RefCLIP: A Universal Teacher for Weakly Supervised Referring Expression Comprehension

Lei Jin [12]*, Gen Luo [1]*, Yiyi Zhou[12], Xiaoshuai Sun[12]†, Guannan Jiang[3], Annan Shu[3], Rongrong Ji[12]
[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China.
[2] Institute of Artificial Intelligence, Xiamen University, 361005, P.R. China.
[3]Intelligent Manufacturing Department, Contemporary Amperex Technology Co. Limited (CATL).
{kingts,luogen}@stu.xmu.edu.cn,{zhouyiyi,xssun,rrji}@xmu.edu.cn,{jianggn, shuan01}@catl.com

Abstract

Referring Expression Comprehension (REC) is a task of grounding the referent based on an expression, and its development is greatly limited by expensive instance-level annotations. Most existing weakly supervised methods are built based on two-stage detection networks, which are computationally expensive. In this paper, we resort to the efficient one-stage detector and propose a novel weakly supervised model called RefCLIP. *Specifically, RefCLIP redefines weakly supervised REC as an anchor-text matching problem, which can avoid the complex post-processing in existing methods. To achieve weakly supervised learning, we introduce anchor-based contrastive loss to optimize RefCLIP via numerous anchor-text pairs. Based on RefCLIP, we further propose the first model-agnostic weakly supervised training scheme for existing REC models, where RefCLIP acts as a mature teacher to generate pseudo-labels for teaching common REC models. With our careful designs, this scheme can even help existing REC models achieve better weakly supervised performance than RefCLIP,* e.g., *TransVG and SimREC. To validate our approaches, we conduct extensive experiments on four REC benchmarks,* i.e., *RefCOCO, RefCOCO+, RefCOCOg and ReferItGame. Experimental results not only report our significant performance gains over existing weakly supervised models,* e.g., *+24.87% on RefCOCO, but also show the 5x faster inference speed. Project: https://refclip.github.io.*

1. Introduction

Referring Expression Comprehension (REC), also known as visual grounding [5, 16], aims to locate the target instance in an image based on a referring expression [25–27, 42, 48]. As a cross-modal recognition task, REC is not limited to a fixed set of object categories and is theoretically capable of any open-ended detection [45]. These appealing properties give REC increasing attention from the community of computer vision [25, 28, 45–48]. However, the expensive instance-level annotation has long plagued its development.

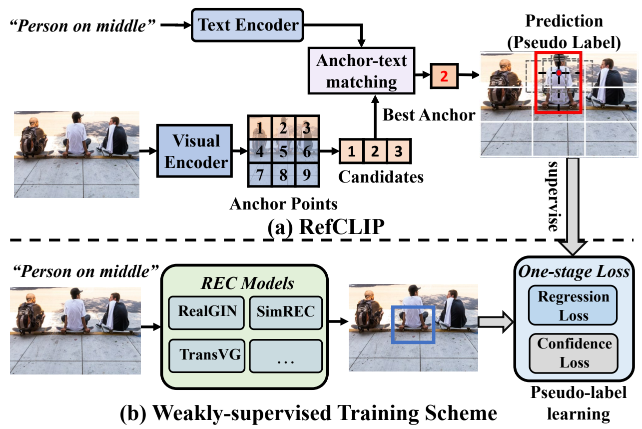

Figure 1. Illustration of the proposed RefCLIP and weakly-supervised training scheme. RefCLIP selects the target bounding box from YOLOv3 via anchor-text matching, which is optimized by anchor-based contrastive learning. Our training scheme uses RefCLIP as a mature teacher to supervise common REC models , which requires no network modifications.

To this end, recent progress has been devoted to the research of weakly supervised REC models, which aim to learn detection based merely on language information [7, 38, 43]. Specifically, existing methods extend the two-stage object detector like Faster-RCNN [37] to a weakly supervised REC model. In terms of methodology, they regard the REC as a region-text ranking problem, where the salient regions of an image are first extracted by Faster-RCNN and then ranked via cross-modal matching. To achieve weakly supervised training, they only use expressions as supervi-

*Equal Contribution.
†Corresponding Author.

sion information and optimize the ranking modules via semantic reconstruction [19,20,38] or cross-modal contrastive learning [7, 43]. However, these methods are often inferior in inference speed due to the use of Faster-RCNN.

To overcome these limitations, we resort to one-stage detectors for weakly supervised REC. Compared with Faster-RCNN, one-stage detectors like YOLOv3 [36] have obvious advantages in efficiency, but it is intractable to directly adapt them to existing weakly supervised schemes. Above all, existing one-stage detectors [17, 36] predict the bounding boxes based on the features of the last few convolution layers, also known as *anchor points* [36]. In terms of multi-scale detection, thousands of bounding boxes will be predicted for an image, so transforming them into region features becomes more time consuming[1]. However, we notice that the receptive field of convolution features will be much larger than the actual areas they represent [29], suggesting that an anchor point in the one-stage detector may contain enough information for recognition.

Motivated by the above observations, we define weakly supervised REC as an anchor-text matching problem and propose a novel weakly supervised model named *RefCLIP*. Specifically, we change the task definition from *which detected region is the referent* to *which anchor point has the target bounding box*. In this case, we can directly rank anchor points without complex post-processing like ROI pooling and NMS [37]. To achieve weakly supervised learning, RefCLIP performs anchor-based contrastive learning inter and intra images, thereby learning vision-language alignments via numerous anchor-text pairs. Notably, this contrastive learning scheme also exhibits superior flexibility in negative sample augmentation, which is not constrained by the batch size.

