Multi-Modal Dynamic Graph Transformer for Visual Grounding

Sijia Chen, Baochun Li
Department of Electrical and Computer Engineering
University of Toronto
sjia.chen@mail.utoronto.ca, bli@ece.toronto.edu

Abstract

Visual grounding (VG) aims to align the correct regions of an image with a natural language query about that image. We found that existing VG methods are trapped by the single-stage grounding process that performs a sole evaluate-and-rank for meticulously prepared regions. Their performance depends on the density and quality of the candidate regions, and is capped by the inability to optimize the located regions continuously. To address these issues, we propose to remodel VG into a progressively optimized visual semantic alignment process. Our proposed multi-modal dynamic graph transformer (M-DGT) achieves this by building upon the dynamic graph structure with regions as nodes and their semantic relations as edges. Starting from a few randomly initialized regions, M-DGT is able to make sustainable adjustments (i.e., 2D spatial transformation and deletion) to the nodes and edges of the graph based on multi-modal information and the graph feature, thereby efficiently shrinking the graph to approach the ground truth regions. Experiments show that with an average of 48 boxes as initialization, the performance of M-DGT on the Flickr30k Entities and RefCOCO datasets outperforms existing state-of-the-art methods by a substantial margin, in terms of both accuracy and Intersect over Union (IOU) scores. Furthermore, introducing M-DGT to optimize the predicted regions of existing methods can further significantly improve their performance. The source codes are available at https://github.com/iQua/M-DGT.

1. Introduction

Visual grounding (VG) is a crucial task in the interdisciplinary subject of computer vision and natural language processing. With a focus on aligning semantically consistent phrase-region pairs from the given image and sentence, the general visual grounding problem can be extended to phrase localization [5, 9, 36] and referring expression comprehension [17, 28]. Then, tasks including translation [24], cross-modal retrieval [14], image caption [5, 39], and visual query answering [1, 38] can benefit from aligned phrase-region pairs.

Although there have been significant breakthroughs in recent years, we noticed that the main bodies of proposed state-of-the-art VG methods [3, 6, 7, 16, 18, 20, 27, 28, 42, 46] follow the one-step evaluate-and-rank matching architecture. Methods with this architecture evaluate regions in the image to select the correct ones, a single-execute process without continuous optimization. Existing works [16, 19, 20, 27, 42] relying on the region proposal build models based on candidate regions and make once predictions from these regions. In some works [6, 18, 43] with neural attention mechanism, the attention scores are assigned to rough regions that are refined once to generate the grounding boxes. The inherent issues in this architecture are that solving a complex nonlinear matching problem with a one-step manner leads to poor region-phrase matching results caused by local optimum, and as a result the matched regions for phrases cannot be further adjusted. For example, if there is a significant deviation between the predicted regions and ground truth regions, existing works cannot adjust the regions to approach the target. This is proven by our experiments, in that the accuracy of these works decreases substantially with the increase of the Intersect over Union (IOU) threshold.

Some more recent works [9, 33, 41] tried to alleviate problems in such matching architecture by introducing the idea of progressive learning. Dogan *et al.* [9] achieved this by performing phrase grounding sequentially while Sun *et al.* [33] proposed the idea of iterative shrinking the detection area through reinforcement learning to locate the target. However, they are still trapped in the one-step matching architecture because the failure of any matching step in the learning process will ultimately produce poor results.

In this work, we break the limits in such design by proposing a search-based visual grounding mechanism. More specifically, we remodel VG into a progressively optimized visual semantic alignment process. By doing this, the simple initialization regions can be continuously optimized to approach the target regions. To achieve this goal,

the main challenge is to pass the information in each region from local to global with the minimum cost required. Motivated by works [16, 20, 40], the highly structured information in the spatial regions and multi-modal input can be modeled by graph theory.

With these insights, we propose a **m**ulti-modal **d**ynamic **g**raph **t**ransformer (M-DGT), a framework that regards regions as nodes and semantic relations as edges, making the process of progressively approaching the ground truth regions equivalent to the transformation of a graph. M-DGT first constructs a graph from a few simple initialization regions and then continuously transforms the nodes and edges according to the multi-modal information and the graph feature, thereby shrinking the graph to the target layout. The image regions corresponding to the nodes of the target graph are most aligned with the semantics in the query. M-DGT with a graph-based progressive search method can continuously correct the deviation of the previous transformation in the subsequent learning to alleviate the impact of a failed transformation on the final result, which improves robustness. In addition, M-DGT naturally exploits the spatial relations between regions and the cross-modal semantic relations by modeling them in a multi-modal graph structure.

The original contributions of this paper are as follows. *First*, we rethink the VG as a progressively optimized visual semantic alignment process, making the VG can be divided into sub-problems that can be easily solved progressively. *Second*, we propose a novel multi-modal dynamic graph transformer (M-DGT) to model the process of searching for a matching region-text as a graph structure transformation. *Third*, M-DGT is fast, accurate, and with high generality. Starting with just a few simple initialization regions, our framework can gradually obtain tighter matching regions for phrases without missing targets, enabling it to work on arbitrary datasets. *Finally*, in two tasks, including phrase localization on the Flickr30k Entities and referring expression comprehension on three RefCOCO datasets, M-DGT not only achieves state-of-the-art accuracies but also produces bounding boxes with high Intersect over Union (IOU) scores.

2. Related Work

The broader definition of Visual Grounding (VG) includes the phrase localization [3, 9, 16, 19, 20, 27, 28, 29, 36, 42, 46] and the referring expression comprehension [6, 16, 21, 24, 37, 43, 45]. The phrase localization aims to extract semantic aligned phrase-region pairs from the given image and sentence. Referring expression comprehension requires a model to respond to a query by specifying a corresponding region in an image.

