Visual Grounding With Joint Multimodal Representation and Interaction

Hong Zhu, Qingyang Lu, Lei Xue, Mogen Xue, Guanglin Yuan, and Bineng Zhong

***Abstract*— This article tackles the challenging yet significant task of grounding a natural language query to the corresponding region onto an image. The main challenge in visual grounding is to model the correspondence between visual context and semantic concept referred by the language expression, i.e., multimodal fusion. Nevertheless, there is an inherent deficiency in the current fusion module designs, which makes visual and linguistic feature embeddings cannot be unified into the same semantic space. To address the issue, we present a novel and effective visual grounding framework based on joint multimodal representation and interaction (JMRI). Specifically, we propose to perform image–text alignment in a multimodal embedding space learned by a large-scale foundation model, so as to obtain semantically unified joint representations. Furthermore, the transformer-based deep interactor is designed to capture intramodal and intermodal correlations, rendering our model to highlight the localization-relevant cues for accurate reasoning. By freezing the pretrained vision-language foundation model and updating the other modules, we achieve the best performance with the lowest training cost. Extensive experimental results on five benchmark datasets with quantitative and qualitative analysis show that the proposed method performs favorably against the state-of-the-arts.**

***Index Terms*— Cross-modal interaction, feature alignment, image–text foundation model, visual grounding.**

I. Introduction

VISUAL grounding aims at predicting the location of an object referred by a language expression onto an image, also known as referring expression grounding or referring expression comprehension. Intersecting with natural language processing and visual understanding, it plays an important role in various applications, such as autonomous driving [1], [2], intelligent detection [3], [4], instrument monitoring [5], vision-language navigation [6], and cross-modality retrieval [7], [8].

Visual grounding poses a significant challenge, which requires not only a comprehensive understanding of intricate language semantics and diverse image contents but also more importantly, the establishment of multimodal semantic correspondences. The key to identifying the target object from others lies in the corresponding relationship between its visual content and the semantic concept described by the language. In other words, the core of visual grounding is multimodal fusion.

Early multimodal fusion methods are simplistic. Fast and accurate one-stage approach (FAOA) [9] employs a straightforward approach by concatenating the language embeddings with visual features for fusion, and the Similarity Net [10] uses an MLP to measure the similarity between vision and language. Although useful, these simple strategies do not yield satisfactory results, especially when dealing with lengthy and intricate language expressions. Subsequent studies have proposed various fusion methods to improve the grounding performance, including modular attention network [11], multimodal tree [12], [13], scene graph [14], [15], and query decomposition [16]. These heuristic works can often be complex and prone to overfitting specific data distributions. To address this, TransVG [17] and TransVG++ [18] employ a simple yet effective strategy by stacking multiple transformer encoder layers together to achieve platitudinous interaction between vision and language contexts. Visual grounding with transformers (VGTR) [19] introduces a text-guided self-attention module to facilitate the learning of visual features under the guidance of the language expression, improving the accuracy of grounding the referred object. Despite using a transformer to enhance multimodal correlation and inference effectiveness, vision and language are treated independently and processed distantly until later fusion. Pretraining of visual and linguistic encoders is completely separate, and each has its own encoding structure and dedicated training data. Although this fusion may easily get two modalities connected, the resulting visual and linguistic features cannot be unified into the same semantic space, leading to a lower and upper bound for grounding. Therefore, we argue it is necessary to perform the intermodal alignment before deep fusion.

In recent years, the foundation model pretraining [20], [21], [22] has been researched for mass downstream tasks,

Manuscript received 5 July 2023; revised 6 September 2023; accepted 26 September 2023. Date of publication 13 October 2023; date of current version 8 November 2023. This work was supported in part by the Natural Science Foundation of Anhui Province under Grant 2008085QF325 and Grant 2008085QF314, in part by the National University of Defense Technology through the Scientific Research Project under Grant zk19-15, in part by the Foundation of Army Artillery and Air Defense Academy of PLA under Grant PFXY220114001, and in part by the Anhui Province Laboratory of Advanced Laser Technology under Grant ky19c604. The Associate Editor coordinating the review process was Dr. Bineng Zhong. *(Corresponding author: Guanglin Yuan.)*

Hong Zhu is with the College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China, and also with the Anhui Key Laboratory of Polarization Imaging Detection Technology, Army Artillery and Air Defense Academy of PLA, Hefei 230031, China (e-mail: candy_zhuhong@126.com).

Qingyang Lu and Guanglin Yuan are with the Army Artillery and Air Defense Academy of PLA, Hefei 230031, China (e-mail: lqy465813@163.com; yuangl_plus@126.com).

Lei Xue is with the College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China (e-mail: eeixuelei@163.com).

Mogen Xue is with the Anhui Key Laboratory of Polarization Imaging Detection Technology, Army Artillery and Air Defense Academy of PLA, Hefei 230031, China (e-mail: xuemogen@126.com).

Bineng Zhong is with the School of Computer Science and Engineering, Guangxi Normal University, Guilin 541004, China (e-mail: bnzhong@gxnu.edu.cn).

Digital Object Identifier 10.1109/TIM.2023.3324362

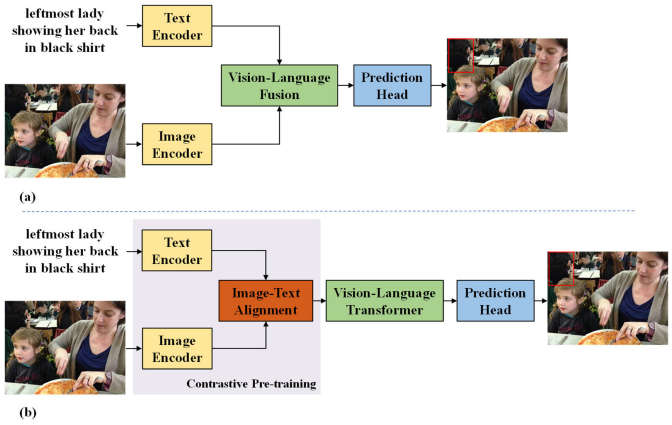


Fig. 1. Comparison of (a) existing works and (b) our JMRI framework. As shown in (a) that the existing works typically adopt the strategy of independent feature extraction and then fusion. In contrast, as shown in (b), JMRI performs feature alignment in a common semantic space learned by jointly contrastively training the dual-encoder and projection layer, before fusion.

aiming to amortize training costs and improve the performance and efficiency of task-specific models. Contrastive language-image pretraining (CLIP) [21] has shown that a dual-encoder model pretrained on large amounts of image–text pairs using contrastive objectives can encode visual and linguistic embeddings to a unified semantic space. The joint representations generated by the pretrained CLIP model exhibit strong semantics, allowing CLIP to seamlessly transfer to downstream zero-shot image classification and image–text retrieval. Nonetheless, it is not directly applicable for visual grounding, due to missing joint components to learn features with localization information.

