Relation-aware Instance Refinement for Weakly Supervised Visual Grounding

Yongfei Liu[1,5,6*] Bo Wan[1,2*] Lin Ma[3] Xuming He[1,4]
[1]School of Information Science and Technology, ShanghaiTech University
[2]Department of Electrical Engineering (ESAT), KU Leuven
[3]Meituan [4]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[5]Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences
[6]University of Chinese Academy of Sciences
{liuyf3, wanbo, hexm}@shanghaitech.edu.cn forest.linma@gmail.com

Abstract

Visual grounding, which aims to build a correspondence between visual objects and their language entities, plays a key role in cross-modal scene understanding. One promising and scalable strategy for learning visual grounding is to utilize weak supervision from only image-caption pairs. Previous methods typically rely on matching query phrases directly to a precomputed, fixed object candidate pool, which leads to inaccurate localization and ambiguous matching due to lack of semantic relation constraints. In our paper, we propose a novel context-aware weakly-supervised learning method that incorporates coarse-to-fine object refinement and entity relation modeling into a two-stage deep network, capable of producing more accurate object representation and matching. To effectively train our network, we introduce a self-taught regression loss for the proposal locations and a classification loss based on parsed entity relations. Extensive experiments on two public benchmarks Flickr30K Entities and ReferItGame demonstrate the efficacy of our weakly grounding framework. The results show that we outperform the previous methods by a considerable margin, achieving 59.27% top-1 accuracy in Flickr30K Entities and 37.68% in the ReferItGame dataset respectively[1].

1. Introduction

Cross-modal understanding of visual scene and natural language description plays a crucial role in bridging human and machine intelligence, and has attracted much interest from AI community [13]. Towards this goal, one core problem is to establish instance-level correspondence between visual regions and its related language entities, which is commonly referred to as visual grounding [15]. Such correspondence serves as a fundamental building-block for many vision-language tasks, such as image captioning [6, 32], visual question answering [43, 23], visual navigation [37, 45] and visual dialog[18, 7].

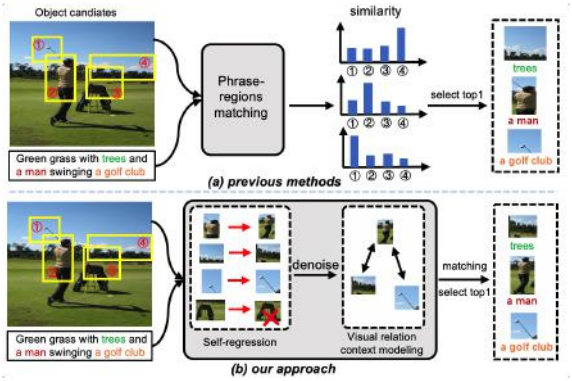

Figure 1. Comparison of visual entities representation with existing weakly-supervised grounding models. (a) Previous methods directly match between noun phrases and a precomputed, fixed object proposals. (b) Our approach is capable of refining the initial object proposals and enriching their representation with visual relation context cues.

Much progress has been made recently in learning visual grounding with strong supervision [22, 39], which requires costly annotations on region-phrase correspondence. A more scalable modeling strategy is to learn from only image-caption pairs, namely *weakly-supervised visual grounding* [29, 40, 2, 8, 20]. Nevertheless, learning from such weak supervision is particularly challenging mainly due to the severe ambiguity in visual object location and in correspondence between diverse noun phrases and object entities during cross-modal learning.

Most existing approaches tackle those challenges via the

*Both authors contributed equally. This work was done when Yongfei Liu was a research intern at Tencent AI Lab, and Bo Wan was a master student in ShanghaiTech University. This work was supported by Shanghai NSF Grant (No. 18ZR1425100).

[1]Code is available at https://github.com/youngfly11/ReIR-WeaklyGrounding.pytorch.git

Multiple Instance Learning (MIL) framework [14] using object candidates generated by a pre-trained object detector [29, 4, 2]. Despite their promising results, these learning pipelines often suffer from the visual and matching ambiguity from several aspects. First, they usually rely on a precomputed object proposal set that contain many distractor or background regions, making it difficult to infer positive matches for learning. In addition, these proposals are typically kept fixed during learning, which leads to inaccurate localization bounded by the external detectors (cf. Fig. 1(a)). Furthermore, these methods often represent noun phrase or visual object context in an implicit manner, using attention-based feature aggregation or encoding predicate triples [20, 21]. Such representations are limited in capturing rich semantic constraints from relations in an image-sentence pair, resulting in cross-modal matching ambiguity in both learning and prediction.

To address the afore-mentioned limitations, we propose a flexible and context-aware object representation for weakly-supervised visual grounding in this work. Unlike previous work, our representation is capable of refining the spatial locations of object proposals using a self-taught mechanism, and incorporates a relation-aware context model by exploiting the language prior (cf. Fig. 1(b)). Such enriched representation alleviates the impact from the inaccurate object detection and the cross-modal matching ambiguity. To achieve this, we develop a coarse-to-fine matching strategy modeled as a two-stage deep network. The first stage of our model consists of a backbone and a coarse-level matching network for proposal generation and refinement, while the second stage builds a visual object graph network and a fine-level matching network for context modeling and final matching prediction.

Specifically, given a pair of image and language description, we first use the backbone network to generate a set of object proposals with their visual features and compute the language embedding for the noun phrases. Then the coarse-level matching network selects a small set of relevant proposals for each phrase and refines their spatial locations. For the second stage, we construct the visual object graph network on the refined proposals by exploiting parsed language structure, which enriches object features with their relations and context. Based on the context-aware representation, the fine-level matching network finally predicts instance-level correspondence between phrases and object proposals, as well as further refined object locations.