The recent works of VG follow the remarkably similar one-step evaluate-and-rank matching architecture. Especially shown by works [16, 27, 29, 36, 46], the model was trained to evaluate the prepared regions and select the correct one for the query. The candidate regions are generated by using object proposal methods or pre-trained object detection methods, such as RCNN [31] and bottom-up attention [1]. Then, most works utilized the rank function [22, 25, 36], such as maximum-margin ranking loss, to find the matched region-query from these candidate ones. [42] proposed a one-stage approach by fusing a text query's embedding into the YOLOv3 [30]. However, such methods still made a single prediction based on high-density initialization boxes. Other attention methods, including [24, 26, 43], extract attention from various kinds of multi-modal data to highlight the target regions and then make one refinement to get the bounding box.

More recent works tend to model the VG based on the idea of progressive learning. The works [4, 6, 9, 11, 18, 33, 41] designed the model to gradually adjust the predicted regions or the attention scores to adjust the bounding box to locate the matched region-text pair. [9] reformulated phrase grounding as a sequence labeling task while [4] was proposed to utilize semantic relationships to iteratively reason and build up estimates. The ones that most related to our top idea are works [11] and [33]. They utilized the progressive bounding boxes refinement architecture to adjust the boxes closer to the ground truth. Specifically, in the work [33], the authors formalize VG as a sequence of image-level shrinking processes, thus adjusting image patches in each iteration to gradually approach the target. Then, our work is also motivated by the work [7]. They presented that a simple stack of transformer encoder layers behaves as an effective way to learn multi-modal correspondence.

In recent years, the widely used graph structure [38, 48] is also introduced to the visual grounding domain in many state-of-the-art methods [16, 20, 37]. As the graph is a natural way to organize the structured relations in the multi-modal information, these works built the visual static graph and language static graph. Then, the advanced tools, including the graph attention mechanism [37], the graph matching technique [16, 20], are proposed to learn the correspondences matching of the entities in the graph representations.

3. Problem Setting and Overview

The target of visual grounding with the text query $\boldsymbol{t}$ and raw image $\mathbf{M} \in R^{H \times W}$ as inputs is to locate correct image regions $\mathbb{L} \subset \mathbf{M}$ that has the same semantic with P phrases in the query. We aim to learn a visual grounding model that can gradually adjust an initial set of bounding boxes to produce the tightest visual regions for the text query. These bounding boxes $B = (\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_Z) \in R^{Z \times 4}$ are roughly generated to cover the whole image without overlap. Thus, large deviations between ground-truth regions and these ini-

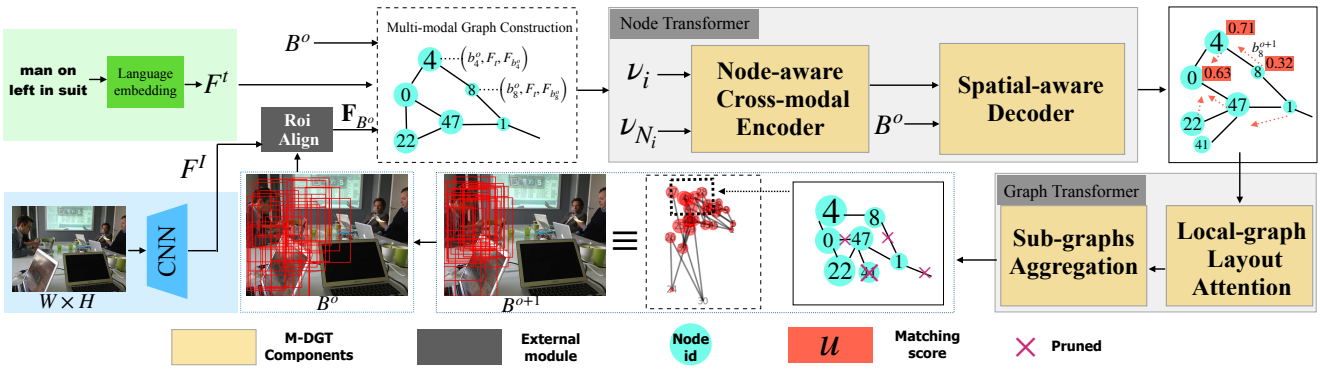

Figure 1. The pipeline overview of our proposed M-DGT for visual grounding. M-DGT can gradually transform the initialization boxes to approach the target regions through the multi-modal dynamic graph structure. This is achieved by the *node transformer*, and the *graph transformer*, which are are smoothly combined for end-to-end training.

tial candidates B make it impossible to predict the matched region-phrase pairs using a sole evaluate-and-rank process. Therefore, we reformulate visual grounding as an iterative search process with index o. Our iterative model is designed to adjust coordinates of B to obtain the correct regions progressively. We refer this to progressively optimized visual-semantic alignment learning. There are three problems in such an iterative learning process. Firstly, the 2D transformation learning process should be modeled to rely on the multi-modal contextual semantic and relative spatial information. Secondly, B adjusted by predicted 2D transformation coefficients should approach corresponding ground-truth regions in each iteration. Thirdly, high efficiency is required.

In this work, we address these problems in the context of multi-modal graph structure $\mathcal{G}$ by explicitly modeling bounding boxes as nodes ν and multi-modal semantic relations as edges E. We instantiate the effective multi-modal region-semantic reasoning for 2D transformation learning in the graph $\mathcal{G}$ by developing the `node transformer`. Our further module, the `graph transformer` supported by the attention mechanism, is to alleviate the grounding ambiguity in the iterative process by pruning nodes and edges of the graph. In addition, the efficiency of our framework also benefits from such an ever-decreasing graph scale. The graph structure, including the layout and information of nodes/edges, changes over iteration o, making our framework's foundation be the multi-modal dynamic graphs.

4. Model Architecture

We now elaborate on components design in our multi-modal dynamic graph transformer (M-DGT). As shown by an overview of the entire model pipeline in Fig. 1, M-DGT models the progressively optimized visual-semantic alignment process by the dynamic graph structure. Specifically, M-DGT adjusts nodes by the *node transformer* and refines the graph layout by the *graph transformer* to obtain the tight bounding boxes for the text query progressively and efficiently. To effectively train the M-DGT, we also propose an iteration-aware training method.