In this paper, we also focus on the model-agnostic training scheme for weakly supervised REC. Including RefCLIP, all existing solutions are model-specific, which can not directly generalize to existing supervised REC models [5, 25, 42, 45]. To this end, we further propose the first model-agnostic weakly supervised training scheme for REC. Specifically, we use RefCLIP as a teacher to produce pseudo-labels, *i.e.*, bounding boxes, to supervise common REC models. Meanwhile, we also alleviate the confirmation bias [1] caused by pseudo-label noise via EMA [39] and data augmentation [13]. In this scheme, existing REC models can be weakly trained without any modification, which makes our work greatly different from the existing ones [7, 18–20, 38].

To validate the proposed RefCLIP and weakly supervised training scheme, we conduct extensive experiments on four REC benchmarks, *i.e.*, RefCOCO [32], RefCOCO+ [32], RefCOCOg [30] and ReferItGame [10], and compare with a bunch of latest weakly supervised REC models [18, 22, 38, 41]. We apply our training scheme to several representative REC models including RealGIN [45], TransVG [5] and SimREC [25]. Experimental results show obvious performance gains of our RefCLIP over existing weakly supervised REC models, *e.g.*, +21.25% on RefCOCO. Meanwhile, with our careful designs, the proposed training scheme can even help these REC models obtain new SOTA performance of weakly supervised REC.

Conclusively, our main contributions are three-fold:

- We propose a novel one-stage contrastive model called RefCLIP, which achieves weakly supervised REC via anchor-based cross-modal contrastive learning and significantly improves the inference speed by 5 times.
- We propose the first generic weakly supervised training scheme for common REC models, which can effectively boost any REC model using pseudo-labels generated by our RefCLIP.
- The proposed RefCLIP outperforms existing apporoaches on four benchmarks, and our training scheme also helps previous REC models obtain new weakly supervised SOTA performance.

2. Related Work

2.1. Referring Expression Comprehension.

Referring Expression Comprehension (REC) [26,42,45], also known as visual grounding [5, 16] or phrase grounding [6], aims to locate the target object in an image based on the given referring expression. The methodology of REC can be divided into two categories, *i.e.*, two-stage and one-stage based ones. Two-stage methods [16, 21, 42] first use the detection networks like Faster-RCNN [37] to generate a set of candidate regions, and then perform region-text ranking to select the target one. Recently, one-stage approaches [14, 24, 26, 45, 48] obtain more attention due to their high inference speed and superior performance. Early one-stage methods [26,45] mainly consist of shallow multi-modal fusion layers. Inspired by the great success of Transformer [40], recent researchers [5,48] resort to deep Transformer architecture for REC.

2.2. Weakly Supervised Referring Expression Comprehension.

Compared with fully supervised REC, weakly supervised REC is more challenging due to the lack of box annotations. Most existing methods [7, 19, 20, 22, 38, 41, 43] are motivated by two-stage supervised REC models and formulate weakly supervised REC as a region-text ranking problem. In these approaches, the main difficulty relies on

[1] With confidence filtering, this processing still requires about 26.6% additional computation on COCO images.

