Confidence-aware Pseudo-label Learning for Weakly Supervised Visual Grounding

Yang Liu[1,2*] Jiahua Zhang[1] Qingchao Chen[3] Yuxin Peng[1]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]National Key Laboratory of General Artificial Intelligence, BIGAI
[3]National Institute of Health Data Science, Peking University
{yangliu, qingchao.chen, pengyuxin}@pku.edu.cn {zhangjiahua}@stu.pku.edu.cn

Abstract

Visual grounding aims at localizing the target object in image which is most related to the given free-form natural language query. As labeling the position of target object is labor-intensive, the weakly supervised methods, where only image-sentence annotations are required during model training have recently received increasing attention. Most of the existing weakly-supervised methods first generate region proposals via pre-trained object detectors and then employ either cross-modal similarity score or reconstruction loss as the criteria to select proposal from them. However, due to the cross-modal heterogeneous gap, these method often suffer from high confidence spurious association and model prone to error propagation. In this paper, we propose Confidence-aware Pseudo-label Learning (CPL) to overcome the above limitations. Specifically, we first adopt both the uni-modal and cross-modal pre-trained models and propose conditional prompt engineering to automatically generate multiple 'descriptive, realistic and diverse' pseudo language queries for each region proposal, and then establish reliable cross-modal association for model training based on the uni-modal similarity score (between pseudo and real text queries). Secondly, we propose a confidence-aware pseudo label verification module which reduces the amount of noise encountered in the training process and the risk of error propagation. Experiments on five widely used datasets validate the efficacy of our proposed components and demonstrate state-of-the-art performance. Code can be found at https://github.com/zjh31/CPL.git

1. Introduction

Visual grounding is an important task with vast potential applications in visual question answering [1], robot manipulation [38, 51], etc. The goal is to find the target object (region) in an image associated with a given free-form natural language query. Fully supervised visual grounding [6, 42, 43, 46, 48, 15, 21] has witnessed remarkable progress recently. However, accurate box annotations for each target object are unfortunately expensive to obtain and thus difficult to scale. Therefore the weakly supervised setting, where only image-level descriptions are available during training, is more practical and draws increasing attention from the community.

*Corresponding author

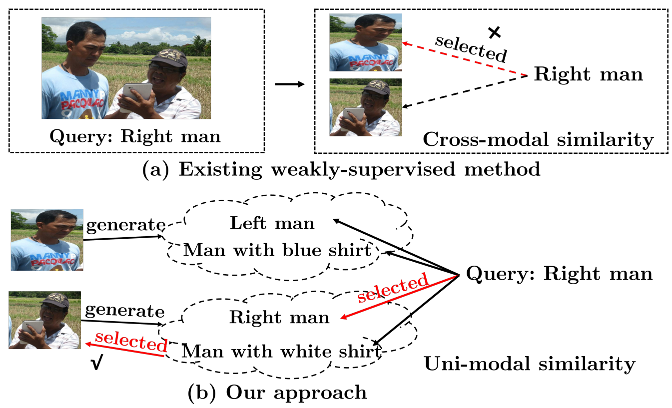


Figure 1. Our method compare with other weakly supervised visual grounding methods. (a) Existing weakly supervised methods. (b) Our approach.

Most existing weakly supervised solutions generate region proposals via pre-trained object detectors and then employ either the contrastive learning-based or reconstruction-based paradigms to select from them. As shown in Figure 1(a), the proposal selection is conducted based on the *cross-modal (region-textual)* (directly compute the matching score between the proposal and query). Specifically, contrastive learning-based methods learn the cross-modal alignment in the image level by maximizing the matching scores of the image and the paired descriptions while suppressing that of the unpaired ones. Reconstruction-based methods perform the proposal selection with the *cross-modal reconstruction loss*, assuming that the proposals that match the text should best reconstruct the entire query.

However, both paradigms have the following two limitations. Firstly, due to the heterogeneous gap between the high-level concepts of text descriptions and the pixel-level contents of the image region, using *cross-modal* matching score or reconstruction query directly for proposal selection is not reliable. Such matching ambiguity often misleads the grounding model to learn spurious association, which greatly hinders the grounding performance. Secondly, existing approaches are trapped by the error propagation and accumulation as they neglect the confidence of the learned cross-modal association and unavoidably keep overfitting to some incorrect ones encountered during the model training. A recent work [13] proposes generating pseudo queries for proposals in an unsupervised method and using them for training a grounding model directly. However, it can only generate short and unreliable descriptions with limited style and structure based on hand-crafted templates.

To address the above limitations, we introduce a novel weakly supervised method for visual grounding by using more reliable *uni-modal* matching for proposal selection and perform association verification before leveraging them in model training. We call it Confidence-aware Pseudo-label Learning (CPL). Firstly, to establish more reliable region-text association for model training, we propose to use three complementary pipeline to automatically generate multiple 'descriptive, realistic and diverse' pseudo language queries for each region proposal and form $<Region - PseudoQuery>$ pairs. As shown in Figure 1(b), our method then perform proposal selection based on the uni-modal similarity score (between real query and pseudo queries) and form $<Region - RealQuery>$ pairs. All region-query pairs are used to train a fully-supervised grounding model. To reduce the contribution of error region-query pairs, we propose an confidence-aware cross-modal verification module that estimates the confidence score of the region-query associations. We propose a selective grounding loss based on the confidence score to rebalance the weight of each sample in the training process.

