Knowledge-guided Pairwise Reconstruction Network for Weakly Supervised Referring Expression Grounding

Xuejing Liu[1,2], Liang Li[1,*], Shuhui Wang[1], Zheng-Jun Zha[3], Li Su[2,1], Qingming Huang[2,1]
[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, China
[2]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China
[3]School of Information Science and Technology, University of Science and Technology of China, Hefei, China
{xuejing.liu,liang.li}@vipl.ict.ac.cn,wangshuhui@ict.ac.cn,zhazj@ustc.edu.cn,{suli,qmhuang}@ucas.ac.cn

ABSTRACT

Weakly supervised referring expression grounding (REG) aims at localizing the referential entity in an image according to linguistic query, where the mapping between the image region (proposal) and the query is unknown in the training stage. In referring expressions, people usually describe a target entity in terms of its relationship with other contextual entities as well as visual attributes. However, previous weakly supervised REG methods rarely pay attention to the relationship between the entities. In this paper, we propose a knowledge-guided pairwise reconstruction network (KPRN), which models the relationship between the target entity (subject) and contextual entity (object) as well as grounds these two entities. Specifically, we first design a knowledge extraction module to guide the proposal selection of subject and object. The prior knowledge is obtained in a specific form of semantic similarities between each proposal and the subject/object. Second, guided by such knowledge, we design the subject and object attention module to construct the subject-object proposal pairs. The subject attention excludes the unrelated proposals from the candidate proposals. The object attention selects the most suitable proposal as the contextual proposal. Third, we introduce a pairwise attention and an adaptive weighting scheme to learn the correspondence between these proposal pairs and the query. Finally, a pairwise reconstruction module is used to measure the grounding for weakly supervised learning. Extensive experiments on four large-scale datasets show our method outperforms existing state-of-the-art methods by a large margin[1].

CCS CONCEPTS

• **Computing methodologies** → *Matching*; **Learning latent representations**; *Neural networks.*

KEYWORDS

referring expression grounding, weakly supervised learning, attention mechanism, language reconstruction

ACM Reference Format:
Xuejing Liu[1,2], Liang Li[1,*], Shuhui Wang[1], Zheng-Jun Zha[3], Li Su[2,1], Qingming Huang[2,1]. 2019. Knowledge-guided Pairwise Reconstruction Network for Weakly Supervised Referring Expression Grounding. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19), Oct. 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351074

1 INTRODUCTION

Referring expression grounding (REG), also known as phrase localization, has been a surge of interest in both computer vision and natural language processing [16, 22, 27, 34, 44, 46]. Given a query (referring expression) and an image, REG is to find the corresponding location in the image of the referential entity. REG can be widely used in interactive applications, such as robotic navigation [1, 36, 40], visual Q&A [6, 10, 21], or photo editing [5, 39]. Traditionally, training the REG model in a supervised manner requires expensive annotated data that explicitly draw the connection between the input query and its corresponding proposal in the image. Besides, supervised REG models can only handle the grounding with certain categories in the limited training data, which cannot meet the demand for real-world applications. Hence we focus on weakly supervised REG task, where only the image-query pairs are used for training without the mapping information between the query and the proposal.

The weakly supervised REG problem can be formulated as follows. Given an image I, a query q and a set of region proposals $\{r_i\}_{i=1}^{N}$, REG aims at selecting the best-matched region r^* according to the query without the ground-truth pair (q, r_i) . To find the correct mapping between the query and proposal under weakly supervised scenario, Rohrbach *et al.* [34] select the best proposal from a set of candidate proposals through attention mechanism, then reconstruct the input query based on the selected proposal. Chen *et al.* [2] design knowledge aided consistency network, which reconstructs both the input query and proposal's information. Xiao *et al.* [42] generate attention mask to localize linguistic query based on image-phrase pairs and language structure. Zhao *et al.* [50] try to find the location of the referential object by searching over the entire image. All the above methods only exploit the visual attribute features of proposals during grounding and reconstruction.

However, in addition to the visual attribute information, people often describe a target entity in terms of its relationship with other contextual entities, as shown in Fig. 1. Recently, Zhang *et al.* [49]

[1]Code is available at https://github.com/GingL/KPRN.

*Corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351074

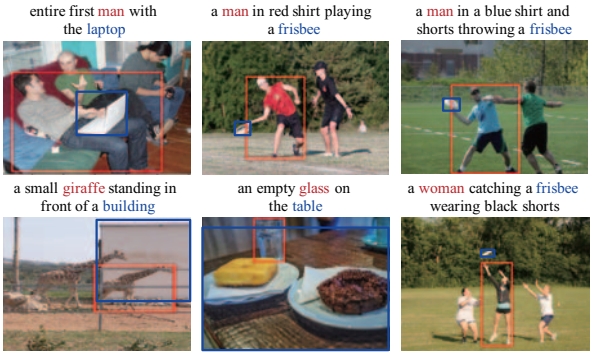

Figure 1: Examples of Referring Expression Grounding (REG). Given a query and an image, REG aims to localize the referential entity. We can observe that besides visual attributes, relationship with other entities is often used to describe the target entity. The target entity and its corresponding proposal are shown in red. The contextual entity and its corresponding proposal are shown in blue.

propose a variational Bayesian method to exploit the reciprocal relation between the referent and context for referring expression grounding. However, their method learns to model the relationship between referent and context based on the annotation of the target entity. Thus the model is not suitable under the weakly supervised setting, where neither the annotations for target nor that for context are available for training.

To address the above challenge, we propose a knowledge-guided pairwise reconstruction network (KPRN) for weakly supervised REG. KPRN learns the mapping between proposal pair (subject, object) and query with the assistance of prior knowledge, and grounds these two entities. KPRN mainly consists of the following three steps.

