2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Weakly Supervised Referring Expression Grounding via Target-Guided Knowledge Distillation

Jinpeng Mi[1], Song Tang[1], Zhiyuan Ma[1], Dan Liu[1], Qingdu Li[1], Jianwei Zhang[2]

Abstract— **Weakly supervised referring expression grounding aims to train a model without the manual labels between image regions and referring expressions during the training phase. Current predominant models often adopt deep structures to reconstruct the region-expression correspondence. A crucial deficiency of the existing approaches lies in that these models neglect to exploit potential valuable information to further improve their grounding performance. To address this issue, we leverage knowledge distillation as a unique scheme to excavate and transfer helpful information for acquiring a better model. Specifically, we propose a target-guided knowledge distillation framework that accounts for region-expression pairs reconstruction and matching. We reactivate the target-related prediction information learned by a pre-trained teacher model and transfer the target-related prediction knowledge from the teacher to guide the training process and boost the performance of the student model. We conduct extensive experiments on three benchmark datasets, i.e., RefCOCO, RefCOCO+, and RefCOCOg. Without bells and whistles, our approach achieves state-of-the-art results on several splits of benchmark datasets. The implementation codes and trained models are available at: https://github.com/dami23/WREG_KD.**

I. INTRODUCTION

Referring expression grounding (REG) aims to ground target regions in images according to given referring expressions. REG is one of the core tasks of artificial intelligence [1], which can be used to test the machine's ability to understand natural language and visual scenes. As the bridge to connect vision detection and natural language processing, REG has a wide range of practical applications, such as visual question answering [2], image retrieval [3], human-robot interaction [4], [5], etc.

The existing work of REG can be divided into fully supervised and weakly supervised methods. The performance of fully supervised approaches is extremely conferred by the large volume of manually labeled datasets, which demand exhaustive annotations of bounding boxes, referring expressions, and corresponding mapping between each bounding box and referring expressions. In order to reduce the cost and labor intensity of the manual annotations and extend the practical applications of REG, plenty of weakly supervised methods [6], [7], [8], [9], [10] have been proposed. The primary motivation of these weakly supervised approaches is to reconstruct the mapping between image regions and referring expressions. These approaches build the mapping via utilizing off-the-shelf information, such as linguistic structure [7], semantics in visual regions features [8], and semantic similarities between image regions [9]. Although these models achieve promising results, they do not attempt to further exploit concealed helpful information to improve the model performance.

Knowledge distillation [11] aims to transfer helpful knowledge from a well-trained teacher model to generate a better student model, and knowledge distillation can be deemed a unique scheme to excavate and transfer helpful information to boost the model performance. Inspired by the primary function of knowledge distillation, we introduce knowledge distillation into weakly supervised REG to explore potential valuable information.

In addition, the language-guided models that mine the effect of expression components for describing target objects acquire better performance on several vision-language tasks, such as REG [12], relational reasoning [13], 3D visual grounding [14]. Such cases inspire us to excavate the advantages of target-related information for weakly supervised REG. Moreover, motivated by the salient attribute of knowledge distillation and the role of referring expression components for grounding target objects, we exploit the benefits of target-related information from the perspective of knowledge distillation.

In this paper, we propose a knowledge distillation-based architecture to address weakly supervised REG. Specifically, we first train a teacher model to learn the region-expression mapping reconstruction and matching. We then reactivate the target-related prediction information learned by the teacher model and distill the target-related prediction knowledge from the teacher to lead the training process of the student model, and further improve the grounding performance of the student. It is worth noting that this work is the first attempt to facilitate weakly supervised REG with knowledge distillation.

The main contributions of this paper are summarized as follows:

- We leverage knowledge distillation as a unique scheme to exploit and transfer valuable information for weakly supervised REG.
- We propose a target-guided knowledge distillation architecture that reactivates the target-related prediction knowledge learned by the teacher model and adopts knowledge distillation as a channel to guide the training process for achieving a better student model.
- We validate the proposed architecture on the benchmark datasets, RefCOCO [15], RefCOCO+ [15], and RefCOCOg [16]. Our proposed approach achieves a

[1]Institute of Machine Intelligence (IMI), University of Shanghai for Science and Technology, China. [2]Technical Aspects of Multimodal Systems (TAMS), Department of Informatics, University of Hamburg, Germany. mi@informatik.uni-hamburg.de

new state-of-the-art (SOTA) on several splits of the benchmark datasets.