To sum up, the main contributions of our work are:

- In contrast to performing proposal selection based on cross-modal matching scores, we propose to generate multiple 'descriptive, realistic and diverse' pseudo language queries for each region proposal, and then establish more reliable cross-modal association for model training based on the uni-modal similarity (between pseudo and real text queries).

- We propose a confidence-aware cross-modal verification module and selective grounding loss to suppress the contribution of spurious association, which reduces the risk of error propagation in the training process.

- Experiments on the RefCOCO [47], RefCOCO+ [47], RefCOCOg [25], ReferItGame [14] and Fliker30K Entities[28] datasets demonstrate the effectiveness of our method in weakly supervised visual grounding.

2. Related Work

2.1. Fully supervised Visual Grounding

Recent advances in visual grounding can be roughly divided into two categories, including two-stage methods [10, 11, 21, 37, 38, 41, 46, 52, 49] and one-stage methods [43, 3, 20, 42, 12]. Two-stage approaches generate a set of candidate objects from images by leveraging uni-modal pretrained models (i.e., off-the-shelf detectors [49]) in the first stage, then compute the matching scores between the candidate objects and referring expression and select the top-ranked one. One-stage methods localize referred objects without generating object proposals in advance. Instead of generating proposals, the visual feature is densely fused with the text feature, and the language-fused feature map is further leveraged to predict the final bounding box. Recently, transformer-based methods [6, 40] achieve remarkable results. Transformer-based methods take the visual and linguistic feature tokens as inputs, then input them into a set of transformer encoder layers to perform cross-modal fusion and predict the target region directly. However, fully supervised methods need laborious manual annotation of target object bounding box in model training thus limiting its scalability and practicability.

2.2. Weakly-supervised Visual Grounding

Different from fully supervised methods, the weakly-supervised aims to learn region-query correspondence with only image-query pairs. Most works employ contrastive learning [36, 9] and reconstruction strategies [22, 23, 33, 31, 2, 24] for the weakly-supervised visual grounding task.

The reconstruction strategies usually generate a set of region proposals from an image with an external object detector, and reconstruct the entire query with the selected proposal. Contrastive learning strategy maximize compatibility of the attention-weighted regions and the query in the corresponding caption, compared to non-corresponding pairs of images and expression. However, all paradigms ignore the heterogeneous gap between the textual descriptions and image regions, and these methods implicitly align language and visual space in the training process, which makes cross-modal matching scores or proposal reconstruction quality unreliable. Besides, these methods do not take the the problem of error-propagation into account because some queries do not have corresponding proposals due to limitations in the number and quality of proposals.

Recently, Pseudo-Q [13] proposes a novel unsupervised method which produces pseudo region-query pairs based on rule-based template for supervised training, in which pseudo query is less realistic. However, Pseudo-Q ignores the distribution shift between the pseudo and real queries and also does not take the problem of incorrect queries which harms final performance. Different from it, we propose three complementary pipeline to generate 'descriptive, realistic and diverse' pseudo language query for each region proposal and a confidence-aware pseudo label verification module to surpass the contribution of error association in the training process.

2.3. Pre-trained Models

Uni-modal pre-trained models have witnessed remarkable progress in vision understanding and natural language understanding tasks. Most of the existing visual grounding methods leverage the Uni-modal pre-trained models (e.g., off-the-shelf detectors [30, 26], sentence encoders [7]). However, in principle, the uni-modal pretraining is sub-optimal for visual grounding tasks as it requires cross-modal region-text semantic alignment.

Vision and language cross-modal pre-training [34, 4, 19, 29, 18, 17, 39] aims to learn multi-modal representations from large-scale image-text pairs to improve downstream vision and language tasks. CLIP [29] uses a separate image and text transformer and a contrastive pre-training objective. BLIP [17] establishes a unified understanding and generation of multi-modal models based on transformers. However, most existing cross-modal models are pretrained from image-text pair without any box-wise region-text pair annotation, thus lacking region-level grounding capability. Recently, a zero-shot approach ReCLIP [32] utilizes the discriminative capability of the cross-modal pre-trained model and simple rules with respect to spatial relation for visual grounding. However, the proposal selection still suffers from the spurious association due to the cross-modal heterogeneous gap. In contrast, to the best of our knowledge, we are the first to utilize both the discriminative and generative capability of the pre-trained model for visual grounding. We propose conditional prompt learning to obtain the object-centric and relation-aware region-level pseudo queries and then perform proposal selection based on the uni-modal similarity score. We also propose a confidence-aware pseudo-label verification module to reduce the risk of error propagation.

3. Method

3.1. Problem Formulation

Given a paired image and natural language query $\{\text{I, t}\}$, by using the detectors to extract some salient regions as the proposals, our objective is to find the target region (object) in image I that is most aligned with query t in semantic. Although image-query pairs are available in training, there is no access to the ground-truth box annotations for the target object. We propose a Confidence-aware Pseudo-label Learning (CPL) framework for this task, as shown in Figure 2. It consists of four main stages: Pseudo-Query Generation, Uni-modal real query propagation, Cross-modal verification and Grounding model training. We discuss each of these stages and their interactions in the following.