First, to fulfill the lack of annotations for target entity (subject) and contextual entity (object), we design a knowledge extraction module to guide the proposal selection of subject and object. Specifically, subject and object are first extracted from the input query. The category for each candidate proposal is obtained through Faster R-CNN [33]. Then, both the proposal category and the subject/object are encoded into embeddings. Subsequently, the prior knowledge is obtained in a specific form of semantic similarities between the proposal and the subject/object.

Second, under the guidance of such knowledge, we design the subject and object attention module to construct the subject-object proposal pairs. Subject attention learns to discard the unrelated candidate subject proposals. Object attention learns to select the best proposal as the contextual proposal.

Third, we introduce pairwise attention to learn the matching score, which represents the correspondence between these proposal pairs and the query. Further, we design an adaptive weighting scheme for refining the correspondence based on the spatial relationship in the subject-object proposal pair. As the measurement of weakly supervised grounding, a pairwise reconstruction module is used to reconstruct the input query based on the subject-object proposal pairs and their matching scores.

KPRN can be trained in an end-to-end manner. At the inference stage, KPRN only utilizes the grounding to localize the referent without reconstruction.

The main contributions of this paper are concluded as follows:

- We propose an end-to-end knowledge-guided pairwise reconstruction network, which models the mapping between the input query and proposal pair (subject, object), and grounds the subject and object. We design a knowledge extraction module to introduce the supervision of prior knowledge.
- We design the subject and object attention to compose the subject-object proposal pairs under the guidance of prior knowledge. The subject attention can exclude the unrelated candidate subject proposals, and the object attention can help find the best contextual proposal.
- Through pairwise attention and an adaptive weighting scheme, the matching scores are learned to represent the correspondence between the subject-object proposal pairs and the query. A pairwise reconstruction is used to reconstruct the input query with attentive pairwise proposals.
- Comparison and ablation experiments on the RefCLEF and three MS-COCO datasets show that the proposed KPRN achieves state-of-the-art results in the weakly supervised REG task.

2 RELATED WORK

Supervised Referring Expression Grounding (REG). REG [4, 9, 17, 27, 28, 43, 47, 48] is also known as referring expression comprehension or phrase localization, which is the inverse task of referring expression generation. REG intends to localize the corresponding object described by a free-form natural language query in an image. Given an image I, a query q and a set of region proposals $\{r_i\}_{i=1}^{N}$, REG selects the best-matched region r^* according to the query. Most REG methods can be roughly divided into two kinds. One is CNN-LSTM based encoder-decoder structure to model $P(q|I, r)$ [16, 26, 27, 30, 35, 47]. The other is the joint vision-language embedding framework to model $P(q, r)$. During training, the supervision is object proposal and referring expression pairs (r_i, q_i) [3, 20, 25, 34, 38, 46]. The relationship between the target entity and context entity is often used to assist grounding the target in supervised REG methods [15, 19, 23, 30, 32, 49]. However, these methods learn to model the relationship between target entity and contextual entity based on the annotation of the target entity, which is not available under weakly supervised scenario.

Weakly Supervised Referring Expression Grounding. Weakly supervised REG only has image-level correspondence, and there is no mapping between image regions and referring expressions. To solve this problem, Rohrbach *et al.* [34] propose a framework which learns to ground by reconstructing the given referring expression through attention mechanism. Based on this framework, Chen *et al.* [2] design knowledge aided consistency network, which reconstructs both the input query and proposal's information. Xiao *et al.* [42] ground arbitrary linguistic phrase in the form of spatial attention mask and propose a network with discriminative and structural loss. Different from selecting the optimal region from

