Transformer-based Visual Grounding with Cross-modality Interaction

KUN LI, JIAXIU LI, and DAN GUO*, Hefei University of Technology, China
XUN YANG*, University of Science and Technology of China, China
MENG WANG, Hefei University of Technology, China

This paper tackles the challenging yet important task of Visual Grounding (VG), which aims to localize a visual region in the given image referred by a natural language query. Existing efforts on the VG task are twofold: 1) *two-stage methods* first extract region proposals and then rank them according to their similarities with the referring expression, which usually leads to suboptimal results due to the quality of region proposals; 2) *one-stage methods* usually predict all the possible coordinates of the target region online by leveraging modern object detection architectures, which pay few attention to cross-modality correlations and have limited generalization ability. To better address the task, we present an effective transformer-based end-to-end visual grounding approach, which focuses on capturing the cross-modality correlations between the referring expression and visual regions for accurately reasoning the location of the target region. Specifically, our model consists of a feature encoder, a cross-modality interactor, and a modality-agnostic decoder. The feature encoder is employed to capture the intra-modality correlation, which models the linguistic context in query and the spatial dependency in image respectively. The cross-modality interactor endows the model with the capability of highlighting the localization-relevant visual and textual cues by mutual verification of vision and language, which plays a key role in our model. The decoder learns a consolidated token representation enriched by multi-modal contexts and further directly predicts the box coordinates. Extensive experiments on five public benchmark datasets with quantitative and qualitative analysis clearly demonstrate the effectiveness and rationale of our proposed method.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Computer vision**; **Computer vision tasks**;

Additional Key Words and Phrases: Visual grounding, Referring expression, Cross-modality interaction

ACM Reference Format:
Kun Li, Jiaxiu Li, Dan Guo, Xun Yang, and Meng Wang. 2023. Transformer-based Visual Grounding with Cross-modality Interaction. *ACM Trans. Multimedia Comput. Commun. Appl.* 1, 1, Article 1 (June 2023), 19 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding authors.

K. Li and J. Li are with School of Computer Science and Information Engineering, School of Artificial Intelligence, Hefei University of Technology (HFUT), Hefei, 230601, China. (e-mail: kunli.hfut@gmail.com; jiaxiuli@mail.hfut.edu.cn). D. Guo and M. Wang are with School of Computer Science and Information Engineering, School of Artificial Intelligence, Hefei University of Technology (HFUT), Hefei, 230601, China, and are with Intelligent Interconnected Systems Laboratory of Anhui Province, Hefei University of Technology (HFUT), and are with Key Laboratory of Knowledge Engineering with Big Data (HFUT), Ministry of Education, and are with the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China. (e-mail: guodan@hfut.edu.cn; eric.mengwang@gmail.com). X. Yang is with School of Information Science and Technology, University of Science and Technology of China, Hefei, 230026, China. (e-mail: xyang21@ustc.edu.cn).

1551-6857/2023/6-ART1 $15.00
https://doi.org/XXXXXXX.XXXXXXX

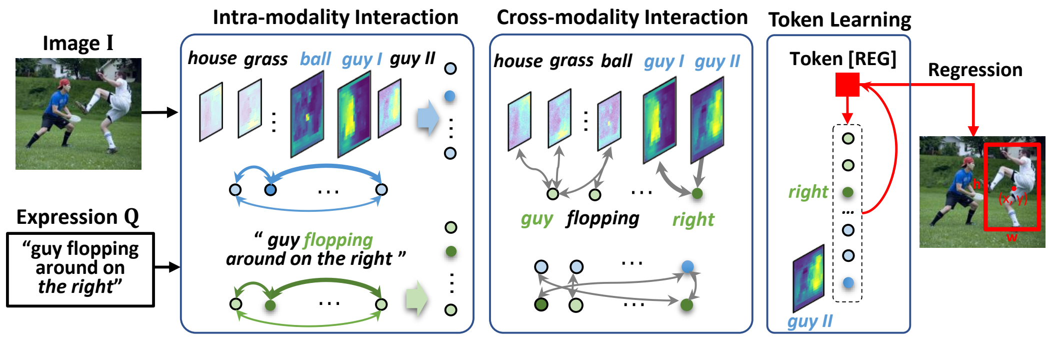

Fig. 1. The basic idea of this work. The goal of the Visual Grounding (VG) task is to localize a visual region in the given image referred by a natural language query. How to effectively model the intra-modality interaction and cross-modality interaction is the key to addressing the VG task.

1 INTRODUCTION

Visual Grounding (VG) is an important and challenging task in the computer vision research community [4, 9, 21, 47, 50]. It has been intersected with natural language processing and multimedia understanding [11], having valuable impacts on the downstream cross-modality tasks, such as video questioning answering (VQA) [16, 38], visual dialog [10], video grounding [7, 20, 45, 48, 56], vision-language navigation [31], cross-modality retrieval [35, 41, 44], video object grounding [46], and multi-modal pre-training [17], *etc*. Specifically, visual grounding is also called Referring Expression Comprehension or Referring Expression Grounding (REC or REG) [4, 15, 25]. As shown in Fig. 1, given a referring expression and an image, the goal of the VG task is to localize a visual region (*e.g.*, object or background) in the image corresponding to the language expression semantically. The referring expressions usually contain diverse language descriptions (*e.g.*, relative position, object context, attributes) about the queried region, making the VG task quite challenging.

Conventional VG methods [23, 28, 55, 57] mostly formulate this problem as an object retrieval task, where an object that best semantically matches the referring expression is retrieved from a set of object proposals. As shown in Fig. 2 (a), these methods are mainly composed of two stages. In the first stage, dense proposals (diverse visual regions in the image) are extracted by various object detection methods (*e.g.*, Faster R-CNN [12], YOLOv3 detector [32]). In the second stage, a common practice [23, 28, 55] is to use CNN and LSTM to encode these candidate objects and the referring expression respectively, and then calculate their similarity to select the best matched visual region as the output. To build the relationship between the visual region and the expression, various attention-based methods [25, 57] and graph-based methods [40, 42] are proposed. Although existing two-stage methods have achieved great advance, there are still some issues: 1) the grounding performance highly depends on the quality of object proposals. It is impossible to localize the target region if it is not accurately detected in the first stage; 2) two-stage methods are usually computationally costly due to the proposal generation and cross-modality similarity computation.

Recently, one-stage solutions [4, 34, 49, 50] are introduced to alleviate the issues of the two-stage methods. As shown in Fig. 2 (b), they are inspired by the one-stage object detection methods, and usually directly predict the bounding box coordinates of the target region. Yang *et al.* [50] follow the whole regression manner in the YOLOv3 detector [32] and replace the last *sigmoid* layer with a *softmax* function for target boundary prediction. Similarly, Sadhu *et al.* [34] propose a zero-shot video grounding approach by introducing the SSD detector [24]. Deng *et al.* [4] propose a transformer [37] based framework, which formulates the visual grounding as a direct coordinates

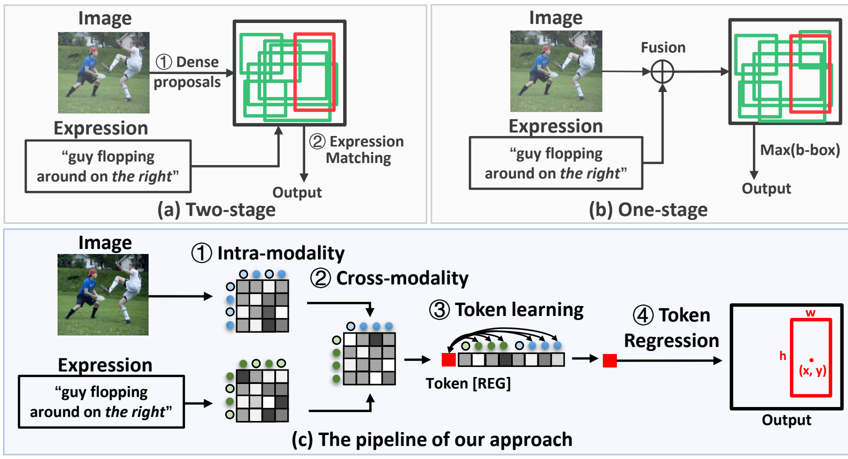

Fig. 2. The pipeline of visual grounding task. Previous work can be grouped into two pipelines: (a) two-stage methods [15, 22, 52] and (b) one-stage (end-to-end) methods [49, 50]. Our method belongs to the latter.

regression problem. Du *et al.* [6] use the transformer for visual feature learning under the guidance of referring expression. Despite the encouraging progress achieved by one-stage methods, they usually pay more attention to the design of visual or textual feature encoder modules while ignoring the importance of modeling cross-modality interaction between referring expression and visual regions. How to effectively understand the contents of the language and vision parts and capture the cross-modality semantic correlation has not been well studied so far for the VG task.