3.2. Pseudo-Query Generation

In this section, the ultimate goal is to form multiple 'descriptive, realistic and diverse' high-quality $<Region - PseudoQuery>$ pairs, which can be safely leveraged in later grounding model training. 'Descriptive' means that the query is highly correlated with the image to avoid errors; 'diversity' means that the generated text is as different as possible to increase the robustness of the model; 'realistic' means that the generated query is as syntactic as possible, so as to be closer to the real query and avoid distribution drift. Therefore, we propose three complementary pipelines to generate multiple 'descriptive, realistic and diverse' plausible pseudo language queries for each region proposal. As shown in Figure 2, the p_{ij} represents j-th pseudo query generated by i-th proposal. The three pseudo-query Generation pipelines are described as follows.

(1) Heuristic+ pipeline

A recent work [13] first proposes to generate pseudo queries for training the grounding model directly. However, it can only generate short descriptions with limited style and structure and also neglects the distribution shift between the pseudo-queries and the real queries. To address the above limitations, we propose **the first pipeline: Heuristic+**, which consist of a series of technical improvement to the model of [13].

Specifically, for **Nouns**, to minimize the influence of the pseudo and real queries distribution shift, different from [13] that select top-N objects with the highest confidence score of the off-the-shelf object detector, we propose to remove the candidate regions (outliers) which the semantic is far away from the vocabulary in the real queries. For **Attributes**, to make pseudo query *more descriptive*, different from [13] that neglect tiny object, we treat some tiny object o_i as the attribute of the bigger object o_j if the ratio between the area of the intersection of boxes i, j compared to the area of box i is above a threshold. For example, we assign "black hair" as attributes for the left person in Fig 2. For **Spatial Relationship**, we observe that there are around 80% images containing more than two instances from the same category; different from [13] describing a simple pair-wise relationship, we add some compound words (*e.g, left top, right bottom*), ordinal numbers (*e.g, leftmost, second right*) by comparing relative coordi-

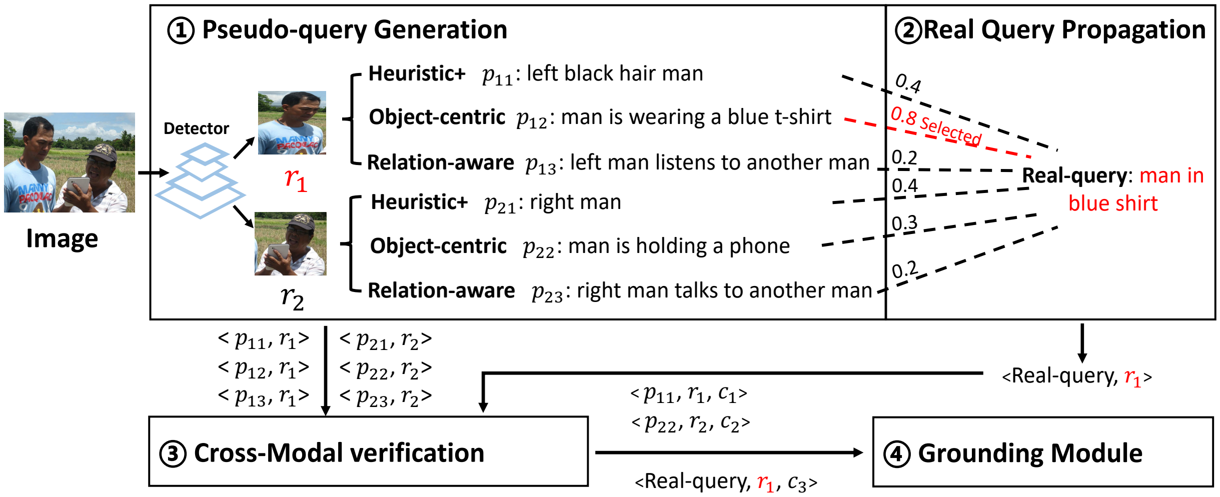
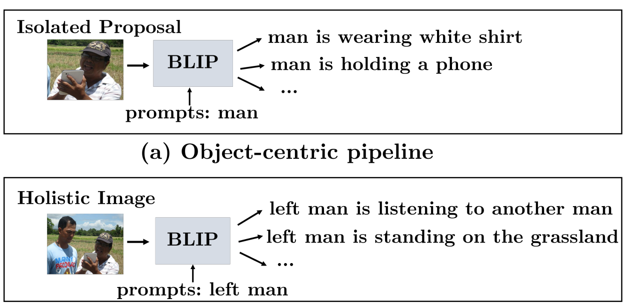


Figure 2. Overview of our CPL method. Our approach consists of a pseudo query generation module, a uni-modal real query propagation module, a cross-modal verification module and a grounding module. The pseudo query generation module generates multiple $<Region-PseudoQuery>$ associations through three pipelines for each proposal, the p_{ij} represents the j-th pseudo query generated by proposal r_i. Then the uni-modal real query propagation selects proposal based on the uni-modal similarity between pseudo and real query, and establish $<Region-RealQuery>$ association. Cross-modal verification module calculate the confident score c_i of all region-query association before leveraging them to train the grounding module. The grounding module trains on region-query pairs.



(a) Object-centric pipeline

(b) Relation-aware pipeline

Figure 3. Conditional Prompt Learning.

nates with multiple boxes to make the description *more accurate*.