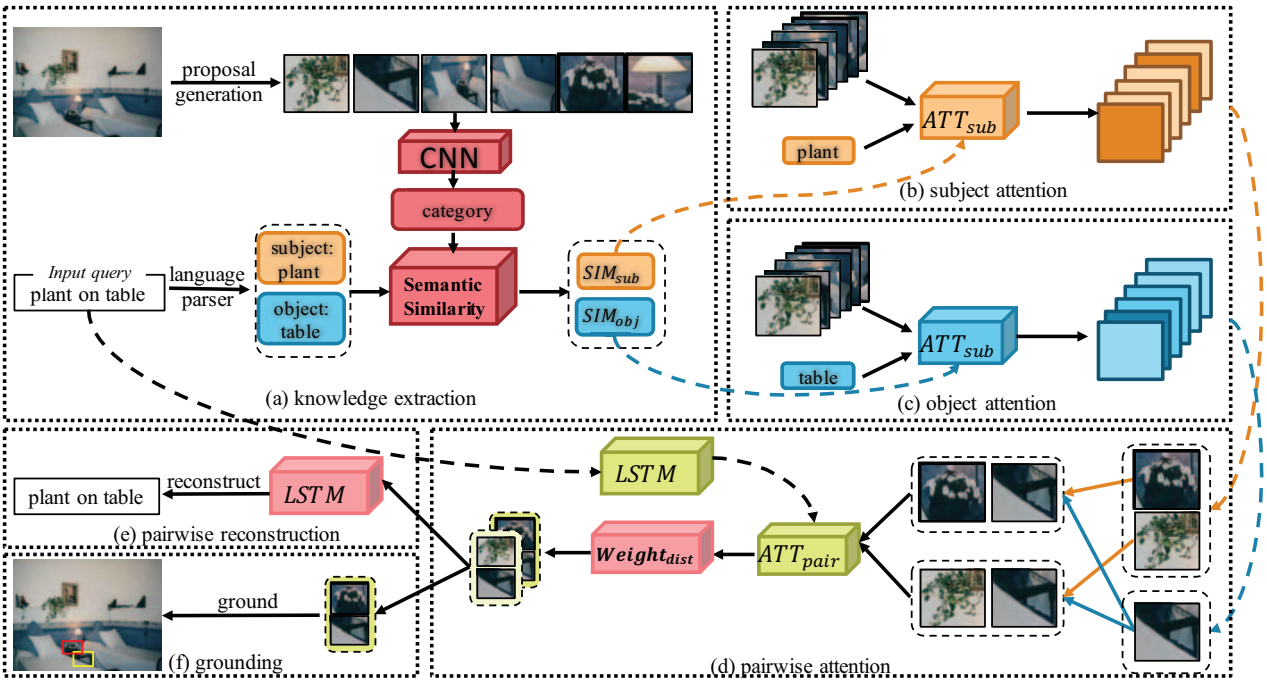

Figure 2: The framework of the proposed KPRN. It consists of (a) knowledge extraction (Section 3.2), (b) subject attention (Section 3.3), (c) object attention (Section 3.3), (d) pairwise attention (Section 3.4) (e) pairwise reconstruction (Section 3.5) and (f) grounding (Section 3.4). The composition of ($a \rightarrow b \rightarrow c \rightarrow d \rightarrow e$) is used for training, while the combination of ($a \rightarrow b \rightarrow c \rightarrow d \rightarrow f$) is used for inference.

a set of region proposals, Zhao *et al.* [50] propose multi-scale anchored transformer network, which can search the entire spatial feature map by taking region proposals as anchors to get a more accurate location.

3 METHODOLOGY

Weakly supervised REG intends to ground the target entity described by the query under the scenario where the region-query correspondence is not available during training. To overcome the lack of supervised information, previous methods usually utilize the selected image regions to reconstruct the corresponding query. Here, we develop our method under such reconstruction mechanism.

Different from previous methods, we reconstruct the input query using subject-object proposal pairs, where subject proposal denotes the target proposal, and object proposal represents the context proposal. Specifically, we propose a knowledge-guided pairwise reconstruction network (KPRN), which learns the mapping between proposal pairs and query with the assistance of prior knowledge, and grounds these two entities. The overall structure is shown in Fig. 2. Initially, through a knowledge extraction module, we introduce the supervision information for the selection of subject/object proposals. The supervision is prior knowledge obtained in a specific form of semantic similarities, representing the uniformity between the proposal category and the subject/object. Secondly, with the supervision information, subject and object attentions are learned to exclude the unrelated proposals from the candidates and compose the proposal pairs. Thirdly, pairwise attention and a weighting scheme are designed to learn the matching score to represent the correspondence between the query and the proposal pairs. Pairwise reconstruction is utilized to measure the grounding under weakly supervised scenario.

In the following, we first introduce the feature encoding of the image and query, and then detail every module of our method.

3.1 Feature Encoding

3.1.1 Visual Features. Given an image with a set of region proposals, obtained by any off-the-shelf proposal generator [51] or object detectors [33], we extract visual features for each proposal by the pre-trained networks and arithmetic operators. Here we use two kinds of visual features, including subject and object features.

Subject feature is the concatenation of CNN features and spatial representations for each proposal. Given an image I and a set of region proposals $\{r_i\}_{i=1}^{N}$, we run the forward propagation of Faster R-CNN based on ResNet [13] for each image I, and crop its C3 and C4 features as the CNN feature $\widetilde{v}_s^i = f_{CNN}(r_i)$ for each proposal. The C3 features represent lower-level features such as colors and shapes while C4 features contain higher-level representations.

Spatial representations consists of absolute position and relative locations with other entities of the same category in the image. Following [46–48], the absolute location feature of each proposal

is decoded as a 5-dim vector $v_l^i = \left[\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w \cdot h}{W \cdot H}\right]$, denoting the top-left, bottom-right position and relative area of the RoI to the whole image. The relative location feature indicates the relative location information between the target proposal and 5 surrounding proposals of the same category. For each surrounding proposal, we calculate its offset and area ratio to the candidate: $\delta v_l^{ij} = \left[\frac{[\Delta x_{tl}]_{ij}}{w_i}, \frac{[\Delta y_{tl}]_{ij}}{h_i}, \frac{[\Delta x_{br}]_{ij}}{w_i}, \frac{[\Delta y_{br}]_{ij}}{h_i}, \frac{w_j h_j}{w_i h_i}\right]$. Then, we concatenate the above absolute and relative location feature as the spatial representations of the proposal, which is a 30-dim vector: $\widetilde{v}_l^i = \left[v_l^i; \delta v_l^i\right]$. Finally, the subject features of each proposal is $v_s^i = [v_s^i; \widetilde{v}_l^i]$

Object feature is extracted for representing the object proposals with CNN and another spatial feature. It is composed of C4 feature $v_{ij} = f_{CNN}(r_j)$ and its relative location feature to subject. The relative location feature is encoded as follows: $\delta m_{ij} = \left[\frac{[\Delta x_{tl}]_{ij}}{w_i}, \frac{[\Delta y_{tl}]_{ij}}{h_i}, \frac{[\Delta x_{br}]_{ij}}{w_i}, \frac{[\Delta y_{br}]_{ij}}{h_i}, \frac{w_j h_j}{w_i h_i}\right]$. The object feature is $v_o^{ij} = \left[v_{ij}; \delta m_{ij}\right]$.

Noting that global visual features are not utilized in our method, as global features might introduce some ambiguity to the grounding task [47].

3.1.2 Referring Expression Features. The language features are extracted through LSTM [14]. Given an query $q = \{w_t\}_{t=1}^T$, first each word in q is one-hot encoded and mapped into a word embedding e_t. Then the word embedding e_t is fed into a bi-directional LSTM. The final representation $h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t]$ is the concatenation of the hidden vectors in both directions.

3.2 Knowledge Extraction

We propose knowledge extraction module to introduce prior knowledge to fulfill the lack of annotations for subject and object. The prior knowledge is in the form of semantic similarity between proposal category and the subject/object, which can guide the selection of subject and object proposals. The details are as follows.

