Language Capability Enhancements in Weakly Supervised Referring Expression Comprehension

BoTong[1], Donglin Qian[2], Gao Han[1] Chang Dong[1]

[1] School of Infomatics, Xiamen University
[2]Artificial Intelligence Research Institute, Xiamen University
31520231154314@stu.xmu.edu.cn, 36920231153227@stu.xmu.edu.cn,
23020231154185@stu.xmu.edu.cn, 23020231154178@stu.xmu.edu.cn,

Abstract

Referring Expression Comprehension (REC) is a crucial task in natural language processing and computer vision, aiming to ground referents based on given expressions. Conventional fully-supervised REC demands extensive instance-level annotations, which are costly and time-consuming to obtain, thus limiting the scalability and applicability of REC models. The adoption of weakly supervised methods significantly reduces the annotation burden, making REC more accessible for a wider range of applications. RefCLIP(Jin et al. 2023) defines weak supervision as an anchor text matching problem, introducing anchor-based contrastive loss, and optimizing RefCLIP with a large number of anchor text pairs, achieving impressive performance. In this model, LSTM(Hochreiter and Schmidhuber 1997) is employed for text feature extraction, and a simple linear mapping is used for feature fusion. We believe that utilizing more advanced language encoders can enhance the model's understanding of natural language descriptions while incorporating a deep feature fusion module facilitates better integration of image features and textual information. Therefore, in this study, we investigate factors that impact the REC task. Through rigorous empirical analysis, we reveal that specific text encoding and feature fusion methods significantly enhance the performance of weakly supervised REC. This research not only advances academic understanding of weakly supervised REC but also provides valuable insights with practical relevance for real-world applications.

Introduction

Referring Expression Comprehension (REC) aims to locate the target instance in an image based on a referring expression (Luo et al. 2020). REC covers the understanding and coordination of multimodal information (text and images) and is a crucial step toward creating more intelligent and interactive human-computer interfaces. These remarkable features have garnered increasing attention from the computer vision community (Luo et al. 2022). However, instance-level annotation often comes at a high cost, and is challenging to collect professional instance-level annotations in fields such as medicine, significantly constraining the development of the REC task.

To overcome this limitation, some researchers have started to explore weakly supervised REC models. Existing weakly supervised methods usually extend two-stage object detectors, such as Faster-RCNN (Ren et al. 2017), to weakly supervised REC models. Specifically, they treat REC as a region-text ranking problem, where they first extract salient regions from images using Faster-RCNN and then rank these regions through cross-modal matching. However, these methods often have slower inference speeds, making them less suitable for real-time applications.

In comparison to Faster-RCNN, one-stage detectors like YOLOv3 (Redmon and Farhadi 2018) offer distinct advantages in efficiency. Yet, the challenge lies in how to adapt them effectively to existing weakly supervised schemes. Existing one-stage detectors (Redmon and Farhadi 2018) typically predict bounding boxes based on features from the last few convolution layers, also known as anchor points. Since one-stage detectors typically make multiple-scale predictions, these anchor point predictions often involve many bounding boxes. These bounding boxes need to be associated with textual descriptions to facilitate weakly supervised tasks. This process can be labor-intensive and time-consuming.

Fortunately, the recently proposed model called RefCLIP redefines weakly supervised REC as an anchor-text matching problem, avoiding the complex post-processing in existing methods. To achieve weakly supervised learning, RefCLIP introduces anchor-based contrastive loss to optimize RefCLIP via numerous anchor-text pairs. It achieves significant performance gains over existing weakly supervised models, +24.87 % on RefCOCO with an inference speed 5x faster than the former method.

However, the text given in REC tends to be more specific and complicated, which asks for a higher text processing ability from the model. Our model uses RefCLIP as a baseline. By replacing the former text decoder with pre-trained BERT, our model gets a better ability to obtain text features and achieves better performance than RefCLIP.

Related Work

1. Referring Expression Comprehension

We used the concept of REC (Referring Expression Comprehension) (Luo et al. 2020) to achieve the purpose of locating the target object in the image based on the given reference expression. Currently, REC implementation methods can be mainly divided into two types: two-stage detection networks and one-stage detection networks. Among them, two-stage detection networks (Liu et al. 2019b) propose possible object bounding boxes in the first stage and determine the final object detection and localization in the second stage. They are suitable for applications that require precise localization and detection of objects in images. Typical two-stage detection networks include Faster R-CNN (Ren et al. 2017) and R-FCN. On the other hand, one-stage detection networks (Luo et al. 2020) perform object detection in a single step by directly classifying and regressing bounding boxes for all positions in the image. They are suitable for applications that require accelerated object detection. Typical one-stage detection networks include YOLO (You Only Look Once) and SSD (Single Shot MultiBox Detector). Due to the high-speed requirements of the application in this paper, a one-stage method, specifically YOLOv3, is adopted.

