RefCLIP: A Universal Teacher for Weakly Supervised Referring Expression Comprehension

Lei Jin 12, Gen Luo 1*, Yiyi Zhou 12, Xiaoshui 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).

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.

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-
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 approaches 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 multimodal 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.
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 innovative 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

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

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:

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

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.

\footnote{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 learns 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 $\mathbf{f}_t \in \mathbb{R}^d$, respectively. Then, an anchor is represented by the corresponding feature in $\mathbf{F}_v$, denoted as $\mathbf{f}_a \in \mathbb{R}^d$.

After anchor selection, we linearly project the selected anchor $\mathbf{f}_a$ and the text feature $\mathbf{f}_t$ onto the same semantic space, and their similarity is calculated by

$$\text{sim}(\mathbf{f}_a, \mathbf{f}_t) = (\mathbf{f}_a \mathbf{W}_a)^T (\mathbf{f}_t \mathbf{W}_t),$$

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:

$$\mathcal{L}_c = - \log \frac{\exp (\text{sim}(\mathbf{f}_{a_n}, \mathbf{f}_i)/\tau)}{\sum_{n=0}^{N} \sum_{j=0}^{M} \mathbb{1}_{(i=j \land n \neq 0)} \exp (\text{sim}(\mathbf{f}_{a_n}, \mathbf{f}_i)/\tau)},$$

where $\mathbf{f}_{a_n}$ are anchors sampled from a batch and $\mathbf{f}_i$ is the positive one of image $i$. $\mathbb{1}_{(i=j \land n \neq 0)}$ 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. $\tau$ 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 by 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

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

where $\theta_s$ 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 \cite{23, 31}, we implement two designs to alleviate this problem.

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

$$\theta^t_s \leftarrow \alpha \theta^{t-1}_{st} + (1 - \alpha)\theta^t_s,$$

where $\alpha$ 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:

$$\theta^t_s = \hat{\theta}_s - \gamma \sum_{k=1}^{t-1} (1 - \alpha^{-k+(t-1)}) \frac{\partial L_s(I, T, b; \theta_s)}{\partial \theta^k_s},$$

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 \cite{19, 20}.

5. Experiments

5.1. Datasets and Metric

**RefCOCO** \cite{32} has 142,210 referring expressions and 50,000 objects from 19,994 MSCOCO \cite{15} images. The expressions of RefCOCO are mainly about absolute spatial information. **RefCOCO+** \cite{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** \cite{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 \cite{30} of RefCOCOG in our experiments. **ReferItGame** \cite{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 $416 \times 416$. 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 \cite{36} as the detector to extract anchor features, which is pre-trained on MS-COCO \cite{15} and the images of val and test set in three datasets above are removed. For fair comparison with \cite{21, 41} in ReferItGame, we use the YOLOv3 pre-trained on Visual Genome \cite{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 \cite{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 $52 \times 52$ or $26 \times 26$ lead to drastic drops in performance, especially the former. As mentioned above, the refers 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 $52 \times 52$ is zero. In this case, the smallest scale, i.e., $13 \times 13$, 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.

<table>
<thead>
<tr>
<th>Anchor Selection</th>
<th>Contrastive Learning</th>
<th>RefCOCO</th>
<th>RefCOCO+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>val</td>
<td>val</td>
</tr>
<tr>
<td>Scale</td>
<td>Conf.</td>
<td>NAA</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>33.71</td>
<td>29.11</td>
</tr>
<tr>
<td>✓</td>
<td>-</td>
<td>50.75</td>
<td>36.65</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>53.30</td>
<td>40.07</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>60.36</td>
<td>40.39</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Anchor Selection</th>
<th>Setting</th>
<th>RefCOCO</th>
<th>RefCOCO+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>val</td>
<td>val</td>
</tr>
<tr>
<td>Scale selection</td>
<td>all</td>
<td>48.75</td>
<td>38.14</td>
</tr>
<tr>
<td></td>
<td>52 × 52</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>26 × 26</td>
<td>11.23</td>
<td>7.19</td>
</tr>
<tr>
<td></td>
<td>13 × 13</td>
<td>60.36</td>
<td>40.39</td>
</tr>
<tr>
<td>Confidence filtering</td>
<td>100%</td>
<td>20.64</td>
<td>39.74</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>59.31</td>
<td>41.06</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>60.36</td>
<td>40.39</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>48.46</td>
<td>39.69</td>
</tr>
</tbody>
</table>

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.

