

RefTeacher: A Strong Baseline for Semi-Supervised Referring Expression Comprehension

Jiamu Su¹, Gen Lu¹, Yiyi Zhou^{1,2*}, Xiaoshuai Su², Guannan Jiang³, Zhiyu Wang³, Rongrong Ji^{2,4}

¹Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Xiamen University

²Institute of Artificial Intelligence, Xiamen University

³Contemporary Ampere Technology, Shenzhen Research Institute of Xiamen University

f sunjiamu, luogen g@stu.xmu.edu.cn, f zhouyiyi, xssun, rrji g@xmu.edu.cn, f jianggn, wangzy13 g@catl.com

Abstract

Referring expression comprehension (REC) often requires a large number of instance-level annotations for fully supervised learning, which are laborious and expensive. In this paper, we present the first attempt of semi-supervised learning for REC and propose a strong baseline method called RefTeacher. Inspired by the recent progress in computer vision, RefTeacher adopts a teacher-student learning paradigm, where the teacher REC network predicts pseudo-labels for optimizing the student one. This paradigm allows REC models to exploit massive unlabeled data based on a small fraction of labeled. In particular, we also identify two key challenges in semi-supervised REC, namely, sparse supervision signals and worse pseudo-label noise. To address these issues, we equip RefTeacher with two novel designs called Attention-based Imitation Learning (AIL) and Adaptive Pseudo-label Weighting (APW). AIL can help the student network imitate the recognition behaviors of the teacher, thereby obtaining sufficient supervision signals. APW can help the model adaptively adjust the contributions of pseudo-labels with varying qualities, thus avoiding confirmation bias. To validate RefTeacher, we conduct extensive experiments on three REC benchmark datasets. Experimental results show that RefTeacher obtains obvious gains over the fully supervised methods. More importantly, using only 10% labeled data, our approach allows the model to achieve near 100% fully supervised performance, only -2.78% on RefCOCO. Project: <https://refteacher.github.io/>.

1. Introduction

Referring Expression Comprehension (REC) [33–35, 45, 49, 53, 60, 62], also called Visual grounding [26, 51, 52] or Phrase localization [20, 39], aims to locate the target

Figure 1. Statistics of the pseudo-label quality in semi-supervised REC with different percentages of labeled information. The REC model only predicts one pseudo-box for each image-text pair and cannot apply filtering, so the pseudo-labels are usually noisy and low-quality during training. objects in an image referred by a given natural language expression. Compared to conventional object detection tasks [4, 9–11, 23, 24, 28, 40–42], REC is not limited to a fixed set of categories and can be generalized to open-vocabulary recognition [31, 60]. However, as a detection task, REC also requires a large number of instance-level annotations for training, which poses a huge obstacle to its practical applications.

To address this issue, one feasible solution is semi-supervised learning (SSL), which has been well studied on various computer vision tasks [1–3, 8, 16, 29, 43, 44, 47, 59] but not yet exploited in REC. In particular, recent advances in semi-supervised object detection (SSOD) [16, 29, 44, 47, 59] has yielded notable progress in practical applications. These SSOD methods apply a training framework consisted of two detection networks with the same configurations, acting as teacher and student, respectively. The teacher network is in charge of generating pseudo-labels to optimize

*Corresponding Author

the student during training, which can exploit massive unlabeled data based on a small amount of labeled information. With the help of this effective training paradigm, the latest SSOD method [37] can even achieve fully supervised performance with only 40% and 25% labeled information on MSCOCO [25] and PASCAL VOC [7], respectively.

However, directly transferring this successful paradigm to REC still suffers from two main challenges due to task gaps. The first one is the extremely sparse supervision signals. In contrast to object detection, REC only ground one instance for each text-image pair. This prediction pattern makes the REC model receive much fewer pseudo-supervisions than SSOD during teacher-student learning, i.e., only one bounding box without class pseudo-labels. For instance, Compared with SSOD that has 6-15 high-quality pseudo-boxes for each image [30, 37], semi-REC only has 0.5 box on average at the infant training stages as shown in Fig. 1. The sparse supervision signals also lead to worse pseudo-label quality, which is the second challenge. SSOD methods [16, 29, 44, 47, 59] can apply NMS [10] and high-threshold filtering to discard the vast majority of noisy pseudo-labels, thereby avoiding the error accumulation issue [29, 37, 44] in SSL. But in REC, a strong filtering is not feasible due to the already sparse pseudo-label information. This results in that most pseudo-labels of REC are of much lower-quality.