To address the aforementioned issues, this article presents a joint multimodal representation and interaction framework for visual grounding, called JMRI, which is based on the image–text foundation model and transformer. As shown in Fig. 1, different from the existing works applying independent feature extraction and then fusion, our framework first performs feature alignment in a multimodal semantic space learned by CLIP. CLIP gains the ability of intermodal alignment by jointly contrastively training the dual-encoder and projection layer for image–text matching. Our multimodal fusion consists of two parts. An early fusion module is designed to encode features from both modalities to the same semantic space for alignment, yielding joint multimodal representations. We further add the transformer-based deep fusion to early fuse information and to capture the cross-modal correlation between the referring expression and visual region for localization. The resulting object-level and language-aware visual features enable our JMRI to locate the referred object more accurately.

In summary, the main contributions of this work are as follows.

1) We propose a novel and effective visual grounding framework by combining early joint representation and deep cross-modal interaction.
2) We propose to use CLIP to extract and align visual and linguistic features, ensuring that the resulting features are in the same semantic space, which is beneficial for the subsequent cross-modal fusion.
3) By freezing the pretrained CLIP and updating the other modules, we achieve the best performance with the least training budget and deployment cost.
4) We conduct comprehensive experiments to verify the benefits of the proposed method and achieve leading results on five prevalent benchmarks.

II. Related Work

A. Visual Grounding

Recent advancements in visual grounding can be broadly classified into two main branches: the propose-and-rank method and the proposal-free method. We give a brief overview of each branch in the following.

The propose-and-rank method first generates a set of region proposals from the input image by unsupervised methods or pretrained object detectors, and then scores the candidates based on the language expression and selects the top-ranked one. It can be roughly grouped into five categories: convolutional neural network-long short term memory (CNN-LSTM)-based model [23], [24], modular network [11], [25], attention-based model [26], [27], graph-based model [14], [15], and language parser model [12], [28]. The proposal-and-rank method benefits from explicit region proposals, which is useful in complex scenes or when dealing with multiple objects. However, it may face challenges in terms of the performance of the proposal generators and the additional ranking mechanisms.

The proposal-free method directly predicts the bounding box of the referred object without generating region proposals. It is still in its infancy, and the related work is much less than the previous one. We classify the present work into four categories: object detection-based model [9], visual feature optimization model [29], [30], textual feature optimization model [16], [31], and transformer-based model [17], [19]. The proposal-free method avoids the dependency on explicit proposals and captures more complicated relationships between vision and language. As a result, this method has demonstrated significant potential in terms of accuracy and efficiency, and it is gradually becoming the major framework in visual grounding.

B. Image–Text Foundation Model

Recently, there has been a growing trend to develop image–text foundation models for vision-language pretraining and visual recognition, where visual models are trained with natural language supervision. Two notable models in this field are CLIP [21] and ALIGN [22], which perform multimodal contrastive learning on hundreds of millions of image–text pairs. By leveraging this approach, they successfully learn powerful image and text representations that can be directly used for intermodal alignment and open-vocabulary image classification. Vision and language knowledge distillation (ViLD) [32] proposes a two-stage detector that incorporates the knowledge distilled from the CLIP/ALIGN model for zero-shot object detection. Grounded language-image pretraining (GLIP) [33] combines object detection with phrase grounding for pretraining. The learned representations exhibit strong transferability to different object-level recognition tasks, even

in zero-shot and few-shot scenarios. In order to improve the classification accuracy, locked-image tuning (LiT) [34] and BASIC [35] adopt a two-step training approach. They first pretrain the models on large numbers of image annotation datasets using cross-entropy loss, and then fine-tune on noisy image–text datasets with contrastive loss. The combination of pretraining and fine-tuning improves the generalization ability of the model and achieves better performance on image classification.

C. Transformer

The first introduction of the transformer in [36] revolutionizes the field of neural machine translation by enabling more effective sequence transformation. In simple terms, a transformer is an architecture that utilizes attention-based encoders and decoders to process the input sequence and generate the output one. The attention mechanism plays a critical role in determining the importance of different parts of the input sequence. Due to its remarkable ability to process long sequences, the transformer has been extended to handle vision and vision-language tasks.

DEtection TRansformer (DETR) [37] is an influential work that treats object detection as a set prediction problem and incorporates an encoder–decoder as the detection head. Similar to natural language processing, vision transformer (ViT) [38] interprets the image as a sequence of patches and processes them using a standard transformer encoder. Its extensions have demonstrated the effectiveness of the pure transformer architecture in numerous vision tasks [39], [40], [41]. The later swin transformer [42] introduces inductive bias to the vision transformer by designing shifted-window attention.

Inspired by the powerful model of bidirectional encoder representations from transformers (BERT) [43], researchers have begun to investigate vision-language pretraining [44], [45], [46], which aims to jointly encode visual and linguistic information in a fusion model. Typically, these models take vision and language as inputs, and design multiple transformer encoder layers for joint representation learning.

III. Proposed Method

This article formulates visual grounding as a bounding box regression problem. Given an image $\boldsymbol{I}$ and a language expression $\boldsymbol{T}$ as inputs, we learn a model F parameterized by Θ to locate a target region from the image with a bounding box $b = [x, y, w, h]$, where x, y, w, and h denote the center point coordinates, width and height of the referred object, corresponding to the expression $\boldsymbol{T}$ semantically

$$b = F(\boldsymbol{I}, \boldsymbol{T}; \Theta). \tag{1}$$

The framework of the proposed JMRI is shown in Fig. 2. Given an image and a language expression, we first feed them into a dual-encoder model to extract feature representations of each modality, and then, linearly project the features into a learned multimodal embedding space for image–text alignment. The resulting joint representations serve as inputs for later fusion. We design a cross-modal interactor to model intramodal and intermodal contexts with the attention mechanism. Finally, the contextual enriched features are leveraged to regress the spatial coordinate of the referred object in the prediction head.

Concretely, JMRI is composed of three modules.

1) *Early Joint Representation (Section III-A):* This module aims to learn joint multimodal representations from each modality in a common embedding space.
2) *Deep Cross-Modal Interaction (Section III-B):* This module focuses on capturing the semantic correlation between vision and language modalities for accurate reasoning.
3) *Prediction Head (Section III-C):* This module is dedicated to performing bounding box coordinates regression with keypoint estimation.

A. Early Joint Representation

Following the CLIP architecture, there are three main components in this module: an image encoder, a text encoder, and a linear projection layer, as shown in Fig. 2 (left). For the image encoder, we first resize the whole image and then forward it into a ViT [38] to yield a set of visual tokens, denoted by $z_I \in R^{N_I \times D_I}$, where N_I is the number of tokens, and one token has size D_I. For the text encoder, we tokenize the text, and then extract the textual tokens by a language transformer [47], $z_T \in R^{N_T \times D_T}$. It is a 12-layer transformer model with eight attention heads.