To train our deep network in a weak supervision setting, we introduce a novel multi-task loss function to exploit both the model prediction and linguistic relation cues. In particular, we first devise a self-taught regression loss for the proposal location refinement, which employs highly confident proposal predictions as pseudo groundtruth for their neighboring proposals. Moreover, we develop a classification loss on visual relation types based on the output of an external language parser. This enables us to generate effective supervision from the noisy language parsing results for learning better entity representations.

We conduct extensive experiments on two public benchmarks: Flickr30K Entities [27] and ReferItGame [16]. The experiment results show that our method outperforms the prior state-of-the-art with a considerable margin. To validate the effectiveness of each model component, we also provide the detailed ablative study on Flickr30K Entities dataset. The main contributions of our work are three-folds:

- We adopt a coarse-to-fine strategy to refine object proposals and alleviate semantic ambiguities by enriching visual feature with relationship constraints.
- We propose a self-taught regression loss to supervise object proposal refinement, and introduce an additional visual relation loss that helps learn a context-aware object representation.
- Our method achieves new state-of-the-art performance on Flickr30K Entities and ReferItGame benchmarks.

2. Related Work

Visual Grounding Visual grounding [26, 42, 22, 25, 41, 3] aims to learn region-phrase correspondence **with bounding box annotation** for each phrase during training stage. In recent years, the deep network is widely used in this task and achieves remarkable success. Plummer et al. [26] devised a single end-to-end network to learn multiple text-conditioned embedding for grounding and DDPN [42] proposed to generate a group of high-quality proposals with a diversified and discriminate network. However, they ignored the semantic context cues and relation constraints in both vision and language. To address this problem, Nagaraja et al. [24] explored to utilize LSTM to encode visual and linguistic context for referring expression, and SeqGROUND [25] adopts chain-structure LSTMs to encode context in cross-domain with a history stack for visual grounding. Besides, Wang et al. [35] took a self-attention mechanism to capture their context in a sentence and build a directed graph over neighbor objects to exploit the visual relations, and Liu et al. [22] aimed to build a cross-modal graph network under the guidance of language structure to learn global context representation for both phrases and visual objects. Although these methods demonstrate superior performance on visual grounding, they highly rely on the strong supervision that is too expensive to obtain in most scenarios. Thus the main focus of this work is on learning cross-modal matching in a weak supervision setting.

Weakly-Supervised Visual Grounding Different from supervised visual grounding, a weakly-supervised setting aims to learn the fine-grained region-phrase correspondence **with only image-sentence association**. Most recent

works [29, 40, 2, 4, 8, 34, 20, 36] take a hypothesis-and-matching strategy for the weakly-supervised visual grounding task, in which they first generate a set of region proposals from an image with an external object detector, and then match between query phrases and those regions. WPT [33] directly computed the cross-modal similarity between noun phrases and detected multi-level visual concepts from amounts of object detectors. GroundR [29] built correspondences by reconstructing phrases with an attention mechanism on visual features. To explore more powerful supervision, KAC-Net [2] took a similar formulation but exploited visual consistency and knowledge from object categories, and [4] adopted a ranking-loss to minimize the distances between associated image-caption and maximize the distance between irrelevant pairs. Besides, [8, 34] introduced a contrastive loss to distillate knowledge from external language [5] and visual models [12].

Although these methods discovered various type of supervision for weakly-supervised visual grounding, they suffered from limited objects recall and all the methods above fail to refine object regions due to the lack of location supervision. MATN [44] solved this problem by introducing a transformation network to search target phrase location over the entire image directly, and such locations were regularized by the precomputed proposals. Only a few work took into account context cues to eliminate semantic ambiguities in a weakly-supervised setting: ARN [20] extracted the linguistic and visual cues on entity, location and context levels separately that enforced multi-level cross-modal consistency. KPRN [21] further exploited linguistic context and required subject & object matching simultaneously. Our focus is to exploit context-aware instance refinement weakly-supervised learning strategy for both limitations.

3. Method

3.1. Problem Setting and Overview

The task of weakly-supervised phrase grounding aims to localize the noun phrases of a language description in an associate image, while the correspondences between the noun phrases and image regions are not available for training. Formally, we aim to learn a visual grounding model $\mathcal{M}$, which takes an input image I and a description D with a group of noun phrases $\mathcal{Q} = \{q_i\}_{i=1}^{N}$ and predicts the corresponding locations $\mathcal{B} = \{b_i\}_{i=1}^{N}$ for the query phrases, i.e., $\mathcal{B} = \mathcal{M}(I, D, \mathcal{Q})$. For the weakly-supervised learning setting, we are only provided a training dataset $\mathcal{X} = \{(I^{(l)}, D^{(l)}, \mathcal{Q}^{(l)})\}_{l=1}^{L}$ of size L, where the corresponding locations $\mathcal{B}^{(l)}$ of phrases $\mathcal{Q}^{(l)}$ are unobserved.

In this work, we adopt a typical grounding strategy that first uses an external object detector (e.g., [28]) to generate a group of visual object proposals $\mathcal{O}$ from I, which are then matched to the phrases [29, 40, 2]. Learning a cross-modal matching with weak supervision, however, is particularly challenging in such a generate-and-match framework due to visual ambiguity caused by inaccurate object detection and the lack of instance-level region-phrase correspondence.

