Shifting More Attention to Visual Backbone: Query-modulated Refinement Networks for End-to-End Visual Grounding

Jiabo Ye[1] Junfeng Tian[2] Ming Yan[2] Xiaoshan Yang[3]
Xuwu Wang[4] Ji Zhang[2] Liang He[1] Xin Lin[1]
[1]East China Normal University, Shanghai, China [2]Alibaba Group, Hangzhou, China
[3]NLPR, CASIA, Beijing, China [4]Fudan University, Shanghai, China
jiabo.ye@stu.ecnu.edu.cn, {xlin,lhe}@cs.ecnu.edu.cn, xwwang18@fudan.edu.cn
xiaoshan.yang@nlpr.ia.ac.cn, {tjf141457,ym119608,zj122146}@alibaba-inc.com

Abstract

Visual grounding focuses on establishing fine-grained alignment between vision and natural language, which has essential applications in multimodal reasoning systems. Existing methods use pre-trained query-agnostic visual backbones to extract visual feature maps independently without considering the query information. We argue that the visual features extracted from the visual backbones and the features really needed for multimodal reasoning are inconsistent. One reason is that there are differences between pre-training tasks and visual grounding. Moreover, since the backbones are query-agnostic, it is difficult to completely avoid the inconsistency issue by training the visual backbone end-to-end in the visual grounding framework. In this paper, we propose a Query-modulated Refinement Network (QRNet) to address the inconsistent issue by adjusting intermediate features in the visual backbone with a novel Query-aware Dynamic Attention (QD-ATT) mechanism and query-aware multiscale fusion. The QD-ATT can dynamically compute query-dependent visual attention at the spatial and channel levels of the feature maps produced by the visual backbone. We apply the QRNet to an end-to-end visual grounding framework. Extensive experiments show that the proposed method outperforms state-of-the-art methods on five widely used datasets. Our code is available at https://github.com/LukeForeverYoung/QRNet.

1. Introduction

Visual grounding [25, 32, 36, 59], i.e., localizing the referent object in an image according to the given natural language query, is a fundamental component of multimodal reasoning system. Compared with conventional object detection methods [38, 39] which can only recognize the restricted categories contained in the training data, visual grounding has the advantage of detecting novel combinations of categories and attributes expressed in free-form text. In recent years, it has attracted much attention in the field of computer vision and machine learning due to its potential applications in many downstream tasks, such as visual question answering [15, 50, 63], visually-grounded language navigation [2, 47] and image captioning [1, 8, 57].

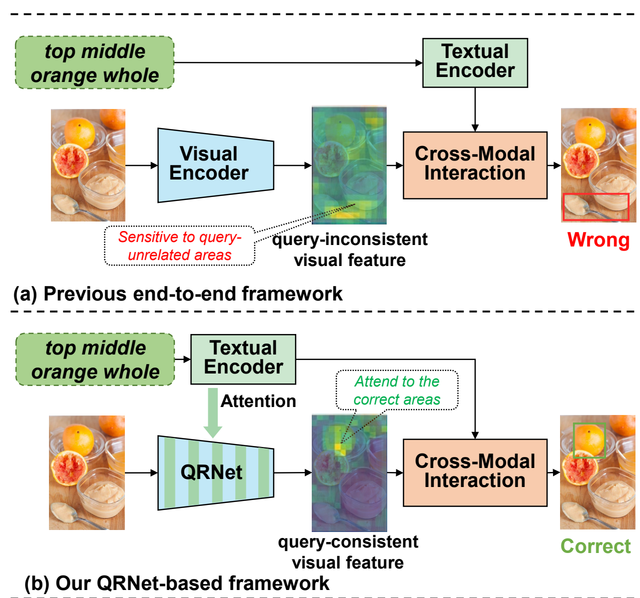

Figure 1. (a) A typical end-to-end visual grounding framework that uses two individual encoders to extract visual and textual features for cross-modal interaction. (b) Our visual grounding framework based on a Query-modulated Refinement Network (QRNet).

The work was done when Jiabo Ye was working as an intern at Alibaba DAMO Academy.

The early methods of visual grounding focus on extending the popularly used one-stage and two-stage object detection architectures. One-stage methods [11, 23, 54, 56] use a pre-trained fully convolutional network (e.g. Darknet53 [38], ResNet [20]) to directly extract pixel-level feature maps and leverage manually-defined dense anchors to return the most likely candidate for the query text. These methods are easy and efficient for learning or inference, but they cannot perform well on complex queries that have various objects and relations. Two-stage methods [52, 53, 58] use an off-the-shelf detector (e.g. Faster R-CNN [39]) to extract the region proposals and return the one that best matches the query text using the modality-shared representations. These methods always have better performance than the one-stage ones by introducing more complicated multimodal fusion and reasoning mechanisms [33, 52, 53]. However, the complicated fusion modules cannot be jointly learned with the detectors, which may limit their ability in multimodal reasoning. More recently, Transformer [46] has been applied in visual grounding [11, 24] to conduct the multimodal reasoning more succinctly based on pixel-level feature maps without region proposals or dense anchors.