To fill the research gap and better address the VG task, we present a transformer-based visual grounding approach, which focuses on capturing the relationship between the referring expression and visual regions for accurately reasoning the location of the target region in an end-to-end fashion. As shown in Fig. 1, the image contains complex background and diverse visual subjects (*i.e.*, *"house", "grass", "ball", "guy I" (left), "guy II" (right)*), and the expression is composed of rich linguistic semantics (*e.g.*, *"guy", "flopping", "right"*). The key is how to effectively align the diverse visual regions and rich linguistic expression, thus closing the semantic gap between vision and language. We attempt to address this issue by explicitly learning relevant cues through multi-modality correlation.

As is shown in Fig. 2 (c), the pipeline of this work is dominated by a progressive correlation strategy for the task, involving three types of interaction, *i.e.*, intra-modality interaction, cross-modality interaction, and visual-linguistic fusion. Firstly, the intra-modality correlation is explored, which merely considers the correlation in single-modality, leading to an overemphasis on salient semantics in each own modality (*e.g.*, visual subjects {*"guy I", "guy II", "ball", "grass", "house"*} and texts {*"guy", "flopping", "around", "right"*} in the example of Fig. 1). This is not conducive enough to location reasoning. Thus, secondly, we leverage the cross-modality correlation to highlight relevant cues in both modalities. Relevant cues (*e.g.*, the co-occurrence subjects - visual *"guy I", "guy II"* and text *"guy"*) are exploited based on the cross-attention mechanism. After that, we employ a learnable token [REG] to interact with the global multi-modal sequence for the final visual-linguistic fusion. Finally, we use the output state of the token [REG] to predict the target location of the queried object.

In this work, as shown in Fig. 3, we first use an image and a language feature encoders to model the intra-modality correlation in each modality of image and expression, respectively. Then, we devise a cross-modality interactor to explore the cross-modality correlation between image and expression. Finally, a modality-agnostic decoder is dedicated to learning a regression token enriched by multi-modal contexts. Merely the contextual enriched token is used to predict the coordinates of the target object. Our main contributions are summarized as follows:

- We propose a transformer-based visual grounding approach that aims to highlight vision and language semantics through cross-modality interaction.
- We propose a progressive interaction approach, covering comprehensive manners of intra-modality interaction, cross-modality interaction, and visual-linguistic fusion.
- Extensive experiments are conducted on five public benchmark datasets to demonstrate the effectiveness of the proposed method. Ablation studies and qualitative analysis clearly validate the rationale of our method.

2 RELATED WORK

In this section, we review previous visual grounding work. According to the core pipeline of methods, previous work can be roughly grouped into two types: two-stage methods and one-stage methods.

2.1 Two-stage methods

As shown in Fig. 2 (a), early work addresses this task with two offline stages. In the first stage, enormous region proposals and visual features are extracted by selective search [36], Edge-box [58], Faster R-CNN [12], and YOLO series [32], *etc.* Then, in the second stage, the best candidate proposal is selected according to the similarity between visual and textual features. We review the existing methods as follows. (1) In ***CNN-LSTM-based model***, CNN is responsible for visual modeling while LSTM is responsible for language modeling. There are some typical methods, such as VC [55] and Attribute [23]. **(2)** ***Modular network*** decomposes the referring expression into different modular components (*e.g.*, subject, location, and relationship to other objects) and then matches each component with the image, such as CMN[15] and MAttNet [52]. **(3)** ***Attention-based model*** focuses on crucial words or image regions, such as ParalAttn [57] and CM-A-E [25]. **(4)** ***Graph-based model*** through the object-relation graph learning discovers the related objects in the expression, such as LGRANS [40] and DGA [42]. **(5)** ***Language parser model*** introduces the external language parser to enhance the representation ability of the expression, such as GroundNet [3] and NMTree [22].

Although the above two-stage methods have achieved encouraging progress, there are still some drawbacks. Firstly, the extraction of rich proposals is computationally expensive, and dense proposals make it difficult for the model to realize real-time referring expression comprehension. Secondly, the quality of the proposal will also affect the accuracy of grounding. For example, if some objects are neglected in the first stage, it is hard to locate the target in the second stage.

2.2 One-stage methods

Inspired by the great success of the one-stage pipeline in object detection, the neat and clear idea of the one-stage pipeline is also applied to the visual grounding task. Different from the two-stage method, the one-stage method (as shown in Fig. 2 (b)) predicts the bounding box of the target region directly. The one-stage method is still in its infancy, and the related work is much less than the two-stage method. The current work can be grouped into four types. **(1)** ***Object detection based model***: Inspired by YOLOv3 [32], Yang *et al.* [50] propose a one-stage framework equipped with DarkNet [32] and BERT [5] for extracting visual and textual features, respectively. **(2)** ***Visual feature optimization model:*** Sadhu *et al.* [34] propose a zero-shot grounding network to improve the quality of visual features. Liao *et al.* [21] present a real-time cross-modality filtering network

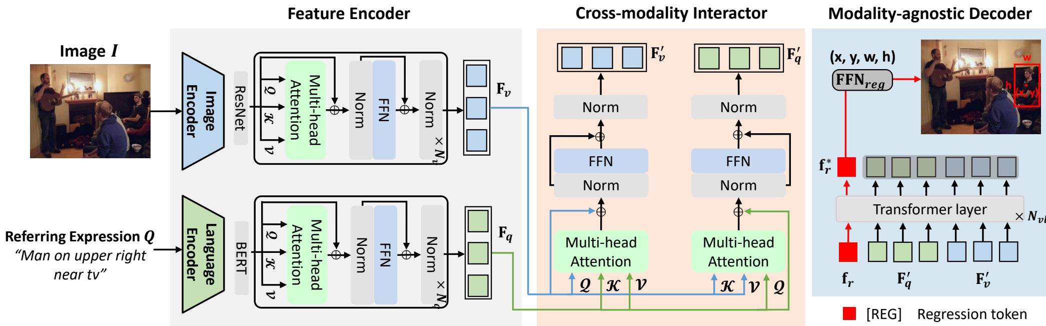

Fig. 3. Overview of the proposed approach for visual grounding. Firstly, image **I** and referring expression **Q** are fed into the *Feature Encoder*. Subsequently, the encoded visual and textual features are fed into the *Cross-modality Interactor* to attend to relevant features via cross-modality correlation. Finally, we concatenate all the cross-modality features with a regression token in a transformer framework and feed only the token to a *Modality-agnostic Decoder* for target regression.

to generate multi-level visual feature maps. **(3)** ***Textual feature optimization model:*** Yang *et al.* [49] propose a recursive sub-query construction module to reduce the referring ambiguity. Ye *et al.* [51] propose a filter-based cross-modality fusion network to select discriminative visual feature maps with explicit textual guidance. **(4)** ***Transformer-based model:*** Deng *et al.* [4] propose a transformer-based framework for object coordinates regression. Du *et al.* [6] propose an encoder-decoder transformer to learn more discriminative visual features under the guidance of textual expression.

In summary, previous work focuses on different aspects of the visual grounding solution, such as feature extraction, feature fusion, proposal modeling, proposal regression, *etc.* In this paper, we explore cross-modality interaction in-depth (*i.e.*, textual to visual, visual to textual, jointly {visual & textual} to {visual & textual}) for visual grounding. The main idea is to enhance the representation ability of image and query to improve grounding accuracy. We hope the idea can inspire future work on visual grounding.

3 OUR APPROACH

This work formulates the visual grounding task as a bounding box regression problem. Given an RGB image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and a referring expression $\mathbf{Q} = \{\mathbf{q}_l\}_{l=1}^{L_q}$, where $\mathbf{q}_l$ is the l-th word, and L_q is the number of words. We learn a model $\mathcal{F}$ parameterized by Θ to locate a target region in the image with a bounding box $b = [x, y, w, h]$, corresponding to the expression **Q** semantically:

$$b = \mathcal{F}(\mathbf{I}, \mathbf{Q}; \Theta). \tag{1}$$

The overview of our approach is shown in Fig. 3. Specifically, the proposed method mainly consists of three components. 1) *Feature Encoder* (Sec. 3.1): this module aims to capture the intra-modality correlation,*i.e.*, linguistic context in the query and spatial dependency in the image. 2) *Cross-modality Interactor* (Sec. 3.2): this module is dedicated to capturing crucial contexts by cross-modality correlation. 3) *Modality-agnostic Decoder* (Sec. 3.3): this module is committed to learning a modality-agnostic token for target bounding box prediction.