However, even with the above modifications, Heuristic+ still suffers from the short description with limited style, which looks unreal and lacks diversity.

(2) Object-Centric pipeline

We draw inspiration from the work of BLIP [17] that achieves strong performance on both understanding and generation tasks. In this work, to address the problem suffered in the Uni-modal Heuristic+, we propose the **second pipeline** that leverages pretrained Cross-modal Model to generate pseudo queries, named **Object-Centric**. Specifically, we propose to crop individual proposals and feed-forward them to the pretrained cross-modal models to generate multiple object-centric free-form natural language queries for each image region (proposal). To make the generated description more *object-centric*, we propose conditional prompt learning. The key idea is to make a prompt conditioned on the uni-modal knowledge captured for each region, rather than a fixed one for all regions. We use the pre-prompt template '*object name* {}' and '*object name* is {}' to the decoder and ask the BLIP model to complete the missing part of the sentence. As shown at the top of Figure 3, the former pre-prompt style focuses more on the attribute of the given object, and the latter leans towards describing the action. Such design can guide the model to generate descriptions based on the prior knowledge given in the pre-prompt (captured from the uni-modal knowledge).

However, this object-centric pipeline cannot perform relationship reasoning among multiple objects.

(3) Relation-aware pipeline

To make the model capable of generating relation-aware description, we propose the **third pipeline** that feed-forward the *holistic image* to the pretrained cross-modal models to generate multiple free-form natural language queries, named **Relation-Aware**. As an image contains many salient regions (concepts) and multiple levels of details, this generation pipeline can generate a variety of captions that express different concepts and details. The challenge in this generation pipeline is that due to the decoder now having a full receptive field of the full image, the generated description might suffer referring ambiguity. To alleviate the problem mentioned above and make the generated description more *region-sensitive* and *relation-aware*, we combined the object name and its corresponding spatial relationship to form the pre-prompt template so as to guide the decoder to complete the missing part of the sentence. It is worth noting that such region-conditional pre-prompt is not only depending on the object name captured from the uni-modal pretrained model, but also on their spatial relationship. For example, as shown in Figure 3, we can input the whole image and the prompt "*left man is*" into BLIP to generate a pseudo-query: "*left man is listening to another man*". It is worth noting that this generation pipeline allows the model not only to describe the spatial relationship among different instances but also to deal with other relationships, i.e., human-object interact, human-human interaction, etc., thus further boosting the diversity of the generated pseudo queries.

Note that the cross-modal model is pretrained on image-text pairs (without any box-wise region-text annotation), the generated $<Region-PseudoQuery>$ pair in this way might still contain some spurious association, we propose a verification model later to address this issue.

3.3. Uni-Modal Real Query Propagation

Most of the existing weakly supervised grounding methods first generate region proposals via pre-trained object detectors and then employ either cross-modal similarity score or reconstruction loss as the criteria to implicitly select proposals from them. In contrast, we propose to explicitly calculate the uni-modal similarity score between the real and pseudo query, and propagate the box of the top-1 most similar pseudo query to the real query to form new training samples. In principle, the better the quality and coverage of the generated pseudo queries is, the higher chance we could establish more reliable < Region-RealQuery > associations. The Uni-Modal Real Query propagation is shown as:

$$r_i = \underset{i}{argmax}\, Sim(t, p_{ij}), \forall i, j \tag{1}$$

where r_i denotes the i^{th} proposal of image, t is the real query, $p_{i,j}$ represents the j^{th} pseudo query generated for the i^{th} proposal. Sim($\cdot$) represents the similarity function as:

$$Sim(t, p_{ij}) = \frac{\phi(t)\phi(p_{i,j})}{|\phi(t)||\phi(p_{i,j})|} \tag{2}$$

where $\phi(\cdot)$ represents the function to transform the queries to its semantic text embedding. In principle, we can use any off-the-shelf pre-trained text embedding,i.e.,word2vec [5] , glove [27], bert [7], etc. In this paper, we use word2vec in all following experiments unless otherwise specified.

3.4. Cross-Modal Verification Module

Since some pseudo queries generated by the unimodal model are not realistic and the cross-modal pretrained model generate incorrect pseudo queries, we propose a confidence-aware cross-modal verification module to verify the quality of the $<Region-PseudoQuery>$ association (obtained from Pseudo-Query Generation) and $<Region-RealQuery>$association (obtained from Uni-Modal Real Query propagation) before leveraging them to train the grounding module.

Specifically, we propose to use image-text matching module BLIP model (pretrained with Image-Text Contrastive Loss or Image-Text Matching Loss) to estimate the confidence score c_i of the i^{th} learned association. Based on the confidence score, we can filter and remove spurious association where the paired pseudo or real queries do not accurately describe the corresponding proposal of the images.

3.5. Grounding Module and Training

We finally use both the $<Region-PseudoQuery>$ association pair and $<Region-RealQuery>$association to train a fully-supervised grounding module. We follow the design of previous work [13], which uses a simple stack of transformer encoder layers (consists of a visual encoder, language encoder, a cross-modal fusion module and a regression head) and formulate the grounding task to a coordinate regression problem. The grounding module takes image and query as input and output the bounding box $b_i = (\hat{x_i}, \hat{y_i}, \hat{w_i}, \hat{h_i})$. the training objective of the i^{th} sample is:

$$\mathcal{L}_i = \mathcal{L}_{smooth-L1}(b_i, \hat{b_i}) + \mathcal{L}_{giou}(b_i, \hat{b_i}) \tag{3}$$

where $b^i = (x_i, y_i, w_i, h_i)$ is the normalised ground truth box, $\mathcal{L}_{smooth-L1}$ and $\mathcal{L}_{giou}$ are the smooth L_1 loss and GIoU loss.