4.1. Multi-modal Graph Construction

Given one image and the text query, the first stage of M-DGT is the multi-modal graph construction module that generates the graph based on bounding boxes and semantic relations. Basically, each node ν_i in the graph corresponds to one bounding box in the image. Thus, the position of the node in the graph is determined by the center $c_i = (y_{c_i}, x_{c_i})$ of its bounding box $\boldsymbol{b}_i^o$. The visual feature $\boldsymbol{F}_{b_i^o} \in R^{d_b}$ of the node is obtained by applying the roiAlign [12] for the single-scale box $\boldsymbol{b}_i^o$ on the backbone visual feature map $\boldsymbol{F}^I \in R^{H' \times W' \times C}$. Besides, we insert the text to each node to facilitate the region-semantic reasoning. The text feature is denoted as $\boldsymbol{F}^t \in R^{P \times d_t}$, where each p phrase feature is obtained by averaging the final-layer token-level vectors yielded by BERT [8]. Besides, with linear projection, the visual and text features are mapped to the common space with dimension d. Finally, each node with id i in the graph $\mathcal{G}$ contains spatial and multi-modal information, i.e., $\left(\boldsymbol{b}_i^o, \boldsymbol{F}^t, \boldsymbol{F}_{b_i^o}\right)$.

When $o = 0$, the initial set of boxes with a single scale 128×128, the stride 128, and constant aspect ratios $(1, 1)$ are generated to cover the image. As for edges, each box is connected to its most nearby neighbors. We present the details in section 3 of the appendix. When $o > 0$, as shown in Fig. 1, these existing nodes and edges are only transformed and pruned to contribute to the visual grounding.

4.2. Node Transformer

Operated upon the built multi-modal graph with node feature $\left(\boldsymbol{b}_i^o, \boldsymbol{F}^t, \boldsymbol{F}_{b_i^o}\right)$, the node transformer aims to trans-

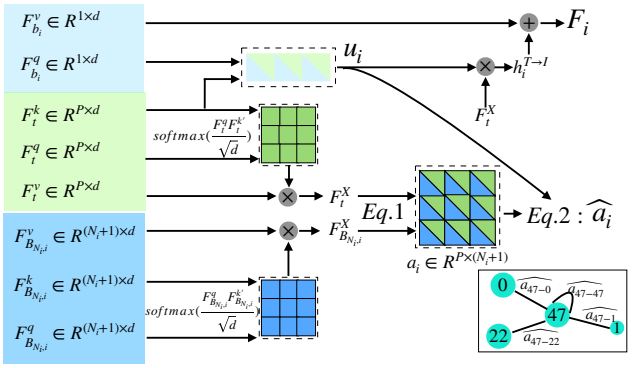

Figure 2. Detailed architecture of the node-aware cross-modal encoder. The positional encoding layer, norm sublayer and FFN sublayer are not shown here for brevity. The graph layout of node 47 and its neighbors N_{47} is presented as an example in the right bottom.

form the spatial positions and coordinates of bounding boxes to approach ground-truth regions. To achieve this, we introduce the *node-aware cross-modal encoder* to fuse the multi-modal information and generate the node attentions based on inter-semantic relations. Further, the *spatial-aware decoder* builds the node and edge feature to achieve node transformation regression. For brevity, we omit the iteration index o in the following discussion without losing generality.

Node-aware cross-modal encoder. As shown by the colored area in Fig. 2, for node ν_i with the visual feature $\boldsymbol{F}_{b_i}$, we compute the corresponding value vector and query vector as $\boldsymbol{F}_{b_i}^v = \boldsymbol{F}_{b_i}\boldsymbol{W}_v^I$ and $\boldsymbol{F}_{b_i}^q = \boldsymbol{F}_{b_i}\boldsymbol{W}_q^I$. For the text query feature $\boldsymbol{F}_t \in R^{P\times d}$, the value, key, and query vectors are $\boldsymbol{F}_t^v \in R^{P\times d}$, $\boldsymbol{F}_t^q \in R^{P\times d}$, and $\boldsymbol{F}_t^k \in R^{P\times d}$, respectively. Then, as shown in Fig. 2, we compute the text-to-visual feature as $\boldsymbol{h}_i^{T\to I} = softmax(\boldsymbol{u}_i)\boldsymbol{F}_t^X$ where $\boldsymbol{u}_i = \frac{\boldsymbol{F}_{b_i}^q \boldsymbol{F}_t^{k\prime}}{\sqrt{d}}$ is the matching score of the node i and P phrases and $\boldsymbol{F}_t^{k\prime}$ is the transpose of $\boldsymbol{F}_t^k$. One example of the matching score is shown by red boxes in Fig. 1. Then, we can directly utilize the residual connection to obtain the multi-modal feature as $\boldsymbol{F}_i = norm(\boldsymbol{h}_i^{T\to I} + \boldsymbol{F}_{b_i}^v)$.

Then, inspired by the graph attention network [35], we compute the attention scores between the node i and its connected neighbor nodes N_i as their responses to the text query. Firstly, we compute the response scores of node i and its neighbors N_i to P phrases as follows:

$$\boldsymbol{a}_i = softmax(\boldsymbol{c}_i), \boldsymbol{c}_i = \frac{\boldsymbol{F}_t^X \boldsymbol{F}_{B_{N_i,i}}^{X\prime}}{\sqrt{d}} \tag{1}$$

where $\boldsymbol{F}_t^X$ and $\boldsymbol{F}_{B_{N_i,i}}^X$ are computed as shown in Fig. 2. The shape of $\boldsymbol{c}_i$ is $P\times(N_i+1)$. Each row p of $\boldsymbol{a}_i \in R^{P\times(N_i+1)}$ presents response scores of nodes i and its neighbors N_i to the phrase p.