Based on this observations, we propose the first semi-supervised approach for REC called RefTeacher with two novel designs, namely Attention-based Imitation Learning (AIL) and Adaptive Pseudo-label Weighting (APW). In principle, RefTeacher also adopts a teacher-student framework, where the teacher predicts the pseudo bounding boxes for the student according to the given expressions. Following the latest SSOD [29, 30, 37, 47], we also use EMA to update the gradients of the teacher network from the student, and introduce data augmentation and in-strategies to improve SSL. To enrich the supervision signals, the proposed AIL helps the student imitate the attention behaviors of the teacher, thereby improving the knowledge transferring. APW is further used to reduce the impact of noisy pseudo-labels, which is achieved via adaptively weighting label information and the corresponding gradient updates.

To validate RefTeacher, we apply RefTeacher to two representative REC models, i.e. RealGIN [60] and TransVG [5], and conduct extensive experiments on three REC benchmark datasets, namely RefCOCO [54], RefCOCO+ [54] and RefCOCOg [38]. Experimental results show that RefTeacher can greatly exceed the supervised baselines, e.g. +18.8% gains on 10% RefCOCO. More importantly, using only 10% labeled data, RefTeacher can help RealGIN achieve near 100% fully supervised performance.

Overall, the contributions of this paper are three-fold:

- We present the first attempt of semi-supervised learn-

ing for REC with a strong baseline method called RefTeacher.

- We identify two challenges of semi-supervised REC, i.e. sparse supervision signals and worse pseudo-label noise, and address them with two novel designs, namely Attention-based Imitation Learning (AIL) and Adaptive Pseudo-label Weighting (APW).
- RefTeacher achieves significant performance gains on RefCOCO, RefCOCO+, and RefCOCOg datasets over the fully supervised methods.

2. Related Work

Referring Expression Comprehension. Referring Expression Comprehension (REC) [5, 12, 13, 19, 27, 33, 36, 49, 51, 53–55, 57, 60] aims to ground the referent in an image according to a natural language expression. Early REC works [12, 27, 36, 49, 53, 54, 57, 61] mainly followed the two-stage pipeline. Concretely, these two-stage methods first generate candidate proposals in an image, and then select the one that best matches with the expression as the referent. More recently, one-stage methods [5, 15, 19, 51, 60] have attracted more attentions due to its superior efficiency. By omitting the step of region proposals, one-stage REC models can achieve real-time inference [60]. To improve the performance of one-stage REC, researchers have put numerous efforts on the study of vision-and-language fusion [15, 45, 51, 52, 60] and detection paradigm [5, 19, 33], which have achieved significant progresses.

Semi-Supervised Learning. Semi-Supervised Learning (SSL) has made great progresses in image classification [2, 3, 14, 43, 46, 56, 58], which can be roughly divided into two categories, i.e. consistency-based [2, 3, 8, 16, 43] and pseudo-label based methods [1, 29, 30, 37, 43, 44, 47, 59]. In particular, consistency-based approaches [2, 3, 8, 16, 43] propose regularization loss to ensure that the teacher and student with different noisy input can have the consistency predictions. Pseudo-label based methods [1, 29, 30, 37, 44, 47, 59] use a well-trained teacher network to predict pseudo-labels for unlabeled data to optimize the student one. Inspired by these progresses, some recent works [16, 17, 29, 44, 59] also apply SSL to object detection. One representative work of semi-supervised object detection (SSOD) is STAC [44]. It first trains a teacher network using a small amount of labeled data and then produces pseudo-labels for unlabeled data. After that, a student network is trained with both labeled and unlabeled data. Based on this paradigm, a bunch of SSOD approaches [16, 17, 29, 30, 37, 44, 59] have been proposed recently and have achieved great success.

Figure 2. The overall semi-supervised REC framework of RefTeacher. RefTeacher includes two REC networks with the same configurations, namely Teacher and Student. The teacher will predict pseudo-labels of unlabeled data, based on which the student is trained together with a few labeled samples. The teacher is updated via EMA [48] from the student. Attention-based Imitation Learning (AIL) and Adaptive Pseudo-label Weighting (APW) are further deployed to address sparse supervision signals and extreme noisy pseudo-labels.

3. Method

3.1. Task Definition

Given a set of labeled data $\mathcal{D}_s = \{f(I_i^s; T_i^s); Y_i^s\}_{i=1}^{N_s}$ and unlabeled data $\mathcal{D}_u = \{f(I_i^u; T_i^u)\}_{i=1}^{N_u}$, the target of semi-supervised REC can be formulated as

$$\min L(\cdot; \mathcal{D}_s; \mathcal{D}_u): \quad (1)$$

Here, L is the semi-supervised objective function. f denotes the REC model, I , T and Y denote the images, expressions and box annotations, respectively. In practice, the numbers of labeled and unlabeled data meet $N_s \ll N_u$.