2. Weakly Supervised Referring Expression Comprehension

Weakly supervised learning is a machine learning approach that involves training models when the quality of labels in the training data is low or uncertain. It is particularly applicable to tasks like Referring Expression Comprehension (REC) with costly instance-level annotation requirements. It allows for reducing label costs while accepting some loss in accuracy. However, implementing weakly supervised REC can be more challenging than fully supervised REC due to the lack of bounding box annotations.

Most existing methods (Liu et al. 2019a; Wang et al. 2021; Zhang et al. 2020) for weakly supervised Referring Expression Comprehension predominantly rely on two-stage supervised REC models. During the early exploration of one-stage models for weakly supervised REC (Zhao et al. 2018), it was found that their performance was not as good as two-stage models. Two-stage supervised REC models frame the REC task as a region-text ranking problem. This task necessitates that the model effectively comprehends the relationship between textual descriptions and image content and ranks textual descriptions or image regions based on relevance. The primary challenge lies in providing effective supervision signals in image-text pairs. Researchers have addressed this issue using methods such as sentence reconstruction (Liu et al. 2019a) and contrastive learning (Zhang et al. 2020). While these methods achieve high accuracy, they are computationally expensive, and Faster R-CNN models often suffer from slower inference speeds.

Different from these approaches, the model used in this paper is based on a one-stage network testing approach, specifically YOLOv3. It employs an anchor-text matching method for text matching and information retrieval.

Proposed Solution

1. Problem Definition

In the current weakly supervised setting (Liu et al. 2019a), Referring Expression Comprehension (REC) aims to locate the target instance within an image, denoted as I, using a textual expression T to define its bounding box b. However, achieving detection solely based on text expressions and images is infeasible.

In this case, existing weakly supervised solutions usually adopt a pre-trained two-stage detection network, e.g., Faster-RCNN (Ren et al. 2017), to provide a set of candidate bounding boxes B, similar to existing two-stage REC methods (Liu et al. 2019b). Then, REC is formulated as a region-text matching problem, defined by:

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

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. Afterward, the model conducts weakly supervised training based on semantic reconstruction (Liu et al. 2019a) or cross-modal contrastive losses (Zhang et al. 2020). Although feasible, this approach necessitates intricate post-processing steps, such as ROI pooling for region feature extraction, resulting in a substantial reduction in its inference speed.

To this end, we turn to the utilization of efficient one-stage detectors such as YOLOv3 (Redmon and Farhadi 2018)in constructing our RefCLIP. RefCLIP capitalizes on the detection capabilities offered by YOLOv3. However, 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:

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

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,one-stage detectors such as YOLOv3 make predictions relying on grid features within output feature maps, referred to as 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.

More importantly, utilizing Eq. 2 allows us to directly employ the convolution backbone for extracting anchor features without intricate post-processing. To accomplish weakly supervised optimization, we extend this by conducting anchor-based contrastive learning both within and outside of images.

2. Anchor Selection

The framework of RefCLIP, presented in Figure 1, follows a similar structure to the widely adopted cross-modal contrastive learning model, CLIP (Radford et al. 2021). Just like CLIP, RefCLIP aligns visual and textual features within a shared semantic space, facilitating the learning of vision-language correspondence across diverse multi-modal pairs.

Within RefCLIP, employing all anchors as candidates will hinder the efficiency and quality of contrastive learning. It is because one-stage detectors (Redmon and Farhadi 2018) are frequently multi-scale, resulting in the generation of thousands of candidate anchor points, a significant portion of which are background or low-quality elements.

Hence, RefCLIP needs to filter out and eliminate a majority of low-value anchors, as depicted in Fig. 1. Firstly,