Then, we obtain the multi-modal semantic matching score $\widehat{a}_{ij}$ for the node i and its neighbor node j as follows:

$$\widehat{a}_{ij} = \frac{exp\left(\sum_{s=1}^{P} u_{is} exp^{[\boldsymbol{c}_i]_{sj}}\right)}{\sum_{k\in(N_i,i)} exp\left(\sum_{s=1}^{P} u_{is} exp^{[\boldsymbol{c}_i]_{sk}}\right)} \tag{2}$$

where we have $j \in (N_i, i)$. $\widehat{\boldsymbol{a}}_i \in R^{(N_i+1)}$.

The quantity $\widehat{a}_{ij}$, works as the attention score between the node i and node j. The solid line box in the lower right corner of Fig. 2 presents a direct example. As nodes i and j corresponds to bounding box b_i and b_j, $\widehat{a}_{ij}$ essentially measure the overall multi-modal semantic similarity of these two bounding boxes on P phrases.

Spatial-aware decoder. Once obtaining all attention factors $\widehat{\boldsymbol{a}}_i$ and cross-modal fusion feature $\boldsymbol{F}_i$ for each node i, the spatial-aware decoder conducts the 2D transformation regression based on the hidden representation $\boldsymbol{h}_i'$ by performing the propagation model in the general GNN method [35].

In our M-DGT, the spatial information of the bounding box in each node is critical to make semantic reasoning and learn the bounding box transformation to approach the ground-truth region. As each node is connected with other nodes in the graph structure, the relative spatial information between nodes can facilitate the learning process. One direct intuition is that the node containing the ground-truth bounding box can pull other nodes to its correct region. Therefore, we propose spatial-aware representation $\boldsymbol{h}_i'$ by introducing the spatial information of each node and modeling the relative spatial information as the edge feature.

For spatial feature of the node, we utilize the spatial coordinates used in the paper [42]. Specifically, for the grid with size H' and W' of the visual feature map, we first compute the correspond position $(m_i, n_i), m_i \in [0, H'], n_i \in [0, W']$ of the node's center c_i in the grid and then obtain the spatial information $\boldsymbol{g}_i = \left(\frac{n_i}{W'}, \frac{m_i}{H'}, \frac{n_i+0.5}{W'}, \frac{m_i+0.5}{H'}, \frac{n_i+1}{W'}, \frac{m_i+1}{H'}, \frac{1}{W'}, \frac{1}{H'}\right)$. We then map this vector to the feature space by using the linear projection $\boldsymbol{h}_{g_i} = \boldsymbol{g}_i\boldsymbol{W}_g$, where $\boldsymbol{W}_g$ is the trainable parameters.

For the edge feature between node i and its neighbor $j \in N_i$, we compute their spatial relationship as $\boldsymbol{g}_{ij} = \left(\frac{|x_{i\,\min} - x_{j\,\min}|}{\triangle w_i}, \frac{|y_{i\,\min} - y_{j\,\min}|}{\triangle h_i}, \frac{\triangle w_i}{\triangle w_j}, \frac{\triangle h_i}{\triangle h_j}\right)$ with $\triangle w_i = x_{imax} - x_{imin}$ and $\triangle h_i = y_{imax} - y_{imin}$. Then, g_{ij} is mapped to a high-dimensional representation shown as $\boldsymbol{e}_{ij} = \boldsymbol{g}_{ij}\boldsymbol{W}_e$. Thus, $\boldsymbol{e}_{ij}$ is the edge feature between the node i and j.

Finally, the node feature is obtained by fusing the spatial information and multi-modal feature as $\boldsymbol{h}_i = \boldsymbol{F}_i \parallel \boldsymbol{h}_{g_i}$.

Then, our proposed spatial-aware representation $\boldsymbol{h}'_i$ is computed as:

$$\boldsymbol{h}'_i = \sigma\left(\sum_{j\in(N_i,i)} \widehat{a}_{ij}\boldsymbol{W}\left[\boldsymbol{W}_f\boldsymbol{h}_i \parallel \boldsymbol{W}_e\boldsymbol{e}_{ij}\right]\right) \tag{3}$$

where $\parallel$ is the concatenate operation and $\sigma(\cdot)$ is the a nonlinear function.

Motivated by the work [15], the node 2D transformation in our M-DGT is defined as the affine transformations with both translation and scaling in 2D space represented by matrix multiplication in homogeneous coordinates. Also, the affine transformation can be formulated in the parameterized form. The details are discussed in the appendix. We have 2D transformation coefficients $[s_1, s_2, r_1, r_2]$ where s_1, s_2 are scaling coefficients and r_1, r_2 are translation coefficients. For the coordinate x, y, the new coordinate is $x' = s_1x + r_1, y' = s_2y + r_2$.

Therefore, with spatial-aware representation as the input, the learning target of our decoder in the node i is modeled as a transformation regression function.

$$r_{1_i}, r_{2_i}, s_{1_i}, s_{2_i} = tanh\left(\boldsymbol{W}_D\boldsymbol{h}'_i\right) \tag{4}$$

where $\boldsymbol{W}_D$ is the trainable parameters of the fully-connected network and the final transformation coefficients are converted to $s'_{1_i}, s'_{2_i} \in [0.5, 1.5]$ and $r'_{1_i}, r'_{2_i} \in [-0.2, 0.2]$, thereby ensuring the stability of the 2D transformation.

4.3. Graph Transformer

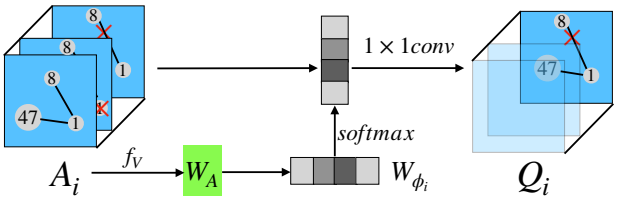

Figure 3. The operation of the graph transformer on the local layout of the node i. The graph transformer layer softly selects the target layout from a set of candidate local graph layouts A_i by performing the layout attention mechanism on the adjacency matrices and nodes.