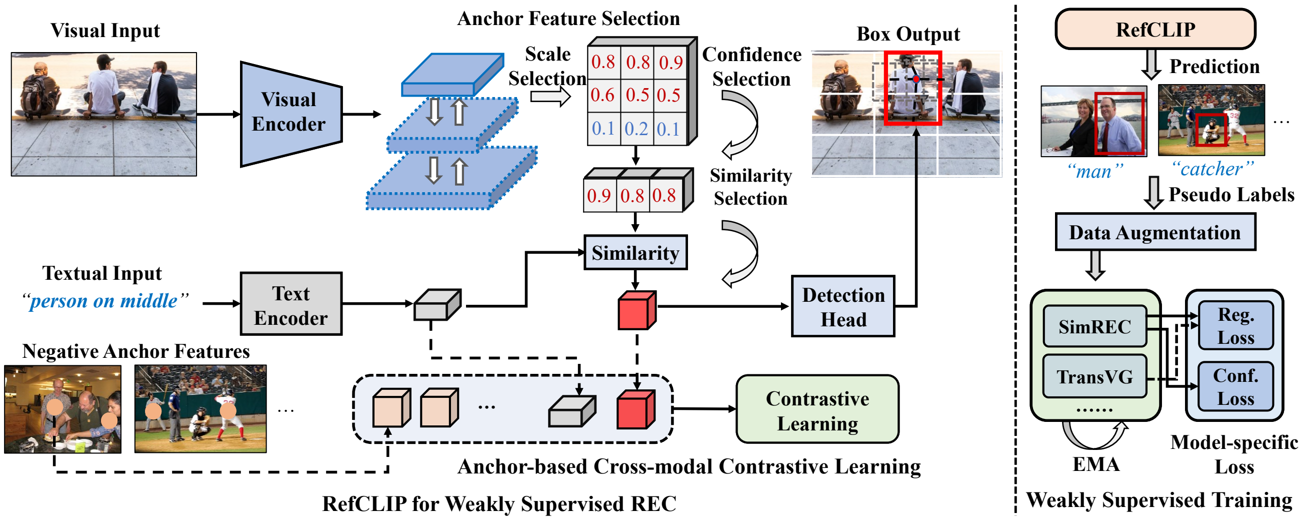

Figure 2. The framework of the proposed RefCLIP (left) and weakly supervised training scheme (right). In RefCLIP, the image and expression are first processed by visual and text encoders. After that, RefCLIP filters anchors of low-value and returns the best-matching one for bounding box prediction. RefCLIP is weakly supervised via anchor-based contrastive learning. In our weakly supervised training scheme, RefCLIP serves as a mature teacher to provide pseudo-labels for training common REC models without any modification.

how to provide effective supervision signal from image-text pairs. To address this issue, researchers adopt methodologies like sentence reconstruction [19,20,38] and contrastive learning [7, 43]. In particular, sentence reconstruction selects the region with the highest ranking score to reconstruct the input expression. Compared to sentence reconstruction, contrastive learning based approaches [7,43] construct positive and negative sample pairs from the selected regions and expressions, and calculate the InfoNCE loss [34]. We also notice that few early work [44] has explored one-stage models for weakly supervised REC, but their performance is still inferior to the two-stage ones. Different from these approaches, RefCLIP is a one-stage model with an innovate weakly supervised formulation, *i.e.,* anchor-text matching. Based on RefCLIP, we propose a new weakly supervised training scheme, *i.e.,* pseudo-label learning, which is applicable to most REC models and does not require any network modification.

3. RefCLIP

3.1. Problem Definition

Given an image I and a text expression T, Referring Expression Comprehension (REC) aims to locate the target instance by a bounding box b. Under the existing weakly supervised setting [19, 20, 38], the model is expected to learn detection based merely on text expressions and images, which is intractable to accomplish.

In this case, existing weakly supervised solutions usually adopt a pre-trained two-stage detection network, *e.g.*, Faster-RCNN [37], to provide a set of candidate bounding boxes B[2], similar to existing two-stage REC methods [16, 21, 42]. Then, REC is formulated as a region-text matching problem, defined by

$$b^* = \arg\max_{b \in B} \Phi(T, I, b), \tag{1}$$

where b^* is the best-matched box, and $\Phi(\cdot)$ is a cross-modal ranking network that returns the similarities between the candidate regions (boxes) and expression. Afterwards, the model conducts weakly supervised training based on semantic reconstruction [19,20,38] or cross-modal contrastive losses [7, 43]. Despite the feasibility, this solution requires complex post-processing, *e.g.*, ROI pooling for region feature extraction, which greatly limits its inference speed.

To this end, we resort to efficient one-stage detectors like YOLOv3 [36] to build our RefCLIP. RefCLIP also leverages the detection capability of YOLOv3. But in practice, we simplify the REC task to an anchor-text matching problem, *i.e.*, which anchor is most likely to have the target box:

$$a^* = \arg\max_{a \in A} \phi(T, I, a), \tag{2}$$

where a^* is the best anchor, A denotes the set of anchor points in YOLOv3, and $\phi(\cdot)$ is a simple linear ranking module. To explain, the prediction of one-stage detectors like YOLOv3 is based on the grid features of the output feature maps, also termed *anchor points*. By knowing which anchor is correct, we can greatly narrow down the range of candidate boxes and finally obtain the most confident box as the prediction.

[2] Some methods use the official annotations of MSCOCO as candidates.

More importantly, through Eq. 2, we can directly use the convolution backbone to extract anchor features without complex post-processing. To achieve weakly supervised optimization, we further perform anchor-based contrastive learning in and out of images.

3.2. Anchor Selection

The framework of RefCLIP is depicted in Fig. 2. Similar to the popular cross-modal contrastive learning scheme, *i.e.*, CLIP [35], RefCLIP also projects visual and text features onto a joint semantic space and learn vision-language alignment via numerous multi-modal pairs.

In RefCLIP, using all anchors as candidates will hinder the efficiency and quality of contrastive learning. It is because that one-stage detectors [17, 36] are often multi-scale, so they have thousands of candidate anchor points, most of which are background or low-quality.

Therefore, RefCLIP needs to filter out most low-value anchors, as illustrated in Fig. 2. Firstly, we only keep the anchors of the last convolution feature map. To explain, in recent REC datasets [10, 30, 32], most objects are relatively large and can be detected by anchors in small-resolution feature maps. Secondly, we filter the remaining anchors according to their confidence scores, *e.g.*, selecting the top 10 percents of anchors.

Afterwards, RefCLIP computes the similarities between these candidate anchors and expression in the joint semantic space, and then returns the best-matching anchor as the positive one for contrastive optimization.

3.3. Anchor-based Contrastive Learning

To achieve weakly supervised learning, we introduce an anchor-based cross-modal contrastive learning scheme. Specifically, given an image I and an expression T, we first use the detection network and language encoder to extract their features, denoted as $\mathbf{F}_v \in \mathbb{R}^{h\times w\times d}$ and $f_t \in \mathbb{R}^d$, respectively. Then, an anchor is represented by the corresponding feature in $\mathbf{F}_v$, denoted as $f_a \in \mathbb{R}^d$.

