–4682.

Liu, H.; Lin, A.; Han, X.; Yang, L.; Yu, Y.; and Cui, S. 2021. Refer-it-in-rgbd: A bottom-up approach for 3d visual grounding in rgbd images. In *CVPR*, 6032–6041.

Liu, X.; Wang, Z.; Shao, J.; Wang, X.; and Li, H. 2019b. Improving referring expression grounding with cross-modal attention-guided erasing. In *CVPR*, 1950–1959.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019c. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; Wu, Z.; and Tóth, R. 2020. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *CVPR Workshops*, 996–997.

Ma, X.; Zhang, Y.; Xu, D.; Zhou, D.; Yi, S.; Li, H.; and Ouyang, W. 2021. Delving Into Localization Errors for Monocular 3D Object Detection. In *CVPR*, 4721–4730.

Mauceri, C.; Palmer, M.; and Heckman, C. 2019. Sun-spot: An rgb-d dataset with spatial referring expressions. In *ICCV Workshops*, 1883–1886.

Park, D.; Ambrus, R.; Guizilini, V.; Li, J.; and Gaidon, A. 2021. Is pseudo-lidar needed for monocular 3d object detection? In *ICCV*, 3142–3152.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, volume 30.

Qi, Y.; Wu, Q.; Anderson, P.; Wang, X.; Wang, W. Y.; Shen, C.; and Hengel, A. v. d. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*, 9982–9991.

Qin, Z.; Wang, J.; and Lu, Y. 2019. Monogrnet: A geometric reasoning network for monocular 3d object localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8851–8858.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 658–666.

Roh, J.; Desingh, K.; Farhadi, A.; and Fox, D. 2022. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, 1046–1056.

Sadhu, A.; Chen, K.; and Nevatia, R. 2019. Zero-shot grounding of objects from natural language queries. In *ICCV*, 4694–4703.

Sun, M.; Suo, W.; Wang, P.; Zhang, Y.; and Wu, Q. 2022. A proposal-free one-stage framework for referring expression comprehension and generation via dense cross-attention. *IEEE Transactions on Multimedia*.

Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; and Hengel, A. v. d. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *CVPR*, 1960–1968.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *ICCV*, 913–922.

Yang, L.; Xu, Y.; Yuan, C.; Liu, W.; Li, B.; and Hu, W. 2022. Improving Visual Grounding with Visual-Linguistic Verification and Iterative Reasoning. In *CVPR*, 9499–9508.

Yang, S.; Li, G.; and Yu, Y. 2019. Dynamic graph attention for referring expression comprehension. In *ICCV*, 4644–4653.

Yang, S.; Li, G.; and Yu, Y. 2020. Relationship-embedded representation learning for grounding referring expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8): 2765–2779.

Yang, Z.; Chen, T.; Wang, L.; and Luo, J. 2020. Improving one-stage visual grounding by recursive sub-query construction. In *ECCV*, 387–404.

Yang, Z.; Gong, B.; Wang, L.; Huang, W.; Yu, D.; and Luo, J. 2019. A fast and accurate one-stage approach to visual grounding. In *ICCV*, 4683–4693.

Yang, Z.; Zhang, S.; Wang, L.; and Luo, J. 2021. Sat: 2d semantics assisted training for 3d visual grounding. In *ICCV*, 1856–1866.

Ye, J.; Tian, J.; Yan, M.; Yang, X.; Wang, X.; Zhang, J.; He, L.; and Lin, X. 2022. Shifting More Attention to Visual Backbone: Query-modulated Refinement Networks for End-to-End Visual Grounding. In *CVPR*, 15502–15512.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018a. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 1307–1315.

Yu, Z.; Yu, J.; Xiang, C.; Zhao, Z.; Tian, Q.; and Tao, D. 2018b. Rethinking diversified and discriminative proposal generation for visual grounding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 1114–1120.

Yuan, Z.; Yan, X.; Liao, Y.; Zhang, R.; Wang, S.; Li, Z.; and Cui, S. 2021. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *ICCV*, 1791–1800.

Zhan, Y.; Xiong, Z.; and Yuan, Y. 2023. RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Zhang, H.; Niu, Y.; and Chang, S.-F. 2018. Grounding referring expressions in images by variational context. In *CVPR*, 4158–4166.

Zhang, R.; Qiu, H.; Wang, T.; Guo, Z.; Xu, X.; Qiao, Y.; Gao, P.; and Li, H. 2022. MonoDETR: depth-guided transformer for monocular 3D object detection. *arXiv preprint arXiv:2203.13310*.

Zhang, Y.; Lu, J.; and Zhou, J. 2021. Objects are different: Flexible monocular 3d object detection. In *CVPR*, 3289–3298.

Zhao, L.; Cai, D.; Sheng, L.; and Xu, D. 2021. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, 2928–2937.