To address those challenges, we propose to learn a flexible context-aware entity representation based on the language prior and a coarse-to-fine matching process, which enables us to mitigate the impact of the matching and localization ambiguity. We instantiate our strategy with a two-stage deep network, as illustrated in Fig. 2. Specifically, the first-stage network extracts the visual and textual features from inputs, and perform a coarse-level matching in which we estimate the similarity scores between each q_i and $\mathcal{O}$ and a refinement of object proposal locations. In the second stage, we select a small group of relevant proposals from $\mathcal{O}$ for each phrase q_i according to the similarity scores, and build a visual object graph network by exploiting parsed language structure. Our second-stage network performs message passing to enhance the visual representation with contextual cues, and finally predicts a fine-level similarity score for each object-phrase pair as well as further refinement of object locations.

To train our model, we develop a joint learning strategy with a multi-task loss, which additionally incorporates a self-taught regression loss to refine the object locations and a language-induced relation classification loss to enforce a relational constraint on the entity matching. Below we first introduce the details of our model architecture in Sec. 3.2 followed by our design of loss functions in Sec. 3.3.

3.2. Model Architecture

We now introduce our two-stage network which consists of four sub-modules and can be divided into two stages. The first stage of our network includes a *Backbone Network* to extract visual and linguistic features (Sec. 3.2.1), and a *Coarse-level Matching Network* to refine object locations and select a subset of relevant proposals for each phrase (Sec. 3.2.2). For the second-stage network, we build a *Visual Object Graph Network* to capture visual context cues by message passing (Sec. 3.2.3), and a *Fine-level Matching Network* to compute the final matching and object locations with context-aware features (Sec. 3.2.4).

3.2.1 Backbone Network

Our first module is a backbone network consisting of a convolutional network for extracting visual features and a recurrent network for encoding language features.

The convolutional network (e.g. ResNet [10]) takes the image I as input and computes a feature map $\mathbf{\Gamma}$. An external object detector (e.g. Faster R-CNN [28]) provides a set of object proposals $\mathcal{O} = \{\langle o_m, c_m \rangle\}_{m=1}^{M}$, where $o_m \in \mathbb{R}^4$ denotes object regions and $c_m \in \{1, 2, ..., C\}$ indicates object category. Similar to [22], for each o_m, we use RoI-

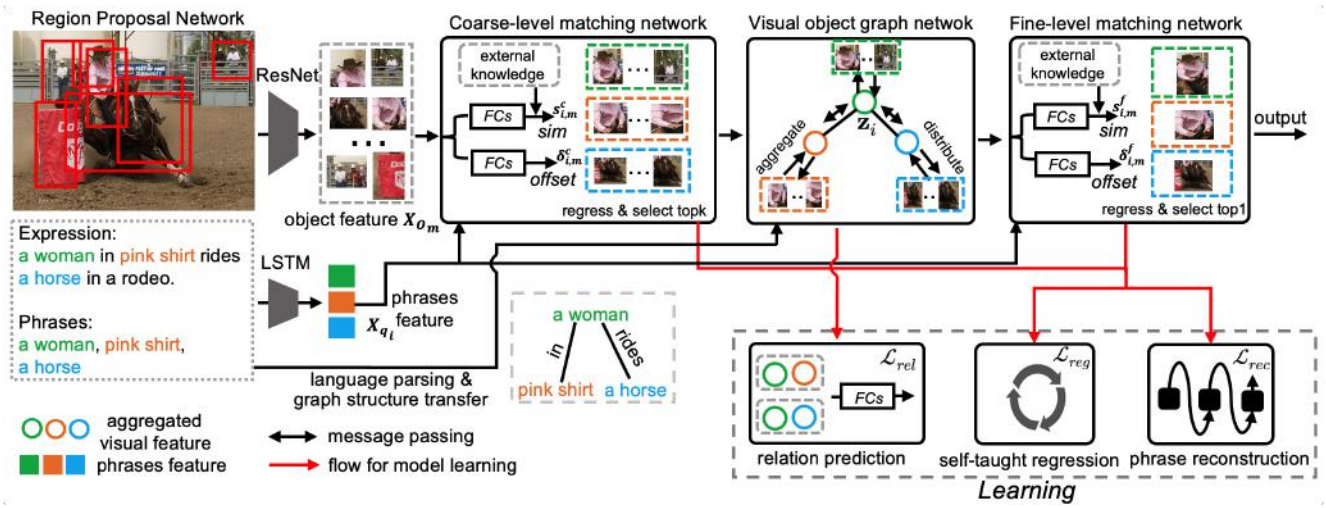

Figure 2. **Model Overview:** There are four modules in our network, the **Backbone Network** prepares basic phrase and visual features; the **Coarse-level Matching Network** selects a small set of objects for each phrase and refines their spatial locations; the **Visual Object Graph Network** enriches the object feature with their context and relations by exploiting language structure, finally the **Fine-level Matching Network** predicts instance-level correspondences and refines their locations further based on the context-aware visual representation. Three main losses are demonstrated to supervise the whole network, like relation classification loss $\mathcal{L}_{rel}$, self-taught regression loss $\mathcal{L}_{reg}$ and phrase reconstruction loss $\mathcal{L}_{rec}$.

Align [9] and global average pooling to compute its conv-feature, which is fused with its spatial feature and embeded to a vector $\mathbf{x}_{o_m} \in \mathbb{R}^d$, where d is the feature dimensions.