We propose a selective grounding loss $\mathcal{L}$ based on the confidence weight to help model distinguish between clean and noisy samples as shown in:

$$\mathcal{L} = \sum_{i \in D} [\frac{\alpha_i}{\sum_{j \in D} \alpha_j}] \mathcal{L}_i, \tag{4}$$

with

$$\alpha_i = \begin{cases} 0 & c_i < \tau \\ c_i & c_i \geq \tau \end{cases} \tag{5}$$

where $D \in 1, ..., N$ indexes the subset of non-zero elements of α. Please note again that the confidence score c_i is predicted from the cross-modal verification module (image-text matching score). We set the weight α_i to 0 if the confidence score is below the threshold τ to remove the noisy or incorrect association. Then re-normalize the remaining verified association by re-normalizing the remaining weights to sum to one (practically conducted in a batch manner). The selective grounding loss highlights the reliable association while suppressing the spurious ones during training. This enables the model to avoid overfitting to some incorrect association (error accumulation).

4. Experiment

4.1. Datasets

RefCOCO/RefCOCO+/RefCOCOg: RefCOCO [47], RefCOCO+ [47] and RefCOCOg [25] are collected from MSCOCO. RefCOCO [47] contains 19,994 images with 142,210 referring expressions for 50,000 referred objects. RefCOCO+ [47] contains 19,992 images with 49,856 referred objects and 141,564 referring expressions. RefCOCOg [25] has 25,799 images with 95,010 referring expressions for 49,856 referred objects. Following previous visual grounding methods [13, 6], we report the performance on the validation, testA and testB splits for RefCOCO and RefCOCO+, validation split for RefCOCOg-google, validation and test splits for RefCOCOg-umd.
ReferItGame: ReferItGame contains 20,000 images collected from the SAIAPR-12 dataset [8]. We follow the previous works [40, 6] to split the dataset into three subsets, including a train set (54,127 referring expressions), a validation set (5,842 referring expressions), and a test set (60,103 referring expressions).
Flickr30K Entities: Flickr30k Entities contains 31,783 images with 427k referred expressions. We follow the same split as in works [6, 40] for train, validation and test.

4.2. Implementation Details

For a fair comparison, we use an existing open-sourced model pretrained on Visual Genome data [16] like other papers. The cross-modal pretrained model BLIP used in our paper is trained on image-query pairs instead of region-query pairs, so the BLIP model itself is lack of region-level grounding capability. we select top-10 objects according to the detection confidence for the cross-modal pre-trained model pipeline. For the uni-modal real query propagation Module, we adopt the word2vec model (300-dim) with the Google-News corpus. We follow the common practice in [20, 42, 43] to perform data augmentation and model initial for model training. Our grounding model is optimized end-to-end with the Adamw optimizer. The initial learning rate is set to 1×10^{-4} except 1×10^{-5} for the visual and language encoder. All the datasets use cosine learning rate schedule and our model is trained with 20 epochs in all datasets.

4.3. Comparisons with State-of-the-art Methods

We show the top-1 accuracy (%) results following previous works [6, 13]. Once the Jaccard overlap between the predicted region and the ground-truth box is above 0.5, the prediction is regarded as a correct one.

In order to enable a fair comparison with different existing approaches, we conduct experiment by using uni-modal pretrained model[1] and cross-modal pretrained model respectively.

RefCOCO/RefCOCO+/RefCOCOg Our method's performances on RefCOCO, RefCOCO+ and RefCOCOg datasets are reported in Table 1. Our method outperforms other unsupervised and weakly supervised methods in all partitions of the three datasets. *Under the set-up of using unimodal pretrained model*, our method can surpass the best unsupervised method Pseudo-Q [13] by a remarkable margin on all three datasets when only. Our method significantly outperforms DTWREG which is the second best weakly supervised method using uni-modal method by more than 23.54%, 7.44%, 11.95% on RefCOCO, RefCOCO+, RefCOCOg, respectively. *Under the set-up of using cross-modal pretrained model,* compared with the second best weakly supervised method ReCLIP, we can still improve the performance by up to 28.48%, 8.24%, 1.11% respectively on RefCOCO, RefCOCO+ and RefCOCOg. These results validate the superiority of our method under different settings. Also, our method performs better in the cross-modal pre-trained model setting than in the uni-modal pre-trained model setting on all three datasets. The phenomenon shows the effective of cross-modal pre-trained model. Finally, there is still a gap between our method and fully-supervised method.

ReferItGame/Flickr30K Entities We also report experimental performance under different setting and show the comparisons with other existing visual grounding methods on ReferItGame and Flickr30K Entities dataset in Table 2. Notably, our method under uni-modal pre-trained model setting achieve 44.07% and 62.96% accuracy which outperforms unspervised method and other weakly supervised method. The experimental results demonstrate the superiority of our proposed method. Also, the performance of method under cross-modal pre-trained model setting is also better than the method under uni-modal pre-trained model setting. Finally, we can observe that the performance of our method is still far from fully supervised methods.