3.1 Feature Encoder

As shown in Fig. 3, this module consists of two modality-specific encoders (the image encoder and the language encoder). **For the image encoder**, given an RGB image **I**, we extract the feature

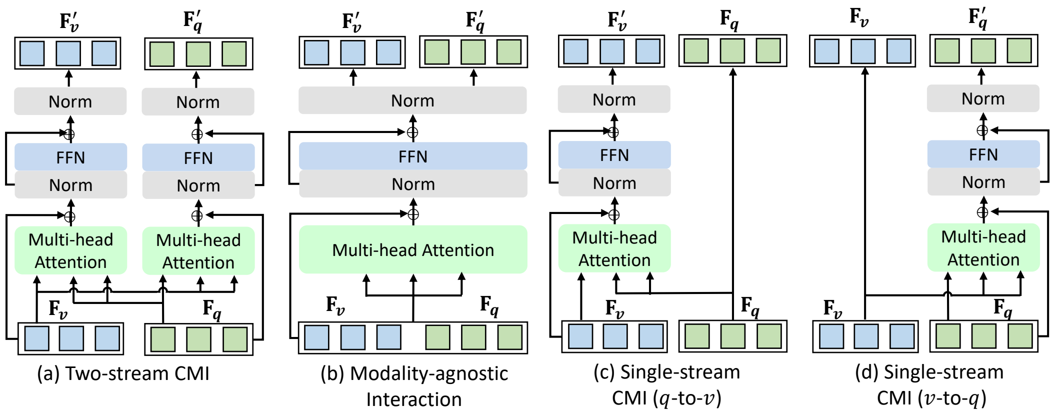

Fig. 4. Different instantiations of cross-modality interaction. The instantiation (a) is adopted in our method.

map $\mathbf{F}_I \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times d_I}$ by ResNet [13], where d_I=2048. Then, we use a 1×1 convolutional layer on $\mathbf{F}_I$ to reduce the feature dimension, and flatten it to a feature sequence $\mathbf{F}'_I \in \mathbb{R}^{L_v\times d}$, where $L_v = \frac{H}{32} \times \frac{W}{32}$ and d is a unified dimension. After that, we feed $\mathbf{F}'_I$ into a transformer block with N_v=6 layers [37] to yield the final visual feature $\mathbf{F}_v \in \mathbb{R}^{L_v\times d}$. **For the language encoder**, we follow the convention [5] to append "[CLS]" and "[SEP]" as head and tail words at the beginning and end of the referring expression $\mathbf{Q}$, and then extract the textual features by BERT, $\mathbf{F}_Q \in \mathbb{R}^{L_q\times 768}$. BERT [5] is a transformer model with N_q=12 layers. To keep the same dimension of visual and textual features, we use a linear projection to convert the feature dimension of $\mathbf{F}_Q$ into d. We get the new textual features $\mathbf{F}_q \in \mathbb{R}^{L_q\times d}$.

In other words, the transformers used in the two encoders are exploited to model the intra-modality interaction. To be specific, in the transformer, a specific sequence $\mathbf{F}_x \in \mathbb{R}^{L_x\times d}$ of modality $\mathbf{X}$ is linearly transformed into *query* $\mathbf{F}^q_x$, *key* $\mathbf{F}^k_x$, and *value* $\mathbf{F}^v_x$, respectively, as the input of a multi-head (self) attention (**MHA**):

$$\begin{cases} \mathbf{MHA}(\mathbf{F}^q_x, \mathbf{F}^k_x, \mathbf{F}^v_x) = \mathbf{W}_1[h_1; h_2; \ldots; h_n], \\ \quad h_i = \mathrm{softmax}(\frac{\mathbf{F}^q_x(\mathbf{F}^k_x)^\top}{\sqrt{d_h}})\mathbf{F}^v_x, \end{cases} \tag{2}$$

where n is the number of attention heads, h_i is the output of the i-th head, d_h is the hidden dimension of each head, and $\mathbf{W}_1 \in \mathbb{R}^{d\times d}$ is a trainable parameter. Up to now, we finish the intra-modality correlation in respective sequences $\mathbf{F}_v$ and $\mathbf{F}_q$.

3.2 Cross-modality Interactor

After the above intra-modality correlation, we get the feature representation of visual and textual modalities $\mathbf{F}_v \in \mathbb{R}^{L_v\times d}$ and $\mathbf{F}_q \in \mathbb{R}^{L_q\times d}$. In this part, we devise to explore the mutually correlated contexts in both sides of $\mathbf{F}_v \in \mathbb{R}^{L_v\times d}$ and $\mathbf{F}_q \in \mathbb{R}^{L_q\times d}$. Specifically, we extend the transformer layer as a Cross-Modality Interaction Module (CMIM) to measure the interaction of modality Y to modality $\mathbf{X}$ (Y-to-X). Given a specific sequence $\mathbf{F}_x$ of modality $\mathbf{X}$ and a sequence $\mathbf{F}_y$ of modality Y, we linearly transform both of them into three new feature sequences: *query* $\mathbf{F}^q_x$, *key* $\mathbf{F}^k_y$, and *value* $\mathbf{F}^v_y$. Subsequently, we use another multi-head attention (**MHA**) to perform the cross-modality correlation between the sequences $\mathbf{F}_x$ and $\mathbf{F}_y$:

$$\begin{cases} \mathbf{MHA}(\mathbf{F}_x^q, \mathbf{F}_y^k, \mathbf{F}_y^v) = \mathbf{W}_2[h_1; h_2; \ldots; h_n], \\ h_i = softmax(\frac{\mathbf{F}_x^q(\mathbf{F}_y^k)^\top}{\sqrt{d_h}})\mathbf{F}_y^v, \end{cases} \tag{3}$$

where n is the number of attention heads, h_i is the output of the i-th head, and $\mathbf{W}_2 \in \mathbb{R}^{d\times d}$ is a trainable parameter of fully-connected layer. To summarize, the CMIM is formulated as:

$$\text{CMIM}(\mathbf{F}_x, \mathbf{F}_y, \mathbf{F}_y) \Leftrightarrow \begin{cases} \hat{\mathbf{F}_x} = \mathbf{LN}(\mathbf{MHA}(\mathbf{F}_x, \mathbf{F}_y, \mathbf{F}_y) \oplus \mathbf{F}_x); \\ \mathbf{F}_x' = \mathbf{LN}(\mathbf{FFN}(\hat{\mathbf{F}_x}) \oplus \hat{\mathbf{F}_x}), \end{cases} \tag{4}$$

where $\oplus$ denotes element-wise addition. **LN** and **FFN** denote the layer norm and feed-forward layer, respectively.

In this work, we leverage the CMIM operation to further develop a **Two-stream Cross-Modality Interaction (CMI)** module, as shown in Fig. 4 (a). If taking $\mathbf{F}_q$ as the guided feature, we can implement the textual to visual interaction (q-to-v) as $\mathbf{F}_v' = \text{CMIM}(\mathbf{F}_v, \mathbf{F}_q, \mathbf{F}_q)$. Similarly, if taking $\mathbf{F}_v$ as the guided feature, the visual to textual interaction (v-to-q) is performed as $\mathbf{F}_q' = \text{CMIM}(\mathbf{F}_q, \mathbf{F}_v, \mathbf{F}_v)$. $\mathbf{F}_v'$ and $\mathbf{F}_q'$ denote the contextualized visual feature and textual feature, respectively, which are used for the latter localization reasoning. The two-stream CMI unifies the two types of cross-modality feature interaction, thus can effectively boost cross-modality reasoning.