After extracting the feature representations of each modality, we employ a linear transformation layer to project them into a learned multimodal embedding space, i.e., $\boldsymbol{p}_I = \text{Embedding}(z_I)$ and $\boldsymbol{p}_T = \text{Embedding}(z_T)$. We denote the projected visual embeddings and textual embeddings as $\boldsymbol{p}_I \in R^{N_I \times D_P}$ and $\boldsymbol{p}_T \in R^{N_T \times D_P}$. The projection space is learned by CLIP using contrastive pretraining. During pretraining, given N image–text pairs, CLIP jointly trains the dual-encoder and projection layer to maximize the cosine similarity between the embeddings of the N real pairs while minimizing the cosine similarity between the embeddings of the $N^2 - N$ incorrect pairs. Consequently, the early fusion between vision and language is achieved by the dot product in a learned contrastive embedding space.

B. Deep Cross-Modal Interaction

In the vision-language literature [33], [45], [49], it has been shown that a deep fusion of visual and linguistic features is necessary to develop a high-performing grounding model. After the previous module fuses vision and language only at the last projection layer, we further introduce deep fusion to enhance the ego-information and cross-modal interaction, as shown in Fig. 2 (middle).

This module includes two types of components: intramodal interaction (IMI) and cross-modal interaction (CMI). As shown in the green bottom box in Fig. 2 (middle), IMI aims to adaptively enhance useful ego-information for the grounding task by multihead self-attention. To be specific, let the specific tokens $\boldsymbol{f}_X \in R^{N_X \times D}$ of modality X as inputs of multihead attention (MHA), namely, queries $\boldsymbol{f}_X^q$, keys $\boldsymbol{f}_X^k$, and values $\boldsymbol{f}_X^v$, the procedure in the IMI is

$$\boldsymbol{f}_{\text{IMI}} = \text{LN}\left(\boldsymbol{f}_X + \text{MHA}\left(\boldsymbol{f}_X^q, \boldsymbol{f}_X^k, \boldsymbol{f}_X^v\right)\right) \tag{2}$$

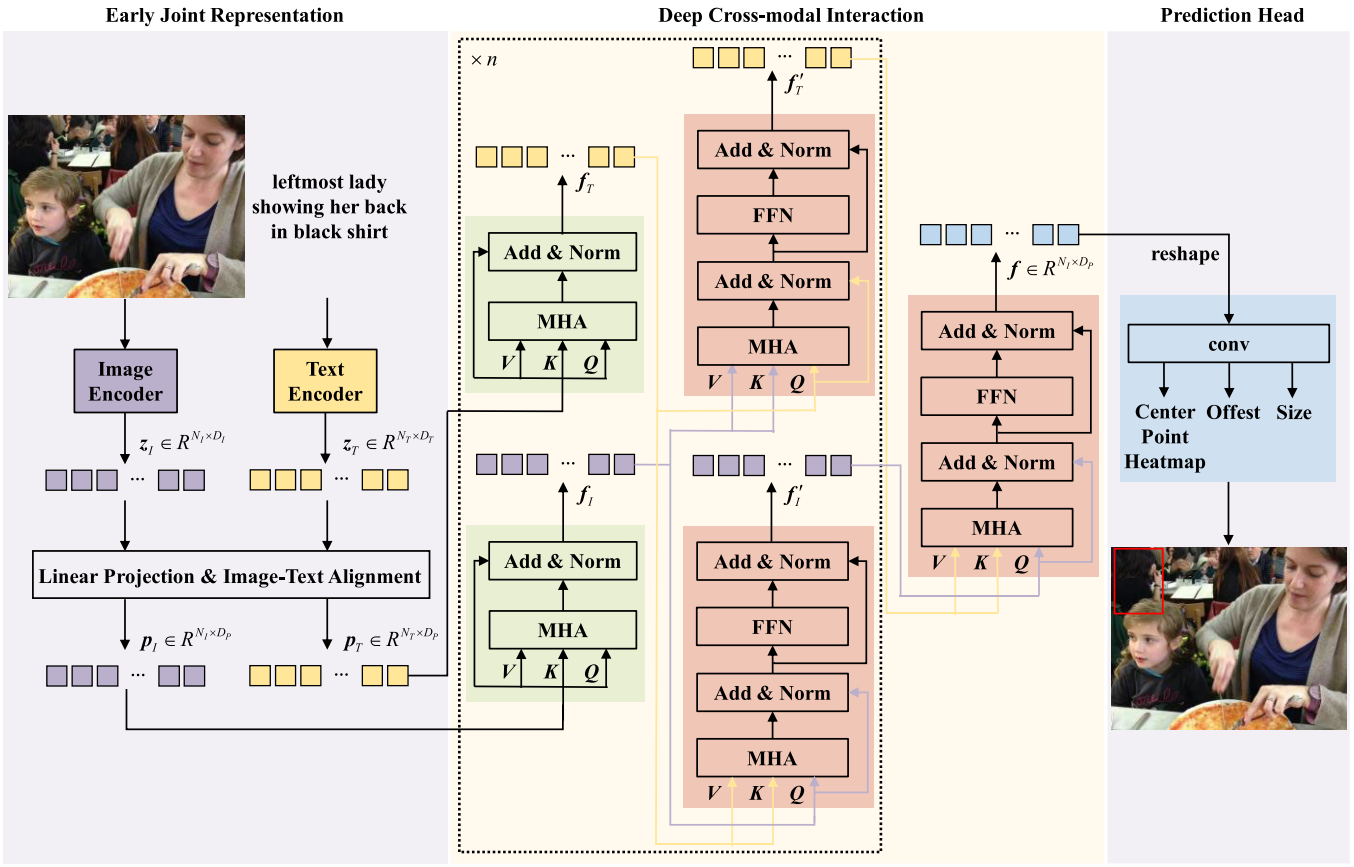


Fig. 2. Overall architecture of the proposed JMRI. It consists of three main components: early joint representation (left), deep cross-modal interaction (middle), and prediction head (right).

where LN(·) indicates the layer normalization [50], MHA(·) is the multihead attention layer. As shown in the red bottom box in Fig. 2 (middle), CMI fuses the features from different modalities by multihead cross-attention to capture the intermodal correlated context. Compared with IMI, CMI adds an additional feed-forward network (FFN) layer to enhance the fitting ability of the model, which is a fully connected network including two linear transformations and a rectified linear unit (ReLU) activation function in the middle. Concretely, we devise CMI to measure the interaction of modality Y to modality X, denoted by $\mathrm{CMI}_{Y\,\mathrm{to}\,X}$. Given the specific tokens $\boldsymbol{f}_X \in R^{N_X \times D}$ of modality X and the tokens $\boldsymbol{f}_Y \in R^{N_Y \times D}$ of modality Y, we set the inputs of MHA as queries $\boldsymbol{f}_X^q$, keys $\boldsymbol{f}_Y^k$, and values $\boldsymbol{f}_Y^v$. To summarize, the $\mathrm{CMI}_{Y\,\mathrm{to}\,X}(\boldsymbol{f}_X^q, \boldsymbol{f}_Y^k, \boldsymbol{f}_Y^v)$ is formulated as follows:

$$\tilde{\boldsymbol{f}}_{\mathrm{CMI}} = \mathrm{LN}\big(\boldsymbol{f}_X^q + \mathrm{MHA}\big(\boldsymbol{f}_X^q, \boldsymbol{f}_Y^k, \boldsymbol{f}_Y^v\big)\big) \tag{3}$$

$$\boldsymbol{f}_{\mathrm{CMI}} = \mathrm{LN}(\tilde{\boldsymbol{f}}_{\mathrm{CMI}} + \mathrm{FFN}(\tilde{\boldsymbol{f}}_{\mathrm{CMI}})) \tag{4}$$

where FFN(·) denotes the feed-forward network layer.