The performance of our model fine-tuned with a small number of labeled samples: We fine-tuned our model with a few labeled training samples. The results in Table 3 show that using just 5% labeled training data narrows the gap with the fully supervised method and even surpasses it by 0.88% on the testA split. With 10% labeled data, our approach outperforms the fully supervised approach.

4.4. Ablation Study

In this section, we empirically investigate how the performance of the proposed method is affected by different

[1]For uni-model preatined model, we first generate pseudo language queries for each region proposal with only Heuristic+ Pipeline. And then propagate the box of the most similar pseudo query to the real query. We do not utilize the BLIP model or cross-model verification in this process to enable a fair comparison.

Method	Sup.	Pre-trained	RefCOCO			RefCOCO+			RefCOCOg		
			val	testA	testB	val	testA	testB	val-g	val-u	test-u
TransVG [6]	Full	Uni-modal	80.32	82.67	78.12	63.50	68.15	55.63	66.56	67.66	67.44
VLTVG [40]		Uni-modal	84.53	87.69	79.22	73.60	78.37	64.53	72.53	74.90	73.88
CPT [44]	No	Uni-model	32.20	36.10	30.30	31.90	35.20	28.80	-	36.70	36.50
Pseudo-Q [13]			56.02	58.25	54.13	38.88	45.06	32.13	49.82	46.25	47.44
VC [49]	Weak	Uni-modal	-	33.29	30.13	-	34.60	31.58	33.79	-	-
ARN [22]			34.26	36.43	33.07	34.53	36.01	33.75	33.75	-	-
KPRN [23]			35.04	34.74	36.98	35.96	35.24	36.96	33.56	-	-
DTWREG [33]			39.21	41.14	37.72	39.18	40.10	38.08	43.24	-	-
Ours			**66.75**	**69.77**	**63.44**	**50.65**	**55.30**	**45.52**	**55.19**	**53.8**	**53.92**
ReCLIP[32]	Weak	Cross-modal	45.78	46.10	47.07	47.87	50.10	45.10	-	59.33	59.01
Ours			**70.67**	**74.58**	**67.19**	**51.81**	**58.34**	**46.17**	**57.04**	**60.21**	**60.12**

Table 1. Comparison with state-of-the-art methods on RefCOCO [47], RefCOCO+ [47], and RefCOCOg [14] datasets in terms of top-1 accuracy (%). "Sup." refers to supervision level: No(unsupervised), Weak(only annotated queries, no box provided) and Full(query-region pairs). "Pre-trained" represents pre-trained model the method utilize. The first and second best results are highlighted in **bold** and underline (excluding the fully supervised approaches), respectively.

Method	Sup.	Pre-trained	ReferIt	Flickr30K
PIN [15]	Full	Uni-modal	59.13	72.83
DDPN [48]			63.00	73.30
FAOA [43]			60.67	68.71
RSC [42]			64.60	69.28
TransVG [6]			69.76	78.47
VLTVG [40]			71.60	79.18
UTG [45]	No	Uni-modal	36.93	20.91
PLM[35]			26.48	50.49
Pseudo-Q [13]			43.32	60.41
KAC [2]	Weak	Uni-modal	33.67	46.61
MATN [50]			33.10	13.61
ARN [22]			26.19	-
CLWPL [9]			-	51.67
RIR[24]			37.68	59.27
CKD [36]			38.39	53.10
Ours			**44.07**	**62.96**
Ours	Weak	Cross-modal	**45.23**	**63.87**

Table 2. Comparison with state-of-the-art methods on ReferItGame and Flickr30K Entities datasets in terms of top-1 accuracy (%). "Sup." refers to supervision level: No(unsupervised), Weak(only annotated queries, no box provided) and Full(query-region pairs). "Pre-trained" represents pre-trained model the method utilize. The first and second best results are highlighted in **bold** and underline, respectively.

Number	val	testA	testB
0%	50.65	55.33	45.52
5%	59.32	69.05	48.59
10%	**64.76**	**70.93**	**55.91**
TransVG	63.50	68.15	55.63

Table 3. Performance of model fine-tuned with different numbers of fully annotated samples on RefCOCO+.

model settings on the RefCOCO+ dataset.

Network Components The method Pseudo-Q [13] serves as a baseline in the comparison. As shown in Table 4, the 5 different components of the proposed model all boost recognition performance compared to the baseline.

Firstly, we observe that the heuristic+ pipeline improves the performance compared with Pseudo-Q, which verifies the effectiveness of our improvement over the original heuristic method. Then adding object-centric and relation-aware pipeline can boost the performance. The result can demonstrate the effectiveness and compatibility of the three pipeline. Also, it is observed that real query propagation contributes to the most performance gain as an individual module under different settings. We attribute this improvement to that our method avoids the distribution shift between the pseudo and real query. And the improvement of the model performance by the Cross-modal verification module verifies that our method can suppress the contribution of the spurious association in the training process.An interesting observation is that the 'H+, O, R, Real-query' approach performs only slightly better than "H+, real-query", sometimes even worse(on the testB split). We conjecture the reason for this is that the BLIP model is trained on image-text pairs without any region-text annotation. This leads to the generation of some erroneous pairs, which have a negative impact on the model's overall performance.

