Mono3D-VLDL: Perception-Aware Vision-Language Dictionary Learning for Multimodal Fusion in Monocular 3D Grounding

Tiantian Wang, Haixiang Hu, Haoxiang Liang*, Zhaoyang Zhang* , Tinglei Jia,
Shuwen Huang, Yongfeng Bu, Xiaowei Qian, Rong Wang, Kaifei Li,
Hanke Luo, Hua Cui
Chang'an University
wangtiantian@chd.edu.cn, lhx@chd.edu.cn

Abstract

*We propose **Mono3D-VLDL**, a novel single-stage framework for language-visual fusion in robotic vision, addressing the limitations of traditional two-stage methods that separately perform image registration and feature fusion. These methods are computationally intensive, hardware-demanding, and struggle with the modality gap between language and visual data, particularly in dynamic environments. **Mono3D-VLDL** integrates image registration and feature fusion into a unified stage, eliminating explicit registration. The framework employs a Cross-Modality Dictionary to compensate for missing textual information while preserving modality-specific features. Additionally, it uses parallel cross-attention mechanisms to effectively integrate depth, text, and visual information for robust 3D object attribute prediction. Experiments on the Mono3DRefer dataset demonstrate that our method achieves superior efficiency and accuracy compared to existing one-stage approaches, making it highly suitable for real-time robotic applications in resource-constrained settings.*

1. Introduction

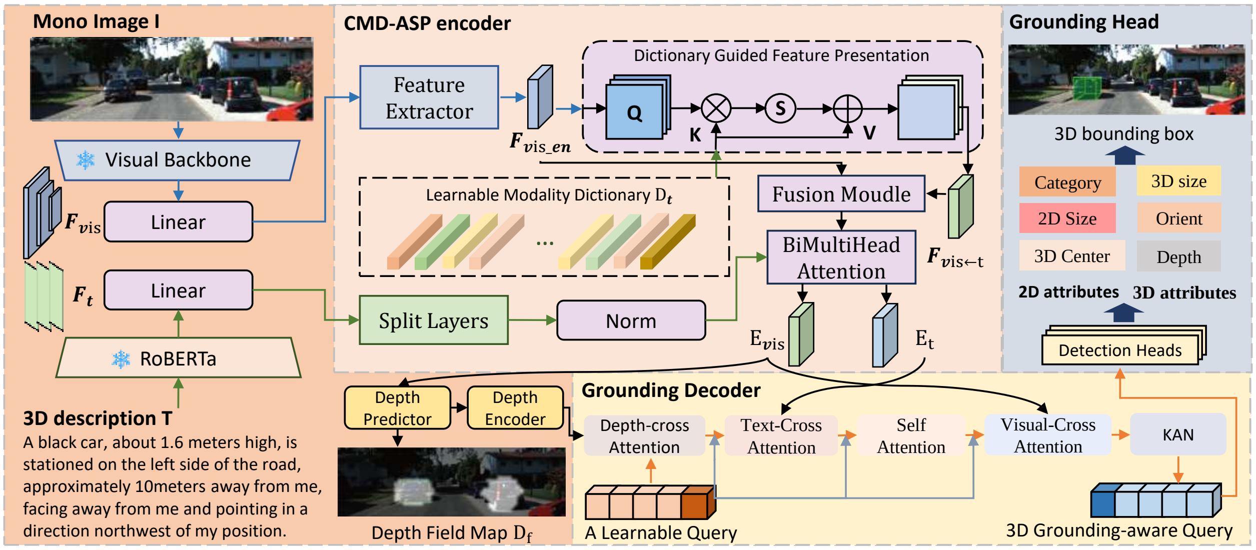
In recent years, 3D vision technology has received increasing attention[12]. For robot perception systems[2], acquiring 3D spatial information of the real world is extremely important. Accompanied by the development of Natural Language Processing[20], collaboration and communication between humans and robots in shared physical spaces become more natural and efficient, allowing robots to more accurately understand task scenarios and human intentions based on verbal instructions.