In this article, IMI and CMI are applied to develop the deep cross-modal interaction module, and the process is as follows.

1) Aligned visual tokens $\boldsymbol{p}_I$ and textual tokens $\boldsymbol{p}_T$ are fed into two IMIs to finish the intramodal correlation in respective tokens. After the IMI, we get the advanced tokens $\boldsymbol{f}_I$ and $\boldsymbol{f}_T$, respectively.
2) Taking $\boldsymbol{f}_I$ as the guided feature, we implement the visual-to-textual interaction, i.e., $\boldsymbol{f}'_T = \mathrm{CMI}_{I\,\mathrm{to}\,T}(\boldsymbol{f}_T^q, \boldsymbol{f}_I^k, \boldsymbol{f}_I^v)$. Similarly, taking $\boldsymbol{f}_T$ as the guided feature, the textual-to-visual interaction is performed as $\boldsymbol{f}'_I = \mathrm{CMI}_{T\,\mathrm{to}\,I}(\boldsymbol{f}_I^q, \boldsymbol{f}_T^k, \boldsymbol{f}_T^v)$.
3) Two IMIs and two CMIs form a fusion layer, which is repeated n times, as shown in the dotted box in Fig. 2 (middle). We employ $n = 6$ by default to balance the computational cost and performance similar to [51].
4) An additional $\mathrm{CMI}_{T\,\mathrm{to}\,I}$ is used to decode a fused feature for the latter localization reasoning.

C. Prediction Head and Training Objective

Given its simplicity and efficiency, we adopt keypoint estimation [52] to find the center point and regress the size of the referred object. Given the contextualized tokens $\boldsymbol{f} \in R^{N_I \times D_P}$ from the deep cross-modal interaction module, we reshape them to the visual feature maps, and then use a prediction head consisting of three branches to produce the center point heatmap $\hat{Y}_{x,y}$, the center offset $\hat{O} = (\hat{o}_x, \hat{o}_y)$, and the size $\hat{S} = (\hat{w}, \hat{h})$, respectively. For each branch, there is a 3×3 convolution, an ReLU activation, and another 1×1 convolution. We can obtain the 4-D coordinates of the object by computing from the keypoint to the bounding box.

The training objective is formulated as a linear combination of three commonly used losses. The first term is a

penalty-reduced pixelwise logistic regression with focal loss L_{center} [53]

$$L_{\text{center}} = -\sum_{xy} \begin{cases} (1-\hat{Y}_{xy})^{\alpha}\log(\hat{Y}_{xy}), & \text{if } Y_{xy}=1 \\ (1-Y_{xy})^{\beta}(\hat{Y}_{xy})^{\alpha}\log(1-\hat{Y}_{xy}), & \text{otherwise} \end{cases} \quad (5)$$

where α and β are the hyper-parameters of the focal loss, $\hat{Y}_{x,y}$ denotes the prediction at the (x, y) position of the heatmap, Y_{xy} denotes its ground truth. The latter two terms are both $L1$ loss L_{size} and L_{off}, and the overall loss is defined as follows:

$$L = L_{\text{center}} + \lambda_{\text{size}} L_{\text{size}} + \lambda_{\text{off}} L_{\text{off}} \quad (6)$$

where λ_{size} and λ_{off} are the hyper-parameters to modulate the effect of L_{size} and L_{off}, respectively. Keypoint estimation is able to directly produce all outputs simply and accurately without intersection over union (IoU)-based nonmaxima suppression (NMS) or other postprocessing.

IV. Experiments

A. Datasets

To comprehensively analyze the performance of our proposed method, we conduct experimental verifications on five benchmark datasets, which are ReferItGame [54], Flickr30K Entities [55], RefCOCO [56], RefCOCO+ [56], and RefCOCOg [57].

1) *ReferItGame:* ReferItGame includes 20 000 images with 120 072 referring expressions. Each image may have one or more regions corresponding to different referring expressions. We follow the standard practice of splitting the dataset into train/validation/test sets with 54 127/5842/60 103 referring expressions.
2) *Flickr30K Entities:* With 31 783 images and 427 000 referring expressions, the Flickr30K Entities dataset is divided into 29 783/1000/1000 images for training/validation/testing.
3) *RefCOCO/RefCOCO+/RefCOCOg:* The images of these three visual grounding datasets are all selected from MSCOCO [58]. RefCOCO consists of 19 994 images with 50 000 referred objects and 142 210 referring expressions. According to the split strategy [9], [16], it is split into train/validation/testA/testB sets with 120 624/10 834/5657/5095 expressions. Similarly, RefCOCO+ consists of 19 992 images with 49 856 referred objects and 141 564 referring expressions. There are 120 191/10 758/5726/4889 expressions for train/validation/testA/testB sets. RefCOCOg consists of 25 799 images with 49 856 referred objects and 95 010 referring expressions. Two commonly used split practices are RefCOCOg-google [57] and RefCOCOg-umd [59]. We report our results on both RefCOCOg-google (val-g) and RefCOCOg-umd (val-u and test-u).

B. Implementation Details

1) Inputs: We take two versions of vision transformer as the image encoder, denoted by JMRI ViT-B/16 (JMRI I) and JMRI ViT-L/14-336 (JMRI II), respectively; thus, there are also two settings for the input image. The input image is resized to 224 × 224 for JMRI I or 336 × 336 for JMRI II. Keeping the original aspect ratio, the longer edge of the image is resized to 224/336, while the shorter one is padded to 224/336 with zeros. Meanwhile, we set the max text sequence length as 77. The text sequence is enclosed with [SOS] token (start of sequence) and [EOS] token (end of sequence), and we pad it with zero tokens after the [EOS] token to achieve a consistent length equal to 77.