Different from the transformers [34] used in computer vision and natural language processing, the action of our proposed graph transformer is to prune the nodes and edges, thus transforming the graph structure.

Adjusting the graph structure by removing nodes and edges to facilitate the learning process is an NP-hard problem. Therefore, instead of processing the whole graph directly, we proposed *local-graph layout attention* mechanism for the graph transformer to prune the local layout with only a few nodes and edges and then integrate multiple adjusted local layouts by using the *sub-graphs aggregation* method to obtain the final transformed graph. As shown in Fig. 3, such local layout used in our method is constructed by each node i and its connected neighbors. Thus, for the graph with $|\nu|$ nodes, there can be $|\nu|$ local layouts.

Motivated by the work [47] that builds a new meta-path graph by selecting edge types from the set of candidates. In each node i with the corresponding local layout, we construct the set A_i containing all possible neighbor nodes and edges combinations. Then, the r-th element of A_i presents one possible local layout A_{ir} that is referred to as one possible transformation of the original local layout.

As shown in Fig. 3, we first obtain the feature of each A_{ir} using the function $f_v(A_{ir}) = concat\left(\overline{\boldsymbol{h}}_{ir}, \overline{\boldsymbol{e}}_{ir}\right)$ that is computed as follows:

$$\overline{\boldsymbol{h}}_{ir} = \frac{1}{N_{ir}+1}\sum_{k\in\{N_{ir},i\}}\boldsymbol{h}_k, \overline{\boldsymbol{e}}_{ir} = \frac{1}{|E_{N_{ir},i}|}\sum_{ij\in E_{N_{ir},i}}\boldsymbol{e}_{ij} \tag{5}$$

where N_{ir} is the nodes in the r-th local layout of the node i while $E_{N_{ir}}$ is the corresponding edges. And we have $\overline{\boldsymbol{h}}_{ir} \in R^{2d}$ and $\boldsymbol{e}_{ij} \in R^{d_s}$.

Then, the score tensor $\boldsymbol{W}_{\phi_i} \in R^{1\times1\times|A_i|}$ is computed by $f_v(A_i)\boldsymbol{W}_A$ where $\boldsymbol{W}_A$ is a trainable weight.

Our graph transformer achieves the softly selection from local layouts A_i by 1×1 convolution with non-negative weights from $softmax(\boldsymbol{W}_{\phi_i})$. The formaulation for final local layout Q_i is:

$$Q_i = \phi\left(A_i, softmax\left(\boldsymbol{W}_{\phi_i}\right)\right) \tag{6}$$

where ϕ is the convolution layer. This can be regarded as the channel attention pooling.

Our sub-graphs aggregation method mainly utilizes the idea of the hard voting ensemble to determine the target transformed graph $\mathcal{G}^{o+1}$. Firstly, we select the Q_{ir} with the highest attention score to be regarded as the vote of the node i on the target transformed local layout. Then, the preliminary transformed graph is the integration of all Q_{ir} for $\nu_i \in \nu$ in the graph $\mathcal{G}^o$. Finally, the node and edge will be pruned from this preliminary transformed graph once over half of its neighbors remove the node/edge in its vote. Therefore, this graph pruning method directly induces the deletion of the abundant and useless bounding boxes and semantic relations to promote efficient and effective learning.

4.4. Training Objective

Our iteration-based M-DGT aims to learn 2D transformation coefficients for each node in each iteration, thus transforming the bounding boxes to approach the ground-truth regions progressively. To this end, we propose an iteration-related training (IRT) method that continuously

optimizes the parameters based on the transformation loss function in each iteration.

For each query phrase p, we compute the Huber loss of predicted transformation coefficients and ground truth coefficients $(\widetilde{r}_1^p, \widetilde{r}_2^p, \widetilde{s}_1^p, \widetilde{s}_2^p)$. Then, for the node i, its loss function l_{ip}^h for p-th query phrase is defined as $l_{ip}^h = Huber(\left(r_{1_i}^p, r_{2_i}^p, s_{1_i}^p, s_{2_i}^p\right), \left(\widetilde{r}_{1_i}^p, \widetilde{r}_{2_i}^p, \widetilde{s}_{1_i}^p, \widetilde{s}_{2_i}^p\right))$. Then, the GIOU score [32] between the 2D transformed bounding box b_i of node i, and the ground-truth region $\mathbb{L}_p$ of phrase p is denoted as $giou_{ip}$. The matching score u_{ip} of the node i and p phrase is desired to close to the $giou_{ip}$. Based on this, the transformation loss function of node i is $l_i = \sum_P^{p=1} u_{ip} l_{ip}^h + l_{smooth-l1}\left(\boldsymbol{u}_i, \boldsymbol{giou}_i\right)$, where $l_{smooth-l1}$ is the general smooth L1 loss.

The main structure of our IRT is motivated by the replay buffer mechanism in reinforcement learning [13]. The core idea is to train the whole M-DGT in each iteration step rather than calculating the gradient once at the end. However, we do not decouple the update in each iteration from the whole sequential learning process but assign an iteration-related attenuation factor $\alpha = \frac{1}{1+e^{-o}}$ to the transformation loss function l_i^o in o-th iteration. Finally, the overall loss for our model in each iteration o is defined as αl_i^o

5. Evaluation

We evaluate our M-DGT method on four publicly available datasets, including Flickr30k Entities [29], and RefCOCO [44], RefCOCO+ [44], and RefCOCOg [23]. The M-DGT with 48 roughly initialized boxes and 7 number of iterations is compared with state-of-the-art methods in terms of the accuracy under the IOU threshold of 0.5 (i.e., Acc@0.5). Then, in ablation experiments, the performance of M-DGT is discussed in terms of the different number of iterations (i.e., $3, 5, 7, 9$), different combinations of components in M-DGT, and different box initialization methods. Finally, the M-DGT is utilized as the plugin to optimize predicted bounding boxes of existing outstanding works. Limited by space, we describe implementation details in subsection 6.3 of the appendix.