Since REC is a language-guided visual recognition task, the absence of text information will make the model unable to make predictions. Besides, such image-level captions are relatively easy to obtain from existing vision-language datasets [21, 39] or online data. In this case, we only discard the bounding box annotations and keep the text captions in our semi-supervised setting.

3.2. RefTeacher

3.2.1 Overall Framework

We first introduce the overall framework of RefTeacher for semi-supervised REC. As shown in Fig. 2, our semi-supervised framework consists of two REC networks with the same configuration, acting as Teacher and Student, respectively.

During training, the teacher network is used to predict pseudo-labels for unlabeled data, and then the student is optimized with both ground-truth and pseudo labels.

In this case, the overall optimization objective for the student is defined by

$$L = L_{\text{sup}} + \alpha L_{\text{unsup}}; \quad (2)$$

where α is a hyper-parameter that controls the weight of the unsupervised loss. L_{sup} is the loss function of REC models, which can be defined by

$$L_{\text{sup}} = \sum_i^N c_i L_{\text{box}}(b_i; \hat{b}_i) + L_{\text{conf}}(c_i; \hat{c}_i); \quad (3)$$

Here, L_{box} and L_{conf} are the regression and confidence losses [33, 51, 52, 60], respectively, and \hat{b} and \hat{c} are the predicted box coordinates and confidence, while b and c are the corresponding ground truths. N denotes the number of the predicted bounding boxes. The unsupervised loss L_{unsup} will be described in detail later.

Teacher-student mutual learning. To improve the efficiency of semi-supervised REC, we also introduce novel designs in SSL [29, 30, 37, 43, 44, 50] to RefTeacher, such as burn-in stage [29, 30, 37], Exponential Moving Average [48] and Strong-weak data augmentations [29, 30, 37].

Concretely, before teacher-student learning, the teacher network will first be trained with labeled data for a short period, which is called burn-in stage [29, 30, 37]. This setup can enable the teacher with a certain detection ability to provide viable pseudo-labels. Afterwards, the parameters of the teacher network are used to initialize the student one.

During semi-supervised training, the student is optimized by both labeled and unlabeled data, as defined in Eq. 2. In contrast, gradient backpropagation is prohibited in the

Algorithm 1 Algorithm of RefTeacher

Input: Labeled data $f(I^s; T^s); Y^s$, unlabeled data $f(I^u; T^u)$, burn-in step k , maximum iteration N

Output: Teacher Model θ_t^i

```

1: for  $i < N$  do
2:   if  $i < k$  then
3:     Update  $\theta_s^i$  by Eq. (3)
4:   if  $i = k$  then
5:     Initialize  $\theta_t^i$  with  $\theta_s^i$ 
6:   if  $k > N$  then
7:     Predict pseudo-labels  $\hat{y}_t^i$  and  $\hat{y}_s^i$  by Teacher  $\theta_t^i$ 
8:     Calculate  $L_{sup}$  with  $(x_t^i; t_t^i; y_t^i)$  by Eq. (3)
9:     Calculate  $L_{imi}$  by Eq. (6)
10:    Calculate  $L_{unsup}$  with  $(x_u^i; t_u^i; y_u^i)$  by Eq. (7)
11:    Update  $\theta_t^i$  and  $\theta_s^i$  by Eq. (4) and Eq. (8)
12: return  $\theta_t^i$ 
  
```

teacher, and its parameters are updated from the student via EMA [48]:

$$\theta_s^i = \alpha \theta_s^{i-1} + (1 - \alpha) \theta_t^i, \quad (4)$$

where θ_t and θ_s are the parameters of the teacher and student, respectively, i denotes the training step, α is the keeping rate. This optimization scheme can prevent the teacher from easily overfitting limited labeled data [48].

In addition, we also apply strong and weak data augmentation schemes to the student and the teacher, respectively. In existing SSL methods [29, 30, 37, 43, 44, 50], this training setup is widely used to prevent the student from overfitting redundant pseudo-labels. However, we also notice that some augmentation techniques, GaussianBlur, will change the semantic consistency between the image and expression, thus leading to negative examples. In this case, we combine RandomHorizontalFlip, RandomCrop and ColorJitter [22] as the strong augmentation. And RandomHorizontalFlip is retained as weak augmentation for the semantic consistency between teacher and student.