Although existing visual grounding methods, especially the Transformer-based methods [11, 24], have achieved promising results, we argue that they do not pay enough attention to the visual backbone which plays a crucial role in effective multimodal reasoning. Since the visual backbone determines whether all integral visual content in the image is successfully extracted for matching the query text. Currently, the most widely used backbones are the CNN model (e.g., ResNet [20]) pre-trained for image classification on ImageNet and the detector (e.g., Faster R-CNN [39] and Mask R-CNN [19]) pre-trained for general object detection of close-set categories. Therefore, the difference between the visual grounding task and the pre-training task of the backbones may lead to an **inconsistency** between the visual features produced by the backbones and the ones really needed for the multimodal reasoning. As shown in Figure 1(a), the pre-trained visual backbone extracts general purposed visual features sensitive to regions that may contain the objects of pre-defined categories. Whereas, the visual grounding requires the backbone to localize a different object referred to by the query. A straightforward way to alleviate the inconsistency is learning the visual grounding model in an end-to-end form as in [11]. However, it still cannot completely avoid the inconsistency because the backbone is query-agnostic. In other words, given the same image, the query-agnostic backbone will always output the same feature map no matter what the query sentences are.

In this paper, we propose a Query-modulated Refinement Network (QRNet) to address the **inconsistency** issue. As shown in Figure 1(b), the proposed QRNet can produce query-consistent features by adjusting the feature maps of the visual backbone with the guidance of the query text, which benefits the cross-modal alignment between the query and the relevant region to make a correct prediction. The QRNet is designed based on Swin-Transformer [31] and a novel Query-aware Dynamic Attention (QD-ATT), which can help the QRNet extract query-refined visual feature maps from the visual backbone and fuse the multiscale features with the query guidance. The QD-ATT dynamically computes textual-dependent visual attentions at the spatial and channel level of the feature maps produced by the visual backbone. The spatial and channel attentions are further multiplied with the original feature maps to obtain query-refined hierarchical visual feature maps. To comprehensively consider the fine-grained visual features of the candidate regions at different scales, we aggregate the query-refined visual feature maps obtained at different stages of the QRNet by a query-aware multiscale fusion scheme.

We instantiate the proposed QRNet by building a flexible visual grounding framework based on the recently proposed TransVG [11]. We adopt the same multi-layer visual-linguistic transformer as in [11] to perform intra- and inter-modal reasoning based on the output token sequence of the QRNet. The complete pipeline significantly outperforms existing methods, e.g., TransVG [11] (3.75% on ReferItGame and 2.85% on Flickr30K Entities). Note that the proposed QD-ATT can be easily applied to other pre-trained visual backbones, e.g., ResNet [20].

The main contributions of this paper are three-fold:

- We propose a query-modulated refinement network to address the inconsistency issue caused by the pre-trained visual backbone through adjusting the visual feature maps with the guidance of query text.
- We propose a novel query-aware dynamic attention mechanism, which can dynamically compute query-dependent spatial and channel attentions for refining visual features.
- We build a flexible visual grounding framework based on the query-modulated refinement network and demonstrate that it achieves significantly better performance than existing methods on five widely used public datasets.

2. Related Work

2.1. Visual Grounding

The visual grounding methods can be categorized into two-stage methods and one-stage methods.

Two-stage methods divide the visual grounding process into two steps: generation and ranking. Specifically, a module such as selective search [44], region proposal network [39], or pre-trained detector [16, 17, 39] is firstly used to generate proposals that contain objects. Then a multi-modal ranking network will measure the similarity between the query sentence and proposals and select the best matching result. Early works [22, 32, 41] only consider sentence-region level similarity. Yu et al. [58] decompose the query sentence and image into three modular components related to the subject, location, and relationship to model fine-grained similarity. Several studies [4, 33, 52, 53] incorporate graph learning to better model cross-modal alignment. The more related work to ours is Ref-NMS [7], which uses the query feature to guide the Non-Maximum Suppression on region proposals to increase the recall of critical objects. However, Ref-NMS [7] can only be integrated into two-stage methods. Moreover, it cannot affect the visual backbone in feature extraction and proposal generation.

One-stage methods extract visual feature maps that maintain the spatial structure and perform cross-modal interaction at pixel level. The fused feature maps are further used to predict bounding boxes. Yang et al. [55] use a Darknet [38] to extract feature maps and broadcast the query embedding to each spatial location. Several recent works [26, 54, 56] regard multimodal interaction as a multi-step reasoning process to better understand the input query. Huang et al. [23] extract landmark features with the guidance of a linguistic description. The more recent work TransVG [11] incorporates DETR encoder to extract visual features and proposes a transformer-based visual grounding framework. Kamath et al. [24] model the visual grounding as a modulated detection task and propose a novel framework MDETR derived from the DETR detector.

2.2. Transformer-based Visual Backbones

Dosovitskiy et al. [13] propose to extract image features by applying a pure Transformer architecture on image patches. This method is called Vision Transformer (ViT) and achieves excellent results compared to state-of-the-art convolutional networks. Touvron et al. [43] introduce improved training strategies to train a data-efficient ViT. Yuan et al. [60] propose a Tokens-to-Token transformation to obtain a global representation and a deep-narrow structure to improve the efficiency. Some recent works [9, 18] modify the ViT architecture for better performance. Liu et al. [31] present a novel hierarchical ViT called Swin Transformer. It computes the representation with shifted windowing scheme to allow for cross-window connection. Due to the fixed window size, the architecture has linear computational complexity with respect to image size.

2.3. Multimodal Interaction

In early studies, multimodal systems use simple interaction methods such as concatenation, element-wise product, summation, and multilayer perception to learn the interaction [5, 61]. Fukui et al. [14] introduce a compact bilinear pooling to learn a more complex interaction. Arevalo et al. [3] present a gated unit that can learn to decide how modalities influence the activation. [34, 35] use the cross-attention mechanism to enable dense, bi-directional interactions between the visual and textual modalities. De Vries et al. [10] introduce Conditional Batch Normalization at channel-level to modulate visual feature maps by a linguistic embedding. Vaezi Joze et al. [45] use squeeze and excitation operations to fuse the multimodal features and recalibrate the channel-wise visual features. There are also methods [3, 5, 10, 45, 61] that only perform interaction at channel-level. However, spatial-level information is also important to some downstream tasks, e.g., visual grounding and visual commonsense reasoning. Other methods [14, 34, 35] maintain the spatial structure to perform interaction but have a high computational cost.

