Iterative Robust Visual Grounding with Masked Reference based Centerpoint Supervision

Menghao Li[1] * Chunlei Wang[1] * Wenquan Feng[1] Shuchang Lyu[1]
Guangliang Cheng[2] ✉ Xiangtai Li[3] Binghao Liu[1] Qi Zhao[1] ✉

[1] Beihang University [2] University of Liverpool [3] S-Lab, Nanyang Technological University
{sy2102227, wcl_buaa, buaafwq, lyushuchang, liubinghao, zhaoqi}@buaa.edu.cn
{Guangliang.Cheng}@liverpool.ac.uk {xiangtai.li}@ntu.edu.sg

Abstract

Visual Grounding (VG) aims at localizing target objects from an image based on given expressions and has made significant progress with the development of detection and vision transformer. However, existing VG methods tend to generate ***false-alarm*** *objects when presented with inaccurate or irrelevant descriptions, which commonly occur in practical applications. Moreover, existing methods fail to capture fine-grained features, accurate localization, and sufficient context comprehension from the whole image and textual descriptions. To address both issues, we propose an Iterative Robust Visual Grounding (****IR-VG****) framework with Masked Reference based Centerpoint Supervision (MRCS). The framework introduces iterative multi-level vision-language fusion (IMVF) for better alignment. We use MRCS to ahieve more accurate localization with point-wised feature supervision. Then, to improve the robustness of VG, we also present a multi-stage false-alarm sensitive decoder (MFSD) to prevent the generation of false-alarm objects when presented with inaccurate expressions. Extensive experiments demonstrate that IR-VG achieves new state-of-the-art (SOTA) results, with improvements of 25% and 10% compared to existing SOTA approaches on the two newly proposed robust VG datasets. Moreover, the proposed framework is also verified effective on five* ***regular*** *VG datasets. Codes and models will be publicly at* https://github.com/cv516Buaa/IR-VG.

1. Introduction

Visual Grounding (VG) is a crucial computer vision task gaining significant attention due to its potential for enabling practical applications such as robot navigation [9] and visual dialog [30, 13]. VG aims to locate a target object within an image based on the given language reference expressions by incorporating information from both textual and visual modalities. However, existing VG methods suffer from false-alarm issues, where they assume that the referred object always exists in the image, leading to inaccurate or wrong targets being detected when irrelevant or inaccurate textual expressions are provided, shown in Fig. 1 (a).

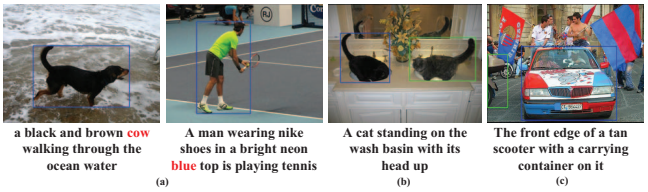

Figure 1. Illustration of the weakness of existing VG approaches. Green and blue boxes represent ground truths and predictions.

Previous works [15, 14, 24] have made significant progress in VG through various techniques. However, the task of cross-modal learning involved in the VG task remains challenging, and current approaches can be broadly divided into two main categories: two-stage methods [11, 19, 4, 32] and one-stage methods [38, 17, 35, 22, 27, 36]. Despite the significant achievements, the VG approaches suffer from some limitations, such as failing to capture the detailed feature representation accurately, resulting in a lack of discrimination between fine-grained objects with reference expressions shown in Fig. 1 (b), and detecting irrelevant or incorrect targets without understanding the whole context shown in Fig. 1 (c).

To address the above issues, this paper proposes a novel iterative robust visual grounding (IR-VG) approach with masked reference based centerpoint supervision. The approach first constructs two new robust VG datasets and proposes a multi-stage false-alarm sensitive decoder (MFSD) module to handle the case when there is no target object from the textual expression, avoiding generating false

*Contribute Equally.

alarms. Secondly, a new masked reference based centerpoint supervision (MRCS) module is proposed to capture the fine-grained feature and enhance the localization capacity from the given reference expressions. Finally, an iterative multi-level vision-language fusion (IMVF) module is leveraged to fuse multi-level visual and textual information that is crucial for vision-language understanding.