5.1. Global Performance

Gloabl accuracy. As shown in Table 1, Table 3, and Fig. 4, the performance of our M-DGT is significantly better than other methods and achieves state-of-the-art accuracy in listed benchmark datasets. Specifically, in the phrase grounding task of the Flickr30K Entities dataset, M-DGT achieves 79.97% top-1 accuracy. Meanwhile, in the RefCOCO dataset with highly semantic reasoning requirements, M-DGT obtains the highest accuracies 85.374%, 70.018%, 79.213% on three datasets, respectively. Besides, M-DGT can produce tighter predicted regions for the text query, which is reflected in maintaining high accuracy under high IOU thresholds, as shown by our detailed performance comparison under the IOU threshold range from 0.35 to 0.9 in subsection 7.2 of the appendix.

In the fair comparison with the two-stage method that relies on region proposals, the accuracy of M-DGT outperforms the best method LCMCG [20] by 3.23% on Flickr30K Entities dataset and is 2.69%, 2.32%, and 3.48% higher than the VS-graph [16] on three RefCOCO datasets, respectively. Besides, our M-DGT is 4.72% higher than the best one-stage method VGTR [10]. This shows that M-

Table 1. Comparisons with state-of-the-art methods on the test set of Flickr30k Entities [29] in terms of top-1 phrase grounding accuracy (%) with IOU threshold 0.5.

Method	Visual Backbone	Region Proposals	Language Embedding	Acc@0.5	Time (ms)
CCA [29]	VGG19	Edgebox N=200	Word2vec, FV	50.89	-
Two-branch [36]	VGG19	Edgebox N=200	Word2vec, FV	51.05	305
SPC+PPC [28]	ResNet101	Edgebox N=200	Word2vec, FV	55.85	-
QRC Net [3]	VGG19	Faster R-CNN [31] N=100	LSTM	65.14	-
SeqGROUND [9]	ResNet50	Faster R-CNN N=200	LSTM	61.06	-
CITE [27]	VGG16	Faster R-CNN N=200	Word2vec, FV	61.89	184
DDPN [46]	ResNet101	Faster R-CNN N=100	LSTM	73.3	196
SL-CCRF [19]	ResNet50	Bottom-Up Attention [1] N=100	LSTM	74.69	-
VS-graph [16]	VGG16	Faster R-CNN N=100	LSTM	76.87	-
LCMCG [20]	ResNet101	Faster R-CNN N=100	Bert	76.74	-
FAOS-FV [42]	Darknet53	None	Word2vec, FV	68.38	**16**
FAOS-Bert [42]	Darknet53	None	Bert	68.69	38
VGTR [10]	ResNet101	None	LSTM	75.25	50
M-DGT_FV	ResNet50	None	Word2vec, FV	78.21	67
M-DGT_LSTM	ResNet50	None	LSTM	77.67	74
M-DGT_Bert	ResNet18	None	Bert	77.02	66
M-DGT_Bert	ResNet50	None	Bert	79.32	91
M-DGT_Bert	ResNet101	None	Bert	**79.97** (↑ 4.72%)	108

Table 2. The corresponding accuracy (%) of categories in the test set of Flickr30k Entities data.

Methods	person	clothing	body-parts	animals	vehicles	instruments	scene	other
CCA [29]	64.73	46.88	17.21	65.83	68.75	37.65	51.39	31.77
SPC+PPC [28]	71.69	50.95	25.24	76.25	66.5	35.8	51.51	35.98
QRC Net [3]	76.32	59.58	25.24	80.5	78.25	50.62	67.12	43.6
SeqGROUND [9]	76.02	56.94	26.18	75.56	66	39.36	68.69	40.6
CITE [27]	. 73.20	52.34	. 30.59	76.25	75.75	. 48.15	55.64	42.83
SL-CCRF [19]	84.41	78.51	46.74	. 88.89	81.41	64.97	75.95	57.57
LCMCG [20]	86.82	79.92	53.54	90.73	84.75	63.58	77.12	58.65
VS-graph [16]	86.57	79.92	52.77	**91.89**	85.25	58.64	78.78	59.04
M-DGT_FV	87.79	79.12	55.21	89.02	88.20	65.97	79.03	59.70
M-DGT_LSTM	89.10	78.06	55.17	88.90	88.28	65.50	78.77	58.12
M-DGT	**89.41**	**80.12**	**56.91**	90.74	**88.74**	**66.32**	**79.8**	**61.22**

Table 3. Comparisons with state-of-the-art methods on RefCOCO [44], RefCOCO+ [44] and RefCOCOg [23] in terms of top-1 accuracy with IOU threshold 0.5. The best reported results of these leading methods are presented.

Methods	RefCOCO			RefCOCO+			RefCOCOg	
	Val	TestA	TestB	Val	TestA	TestB	Val	Test
CGR [21]	-	74.04	73.43	-	60.26	55.03	55.03	55.03
SLR [45]	77.48	76.58	78.94	60.5	61.39	58.11	69.93	69.03
VGTR [10]	79.20	82.32	73.78	63.91	70.09	56.51	62.28	67.23
MAttNet [43]	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
ParallelAttn [24]	81.67	80.81	81.32	64.18	66.31	61.46	-	-
AccumulateAttn [6]	81.27	81.17	80.01	65.56	68.76	60.63	-	-
LGRANs [37]	82	81.20	84.00	66.6	67.6	65.5	75.4	74.7
VS-graph [16]	82.68	82.06	84.24	67.70	69.34	65.74	75.73	75.31
TransVG [7]	81.02	82.72	78.35	64.82	70.70	56.94	67.02	67.73
ResNet50:								
M-DGT_FV	83.14	82.4	84.26	68.47	69.31	66.76	76.91	76.3
M-DGT_LSTM	82.95	81.84	84.13	67.01	69.16	66.03	75.93	75.54
M-DGT_Bert	84.05	83.6	85.86	68.91	70.76	67.33	77.91	77.16
ResNet101:								
M-DGT_LSTM	83.98	83.01	85.24	68.31	70.06	67.14	76.83	76.34
M-DGT_Bert	**85.37**	**84.82**	**87.11**	**70.02**	**72.26**	**68.92**	**79.21**	**79.06**