For each proposal, we use the pre-trained Faster R-CNN to predict its category C_i. For the referring expression, we parse it into seven attributes: category name, color, size, absolute location, relative location, relative object, and generic attribute based on [17]. The category name is considered as the subject W_s, and the relative object is viewed as the object W_o for each referring expression.

Next, we utilize an English vocabulary of 72,704 words contained in the GloVe pre-trained word vectors [31] to encode the proposal category(C_i) and subject(W_s)/object(W_o) into a vector. "unk" is used to indicate for the word which is out of the vocabulary. We calculate the cosine distance of the category and subject/object as semantic similarities.

$$\begin{aligned} emb_s &= GloVe(W_s) \\ emb_o &= GloVe(W_o) \\ emb_c &= GloVe(C_i) \\ SIM_s &= cos(emb_c, emb_s) \\ SIM_o &= cos(emb_c, emb_o) \end{aligned} \tag{1}$$

The visualization of the semantic similarity is shown in Fig. 3

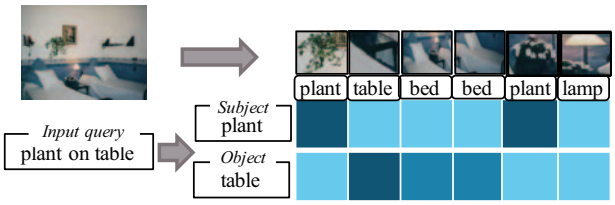

Figure 3: The knowledge extraction module. Subject and object are extracted from the query. Categories are obtained for each proposal. We calculate the semantic similarity as the prior knowledge to guide the training of subject and object attention.

3.3 Subject and Object Attention

Under the guidance of prior knowledge, we design subject and object attention to construct the subject-object proposal pairs. Through subject attention, the candidate proposals with little probability being the target, are assigned to lower weights or even excluded in the next processor. Through object attention, the candidate proposal with the highest probability being the contextual entity is chosen as the object.

This process is shown in Fig. 2 (b) and (c). $\widetilde{v}_s^i$, v_{ij} are the subject and object features extracted from the proposals in the image. emb_s and emb_o represent the embedding of subject and object from the query. Taking the subject as an example, $\widetilde{v}_s^i$ and emb_s are first concatenated into one vector. Then the vector is fed into the subject attention, which is a two layer perceptron, to get the corresponding matching score between proposals and subject. The biases are omitted in Eq. (1).

$$\begin{aligned} \overline{Score}_s^i &= f_{ATT}\left(\widetilde{v}_s^i, emb_s\right) &&= W_2 \phi_{\mathrm{ReLU}}\left(W_1[\widetilde{v}_s^i, emb_s]\right) \\ \overline{Score}_o^{ij} &= f_{ATT}\left(v_{ij}, emb_o\right) &&= W_2 \phi_{\mathrm{ReLU}}\left(W_1[v_{ij}, emb_o]\right) \end{aligned} \tag{2}$$

We normalize the scores using softmax.

$$\begin{aligned} Score_s^i &= \mathrm{softmax}_i\left(\overline{Score}_s^i\right) \\ Score_o^{ij} &= \mathrm{softmax}_i\left(\overline{Score}_o^{ij}\right) \end{aligned} \tag{3}$$

Furthermore, the semantic similarity of subject and object are considered as supervision for the two attention networks. We use Mean Squared Error (MSE) criterion to minimize the distance between the matching score and the semantic similarity.

$$\begin{aligned} Loss_{sub} &= \mathrm{MSE}\left(Score_s, SIM_s\right) \\ Loss_{obj} &= \mathrm{MSE}\left(Score_o, SIM_o\right) \end{aligned} \tag{4}$$

Based on subject attention score, we have two different methods to obtain the candidate subject proposals. One is called as **soft filter**, which assigns different weights to the composed proposal pairs when grounding the query. The other is named as **hard filter**, which discards the unrelated subject proposals if their subject attention score is under the threshold.

Then, we choose the proposal with the maximum object attention score as the object proposal. This is due to the syntax habit that people tend to use the entity with a small number of occurrences in the image as a context to describe the target entity.

Finally, the proposal pairs are constituted based on the above subject proposals and object proposal.

3.4 Pairwise Attention

Different from previous methods, we try to find the best-matched pair $\{q, (r_s, r_o)\}$. q denotes the query, r_s denotes the subject proposal, and r_o denotes the object proposal. For each proposal pair (r_s, r_o), a matching score between it and its corresponding query is learned through pairwise attention and an adaptive weighting scheme. The pair with the maximum matching score will be viewed as the final grounding result.

The procedure of pairwise attention is shown in Fig. 2 (d). v_s^i and v_o^{ij} are the subject and object features of proposals. h_t indicates the language feature extracted from the query through bi-directional LSTM. v_s^i, v_o^{ij} and $h_t$are first concatenated into one vector. Then the vector is fed into the proposal pair attention to get the corresponding matching score between (r_s, r_o) and q. The biases are omitted in Eq. (5).

$$Score_{pair}^{i} = f_{ATT}\left(h_t, v_s^i, v_o^{ij}\right) = W_2 \phi_{\mathrm{ReLU}}\left(W_1[h_t, v_s^i, v_o^{ij}]\right) \quad (5)$$

Further, because people are used to describing the target with the near contextual entity, we designed an adaptive weighting scheme for refining the correspondence based on the spatial relationship in the subject-object proposal pair. The weights are calculated as follows:

$$\omega_{dis}^{ij} = 100/(dist_M^{ij} + 100), \quad (6)$$

where $dist_M$ denotes the Manhattan distance between subject and object proposal in each proposal pair.

The final score S_t^i indicates the matching probability of proposal pair (r_s, r_o) and query q.