3. Approach

3.1. Architecture

In this section, we first formulate the visual grounding task and then present the architecture of the adopted framework.

Visual grounding task aims at grounding a query onto a region of an image. The query refers to an object in the image and can be a sentence or a phrase. It can be formulated as: given an image I and a query q, the model needs to predict a bounding box $\mathbf{b} = \{x, y, w, h\}$ which exactly contains the target object expressed by the query.

As shown in Figure 2 (a), our framework is designed based on a typical end-to-end visual grounding architecture TransVG [11][1]. Given an image and a query sentence, there is a linguistic backbone, typically a pre-trained BERT model, extracting a sequence of 1D feature $\mathbf{T} \in \mathbb{R}^{D_l \times N_l}$. For the visual backbone, we adopt a new Query-modulated Refinement Network (described in Section 3.2) to extract a flattened sequence of visual feature $\mathbf{V} \in \mathbb{R}^{D_v \times N_v}$. The major difference between the proposed Query-modulated Refinement Network and existing visual backbones is that we take the contextual textual feature from the linguistic backbone to guide the feature extraction and multiscale fusion with the help of Query-aware Dynamic Attention (described in Section 3.2.1). Two projection layers map the visual and linguistic features into the same feature space $\mathbb{R}^D$. The projected visual and linguistic features are denoted by $\mathbf{p}_v \in \mathbb{R}^{D \times N_v}$ and $\mathbf{p}_l \in \mathbb{R}^{D \times N_l}$. Then $\mathbf{p}_v$ and $\mathbf{p}_l$ are concatenated by inserting a learnable embedding (i.e., a [REG] token) at the beginning of the concatenated sequence. The joint sequence is formulated as:

$$\boldsymbol{X} = [\mathbf{p}_r, \underbrace{\mathbf{p}_v^1, \mathbf{p}_v^2, \cdots, \mathbf{p}_v^{N_v}}_{\text{visual tokens } \boldsymbol{p}_v}, \underbrace{\mathbf{p}_l^c, \mathbf{p}_l^1, \cdots, \mathbf{p}_l^{N_l}}_{\text{linguistic tokens } \boldsymbol{p}_l}], \quad (1)$$

[1] https://github.com/djiajunustc/TransVG

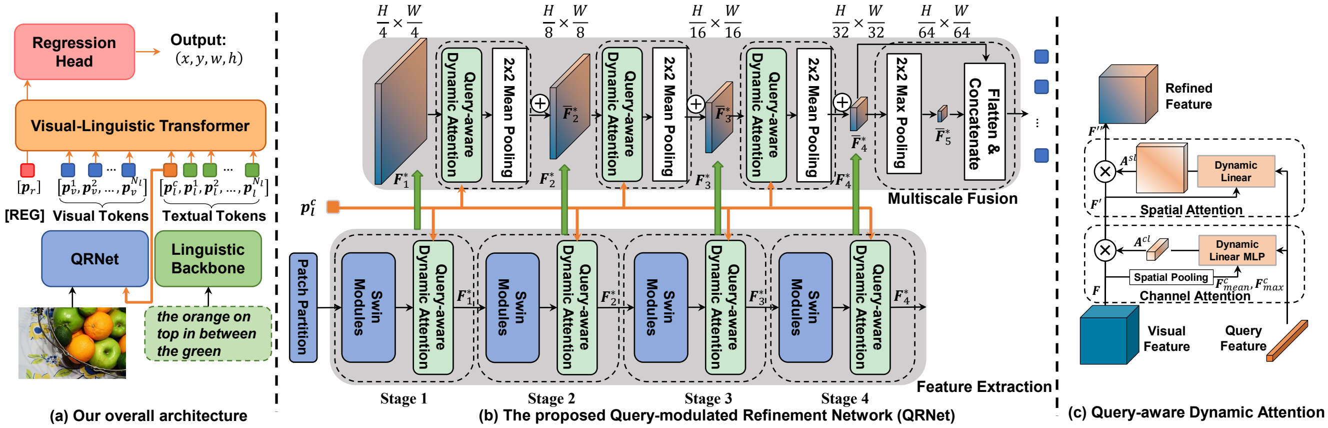

Figure 2. (a) The QRNet-based visual grounding framework used in this paper. (b) An overview of our Query-modulated Refinement Network. (c) Illustration of Query-aware Dynamic Attention.

where $\mathbf{p}_r$ is the learnable embedding of [REG] token. $\mathbf{p}_l^c$ is the [CLS] token's representation which is regarded as the contextual textual feature.

Next, a multi-layer visual-linguistic transformer is applied to perform intra- and inter-modal reasoning on the joint sequence. Finally, a prediction head takes the output representation of the [REG] token to predict the bounding box coordinates $\mathbf{b}$. The smooth L1 loss [16] and giou loss [40] are used to train the framework. The training objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_{smooth-l1}(\mathbf{b}, \hat{\mathbf{b}}) + \mathcal{L}_{giou}(\mathbf{b}, \hat{\mathbf{b}}), \tag{2}$$

where $\hat{\mathbf{b}}$ is the ground-truth box.