Figure 4. Success cases of our M-DGT on challenging instances from three datasets. The ground truth regions are drawn by black boxes. For instances from the Flickr30k Entities, from left to right, the queries are *A baseball player wearing white with blue sleeves and a gold helmet is swinging to hit a ball*, *Group of tourist posing for a photo* , *People outside an ice cream shop that has summer decorations for sale*, *Three girls are running on a field in front of a fence*. For the instances to the right of the solid line, the corresponding queries are *tv in front of middle boy* and *kid in gray t shirt* for the ReferCOCO, *the man in the background* and *guy at net* for the ReferCOCO+, and *a woman with brown hair and a burgundy shirt standing with a wii - mote in her hand infront of a fireplace* and *scissors positioned second from left* for the ReferCOCOg.

DGT with progressive search can continuously optimize the bounding boxes, thereby preventing the grounding process from being hindered by the failure of the previous stage, such as bad candidate regions or inappropriate attention assignment. Besides, as shown by Fig. 5, M-DGT can sufficiently model the spatial and semantic relations by the multi-modal graph transformer structure, boosting the detection and meticulous adjustments of the bounding boxes. This is further verified by the highest accuracy in many categories of the Flickr30K Entities dataset, especially the person and body-parts shown in Table 2.

Efficiency and generality. Compared with other methods that rely on hundreds of well-prepared regions provided by an external trained model, our framework can be trained directly based on 48 roughly initialized boxes to achieve the best performance. Further, as shown in Table 1, our M-DGT costs only $108ms$ for the inference, which presents a competitive efficiency compared to those methods that also do not rely on region proposals. In addition, as we can see in Fig. 5, the M-DGT continuously remove the abundant or useless edges and nodes to reduce the required computation. Then, since the only box initialization method utilized by our M-DGT is to generate single-scale boxes that can cover the image, our model can be easily adapted to any dataset. This contributes to the high generality of the M-DGT.

5.2. Ablation Experiments

Shown in Table 4, we first study the performance of M-DGT working directly on the region proposals provided by Faster R-CNN [31] (FR) and Bottom-Up Attention [1] (BA). Compared with M-DGT that builds the graph from boxes simply initialized to cover the image, both methods M-DGT_FR and M-DGT_BA present relatively lower accuracy and higher inference time. The main reason is that the further grounding process can be hindered by the bad region candidates that do not cover the target regions. This shows the necessity of our idea that gradually search the target region from simple initialized boxes in a progressive manner. We then conduct experiments on components ablation. The general transformer in work [2] is utilized as a replacement for our proposed node transformer (NT). Without the iteration-related training (IRT), the M-DGT is only optimized once in the final iteration. The results empirically demonstrate the effectiveness of the node transformer on performance improvement. The graph transformer contributes to accuracy and efficiency as it reduces the graph scale and eliminates semantic ambiguity. Specifically, training our M-DGT with IRT is important for maintaining high-level performance. However, the efficiency contribution of IRT is unclear. Finally, more iterations significantly boost the grounding accuracy, but there is no improvement or even a decrease in the performance after the 7-th iteration.

5.3. Plugin Experiments

We further utilize M-DGT as the plugin at the end of existing state-of-the-art methods. More specifically, the boxes B are initialized with the bounding boxes predicted by these methods. The results in Table 5 present that our M-DGT can adjust outputs of these methods to improve the accuracy. The accuracy of the listed methods is increased by 1.12, 1.284, and 0.964 on Flickr30k Entities, RefCOCO, and RefCOCO+ datasets, respectively. This proves that our M-

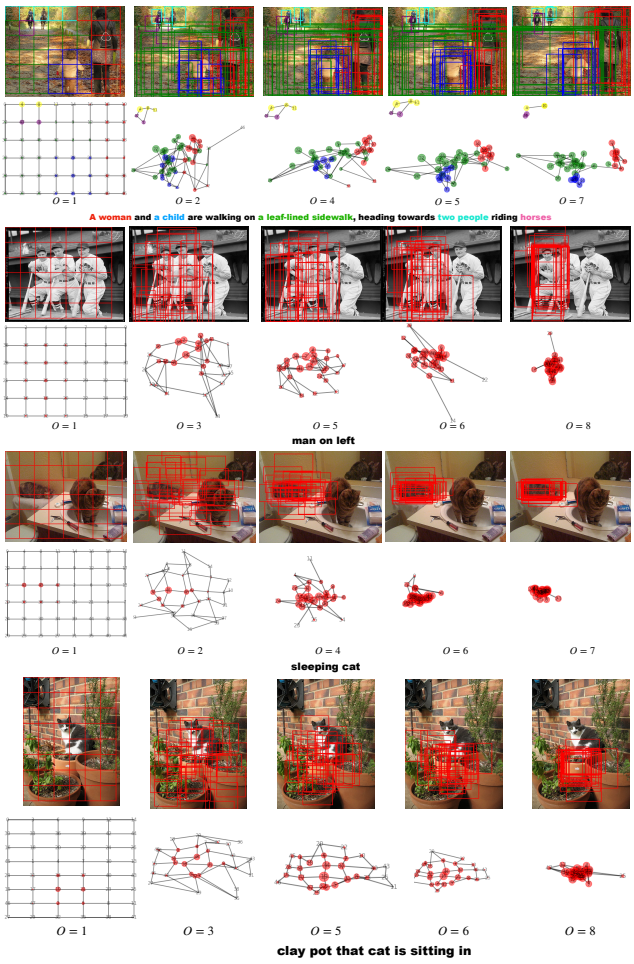

Figure 5. The graph transformation in each iteration of the M-DGT on instances from three datasets. M-DGT achieves the efficient search-based progressive visual grounding by implementing the graph transformation. Thus, the dynamic graphs, shown in the second row of each subfigure, are equivalent to the bounding boxes' continuous adjustment process. The first, second, third, and fourth are the instances from the Flickr30k Entities, RefCOCO, RefCOCO+, and RefCOCOg, respectively.