Overall, the complete procedure of RefTeacher is described in Algorithm 1.

3.2.2 Semi-supervised Learning

As shown in Eq. 2, semi-supervised REC consists of two main optimization targets, i.e., the supervised and the unsupervised ones. In terms of the unsupervised loss, a natural choice is to use the predicted pseudo bounding boxes to supervise the student network, which is also widely used in existing semi-supervised object detection (SSOD) [16, 17, 29, 30, 37, 44, 59]. It can be defined by

$$L_{box} = \sum_i^N L_{conf}(\hat{q}_i; q_i); \quad (5)$$

where \hat{q}_i denotes the pseudo-labels predicted by the teacher. Following the recent progress in SSL [29, 37], we ignore the box regression in L_{box} , which can be achieved based merely on L_{sup} [14, 29, 37].

However, only using L_{box} as the unsupervised loss will make the student cannot be sufficiently trained. To explain, in SSOD, the detection network often predicts a set of pseudo-boxes for teacher-student learning, where filtering rules can also be applied to select pseudo labels of high quality. In stark contrast, the REC model can only provide one pseudo-box for each expression, which also makes label filtering infeasible.

This case results in two key issues for semi-supervised REC. The first one is that the pseudo-label information is not sufficient for teacher-student learning. For instance, the student can only learn location information from the teacher, while other cross-modal information is in absence. Meanwhile, without sufficient pseudo-labels and filtering process, the supervised information used is often very noisy, as illustrated in Fig. 1.

Attention-based Imitation Learning (AIL). We first propose AIL to address the issue of sparse supervision signals. Its main principle is to let the student learn more comprehensive knowledge from the teacher, including cross-modal alignments and prediction patterns.

To learn the cross-modal behavior of the teacher, we introduce an attention-based imitation loss, termed L_{imi} . Specifically, given the input image $I \in \mathbb{R}^{H \times W \times 3}$ and text $T \in \mathbb{R}^l$, existing one-stage REC approaches [5, 15, 19, 33, 45, 51, 52, 60] first extract features of two modalities, which are fused to obtain multi-modal features $F \in \mathbb{R}^{h \times w \times d}$. For recently proposed one-stage REC models, attention modules are usually applied based on F to facilitate the vision-language alignment, which can be denoted by $F_{att} \in \mathbb{R}^{h \times w \times d}$. L_{imi} aims to encourage the consistency of F_{att} between the teacher and student, which is denoted as

$$L_{imi} = \sum_{i=1}^N \frac{1}{n} (F_{att}^i - F_{att}^i)^2; \quad (6)$$

Here, $n = h \times w \times d$, and F_{att}^t and F_{att}^s denote the attention features of the teacher and student, respectively.

Adaptive Pseudo-label Weighting (APW) AIL can effectively improve the efficiency of semi-supervised optimization. However, the numerous noisy pseudo-labels still greatly hurt the effectiveness of SSL. To address this issue, one direct solution from SSOD is to filter the low-confidence samples in SSL. However, this also causes a lot of useful training samples to be discarded, which inevitably hurts the training process.

To this end, we propose APW, which uses the predicted confidence to weight the unsupervised loss of each sample. In particular, APW can be defined as

$$L_{\text{box}}^{\text{apw}} = \sum_i^N c_i L_{\text{conf}}(c_i; c_i) : \quad (7)$$

When the teacher predicts an uncertain pseudo-label usually of low-confidence, APW will adaptively lower the contribution of its unsupervised loss. By this way, APW can greatly alleviate the error accumulation issue.

Based on these two designs, the final optimization objective for semi-supervised REC can be re-written by

$$L = L_{\text{box}}^{\text{sup}} + L_{\text{conf}}^{\text{sup}} + \alpha_u L_{\text{box}}^{\text{apw}} + \alpha_{\text{imi}} L_{\text{imi}} : \quad (8)$$

Here, α_u and α_{imi} are hyper-parameters to adjust the contributions of loss.

4. Experiment

4.1. Datasets and Metric

RefCOCO & RefCOCO+ [54] contain 14k referring expressions for 50 bounding boxes in 10 images from MS-COCO [25]. There are four splits in RefCOCO and RefCOCO+, i.e., train, val, testA and testB. The expressions of RefCOCO are mainly about absolute position, while the ones of RefCOCO+ contain more contents about attributes. RefCOCOg [38] has 10k expressions for 54 objects from 26k images. By default, we use its 1MD split [38], which contains three splits, i.e., training, validation and testing. Compared to RefCOCO and RefCOCO+, the expressions of RefCOCOg are more complex and longer, which have 8.4 words on average and contain both attribute and localization information. For semi-supervised learning, we randomly select 0.1%, 1%, 5%, 10% samples as the labeled data, and the rest are used for the unlabeled data.