3.2. Query-modulated Refinement Network

In this section, we introduce the proposed visual backbone of Query-modulated Refinement Network (QRNet). An overview of the network is shown in Figure 2 (b). The network consists of two phases: (1) Query-refined Feature Extraction for extracting query-refined visual feature maps with a hierarchical structure, (2) Query-aware Multiscale Fusion for fusing the extracted feature maps at different scales with the guidance of the query feature. The two phases both rely on a novel Query-aware Dynamic Attention (QD-ATT), which dynamically computes textual-dependent visual attentions at spatial and channel levels, to enable to compute features with query guidance. In the following, we will first introduce the implementation of the Query-aware Dynamic Attention. Then we detail the Query-refined Feature Extraction and Multiscale Fusion.

3.2.1 Query-aware Dynamic Attention

Now, we will describe the details of our Query-aware Dynamic Attention (shown in Fig. 2(c)). We first introduce the dynamic linear layer, which learns a linear transformation that can be applied to the visual features to compute the query-aware attentions. Different from the conventional trainable linear layer, the parameters of the dynamic linear layer are generated dynamically based on textual features. Next, we will illustrate how to use the dynamic linear layer to compute query-aware channel and spatial attentions and obtain the query-consistent visual features. A pseudo-code is presented in the Appendix for better comprehension.

Dynamic Linear Layer. Existing visual backbones use modules with static parameters to compute the visual feature maps, where feeding in the same image will output the same feature maps. However, in visual grounding, different queries for a single image may reveal different semantic information and intentions, which require different visual features. We present a dynamic linear layer which can leverage the contextual textual feature $\mathbf{p}_l^c \in \mathbb{R}^{D_l}$ to guide the mapping from an given input vector $\mathbf{z}_{in} \in \mathbb{R}^{D^{in}}$ to the output $\mathbf{z}_{out} \in \mathbb{R}^{D^{out}}$. The dynamic linear layer is formalized as follow:

$$\mathbf{z}_{out} = \mathrm{DyLinear}_{\mathcal{M}_l}(\mathbf{z}_{in}) = \mathbf{W}_l^\intercal \mathbf{z}_{in} + \mathbf{b}_l \tag{3}$$

where $\mathcal{M}_l = \{\mathbf{W}_l, \mathbf{b}_l\}$, $\mathbf{W}_l \in \mathbb{R}^{D^{in} \times D^{out}}$, $\mathbf{b}_l \in \mathbb{R}^{D^{out}}$. D^{in} and D^{out} are the dimensions of input and output, respectively.

We use a plain linear layer to generate the $\mathcal{M}_l$. The generator is denoted by $\mathcal{M}_l' = \Psi(\mathbf{p}_l^c)$, where $\mathcal{M}_l' \in \mathbb{R}^{(D^{in}+1)*D^{out}}$. In detail, we predict a $(D^{in}+1)*D^{out}$ vector which can be reshaped to $\mathcal{M}_l$. However, it is easy to find that the number of parameters in the generator, i.e., $D_l * ((D^{in}+1) * D^{out})$, is too large. Such large-scale parameters will slow down the speed of the network and make it easier to be overfitting. Inspired by matrix factorization, we consider decomposing the $\mathcal{M}_l$ into two factors

$\mathbf{U} \in \mathbb{R}^{(D^{in}+1)\times K}$ and $\mathbf{S} \in \mathbb{R}^{K\times D^{out}}$, where $\mathbf{U}$ is a matrix generated from $\mathbf{p}_l^c$, $\mathbf{S}$ is a static learnable matrix, K is a hyper-parameter denoting the factor dimension. The factor generator $\{\mathbf{W}_l, \mathbf{b}_l\} = \Psi^*(\mathbf{p}_l^c)$ can be formulated as follow:

$$\begin{aligned} &\mathbf{U} = \mathrm{Reshape}(\mathbf{W}_g{}^{\intercal}\mathbf{p}_l^c + \mathbf{b}_g), \\ &\mathcal{M}_l = \mathbf{U}\mathbf{S}, \\ &\{\mathbf{W}_l, \mathbf{b}_l\} = \mathrm{Split}(\mathcal{M}_l), \end{aligned} \tag{4}$$

where $\mathbf{W}_g \in \mathbb{R}^{D_l\times(D^{in}+1)*K}$ and $\mathbf{b}_g \in \mathbb{R}^{(D^{in}+1)*K}$ are trainable parameters of the factor generator. The parameter matrix $\mathcal{M}_l$ of the dynamic linear layer can be reconstructed by multiplying $\mathbf{U}$ and $\mathbf{S}$. The $\mathbf{W}_l$ and $\mathbf{b}_l$ can be split from the matrix $\mathcal{M}_l$ along the first dimension. Finally, we reformulate the dynamic linear layer as follows:

$$\mathbf{z}_{out} = \mathrm{DyLinear}_{\mathcal{M}_l}(\mathbf{z}_{in}) = \mathrm{DyLinear}_{\Psi^*(\mathbf{p}_l^c)}(\mathbf{z}_{in}). \tag{5}$$

Unless otherwise specified, we use $\mathrm{DyLinear}(\mathbf{z}_{in})$ to represent a dynamic linear layer with $\mathbf{p}_l^c$ for simplify. And the different dynamic linear layers do not share parameters. When the input is a multidimensional tensor, the dynamic linear layer transforms the input at the last dimension.