After anchor selection, we linearly project the selected anchor f_a and the text feature f_t onto the same semantic space, and their similarity is calculated by

$$\text{sim}(f_a, f_t) = (f_a\mathbf{W}_a)^T(f_t\mathbf{W}_t), \tag{3}$$

where $\mathbf{W}_a$ and $\mathbf{W}_t$ are projection matrices, and $\text{sim}(\cdot)$ can be regarded as the lightweight ranking module in Eq. 2.

In REC, the target instance and expression in an image are usually matched one-to-one. Theoretically, only one anchor is the positive example, and the rest ones are negative, especially those that are filtered out. Therefore, we define the contrastive loss inter and intra images:

$$\mathcal{L}_c = -\log \frac{\exp\left(\text{sim}(f^i_{a_0}, f^i_t)/\tau\right)}{\sum\limits_{n=0}^{N}\sum\limits_{j=0}^{M}\mathbb{I}_{\neg(i=j\,\wedge\,n\neq 0)}\exp\left(\text{sim}(f^j_{a_n}, f^i_t)/\tau\right)}, \tag{4}$$

where $f^j_{a_n}$ are anchors sampled from a batch and $f^i_{a_0}$ is the postive one of image i. $\mathbb{I}_{\neg(i=j\,\wedge\,n\neq 0)}$ is the indicator function, which is equal to 0 when $i=j$ and $n\neq 0$. N and M denote the number of negative anchors per image and batchsize, respectively. τ is the temperature [9]. In terms of N, we select the negative anchors based on their confidence scores.

From Eq. 4, we can see the flexibility of RefCLIP in augmenting negative samples. In principle, more negative samples can better facilitate optimization. However, in existing image-level contrastive learning schemes, the number of negative examples is limited to the batch size [4] or relies on external stacks [8]. In our anchor-based scheme, the number of negative samples can be multiple times the batch size, greatly improving the training efficiency.

3.4. Network Settings

As shown in Fig. 2, RefCLIP consists of a pre-trained one-stage detector, *i.e.,* YOLOv3 [36], a language encoder and a multi-scale fusion module [25, 26]. The language encoder is a bidirectional GRU [2] followed by a self-attention layer [40]. Before cross-modal matching, we employ a multi-scale fusion module [26] to fuse the semantic information of three scales.

During inference, RefCLIP first selects the best-matching anchor point, based on which the detection head is used to predict the bounding boxes. Since an anchor point may yield several boxes [36], we use the one with the highest confidence score as the prediction.

4. Pseudo-label based weakly supervised training Scheme

In this section, we introduce a novel pseudo-label based training scheme for arbitrary REC models, which is also the first attempt in REC. In this scheme, RefCLIP plays a role of teacher to teach common REC models via its pseudo-labels, which can help them generalize to weakly supervised REC without any modification.

Given an image-text pair (I, T), we first use RefCLIP to generate the pseudo-label b. After that, we construct a triplet (I, T, b) to supervise the common REC model, and the objective can be defined by

$$\min \mathcal{L}_s(I, T, b; \theta_s), \tag{5}$$

where θ_s denotes the model parameters, and $\mathcal{L}_s$ is the loss function, which can be the ranking loss for two-stage models [42] or the regression one for one-stage models [5, 45].

The pseudo labels generated by RefCLIP are still likely to be noisy and of low quality, leading to a critical issue called *confirmation bias* [1]. This issue means that the training signal may be dominated by noisy samples, and the accumulated errors will eventually limit the performance ceiling. Drawing on the latest research progress [23, 31], we implement two designs to alleviate this problem.

Specifically, we conduct data augmentation on the input image, *e.g.*, *random resize* [13], to prevent the model from prematurely overfitting the pseudo-labeled data. In addition, we adopt Exponential Moving Average (EMA) [39] to the REC model, defined by

$$\theta_s^t \leftarrow \alpha\theta_s^{t-1} + (1-\alpha)\theta_s^t, \quad (6)$$

where α is the EMA coefficient and t is the training step. As defined in Eq. 6, EMA will gradually ensemble the REC models at different training statuses, thereby preventing the decision boundary from moving towards noisy samples.

Lastly, the gradient update in our training scheme is:

$$\theta_s^t = \hat{\theta}_s - \gamma \sum_{k=1}^{t-1} (1-\alpha^{-k+(t-1)}) \frac{\partial \mathcal{L}_s(I,T,b;\theta_s)}{\partial \theta_s^k}, \quad (7)$$

where $\hat{\theta}_s$ denotes the initial model weights.

Although the proposed scheme is similar to fully supervised training, it does not use any ground-truth bounding boxes during training, which is consistent with the definition of weakly supervised REC [19, 20].