For the language features, we compute an embedding of noun phrases $q_i \in \mathcal{Q}$ as follows. We first embed the words in description D into a sequence of vectors $\mathbf{H} = \{\mathbf{h}_t\}_{t=1}^T$ via an encoding LSTM [31], where T is the word length in D. The language feature $\mathbf{x}_{q_i}$ of each phrase q_i is computed by taking average pooling on its word representations:

$$\mathbf{H} = \text{LSTM}_{enc}(D), \quad \mathbf{x}_{q_i} = \frac{1}{|q_i|}\sum_{t \in q_i} \mathbf{h}_t \tag{1}$$

where $|q_i|$ indicates the phrase length in words, and the embeddings $\mathbf{h}_t, \mathbf{x}_{q_i} \in \mathbb{R}^d$, which have the same dimensionality as the visual features.

3.2.2 Coarse-level Matching Network

Our second module performs a coarse-level matching between phrases and the initial object proposals, aiming to select a small set of relevant proposals and refine their spatial locations. To this end, for each phrase-boxes pair $\{q_i, o_m\}$, we compute a similarity score $\hat{s}^c_{i,m}$ and regression offsets $\delta^c_{i,m} \in \mathbb{R}^4$ according to the phrase feature $\mathbf{x}_{q_i}$ and object feature $\mathbf{x}_{o_m}$ as follows:

$$\hat{s}^c_{i,m} = F_{cls}(\mathbf{x}_{q_i}, \mathbf{x}_{o_m}), \quad \delta^c_{i,m} = F_{reg}(\mathbf{x}_{q_i}, \mathbf{x}_{o_m}) \tag{2}$$

where F_{cls} and F_{reg} are two fully-connected networks.

Following [2], we further utilize object categories as a semantic cue to compute an additional similarity score $s^a_{i,m}$ in the linguistic space:

$$s^a_{i,m} = \langle E_{ext}(q_i), E_{ext}(c_m) \rangle \tag{3}$$

where $\langle \cdot, \cdot \rangle$ indicates the inner product, and E_{ext} represents an off-the-shelf language embedding (e.g., Skip-thoughts [17]). Finally, we fuse the above two similarity scores and compute an attention weight by taking the Softmax over all the proposals:

$$s^c_{i,m} = \hat{s}^c_{i,m} \cdot s^a_{i,m}, \quad \alpha^c_{i,m} = \underset{m \in [1:M]}{\text{Softmax}}(s^c_{i,m}) \tag{4}$$

To refine the proposal set, we apply the regression offsets $\boldsymbol{\delta}^c_i = \{\delta^c_{i,m}\}_{m=1}^M$ and select the top $K (K \ll M)$ proposals for each phrase q_i based on the similarity scores $\mathbf{s}^c_i = \{s^c_{i,m}\}_{m=1}^M$. This generates a set of refined proposals $\mathcal{V}_i = \{o_{i,k}\}_{k=1}^K$ for phrase q_i, and we denote the proposal set for all the noun phrases as $\mathcal{V} = \{\mathcal{V}_i\}_{i=1}^N$.

3.2.3 Visual Object Graph Network

Given the refined proposal sets, we introduce a graph neural network, dubbed as the Visual Object Graph Network, to capture the visual context with the guidance of language structure. Specifically, we first extract a set of relation phrases from the description D with an external language parser[2] [30, 38, 22]. We then build a graph network with N nodes and its i-th node, denoted by $\mathbf{z}_i$, encodes the visual feature for phrase q_i. Two nodes $\mathbf{z}_i$ and $\mathbf{z}_j$ are connected if a relation phrase exists between two phrases q_i and q_j.

[2] 1 https://github.com/vacancy/SceneGraphParser

We initialize the node feature $\mathbf{z}_i$ based on the phrase proposal set $\mathcal{V}_i$ and its attention scores $\{\alpha^c_{i,k}\}_{k=1}^K$, which represents an estimation of the visual feature for q_i:

$$\mathbf{z}_i = \sum_{k=1}^{K} \alpha^c_{i,k} \cdot \mathbf{x}_{o_{i,k}} \tag{5}$$

where $\mathbf{x}_{o_{i,k}}$ are the object proposal features, and $\alpha^c_{i,k}$ are from Eq. 4. Subsequently, our graph network refines the visual features $\{\mathbf{z}_i\}$ by a message passing step as below:

$$\mathbf{z}'_i = \mathbf{z}_i + \sum_{j \in \mathcal{N}(i)} \omega_{i,j} F_M(\mathbf{z}_j) \tag{6}$$

$$\omega_{i,j} = \underset{j \in \mathcal{N}(i)}{\text{Softmax}}(F_M(\mathbf{z}_i)^\intercal F_M(\mathbf{z}_j)) \tag{7}$$

where $\mathbf{z}'_i$ denotes the updated visual features, $\mathcal{N}(i)$ is the neighborhood of node i, F_M is a multi-layer network for computing messages, and $\omega_{i,j}$ is an attention weight between node i and j. Finally, we update each object proposal feature, denoted by $\{\mathbf{x}'_{o_{i,k}}\}$, with the visual features of its corresponding phrase q_i as follows:

$$\mathbf{x}'_{o_{i,k}} = \mathbf{x}_{o_{i,k}} + \alpha^c_{i,k} \cdot \mathbf{z}'_i, \tag{8}$$

where the visual context in $\mathbf{z}'_i$ is distributed to $o_{i,k}$ with attention weighting.