II. Related Work

A. Fully Supervised Referring Expression Grounding

Fully supervised REG methods train models using a large volume of manual mapping labels between image regions and referring expressions. The pioneer approaches of REG [15], [16] ground targets by learning the semantics relatedness between visual region features and textual representations. Based on these methods, plenty of models are introduced. According to the target object grounding pattern, the existing models can be divided into two-stage and one-stage paradigms. Two-stage models [5], [17], [18], [19], [20], [21], [22] first adopt a pre-trained object detection model, such as Faster R-CNN [23], to detect region candidates and extract deep features for the detected candidates, and then ground target objects via modeling the matching between referring expressions and image regions.

While most one-stage methods directly fuse the target grounding in the object detection. For example, Yang et al. [24] inject the textual representations of expressions into the YOLO v3 [25] to predict the target objects. Some newly proposed one-stage models improve the performance by introducing multiple strategies, such as augmenting the visual representations of candidate regions [12], [26], [27], mitigating the difference between the textual and visual representations [28], leveraging attention-based multimodal data fusion to acquire better target reasoning clues [29], etc.

B. Weakly Supervised Referring Expression Grounding

Weakly supervised REG methods learn to ground target objects without the manual annotations between region proposals and corresponding referring expressions. Thus the predominant training strategy is to reconstruct the region-expression mapping. For instance, Rohrbach et al. [6] learn the mapping by an attention mechanism, Niu et al. [30] leverage variational context to construct the relationship between visual regions and textual expressions, Liu et al. [8] utilize a hierarchical attention mechanism to build the corresponding relationship, and Liu et al. [9] draw support from the prior knowledge acquired from pre-trained Faster RCNN to model the mapping.

Different from the methods mentioned above, Zhang et al. [31] explore the benefits of counterfactual results in the training process by leveraging Counterfactual Contrastive Learning (CCL). Sun et al. [10] parse expressions into discriminative triads that describe the target objects, the related subjects, and the relationship between the targets and subjects, respectively. Liu et al. [32] improve the model performance via an entity augmentation strategy to filter the unrelated image regions. Sun et al. [33] attempt to generate a unique textual description for each region proposal and optimize the deviation issue during the training phase by a non-cyclic framework.

Unlike the aforementioned approaches, we further exploit the valuable information and leverage knowledge distillation as a unique channel to excavate the role of target-related prediction information for weakly supervised REG.

C. Knowledge Distillation

Knowledge distillation is initially introduced in [11], which transfers helpful knowledge from a well-trained teacher model to a student model in an interactive pattern. Existing knowledge distillation approaches [34], [35], [36] train the student network under the supervision of ground truths and the softened outputs of the pre-trained teacher network. In this way, the models' performance and training efficiency are significantly improved. Recently, plenty of approaches successfully employ knowledge distillation to tackle challenging tasks, such as object detection [37], semantic segmentation [38], neural machine translation [39], and so on.

Inspired by the primary function of knowledge distillation, we introduce knowledge distillation to weakly supervised REG. On the other hand, we adopt knowledge distillation as a unique strategy to exploit the valuable information in the pre-trained teacher model, and transfer the helpful knowledge to guide the training procedure of the student model as well as boost the performance of the student.

III. Proposed Approach

In this paper, we propose a target-guided knowledge distillation architecture to boost the weakly supervised REG. Specifically, we re-attend the target-related prediction information learned by the pre-trained teacher model, and leverage knowledge distillation as a unique scheme to excavate the effect of the target-related information. We then transfer the target-related prediction knowledge to achieve a better student model. The proposed architecture is independent of the backbone grounding framework and does not introduce additional model parameters. We illustrate the details of the proposed architecture in Figure 1.

A. Problem Formulation

REG aims to ground a target object r^* via modeling the relationship between a given referring expression E and the region r_i in an image I with N regions of interest $R = \{r_i\}_{i=1}^{N}$, that are detected by a pre-trained object detector, e.g., Faster-RCNN [23]. While the weakly supervised REG grounds r^* by learning the matching score between r_i and E without the mapping annotations between r_i and E during the training phase. The weakly supervised REG selects r_i with the maximum matching score as the target r^*:

$$r^* = \arg\max_{r_i \in R} \mathrm{S}(r_i, E) \tag{1}$$

where S represents the matching score learning operation.