Intersection-over-Union (IoU) is the metric used in REC, which measures the overlap degree between the prediction and ground truth. Following previous works [33, 51, 52, 60], we use IoU@0.5 to evaluate the model performance.

4.2. Experimental Settings

4.2.1 Implementation Details

All models are trained by Adam [6] with a constant learning rate of $1e-4$. The batch size is set to 16, which consists of 8 labeled and 8 unlabeled image-text pairs. The total training steps are 70k where the burn-in steps are 6k by default and 7.5k for 10% semi-supervised settings. The hyper-parameters of α_{imi} and α_u are set to 0.9996, 0.05 and 0.5, respectively. For strong-weak augmentations we use RandomResize, RandomSizeCrop, RandomHorizontalFlip, ColorJitter, AugTranslate as strong augmentations, and RandomHorizontalFlip as weak augmentations.

For other network configurations, such as input resolutions and visual backbones, we follow the default settings of previous works [5, 60].

We use the fully supervised method and SSOD approach, i.e., STAC [44], as the baselines. For all supervised baselines, we use RealGIN [60] as the REC model. For the fully supervised method, we follow the default setups of RealGIN to train model on available labeled data. For STAC, the training can be divided into two stages. Firstly, we train a teacher model on the labeled data, which is used to predict the pseudo-labels for unlabeled data. Then, a student model is trained with both labeled and unlabeled data. Following STAC, we also apply strong data augmentations to the input images of the student. For all the semi-supervised methods, both teacher and student share the same REC network.

4.3. Experimental Results

4.3.1 Comparison with Baselines

Results of RefTeacher and baselines in Tab. 1, we first compare RefTeacher with the fully supervised baseline and STAC. The first observation is that all semi-supervised approaches can obtain obvious gains over the fully supervised baseline. STAC outperforms the ‘‘Supervised’’ by 5.8% on 1% RefCOCOval. Compared with STAC, the performance gains by RefTeacher are more significant, +19.5% on 1% RefCOCOval. With only 10% labels, RefTeacher can even achieve comparable performance to the 100% fully supervised result, i.e., 72.22vs 75.00. Besides, the benefits of RefTeacher are also obvious to the expressions about objects, i.e., +23.02% on 1% RefCOCO. On RefCOCOg, STAC performs closely to the fully supervised method, while RefTeacher brings consistent gains on different proportions of labeled data, i.e., +17.16% on 1% RefCOCOg val.

In Fig. 3, we compare the training curves of RefTeacher and baselines. It can be seen that the fully supervised baseline and STAC will not further improve after 10k training setups, while RefTeacher can consistently benefit the model during the whole training period. These results greatly validate the effectiveness of RefTeacher.

Results of RefTeacher using additional unlabeled data. In Tab. 2, with the base model, i.e., RealGIN we train RefTeacher with 100% labeled data of RefCOCO and examine whether additional unlabeled data can further improve the performance. From this table, we can see that the effectiveness of RefTeacher is still obvious on 100% RefCOCO. By using RefCOCO+ as unlabeled data, RefTeacher can outperform the supervised one, +4.1% on RefCOCOval. We also use the out-of-domain data as the unlabeled data, i.e., ReferIt, the performance gains by RefTeacher can be +2.94% on RefCOCO val. More im-

¹The images in RefCOCO are removed.

Table 1. Comparison of RefTeacher and baselines on RefCOCO, RefCOCO+ and RefCOCOg. For all approaches, we use RealGIN [60] as the REC model. “Supervised” denotes the fully supervised training. STAC [44] is the semi-supervised approach from SSOD.