3.2.4 Fine-level Matching Network

Given the context-aware features, our final module performs a fine-level matching between phrases and the refined object proposal subsets. Similar to the coarse-level matching, we predict a similarity score $s^f_{i,k}$, an attention weight $\alpha^f_{i,k}$, and a phrase-specific regression offset $\delta^f_{i,k} \in \mathbb{R}^4$ for each phrase-proposal pair $\{q_i, o_{i,k}\}$ as below:

$$s^f_{i,k} = F_{cls}(\mathbf{x}_{q_i}, \mathbf{x}'_{o_{i,k}}) \cdot s^a_{i,k}, \quad \alpha^f_{i,k} = \underset{k \in [1:K]}{\text{Softmax}}(s^f_{i,k}) \tag{9}$$

$$\delta^f_{i,k} = F_{reg}(\mathbf{x}_{q_i}, \mathbf{x}'_{o_{i,k}}) \tag{10}$$

Model Inference During the model inference, for each query q_i, we first compute an overall matching scores $\{s_{i,k}\}_{k=1}^K$ for the candidate proposals in $\mathcal{V}_i$ by fusing the coarse-level and fine-level scores:

$$s_{i,k} = s^c_{i,k} \cdot s^f_{i,k}, \tag{11}$$

followed by applying the estimated offset $\delta^f_{i,k}$ to their locations. Finally, we take the proposal $o^*_{i,k}$ with the maximum similarity score $s^*_{i,k}$ as the grounding result of phrase q_i.

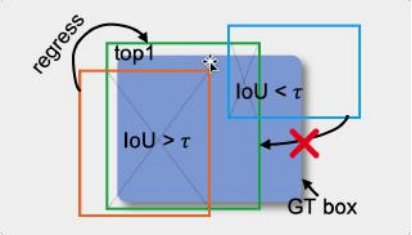

Figure 3. Illustration of self-taught regression. The green box has the highest confident matching score. The orange box overlapping with the green box with IoU $> \tau$ will be regressed toward the green box, while the blue box overlapping with the green box with IoU $< \tau$ will stay unchanged. The GT box region here is unobserved during training.

3.3. Learning with Weak Supervision

We now introduce our weakly supervised learning strategy for training the two-stage deep network. To this end, we develop a multi-task loss that incorporates two novel supervision signals from a partially trained model itself and a language prior on entity relations, respectively.

Specifically, our overall loss function $\mathcal{L}$ consists of four terms, including a reconstruction loss $\mathcal{L}_{rec}$ for noun phrases, a self-taught regression loss $\mathcal{L}_{reg}$ for refining object proposal locations, a relation classification loss $\mathcal{L}_{rel}$ for language-induced relation cues, and a ranking loss $\mathcal{L}_{rank}$ for image-caption pairs. Formally, this weakly-supervised learning loss can be written as follows,

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_1 \cdot \mathcal{L}_{reg} + \lambda_2 \cdot \mathcal{L}_{rel} + \lambda_3 \cdot \mathcal{L}_{rank} \tag{12}$$

where $\{\lambda_1, \lambda_2, \lambda_3\}$ are weight coefficients to balance the loss terms. In this work, we adopt a similar ranking loss as [4], and leave its details to the Suppl. We will instead focus on the remaining three loss terms below, which are defined for each image-caption pair.

Phrase reconstruction loss $\mathcal{L}_{rec}$ Given the noun phrases, we adopt a phrase reconstruction loss [2, 29] at both stages of our deep network to provide model supervision. As those two reconstruction loss terms are similar, we will use the coarse-level as an example below.

To apply the reconstruction loss, we first generate a visual representation $\mathbf{z}^c_i$ for phrase q_i. Specifically, we aggregate the object features $\{\mathbf{x}_{o_m}\}_{m=1}^M$ with the attention weights $\{\alpha^c_{i,m}\}_{m=1}^M$ in Eq. 4, which can be written as $\mathbf{z}^c_i = \sum_{m=1}^M \alpha^c_{i,m} \cdot \mathbf{x}_{o_m}$. We then reconstruct the noun phrase q_i using the visual feature $\mathbf{z}^c_i$ based on a decoding LSTM. Concretely, we compute a sequence of word distribution $\hat{\mathbf{y}}^c_i = \{\hat{\mathbf{y}}^c_{i,w}\}_{w=1}^{|q_i|}$ as below:

$$\hat{\mathbf{y}}^c_i = \text{LSTM}_{dec}([\mathbf{z}^c_i, q_i]) \tag{13}$$

Similarly, we also predict $\hat{\mathbf{y}}^f_i$ in the fine-level based on a context-aware feature for each phrase q_i as $\mathbf{z}^f_i =$

$\sum_{k=1}^{K} \alpha_{i,k}^{f} \cdot \mathbf{x}'_{o_{i,k}}$. Both stages share the same parameters for the decoder LSTM$_{dec}$. For each phrase, we adopt the standard sequence log loss $L_{\log}$ and the overall reconstruction loss can be written as:

$$\mathcal{L}_{rec} = \sum_{i=1}^{N} \Big(L_{\log}(\hat{\mathbf{y}}_i^c, q_i) + L_{\log}(\hat{\mathbf{y}}_i^f, q_i) \Big) \tag{14}$$

Self-taught regression loss $\mathcal{L}_{reg}$ As phrase locations are not annotated, we introduce a self-taught regression loss for training the location regressors. Particularly, we observe that the proposals with high matching scores often have an accurate localization after several rounds of training without $\mathcal{L}_{reg}$. This motivates us to use the confident proposals from partially-trained models to supervise the location refinement of their neighboring proposals, as shown in Fig. 3. Concretely, for phrase q_i, we denote $\boldsymbol{\delta}_i^{c*} = \{\delta_{i,m}^{c*}\}$ in which $\delta_{i,m}^{c*}$ is the offset between proposal o_m and the most confident proposal if their overlaps are larger than a threshold τ, and otherwise the predicted $\delta_{i,m}^{c}$, which means they will stay unchanged. We then adopt the smooth-L1 loss L_{sm} [28, 9] for the offset regression:

$$\mathcal{L}_{reg} = \sum_{i=1}^{N} \Big(L_{sm}(\boldsymbol{\delta}_i^c, \boldsymbol{\delta}_i^{c*}) + L_{sm}(\boldsymbol{\delta}_i^f, \boldsymbol{\delta}_i^{f*}) \Big) \tag{15}$$