5. Experiments

5.1. Datasets and Metric

RefCOCO [32] has 142,210 referring expressions and 50,000 objects from 19,994 MSCOCO [15] images. The expressions of RefCOCO are mainly about absolute spatial information. **RefCOCO+** [32] contains 141,564 referring expressions for 49,856 bounding boxes from 19,992 MSCOCO images. The data splits of RefCOCO+ are the same as RefCOCO. However, the descriptions of RefCOCO+ are about relative spatial information and appearance, *e.g.,* color and texture. **RefCOCOg** [30, 32] has 104,560 referring expressions for 54,822 bounding boxes in 26,711 images. Compared with RefCOCO and RefCOCO+, the expressions of RefCOCOg are longer and more complex. Here, we use the *google* split [30] of RefCOCOg in our experiments. **ReferItGame** [10] has 19,997 images from the SAIAPR-12 dataset, 99,220 bounding boxes and 120,072 referring expressions.We partition the dataset into $train$,val,$test$ according to berkeley split. We use **IoU@0.5** as the metric. If IoU between the predicted and the ground-truth box is larger than 0.5, the prediction is correct.

5.2. Implementation Details

We resize the input image to 416×416. The maximum length of the input text is set to 15 for RefCOCO, RefCOCO+ and RefCOCOg and 20 for ReferItGame. For RefCLIP, we use YOLOv3 [36] as the detector to extract anchor features, which is pre-trained on MS-COCO [15] and the images of *val* and *test* set in three datasets above are removed. For fair comparison with [21, 41] in ReferItGame, we use the YOLOv3 pre-trained on Visual Genome [12] as the detector of our RefCLIP. During training, the parameters of YOLOv3 are fixed. The dimension of the language encoder is set to 512. The anchor features are projected to 512 by the multi-scale fusion. In anchor-based contrastive learning, the dimension of linear projection is 512, and 2 negative anchors per image are used by default. All models are trained by *Adam* [11] optimizer with a constant learning rate of 1e-4. The training epochs and the batch size are set to 25 and 64, respectively. For the weakly supervised training scheme, we apply random resize as the data augmentation to the input image. The EMA coefficient is set to 0.9997. Other configurations of RealGIN, SimREC and TransVG remain the same as their default settings.

5.3. Quantitative Analysis

Ablation of RefCLIP. Tab. 1 shows the ablation results of two main designs in RefCLIP, *i.e.*, *anchor selection* and *negative anchor augmentation* (NAA). NAA denotes that adding negative samples intra images without changing the batch size. We can first observe that anchor filtering is critical for RefCLIP. In the absence of any filtering rules, the performance of RefCLIP is actually far from satisfactory, which confirms our motivation about anchor noise. In this case, a simple scale selection can improve the performance to a large extent, *e.g.*, +17% on RefCOCO. When combined with the confidence-based filtering, the performance can be further improved on both datasets. The results of the last row, *i.e.*, NAA, reflect that adding negative anchors intra images is also beneficial for REC performance, which can improve contrastive learning with very limited additional cost.

Tab. 2 shows the effect of different settings of anchor selection. We first notice that the scales of 52×52 or 26×26 lead to drastic drops in performance, especially the former. As mentioned above, the referents in existing REC datasets are relatively large, so the target bounding boxes are barely distributed on the predictions at these scales, which also explains why the accuracy of 52×52 is zero. In this case, the smallest scale, *i.e.*, 13×13, is the best choice. Even so, the anchor points of YOLOv3 are still redundant. As shown in Tab. 2, by filtering up to 80% or 90% anchors based on confidence, the performance can still be improved slightly. These results well confirm our assumption about the anchor redundancy for contrastive learning.

In Tab. 4, we examine the effect of negative sample size for contrastive learning. Specifically, we adjust the number of negative anchors per image and batch size for controlled experiments, *i.e*, N and M defined in Eq. 4. We first observe that a larger batch size is beneficial for contrastive

Table 1. Ablation study of RefCLIP. " Scale" refers to scale selection. " Conf." is confidence filtering. " NAA" denotes negative anchor augmentation.

Anchor Selection		Contrastive Learning	RefCOCO	RefCOCO+
Scale	Conf.	NAA	val	val
-	-	-	33.71	29.11
✓	-	-	50.75	36.65
✓	✓	-	53.30	40.07
✓	✓	✓	**60.36**	**40.39**

Table 2. The impact of anchor selection settings for RefCLIP.

Anchor Selection	Setting	RefCOCO val	RefCOCO+ val
Scale selection	all	48.75	38.14
	52×52	0.00	0.00
	26×26	11.23	7.19
	13 $\times$ **13**	**60.36**	**40.39**
Confidence filtering	100%	20.84	39.74
	20%	59.31	**41.06**
	10%	**60.36**	40.39
	5%	48.46	39.69

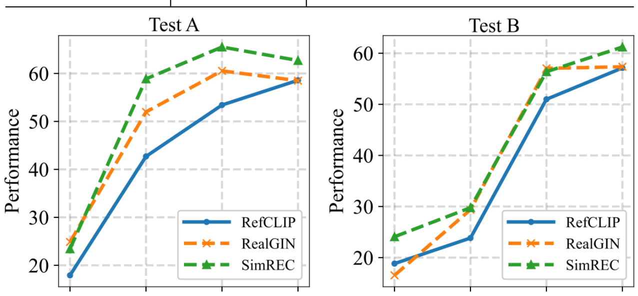

Figure 3. The impact of RefCLIP's performance on common REC models, *i.e.*, RealGIN and SimREC, on RefCOCO test splits.

learning, but the merit will become marginal as the batch size increases. Therefore, we only test the max batch size of 64. The second block shows the effect of negative anchors within images. We can observe that $N = 2$ does not lead to much additional cost, but its performance gain is significant, suggesting our advantage in negative anchor augmentation. We also notice that using more negative anchors is counterproductive, *e.g.*, $N = 3$, which is not consistent with existing contrastive learning study [8]. A potential reason is that RefCLIP only needs to optimize the language encoder and the joint semantic space, which makes it easy to overfit at the existing data scale.