The contributions of our works are summarized as follows: firstly, the proposed approach handles the false-alarm issue in the VG task for the **first time** by constructing two new robust VG benchmarks and introducing a multi-stage false-alarm sensitive decoder (MFSD) module. Secondly, a new masked reference based centerpoint supervision (MRCS) module is proposed to achieve more accurate fine-grained features and better localization capacity from fully visual-textual comprehension. Lastly, the iterative multi-level vision-language fusion (IMVF) module is introduced to comprehensively fuse multi-level visual and textual information for better vision-language understanding and alignment. Extensive experiments on five *regular* VG benchmarks and two newly constructed *robust* VG benchmarks demonstrate the effectiveness of the proposed approach, achieving **10%** improvement on robust datasets.

2. Related Work

Visual Grounding. The Visual Grounding task is an important problem in computer vision that aims to localize an object within an image based on a given language reference expression. The existing approaches typically extend the object detection framework, such as YOLOV3 [25], Faster-RCNN [26], RetinaNet [18], CenterNet [8], and DETR [3], by incorporating a visual-linguistic fusion module. These approaches can be categorized into two main categories: two-stage methods [11, 19, 4, 32, 39] and one-stage methods [38, 17, 35, 22, 27, 36]. Two-stage approaches, including CMN [11], NMTree [19] and RefNMS [4], Two-branch Network [32] and MAttNet [39], due to the large number of proposals and matching process may slow down the inference speed. On the other hand, one-stage approaches [38, 17, 35, 22, 27, 36] directly incorporate the linguistic context into visual features to predict the object's location. However, it may not be flexible enough to achieve a global context understanding due to the pointwise feature representations. *Recently*, transformer-based Visual Grounding approaches have gained popularity due to their attention capacity and efficiency. For instance, TransVG [6] captures intra- and inter-modal contexts using transformers in a uniform manner, while VLTVG [34] builds discriminative feature maps and detects the target object through a multi-stage decoder.

Robustness in Visual Grounding. In terms of robustness in Visual Grounding, recent studies have explored CNN robustness in various benchmarks [10] [23], and some works have evaluated and improved CNN robustness for practical applications [29] [28] [1] [33]. RefSegformer [33] incorporates negative sentence inputs to handle false-alarm issues in referring segmentation tasks. However, to the best of our knowledge, no existing benchmarks or approaches have explored the robustness of the Visual Grounding task. This paper takes a further step by proposing a new iterative *robust* VG framework and building two robust VG datasets to address this research problem. It is important to note that, within the context of this paper, the term "**robust**" refers to the ability of the proposed method to produce accurate results and avoid false-alarm predictions even when provided with irrelevant and incorrect expressions.

3. Method

In this section, we present the architecture of the proposed robust VG pipeline and its components. Fig. 2 illustrates the pipeline.

3.1. Masked Reference

Masked reference augmentation. As illustrated in the down-left part of Fig. 2, we propose a text augmentation approach to generate diversified textual information given an input language expression. We employ the NLTK [2] tokenization strategy to extract lexical properties for each word, followed by masking one word in the text according to the well-designed rules (shown in the supplementary materials). This masking process is repeated thrice, achieving one full text and three masked texts in total. BERT [7] is then utilized to generate different textual embeddings for these sentences.

Visual-linguistic alignment. The proposed model, illustrated in Fig. 3, incorporates a visual-linguistic alignment module with two consecutive multi-head attention (MHA) [31] layers. The visual feature map F_v is input as the *Query*, and the textual embeddings are input as *Key* and *Value* to the first MHA. This process produces an enhanced feature map that gathers relevant semantic information from the corresponding linguistic representation. Subsequently, the enhanced feature map undergoes another MHA operation that performs self-attention on the visual features to encode the involved visual contexts. The features from the two MHAs are element-wisely summed in a residual manner for the centerpoint supervision component.

Centerpoint supervision. To obtain the final centerpoint heatmaps, the summed feature map obtained from each language expression is processed through two consecutive convolutional layers. Multiple centerpoint heatmaps (one from the full text and three from the masked text) are then fused by performing a *maxpooling* process, with the centerpoint coordinates determined by performing a *max* operation on the resulting heatmap. The cross entropy loss is then utilized as the supervision loss between the centerpoint heatmap and the corresponding ground truth.