Relation classification loss $\mathcal{L}_{rel}$ We further introduce a pairwise loss on the context-aware features of phrases, which imposes a relational constraint for the context encoding and fine-level matching. Specifically, we first extract relation phrases on the entire dataset by the language parser as described in Sec. 3.2.3, and select C_r most frequent relations to form a set of relationship categories $\mathcal{R} = \{0, 1, 2, \cdots, C_r\}$, where 0 indicates no relation. Then we predict the relation type $\hat{\mathbf{y}}_{i,j}^r \in \mathbb{R}^{C_r}$ between a pair of $\{q_i, q_j\}$ according to their fine-level context-aware object features with a multi-layer network F_{rel} as follows,

$$\hat{\mathbf{y}}_{i,j}^r = F_{rel}(\mathbf{z}_i^f, \mathbf{z}_j^f). \tag{16}$$

Denote the relation labels of $\{q_i, q_j\}$ as $r_{ij} \in \mathcal{R}$, we use the cross entropy loss for the relation classification if $r_{ij} > 0$:

$$\mathcal{L}_{rel} = \sum_{i,j} L_{ce}(\hat{\mathbf{y}}_{i,j}^r, r_{ij}). \tag{17}$$

4. Experiments

In this section, we first depict the experimental setup and implementation details; then compare our model with previous arts. Detailed ablation studies are also conducted to validate each components in our model. Finally, we demonstrate several qualitative results to show model efficacy.

4.1. Datasets and evaluation metric

Flickr30K Entities [27] contains 29783 images for training, 1000 images for validation and 1000 images for test. Each image is associated 5 captions. For **ReferItGame** [16] dataset, there are around 20k images and 130k query phrases. Each object is referred by 1-3 query phrase and we follow the standard dataset split of [29]. It is worthed noting that we ignore the box annotations for the noun phrases on both datasets during the training stage.

Evaluation Metric: We consider a noun phrase grounded correctly when its predicted box has at least 0.5 IoU with its ground-truth location. The grounding accuracy *Acc* (i.e., Recall@1) is the fraction of correctly grounded noun phrases. We also report the point game metric *PointIt* for a clear comparison with previous methods [11, 1, 4]. Following [4], we define a hit if the center of the predicted bounding box lies in anywhere inside the ground-truth region, and *PointIt* is the percentage of these hits.

4.2. Implementation Details

We generate an initial set of M=50 object proposals with an external RPN [28] pre-trained on Visual Genome [19] dataset, and predict their object categories with the classification head of Faster-RCNN [28]. Then we apply RoI-Align [9] to extract the object visual representation on feature map $\mathbf{\Gamma}$, which is the output of C4 block in ResNet-101 with channel dimension d=2048. In coarse-level matching network, we set K=5 to filter out most irrelevant proposals for each noun phrase. In addition, we select C_r=88 relations, of which frequency are greater than 100.

For model learning, we train the entire network with SGD optimizer with an initial learning rate of 1e-3 and weight decay of 5e-4. The training iterations are up to 80k and the batch size of each is 40. We decay the learning rate by 10 times in 32k and 40k respectively. The hyper-parameters $(\lambda_1, \lambda_2, \lambda_3)$ are set as $(0.1, 1, 1)$ in loss function. The threshold τ=0.6 in self-taught regression and $\mathcal{L}_{reg}$ is applied for training after 7.5k iterations. All optimal hyper-parameters are selected by conducting a grid search on validation set and applied to test set directly once fixed. More details on ReferItGame are described in Suppl.

4.3. Quantitative Results

We compare our model with several previous works in terms of *Acc* and *PointIt* on both Flickr30K Entities [27] and ReferItGame [16] datasets.

Flickr30K Entities: As shown in Tab. 1, our approach outperforms the prior methods by a considerable margin in both evaluation metric, achieving 59.27% on *Acc* and 78.60% on *PointIt*. Compared with reconstruction-based methods, we can outperform KAC Net* [2] by 12.66% on *Acc* and 4.43% on *PointIt*, which demonstrates that our

Methods	Backbone	Language	Proposals	Det Label	Flickr30k		ReferItGame	
					Acc%	*PointIt%*	*Acc%*	*PointIt%*
SSS [11]	VGG18	N/A	-	×	-	49.10	-	49.90
MultiGrounding [1]	PNAS Net	Elmo	-	×	-	69.19	-	48.42
GroundR [29]	VGG16	LSTM	SS	×	28.94	-	10.70	-
MATN [44]	VGG16	LSTM	SS	×	33.10	-	13.61	-
KAC Net [2]	VGG16	LSTM	SS	✓	38.71	-	15.83	-
Align2Ground [4]	RN152	LSTM	FRCNN(VG)	×	11.20	71.00	-	-
UTG [40]	N/A	Glove	YOLOV2(COCO)	✓	36.93	-	20.91	-
ARN [20]	RN101	LSTM	MaskRCNN(COCO)	✓	-	-	26.19	-
KAC Net* [2]	RN101	LSTM	FasterRCNN(VG)	✓	46.61	74.17	33.67	56.57
KPRN[21]	RN101	LSTM	MaskRCNN(COCO)	✓	-	-	33.87	-
Contr. Learning [8]	RN101	Bert	FasterRCNN(VG)	×	51.67	76.74	-	-
Contr. Dist. [34]	RN101	LSTM	FasterRCNN(OI)	✓	50.96	-	27.59	-
ours	RN101	LSTM	FasterRCNN(VG)	✓	**59.27**	**78.60**	**37.68**	**58.96**

Table 1. Comparison of phrases grounding accuracy on Flickr30K Entities and ReferitGame test sets. * denotes the re-implementation using the same backbone and object proposals as ours. SS denotes the selective search and (VG),(COCO),(OI) denote the object detector pre-trained on Visual Genome, MSCOCO, OpenImage dataset.