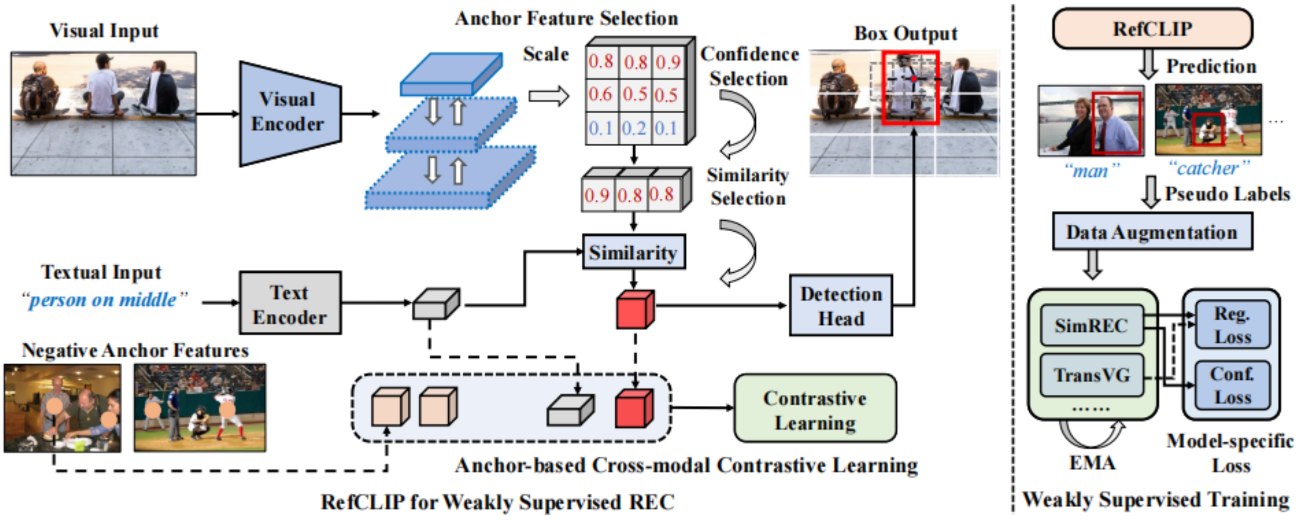

Figure 1: The framework of the proposed RefCLIP (left) and weakly supervised training scheme (right).

we retain solely the anchors from the final convolution feature map. To explain, in recent REC datasets (Mao et al. 2016; Nagaraja, Morariu, and Davis 2016), 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 percent of anchors.

Following this, RefCLIP calculates the similarities between these candidate anchors and expressions in the joint semantic space, eventually returning the best-matching anchor as the positive one for contrastive optimization.

3. Anchor-based Contrastive Learning

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

Following anchor selection, we perform a linear projection of the chosen anchor f_a and the text feature f_t into a shared semantic space. Their similarity is subsequently computed by

$$sim\left(f_a, f_t\right) = \left(f_a \mathbf{W}_a\right)^T \left(f_t \mathbf{W}_t\right) \qquad (3)$$

where W_a and W_t are projection matrices, and sim($\cdot$) can be regarded as the lightweight ranking module in Eq. 2.

Within REC, typically, the target instance and expression in an image form a one-to-one match. In theory, only one anchor serves as a positive example, while the remainder, particularly those filtered out, act as negatives. Hence, we establish the definition of inter and intra images contrastive loss as follows:

$$\mathcal{L}_c = -\log \frac{\exp\left(sim\left(f_{a_0}^i, f_t^i\right)/\tau\right)}{\sum_{n=0}^{N}\sum_{j=0}^{M} I_{\neg(i=j\wedge n\neq 0)} \exp\left(sim\left(f_{a_n}^j, f_t^i\right)/\tau\right)}, \qquad (4)$$

where $f_{a_n}^j$ are anchors sampled from a batch and $f_{a_0}^i$ is the positive one of image i. $I_{\neg(i=j\wedge n\neq 0)}$ is the indicator function, which is equal to 0 when $i = j$ and $n = 0$. N and M denote the number of negative anchors per image and batch size, respectively. τ is the temperature (Hinton, Vinyals, and Dean 2015). In terms of N, we select the negative anchors based on their confidence scores.

Eq. 4 highlights RefCLIP's adaptability in augmenting negative samples. In principle, more negative samples can better facilitate optimization. Yet, existing image-level contrastive learning approaches typically confine the count of negative instances to the batch size or depend on external repositories. Contrastingly, within our anchor-based framework, the pool of negative samples can surpass the batch size by multiple folds, significantly enhancing training efficiency.

4. Network Settings

RefCLIP, depicted in Figure 1, comprises a pre-trained one-stage detector (YOLOv3) (Redmon and Farhadi 2018), a language encoder, and a multi-scale fusion module (Luo et al. 2020). The language encoder consists of a bidirectional GRU (Bahdanau, Cho, and Bengio 2015) followed by a self-attention layer (Vaswani et al. 2017). Before cross-modal matching, we utilize a multi-scale fusion module (Luo et al. 2020)to amalgamate semantic information across three scales.