B. Feature Encoding

In our experiments, we utilize the modular setting of DTMR [10] as the backbone. DTMR parses the expressions into discriminative triads, which represent the target objects, the related subjects, and the discriminative relationship between the target and the subject objects, respectively. Each

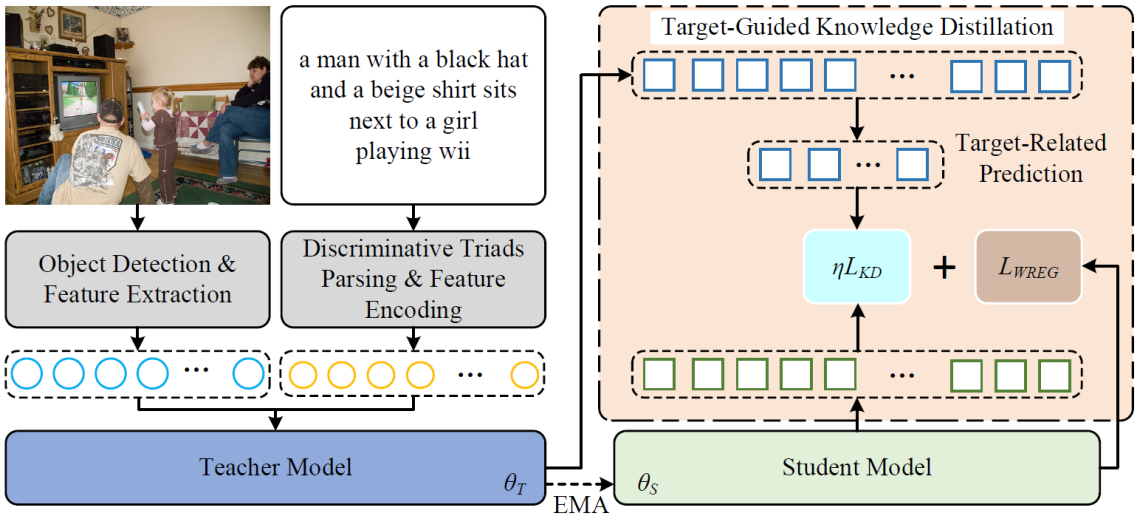

Fig. 1. Overview of the target-guided knowledge distillation for weakly supervised referring expression grounding. The target-related prediction information learned by the teacher model is reactivated and transferred as guidance for achieving a better student model. The student parameters are updated via exponential mean average (EMA) manner, and the final loss for training the student is the sum of weakly supervised referring expression grounding L_{WREG} and the knowledge distillation loss L_{KD} with trade-off coefficient η.

triad contains a target element u_t, a subject element u_s, and a relation element u_r. And then GloVe [40] is adopted to embed textual representations f_t^t , f_t^s, $f_t^r \in \mathbb{R}^{1\times 300}$ for u_t, u_s, and u_r, respectively.

For the visual images, we adopt Faster R-CNN [23] to detect candidate region r_i in each image I, and employ ResNet-101 [41] to extract the deep visual feature $f_v^i \in \mathbb{R}^{7\times 7\times 2048}$ for each r_i. In order to represent the spatial relationships between image regions, we follow the idea in [15] and employ a 5-D vector $f_s^i = [\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w\cdot h}{W\cdot H}]$ to encode the spatial relation feature for each r_i, where x_{tl}, y_{tl}, x_{br}, y_{br} denote the top left and bottom right position of r_i, W and H represent the width and height of I, and w and h are the width and height of r_i.

C. Teacher Model Training

DTMR grounds the target objects via triad-level reconstruction and matching. In order to better transfer the target-related prediction knowledge, we make modifications to the original triad-level matching of DTMR. Specifically, we model the matching score between the image region pair (r_i, r_j) and the three triad elements via:

$$\begin{aligned} s_i^t &= \text{Softmax}(w_2^t\Psi(w_1^t\Psi(f_v^i \oplus f_t^t) + b_1^t) + b_2^t) \\ s_j^s &= \text{Softmax}(w_2^s\Psi(w_1^s\Psi(f_v^j \oplus f_t^s) + b_1^s) + b_2^s) \\ s_{i,j}^r &= \text{Softmax}(w_2^r\Psi(w_1^r\Psi(\bar{f}_v \oplus f_t^r) + b_1^r) + b_2^r) \end{aligned} \tag{2}$$

where w_1^A, b_1^A, w_2^A, b_2^A, $A \in \{\text{t, s, r}\}$ are the parameters for the target, the subject, and the discriminative relationship matching learning module, respectively. Ψ represents the ReLU activation function, and $\oplus$ denotes the concatenation operation. f_v^j and f_s^j are the visual and spatial features of related subject r_j, and $\bar{f}_v = f_v^i \oplus f_s^i \oplus f_v^j \oplus f_s^j$.