However, existing 3D vision-language fusion methods typically use a two-stage process: image registration first, followed by feature fusion[4]. This approach not only increases computational complexity and hardware demands but also significantly limits its application in robotic scenarios[7]. Current solutions, relying on high-quality synthetic data, often fail to achieve robust performance in dynamic environments. To address these issues, we propose Mono3D-VLDL, a novel single-stage framework. It integrates image registration and feature fusion into one stage, achieving SOTA performance on specific evaluation components of the Mono3DRefer test set[20], where the vehicle-mounted perspective data align with outdoor robot visual perception images, further demonstrating its superiority in multimodal fusion and 3D object localization. Moreover, it designs a learnable modality dictionary for cross-modal alignment and uses parallel cross-attention mechanisms to integrate depth, text, and visual information, achieving efficient and robust 3D object attribute prediction[1]. Our key contributions are as follows:

- **Cross-Modality Dictionary Alignment**: A CMD-ASP encoder with a learnable dictionary compensates for missing textual information while preserving modality-specific features.
- **3D Grounding-Aware Query**: Parallel cross-attention mechanisms integrate depth, text, and visual information for accurate object attribute prediction.
- **SOTA Performance**: Achieving state-of-the-art (SOTA) performance on specific evaluation components of the Mono3DRefer test set, demonstrating superior capability in multimodal fusion and 3D object localization tasks.

In summary, the application prospects of 3D vision-language fusion technology in robotic perception systems are broad. However, existing methods have many limitations. The proposed Mono3D-VLDL framework[22], with its innovative single stage design, effectively addresses these issues and demonstrates superior performance in multimodal fusion and 3D object localization.

[*] Corresponding author



A black car, about 1.6 meters high, is stationed on the left side of the road, approximately 10meters away from me, facing away from me and pointing in a direction northwest of my position.

Figure 1. The proposed framework processes monocular images (blue lines) and 3D text descriptions (green lines), with black lines indicating fusion pathways. Initially, visual and textual inputs are encoded using Swin Transformer[14] and RoBERTa[13] backbones, respectively, extracting spatial-visual and geometric-semantic features. The CMD-ASP encoder then aligns these modalities via a learnable cross-modal dictionary, enabling fusion-aware feature integration. During decoding, depth embeddings from depth maps are fused with cross-modal features, followed by cross- and self-attention operations on a trainable query to produce a 3D grounding-aware query. Finally, the Grounding Head processes this query through a detection module to jointly predict 2D and 3D object attributes.

2. Method

In Figure 1, we illustrate that Mono3D-VLDL comprises three primary components: the CMD-ASP encoder, the Grounding decoder, and the Grounding head.

2.1. CMD-ASP encoder

The Cross-Modality Dictionary Alignment Fusion Perception Encoder comprises multiple modules. The Feature Extractor module consists of three distinct feature extraction blocks: a Base block, a Visual Feature Extraction block, and a Perceptual Alignment Feature Extraction (EN) block. These blocks share the same architecture but differ in the number of layers. Each layer includes a convolutional layer with a kernel size of 3×3 and a stride of 1, followed by a batch normalization layer and a ReLU activation function layer. The visual features $F_{\text{vis}} \in \mathbb{R}^{B\times(H\times W)\times C}$ obtained through the Swin Transforme[14] are reshaped into $F_{\text{vis}} \in \mathbb{R}^{B\times H\times W\times C}$ to pass through the Feature Extractor module, obtaining features F_{vis_en} that complement shallow and deep features. Due to the inherent modality differences between text and images, aligning image text at the pixel level is significantly challenging. We employ a learnable modality dictionary[9] to compensate for the lack of textual modality information. It aims to preserve the unique characteristics of each specific modality as much as possible while ensuring that the features extracted from one modality contain relevant information from the other modality. As shown in the figure 1, D_t is a learnable modality dictionary, and W_{vis}^{Q} , W_{t}^{K} , W_{t}^{V} are linear mappings performed on F_{vis_en}. Therefore, we have $Q_{vis} = W_{\text{vis}}^{Q} F_{vis\_en}$, $K_t = W_{\text{t}}^{K} D_t$, and $V_t = W_{\text{t}}^{V} D_t$. With the assistance of D_t, $F_{vis\leftarrow t}$ can be expressed as:

$$F_{vis\leftarrow t} = \text{softmax}\left(\frac{Q_{\text{vis}} K_{\text{t}}}{\sqrt{d}}\right) V_{\text{t}} \qquad (1)$$

where $vis \leftarrow t$ denotes injecting textual information into visual features. After compensating with the modality dictionary, the generated features $F_{vis\leftarrow t}$ inject specific information missing from the textual modality. Subsequently, to prevent the generated features from losing the original features, we design a Fusion Model module that simply achieves a residual effect by fusing with the original features F_{vis}. In the proposed approach, the two image features are first combined through layer-wise summation, followed by a Layer LayerNorm operation to achieve the desired feature fusion effect. Finally, the obtained features are processed with BiAttention[11] along with the text feature input to generate fused-visual feature E_{vis} and fused-text feature E_t outputs that provide more comprehensive knowledge for the query.In this block, we employed BiMultiHeadAttention six times to enhance the fine-grained correlation between fused-text features and fused-visual features.