Here, we also introduce some different instantiations for implementing cross-modality interaction, as briefly described as follows:

- **Modality-agnostic Interaction ($[v;q]$-to-$[v;q]$)).** It is a standard transformer layer. As shown in Fig. 4 (b), given the visual and textual features $\mathbf{F}_v$ and $\mathbf{F}_q$, we concatenate the features together $[\mathbf{F}_v; \mathbf{F}_q]$ as the input of the transformer layer. Then, the modality-agnostic interaction is performed by applying the multi-head self-attention operation. We split the output of the last layer normalization to get the contextualized visual feature $\mathbf{F}_v'$ and textual feature $\mathbf{F}_q'$.
- **Single-stream CMI (q-to-v).** As shown in Fig. 4 (c), we only consider the one-side CMI in Fig. 4 (a). That is, we use the visual feature as the *query* input and the textual feature as the *key* and *value* inputs. Then, the single-stream CMI is performed by the cross-attention, as shown in Fig. 4 (c). The contextualized visual feature is obtained by $\mathbf{F}_v' = \text{CMIM}(\mathbf{F}_v, \mathbf{F}_q, \mathbf{F}_q)$. The textual feature $\mathbf{F}_q$ is kept unchanged.
- **Single-stream CMI (v-to-q).** Similar to the single-stream CMI (q-to-v), we also consider another side CMI in Fig. 4 (a) and perform the single-stream CMI (v-to-q) by the cross-attention, as shown in Fig. 4 (d). The contextualized textual feature is obtained by $\mathbf{F}_q' = \text{CMIM}(\mathbf{F}_q, \mathbf{F}_v, \mathbf{F}_v)$.

Compared with the two-stream CMI, the modality-agnostic interaction in Fig. 4 (b) cannot effectively capture the correlation between natural language expression and the given image. It cannot accommodate the differing processing needs of each modality. The two single-stream CMIs in Fig. 4 (c) and (d) can just capture the unidirectional vision and language interaction, which are insufficient to solve the complex language-based VG task. More discussion and analysis about the cross-modality interaction are given in Sec. 4.2.

3.3 Modality-agnostic Decoder

Here, we utilize a decoder to learn a modality-agnostic regression token to predict the bounding box coordinates. Specifically, we employ a learnable token [REG] to capture the global context in the whole visual and textual sequences. Let the regression token be denoted as $\mathbf{f}_r \in \mathbb{R}^d$, we concatenate

Table 1. Statistics of the Visual Grounding datasets. #Avg. denotes the average words of referring expression.

Datasets	#Images	#Expressions	#Avg. words
RefCOCO	19,994	142,210	3.61
RefCOCO+	19,992	141,564	3.53
RefCOCOg	25,799	95,010	8.43
Flicke30K Entities	31,783	427,000	-
ReferItGame	20,000	120,072	3.61

it with $\mathbf{F}'_q, \mathbf{F}'_v$ to build a new sequence $\mathbf{Z}_0 \in \mathbb{R}^{(1+L_q+L_v)\times d}$. Then, $\mathbf{Z}_0$ is fed to a transformer block with N_{vl} layers as follows:

$$
\begin{aligned}
&[\mathbf{f}^*_r, \mathbf{F}^*_q, \mathbf{F}^*_v] = \text{Transformer}(\mathbf{f}_r, \mathbf{F}'_q, \mathbf{F}'_v) \\
\Leftrightarrow &\begin{cases}
\mathbf{Z}_0 = [\mathbf{f}_r; \mathbf{F}'_q; \mathbf{F}'_v] + \mathbf{F}_{pos}; \\
\mathbf{Z}'_n = \mathbf{LN}(\mathbf{MHA}(\mathbf{Z}_{n-1})) + \mathbf{Z}_{n-1}; \quad 1 \le n \le N_{vl}, \\
\mathbf{Z}_n = \mathbf{LN}(\mathbf{FFN}(\mathbf{Z}'_n)) + \mathbf{Z}'_n, \\
\mathbf{f}^*_r = \mathbf{Z}^0_n,
\end{cases}
\end{aligned}
\tag{5}
$$

where $\mathbf{f}_r$ is randomly initialized before training, $\mathbf{F}_{pos} \in \mathbb{R}^{(1+L_q+L_v)\times d}$ denotes positional embedding as stated in [37], and N_{vl} is set to 6 empirically.

In this work, only the token [REG] $\mathbf{f}^*_r \in \mathbb{R}^d$ is used to predict the target bounding box, rather than multi-modal features $[\mathbf{F}^*_q, \mathbf{F}^*_v]$ as in previous work [50, 52, 54, 55]. Based on token $\mathbf{f}^*_r$, we adopt a feed-forward module named $\mathbf{FFN}_{reg}$ to predict the target bounding box $b = [x, y, w, h] \in [0, 1]^4$, where $[x, y]$ denotes the center coordinates of the bounding box, w and h are width and height of the bounding box, respectively. The ground truth is denoted as $\hat{b} \in [0, 1]^4$. $\hat{b}$ is a normalized ratio of the entire image region. As shown in Fig. 3, the predicted bounding box b is calculated by $b = \mathbf{FFN}_{reg}(\mathbf{f}^*_r)$.

3.4 Training

To optimize the proposed approach, we use a combination of two widely used losses as the objective function. The first term is the smooth-l1 loss $\mathcal{L}_{s\text{-}l_1}$ [8], and the second term is the Generalized IoU loss $\mathcal{L}_{giou}$ [33]. The overall loss is formulated as below:

$$
\mathcal{L} = \mathcal{L}_{s\text{-}l_1}(b, \hat{b}) + \lambda \cdot \mathcal{L}_{giou}(b, \hat{b}), \tag{6}
$$

where λ is a hyper-parameter to modulate the effect of the $\mathcal{L}_{giou}$ loss. During Inference, we take the predicted bounding box b as the final output.

4 EXPERIMENTS

4.1 Experimental Setup

Datasets. We experiment on five benchmark datasets to evaluate the effectiveness of the proposed method. The statistics of datasets are summarized in Table 1. **(1) RefCOCO/RefCOCO+/RefCOCOg.** The RefCOCO [53] dataset consists of 19,994 images with 142,210 referring expressions. According to the split strategy [49, 50], it is split into train/val with 120,624/10,834 expressions, respectively. The test part consists of testA/testB with 5,657/5,095 expressions, respectively. The RefCOCO+ [53] dataset consists of 19,992 images with 141,564 referring expressions. The RefCOCOg [27] dataset consists of 25,799 images with 95,010 referring expressions. There are two commonly split practices - RefCOCOg-google [27] and RefCOCO-umd [28]. For a fair comparison, we report the results on the validation set of both RefCOCOg-google and RefCOCOg-umd (val-u and test-u). **(2) ReferItGame.**

Table 2. Ablation Studies of different instantiations of *Cross-Modality Interactor* on the RefCOCO and ReferItGame datasets with ResNet-50 features in term of Top-1 Accuracy (%).

CMI	RefCOCO			ReferItGame	
	val↑	testA↑	testB↑	val↑	test↑
(a) Two-stream CMI (Ours)	**81.14**	**84.05**	**76.45**	**73.17**	**70.72**
(b) Modality-agnostic Interaction	79.25	82.28	75.12	72.23	70.06
(c) Single-stream CMI (q-to-v)	79.79	82.23	75.51	72.06	69.27
(d) Single-stream CMI (v-to-q)	78.78	80.16	74.57	71.44	69.08

The ReferItGame dataset [18] consists of 31,783 images with 427,000 referring expressions. The dataset is split into train/validation/test sets with 54,127/5,842/60,103 referring expressions [4, 49]. **(3) Flick30K Entities.** The Flickr30K Entities [30] consists of 31,783 images with 427,000 referring expressions. There are 29,783 /1,000/1,000 images for train/validation/test sets. Following common practice, we report the experimental results on the test set.
Evaluation Metric. Following the convention [4, 50], we adopt IoU as an evaluation metric, *i.e.*, Intersection over Union (IoU) between the predicted bounding box and the ground-truth. If the IoU value is larger than 0.5, we treat the predicted result as a true positive, otherwise a false positive. The Top-1 accuracy (%) is the ratio of predicted positive results at the Top-1 rank.
Implementation Details. (1) Data preparation. Each input image is scaled into 640×640. We truncate each referring expression with the maximum length of 20 on all datasets except RefCOCOg with 40. **(2) Module setting.** We use the encoder of DETR [1] to initialize the transformer layer in the image encoder and take the uncased version of BERT [5] as the language encoder. As for the parameters of the other components, we initialize them with Xavier init. The $\mathbf{FFN}_{reg}$ module is implemented by three fully-connected layers and a *sigmoid* activation layer. **(3) Training details.** We use the AdamW [26] optimizer and set the dropout rate to 0.1 in the whole model. The batch size is set to 8 in all our experiments. During training, the image and language encoders are jointly optimized with the entire model. The initial learning rate of the modality-agnostic decoder is set to 10^{-4}, and the learning rate of other modules is set to 10^{-5}. For the RefCOCO, RefCOCOg, and ReferItGame datasets, we train the model in 90 epochs with a learning rate dropped by a factor of 10 after 60 epochs. For the RefCOCO+ dataset, the training epoch is set to 180, and the epoch of the learning rate drop is set to 120. For the Flickr30K entities dataset, we train the model in 60 epochs with a learning rate that drops after 40 epochs. The unified dimension d is set to 256. The hyper-parameter λ in Equation 6 is set to 1 on all datasets. **(4) Inference:** The proposed method predicts the target bounding box with one-stage token regression. There is no extra post-processing.