Methods	RefCOCO											
	0.1%			1%			5%			10%		
	val	testA	testB	val	testA	testB	val	testA	testB	val	testA	testB
Supervised	14.64	20.31	9.36	37.93	40.02	33.09	53.90	55.68	49.75	58.63	59.52	55.66
STAC [44]	18.68	22.11	14.27	43.69	47.75	37.29	58.14	59.64	53.64	62.35	64.56	58.06
RefTeacher	34.05	35.41	30.25	59.25	60.47	56.11	68.96	71.04	63.18	72.22	74.47	66.69

Methods	RefCOCO+											
	0.1%			1%			5%			10%		
	val	testA	testB	val	testA	testB	val	testA	testB	val	testA	testB
Supervised	13.21	17.22	8.14	26.11	27.70	20.80	34.53	36.33	27.51	37.27	40.38	33.18
STAC [44]	17.96	22.13	11.45	30.48	32.68	23.71	38.88	40.97	31.70	40.83	43.42	34.51
RefTeacher	22.10	28.66	12.76	39.45	41.95	32.17	49.47	53.27	41.52	52.50	56.76	44.69

Methods	RefCOCOg							
	0.1%		1%		5%		10%	
	val-u	test-u	val-u	test-u	val-u	test-u	val-u	test-u
Supervised	16.50	16.21	26.86	26.36	39.44	37.94	44.71	45.02
STAC [44]	18.28	18.18	31.52	30.77	43.18	41.65	47.86	47.74
RefTeacher	29.06	29.76	44.02	42.13	51.20	50.79	51.20	56.80

Table 2. Results of RefTeacher on RefCOCO with additional data. We use RefCOCO as the labeled data and other datasets, RefCOCO+, as the unlabeled data

Model	Unlabeled	val	testA	testB
FAOA [52]	-	72.54	74.35	68.50
ReSC [51]	-	77.63	80.45	72.30
MCN [33]	-	80.08	82.29	74.98
Iter-Shrinking [45]	-	-	74.27	68.10
LBYL-Net [15]	-	79.67	82.91	74.15
TransVG [5]	-	81.02	82.72	78.35
	-	75.00	77.53	69.17
RefTeacher _{RealGIN} [60]	RefCOCO+	79.06	80.80	72.80
	RefCOCOg	79.00	81.51	73.64
	ReferIt	77.94	81.24	72.33

portantly, with unlabeled data, RefTeacher can help RealGIN achieve competitive performance against the SOTAs [5, 15, 33, 45, 51, 52].

our RefTeacher can help the REC model RealGIN exploit additional unlabeled samples to obtain competitive performance against the SOTA REC methods.

4.3.2 Ablation Study

The impact of semi-supervised framework. In Table 3, we first ablate the semi-supervised REC framework of RefTeacher under the setting of 5% labeled data. We can see that the pseudo-label learning can bring obvious performance gains on three REC datasets, +4.2% on 5% RefCOCO. Then, the teacher-student mutual training can also improve the performance. With the semi-supervised optimization, e.g., strong-weak augmentations, the perfor-

Figure 3. Training curves of RefTeacher and baselines. RefTeacher can consistently improve performance throughout the whole training period more effective than baselines.

formance on 5% RefCOCO boosts from 58.9 to 69.0. Such gains also validate the issue of worse pseudo-labels, which is alleviated by the semi-supervised optimization.

The impact of APW and AIL. In Tab. 4, we ablate the effectiveness of APW and AIL in RefTeacher. It can be seen that both designs can improve performance, and AIL can bring more gains than APW, e.g., +3.7% on 1% RefCOCOg. When combining APW with AIL, the performance of RefTeacher can further improve from 58.2 to 59.2 on 1% RefCOCO. These results well validate the effectiveness of APW.

Comparison of APW and AIL with other designs. In Tab. 5, we compare APW and AIL with other viable solutions for worse pseudo-labels and sparse supervision significantly. For worse pseudo-labels, confidence Itering is a direct solution. Listen2student [30] is an approach of SSOD,

Table 3. Ablation study of semi-supervised REC framework on 5% RefCOCO [54], RefCOCO+ [12] and RefCOCOg [38].

Settings	Ref	Ref+	Refg
Supervised	53.9 ₀	34.5 _{0.0}	39.4 _{0.0}
+ pseudo-label learning	58.1 ₀	38.9 _{4.4}	43.2 _{3.8}
+ T-S mutual training	58.9 ₀	39.1 _{4.6}	44.8 _{5.4}
+ semi-supervised opt.	69.0 _{15.1}	49.5 _{15.0}	51.2 _{11.8}

Table 4. Ablation study of adaptive pseudo-label weighting (APW) and attention-based imitation learning (AIL) on 1% RefCOCO, RefCOCO+ and RefCOCOg.

APW	AIL	Ref	Ref+	Refg
		55.4 _{0.0}	38.1 _{0.0}	40.2 _{0.0}
X		56.6 _{1.2}	39.0 _{0.9}	42.6 _{2.4}
	X	58.2 _{2.8}	39.7 _{1.6}	43.9 _{3.7}
X	X	59.2 _{3.8}	39.4 _{1.3}	44.0 _{3.8}

Table 5. Comparison of different methods for pseudo-label filtering and dense learning on 1% RefCOCO, RefCOCO+ and RefCOCOg. Our settings are in gray.