2) Training Details: We utilize the pretrained CLIP model to initialize the early joint representation module and freeze its parameters during training, and the other modules are optimized with AdamW optimizer [60]. The initial learning rate is set to 10^{-4}. On the ReferItGame, Flickr30K Entities, RefCOCO, and RefCOCOg datasets, we train our model in 100 epochs, and decrease the learning rate by a factor of 10 after completing 80 epochs. On the RefCOCO+ dataset, the training epoch is set to 180, and the epoch of the learning rate drop is set to 120. To avoid overfitting, we set the dropout rate to 0.1 in the MHA and FFN of each transformer layer. In all our experiments, the hyperparameters α and β in (5) are set to 2 and 4, while λ_{size} and λ_{off} in (6) are set to 0.1 and 1. Following common practice, we perform data augmentation at the training stage. We train and test the proposed model

TABLE I
Ablation Study on RefCOCO in Terms of Top-1 Accuracy (%). √ Denotes the Module Is Enabled, and Empty Denotes the Module Is Disabled

Feature Encoding		Fusion Layer		RefCOCO testA ↑
w/o Alignment	w/ Alignment	IMI	CMI	
√				33.05
√		√	√	82.83
	√			82.09
	√	√		83.41
	√		√	85.95
	√	√	√	87.30

TABLE II
Comparison With Different Visual Backbones on RefCOCO in Terms of Top-1 Accuracy (%)

Visual Backbone	RefCOCO val ↑	RefCOCO testA ↑	RefCOCO testB ↑
ResNet-50	57.62	64.11	50.64
ResNet-101	59.92	65.26	51.55
RN50×64	62.97	69.66	55.38
ViT-B/32	74.46	79.69	65.55
ViT-B/16	82.97	87.30	74.62
ViT-L/14-336	86.24	89.55	79.08

TABLE III

COMPARISON WITH STATE-OF-THE-ART METHODS ON REFERITGAME AND FLICKR30K ENTITIES IN TERMS OF TOP-1 ACCURACY (%). ↑ STANDS FOR THE HIGHER THE BEST. THE BEST THREE RESULTS ARE SHOWN IN RED, GREEN, AND BLUE FONTS, RESPECTIVELY

Model	Venue	ReferItGame test ↑	Flickr30K Entities test ↑
Proposal-based methods			
CMN [25]	*CVPR'17*	28.33	-
VC [23]	*CVPR'18*	31.13	-
MAttNet [11]	*CVPR'18*	29.04	-
Similarity Net [10]	*TPAMI'18*	34.54	60.89
CITE [61]	*ECCV'18*	35.07	61.33
PIRC [62]	*ACCV'18*	59.13	72.83
DDPN [63]	*IJCAI'18*	63.00	73.30
Proposal-free methods			
SSG [64]	*arXiv'18*	54.24	-
ZSGNet [29]	*ICCV'19*	58.63	63.39
FAOA [9]	*ICCV'19*	60.67	68.71
RCCF [30]	*CVPR'20*	63.79	-
ReSC [16]	*ECCV'20*	64.33	69.04
ReSC-Large [16]	*ECCV'20*	64.60	69.28
LSPN [65]	*ECCV'20*	-	69.53
SAFF [31]	*ACM MM'21*	66.01	70.17
Transformer-based methods			
TransVG [17]	*ICCV'21*	70.73	79.10
VGTR [19]	*ICME'22*	-	75.32
SeqTR [66]	*ECCV'22*	69.66	81.23
VLTVG [67]	*CVPR'22*	71.98	79.84
CMI [68]	*ACM TMCCA'23*	71.07	79.15
JMRI I	-	68.23	79.90
JMRI II	-	71.65	82.11

on an NVIDIA RTX3090 GPU, and for JMRI I/II, the whole training process consumes about 39/88 h with 66.7M/67.8M tunable parameters and 19.4G/97.1G floating point operations (FLOPs).

C. Ablation Study

1) Contribution of Each Part: In this section, we explore the effect and significance of each component in JMRI, and the results are shown in Table I. When CLIP is not used, i.e., w/o alignment in Table I, we adopt the alternative of two independent transformer encoders without contrastive pretraining. The whole architecture of this variant is trained end-to-end, where the vision encoder is initialized with the ViT model [38], and the language encoder is initialized with the transformer model [47]. When CLIP is used, i.e., w/ Alignment in Table I, all experimental settings are as in the JMRI I. Without early alignment and deep fusion, the grounding performance decreases dramatically. Using either our proposed early alignment or deep fusion alone will result in substantial performance gains, but it is clear that the highest accuracy is achieved by combining the two modules. In addition, compared with completely disabling the fusion layer, using only IMI improves the performance from 82.09% to 83.41%, while using only CMI also has an increase in performance (improves from 82.09% to 85.95%). In summary, the experimental results prove that the cross-modal interaction plays a more critical role than the IMI for grounding, and also demonstrate the necessity of combining early alignment and deep fusion.

2) Visual Backbone in Basic Encoder: Table II shows the ablation study on the visual backbone in JMRI. The results demonstrate that choosing a visual backbone as a convolutional network does not achieve the best performance since both the language encoder and fusion module are all transformers. A unified overall framework leads to prominent improvements in accuracy.

D. Comparison With State-of-the-Arts

To validate the merits of the proposed JMRI, we conduct evaluations on five public benchmark datasets and compare its performance against the state-of-the-art methods. Following the convention, the evaluation metric used to measure the performance is Top-1 accuracy (%). If the IoU between the ground truth and the predicted box is greater than 0.5, we treat the predicted result as a true positive, otherwise, a false positive. The Top-1 accuracy (%) is the ratio of the positive results in the Top-1 ranking.

1) ReferItGame/Flickr30K Entities: Table III shows the result comparison on the test sets of ReferItGame and Flickr30K Entities. As shown in Table III, we divide the state-of-the-art into proposal-based methods, proposal-free methods, and transformer-based methods. On the ReferItGame dataset, JMRI II obtains the second-best accuracy among all the approaches. Diversified and discriminative proposal network (DDPN) [63] ranks first in the proposal-based methods, and our method outperforms it by a large improvement of 8.65% points. Compared with the state-of-the-art proposal-free method semantic-aware feature

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS ON REFCOCO, REFCOCO+, AND REFCOCOG IN TERMS OF TOP-1 ACCURACY (%). THE BEST THREE RESULTS ARE SHOWN IN RED, GREEN, AND BLUE FONTS, RESPECTIVELY