DGT has the sustainable adjustment ability, thereby adjusting the inaccurate bounding boxes to reach the ground-truth regions. However, the improved accuracy of these methods is still lower than our full M-DGT. These limited improvements demonstrate that one shortcoming of M-DGT is unable to obtain the correct region for the query once the initial set of boxes does not cover the ground-truth region.

6. Conclusion

We have proposed the search-based mechanism to remodel the visual grounding into a progressively optimized visual semantic alignment process. Taking visual regions and query semantics as nodes and spatial relationships as edges, our proposed multi-modal dynamic graph transformer (M-DGT) can model this process as graph transformation. M-DGT can continuously adjust the nodes and edges to shrink the dynamic graph to the target layout, making the corresponding boxes progressively approach the ground truth regions. With an average of 48 simple initialization boxes, the performance of M-DGT, in terms of the accuracy and the IOU scores, on Flickr30k Entities and three RefCOCO datasets significantly outperforms the alternative state-of-the-art methods. Besides, our analyses reveal that M-DGT can greatly optimize the predicted bounding boxes of existing methods. In future work, we plan to model relations between graphs generated in the learning process.

Table 4. Ablation studies of M-DGT in Flickr30k Entities dataset. The M-DGT using different combinations of components, including node transformer (NT), graph transformer (GT), and iteration-related training (IRT). The performance of M-DGT with different iteration numbers.

Boxes initialization using region proposals				
Methods		Region Proposals	Acc@0.5	Time (ms)
M-DGT_FR		Faster R-CNN	74.901	268
M-DGT_BA		Bottom-Up Attention	76.162	256
Ablation of components				
NT	GT	IRT	Acc@0.5	Time (ms)
✓			71.47	177
	✓		72.21	71
		✓	69.32	84
✓	✓		77.9	101
✓		✓	76.526	175
	✓	✓	76.803	57
M-DGT under different #iterations				
M-DGT (3)			69.815	61
M-DGT (5)			75.552	75
M-DGT (7)			79.317	91
M-DGT (9)			79.315	129

Table 5. The accuracy under the IOU threshold 0.5 of using M-DGT as the plugin to further adjust the predicted bounding boxes of leading methods.

Datasets		SL-CCRF [19]	LCMCG [20]	VS-graph [16]
Flickr30k	Val	↑1.13	↑1.26	↑0.98
Datasets		AccumulateAttn [6]	LGRANs [37]	VS-graph [16]
RefCOCO	Val	↑1.31	↑1.48	↑0.96
	TestA	↑1.03	↑1.21	↑0.93
	TestB	↑1.32	↑1.06	↑1.2
RefCOCO+	Val	↑0.96	↑0.83	↑1.05
	TestA	↑ 0.89	↑ 0.77	↑ 1.06
	TestB	↑ 1.04	↑ 0.98	↑ 1.1

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, 6, 7

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 7

[3] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 824–832, 2017. 1, 2, 6

[4] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7239–7248, 2018. 2

[5] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2601–2610, 2019. 1

[6] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7746–7755, 2018. 1, 2, 6, 8

[7] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 1, 2, 6

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 3

[9] Pelin Dogan, Leonid Sigal, and Markus Gross. Neural sequential phrase grounding (seqground). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4175–4184, 2019. 1, 2, 6

[10] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Visual grounding with transformers. *arXiv preprint arXiv:2105.04281*, 2021. 6

[11] Zicong Fan. *Visual grounding through iterative refinement*. PhD thesis, University of British Columbia, 2020. 2

[12] 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. 3

[13] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 6

[14] 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. 1

[15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*, 2015. 5

[16] Chenchen Jing, Yuwei Wu, Mingtao Pei, Yao Hu, Yunde Jia, and Qi Wu. Visual-semantic graph matching for visual grounding. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4041–4050, 2020. 1, 2, 6, 8

[17] 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

[18] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 1, 2

[19] Jiacheng Liu and Julia Hockenmaier. Phrase grounding by soft-label chain conditional random field. In *EMNLP/IJCNLP*, 2019. 1, 2, 6, 8

[20] 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. 1, 2, 6, 8

[21] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7102–7111, 2017. 2, 6

[22] 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. 2

[23] 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. 6

[24] Julian Richard Medina and Jugal Kalita. Parallel attention mechanisms in neural machine translation. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 547–552. IEEE, 2018. 1, 2, 6

[25] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016. 2

[26] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017. 2

[27] Bryan A Plummer, Paige Kordas, M Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–264, 2018. 1, 2, 6

[28] Bryan A Plummer, Arun Mallya, Christopher M Cervantes,

Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1928–1937, 2017. 1, 2, 6

[29] 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. 2, 6

[30] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 6, 7

[32] 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. 6

[33] 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. 1, 2

[34] 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. 5

[35] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR 2018*, 2017. 4

[36] 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. 1, 2, 6

[37] 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 Conference on Computer Vision and Pattern Recognition*, pages 1960–1968, 2019. 2, 6, 8

[38] Peixi Xiong, Huayi Zhan, Xin Wang, Baivab Sinha, and Ying Wu. Visual query answering by entity-attribute graph matching and reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8357–8366, 2019. 1, 2

[39] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2193–2202, 2017. 1

[40] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4145–4154, 2019. 2

[41] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 387–404. Springer, 2020. 1, 2

[42] 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 International Conference on Computer Vision*, pages 4683–4693, 2019. 1, 2, 4, 6

[43] 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. 1, 2, 6

[44] 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. 6

[45] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7282–7290, 2017. 2, 6

[46] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. In *IJCAI*, 2018. 1, 2, 6

[47] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. 2019. 5

[48] Li Zhang, Dan Xu, Anurag Arnab, and Philip HS Torr. Dynamic graph message passing networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3726–3735, 2020. 2