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 RefCLIP_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 test B. In terms of these results, our hypothesis is that two-stage REC models do not need to learn bounding box regression, which re-

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>RefCOCO</th>
<th>RefCOCO+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aug</td>
<td>val</td>
<td>val</td>
</tr>
<tr>
<td>RefCLIP</td>
<td>-</td>
<td>60.36</td>
<td>58.58</td>
</tr>
<tr>
<td>RealGIN</td>
<td>-</td>
<td>57.36</td>
<td>57.34</td>
</tr>
<tr>
<td></td>
<td>EMA</td>
<td>testA</td>
<td>testB</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>58.99</td>
<td>58.51</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>59.43</td>
<td>58.49</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>60.36</td>
<td>58.58</td>
</tr>
</tbody>
</table>

Table 4. Ablation of negative sample size in RefCLIP. \( N \) and \( M \) denote the numbers of negative anchors per image and batch size.

<table>
<thead>
<tr>
<th>Constrastive Learning</th>
<th>Setting</th>
<th>Neg. Number</th>
<th>RefCOCO</th>
<th>RefCOCO+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>val</td>
<td>val</td>
</tr>
<tr>
<td>M</td>
<td>16</td>
<td>15</td>
<td>48.98</td>
<td>40.08</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>31</td>
<td>52.74</td>
<td>40.98</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>63</td>
<td>53.30</td>
<td>40.07</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>126</td>
<td>60.36</td>
<td>40.39</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>189</td>
<td>44.41</td>
<td>38.66</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>315</td>
<td>42.98</td>
<td>38.46</td>
</tr>
</tbody>
</table>
Table 5. Comparisons with state-of-the-art methods on four REC benchmark datasets. 

<table>
<thead>
<tr>
<th>Method</th>
<th>Ground-truth Proposals:</th>
<th>RefCOCO</th>
<th>RefCOCO+</th>
<th>RefCOCOg</th>
<th>ReferItGame</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>val</td>
<td>testA</td>
<td>testB</td>
<td>val-g</td>
</tr>
<tr>
<td>VC [33]CVPR18</td>
<td></td>
<td>-</td>
<td>33.29</td>
<td>30.13</td>
<td>34.60</td>
</tr>
<tr>
<td>ARN [19]ICCV19</td>
<td></td>
<td>-</td>
<td>38.05</td>
<td>36.43</td>
<td>36.47</td>
</tr>
<tr>
<td>KPRN [20]MM19</td>
<td></td>
<td>-</td>
<td>36.34</td>
<td>35.28</td>
<td>37.72</td>
</tr>
<tr>
<td>DTWREG [38]TPAMI21</td>
<td></td>
<td>-</td>
<td>39.21</td>
<td>41.14</td>
<td>37.72</td>
</tr>
<tr>
<td>EARN [18]TPAMI22</td>
<td></td>
<td>-</td>
<td>38.08</td>
<td>38.25</td>
<td>38.59</td>
</tr>
<tr>
<td>RefCLIP_MAttNet [42] (ours)</td>
<td></td>
<td>69.31</td>
<td>67.23</td>
<td>71.27</td>
<td>43.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Detected Proposals:</th>
<th></th>
<th>RefCOCO</th>
<th>RefCOCO+</th>
<th>RefCOCOg</th>
<th>ReferItGame</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>val</td>
<td>testA</td>
<td>testB</td>
<td>val-g</td>
</tr>
<tr>
<td>VC [33]CVPR18</td>
<td></td>
<td>-</td>
<td>32.68</td>
<td>27.22</td>
<td>-</td>
</tr>
<tr>
<td>KAC Net [3]CVPR18</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MATN [44]CVPR18</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ARN [19]ICCV19</td>
<td></td>
<td>-</td>
<td>32.17</td>
<td>35.25</td>
<td>30.28</td>
</tr>
<tr>
<td>IGN [45]NeurIPS20</td>
<td></td>
<td>-</td>
<td>34.78</td>
<td>37.64</td>
<td>32.59</td>
</tr>
<tr>
<td>DTWREG [38]TPAMI21</td>
<td></td>
<td>-</td>
<td>38.35</td>
<td>39.51</td>
<td>37.01</td>
</tr>
<tr>
<td>ReIR [22]CVPR21</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NCE+Distillation [41]CVPR21</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RefCLIP (ours)</td>
<td></td>
<td>60.36</td>
<td>58.58</td>
<td>57.13</td>
<td>40.39</td>
</tr>
<tr>
<td>RefCLIP_RealGIS [45] (ours)</td>
<td></td>
<td>59.43</td>
<td>58.49</td>
<td>57.36</td>
<td>37.08</td>
</tr>
<tr>
<td>RefCLIP_SimREC [25] (ours)</td>
<td></td>
<td>62.57</td>
<td>62.70</td>
<td>61.22</td>
<td>39.13</td>
</tr>
<tr>
<td>RefCLIP_TransVG [5] (ours)</td>
<td></td>
<td>64.08</td>
<td>63.67</td>
<td>63.93</td>
<td>39.32</td>
</tr>
</tbody>
</table>

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.

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


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


[38] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y Goulgermas. 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


[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