Model	Venue	RefCOCO			RefCOCO+			RefCOCOg		
		val ↑	testA ↑	testB ↑	val ↑	testA ↑	testB ↑	val-g ↑	val-u ↑	test-u ↑
Proposal-based methods										
Neg Bag [59]	*ECCV'16*	57.30	58.60	56.40	-	-	-	39.50	-	49.50
CMN [25]	*CVPR'17*	-	71.03	65.77	-	54.32	47.76	57.47	-	-
VC [23]	*CVPR'18*	-	73.33	67.44	-	58.40	53.18	62.30	-	-
ParalAttn [24]	*CVPR'18*	-	75.31	65.52	-	61.34	50.86	58.03	-	-
MAttNet [11]	*CVPR'18*	76.65	81.14	69.99	65.33	71.62	56.02	-	66.58	67.27
LGRANs [14]	*CVPR'19*	-	76.60	66.40	-	64.00	53.40	61.78	-	-
DGA [15]	*ICCV'19*	-	78.42	65.53	-	69.07	51.99	-	-	63.28
RvG-Tree [13]	*TPAMI'19*	75.06	78.61	69.85	63.51	67.45	56.66	-	66.95	66.51
NMTree [12]	*ICCV'19*	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44
CM-A-E [27]	*CVPR'19*	78.35	83.14	71.32	68.09	73.65	58.03	-	67.99	68.67
Proposal-free methods										
SSG [64]	*arXiv'18*	-	76.51	67.50	-	62.14	49.27	47.47	58.80	-
FAOA [9]	*ICCV'19*	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.36
RCCF [30]	*CVPR'20*	-	81.06	71.85	-	70.35	56.32	-	-	65.73
ReSC [16]	*ECCV'20*	76.59	78.22	73.25	63.23	66.64	55.53	60.96	64.87	64.87
ReSC-Large [16]	*ECCV'20*	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20
SAFF [31]	*ACM MM'21*	79.26	81.09	76.55	64.43	68.46	58.43	-	68.94	68.91
Transformer-based methods										
TransVG [17]	*ICCV'21*	81.02	82.72	78.35	64.82	70.70	56.94	67.02	68.67	67.73
VGTR [19]	*ICME'22*	79.30	82.16	74.38	64.40	70.85	55.84	64.05	66.83	67.28
SeqTR [66]	*ECCV'22*	83.72	86.51	81.24	71.45	76.26	64.88	71.50	74.86	74.21
VLTVG [67]	*CVPR'22*	84.77	87.24	80.49	74.19	78.93	65.17	72.98	76.04	74.18
CMI [68]	*ACM TMCCA'23*	81.92	83.40	77.37	68.49	72.18	60.30	68.39	69.08	69.04
JMRI I	-	82.97	87.30	74.62	71.17	79.82	57.01	69.32	71.96	72.04
JMRI II	-	86.24	89.55	79.08	77.29	85.09	64.96	77.22	78.12	77.50

filter (SAFF) [31], the proposed method also performs significantly better than it (71.65% versus 66.01%). Among the transformer-based methods, our method shows comparable results to their best model visual-linguistic transformer-based framework for visual grounding (VLTVG) [67]. On the Flick30K Entities dataset, JMRI with two versions obtained the first and the third best results, respectively. Compared with the best proposal-based method DDPN and the best proposal-free method SAFF, JMRI I/II performs remarkable improvements (6.60-/8.81-point improvement over DDPN and 9.73-/11.94-point improvement over SAFF). The transformer-based method SeqTR [66] ranks second among all the methods, our JMRI II improves it from 81.23% to 82.11%.

2) RefCOCO/RefCOCO+/RefCOCOg: We also conduct experiments on the RefCOCO, RefCOCO+, and RefCOCOg datasets, and the results of our method, as well as other state-of-the-art, are reported in Table IV. As shown in Table IV, our JMRI, VLTVG, and SeqTR, which are all the transformer-based methods, rank in the top three in accuracy, better than the other methods. On the RefCOCO dataset, JMRI II outperforms all other methods on val and testA, and it obtains the third-best accuracy on testB. Compared with VLTVG and SeqTR, JMRI II achieves improvements by a performance gain (1.41-/2.31-point improvement over VLTVG on val/testA, 2.52-/3.04-point improvement over SeqTR on val/testA). On the RefCOCO+ dataset, our larger version achieves the highest accuracy on both val and testA, and it obtains the second-best accuracy on testB. Compared with VLTVG, JMRI II outperforms it by a significant margin of 3.10/6.16 points on val/testA. Compared with

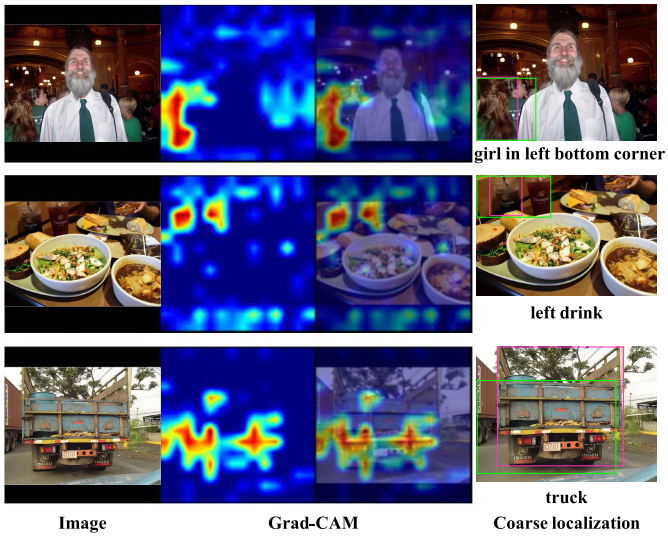


Fig. 3. Visualization of early joint representation. Grad-CAM maps highlight image regions considered to be important and produce a coarse localization considered to be consistent with the text expression. Green and rose red boxes denote the predicted region and the ground truth, respectively.

Fig. 4. Visualization of the predicted results on RefCOCO and RefCOCOg. Green and rose red boxes denote the predicted region and the ground truth, respectively. (a) Woman carrying a gray bag. (b) Painting of scissors, that is, sitting on a chair. (c) One lambs sitting near its mom has two teal spots on its back. (d) Part of a cellphone to the far left. (e) Pizza in front of man in green. (f) Back left sandwitch. (g) Donut with sprinkles to the top right of the other donuts. (h) Teddy bear doll, second from left. (i) Hotdog being held in front of a man in a black shirt. (j) Lady in a red jacket holding a striped kite. (k) Track that is yellow on the top and white on the bottom half. (l) White chair with a brown haired woman wearing a red shirt and blue jeans.

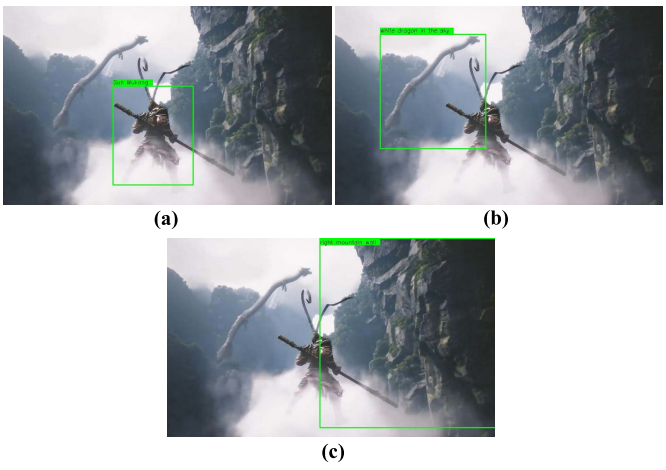


Fig. 5. Visualization of zero-shot prediction. Green box denotes the predicted region. (a) Sun Wukong. (b) White dragon in the sky. (c) Right mountain wall.

SeqTR, JMRI II also achieves significant improvements (5.84/8.83/0.08 points on val/testA/testB). On the RefCOCOg dataset, our method obtains the best accuracy on both RefCOCOg-google and RefCOCOg-umd splits. JMRI II surpasses the previous state-of-the-art VLTVG/SeqTR by an improvement of 4.24/5.72 points on val-g, 2.08/3.26 points on val-u, and 3.32/3.29 points on test-u.