Channel and Spatial Attention. As explained above, the pre-trained visual backbone is sensitive to all the objects it learned in the pre-training task. However, only the object which is referred to by the query text is useful. Besides, the features useful for the bounding box prediction highly depend on the semantics contained in the query sentence (e.g. entities, descriptions of attributes, and relationships). In other words, the importance of each channel or each region of the feature map should change dynamically according to the query sentence. Inspired by Convolutional Block Attention Module [51], we consider inferring channel and spatial attentions, i.e., $\mathbf{A}^{cl}$ and $\mathbf{A}^{sl}$, along the separate dimensions of the feature map to obtain adaptively refined features for a better cross-modal alignment.

Specifically, for a given visual feature map $\mathbf{F} \in \mathbb{R}^{H\times W\times D_v}$, we first aggregate its spatial information by average and maximum pooling and produce $\mathbf{F}^c_{max}$, $\mathbf{F}^c_{mean} \in \mathbb{R}^{1\times 1\times D_v}$. Then, a dynamic multilayer perception consisting of two dynamic linear layers with a ReLU activation in the middle is built to process the pooled features. The output dimension is the same as the input dimension. To reduce the number of parameters, we set the dimension of multilayer perception's hidden states as D_v/r, where $r = 16$ is a reduction ratio. By feeding the $\mathbf{F}^c_{max}$ and $\mathbf{F}^c_{mean}$ to the dynamic multilayer perception, a Sigmoid function is applied to the summation of the two output features. The channel attention map $\mathbf{A}^{cl}$ can be captured as follow:

$$\begin{aligned} \mathbf{F}^{cl}_{mean} &= \mathrm{DyLinear}_1(\mathrm{ReLU}(\mathrm{DyLinear}_2(\mathbf{F}^c_{mean})) \\ \mathbf{F}^{cl}_{max} &= \mathrm{DyLinear}_1(\mathrm{ReLU}(\mathrm{DyLinear}_2(\mathbf{F}^c_{max})) \\ \mathbf{A}^{cl} &= \mathrm{Sigmoid}(\mathbf{F}^{cl}_{mean} + \mathbf{F}^{cl}_{max}). \end{aligned} \tag{6}$$

We perform element-wise multiplication between $\mathbf{F}$ and $\mathbf{A}^{cl}$ to form the channel-wise refined visual feature, where the $\mathbf{A}^{cl}$ is broadcast along the spatial dimension:

$$\mathbf{F}^{'} = \mathbf{A}^{cl} \otimes \mathbf{F}. \tag{7}$$

To generate a spatial attention map, instead of squeezing the channel dimension, we leverage another dynamic linear layer to reduce the channel dimension to learn areas of interest to the query and apply an activation function Sigmoid to generate the attention map. In short, the computing process is formulated as follows:

$$\begin{aligned} \mathbf{A}^{sl} &= \mathrm{Sigmoid}(\mathrm{DyLinear}_3(\mathbf{F}^{'})) \\ \mathbf{F}^{''} &= \mathbf{A}^{sl} \otimes \mathbf{F}^{'} \end{aligned} \tag{8}$$

where $\mathbf{A}^{sl} \in \mathbb{R}^{H\times W\times 1}$ is the spatial attention map and $\mathbf{F}^{''}$ is the spatially refined visual feature and also the output of our Query-aware Dynamic Attention.

3.2.2 Query-refined Feature Extraction

To extract refined feature maps with the guidance of the query feature, we extend the Swin-Transformer[2] to a modulated visual feature extractor. As shown in Figure 2 (b), for a given image $I \in \mathbb{R}^{H\times W\times 3}$, a patch partition operation is firstly adopted to embed I to $\mathbf{F}_0 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$, where C is the embed dimension. Then $\mathbf{F}_0$ is fed into four cascaded stages, where each stage consists of multiple Swin-Transformer blocks and a QD-ATT module. In this work, we take the [CLS] representation of the input query sentence from BERT [12] as the contextual query representation $\mathbf{p}_l^c$ to compute the query-aware dynamic attention. The k-th stage receives the visual feature map $\mathbf{F}^*_{k-1}$ (or $\mathbf{F}_0$ if $k = 1$) obtained in the previous stage and generates a transformed feature map $\mathbf{F}_k$ through the Swin-Transformer blocks. Then, the QD-ATT module takes the transformed feature $\mathbf{F}_k$ and produces a query-aware feature $\mathbf{F}^*_k$ that will be further used in the next stage. The output of the query-refined feature extractor is a list of hierarchical features $[\mathbf{F}^*_1, \mathbf{F}^*_2, \mathbf{F}^*_3, \mathbf{F}^*_4]$.

3.2.3 Query-aware Multiscale Fusion

Multiscale features are beneficial to detecting objects at different scales [6]. However, because visual grounding requires fine-grained interaction after visual backbone (*e.g.*

[2] https://github.com/SwinTransformer/Swin-Transformer-Object-Detection

Cross Attention), high-resolution features will dramatically increase the computation. Therefore, previous works always use the low-resolution features or fuse the multiscale features in a query-agnostic manner, which will lose scale information or incorporate noise. Thanks to the hierarchical structure of the modulated Swin-Transformer, we can obtain multiscale features. We fuse the features obtained from different stages with the help of Query-aware Dynamic Attention mechanism and pooling operation. We flatten and concatenate the low-resolution features as the output token sequence of the backbone. Specifically, except for the first stage, each stage reduces the resolution of the feature map by a patch merging operator. In other words, the resolutions of the output feature maps at four stages are $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$, respectively. Besides, the patch merging operator fuses the patch features into the channel which doubles the channel dimension. Thus the channel dimensions of the output at the four stages are C, $2C$, $4C$ and $8C$, respectively. Next, we will introduce how to efficiently fuse these multiscale features with the guidance of the query text.