We also investigated the performance on visual grounding tasks using only BLIP model. As shown in Table 5, we can observe that the performance is poor, well below baseline [13]. This is because the cross-modal model we use is trained with image-text pairs instead of region-query pairs, which makes it difficult to be utilized directly on the more fine-grained visual grounding task.

Number of proposal The number of proposals is an im-

H	H+	O	R	Real-Query	Verification	Pre-trained	val	testA	testB
✓						Uni-modal	38.88	45.06	32.13
	✓					Uni-modal	40.36($\uparrow$ 1.48)	45.28($\uparrow$ 0.22)	34.83 ($\uparrow$ 2.70)
	✓			✓		Uni-modal	50.65($\uparrow$ 11.77)	55.30($\uparrow$ 10.24)	45.52 ($\uparrow$ 13.39)
	✓	✓				Cross-modal	42.21 ($\uparrow$ 3.33)	45.62($\uparrow$ 0.56)	38.14 ($\uparrow$ 6.01)
	✓	✓	✓			Cross-modal	44.32($\uparrow$ 5.44)	48.38($\uparrow$ 3.32)	38.79($\uparrow$ 6.66)
	✓	✓	✓	✓		Cross-modal	51.08($\uparrow$ 6.76)	56.51($\uparrow$ 8.13)	44.46($\uparrow$ 5.67)
	✓	✓	✓	✓	✓	Cross-modal	**51.81**($\uparrow$ 0.73)	**58.34**($\uparrow$ 1.83)	**46.17**($\uparrow$ 1.71)

Table 4. Ablations of each component. "*H*" represents Pseudo-Q method. "*H+*", "*O*" and "*R*" denote three pipeline of pseudo query generation respectively. "*Real-Query*" represents the uni-modal real query propagation. "*Verification*" means the confidence-aware cross-modal verification module. "Pre-trained" represents pre-trained model the method utilize.

ReBLIP	Object	Relation	val	testA	testB
✓			12.07	13.20	12.07
	✓		36.96	41.92	31.31
		✓	31.92	34.71	28.94

Table 5. Ablations of cross-modal pre-trained model. "*ReBLIP*" means that utilize BLIP model directly select proposal for model training. "*Object*" and "*Relation*" denote two pipeline using cross-modal of pseudo query generation respectively.

Heuristic+	BLIP	val	testA	testB
4	4	40.83	39.49	40.22
6	6	43.93	46.84	40.43
8	8	47.12	49.55	43.17
10	10	48.77	50.89	44.11
All	10	**51.81**	**58.34**	**46.17**

Table 6. Ablation of the number of object proposals. "*Heuristic+*" denotes Heuristic+ pipeline, "*BLIP*" represents the others pipeline. "*All*" means to filter only tiny objects.

Pseudo-query	val	testA	testB
200	49.94	55.20	44.93
400	50.73	56.20	44.95
800	51.60	56.28	45.56
All	**51.81**	**58.34**	**46.17**

Table 7. Ablation of pseudo-query number. "*All*" means sampling all pseudo queries.

Method	val	testA	testB
Pseudo-Q	38.88	45.06	32.13
Pseudo-Q (Our detectors)	38.03	42.88	37.20
Ours (Our detectors)	**50.65**	**55.30**	**45.52**

Table 8. Performance of Pseudo-Q with different detectors.

Method	Pretrained	Training	val	testA	testB
DTWREG	81M+	29M	39.18	40.10	38.08
Pseudo-Q	210M	155.5M	38.88	45.06	32.13
Ours (Frozen BERT)	230M	45M	46.19	51.09	40.44
Ours (Uni-modal)	230M	155.5M	**50.65**	**55.30**	**45.52**

Table 9. Pretrained and training parameters of different methods.

portant variable that limits many weakly supervised methods. We therefore investigated the effect of using different number of proposals. As shown in Table 6, we easily observe that increasing the number of proposals can improve the performance of our model. This is because the number of proposals result in recall of the referred object.

Numbers of pseudo-query Another important factor is the number of pseudo queries in image. We study the influence of sampling different number of pseudo-queries in Table 7. It can be seen that a large number of pseudo-query can reduce the distribution shift between pseudo and real query thus improving the performance of our model. Note that regardless of the number of pseudo queries generated, after real-query propagation, the number of samples used for model training is comparable to the size of original dataset to achieve a fair comparison.

Effectiveness of different detector: We compare the sensitivity of the detectors used in Pseudo-Q with our detectors in Table 8. Observations: (1) Comparing the first 2 rows, Pseudo-Q shows small variations (some splits better while some splits worse) when different detectors are used, indicating comparable detector accuracy. (2) Comparing the last 2 rows, our approach consistently outperforms Pseudo-Q with the same detectors, validating the effectiveness of our method.

The parameters of different methods: We compared the parameter sizes of previous SOTA models with our model, including pre-training and training parameter sizes. In Table 9, since the parameters of Stanford CoreNLP model is difficult to count, it is replaced by "+" signs. Our parameters in uni-modal setting is comparable to Pseudo-Q, but the performance is improved by 11.77% on RefCOCO+ val. When freezing the BERT model, we have a similar number of training parameters (line 3 of Table 9) to DTWREG, but with a performance improvement of 7.01% on RefCOCO+ val. These experimental results demonstrate the superiority of our method in settings with comparable amount of parameters.