Ablation of weakly supervised training. We further examine the effect of EMA and data augmentation in our scheme in Tab. 3. We can first observe that this training scheme is valid for weakly supervised REC. On all three splits, the performance gap between the weakly supervised RealGIN and RefCLIP is not obvious. Meanwhile, with the help of data augmentation and EMA, the performance of RealGIN is comprehensively improved, suggesting their effectiveness for model training.

Table 3. Ablation study of the proposed weakly-supervised training scheme. RealGIN is the base model and RefCLIP is used for reference.

Model	Method		RefCOCO		
	Aug	EMA	val	testA	testB
RefCLIP	-	-	**60.36**	**58.58**	57.13
RealGIN	-	-	57.36	57.34	56.33
	✓	-	58.99	58.51	55.66
	✓	✓	59.43	58.49	**57.36**

Table 4. Ablation of negative sample size in RefCLIP. N and M denote the numbers of negative anchors per image and batchsize.

Constrastive Learning	Setting	Neg. Number	RefCOCO val	RefCOCO+ val
M	16	15	48.98	40.08
	32	31	52.74	40.98
	64	63	53.30	40.07
N	1	63	53.30	40.07
	2	**126**	**60.36**	**40.39**
	3	189	44.41	38.66
	5	315	42.98	38.46

Fig. 3 illustrates the impact of RefCLIP's performance on the tested REC models. The first observation is that the quality of RefCLIP greatly affects the weakly supervised performance of these common REC models. However, we can also see that RefCLIP's performance is not always the performance upper-bound of our training scheme. When the tested model has a better multi-modal reasoning ability or more advanced designs for REC, their performance can easily exceed RefCLIP under different settings, *e.g.*, SimREC and RealGIN. These results greatly validate the generalization of our scheme for existing REC models.

Comparison to the state-of-the-arts. We examine our weakly supervised training scheme and RefCLIP by comparing to a set of weakly supervised REC models in Tab. 5. In Tab. 5, we compare the proposed RefCLIP and common REC models including both one-stage REC models [5, 25, 45] and two-stage REC models [42] weakly trained by our scheme with more weakly supervised methods. The previous best performance is held by the methods [18, 20, 38] under the settings of using manually annotated boxes as region candidates. Even so, RefCLIP can outperform these methods on most splits, which can be up to 21.1% on RefCOCO *val*.

Tab. 5 also shows the results of existing REC models trained by our weakly supervised training scheme, which are denoted as $\text{RefCLIP}_{\text{ModelName}}$. It can be seen that our training scheme can help common REC models easily surpass the existing SOTA performance on multiple splits, *e.g.*, 71.27 on RefCOCO *test B*. We also observe that the performance gains of MAttNet are more obvious than the one-stage ones, *e.g.,* +14.14% on RefCOCO *testB*. In terms of these results, our hypothesis is that two-stage REC models do not need to learn bounding box regression, which re-

Table 5. Comparisons with state-of-the-art methods on four REC benchmark datasets. *Ground-truth proposals* means using the official annotations of MSCOCO as candidates. For a fair comparison, the inference speeds of these methods are not compared. RefCLIP$_{\text{-ModelName}}$ represents the common REC models trained by RefCLIP in our weakly supervised training scheme.

Method	RefCOCO			RefCOCO+			RefCOCOg	ReferItGame	Inference
	val	testA	testB	val	testA	testB	val-g	test	Speed
Ground-truth Proposals:									
VC [33] CVPR18	-	33.29	30.13	-	34.60	31.58	30.26	-	-
ARN [19] ICCV19	38.05	36.43	36.47	34.53	36.40	36.12	39.62	-	-
KPRN [20] MM19	36.34	35.28	37.72	37.16	36.06	39.29	38.37	33.87	-
DTWREG [38] TPAMI21	39.21	41.14	37.72	39.18	40.01	38.08	43.24	-	-
EARN [18] TPAMI22	38.08	38.25	38.59	37.54	37.58	37.92	45.33	36.86	-
RefCLIP$_{\text{-MAttNet}}$ [42] (ours)	**69.31**	**67.23**	**71.27**	**43.01**	**44.80**	**41.09**	**51.31**	-	-
Detected Proposals:									
VC [33] CVPR18	-	32.68	27.22	-	34.68	28.10	29.65	14.50	-
KAC Net [3] CVPR18	-	-	-	-	-	-	-	15.83	-
MATN [44] CVPR18	-	-	-	-	-	-	-	13.61	-
ARN [19] ICCV19	32.17	35.25	30.28	32.78	34.35	32.13	33.09	26.19	5.7fps
IGN [43] NeurIPS20	34.78	37.64	32.59	34.29	36.91	33.56	34.92	-	-
DTWREG [38] TPAMI21	38.35	39.51	37.01	38.91	39.91	37.09	42.54	-	5.9fps
RelR [22] CVPR21	-	-	-	-	-	-	-	37.68	-
NCE+Distillation [41] CVPR21	-	-	-	-	-	-	-	38.39	-
RefCLIP (ours)	60.36	58.58	57.13	**40.39**	40.45	**38.86**	**47.87**	39.58	31.3fps
RefCLIP$_{\text{-RealGIN}}$ [45] (ours)	59.43	58.49	57.36	37.08	38.70	35.82	46.10	37.56	51.7fps
RefCLIP$_{\text{-SimREC}}$ [25] (ours)	62.57	62.70	61.22	39.13	**40.81**	36.59	45.68	42.33	54.8fps
RefCLIP$_{\text{-TransVG}}$ [5] (ours)	**64.08**	**63.67**	**63.93**	39.32	39.54	36.29	45.70	**42.64**	19.3fps