3.2. Iterative Multi-level Vision-language Fusion

Multi-level textual feature enhancement. The multi-level textual feature enhancement (MTFE) module improves tex-

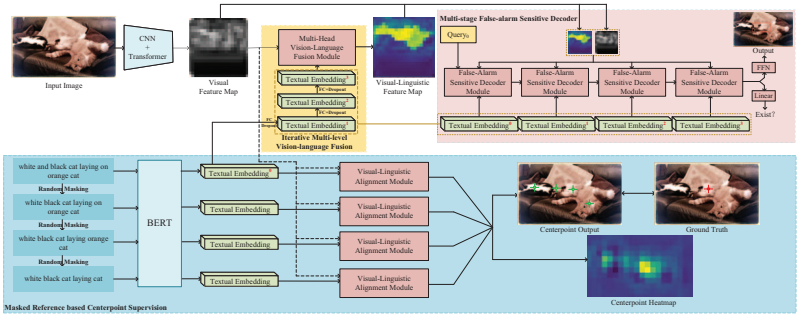

Figure 2. An overview of our proposed IR-VG framework, which comprises Masked Reference based Centerpoint Supervision, Iterative Multi-Level Vision-Language Fusion, and Multi-Stage False-Alarm Sensitive Decoder.

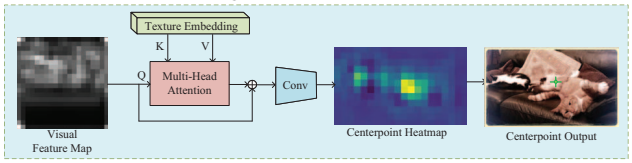

Figure 3. The architecture of visual-linguistic alignment module.

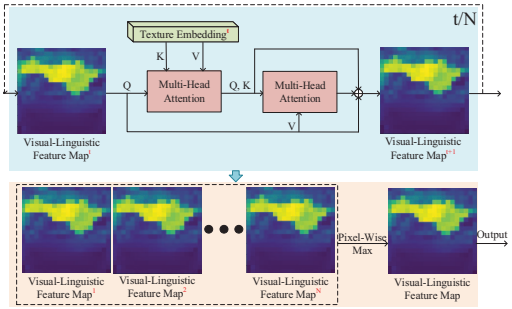

Figure 4. The architecture of IMVF.

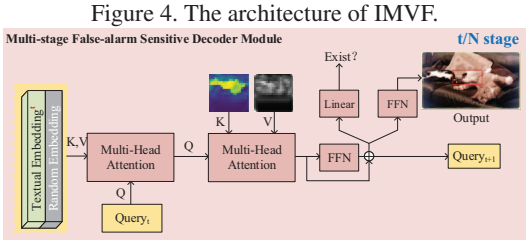

Figure 5. The architecture of MFSD.

tual embedding representation by performing two consecutive fully-connected layers with 768 nodes in each stage. Specifically, as highlighted with yellow color in Fig. 2, the IMVF comprises four stages, and each stage contains an MTFE module. The MTFE module consists of two fully connected layers and a corresponding dropout layer with a 0.1 ratio, aimed at obtaining multi-level textual features that match the multi-level visual features. This enables the model to focus on different key descriptions in the referring expressions and obtain more complete and reliable features for the referred object.

Iterative multi-level vision-language fusion. Fig. 4 illustrates the IMVF module, which is based on MHA and consists of four iterative stages. Each stage includes two MHA layers. The first layer uses the visual feature map $F_v \in \mathcal{R}^{C\times H\times W}$ as the *Query* and the textual embeddings $F_l \in \mathcal{R}^{C\times L}$ from the multi-level textual feature enhancement module as the *Key* and *Value*. Multi-head cross-attention enables the comprehensive incorporation of textual information into the visual feature map $F_g \in \mathcal{R}^{C\times H\times W}$. In the second layer, F_g serves as both the *Query* and *Key*, while F_v serves as the *Value*. This self-attention operator allows the model to gather crucial context features for the referred object based on the textual descriptions provided, and the final feature is $F_c \in \mathcal{R}^{C\times H\times W}$. We sum the F_v, F_g, and F_c element-wisely to obtain the final visual feature map F_m. In each iteration, the i-th visual feature map F_m^i becomes the initial feature map (i.e., F_v^{i+1}). Our experiments include four iterations, and we use element-wise *max* strategy to obtain the final fusion feature $F = \max(F_m^1, F_m^2, F_m^3, F_m^4)$.