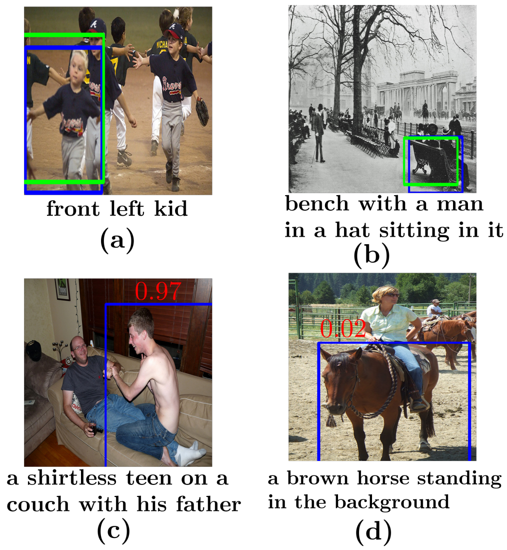


Figure 4. Four visualization examples. Sub-figure(a)(b) demonstrates the effectiveness of the Uni-Modal Real Query propagation module.The green and blue bounding boxes represent the ground truth and those selected proposal by Uni-Modal Real Query propagation module respectively. Sub-figure(c)(d) demonstrates the effectiveness of the verification module.

4.5. Qualitative Analysis

In order to further figure out the importance of uni-modal real query propagation module, we show the qualitative results of two examples from the RefCOCO train set in Figure 4(a)(b). We observe that our approach can successfully select proposals that are close to ground-truth. We also show the qualitative results of cross-modal verification module from the RefCOCOg train set in Figure 4(c)(d). In the first example, we can easily observe that the query is highly consistent with the region in the image and our method also gives high similarity scores. In the last examples, the query does not match the region in the image and our method correspondingly gives low similarity scores. The above examples demonstrate that our method can well select the correct proposal and suppress the contribution of the spurious association.

5. Conclusion

In this paper, we propose Confidence-aware Pseudo-label Learning (CPL) for weakly supervised visual grounding task. Firstly, we propose a pseudo-query generation module to automatically produce pseudo region-query pairs for supervised training. The pseudo-query generation module contains three complementary pipelines that can generate diverse pseudo-queries which makes up for previous work. Secondly, we present an uni-modal real query propagation which can solve the distribution shift between the pseudo and real queries. Finally, to reduce the risk of confirmation bias, we propose a confidence-aware cross-modal verification module that estimates the uncertainty of the region-text association, and propose a selective grounding loss based on the uncertainty weight to suppress the contribution of the spurious association in the training process. Extensive experiments show that our method achieves state-of-the-art methods on five datasets under weak supervision.

6. Acknowledgements

This work was supported by the grants from the Zhejiang Lab (NO.2022NB0AB05), National Natural Science Foundation of China (62132001,61925201), CAAI-Huawei MindSpore Open Fund. We thank MindSpore[2] for the partial support of this work, which is a new deep learning computing framework.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[2] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4050, 2018. 2, 7

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

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3

[5] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017. 5

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

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 5

[8] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villasenor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer vision and image understanding*, 114(4):419–428, 2010. 6

[2]https://www.mindspore.cn/

[9] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 2, 7

[10] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[11] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1115–1124, 2017. 2

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

[13] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15513–15523, 2022. 2, 3, 5, 6, 7

[14] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 2, 7

[15] Rama Kovvuri and Ram Nevatia. Pirc net: Using proposal indexing, relationships and context for phrase grounding. In *Asian Conference on Computer Vision*, pages 451–467. Springer, 2018. 1, 7

[16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 6

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 3, 4

[18] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3

[19] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 3

[20] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020. 2, 6

[21] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4673–4682, 2019. 1, 2

[22] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2611–2620, 2019. 2, 7

[23] 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 *Proceedings of the 27th ACM International Conference on Multimedia*, pages 539–547, 2019. 2, 7

[24] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5612–5621, 2021. 2, 7

[25] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 2, 6

[26] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2701–2710, 2017. 3

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 5

[28] B. A. Plummer and L Wang.... Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123(1):1–20, 2017. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[31] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016. 2

[32] Sanjay Subramanian, Will Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A

strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022. 3, 7

[33] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y Goulermas. Discriminative triad matching and reconstruction for weakly referring expression grounding. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4189–4195, 2021. 2, 7

[34] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

[35] Josiah Wang and Lucia Specia. Phrase localization without paired training examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4663–4672, 2019. 7

[36] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14090–14100, 2021. 2, 7

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

[38] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1960–1968, 2019. 1, 2

[39] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 3

[40] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9499–9508, 2022. 2, 6, 7

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

[42] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive subquery construction. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020. 1, 2, 6, 7

[43] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019. 1, 2, 6, 7

[44] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 7

[45] Raymond A Yeh, Minh N Do, and Alexander G Schwing. Unsupervised textual grounding: Linking words to image concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6125–6134, 2018. 7

[46] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 1, 2

[47] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 2, 6, 7

[48] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. *arXiv preprint arXiv:1805.03508*, 2018. 1, 7

[49] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018. 2, 7

[50] Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. Weakly supervised phrase localization with multi-scale anchored transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5696–5705, 2018. 7

[51] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10012–10022, 2020. 1

[52] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4252–4261, 2018. 2