If we use soft filter for the selection of subject proposals, the final score is

$$S_t^i = \mathrm{softmax}_i\left(\omega_{dis}^{ij} \times Score_s^i \times Score_{pair}^i\right), \quad (7)$$

where $Score_s^i$ is the subject attention score learned according to prior knowledge.

If using hard filter, we directly discard the candidate pairs (r_s, r_o), whose subject similarity is under the setted threshold. The matching score for the remaining proposal pairs is calculated as follows.

$$S_t^i = \mathrm{softmax}_i\left(\omega_{dis}^{ij} \times Score_{pair}^i\right) \quad (8)$$

3.5 Pairwise Reconstruction

We introduce a pairwise reconstruction to formulate the measurement of the grounding in the weakly supervised training stage, where the attentive proposal pairs are exploited to reconstruct the input query.

First, the concatenated subject and object features of the proposal pair (r_s, r_o) are fed into a one-layer perceptron.

$$r_{vis}^i = \phi_{\mathrm{ReLU}}\left(W_v(\left[v_s^i, v_o^{ij}\right]) + b_v\right) \quad (9)$$

Then we obtain the fused features f_{vis} according to the final matching score S_t^i.

$$f_{vis} = \sum_{i=1}^{N} S_t^i r_{vis}^i \quad (10)$$

Finally, inspired by the query generation methods [7, 37], we reconstruct the input query through LSTM.

$$P(q|f_{vis}) = f_{\mathrm{LSTM}}(f_{vis}) \quad (11)$$

The language reconstruction network aims to maximize the likelihood of the ground-truth query $\hat{q}$ generated by LSTM,

$$Loss_{lan} = -\frac{1}{B}\sum_{b=1}^{B} \log(P(\hat{q}|f_{alan})) \quad (12)$$

where B is the batch size.

3.6 Attribute Classification Loss

As mentioned in previous methods [41, 45, 46], attribute information is important on distinguish object of the same category. Thus we also add an attribute classification branch in our model. The attribute label is extracted through an external language parser [17] according to [46]. Subject feature $\widetilde{r}_s^i$ of proposal is used for attribute classification. As each query has multiple attribute labels, we use the binary cross-entropy loss for the multi-label classification.

$$Loss_{att} = f_{\mathrm{BCE}}\left(y_{ij}, p_{ij}\right) \quad (13)$$

We use the reciprocal of the frequency that attribute labels appears as weights in this loss to ease unbalanced data.

3.7 Network Training and Inference

The network is trained with an end-to-end strategy. During training, only query with attribute words goes through attribute classification branch. At inference, the reconstruction module is not needed anymore. We feed the image and query into the network and get the most related proposal pair whose final score is the maximal in the grounding module.

$$j = \arg\max_s f\left(p, (r_s, r_o)\right) \quad (14)$$

The final loss function is:

$$Loss = Loss_{sub} + Loss_{obj} + Loss_{lan} + Loss_{att} \quad (15)$$

4 EXPERIMENTS

4.1 Datasets

We evaluate our method on four popular benchmarks of referring expression grounding.

RefCOCO [47]. It is also called UNC RefExp. The dataset contains 142,209 queries for 50,000 objects in 19,994 images from MSCOCO [24]. It is collected through ReferitGame [17]. The dataset is split into train, validation, Test A, and Test B, which has 16,994, 1,500, 750 and 750 images, respectively. Test A contains multiple people while Test B contains multiple objects. Each image contains at least 2 objects of the same object category. The average length of the queries in this dataset is 3.61.

RefCOCO+ [47]. It has 141,564 queries for 49,856 referents in 19,992 images from MSCOCO [24]. It is also collected through ReferitGame [17]. Different from RefCOCO, the queries in this dataset are disallowed to use locations to describe the referents. Thus, this dataset focus on the appearance of the referents. The split is 16,992, 1,500, 750 and 750 images for train, validation, Test A, and Test B respectively. Each image contains 2 or more objects

Table 1: Accuracy (IoU > 0.5) on RefCOCO, RefCOCO+ and RefCOCOg. Bond: best result. Blue: best result of VC.

Methods	Settings	RefCOCO			RefCOCO+			RefCOCOg
		val	testA	testB	val	testA	testB	val
VC	w/o reg	-	13.59	21.65	-	18.79	24.14	25.14
VC	-	-	17.34	20.98	-	23.24	24.91	33.79
VC	w/o α	-	33.29	30.13	-	34.60	31.58	30.26
VC (det)	w/o reg	-	17.14	22.30	-	19.74	24.05	28.14
VC (det)	-	-	20.91	21.77	-	25.79	25.54	33.66
VC (det)	w/o α	-	32.68	27.22	-	34.68	28.10	29.65
KPRN	soft	34.43	33.82	35.45	35.96	35.24	36.96	33.56
KPRN	soft + attr	**36.34**	**35.28**	**37.72**	**37.16**	**36.06**	**39.29**	36.65
KPRN	hard	35.04	34.74	36.53	35.10	32.75	36.76	35.44
KPRN	hard +attr	34.93	33.76	36.98	35.31	33.46	37.27	**38.37**

of the same object category in this dataset. The average length of the queries in this dataset is 3.53.

RefCOCOg [27]. It is also called Google Refexp. It has 95,010 queries for 49,822 objects in 25,799 images from MSCOCO [24]. This dataset is collected in a non-interactive setting on Amazon Mechanical Turk. Thus It has longer queries containing appearance and location to describe the referents. The split is 21,149 and 4,650 images for training and validation. There is no open test split for RefCOCOg. Images were selected to contain between 2 and 4 objects of the same category. The average length of the queries is 8.43.