duces the difficulty of weakly supervised REC to a large extent. More importantly, the inference speed of either RefCLIP or our one-stage base models is much faster than existing weakly supervised models, *e.g.*, RefCLIP improves the inference speed by an order of magnitude compared to DTWREG [38]. These results well confirm the effectiveness of RefCLIP and our training scheme.

5.4. Qualitative Analysis

To obtain deep insight into the proposed RefCLIP and training scheme, we further visualize the predictions under different settings in Fig. 4. From Fig. 4-a, we can see that without any filtering, the vision-language alignment ability of RefCLIP is very limited. Meanwhile, the model is easy to select the boxes of inappropriate sizes, *e.g.*, the 2-th and 4-th examples. Such cases can be well alleviated by scale selection, *i.e.*, "+scale". With confidence filtering, *i.e.*, "+confidence", the prediction accuracy of RefCLIP is further improved, validating our concerns about anchor redundancy. Fig. 4-b shows the predictions of RefCLIP with different negative sample sizes. It can be seen that a proper increase in negative anchors can greatly improve contrastive learning, making anchor-text matching more accurate, *e.g.*, the 1-st example. Lastly, we compare RefCLIP with the base REC models trained by it in Fig. 4-c. It can be seen that the predictions of these common REC models do not always agree with their teacher RefCLIP. When these models have a stronger reasoning ability, *e.g.*, SimREC, they can even show better cross-modal alignment than RefCLIP, *e.g.*, the 7-th and 8-th examples. These results also well confirm the generalization and superiority of our training scheme.

6. Limitation and Future Work

The detection scale of RefCLIP is designed for REC tasks, which may limit its performance in small object detection. Additionally, our weakly training scheme may result in the student model performing better on easier samples, leading to lower teaching quality on more challenging datasets. Future research will focus on addressing these limitations and expanding the application of our approach to other multi-modal tasks.

7. Conclusions

In this paper, we focus on efficient and general weakly supervised REC. Specifically, we first propose a novel weakly supervised model called RefCLIP. To avoid complex region feature extraction, RefCLIP redefines REC as an anchor-text matching problem and achieves weakly su-