4.2 Ablation Studies

Effect of different CMI instantiations. In this section, we investigate the effect of different instantiations of cross-modality interaction as shown in Fig. 4. Table 2 shows the performance comparison on the RefCOCO and ReferItGame datasets. We have the following observations from Table 2:

- Overall, we can find from Table 2 that the proposed two-stream CMI (q-to-v&v-to-q) in Fig. 4 **(a)** achieves the best performance on both RefCOCO (76.45% on testB set) and ReferItGame (70.72% on test set) datasets, which validates the effectiveness of the two-stream CMI on capturing the cross-modality correlation. Therefore, the two-stream CMI is finally adopted in our framework.
- The single-stream CMIs (c) q-to-v and (d) v-to-q are sub-optimal, *e.g.*, reporting 69.27% and 69.08% vs. the best 70.72% on the ReferItGame test set, respectively. This is because such

Table 3. Ablation Studies of Feature Encoder (with *intra-modality Interaction*) on the ReferItGame dataset with ResNet-50 features in term of Top-1 Accuracy (%). ✓ denotes the module is enable, and – denotes disabled.

Transformer Layer		ReferItGame	
Image Encoder	Language Encoder	val↑	test↑
-	-	69.17	66.10
-	✓	70.98	67.99
✓	-	71.00	69.07
✓	✓	**73.17**	**70.72**

single-stream interaction only updates the one-side modality based on the unidirectional cross-attention mechanism, which hinders the internal information exchange between vision and language.

- The modality-agnostic interaction in Fig. 4 (b) performs worse than the two-stream CMI. This reflects that a simple concatenation of v and q as the input of multi-head self-attention module cannot effectively accommodate the differing processing needs of each modality, thus resulting in sub-optimal cross-modality representation. The fine-grained cross-modality interaction of visual to textual (v-to-q) and textual to visual (q-to-v) should be considered simultaneously.

In Fig. 5, we visualize the predicted results of three samples to validate the effectiveness of cross-modality interaction strategies ((a) ~ (d) in Fig. 4). Taking the expression **(1)** *"top right donut"* as an example, the proposed two-stream CMI obtains the best grounding results, while other interaction strategies fail due to the complex background and subjects (*e.g.*, other donuts). Similarly, for the expression **(2)**, our method correctly locates the target *"checkered phone"* by the mutual verification of inter-modality interaction (v-to-q & q-to-v). For the expression **(3)** *"first potted plant"*, only the proposed interaction correctly locates the target. The instantiations (b) and (d) merely locate the *"pot"* and the instantiation (c) locates the target yet with too excessive background.

Effect of transformer layer in feature encoder. Both the image and language encoders use the transformer layer [37]. As shown in Table 3, compared with the complete disablement of transformer in both image and language feature encoders, using only the visual transformer improves the performance from 66.10% to 69.07% on the test set. Using only the text transformer also has an increase in performance (*i.e.*, 66.10% to 67.99%), but the improvement is less than using visual transformer. Ultimately, the model achieves the best performance by combining the two transformers. These results validate that the visual feature is more crucial than the textual feature for the visual grounding task, and also validate the necessity of transformers (actually the effect of intra-modality interaction in the transformer) for feature encoding.

4.3 Comparison with State-of-the-arts

To verify the effectiveness of our model, we evaluate our model on five public benchmark datasets and compare it with the state-of-art methods as follows: two-stage methods [15, 19, 22, 25, 28, 29, 39, 52, 54, 55], one-stage methods [2, 21, 34, 43, 49–51], and transformer-based methods [4, 6].

Results on RefCOCO, RefCOCO+ and RefCOCOg. As shown in Table 4, on the **RefCOCO** dataset, the proposed method achieves the best accuracy on the sets of val and testA (*i.e.*, 81.92% and 84.05%), and obtains the second best on the testB set (*i.e.*, 77.3%). In addition, our method outperforms all other previous works of two-stage and one-stage methods [22, 25, 43, 51]. On the **RefCOCO+** dataset, our proposed method achieves comparable results. Compared with the

Table 4. Comparison with state-of-the-art methods on RefCOCO, RefCOCO+ and RefCOCOg in term of Top-1 Accuracy (%).

Models	Venue	Visual Feature	RefCOCO			RefCOCO+			RefCOCOg		
			val↑	testA↑	testB↑	val↑	testA↑	testB↑	val-g↑	val-u↑	test-u↑
Two-stage methods											
Neg Bag [28]	*ECCV'16*	VGG-16	57.3	58.6	56.4	-	-	-	39.5	*	49.5
CMN [15]	*CVPR'17*	VGG-16	-	71.03	65.77	-	54.32	47.76	57.47	-	-
VC [55]	*CVPR'18*	VGG-16	-	73.33	67.44	-	58.40	53.18	62.30	-	-
ParalAttn [57]	*CVPR'18*	VGG-16	-	75.31	65.52	-	61.34	50.86	58.03	-	-
MAttNet [52]	*CVPR'18*	ResNet-101	76.65	81.14	69.99	65.33	71.62	56.02	-	66.58	67.27
LGRANs [40]	*CVPR'19*	VGG-16	-	76.60	66.40	-	64.00	53.40	61.78	-	-
DGA [42]	*ICCV'19*	VGG-16	-	78.42	65.53	-	69.07	51.99	-	-	63.28
RvG-Tree [14]	*TPAMI'19*	ResNet-101	75.06	78.61	69.85	63.51	67.45	56.66	-	66.95	66.51
NMTree [22]	*ICCV'19*	ResNet-101	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44
CM-A-E [25]	*CVPR'19*	ResNet-101	78.35	83.14	71.32	68.09	**73.65**	58.03	-	67.99	68.67
One-stage methods											
SSG [2]	*arXiv'18*	DarkNet-53	-	76.51	67.50	-	62.14	49.27	47.47	58.80	-
FAOA [50]	*ICCV'19*	DarkNet-53	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.36
RCCF [21]	*CVPR'20*	DLA-34	-	81.06	71.85	-	70.35	56.32	-	-	65.73
ReSC [49]	*ECCV'20*	DarkNet-53	76.59	78.22	73.25	63.23	66.64	55.53	60.96	64.87	64.87
ReSC-Large [49]	*ECCV'20*	DarkNet-53	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20
SAFF [51]	*ACM MM'21*	DarkNet-53	79.26	81.09	76.55	64.43	68.46	58.43	-	68.94	68.91
Transformer-based methods											
TransVG [4]	*ICCV'21*	ResNet-50	80.32	82.67	**78.12**	63.50	68.15	55.63	66.56	67.66	67.44
VGTR [6]	*ICME'22*	ResNet-50	78.70	82.09	73.31	63.57	69.65	55.33	62.88	65.62	65.30
Ours	–	ResNet-50	**81.14**	**84.05**	76.45	**66.81**	**71.99**	**58.70**	**68.73**	**68.06**	**68.87**
TransVG [4]	*ICCV'21*	ResNet-101	81.02	82.72	**78.35**	64.82	70.70	56.94	67.02	68.67	67.73
VGTR [6]	*ICME'22*	ResNet-101	79.30	82.16	74.38	64.40	70.85	55.84	64.05	66.83	67.28
Ours	–	ResNet-101	**81.92**	**83.40**	77.37	**68.49**	72.18	**60.30**	**68.39**	**69.08**	**69.04**

one-stage state-of-the-art methods **SAFF** [51], our method surpasses it by a large margin (*i.e.*, 68.49% vs. 64.43% on val set, 72.18% vs. 68.46% on the testA set, 60.30% vs. 58.43% on the testB set). When compared with the best two-stage method **CM-A-E** [25], our method surpasses it on the val and testB sets except for the testA set. The **RefCOCOg** dataset consists of two split strategies (*i.e.*, google-split and umd-split); the proposed method achieves state-of-the-art performance on both two split strategies. Compared with the best two-stage method **SAFF**, even though **SAFF** uses a scene graph to build a semantic filter for visual feature mining, our method surpasses it. About **CM-A-E** with best performance in one-stage methods, our method also outperforms it (*i.e.*, 69.08% vs. 67.99% on val set, 69.41% vs. 68.47% on test set.)