2.2. Grounding decoder

As shown in Figure 1, we define a learnable query $Q \in \mathbb{R}^{1\times C}$ as the 3D grounding-aware query for detection. We also design a sequence consisting of four decoder blocks, including depth-based cross-attention, text-based cross-attention, self-attention, vision-based cross-attention, and the final KAN[15]. To enable the query to better integrate depth, text, and visual information, we modify the structure to use these four decoder sequences in parallel. Specifically, the query first collects basic geometric features from depth prediction information through a depth-based cross-attention layer, and then combines with the original query in a residual-like manner. Subsequently, it integrates 3D description information through a text-based cross-attention layer to achieve the parallel use of the four decoders. Finally, in the same manner, the query goes through self-attention and vision-based cross-attention layers to combine positional and content queries, acquiring visual semantics from multiscale visual embeddings. In the output fusion phase of the four-way parallel decoder architecture, we introduce a state-of-the-art KAN module to replace the conventional MLP. This innovative substitution leverages the KAN module's learnable nonlinear functional structure to achieve more adaptive integration of multimodal features, specifically incorporating geometric, semantic, and visual modalities.

2.3. Grounding heads and loss

Referencing MonoDETR[21], The grounding head is based on learnable 3D grounding-aware queries that regress target attributes through a multibranchprediction head. As shown in Figure 1, the query vector $Q \in \mathbb{R}^{1\times C}$ output by the decoder is fed into the following prediction branches: 1) The 2D attribute branch includes a linear classification layer (for target category prediction), a 3-layer MLP (for regressing 2D bounding box parameters (l, r, t, b) representing the distance from the projected center to the four sides), and a 2-layer MLP (for predicting the projected 3D center coordinates $(x^{\mathrm{3D}}, y^{\mathrm{3D}})$); 2) The 3D attribute branch includes a 2-layer MLP (for regressing 3D dimensions $(h^{\mathrm{3D}}, w^{\mathrm{3D}}, l^{\mathrm{3D}})$), a 2-layer MLP (for orientation angle θ estimation), and a depth prediction module based on Laplace uncertainty modeling (refer to the method[21] to calculate the final depth value d). By fusing the above parameters with camera intrinsics, a complete 3D bounding box can be reconstructed. The supervision signal adopts a component-wise loss mechanism: the 2D supervision term L_{2D} integrates Focal Loss[10] for classification, L_1 loss for projected center, and L_1 + GIoU loss[16] for 2D calibration box; the 3D supervision term L_{3D} includes IoU alignment loss[16] for 3D dimension, Multi-Bin orientation loss[3], and depth uncertainty loss[5], achieving fine-grained optimization of geometric constraints.

3. Experiments

3.1. Dataset and metrics

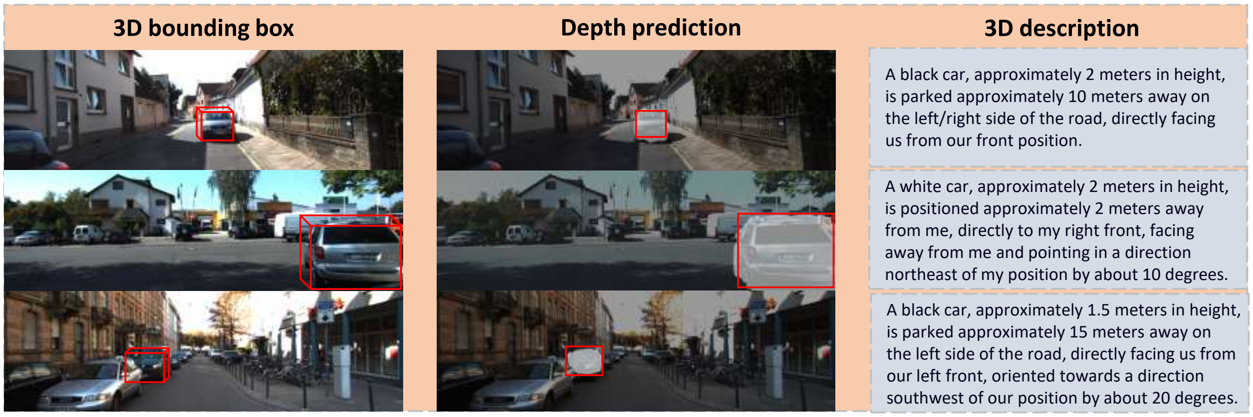
Our experimental evaluation is conducted on the widely adopted Mono3DRefer[20] benchmark, strictly following its official split strategy: the training set contains 29,990 samples, while the validation and test sets contain 5,735 and 5,415 samples, respectively. This dataset is derived from 2,025 frames of images from the KITTI[8] benchmark, covering a total of 41,140 annotated referring expressions and constructing a semantic dictionary containing 5,271 words. The experimental evaluation employs a multidimensional analysis framework: 1) The object distribution subset includes a "unique" subset (containing only a single instance within a category) and a "multiple" subset (containing multiple instances of the same category). 2) The spatial distribution stratification includes depth stratification (near (0-15 m), medium (15-30 m), far (30 m +)) and difficulty stratification (easy/medium/hard based on the degree of occlusion/truncation). The evaluation protocol is strictly aligned with the Mono3DRefer standard, using precision metrics with 3D Intersection over Union (3D IoU) thresholds of 0.25 and 0.50, denoted as Acc@0.25 and Acc@0.5, respectively. This stratified evaluation mechanism can comprehensively reflect the fine-grained performance of the model under different complexities of the scene.