In the inference phase, RefCLIP initially identifies the best-matching anchor point, utilizing the detection head to forecast bounding boxes. Given that an anchor point might correspond to multiple boxes (Redmon and Farhadi 2018), we select the one with the highest confidence score as the prediction.

5. Pseudo-label based weakly supervised training Scheme

Within this section, we introduce an innovative pseudo-label-based training scheme designed for arbitrary REC models, which is also the first attempt in REC. In this scheme, RefCLIP assumes the role of a teacher, imparting knowledge to conventional REC models through its pseudo-labels. This transfer of information aids these models in adapting to weakly supervised REC without necessitating any alterations.

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

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

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 or the regression one for one-stage models.

The pseudo labels produced by RefCLIP may still contain noise and be of inferior quality, consequently resulting in a significant concern known as confirmation bias (Arazo et al. 2020). This concern implies that the training signal might be excessively influenced by noisy samples, ultimately restricting the performance ceiling due to accumulated errors. Drawing on the latest research progress (Mi et al. 2022), we implement two designs to alleviate this problem.

More specifically, we implement data augmentation techniques on the input image, such as random resizing (Krizhevsky, Sutskever, and Hinton 2017), to deter the model from prematurely overfitting to the pseudo-labeled data. In addition, we adopt Exponential Moving Average (EMA) (Tarvainen and Valpola 2017) to the REC model, defined by

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

where α is the EMA coefficient and t is the training step. As described in Eq. 6, EMA will gradually ensemble the REC models at different training statuses, effectively preventing the decision boundary from being influenced by noisy samples.

Lastly, the gradient update in our training scheme is:

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

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

While resembling fully supervised training, the proposed scheme operates without utilizing any ground-truth bounding boxes in its training process, aligning with the definition of weakly supervised REC(Liu et al. 2019a).

Experiments

1. Datasets and Metric

RefCOCO (Nagaraja, Morariu, and Davis 2016) has 142,210 referring expressions and 50,000 objects from 19,994 MSCOCO (Lin et al. 2015) images. The expressions of RefCOCO are mainly about absolute spatial information. **RefCOCO+**(Nagaraja, Morariu, and Davis 2016) 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** (Mao et al. 2016; Nagaraja, Morariu, and Davis 2016) 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 (Mao et al. 2016) of RefCOCOg in our experiments. **ReferItGame** (Kazemzadeh et al. 2014) 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, and 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.

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 (Redmon and Farhadi 2018) as the detector to extract anchor features, which is pre-trained on MS-COCO (Lin et al. 2015), and the images of val and test set in the three datasets above are removed. For a fair comparison with (Liu et al. 2019b; Wang et al. 2021) in ReferItGame, we use the YOLOv3 pre-trained on Visual Genome (Krishna et al. 2017) 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 (Kingma and Ba 2017) 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.

3. Quantitative Analysis

Table 1 presents the comparative results for two crucial components in the RefCLIP model: the choice of language encoder and the method of feature fusion. The comparison focuses on the impact of these design choices on the REC task's performance.

Language Encoder Comparison: Initially, we observed that the choice of language encoder plays a significant role in the REC task's success. The basic LSTM encoder, while adequate, did not offer optimal results. For instance, using LSTM, the accuracy on the RefCOCO dataset was satisfactory but not exceptional. In contrast, the adoption of the BERT encoder markedly improved performance. With BERT, there was an impressive improvement. This enhancement is attributable to BERT's superior ability to understand the context and nuances in natural language, confirming our hypothesis about the importance of advanced language processing in REC.

Feature Fusion Strategies: Furthermore, we explored the impact of different feature fusion strategies within the RefCLIP model. Despite experimenting with two fusion methods, namely simple linear mapping and Cross-Scale Feature Fusion Module(CCFM), the deep feature fusion module failed to deliver substantial enhancements over the linear mapping approach. This limited improvement may be attributed to several factors.

Firstly, it's possible that the complexity introduced by the deep feature fusion module did not align with the specific characteristics of our REC task. The task may not inherently require the intricate feature interactions that the deep fusion module was designed to capture. Additionally, the dataset used in our experiments may not have contained sufficiently diverse and complex examples to fully leverage the capabilities of the deep fusion module.

article multirow

Table 1: The ablation results for two crucial components in the RefCLIP model

Model	RefCOCO		
	val	testA	testB
+LSTM	60.36	58.58	57.13
+LSTM & CCFM	60.15	58.60	57.10
+BERT	63.20	60.02	59.23
+BERT & CCFM	63.22	60.00	59.20

Conclusion

This study has made significant strides in advancing the field of Referring Expression Comprehension (REC) by exploring the efficacy of weakly supervised methods, specifically through the RefCLIP model. Our findings underscore the limitations inherent in fully-supervised REC methods, primarily due to their extensive demands for instance-level annotations. These requirements not only escalate the cost but also consume considerable time, thereby hampering the scalability and practical applicability of REC models.