Results on ReferItGame and Flick30K Entities. As shown in Table 5, our method outperforms all the other methods. On the **ReferItGame** dataset, **DDPN** [54] is the best model in two-stage methods, our method surpasses it by a large margin (*i.e.*, 71.07% vs. 63.00%). **SAFF** [51] achieves the best performance in the one-stage methods. Our proposed method performs significantly better than it, *i.e.*, 70.72% vs. 66.01% with ResNet-50, 71.07% vs. 66.01% with ResNet-101. Compared with the best transformer-based method **TransVG** [4], our method improves the best accuracy from 70.73% to 71.07%. On the larger-scale **Flick30K Entities** dataset, the proposed method also achieves better results than existing models, including **VGTR** [6] (improving the accuracy from 75.32% to 79.15% with ResNet-101, from 74.17% to 79.12% with ResNet-50), where **VGTR** also is a transformer architecture model. Compared with the state-of-the-art one-stage network **SAFF** which uses semantic-aware textual features to filter visual features, our method also performs significant improvements (*i.e.*, improves from 70.17% to 79.15%). Compared with the best two-stage method **DDPN** that dedicates to generating high-quality proposals, our method also achieves improvements by a large margin (*e.g.*, lifts accuracy from 73.30% to 79.15%). These results provide

Table 5. Comparison with state-of-the-art methods on the test set of ReferItGame and Flickr30K Entities in term of Top-1 accuracy (%). The best and second best performance with **Bold** and Underline.

Models	Venue	Visual Feature	ReferItGame test↑	Flickr30K test↑
Two-stage methods				
CMN [15]	*CVPR'17*	VGG-16	28.33	-
VC [55]	*CVPR'18*	VGG-16	31.13	-
MAttNet [52]	*CVPR'18*	ResNet-101	29.04	-
Similarity Net [39]	*TPAMI'18*	ResNet-101	34.54	60.89
CITE [29]	*ECCV'18*	ResNet-101	35.07	61.33
PIRC [19]	*ACCV'18*	ResNet-101	59.13	72.83
DDPN [54]	*IJCAI'18*	ResNet-101	63.00	73.30
One-stage methods				
SSG [2]	*arXiv'18*	DarkNet-53	54.24	-
ZSGNet [34]	*ICCV'19*	ResNet-50	58.63	63.39
FAOA [50]	*ICCV'19*	DarkNet-53	60.67	68.71
RCCF [21]	*CVPR'20*	DLA-34	63.79	-
ReSC [49]	*ECCV'20*	DarkNet-53	64.33	69.04
ReSC-Large [49]	*ECCV'20*	DarkNet-53	64.60	69.28
LSPN [43]	*ECCV'20*	DarkNet-53	-	69.53
SAFF [51]	*ACM MM'21*	DarkNet-53	66.01	70.17
Transformer-based methods				
TransVG [4]	*ICCV'21*	ResNet-50	69.76	78.47
VGTR [6]	*ICME'22*	ResNet-50	-	74.17
Ours	–	ResNet-50	**70.72**	**79.12**
TransVG [4]	*ICCV'21*	ResNet-101	70.73	79.10
VGTR [6]	*ICME'22*	ResNet-101	-	75.32
Ours	–	ResNet-101	**71.07**	**79.15**

strong evidence that the effectiveness of cross-modality interaction in one-stage visual grounding. Its superior performance also provides more insights for researchers to consider cross-modality interaction when designing a novel one-stage network for multi-modal tasks.

4.4 Qualitative Visualization and Analysis

Visualization of cross-modality interaction. To demonstrate the interpretability of the cross-modality interaction, we visualize some cases in Fig. 6, which illustrate the cross-modality interaction on both visual and textual sequences. In each example illustration block of Fig. 6, we display the original image and referring expression in the left column, and display the attention map of textual to visual (q-to-v) in the right region. The color bar appearing above the query words is the attention map of visual to textual (v-to-q). **The textual to visual interaction** usually pays attention to the relevant cues in the image. Taking the expression **(b)** as an example, the token *"elephant"* highlights the surrounding of two elephants. The other tokens usually focus on the surrounding of the target object (*e.g.*, tokens *"in"* and *"front"* highlight the surrounding of the front pizza in expression **(c)**). The tokens "[CLS]" and "[SEP]" typically represent the global semantics of the sentence, and these tokens highlight the target region. Taking the expression **(a)** "*right upper bear*" as an example, both tokens "[CLS]" and "[SEP]" highlight the target *"bear"* region. The subject token typically focuses on the surrounding of the subject. **The visual to textual interaction** is more intuitive than textual

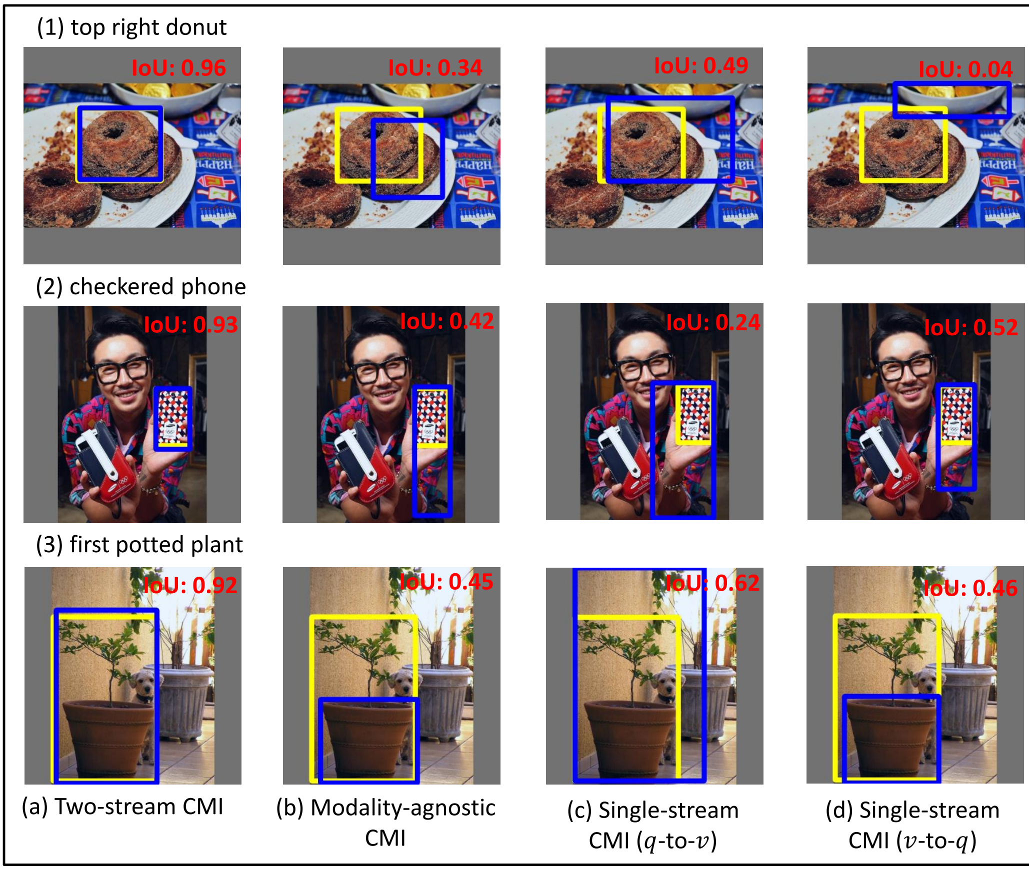

Fig. 5. Visualization of prediction results on the RefCOCO test set with different instantiations of cross-modality interaction. The instantiation (a) is adopted in our method. The bounding boxes of predicted region and ground-truth are marked with blue and yellow, respectively.

to visual attention. If the expression contains obvious position information (*i.e.*, *"upper right"* in expression **(a)** and *"in front"* in expression **(c)**), the visual feature typically guides the model to focus on the subject word, which is essential to locate the object. Then, the model is turned to focus on the words about position relation (*i.e.*, *"upper right"* and *"front"*). If the expression is a sentence that requires an overall understanding of the semantics (*i.e.*, expression **(b)** and **(d)**), the model typically first focuses on the token "[CLS]" containing global semantics, and then focuses on the words of important attributes (*i.e.*, *"sun"* and *"white"*).