As shown in Figure 2 (b), we propose to fuse the output features at different stages with the help of QD-ATT. Specifically, we first use four 1×1 convolutional layers to unify the channel dimensions to D. To filter out noisy signals in the feature maps, we build QD-ATT modules for the feature maps $\{\mathbf{F}^*_k | \, k = 1, 2, 3\}$ of the first three stages. From $\mathbf{F}^*_k$, the QD-ATT module produces a weighted feature map with the same size as the input. We apply a 2×2 mean pooling with stride=2 to reduce the resolution to be the same as the next feature map $\mathbf{F}^*_{k+1}$ and compute the average of them to obtain $\bar{\mathbf{F}}^*_{k+1}$. Finally, the last feature map $\bar{\mathbf{F}}^*_4$ contains features of interest to the query from all scales. To detect very large objects, we also apply a 2×2 max pooling to obtain a $\frac{H}{64} \times \frac{W}{64}$ feature map $\bar{\mathbf{F}}^*_5$. We flatten and concatenate the $\bar{\mathbf{F}}^*_4$ and $\bar{\mathbf{F}}^*_5$ as the output token sequence $\mathbf{V}$.

4. Experiments

4.1. Datasets and Evaluation

We evaluate our method on phrase grounding dataset Flickr30K Entities[36] and referring expression grounding datasets RefCOCO[59], RefCOCO+[59], RefCOCOg[32] and ReferItGame[25]. The details and statistics are summarized in Appendix. In phrase grounding the queries are phrases and in referring expression grounding, they are referring expressions corresponding to the referred objects. We follow the same metric used in [11]. Specifically, a prediction is right if the IoU between the grounding-truth box and the predicted bounding box is larger than 0.5.

4.2. Implementation Details

The QRNet is built based on Swin-Transformer, i.e., Swin-S pre-trained with Mask-RCNN on MSCOCO [27][3]. We use $\mathrm{BERT}_{\mathrm{base}}$(uncased) for linguistic feature extraction. We set the intermediate dimension $D = 256$ and the factor dimension $K = 30$. We follow the TransVG [11] to process the input images and sentences. We also follow the training setting used in TransVG, which uses a AdamW optimizer with weight decay 10^{-4}, sets the batch size to 64 and the dropout ratio to 0.1 for FFN in Transformer. The learning rate is set to 10^{-5} for pre-trained parameters and 10^{-4} for other parameters. The parameters without pre-training are randomly initialized with Xavier. We train our model for 160 epochs. The learning rate is multiplied by a factor of 0.1 at epoch 40 for Flickr30K Entities and at epoch 60 for other datasets. We also follow the commonly used data augmentation strategies in [26, 54, 55].

4.3. Quantitative Results

We show the comparison results on ReferItGame and Flickr30k Entities in Table 1. The ReferItGame collects queries by a cooperative game, requiring one player to write a referring expression for a specified object. The other one needs to click on the correct object. Queries in Flickr30k Entities are phrases in the caption. We observe that the proposed QRNet outperforms previous works. We replace the visual branch of TransVG with Swin-Transformer and denote it by TransVG(Swin) to explore the impact of backbone. The performance is similar to the original TransVG, indicating that the accuracy gains are not from Swin-Transformer.

We also present the accuracy comparison with state-of-the-art methods on ReferCOCO, ReferCOCO+, and ReferCOCOg in Table 2. In ReferCOCO and ReferCOCO+ datasets, referred objects are people in "testA" and could be common objects in "testB". The expressions in ReferCOCOg are much longer than the ones in other datasets. Note that TransVG uses a ResNet-101 and a DETR encoder split from a pre-trained DETR framework. Its backbone is more powerful than a single ResNet-101.

We observe that our QRNet greatly outperforms all two-stage and one-stage state-of-the-art methods. On RefCOCO and RefCOCO+ datasets, our method gains 1.84%∼2.52% absolute improvement in "testA" and 2.51%∼4.32% in "testB". When the referred objects are arbitrary, the inconsistent issue will be more serious, and our modulated backbone can better filter irrelevant objects and correct the representation with textual guidance. In the RefCOCOg test split, we notice that ISRL [42] performs better than TransVG [11]. It models the visual grounding as a Markov

[3] We excluded images in the validation and test set of the RefCOCO series from MSCOCO training set and retrained Swin-Transformer for fair comparison.

Module	Backbone	ReferItGame test	Flickr30K test
Two-stage			
VC [62]	VGG16	31.13	-
MAttNet [58]	ResNet-101	29.04	-
Similarity net [48]	ResNet-101	34.54	50.89
LCMCG [30]	ResNet-101	-	76.74
DIGN [33]	VGG-16	65.15	78.73
One-stage			
FAOA [55]	DarkNet-53	60.67	68.71
RCCF [26]	DLA-34	63.79	-
ReSC-Large [54]	DarkNet-53	64.60	69.28
SAFF [56]	DarkNet-53	66.01	70.71
TransVG [11]	ResNet-101	70.73	79.10
TransVG (Swin)	Swin-S	70.86	78.18
QRNet (ours)	Swin-S	**74.61**	**81.95**

Table 1. The performance comparisons (Acc@0.5) on ReferItGame and Flickr30K Entities.

decision process that handles long expressions iterative by filtering out irrelevant areas. However, the limited action space makes it easy to converge to a local optimum. Our model does not modify the Visual-Linguistic Transformer in the TransVG and only provides query-consistent visual features. The performance is greatly improved, which demonstrates that our query-aware refinement is more effective to model the representation for visual grounding.

4.4. Ablation Study

We conduct ablation studies on ReferItGame and Flickr30k to reveal the effectiveness of the proposed QRNet.