The experimental results provide empirical evidence of the effectiveness of combining early joint representation and deep cross-modal interaction in visual grounding. The superior performance of our approach offers valuable insights for researchers, highlighting the potential benefits of applying the large-scale pretrained foundation model to the multimodal tasks.

E. Qualitative Analysis

1) Visualization of Early Joint Representation: To demonstrate the interpretability of early joint representation, we use Grad-CAM [48] to visualize some cases in Fig. 3. Given the early joint representation module, we compute the gradient by its contrastive loss and generate Grad-CAM visualization as shown in the middle column of Fig. 3. In the right column, we display the coarse bounding box corresponding to high activation values. As can be seen from the figure, Grad-CAM maps usually pay attention to the relevant cues and highlight image regions corresponding to the target object, even if not precise enough. For example, in the second row, the category "drink" is well distinguished, except that it cannot be accurately located in the two drinks. These results further prove that the early joint representations have strong class-discriminative ability, lacking of localization information.

2) Visualization of Grounding Results: As shown in Fig. 4, we show some challenging examples in the RefCOCO and RefCOCOg datasets. We observe that the proposed approach is able to obtain high IoU in the following challenging cases.

1) *Multiple Similar Objects:* As shown in Fig. 4(a)–(d), these images contain multiple objects of the same type

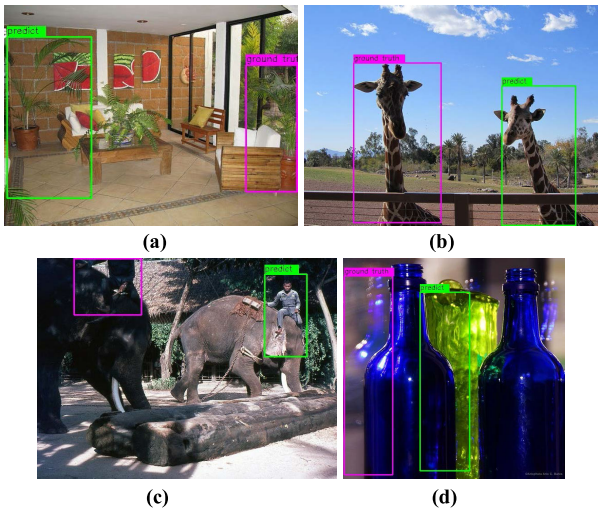


Fig. 6. Analysis of the limitations of JMRI. Green and rose red boxes denote the predicted region and the ground truth, respectively. (a) Plant inside a red vase next to a wooden chair. (b) Giraffe stands next to another giraffe and puts his head over railing. (c) Person in the shadows riding an elephant. (d) First shadow of bottle on left.

(such as woman, scissors, and cellphones), which poses a challenge for the model to accurately distinguish the target from a cluster of visually similar objects.

2) *Complicated Location Relationships:* When the expressions contain the descriptions of complicated location relationships, such as "donut with sprinkles to the top right of the other donuts" in Fig. 4(g) and "teddy bear doll, second from left" in Fig. 4(h), the model should be able to capture the correct cross-modal semantic correlation.

3) *Complex Background/Expression:* Fig. 4(i) and (j) contain a wide variety of objects and backgrounds, and the expressions in Fig. 4(k) and (l) are long and complex, which bring great difficulties for accurate grounding.

In the above challenging cases, the results fully demonstrate the effectiveness of our approach.

3) Visualization of Zero-Shot Prediction: Herein we make an exploratory attempt to test our method on the data that is not part of the five aforementioned datasets, and the results are shown in Fig. 5. The results show that the proposed model can perform zero-shot grounding on certain new visual concepts in the open world, such as Sun Wukong, white dragon, mountain wall, and even abstract words. We believe the reason is that CLIP learned by natural language supervision has flexible zero-shot transfer capability.

F. Limitations

Finally, we discuss the limitations of our method. JMRI is designed for grounding the target object referred to by the natural language. Inevitably, it relies on the explicitness of language expression to some extent. As shown in Fig. 6(a) and (b), the given language expressions do not clearly specify which "plant" and "giraffe" are the target ones, as there are multiple similar objects that fit the descriptions. For such ambiguous queries, the model is difficult to predict the right target. As shown in Fig. 6(c) and (d), some mislocalizations have occurred because our model does not understand the semantics of "shadow." Mitigating the adverse effects of language bias on the grounding task will be our focus in future work.

V. Conclusion

In this article, we present JMRI, a novel visual grounding approach by combining early joint representation and deep cross-modal interaction. We propose to use the large-scale vision-language foundation model for early alignment and transformer for deep fusion to establish multi-modal correspondence, resulting in high-quality language-aware visual representations for localization reasoning. Experimental results on five benchmark datasets demonstrate the effectiveness of the proposed method against the state-of-the-arts. Our JMRI introduces as a novel grounding framework and shows great potential in future research.

References

[1] A. Motroni, A. Buffi, P. Nepa, and B. Tellini, "Sensor-fusion and tracking method for indoor vehicles with low-density UHF-RFID tags," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[2] X. Zhang et al., "RI-fusion: 3D object detection using enhanced point features with range-image fusion for autonomous driving," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[3] Q. Xie, D. Li, Z. Yu, J. Zhou, and J. Wang, "Detecting trees in street images via deep learning with attention module," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 8, pp. 5395–5406, Aug. 2020.

[4] J. Ren et al., "Deep texture-aware features for camouflaged object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1157–1167, Mar. 2023.

[5] H. Yin, Z. Liu, Z. Xu, and L. Gao, "An automatic visual monitoring system for expansion displacement of switch rail," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 6, pp. 3015–3025, Jun. 2020.

[6] W. Zhang, C. Ma, Q. Wu, and X. Yang, "Language-guided navigation via cross-modal grounding and alternate adversarial learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3469–3481, Sep. 2021.

[7] B. Qi et al., "An image-text dual-channel union network for person re-identification," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–16, 2023.

[8] S. Wang, D. Guo, X. Xu, L. Zhuo, and M. Wang, "Cross-modality retrieval by joint correlation learning," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 15, no. 2, pp. 1–16, 2019.

[9] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo, "A fast and accurate one-stage approach to visual grounding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4682–4692.

[10] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 394–407, Feb. 2019.

[11] L. Yu et al., "MAttNet: Modular attention network for referring expression comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1307–1315.

[12] D. Liu, H. Zhang, Z.-J. Zha, and F. Wu, "Learning to assemble neural module tree networks for visual grounding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4672–4681.

[13] R. Hong, D. Liu, X. Mo, X. He, and H. Zhang, "Learning to compose and reason with language tree structures for visual grounding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 684–696, Feb. 2022.

[14] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. van den Hengel, "Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1960–1968.

[15] S. Yang, G. Li, and Y. Yu, "Dynamic graph attention for referring expression comprehension," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4643–4652.

[16] Z. Yang, T. Chen, L. Wang, and J. Luo, "Improving one-stage visual grounding by recursive subquery construction," in *Proc. Eur. Conf. Comput. Vis.*, vol. 2020, pp. 387–404.