In other words, for the textual to visual attention, we notice that the attention of "[CLS]" token is typically distributed on the target object. The attention region of abstract word (*e.g.*, "in" in query (b), "in" and "the" in query (c), and "a" in query (d)) is usually ambiguous. We consider there is no specific visual appearance of the abstract word, so it is hard for the model to give a corresponding response. For visual to textual attention, the image usually contains complex objects, resulting in always the subject word being highlighted. In summary, the proposed method incorporates the above cross-modality interaction, boosting the alignment of visual and language semantics. Here, the model will relieve strong intra-modality correlation cues in the feature encoding stage in the

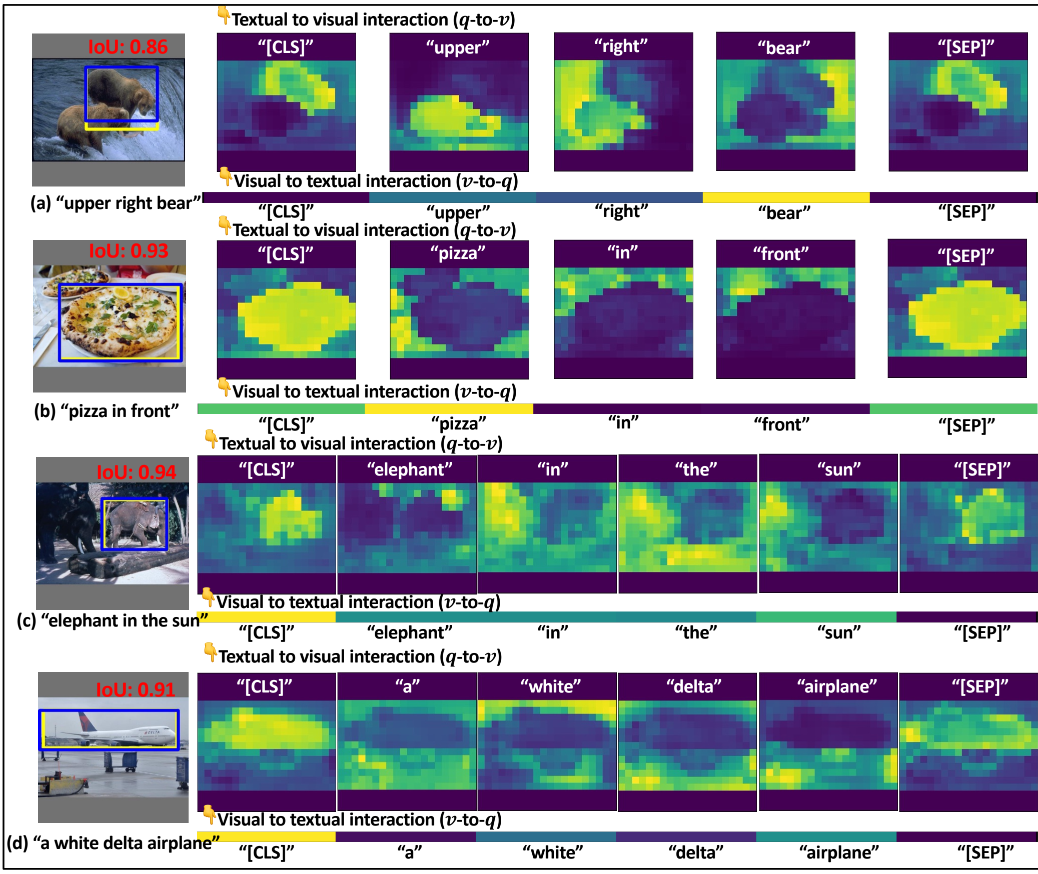

Fig. 6. Visualization of attention map in cross-modality interaction (better viewed in color).

way of mutual verification. In addition, this result also validates that cross-modality interaction can discover crucial visual regions and word tokens for visual grounding.

Visualization of token's attention on image and query expression. As shown in Fig. 7, we visualize the regression token's attention map of the visual and textual sequences in the modality-agnostic decoder step by step. The leftmost column is the original image and expression, while the right columns are the attention maps from layer 0 to layer 5 in the modality-agnostic decoder. For the expression *"left dog"*, the model first attends to textual words *"dog"* and the visual regions related to the two *"dog"* subjects at $\{0, 1\}$-th layers. Next, it captures the global semantic of expression *"left dog"* at $\{2, 3, 4\}$-th layers, and the attention also shifted to the visual region of *"left dog"*. Eventually, the model locates the correct region of *"left dog"* at the last layer.

Visualization of grounding results. To further display the effectiveness of our method, we visualize some challenging examples in RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame datasets. **(1) Redundant objects.** As shown in Fig. 8 (a) ~ (d), the images contain more than one objects of the same type (*i.e.*, *baby*, *giraffe*, *guy*, *pizza*), which requires the model to identify the target object from multiple similar objects. **(2) Indistinguishable objects.** Taking the expression (e) *"a man in a black hat"* as an example, there are two men with different color hats (*i.e.*, black and white) in the image, and the visual region occupied by the hat is very small and indistinguishable. This

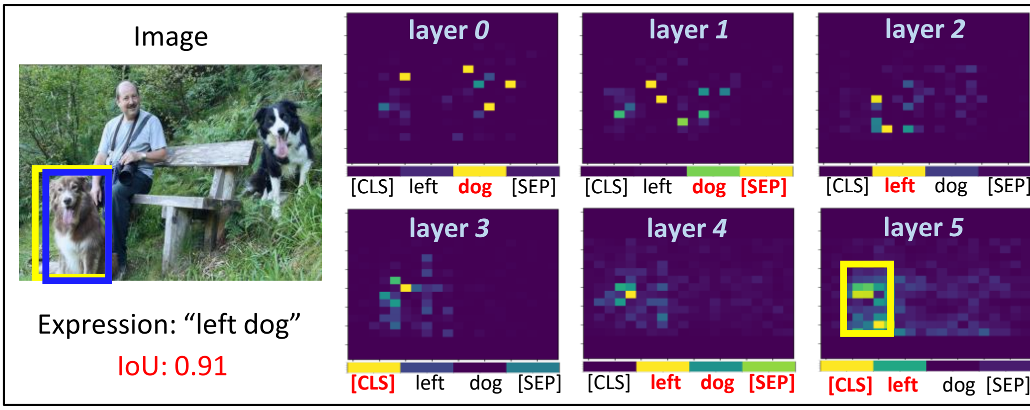

Fig. 7. Visualization of token's attention to both image and referring expression in the modality-agnostic decoder on the RefCOCO test set.

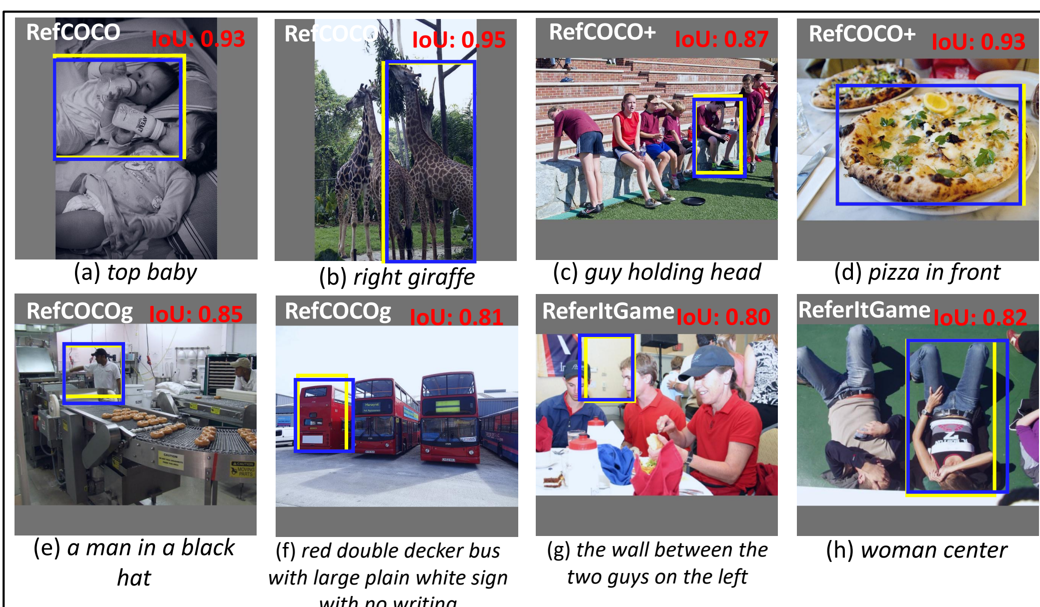

Fig. 8. Visualization of predicted results on RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame datasets.

brings a big difficulty to correctly locate the target "man". **(3) Complex expression/background.** Taking expression (f) as an example, the expression is long and complex, and it is hard to understand which object needs to be grounded. For the expressions (g) and (h) in the ReferItGame dataset, the image contains a variety of complex objects and backgrounds. In the above challenging cases, our model locates the target with a large IoU value, and these results validate the effectiveness of our method.