3.3. Multi-stage False-alarm Sensitive Decoder

Multi-stage false-alarm sensitive decoder. As shown in Fig. 5, the MFSD module consists of several iterative stages, each containing two consecutive MHA layers. In the first stage, we randomly initialize a series of learnable queries. We introduce a random embedding with the same size as textual embedding from the IMVF module to handle the false-alarm case. We concatenate the textual and random embedding in the batch dimension, termed *mixture embedding*. For the first MHA layer, the learnable queries serve as *Query*, and the mixture embedding acts as *Key* and *Value*. With this layer, the textual embedding can be more easily attended to the target tokens, thus achieving enhanced textual embedding. For the second MHA layer, the enhanced textual embedding is treated as *Query*, and the visual-linguistic feature map from the IMVF module as well as the visual feature map F_v are employed as *Key* and *Value*. Through the second MHA layer, the textual information can be comprehensively fused with the visual feature map to achieve an enhanced vision-language feature, which is then taken into a feed-forward network (FFN). We fuse the enhanced vision-language feature and the feature from the second MHA in a residual manner, termed as R_feature, which serves as the *Query* in the next iteration. Then, R_feature is taken into two decoupled heads: one for classification to indicate whether there exist false-alarm results and another for regression to generate the predicted bounding boxes (bbox).

4. Experiment

4.1. Experimental Settings

Datasets. To validate the efficacy of the proposed approach, we assess its performance on two distinct types of datasets, namely the *regular* VG datasets and the *robust* VG datasets. The experimental settings of regular VG datasets are identical to [34]. Given the absence of existing robust VG datasets for evaluating the robustness of existing approaches, we create two new robust VG datasets, RefCOCOg_F and ReferItGame_F, by leveraging existing benchmarks [21, 16]. Our

Method	RefCOCO			RefCOCO+			RefCOCOg		ReferItGame	Flickr30k
	val	testA	testB	val	testA	testB	val-u	test-u	test	test
CMN [11]	-	71.03	65.77	-	54.32	47.76	-	-	28.33	-
VC [40]	-	73.33	67.44	-	58.40	53.18	-	-	31.13	-
NMTree [19]	76.41	81.21	70.09	66.46	72.02	57.52	65.87	66.44	-	-
Ref-NMS [4]	80.70	84.00	76.04	68.25	73.68	59.42	70.55	70.62	-	-
FAOA [38]	72.54	74.35	68.50	56.81	60.23	49.60	61.33	60.36	60.67	68.71
LBYLNet [12]	79.67	82.91	74.15	68.64	73.38	59.49	-	-	67.47	-
TransVG [6]	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73	70.73	79.10
VLTVG [34]	84.77	87.24	80.49	74.19	78.93	65.17	76.04	74.98	71.98	79.84
IR-VG (Ours)	**86.82**	**88.75**	**82.60**	**76.22**	**80.75**	**67.33**	**77.86**	**76.24**	**74.03**	**81.45**

Table 1. Comparisons with SOTA visual grounding methods.

Methods	RefCOCOg_F		ReferItGame_F	
	R_{fad}	R_{mix}	R_{fad}	R_{mix}
CMN [11]	27.10	65.10	24.75	21.41
VC [40]	42.45	68.85	31.03	25.69
SSG [5]	34.15	61.25	32.44	46.43
Ref-NMS [4]	43.90	62.40	41.39	48.15
ReSC-Large [35]	37.35	60.55	32.54	59.89
LBYLNet [12]	45.40	63.32	45.40	60.57
IR-VG (Ours)	**67.32**	**73.61**	**69.44**	**72.03**

Table 2. Comparisons with SOTA approaches on ***robust*** VG datasets.

Methods		RefCOCO			RefCOCO+			RefCOCOg	
I	M	val	testA	testB	val	testA	testB	val-u	test-u
-	-	84.77	87.24	80.49	74.19	78.93	65.17	76.04	74.98
✓	-	85.92	88.41	81.77	75.27	80.66	66.33	77.10	76.06
-	✓	85.53	88.09	81.23	75.34	79.97	66.18	77.21	75.75
✓	✓	**86.82**	**88.75**	**82.60**	**76.22**	**80.75**	**67.33**	**77.86**	**76.24**

Table 3. Ablation studies on three benchmarks, "I" and "M" denote IMVF and the MRCS.

train set is composed of two parts: the first being the original train set of the respective dataset and the second being a random matching dataset that disrupts the correspondence between image information and language descriptions.