3.2. Implementation details

We initialize the model using pre-trained weights from Mono3DVG-TR. A small portion of the weights from the unmodified sections are loaded to facilitate faster learning and convergence of the parameters of the modified encoder and decoder components. For the feature extraction part, the parameters of the RoBERTa[13] text encoder and SwinL[14] visual backbone are directly used as pre-trained weights without any further learning, which highlights the effect of the modified encoder component. On a single RTX 3090 GPU, we train the model for over 60 epochs with a batch size of 10 and an initial learning rate of 10^{-4}. We also set an initial warm-up period of 500 steps, with the learning rate decreasing according to a cosine function. This strategy stabilizes training by preventing large gradient updates during the early stages when the model is still highly unstable, ensuring smoother convergence.

3.3. Comparison With One-stage Methods

As shown in Tables 1 and 2, Mono3DVG-VLDL achieves state-of-the-art (SOTA) performance on specific evaluation components of the Mono3DRefer test set. Specifically, it demonstrates significant improvements in accuracy across various subsets, with increases of +9.08%, +0.91%, and +2.45% in Acc@0.25 for the "unique", "multiple", and "overall" subsets, respectively. Under Acc@0.5,



A black car, approximately 2 meters in height, is parked approximately 10 meters away on the left/right side of the road, directly facing us from our front position.

A white car, approximately 2 meters in height, is positioned approximately 2 meters away from me, directly to my right front, facing away from me and pointing in a direction northeast of my position by about 10 degrees.

A black car, approximately 1.5 meters in height, is parked approximately 15 meters away on the left side of the road, directly facing us from our left front, oriented towards a direction southwest of our position by about 20 degrees.

Figure 2. visualization of the three spatial partitions task (near/middle/far). Based on the 3D description, a batch of target objects can be screened out from depth map, and finally, the objects are locked on the monocular image in combination with the text information.

Table 1. Comparison with one-stage baselines. Best results are **bolded**.

Method	**Unique**		**Multiple**		**Overall**	
			Acc@0.25	**Acc@0.5**		
ZSGNet[17]+backproj	9.02	0.29	16.56	2.23	15.14	1.87
FAOA[18]+backproj	11.96	2.06	13.79	2.12	13.44	2.11
ReSC[19]+backproj	11.96	0.49	23.69	3.94	21.48	3.29
TransVG[6]+backproj	15.78	4.02	21.84	4.16	20.70	4.14
Mono3DVG-TR[20]	57.65	33.04	65.92	**46.85**	64.36	44.25
Mono3D-VLDL (Ours)	**66.73**	**41.37**	**66.83**	44.94	**66.81**	**44.26**

Table 2. Evaluation across three spatial partitions (near/medium/far) and three difficulty levels (easy/normal/hard). Best results are **bolded**.

Method	**Near/easy**		**Med/normal**		**Far/hard**	
	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
ZSGNet	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	18.03/17.51	0.53/3.43	15.64/12.18	3.95/1.34	4.86/8.83	0.62/0.09
ReSC	33.68/27.90	0.59/5.71	24.03/19.23	6.15/1.97	4.24/14.4	1.25/1.02
TransVG	29.34/28.88	0.86/6.95	25.05/16.41	8.02/2.75	4.17/12.9	0.97/1.38
3DVG	64.74/72.36	**53.5**/51.8	**75.44/69.23**	**55.5/48.7**	45.1/49.0	15.3/**29.9**
(Ours)	**69.28/75.97**	52.3/**53.9**	73.28/66.27	51.8/45.7	**53.2/54.4**	**23.0**/29.5