We first study the effectiveness of QD-ATT in query-refined feature extraction and query-aware multiscale fusion. As shown in Table 3, we use a checkmark to denote enabling QD-ATT in the corresponding module, and a cross mark to denote disabling QD-ATT and passing the feature without modification. When disabling QD-ATT in multiscale fusion, the performance is dropped by 2.52% in ReferItGame-*test* and 0.79% in Flickr30k-*test*, respectively. When disabling QD-ATT in query-refined feature extraction, the performance is dropped by 3.22% and 1.51%. When completely disabling QD-ATT, the performance is dropped by 3.75% and 3.77%. We find that QD-ATT is more important in query-refined feature extraction than in multiscale fusion. We also notice that, when compared with completely disabling QD-ATT, only enabling QD-ATT in multiscale fusion gains little improvement because the features from the transformer are still noise and query-inconsistent.

We further study the effectiveness of spatial attention and

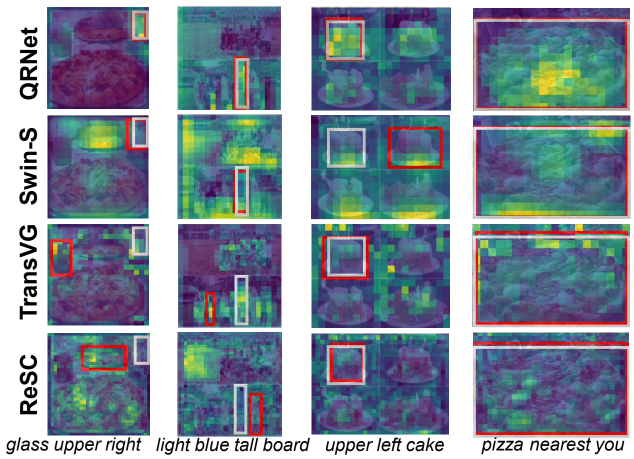

Figure 3. Visualization of activation maps from the backbone of our QRNet and other popular models. Red: the predicted box. White: the ground-truth box.

channel attention in the QD-ATT. Spatial attention can filter out irrelevant areas. And channel attention can redistribute the importance of the features to fit different query sentences. The results shown in the Table 4 demonstrate that spatial and channel attention are both effective.

4.5. Qualitative Results

We show the qualitative comparison between our QRNet and three popular methods in Figure 3. The feature maps are extract from the backbone of each model. We can see that previous methods are sensitive to many query-unrelated areas, which may lead to wrong predictions, e.g., the results of TransVG and ReSC for the queries of "glass upper righ" and "light blue tall board", and the result of Swin-S for the query of "upper left cake". In contrast, our QRNet generates query-consistent features and makes more accurate predictions. More results can be found in the supplementary.

5. Online Deployment

Previous experimental results have proven the advantages of our proposed QRNet, so we deploy it in a real-world online environment to test its practical performance. Specifically, we apply QRNet to enhance the search engine in Pailitao at Alibaba and perform A/B testing to evaluate the impact of our model. Details can be found in supplementary materials. We observe that QRNet decreases the no click rate by 1.47% and improves the number of transactions by 2.20% over the baseline. More specifically, the decrease of no click rate implies that QRNet can generate more accurate target boxes so that users are more likely to click. The improvement of the number of transactions means that the clicked item is exactly the users want to purchase, which also reveals the great performance of QRNet.

Models	Backbone	RefCOCO			RefCOCO+			RefCOCOg		
		val	testA	testB	val	testA	testB	val-g	val-u	test-u
Two-stage										
VC [62]	VGG16	-	73.33	67.44	-	58.40	53.18	62.30	-	-
MAttNet [58]	ResNet-101	76.65	81.14	69.99	65.33	71.62	56.02	-	66.58	67.27
Ref-NMS [7]	ResNet-101	78.82	82.71	73.94	66.95	71.29	58.40	-	68.89	68.67
LGRANs [49]	VGG16	-	76.60	66.40	-	64.00	53.40	61.78	-	-
RvG-Tree [21]	ResNet-101	75.06	78.61	69.85	63.51	67.45	56.66	-	66.95	66.51
CM-Att-Erase [29]	ResNet-101	78.35	83.14	71.32	68.09	73.65	58.03	68.67	-	-
NMTree [28]	ResNet-101	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44
One-stage										
FAOA [55]	DarkNet-53	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.36
RCCF [26]	DLA-34	-	81.06	71.85	-	70.35	56.32	-	-	65.73
ReSC-Large [54]	DarkNet-53	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20
SAFF [56]	DarkNet-53	79.26	81.09	76.55	64.43	68.46	58.43	-	68.94	68.91
HFRN [37]	ResNet-101	79.76	83.12	75.51	66.80	72.53	59.09	-	69.71	69.08
ISRL [42]	ResNet-101	-	74.27	68.10	-	71.05	58.25	-	-	70.05
LBYL-Net [23]	DarkNet-53	79.67	82.91	74.15	68.64	73.38	59.49	62.70	-	-
TransVG [11]	ResNet-101	81.02	82.72	78.35	64.82	70.70	56.94	67.02	68.67	67.73
TransVG (Swin)	Swin-S	82.33	84.01	79.83	64.94	70.19	56.47	67.81	69.34	68.99
QRNet (ours)	Swin-S	**84.01**	**85.85**	**82.34**	**72.94**	**76.17**	**63.81**	**71.89**	**73.03**	**72.52**

Table 2. The performance comparisons (Acc@0.5) on ReferCOCO, ReferCOCO+, and ReferCOCOg. The results of previous best two-stage and one-stage methods are highlighted with underlines. We highlight our results in bold. The results demonstrate that our method outperforms all state-of-the-art one-stage and two-stage methods.