RefCLEF [17]. It is also called ReferIt. It contains 20,000 annotated images from IAPR TC-12 dataset [11] and SAIAPR-12 dataset [8]. The dataset includes some ambiguous queries, such as anywhere. It also has some mistakenly annotated image regions. The dataset is split into 9,000, 1,000 and 10,000 images for training, validation and test for fair comparison with [34]. 100 bounding box proposals [16] are provided for each image using Edge Boxes [51]. Images contain between 2 and 4 objects of the same object category. The maximum length of all the queries is 19 words.

4.2 Experimental Setup

4.2.1 Implementation details. We train KPRN through Adam [18] with an initial learning rate 4e-4, which drops by 10 after every 8,000 iterations. The training iterations are 30,000 with a batch size of a single image. Each image has an indefinite number of annotated queries. The rectified linear unit (ReLU) [29] is used as the non-linear activation function. Batch normalization operations are not used in our framework. ResNet is the main feature extractor for RoI visual features. We adopt EdgeBoxe [51] to generate 100 region proposals for RefCLEF dataset for fair comparison with [2, 34].

4.2.2 Metrics. The Intersection over Union (IoU) between the selected region and the ground-truth are calculated to evaluate the network performance. If the IoU score is greater than 0.5, the predicted region is considered as the right grounding.

4.3 Results on RefCOCO, RefCOCO+ and RefCOCOg datasets

4.3.1 Performance Analysis: We compared the proposed KPRN with the only published unsupervised results on these datasets [49]. "soft" denotes that we use soft filter for the selection of subject proposals. "hard" denotes that we discard the candidate subject proposals whose attention score is under the threshold. "attr" denotes that whether to use attribute classification loss in the model.

Table 1 reports the comparison results on RefCOCO, RefCOCO+ and RefCOCOg datasets. We can have the following findings. First, our method performs better on testB, which can outperform VC by a large margin on RefCOCO and RefCOCO+ datasets. While on testA, the promotion is less. The difference between testA and testB is that testA contains multiple people and testB contains multiple other objects. Second, the KPRN with soft filter achieves better results on RefCOCO and RefCOCO+ datasets, while KPRN with hard filter performs better on RefCOCOg dataset. Third, the attribute classification loss can improve the performance by about 2% on average on these datasets.

4.3.2 Ablation Study: We study the benefits of each module by running ablation experiments. Table 2 reports the results on RefCOCO, RefCOCO+ and RefCOCOg datasets with different settings. "attr", "loc", "obj", "soft", "hard" and "dist" denotes the visual attribute features, location features, context features, soft filter for subject attention, hard filter for subject attention and weighting scheme for spatial distance, respectively.

Some observation can be obtained as follows. First, location features play a critical role on RefCOCO dataset. On RefCOCO+ dataset, the performance does not increase significantly with location features, that is probably because this dataset is disallowed to use locations to describe the referents. Second, the performance increases by a large margin on RefCOCO+ dataset when introducing the object features of proposal. This indicates the effectiveness of proposed object attention. Third, subject attention can increases the performance of RefCOCOg dataset. Fourth, the adaptive weighting scheme for spatial distance is also helpful for the grounding results.

Table 2: Ablation study on RefCOCO, RefCOCO+ and RefCOCOg datasets.

Methods	RefCOCO			RefCOCO+			RefCOCOg
	val	testA	testB	val	testA	testB	val
KPRN (attr)	15.58	8.34	25.57	16.25	14.15	18.41	33.64
KPRN (attr+loc)	36.47	35.11	37.74	20.33	17.34	24.38	36.13
KPRN (attr+loc+obj)	36.73	35.60	37.11	36.49	35.92	38.41	37.17
KPRN (attr+loc+obj+hard)	35.31	33.87	36.17	34.78	33.34	37.43	40.24
KPRN (attr+loc+obj+soft)	35.28	35.41	36.60	35.36	35.16	36.29	38.45
KPRN (attr+loc+obj+hard+dist)	35.98	35.28	37.39	35.66	33.67	38.52	43.16
KPRN (attr+loc+obj+soft+dist)	36.34	35.28	37.72	37.16	36.06	39.29	36.65

Table 3: Albation study on different hard filter threshold for RefCOCO, RefCOCO+ and RefCOCOg datasets.

expID	thr	RefCOCO			RefCOCO+			RefCOCOg
		val	testA	testB	val	testA	testB	val
exp1	0.05	35.92	35.28	36.76	36.60	34.65	38.35	36.74
exp2	0.10	34.93	33.76	36.98	35.31	33.46	37.27	38.37
exp3	0.15	33.96	32.31	36.78	32.56	28.94	36.51	37.35
exp4	0.20	34.68	33.07	37.90	32.60	29.60	36.53	39.26
exp5	0.25	34.89	33.83	37.45	32.66	30.53	36.74	41.13
exp6	0.30	35.53	35.23	37.10	33.71	32.19	36.74	42.03
exp7	0.35	35.98	35.28	37.39	35.66	33.67	38.52	43.16
exp8	0.40	35.24	34.17	37.84	35.84	34.13	39.58	40.93

Table 4: Accuracy (IoU > 0.5) on RefCLEF dataset.

Method	IoU
LRCN [7]	8.59
Caffe-7K [12]	10.38
GroundeR [34]	10.70
MATN [50]	13.61
VC [49]	14.11
VC w/o α [49]	14.50
KAC Net [2]	15.83
KPRN (attr+loc+obj)	20.99
KPRN (attr+loc+obj+soft)	18.35
KPRN (attr+loc+obj+hard)	32.32
KPRN (attr+loc+obj+hard+dist)	**33.87**

Table 3 represents results of hard filter for the selection of subject proposals with different filter threshold. The performance increases with the increase of the threshold on RefCOCOg, while the threshold has weak impact on RefCOCO and RefCOCO+ datasets.