[17] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, "TransVG: End-to-end visual grounding with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1749–1759.

[18] J. Deng et al., "TransVG++: End-to-end visual grounding with language conditioned vision transformer," 2022, *arXiv:2206.06619*.

[19] Y. Du, Z. Fu, Q. Liu, and Y. Wang, "Visual grounding with transformers," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

[20] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive captioners are image-text foundation models," 2022, *arXiv:2205.01917*.

[21] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[22] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[23] H. Zhang, Y. Niu, and S.-F. Chang, "Grounding referring expressions in images by variational context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4158–4166.

[24] J. Liu, L. Wang, and M.-H. Yang, "Referring expression generation and comprehension via attributes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4866–4874.

[25] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling relationships in referential expressions with compositional modular networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4418–4427.

[26] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. van den Hengel, "Parallel attention: A unified framework for visual object discovery through dialogs and queries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4252–4261.

[27] X. Liu, Z. Wang, J. Shao, X. Wang, and H. Li, "Improving referring expression grounding with cross-modal attention-guided erasing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1950–1959.

[28] V. Cirik, T. Kirkpatrick, and L. Morency, "Using syntax to ground referring expressions in natural images," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.

[29] A. Sadhu, K. Chen, and R. Nevatia, "Zero-shot grounding of objects from natural language queries," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4693–4702.

[30] Y. Liao et al., "A real-time cross-modality correlation filtering method for referring expression comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10877–10886.

[31] J. Ye, X. Lin, L. He, D. Li, and Q. Chen, "One-stage visual grounding via semantic-aware feature filter," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1702–1711.

[32] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," 2021, *arXiv:2104.13921*.

[33] L. H. Li et al., "Grounded language-image pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10955–10965.

[34] X. Zhai et al., "LiT: Zero-shot transfer with locked-image text tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18102–18112.

[35] H. Pham et al., "Combined scaling for zero-shot transfer learning," 2021, *arXiv:2111.10050*.

[36] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[37] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[38] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–12.

[39] Z. Huang, J. Li, Z. Hua, and L. Fan, "Underwater image enhancement via adaptive group attention-based multiscale cascade transformer," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–18, 2022.

[40] Y. Tian, H. Meng, F. Yuan, Y. Ling, and N. Yuan, "Vision transformer with enhanced self-attention for few-shot ship target recognition in complex environments," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.

[41] H. Yang, D. Zhang, A. Hu, C. Liu, T. J. Cui, and J. Miao, "Transformer-based anchor-free detection of concealed objects in passive millimeter wave images," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–16, 2022.

[42] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[44] Y. Chen et al., "UNITER: Universal image-text representation learning," in *Proc. Eur. Comput. Vis.*, 2020, pp. 104–120.

[45] X. Li et al., "OSCAR: Object-semantics aligned pre-training for vision-language tasks," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 121–137.

[46] Z. Yang et al., "TAP: Text-aware pre-training for text-VQA and text-caption," 2020, *arXiv:2012.04638*.

[47] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[48] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[49] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pretraining for image captioning and VQA," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13041–13049.

[50] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[51] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8122–8131.

[52] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2019, pp. 1–12.

[53] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[54] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "ReferItGame: Referring to objects in photographs of natural scenes," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 787–798.

[55] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 74–93, May 2017.

[56] L. Yu, P. Poirson, S. Yang, A. Berg, and T. Berg, "Modeling context in referring expressions," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–85.

[57] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 11–20.

[58] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[59] V. Nagaraja, V. Morariu, and L. Davis, "Modeling context between objects for referring expression understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 792–807.

[60] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[61] B. Plummer, P. Kordas, M. Kiapour, S. Zheng, R. Piramuthu, and S. Lazebnik, "Conditional image-text embedding networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 249–264.

[62] R. Kovvuri and R. Nevatia, "PIRC Net: Using proposal indexing, relationships and context for phrase grounding," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 451–467.

[63] Z. Yu, J. Yu, C. Xiang, Z. Zhao, Q. Tian, and D. Tao, "Rethinking diversified and discriminative proposal generation for visual grounding," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1114–1120.

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

[65] S. Yang, G. Li, and Y. Yu, "Propagating over phrase relations for one-stage visual grounding," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 589–605.

[66] C. Zhu et al., "SeqTR: A simple yet universal network for visual grounding," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 598–615.

[67] L. Yang, Y. Xu, C. Yuan, W. Liu, B. Li, and W. Hu, "Improving visual grounding with visual-linguistic verification and iterative reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9489–9498.

[68] K. Li, J. Li, D. Guo, X. Yang, and M. Wang, "Transformer-based visual grounding with cross-modality interaction," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 6, pp. 1–19, Nov. 2023.

Hong Zhu received the B.S. degree from the National University of Defense Technology, Changsha, China, in 2010, and the M.S. degree from the Army Artillery and Air Defense Academy of PLA, Hefei, China, in 2015. She is currently pursuing the Ph.D. degree with the National University of Defense Technology.

Her current research interests include visual grounding and visual tracking.

Qingyang Lu received the B.S. degree from the Army Officer Academy of PLA, Hefei, China, in 2016. He is currently pursuing the M.S. degree with the Army Artillery and Air Defense Academy of PLA, Hefei.

His research interests include multimodal learning and visual grounding.

Lei Xue received the B.S. and M.S. degrees from the College of Electronic Engineering, Hefei, China, in 1983 and 1990, respectively.

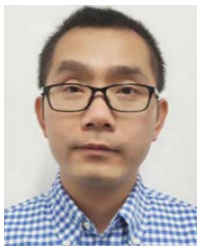
Since July 2017, he has been with the College of Electronic Engineering, National University of Defense Technology, Hefei, where he is currently a Full Professor. His research interests include signal processing, data fusion, and intelligent information processing.

Mogen Xue received the Ph.D. degree from the Second Artillery Engineering University, Xi'an, China, in 2007.

He is currently a Full Professor with the Army Artillery and Air Defense Academy of PLA, Hefei, China, where he is also the Director of the Anhui Key Laboratory of Polarization Imaging Detection Technology. His research interests include image processing, photoelectric detection, and object tracking.

Guanglin Yuan received the Ph.D. degree from the Hefei University of Technology, Hefei, China, in 2011.

He is currently a Professor with the Army Artillery and Air Defense Academy of PLA, Hefei. His research interests include image processing, multimodal fusion, and object tracking.



Bineng Zhong received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2010.

From 2007 to 2008, he was a Research Fellow with the Institute of Automation and Institute of Computing Technology, Chinese Academy of Science. From September 2017 to September 2018, he was a Visiting Scholar with Northeastern University, Boston, MA, USA. From November 2010 to October 2020, he was a Professor with the School of Computer Science and Technology, Huaqiao University, Xiamen, China. He is currently a Professor with the School of Computer Science and Engineering, Guangxi Normal University, Guilin, China. His current research interests include pattern recognition, machine learning, and computer vision.