Settings	Ref	Ref+	Refg
Baseline	55.4 ₀	38.1 _{0.0}	40.2 _{0.0}
Worse pseudo-labels:			
condence ltering	55.9 _{0.5}	39.0 _{0.9}	41.1 _{0.9}
listen2student [30]	55.2	38.5 _{0.4}	40.0 _{0.2}
APW	56.6 _{1.2}	39.0 _{0.9}	42.6 _{2.4}
Sparse supervision signal:			
soft label training	55.8 ₄	38.2 _{0.1}	40.3 _{0.1}
AIL	58.2 _{2.8}	39.7 _{1.6}	43.9 _{3.8}

where the low-quality pseudo-labels are ignored after the comparisons between the teacher and student. Both approaches can reduce the training samples, so their performance gains are not obvious. Instead, our APW can effectively use all training samples and also reduce the impact of noisy pseudo-labels, e.g., +2.4% on 1% RefCOCOg. For sparse supervision signal, we compare AIL with the soft label training, which replaces the one-hot ground truth with a soft distribution. From the results we can see that AIL greatly exceeds the soft label training by +3.7 on 1% RefCOCOg, suggesting more informative supervision provided by attention-based imitation. These results validate AIL and APW again.

4.3.3 Generalization Experiments

Results of RefTeacher on more REC models We generalize RefTeacher to more REC models, Transformed-based model and CNN-based model with anchor free head. Considering the absence of condence in TransVG [5], we only apply AIL to it and directly learning regression on RefCOCO dataset. Our proposed AIL significantly improves

Table 6. Results of RefTeacher on different REC models on 10% labeled data.

Models	Settings	Ref	Ref+	Refg
RealGIN [60]	Supervised	58.6	37.3	44.7
	RefTeacher	72.2	52.5	56.5
TransVG [5]	Supervised	67.2	43.7	47.9
	RefTeacher	70.3	46.4	51.0
SimREC [32]	Supervised	69.9	53.9	54.1
	RefTeacher	73.5	57.6	57.7

Table 7. Generalization of RefTeacher to unsupervised settings. Pseudo-Q can generate pseudo-expressions and pseudo-boxes for images. For RefTeacher, we use 40% samples annotated by Pseudo-Q as the labeled data, and the rest are unlabeled data.

Methods	Ref	Ref+	Refg
Pseudo-Q [18]	52.09	32.05	46.61
Pseudo-Q+RefTeache	54.20	32.94	48.04

the performance. With regard to anchor-free methods, we deploy both the condence and attention information to use AIL and APW. The gains on RefCOCO+ is up to 3.7%, proving the effectiveness of RefTeacher.

Results of RefTeacher under unsupervised settings. In Tab. 7, we generalize RefTeacher to unsupervised settings. Specifically, we use an unsupervised approach, Pseudo-Q [18], to generate pseudo-expressions and boxes for images of RefCOCO, RefCOCO+ and RefCOCOg. For RefTeacher, we only use 40% samples with box annotations as the labeled data, and the rest are unlabeled data. From Tab. 7, we can see that RefTeacher can also benefit

+2.1 % against Pseudo-Q on 1% RefCOCO. Since unsupervised training also faces the challenge of worse pseudo-labels and sparse training signal, we believe RefTeacher can also be a good complement to existing unsupervised REC methods.

4.3.4 Qualitative analysis

To gain more insights into RefTeacher, we conduct extensive visualizations in Fig. 4. From Fig. 4 (a), we observe that the predictions of RefTeacher are more accurate than the fully supervised method on small objects and complex expressions, e.g., Exp-2 and Exp-3. In Fig. 4 (b), we further compare the pseudo-labels with and without AIL and APW. From these visualizations, we can see that both designs can obviously improve the quality of pseudo-labels. Besides, AIL can generate more accurate pseudo-labels in crowded scene, greatly validating the benefits of the attention imitation learning. We also visualize the failure cases in Fig. 4 (c). We observe that RefTeacher tends to fail when the objects are occluded, e.g., Exp-11, or the expressions are very abstract, e.g., Exp-13.

Figure 4. Visualizations of RefTeacher and fully supervised baselines. Sub-figure (a) shows the predictions of RefTeacher are much better than the fully-supervised baseline. Sub-figure (b) indicates that both AIL and APW of RefTeacher can obviously improve the quality of pseudo-labels. Sub-figure (c) demonstrates that RefTeacher still fails in some hard examples.