Evaluation Metrics. For the ***regular*** VG dataset, following previous works [6] [37], we adopt the commonly used top1 accuracy (acc-1) as the evaluation metric. For the ***robust*** VG dataset, we propose two novel evaluation metrics, i.e., false alarm discovery rate R_{fad} with only false-alarm data, and correct rate among the mixed data R_{mix} with both false-alarm and regular data, which are defined as,

$$R_{\text{fad}} = \frac{\text{FA}^{\text{acc}}}{\text{FA}^{\text{all}}}, \quad R_{\text{mix}} = \frac{\text{FA}^{\text{acc}} + \text{Regular}^{\text{acc}}}{\text{FA}^{\text{all}} + \text{Regular}^{\text{all}}}, \tag{1}$$

where FA denotes the false-alarm data with irrelevant or inaccurate descriptions, and Regular means the regular data with accurate descriptions. The superscript **acc** and **all** represent the number of accurate predictions and the total number of data. The detailed dataset descriptions, training loss, and other experiment implementation details will be shown in the supplementary materials.

4.2. Comparisons with Existing SOTA Methods

Results on VG benchmarks. As presented in Tab. 1, we evaluate the proposed approach against other SOTA VG methods. We improve over the best SOTA approaches by about 2% in all five benchmarks, indicating the effectiveness of our proposed method.

Results on robust VG benchmarks. Tab. 2 demonstrates the numerical comparisons on the ***robust*** VG datasets. In particular, we improve over the SOTA approaches by a nontrivial margin in competitive benchmarks of RefCOCOg_F and ReferItGame_F. Specifically, on ReferItGame_F dataset, we achieve about 25% and 10% improvement in R_{fad} and R_{mix} metrics, respectively.

4.3. Ablation Study

Numerical Component Analysis. Tab. 3 shows the effectiveness of each component on the *regular* VG datasets.

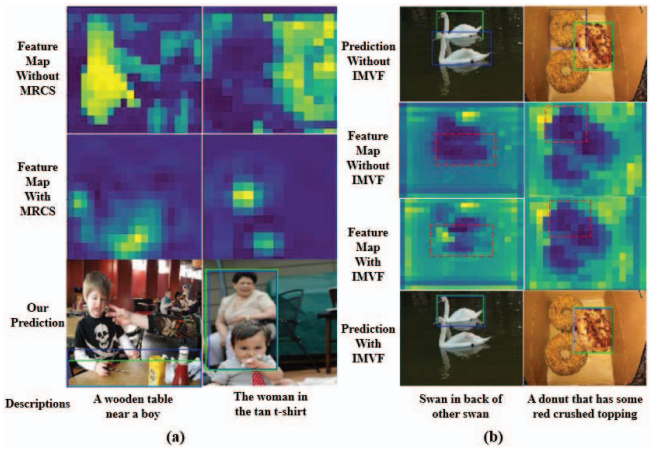

Figure 6. Visualization Results. (a) Feature map with/without MRCS module. (b) Feature map with/without IMVF module, especially for the red rectangle areas.

The proposed approach outperforms the baseline by 2.1% top1 accuracy in the RefCOCO testB dataset. Specifically, IMVF improves by 1.3%, and MRCS improves by 0.7%. Similar conclusions can be drawn from other *regular* VG datasets. Tab. 2 illustrates the effectiveness and robustness of the proposed MFSD module, which significantly improves two robust VG benchmarks.

Qualitative Component Analysis. *Qualitative analysis of MRCS.* Fig. 6(a) presents the visual-linguistic feature map with or without MRCS module. We intuitively observe that the MRCS enables the feature map to attend more accurately to the target object's location and generates a more precise foreground map. To avoid interactions with IMVF module, we conduct this experiment only with MRCS module and MFSD module. *Qualitative analysis of IMVF.* Fig. 6(b) illustrates the visual-linguistic feature map with or without IMVF module. The figure indicates that the IMVF module reduces interference and allows the model to concentrate more on the target by better understanding visual and textual information. To ensure fairness, we performed this experiment only with IMVF and MFSD modules.