4.3.3 Qualitative Results. Fig. 4 shows qualitative example predictions on RefCOCO, RefCOCO+ and RefCOCOg datasets. The query is shown under the corresponding image. The ground truth, grounding proposal and context proposal are denoted as solid white, dashed red and dashed blue, respectively.

Table 5: Ablation study on different hard filter threshold for RefCLEF dataset.

expID	exp1	exp2	exp3	exp4	exp5	exp6
thr	0.1	0.2	0.3	0.4	0.5	0.6
IoU	19.48	23.32	33.17	33.87	35.98	35.78

4.4 Results on RefCLEF Dataset

4.4.1 Performance Analysis: We compare our KPRN with state-of-the-art weakly supervised referring expression grounding methods. The LRCN use the image captioning to score how likely the query phrase is to be generated for the proposal box. CAFFE-7K predicts a class for each proposal box and then compared to the query phrase after both are projected to a joint vector space. Other methods are all introduced in the related work.

Table 4 reports the results on RefCLEF dataset. We can have the following observations. First, the proposed KPRN outperforms state-of-the-art result with a large margin of 18.04%. Second, by introducing extra location features and object proposal features, our method improves the IoU by 5.16%. Third, when using soft filter for subject proposals, the performance degrades for grounding. When using the hard filter, the performance obtains an improvement of 12.88%. This is probably because there are too many candidate proposals in this dataset. With the power of subject attention, many unrelated proposals can be excluded. Thus the number of candidate proposal pairs for grounding can be decreased by a large margin. Fourth, Through adding the adaptive weighting scheme, the performance of KPRN gets a further promotion of 1.55%, which benefits from the modeling of spatial relationship in the subject-object proposal pair.

4.4.2 Ablation Study: We also verify the impact of the threshold for hard filter. The results in Table 5 demonstrate that the IoU gets the best result when the threshold is 0.5.

5 CONCLUSION

To address the weakly supervised REG, we propose a knowledge-guided pairwise reconstruction network (KPRN). The KPRN can model the relationship between subject and object as well as ground them under the guidance of prior knowledge. The prior knowledge is obtained in the form of semantic similarities between each proposal and the subject/object. Then, we design the subject and object

Figure 4: Qualitative results on RefCOCO, RefCOCO+ and RefCOCOg datasets. The denotations of the bounding box colors are as follows. Solid white: ground truth; dashed red: predicted subject proposal; dashed blue: predicted object proposal.

attention to construct the subject-object proposal pairs with the supervision of such knowledge. Further, Pairwise attention and a weighting scheme are used to learn the final matching score between proposal pairs and query. Finally, pairwise reconstruction measures the grounding performance under weakly supervised setting. Experiments demonstrate that the proposed method provides a significant improvement of performance on four datasets.

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China: 61771457, 61732007, 61772494, 61672497, 61622211, 61836002, 61472389, 61620106009 and U1636214, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013, and Fundamental Research Funds for the Central Universities under Grant WK2100100030.

REFERENCES

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In *CVPR*. IEEE Computer Society, 3674–3683.

[2] Kan Chen, Jiyang Gao, and Ram Nevatia. 2018. Knowledge Aided Consistency for Weakly Supervised Phrase Grounding. In *CVPR*. IEEE Computer Society, 4042–4050.

[3] Kan Chen, Rama Kovvuri, and Ram Nevatia. 2017. Query-Guided Regression Network with Context Policy for Phrase Grounding. In *ICCV*. IEEE Computer Society, 824–832.

[4] Xinpeng Chen, Lin Ma, Jingyuan Chen, Zequn Jie, Wei Liu, and Jiebo Luo. 2018. Real-Time Referring Expression Comprehension by Single-Stage Grounding Network. *CoRR* abs/1812.03426 (2018).

[5] Ming-Ming Cheng, Shuai Zheng, Wen-Yan Lin, Vibhav Vineet, Paul Sturgess, Nigel Crook, Niloy J. Mitra, and Philip H. S. Torr. 2014. ImageSpirit: Verbal Guided Image Parsing. *ACM Trans. Graph.* 34, 1 (2014), 3:1–3:11.

[6] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied Question Answering. In *CVPR*. IEEE Computer Society, 1–10.

[7] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*. IEEE Computer Society, 2625–2634.

[8] Hugo Jair Escalante, Carlos A. Hernández, Jesús A. González, Aurelio López-López, Manuel Montes-y-Gómez, Eduardo F. Morales, Luis Enrique Sucar, Luis Villaseñor Pineda, and Michael Grubinger. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding* 114, 4 (2010), 419–428.

[9] Nicholas FitzGerald, Yoav Artzi, and Luke S. Zettlemoyer. 2013. Learning Distributions over Logical Forms for Referring Expression Generation. In *EMNLP*. ACL, 1914–1925.

[10] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. IQA: Visual Question Answering in Interactive Environments. In *CVPR*. IEEE Computer Society, 4089–4098.

[11] Michael Grubinger, Paul Clough, Henning MÃijller, and Thomas Deselaers. 2006. The IAPR TC12 Benchmark: A New Evaluation Resource for Visual Information Systems. *Workshop Ontoimage* (10 2006).

[12] Sergio Guadarrama, Erik Rodner, Kate Saenko, Ning Zhang, Ryan Farrell, Jeff Donahue, and Trevor Darrell. 2014. Open-vocabulary Object Retrieval. In *Robotics: Science and Systems*.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE Computer Society, 770–778.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[15] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling Relationships in Referential Expressions with Compositional Modular Networks. In *CVPR*. IEEE Computer Society, 4418–4427.

[16] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural Language Object Retrieval. In *CVPR*. IEEE Computer Society, 4555–4564.

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*. ACL, 787–798.

[18] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).

[19] Ranjay Krishna, Ines Chami, Michael S. Bernstein, and Li Fei-Fei. 2018. Referring Relationships. In *CVPR*. IEEE Computer Society, 6867–6876.