5. Conclusion

6. Acknowledgements

In this paper, we present the first semi-supervised framework for referring expression comprehension (REC), Program of China (No.2022ZD0118201), the National Natural Science Foundation of China (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).

This work was supported by National Key R&D Program of China (No.2022ZD0118201), the National Natural Science Foundation of China (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).

Extensive experiments validate the superiority of RefTeacher than the fully and semi-supervised baselines.

References

- [1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems* 37, 2014. 1, 2
- [2] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785* 2019. 1, 2
- [3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems* 32, 2019. 1, 2
- [4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 1
- [5] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 2, 4, 5, 6, 7
- [6] John Duchi, Elad Hazan, and Yoram Singer. Adaptive sub-gradient methods for online learning and stochastic optimization. *Journal of machine learning research* 12(7), 2011. 5
- [7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88(2):303–338, 2010. 2
- [8] Mingfei Gao, Zizhao Zhang, Guo Yu, Serca Ar k, Larry S Davis, and Tomas P ster. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pages 510–526. Springer, 2020. 1, 2
- [9] Ross Girshick. Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1
- [10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1, 2
- [11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1
- [12] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1115–1124, 2017. 2, 7
- [13] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4555–4564, 2016. 2
- [14] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. Simple: similar pseudo label exploitation for semi-supervised classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15099–15108, 2021. 2, 4
- [15] Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. Look before you leap: Learning landmark features for one-stage visual grounding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16888–16897, 2021. 2, 4, 6
- [16] Jisoo Jeong, Seungeui Lee, Jeessoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems* 32, 2019. 1, 2, 4
- [17] Jisoo Jeong, Vikas Verma, Minsung Hyun, Juho Kannala, and Nojun Kwak. Interpolation-based semi-supervised learning for object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11602–11611, 2021. 2, 4
- [18] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15513–15523, 2022. 7
- [19] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 2, 4
- [20] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 1
- [21] 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. 3
- [22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM* 60(6):84–90, 2017. 4
- [23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1
- [24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1
- [25] 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. 2, 5
- [26] 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
- [27] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864, 2017. 2
- [28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. *European conference on computer vision*, pages 21–37. Springer, 2016. 1
- [29] 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. 1, 2, 3, 4
- [30] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, 2022. 2, 3, 4, 6, 7
- [31] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1274–1282, 2020. 1
- [32] 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. 7
- [33] 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, 3, 4, 5, 6
- [34] 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
- [35] 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
- [36] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7102–7111, 2017. 2
- [37] 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. 2, 3, 4
- [38] 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, 5, 7
- [39] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1, 3
- [40] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1
- [41] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 1
- [42] 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
- [43] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 1, 2, 3, 4
- [44] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 1, 2, 3, 4, 5, 6
- [45] Mingjie Sun, Jimin Xiao, and Eng Gee Lim. Iterative shrinking for referring expression grounding using deep reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14060–14069, 2021. 1, 2, 4, 6
- [46] Hui Tang and Kui Jia. Towards discovering the effectiveness of moderately confident samples for semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14658–14667, 2022. 2
- [47] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3132–3141, 2021. 1, 2
- [48] 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. 3, 4
- [49] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1960–1968, 2019. 1, 2
- [50] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020. 3, 4
- [51] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-

- query construction. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020. 1, 2, 3, 4, 5, 6
- [52] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. *Proceedings of the IEEE/CVF International Conference on Computer Vision* pages 4683–4693, 2019. 1, 2, 3, 4, 5, 6
- [53] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pages 1307–1315, 2018. 1, 2
- [54] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016b. 2, 5, 7
- [55] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pages 7282–7290, 2017. 2
- [56] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems* 34:18408–18419, 2021. 2
- [57] Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, and Honglak Lee. Discriminative bimodal networks for visual localization and detection with natural language queries. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pages 557–566, 2017. 2
- [58] Zhen Zhao, Luping Zhou, Yue Duan, Lei Wang, Lei Qi, and Yinghuan Shi. Dc-ssl: Addressing mismatched class distribution in semi-supervised learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 9757–9765, 2022. 2
- [59] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 4081–4090, 2021. 1, 2, 4
- [60] 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, 3, 4, 5, 6, 7
- [61] 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* 2019. 2
- [62] 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. *Proceedings of the IEEE/CVF International Conference on Computer Vision* pages 2074–2084, 2021. 1