The introduction of weakly supervised methods marks a pivotal shift in this landscape. By significantly reducing the annotation burden, these methods have made REC more accessible and versatile for a broader spectrum of applications. The RefCLIP model, as proposed by Jin et al. in 2023, exemplifies this approach. By conceptualizing weak supervision as an anchor text matching problem and integrating an anchor-based contrastive loss, RefCLIP harnesses a large corpus of anchor text pairs, culminating in remarkable performance enhancements.

A notable aspect of the RefCLIP model is its use of LSTM (Hochreiter and Schmidhuber, 1997) for extracting text features, complemented by a straightforward linear mapping for feature fusion. Our research posits that the incorporation of more sophisticated language encoders could significantly bolster the model's proficiency in interpreting natural language descriptions. Moreover, the integration of a deep feature fusion module is anticipated to enhance the synergy between image features and textual data, thereby enriching the overall REC process.

Our empirical investigations into various factors influencing the REC task have yielded critical insights. Specifically, we discovered that the choice of text encoding and feature fusion methods exert a substantial impact on the performance of weakly supervised REC systems. These findings not only contribute to the academic discourse by enhancing the understanding of weakly supervised REC but also hold immense practical value. They pave the way for more refined and efficient REC applications in real-world settings, thus bridging the gap between theoretical research and practical utility.

In conclusion, this study reaffirms the transformative potential of weakly supervised methods in REC. By delving into the nuances of text encoding and feature fusion, we have illuminated pathways for future research and application enhancements. Our work catalyzes further innovations in REC, promising to expand its reach and effectiveness across various domains.

References

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2020. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Jin, L.; Luo, G.; Zhou, Y.; Sun, X.; Jiang, G.; Shu, A.; and Ji, R. 2023. RefCLIP: A Universal Teacher for Weakly Supervised Referring Expression Comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2681–2690.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. L. 2014. ReferIt Game: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vision*, 123(1): 32–73.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM*, 60(6): 84–90.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Liu, X.; Li, L.; Wang, S.; Zha, Z.-J.; Meng, D.; and Huang, Q. 2019a. Adaptive Reconstruction Network for Weakly Supervised Referring Expression Grounding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2611–2620.

Liu, X.; Wang, Z.; Shao, J.; Wang, X.; and Li, H. 2019b. Improving Referring Expression Grounding With Cross-Modal Attention-Guided Erasing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1950–1959.

Luo, G.; Zhou, Y.; Sun, X.; Cao, L.; Wu, C.; Deng, C.; and Ji, R. 2020. Multi-Task Collaborative Network for Joint Referring Expression Comprehension and Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Luo, G.; Zhou, Y.; Sun, X.; Wang, Y.; Cao, L.; Wu, Y.; Huang, F.; and Ji, R. 2022. Towards Lightweight Transformer Via Group-Wise Transformation for Vision-and-Language Tasks. *IEEE Transactions on Image Processing*, 31: 3386–3398.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.; and Murphy, K. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11–20.

Mi, P.; Lin, J.; Zhou, Y.; Shen, Y.; Luo, G.; Sun, X.; Cao, L.; Fu, R.; Xu, Q.; and Ji, R. 2022. Active Teacher for Semi-Supervised Object Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14462–14471.

Nagaraja, V.; Morariu, V.; and Davis, L. 2016. Modeling Context Between Objects for Referring Expression Understanding.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Redmon, J.; and Farhadi, A. 2018. YOLOv3: An Incremental Improvement. arXiv:1804.02767.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137–1149.

Tarvainen, A.; and Valpola, H. 2017. Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 1195–1204. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6000–6010. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Wang, L.; Huang, J.; Li, Y.; Xu, K.; Yang, Z.; and Yu, D. 2021. Improving Weakly Supervised Visual Grounding by Contrastive Knowledge Distillation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14085–14095.

Zhang, Z.; Zhao, Z.; Lin, Z.; Zhu, J.; and He, X. 2020. Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Zhao, F.; Li, J.; Zhao, J.; and Feng, J. 2018. Weakly Supervised Phrase Localization with Multi-scale Anchored Transformer Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5696–5705.