Methods	TSD	STR	VOGN&RC	*Acc%*
baseline	-	-	-	48.18
	✓	-	-	50.80
	✓	✓(w/o $\mathbf{x}_{q_i}$)	-	54.05
	✓	✓	-	56.88
	✓	-	✓	52.48
	✓	✓	✓(w/o attention)	55.60
ours	✓	✓	✓	58.30

Table 2. Ablation Study on Flickr30K Entities val set.

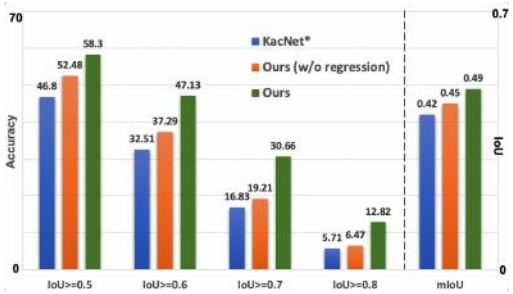

Figure 4. Comparison of grounding accuracy in different IoU threshold and overall mean IoU on Flickr30K Entities val set.

carefully-designed regression loss and visual object graph can solve the spatial and semantic ambuigities simultaneously. When compared with recently proposed contrastive learning based methods [34, 8], we can still improve the performance by 7.60% on *Acc* and 1.96% on *PointIt*, although Contrastive Learning [8] takes more powerful Bert as their language model.[3] In addition, we demonstrate more detailed performance comparisons of per coarse class in Suppl.

ReferItGame: In ReferItGame dataset, we achieve 37.68% on *Acc* and 58.96% on *PointIt*. Our method outperform KPRN [21] by 3.81% on *Acc* and KAC Net* [2] by 3.39% on *PointIt*, which further validates the effectiveness of our method.

4.4. Ablation Study

In this section, we conduct extensive ablation studies on Flickr30K Entities validation set to show effectiveness of each component in our method (Tab. 2).

Baseline: We take the direrctly matching strategy with only Backbone Network and Coarse-level Matching Network (w/o box regression) as our baseline model, which is only supervised by the phase reconstruction loss $\mathcal{L}_{rec}$ and and the ranking loss $\mathcal{L}_{rank}$.

[3]Comparison with concurrent work [36] which is trained with additional supervision from object attributes refers to Suppl.

Two-stage denoising (TSD): As shown in Tab. 2, our two-stage denoising strategy can bring the performance gain of 2.64% on *Acc* compared with the baseline model. Because such strategy helps to filter out most of background distractors and irrelevant objects and thus alleviate difficulties in establishing cross-modal correspondence.

Self-taught regression (STR): Different from the previous work [2, 21, 8, 4, 20] whose performance is directly restricted to the quality of generated object proposals, we improve the grounding accuracy from 50.8% to 56.88% by refining the location of object proposals and thus reduce spatial ambiguities under the supervision of self-taught box regression. In addition, we explore to remove noun phrase feature $\mathbf{x}_{q_i}$ in Eq. 2 & 10 when estimating the proposal offsets based on visual features only, and find *Acc* drop from 56.88% to 54.05%, which suggests the language feature provides semantic-aware guidance for box regression.

To validate its effectiveness further, we remove such self-taught regression loss on the final model and observe a steep accuracy decrease in different IoU threshold across from 0.5 to 0.8, as shown in Fig. 4 (left); and a mean IoU drop shown in Fig. 4 (right).

Visual object graph network (VOGN) & Relation constrain (RC): Visual object graph network enriches each

Figure 5. Visualization of weakly grounding results on Flickr30K validation set. The colored boxes in image correspond to the noun phrases with the same color in sentence. **(a)** demonstrates grounding results when the input sentences are complex, while **(b)** shows results when visual scene is complex. **(c)** shows the effects of self-taught regression and **(d)** illustrates the results with the help of context cues. We denote that the dashed boxes in **(c)** are initial proposals from external detectors, the solid boxes are our predictions after regression. In **(d)**, the dashed boxes are the predictions from our model without visual object graph and relation constraints.

K	3	5	10
Acc%	57.47	58.30	57.15

Table 3. Ablation study of K on Flickr30K Entities val set.

visual representation with their context cues and relation constrain provides additional supervision to learn such representation, as a result it suppresses semantic ambiguity and improve the accuracy from 56.88% to 58.30%. It worth noting that VOGN incorporates with RC to work as a whole in our model, and we find using any separate component will result in a limited contribution to the final results. More details refer to the Suppl.

To further explore the graph structure, we replace the attention weights ω_{ij} in Eq. 7 with non-parametric uniform values by averaging the number of edges. We observe a significant performance drop from 58.30% to 55.60%, which suggests that it is non-trivial to flexiblely learn the graph weights over the whole dataset.