5. Conclusions

Our work introduces the IR-VG framework, which comprises IMVF, MRCS, and MFSD. It outperforms existing approaches in terms of context features, fine-grained features, and localization accuracy while addressing robustness issues when faced with irrelevant or inaccurate reference expressions. Our experiments demonstrate the effectiveness of each module, achieving new SOTA performance.

Limitation and future work. Notably, IR-VG builds a new research direction for robust VG. Future work includes developing a more elegant framework to handle false alarms. In addition, we will explore the false-alarm problems with irrelevant expressions for some foundation models (e.g. Grounding DINO [20]).

References

[1] Said Fahri Altindis, Yusuf Dalva, and Aysegul Dundar. Benchmarking the robustness of instance segmentation models. *CoRR*, 2021. 2

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly, 2009. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[4] Long Chen, Wenbo Ma, Jun Xiao, and et al. Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding. In *AAAI*, 2021. 1, 2, 4

[5] Xinpeng Chen, Lin Ma, Jingyuan Chen, and et al. Real-time referring expression comprehension by single-stage grounding network. *CoRR*, 2018. 4

[6] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, and et al. Transvg: End-to-end visual grounding with transformers. In *CVPR*, 2021. 2, 4

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 2

[8] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, 2019. 2

[9] Zipeng Fu, Ashish Kumar, Ananye Agarwal, and et al. Coupling vision and proprioception for navigation of legged robots. In *CVPR*, 2022. 1

[10] Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, and et al. Generalisation in humans and deep neural networks. In *NeurIPS*, 2018. 2

[11] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, and et al. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 2017. 1, 2, 4

[12] Binbin Huang, Dongze Lian, Weixin Luo, and et al. Look before you leap: Learning landmark features for one-stage visual grounding. In *CVPR*, 2021. 4

[13] Xiaoze Jiang, Jing Yu, Yajing Sun, and et al. DAM: deliberation, abandon and memory networks for generating detailed and non-repetitive responses in visual dialogue. 2020. 1

[14] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1

[15] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. 2014. 1

[16] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 3

[17] 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 *CVPR*, 2020. 1, 2

[18] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2

[19] Daqing Liu, Hanwang Zhang, Zheng-Jun Zha, and et al. Learning to assemble neural module tree networks for visual grounding. In *CVPR*, 2019. 1, 2, 4

[20] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4

[21] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 3

[22] Yue Ming, Nannan Hu, Chunxiao Fan, and et al. Visuals to text: A comprehensive review on automatic image captioning. *IEEE/CAA Journal of Automatica Sinica*, 2022. 1, 2

[23] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and et al. Exploring generalization in deep learning. In *NeurIPS*, 2017. 2

[24] Ahmad Ostovar, Suna Bensch, and Thomas Hellström. Natural language guided object retrieval in images. *Acta Informatica*, 2021. 1

[25] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, 2018. 2

[26] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *PAMI*, 2017. 2

[27] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *ICCV*, 2019. 1, 2

[28] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 2018. 2

[29] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *CVPR*, 2019. 2

[30] Kaili Sun, Chi Guo, Huyin Zhang, and et al. HVLM: exploring human-like visual cognition and language-memory network for visual dialog. *Information Processing & Management*, 2022. 1

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[32] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *PAMI*, 2019. 1, 2

[33] Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards robust referring image segmentation. *CoRR*, 2022. 2

[34] Li Yang, Yan Xu, Chunfeng Yuan, and et al. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *CVPR*, 2022. 2, 3, 4

[35] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and et al. Improving one-stage visual grounding by recursive sub-query construction. In *ECCV*, 2020. 1, 2, 4

[36] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and et al. Improving one-stage visual grounding by recursive sub-query construction. In *ECCV*, 2020. 1, 2

[37] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020. 4

[38] Zhengyuan Yang, Boqing Gong, Liwei Wang, and et al. A fast and accurate one-stage approach to visual grounding. In *CVPR*, 2019. 1, 2, 4

[39] 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 *CVPR*, 2018. 2

[40] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *CVPR*, 2018. 4