[20] Liang Li, Shuqiang Jiang, and Qingming Huang. 2012. Learning Hierarchical Semantic Description Via Mixed-Norm Regularization for Image Understanding. *IEEE Trans. Multimedia* 14, 5 (2012), 1401–1413.

[21] Liang Li, Shuqiang Jiang, Zheng-Jun Zha, Zhipeng Wu, and Qingming Huang. 2013. Partial-Duplicate Image Retrieval via Saliency-Guided Visual Matching. *IEEE MultiMedia* 20, 3 (2013), 13–23.

[22] Liang Li, Shuhui Wang, Shuqiang Jiang, and Qingming Huang. 2018. Attentive Recurrent Neural Network for Weak-supervised Multi-label Image Classification. In *ACM Multimedia*. 1092–1100.

[23] Liang Li, Chenggang Clarence Yan, Xing Chen, Chunjie Zhang, Jian Yin, Baochen Jiang, and Qingming Huang. 2016. Distributed image understanding with semantic dictionary and semantic expansion. *Neurocomputing* 174 (2016), 384–392.

[24] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV (5) (Lecture Notes in Computer Science)*, Vol. 8693. Springer, 740–755.

[25] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. 2017. Referring Expression Generation and Comprehension via Attributes. In *ICCV*. IEEE Computer Society, 4866–4874.

[26] Ruotian Luo and Gregory Shakhnarovich. 2017. Comprehension-Guided Referring Expressions. In *CVPR*. IEEE Computer Society, 3125–3134.

[27] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *CVPR*. IEEE Computer Society, 11–20.

[28] Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2013. Generating Expressions that Refer to Visible Objects. In *HLT-NAACL*. The Association for Computational Linguistics, 1174–1184.

[29] Guido F. Montúfar, Razvan Pascanu, KyungHyun Cho, and Yoshua Bengio. 2014. On the Number of Linear Regions of Deep Neural Networks. In *NIPS*. 2924–2932.

[30] Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. 2016. Modeling Context Between Objects for Referring Expression Understanding. In *ECCV (4) (Lecture Notes in Computer Science)*, Vol. 9908. Springer, 792–807.

[31] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*. ACL, 1532–1543.

[32] Bryan A. Plummer, Arun Mallya, Christopher M. Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Phrase Localization and Visual Relationship Detection with Comprehensive Image-Language Cues. In *ICCV*. IEEE Computer Society, 1946–1955.

[33] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*. 91–99.

[34] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of Textual Phrases in Images by Reconstruction. In *ECCV (1) (Lecture Notes in Computer Science)*, Vol. 9905. Springer, 817–834.

[35] Guoli Song, Shuhui Wang, Qingming Huang, and Qi Tian. 2017. Multimodal Similarity Gaussian Process Latent Variable Model. *IEEE Trans. Image Processing* 26, 9 (2017), 4168–4181.

[36] Jesse Thomason, Jivko Sinapov, and Raymond J. Mooney. 2017. Guiding Interaction Behaviors for Multi-modal Grounded Language Learning. In *RoboNLP@ACL*. Association for Computational Linguistics, 20–24.

[37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*. IEEE Computer Society, 3156–3164.

[38] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *CVPR*. IEEE Computer Society, 5005–5013.

[39] Shuhui Wang, Yangyu Chen, Junbao Zhuo, Qingming Huang, and Qi Tian. 2018. Joint Global and Co-Attentive Representation Learning for Image-Sentence Retrieval. In *ACM Multimedia*. 1398–1406.

[40] Shuhui Wang, Shuqiang Jiang, Qingming Huang, and Qi Tian. 2012. Multi-feature metric learning with knowledge transfer among semantics and social tagging. In *CVPR*. IEEE Computer Society, 2240–2247.

[41] Qi Wu, Chunhua Shen, Peng Wang, Anthony R. Dick, and Anton van den Hengel. 2018. Image Captioning and Visual Question Answering Based on Attributes and External Knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 6 (2018), 1367–1381.

[42] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. 2017. Weakly-Supervised Visual Grounding of Phrases with Linguistic Structures. In *CVPR*. IEEE Computer Society, 5253–5262.

[43] Shijie Yang, Liang Li, Shuhui Wang, Weigang Zhang, and Qingming Huang. 2017. A Graph Regularized Deep Neural Network for Unsupervised Image Representation Learning. In *CVPR*. IEEE Computer Society, 7053–7061.

[44] Shijie Yang, Liang Li, Shuhui Wang, Weigang Zhang, Qingming Huang, and Qi Tian. 2019. SkeletonNet: A Hybrid Network with a Skeleton-Embedding Process for Multi-view Image Representation Learning. *IEEE Transactions on Multimedia* (2019).

[45] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting Image Captioning with Attributes. In *ICCV*. IEEE Computer Society, 4904–4912.

[46] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. 2018. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *CVPR*. IEEE Computer Society, 1307–1315.

[47] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling Context in Referring Expressions. In *ECCV (2) (Lecture Notes in Computer Science)*, Vol. 9906. Springer, 69–85.

[48] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. 2017. A Joint Speaker-Listener-Reinforcer Model for Referring Expressions. In *CVPR*. IEEE Computer Society, 3521–3529.

[49] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. 2018. Grounding Referring Expressions in Images by Variational Context. In *CVPR*. IEEE Computer Society, 4158–4166.

[50] Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. 2018. Weakly Supervised Phrase Localization With Multi-Scale Anchored Transformer Network. In *CVPR*. IEEE Computer Society, 5696–5705.

[51] C. Lawrence Zitnick and Piotr Dollár. 2014. Edge Boxes: Locating Object Proposals from Edges. In *ECCV (5) (Lecture Notes in Computer Science)*, Vol. 8693. Springer, 391–405.