Models	Feature Extraction	Multiscale Fusion	ReferItGame		Flickr30k	
			val	test	val	test
QRNet	✔	✔	**76.84**	**74.61**	**80.83**	**81.95**
(a)	✔	✗	74.31	72.09	80.09	81.16
(b)	✗	✔	73.63	71.39	79.35	80.44
(c)	✗	✗	73.25	70.86	77.17	78.18

Table 3. Ablation studies of QD-ATT in two phases of QRNet.

Models	Channel Attention	Spatial Attention	ReferItGame		Flickr30k	
			val	test	val	test
QRNet	✔	✔	**76.84**	**74.61**	**80.83**	**81.95**
(d)	✔	✗	74.35	72.02	80.22	81.35
(e)	✗	✔	74.41	71.80	80.67	81.55

Table 4. Ablation studies of different attentions in QD-ATT.

6. Conclusion

In this paper, we argue that pre-trained visual backbones cannot produce visual features that are consistent with the requirement of visual grounding. To overcome the weaknesses, we propose a Query-modulated Refinement Network (QRNet) to adjust the visual feature maps with the guidance of query text. The QRNet is designed based on a novel Query-aware Dynamic Attention mechanism, which can dynamically compute query-dependent spatial and channel attentions for refining visual features. Extensive experiments indicate that the improved framework significantly outperforms the state-of-the-art methods. The proposed QRNet exhibits vast potential to improve multimodal reasoning. In future work, we plan to improve the fine-grained interaction ability of QRNet and discard the post-interaction modules to simplify the existing end-to-end visual grounding frameworks.

Acknowledgement We thank Zenghui Sun, Quan Zheng and Weilin Huang at Alibaba Group for their help on online deployment and the anonymous reviewers for their valuable suggestions. This work was supported by the National Innovation 2030 Major S&T Project of China (No.2020AAA0104200 & 2020AAA0104205), the Science and Technology Commission of Shanghai Municipality (No. 19511120200 & 21511100100 & 18DZ2270800), the Open Fund of PDL, and Alibaba Group through Alibaba Research Intern Program.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 1

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018. 1

[3] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017. 3

[4] Mohit Bajaj, Lanjun Wang, and Leonid Sigal. G3raphground: Graph-based language grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4281–4290, 2019. 3

[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 3

[6] Zhaowei Cai, Quanfu Fan, Rogério Schmidt Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908, pages 354–370. Springer, 2016. 5

[7] Long Chen, Wenbo Ma, Jun Xiao, Hanwang Zhang, and Shih-Fu Chang. Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1036–1044, 2021. 3, 8

[8] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9962–9971, 2020. 1

[9] Xiangxiang Chu, Bo Zhang, Zhi Tian, Xiaolin Wei, and Huaxia Xia. Do we really need explicit position encodings for vision transformers? *arXiv e-prints*, pages arXiv–2102, 2021. 3

[10] Harm De Vries, Florian Strub, Jeremie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3

[11] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. *arXiv preprint arXiv:2104.08541*, 2021. 2, 3, 6, 7, 8

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

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[14] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, 2016. 3

[15] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1811–1820, 2017. 1

[16] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2, 4

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[18] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021. 3

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[21] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 8

[22] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016. 3

[23] Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. Look before you leap: Learning landmark features for one-stage visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16888–16897, 2021. 2, 3, 8

[24] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 2, 3

[25] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and

Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 1, 6

[26] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020. 3, 6, 7, 8

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[28] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4673–4682, 2019. 8

[29] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1950–1959, 2019. 8

[30] Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. Learning cross-modal context graph for visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11645–11652, 2020. 7

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[32] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1, 3, 6

[33] Zongshen Mu, Siliang Tang, Jie Tan, Qiang Yu, and Yueting Zhuang. Disentangled motif-aware graph learning for phrase grounding. In *Proc 35 AAAI Conf on Artificial Intelligence*, 2021. 2, 3, 7

[34] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *arXiv preprint arXiv:2107.00135*, 2021. 3

[35] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096, 2018. 3

[36] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1, 6

[37] Heqian Qiu, Hongliang Li, Qingbo Wu, Fanman Meng, Hengcan Shi, Taijin Zhao, and King Ngi Ngan. Language-aware fine-grained object representation for referring expression comprehension. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4171–4180, 2020. 8

[38] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1, 2, 3

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 1, 2

[40] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019. 4

[41] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016. 3

[42] Mingjie Sun, Jimin Xiao, and Eng Gee Lim. Iterative shrinking for referring expression grounding using deep reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14060–14069, 2021. 6, 8

[43] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021. 3

[44] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 2

[45] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[47] Adam Vogel and Dan Jurafsky. Learning to follow navigational directions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 806–814, 2010. 1

[48] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018. 7

[49] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1960–1968, 2019. 8

[50] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den

Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10126–10135, 2020. 1

[51] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 5

[52] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4644–4653, 2019. 2, 3

[53] Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9952–9961, 2020. 2, 3

[54] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive subquery construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 387–404. Springer, 2020. 2, 3, 6, 7, 8

[55] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019. 3, 6, 7, 8

[56] Jiabo Ye, Xin Lin, Liang He, Dingbang Li, and Qin Chen. One-stage visual grounding via semantic-aware feature filter. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1702–1711, 2021. 2, 3, 7, 8

[57] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 1

[58] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 2, 3, 7, 8

[59] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 1, 6

[60] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 3

[61] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. 3

[62] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018. 7, 8

[63] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016. 1