apart from a decrease in the Multiple metric, there were increases of 8.33% and 0.01% in the Unique and Overall metrics, respectively. Notably, in distant scenarios ($> 30m$), Mono3DVG-VLDL demonstrated significant performance improvements compared to Mono3DVG-TR, increasing from 45.1%/49% to 53.2%/54.4%. However, depth-sensitive analysis reveals that performance degrades for medium objects ($15 - 30m$), primarily due to accumulated errors in monocular depth estimation. In contrast, performance improves remarkably for near-distance objects ($< 15m$), with increases of +4.54% and +3.61% in Acc@0.25 and Acc@0.5, respectively. For near-distance, detection accuracy plays a dominant role, and depth prediction contributes minimally. The advantages of multimodal fusion are evident in challenging scenarios with severe occlusion (hard), where visual-textual feature fusion enhances Acc@0.25 by +7.7%, respectively. Key factors contributing to these improvements include a cross-modal enhancement mechanism that reduces depth prediction errors, through collaborative encoding of independent visual features and text-guided features, as well as a multimodal temporal attention module that effectively alleviates depth ambiguity. These results indicate that a joint visual-textual encoding architecture designed for complex monocular scenes can significantly enhance 3D spatial reasoning capabilities.

4. Conclusion

In this study, we introduce Mono3D-VLDL, a novel single-stage framework for language-visual fusion in robotic vision. By integrating image registration and feature fusion into one stage, it is highly suitable in resource-constrained scenarios. Our framework's learnable modality dictionary and parallel cross-attention mechanisms enable efficient and robust 3D object attribute prediction.

To validate its effectiveness, we conducted extensive experiments on the Mono3DRefer dataset, which closely resembles outdoor robot visual perception data. The results show that Mono3D-VLDL outperforms existing two-stage methods in efficiency and accuracy, particularly in dynamic and complex environments.

Looking ahead, we plan to enhance the framework's adaptability to dynamic environments by refining cross-modal alignment. Future work will focus on expanding the framework to support more complex robotic applications, including multi-object manipulation and scene understanding.

Acknowledgements

This work was supported by the Postdoctoral Fellowship Program of the China Postdoctoral Science Foundation (CPSF) under Grant No. GZC20241447, the National Key R&D Program of China under Grant No. 2023YFB4301800, and the Fundamental Research Funds for the Central Universities, Chang'an University (CHD) under Grant No. 300102244202.

References

[1] Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446):eaat8414, 2019. 1

[2] Torgny Brogårdh. Present and future robot control development—an industrial perspective. *Annual Reviews in Control*, 31(1):69–79, 2007. 1

[3] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision – ECCV 2020*, pages 202–221, Cham, 2020. Springer. 3

[4] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 172–181, 2023. 1

[5] Yongjian Chen, Lei Tai, Kaiyue Sun, and Ming Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12093–12102. IEEE, 2020. 3

[6] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1769–1779. IEEE, 2021. 4

[7] Zhizhao Duan, Hao Cheng, Duo Xu, Xi Wu, Xiangxie Zhang, Xi Ye, and Zhen Xie. Cityllava: Efficient fine-tuning for vlms in city scenario. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7180–7189, 2024. 1

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012. 3

[9] Huafeng Li, Zengyi Yang, Yafei Zhang, Wei Jia, Zhengtao Yu, and Yu Liu. Mulfs-cap: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3673–3690, 2025. 2

[10] 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 (ICCV)*, pages 2980–2988. IEEE, 2017. 3

[11] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jiao Yang, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[12] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, and Jie Yang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. arXiv preprint. 1

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2, 3

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 2, 3

[15] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov–arnold networks. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[16] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666. IEEE, 2019. 3

[17] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4694–4703. IEEE, 2019. 4

[18] 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/CVF International Conference on Computer Vision (ICCV)*, pages 4683–4693. IEEE, 2019. 4

[19] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision – ECCV 2020*, pages 387–404. Springer, 2020. 4

[20] Yi Zhan, Yizhen Yuan, and Zixiong Xiong. Mono3dvg: 3d visual grounding in monocular images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6988–6996, 2024. 1, 3, 4

[21] Rui Zhang, Haoyang Qiu, Tai Wang, Zhi Guo, Zhaoxiang Cui, Yu Qiao, et al. Monodetr: Depth-guided transformer for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9155–9166, 2023. 3

[22] Yunsong Zhou, Hongzi Zhu, Quan Liu, Shan Chang, and Minyi Guo. Monoatt: Online monocular 3d object detection with adaptive token transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17493–17503, 2023. 1