Visualization of failure cases. As shown in Figure 9, we visualize some failure cases of our proposed model. (1) For query (a), five buses appear in the figure, and the bus in the middle is

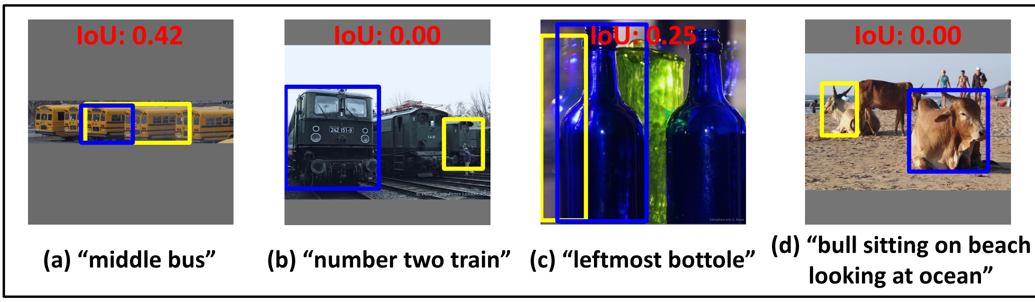

Fig. 9. Some challenging but failed cases of the proposed model. Yellow boxes mark the ground-truth, and blue boxes mark the prediction results.

obscured by other buses. It is hard to predict the right car. Our model predicts the middle region of the image that covers three cars; (2) for query (b), the model is required to locate the train with "number two", our model focuses on the text – number "2"; (3) for query (c), the ground truth is the shadow of the leftmost bottle, and this is ambiguous for the model to predict; (4) at last, for query (d), two bulls are sitting on the beach, and "looking" is hard to understand for the model. Therefore, the cases such as occluded objects, blurred images, interference from OCR markers, and unintelligible queries still remain challenging for visual grounding.

5 CONCLUSIONS

In this work, we develop a transformer-based end-to-end visual grounding approach. It mainly consists of a feature encoder, a cross-modality interactor, and a modality-agnostic decoder, which can effectively and progressively capture the intra-modality and inter-modality correlation, thus boosting the cross-modality reasoning for the visual grounding task. We conduct extensive experimental evaluations, including qualitative and quantitative ablation studies and analyses, on five benchmark datasets. Experimental results clearly demonstrate the effectiveness of our approach. In the future, we will explore stronger and more stable modality interaction structure for visual grounding. Besides, we will try to address the confounding bias, *e.g.*, language bias, in our end-to-end visual grounding framework to improve the generalization ability of the model.

ACKNOWLEDGMENTS

We would like to thank the reviewers for their constructive suggestions. This work was supported by the National Natural Science Foundation of China (62272144, U20A20183, 62020106007, and 72188101) and the Major Project of Anhui Province (202203a05020011).

REFERENCES

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*. 213–229.

[2] Xinpeng Chen, Lin Ma, Jingyuan Chen, Zequn Jie, Wei Liu, and Jiebo Luo. 2018. Real-time referring expression comprehension by single-stage grounding network. *arXiv preprint arXiv:1812.03426* (2018).

[3] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. 2018. Using syntax to ground referring expressions in natural images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[4] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. TransVG: End-to-End Visual Grounding With Transformers. In *Proceedings of the IEEE International Conference on Computer Vision*. 1769–1779.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805* (2018).

[6] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. 2022. Visual Grounding with Transformers. In *IEEE International Conference on Multimedia and Expo.*

[7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision.* 5267–5275.

[8] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision.* 1440–1448.

[9] Dan Guo, Kun Li, Zheng-Jun Zha, and Meng Wang. 2019. Dadnet: Dilated-attention-deformable convnet for crowd counting. In *Proceedings of the 27th ACM international conference on multimedia.* 1823–1832.

[10] Dan Guo, Hui Wang, Hanwang Zhang, Zheng-Jun Zha, and Meng Wang. 2020. Iterative context-aware graph inference for visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 10055–10064.

[11] Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2017. Online early-late fusion based on adaptive hmm for sign language recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 1 (2017), 1–18.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision.* 2961–2969.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 770–778.

[14] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. 2022. Learning to compose and reason with language tree structures for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), 684–696.

[15] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 1115–1124.

[16] Weike Jin, Zhou Zhao, Yimeng Li, Jie Li, Jun Xiao, and Yueting Zhuang. 2019. Video question answering via knowledge-based progressive spatial-temporal attention network. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 2s (2019), 1–22.

[17] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. MDETR-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE International Conference on Computer Vision.* 1780–1790.

[18] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.* 787–798.

[19] Rama Kovvuri and Ram Nevatia. 2018. Pirc net: Using proposal indexing, relationships and context for phrase grounding. In *Asian Conference on Computer Vision.* 451–467.

[20] Kun Li, Dan Guo, and Meng Wang. 2021. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1902–1910.

[21] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. 2020. A Real-Time Cross-modality Correlation Filtering Method for Referring Expression Comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 10880–10889.

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

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

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision.* 21–37.

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

[26] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[27] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 11–20.

[28] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *Proceedings of the European Conference on Computer Vision.* 792–807.

[29] Bryan A. Plummer, Paige Kordas, M. Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. 2018. Conditional Image-Text Embedding Networks. In *Proceedings of the European Conference on Computer Vision.* 249–264.

[30] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of*

the IEEE International Conference on Computer Vision. 2641–2649.
[31] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9982–9991.
[32] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv:1804.02767* (2018).
[33] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 658–666.
[34] Arka Sadhu, Kan Chen, and Ram Nevatia. 2019. Zero-shot grounding of objects from natural language queries. In *Proceedings of the IEEE International Conference on Computer Vision*. 4694–4703.
[35] Ling Shen, Richang Hong, Haoran Zhang, Xinmei Tian, and Meng Wang. 2019. Video retrieval with similarity-preserving deep temporal hashing. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 4 (2019), 1–16.
[36] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *International Journal of Computer Vision* (2013), 154–171.
[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[38] Hui Wang, Dan Guo, Xian-Sheng Hua, and Meng Wang. 2021. Pairwise VLAD Interaction Network for Video Question Answering. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5119–5127.
[39] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2018. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018), 394–407.
[40] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1960–1968.
[41] Shuo Wang, Dan Guo, Xin Xu, Li Zhuo, and Meng Wang. 2019. Cross-modality retrieval by joint correlation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 2s (2019), 1–16.
[42] Sibei Yang, Guanbin Li, and Yizhou Yu. 2019. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE International Conference on Computer Vision*. 4644–4653.
[43] Sibei Yang, Guanbin Li, and Yizhou Yu. 2020. Propagating over phrase relations for one-stage visual grounding. In *Proceedings of the European Conference on Computer Vision*. 589–605.
[44] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. 2020. Tree-augmented cross-modal encoding for complex-query video retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 1339–1348.
[45] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–10.
[46] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. 2020. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *Proceedings of the 28th ACM international conference on multimedia*. 1939–1947.
[47] Xun Yang, Meng Wang, Richang Hong, Qi Tian, and Yong Rui. 2017. Enhancing person re-identification in a self-trained subspace. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 3 (2017), 1–23.
[48] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. 2022. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing* 31 (2022), 1204–1216.
[49] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. 2020. Improving one-stage visual grounding by recursive sub-query construction. In *Proceedings of the European Conference on Computer Vision*. 387–404.
[50] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*. 4683–4693.
[51] Jiabo Ye, Xin Lin, Liang He, Dingbang Li, and Qin Chen. 2021. One-Stage Visual Grounding via Semantic-Aware Feature Filter. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1702–1711.
[52] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1307–1315.
[53] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision*. 69–85.
[54] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. 2018. Rethinking diversified and discriminative proposal generation for visual grounding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 1114–1120.

[55] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. 2018. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4158–4166.

[56] Qi Zheng, Jianfeng Dong, Xiaoye Qu, Xun Yang, Yabing Wang, Pan Zhou, Baolong Liu, and Xun Wang. 2021. Progressive Localization Networks for Language based Moment Localization. *ACM Transactions on Multimedia Computing, Communications, and Applications* (2021).

[57] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. 2018. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4252–4261.

[58] C Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision*. 391–405.