Hyper-parameter K: As shown in Tab. 3, our approach achieves the highest performance when K=5. The performance will drop when K=3, mainly due to lower proposals recall. When K=10, the the performance will drop from 58.30% to 57.15% because of bringing much noisy proposal candidates.

4.5. Qualitative Results

Fig. 5 visualizes a variaty of grounding cases of our final results. We can observe that our model is capable of predicting accurate grounding results when the language description (Fig. 5 **a**) and visual scene (Fig. 5 **b**) are complex. To better understand the capacity of self-taught regression, we also visualize the refined proposals (solid boxes) compared with their initial proposals (dashed boxes) in Fig. 5 **c**, and find that object regions can be regressed to a more accurate location, e.g., in the upper image the initial proposal of *a red ball* is inaccurate with IoU=0.30, and is refined to a precise region with IoU=0.83. In Fig. 5 **d**, we observe the relation constrain can distinguish the target object from similar candidates, demonstrating the effectiveness of such relation-based context information.

5. Conclusion

In this paper, we propose a flexible context-aware instance representation for weakly supervised visual grounding by incorporating coarse-to-fine object refinement and entity relation modeling into a two-stage deep network. Specifically, we develop a coarse-to-fine denoising strategy, which contains a self-taught regression operation to refine object proposals and reduce location ambiguities, and adopt a relation constraint by exploiting language structure to alleviate semantic ambiguities. As a result, we achieve state-of-the-art performance on the public Flickr30K Entities and ReferItGame benchmarks, outperforming previous work with a sizeable margin.

References

[1] Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. In *CVPR*, 2019. 6, 7

[2] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *CVPR*, 2018. 1, 2, 3, 4, 5, 6, 7

[3] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *ICCV*, 2017. 2

[4] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *ICCV*, 2019. 2, 3, 5, 6, 7

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

[6] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *CVPR*, pages 4125–4134, 2019. 1

[7] Dan Guo, Hui Wang, Hanwang Zhang, Zheng-Jun Zha, and Meng Wang. Iterative context-aware graph inference for visual dialog. In *CVPR*, pages 10055–10064, 2020. 1

[8] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *ECCV*, 2020. 1, 3, 7

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 4, 6

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[11] Syed Ashar Javed, Shreyas Saxena, and Vineet Gandhi. Learning unsupervised visual grounding through semantic self-supervision. In *IJCAI*, 2019. 6, 7

[12] Redmon Joseph, Divvala Santosh, Girshick Ross, and Farhadi Ali. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 3

[13] Kushal Kafle, Robik Shrestha, and Christopher Kanan. Challenges and prospects in vision and language research. *Frontiers in Artificial Intelligence*, 2019. 1

[14] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 2

[15] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NeuIPS*, pages 1889–1897, 2014. 1

[16] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 2, 6

[17] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *NeuIPS*, pages 3294–3302, 2015. 4

[18] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*, pages 153–169, 2018. 1

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, page 32–73, 2017. 6

[20] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. In *ICCV*, 2019. 1, 2, 3, 7

[21] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Li Su, and Qingming Huang. Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding. In *ACMMM*, 2019. 2, 3, 7

[22] Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. Learning cross-modal context graph for visual grounding. In *AAAI*, 2020. 1, 2, 3, 4

[23] Jonghwan Mun, Kimin Lee, Jinwoo Shin, and Bohyung Han. Learning to specialize with knowledge distillation for visual question answering. In *NeuIPS*, pages 8081–8091, 2018. 1

[24] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 2

[25] Dogan Pelin, Sigal Leonid, and Gross Markus. Neural sequential phrase grounding (seqground). In *CVPR*, 2019. 2

[26] Bryan A Plummer, Paige Kordas, M Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. In *ECCV*, 2018. 2

[27] 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 *ICCV*, 2015. 2, 6

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Workshop on NeuIPS*, 2015. 3, 6

[29] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 1, 2, 3, 5, 6, 7

[30] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Workshop on Vision and Language (VL15)*, 2015. 4

[31] Hochreiter Sepp and Schmidhuber Jürgen. Long short-term memory. *Neural Computation*, (8):1735–1780, 1997. 4

[32] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, and Wei Liu. Controllable video captioning with pos sequence guidance based on gated fusion network. In *ICCV*, pages 2641–2650, 2019. 1

[33] Josiah Wang and Lucia Specia. Phrase localization without paired training examples. In *ICCV*, 2019. 3

[34] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. *arXiv preprint arXiv:2007.01951*, 2020. 3, 7

[35] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and van den Hengel Anton. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *CVPR*, 2019. 2

[36] Qinxin Wang, Hao Tan, Sheng Shen, Michael W Mahoney, and Zhewei Yao. Maf: Multimodal alignment framework for weakly-supervised phrase grounding. *arXiv preprint arXiv:2010.05379*, 2020. 3, 7

[37] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*, pages 6629–6638, 2019. 1

[38] Yu-Siang Wang, Chenxi Liu, Xiaohui Zeng, and Alan Yuille. Scene graph parsing as dependency parsing. In *NAACL*, 2018. 4

[39] Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *CVPR*, pages 9952–9961, 2020. 1

[40] Raymond A Yeh, Minh N Do, and Alexander G Schwing. Unsupervised textual grounding: Linking words to image concepts. In *CVPR*, 2018. 1, 3, 7

[41] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 2

[42] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. In *IJCAI*, 2019. 2

[43] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019. 1

[44] Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. Weakly supervised phrase localization with multi-scale anchored transformer network. In *CVPR*, 2018. 3, 7

[45] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *CVPR*, pages 10012–10022, 2020. 1