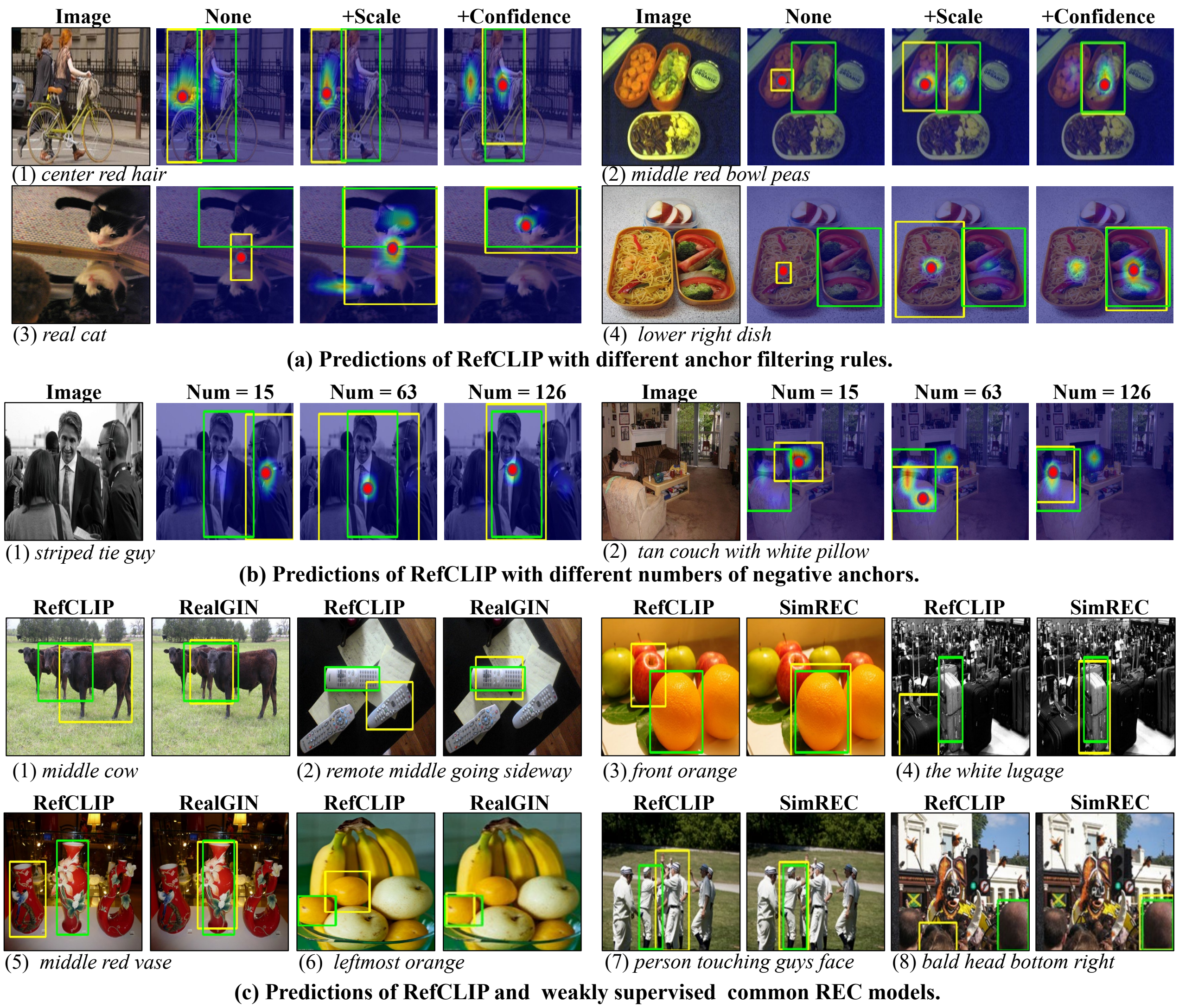

Figure 4. Visualizations of RefCLIP and common REC models trained by our weakly supervised learning scheme. The yellow and green boxes are the predicted and ground truth ones, respectively. Sub-figure (a) shows that scale selection and confidence filtering can help RefCLIP better select the target boxes. The examples in sub-figure (b) reflect the benefit of a larger negative sample size to anchor-text matching. In sub-figure (c), we can see the predictions of common REC models weakly trained by our scheme are not always consistent with their teacher RefCLIP, and they are sometimes even better.

pervised optimization via anchor-based contrastive learning. Based on RefCLIP, we further propose the first model-agnostic weakly supervised training scheme for common REC models, where RefCLIP acts as a teacher for pseudo-label learning. This scheme is applicable to most existing REC models without any network modification. Experimental results on four benchmarks not only show the performance gains of RefCLIP over existing weakly supervised REC models, but also confirm the effectiveness and generalization ability of our training scheme.

Acknowledgements. This work was supported by National Key R&D Program of China (No.2022ZD0118201) , the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001), and the China Fundamental Research Funds for the Central Universities (Grant No. 20720220068).

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 2, 4

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 4

[3] 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. 7

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4

[5] 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, 4, 6, 7

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

[7] 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. 1, 2, 3

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4, 6

[9] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 4

[10] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2, 4, 5

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[12] 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. 5

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2, 5

[14] 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

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[16] 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, 3

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2, 4

[18] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 6, 7

[19] 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, 3, 5, 7

[20] 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, 3, 5, 6, 7

[21] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1950–1959, 2019. 2, 3, 5

[22] 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

[23] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021. 5

[24] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1274–1282, 2020. 2

[25] Gen Luo, Yiyi Zhou, Jiamu Sun, Shubin Huang, Xiaoshuai Sun, Qixiang Ye, Yongjian Wu, and Rongrong Ji. What

goes beyond multi-modal fusion in one-stage referring expression comprehension: An empirical study. *arXiv preprint arXiv:2204.07913*, 2022. 1, 2, 4, 6, 7

[26] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020. 1, 2, 4

[27] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Xinghao Ding, Yongjian Wu, Feiyue Huang, Yue Gao, and Rongrong Ji. Towards language-guided visual recognition via dynamic convolutions. *arXiv preprint arXiv:2110.08797*, 2021. 1

[28] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Yan Wang, Liujuan Cao, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Towards lightweight transformer via group-wise transformation for vision-and-language tasks. *IEEE Transactions on Image Processing*, 31:3386–3398, 2022. 1

[29] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016. 2

[30] 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, 4, 5

[31] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. Active teacher for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14482–14491, 2022. 5

[32] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016. 2, 4, 5

[33] Yulei Niu, Hanwang Zhang, Zhiwu Lu, and Shih-Fu Chang. Variational context: Exploiting visual and textual context for grounding referring expressions. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):347–359, 2019. 7

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[35] 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. 4

[36] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2, 3, 4, 5

[37] 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. 1, 2, 3

[38] 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. 1, 2, 3, 6, 7

[39] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2, 5

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[41] 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, 5, 7

[42] 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, 3, 4, 6, 7

[43] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33:18123–18134, 2020. 1, 2, 3, 7

[44] 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. 3, 7

[45] Yiyi Zhou, Rongrong Ji, Gen Luo, Xiaoshuai Sun, Jinsong Su, Xinghao Ding, Chia-Wen Lin, and Qi Tian. A real-time global inference network for one-stage referring expression comprehension. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 1, 2, 4, 6, 7

[46] Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Deyu Meng, Yue Gao, and Chunhua Shen. Plenty is plague: Fine-grained learning for visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):697–709, 2019. 1

[47] Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. Trar: Routing the attention spans in transformer for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2074–2084, 2021. 1

[48] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. *arXiv preprint arXiv:2203.16265*, 2022. 1, 2