In order to reconstruct the triad elements, we first calculate the weighted sum of the region features and the associated triad element matching scores:

$$\begin{aligned} g^t &= \sum_{i=1}^{N} s_i^t f_v^i \\ g^s &= \sum_{j=1}^{N} s_j^s f_v^j \\ g^r &= \sum_{i,j=1}^{N} s_{i,j}^r \bar{f}_v \end{aligned} \tag{3}$$

We then utilize a multi-layer perceptron (MLP) to generate the linguistic embeddings for the triad elements via:

$$\bar{e}^A = \text{MLP}(g^A) \tag{4}$$

For the target inference, we calculate the triad-level attention score Att_{triad}^T and select the candidate region with the maximum Att_{triad}^T as the grounded target r^*:

$$\begin{aligned} Att_{triad}^T(r_i) &= \delta_1 s_i^t + \delta_2 s_j^s + \delta_3 s_{i,j}^r \\ r^* &= \arg\max_{r_i \in R} Att_{triad}^T(r_i) \end{aligned} \tag{5}$$

We train the teacher model by minimizing the L_2 distances between the predicted triad elements and the original corresponding elements by:

$$L_{WREG} = \sum_{A\in t,s,r} \left\| f_t^A - \bar{e}^A \right\|_2^2 \tag{6}$$

D. Target-Guided Knowledge Distillation

The prevailing weakly supervised REG methods select the checkpoint with the best grounding accuracy on the validation set as the learned model. In contrast, we utilize knowledge distillation as a unique scheme to exploit concealed valuable information in the pre-trained teacher model, and distill the helpful knowledge from the teacher model

to guide the training routine for acquiring a better student model, which has the same structure as the teacher model. Concretely, we first reactivate the target-prediction information learned by the teacher, and then utilize knowledge distillation as a unique channel to transfer the target-related knowledge for leading the training process of the student to achieve better grounding performance.

In this paper, we aim to explore valuable information in the pre-trained teacher model, and improve the model performance by distilling the target-related prediction knowledge from the teacher to the student. Generally, the Kullback-Leibler (KL) divergence loss is utilized to measure the distance between the softened probability distributions of the teacher and the student. By comparison, the mean squared error (MSE) loss encourages the student to directly learn the logits from the teacher model. Thus, we select MSE to distill the target-related prediction knowledge. We reactivate the target attention score s_i^t in the pre-trained teacher model and distill the target-related prediction score from the teacher to the student via:

$$L_{KD} = \left\| \varphi s_i^t - Att_{triad}^S(r_i) \right\|_2^2 \tag{7}$$

where $Att_{triad}^S(r_i)$ is the aggregated triad-level attention score during the student model training, φ represents the hyper-parameter to balance the target-related prediction for the knowledge distillation.

We distill knowledge from the teacher to lead the training procedure of the student model. If the checkpoint obtains the best accuracy on the validation set, we select the checkpoint as the learned student model. The final loss for the training routine with target-guided knowledge distillation is defined as follows:

$$L_{final} = L_{WREG} + \eta L_{KD} \tag{8}$$

where η denotes the trade-off coefficient of the knowledge distillation.

Moreover, inspired by the Mean Teacher [42], which updates model parameters via exponential moving average (EMA) to generate ensemble models and improve training efficiency, we update the parameters of the student model via EMA manner. The target-guided knowledge distillation process is summarized in Algorithm 1.

Algorithm 1: Target-Guided Knowledge Distillation

Input: pre-trained teacher model T with parameters θ_T; total iteration number K; EMA decay weight γ
Initialization: training iteration count i = 0;
student model S with parameters θ_S
Att_{triad}^T, Att_{triad}^S : triad-level attention scores learned by the teacher T and the student S
s_i^t: the target attention score learned by T
def KD(t, s):
    t = t.detach()
    return (t-s).square().mean()
for *i in range K* **do**
    for r_i, *E in dataloader* **do**
        T = torch.load('pre-trained checkpoint')
        s_i^t, Att_{triad}^T = $T(r_i$, E)
        _, Att_{triad}^S = $S(r_i$, E)
        L_{final} = L_{WREG} + η KD(φs_i^t, Att_{triad}^S)
    update(S)
    $\theta_S \leftarrow \gamma\theta_T + (1-\gamma)\theta_S$

IV. Experiments

A. Datasets and Metric

We train and evaluate our proposed approach on RefCOCO [15], RefCOCO+ [15], and RefCOCOg [16]. RefCOCO comprises 19,994 images with 142,210 expressions for 50,000 referents, and RefCOCO+ includes 19,992 images with 141,564 expressions for 49,856 referents. While RefCOCOg contains 25,799 images with 95,010 expressions for 49,822 referents, the expressions in RefCOCOg are longer than those in RefCOCO and RefCOCO+.

We evaluate the grounding accuracy by calculating the Intersection over Union (IoU) between the predicted target region and the ground-truth bounding box, if the predicted region with IoU larger than 0.5 is regarded as a correct grounding.

B. Implementation Details

We set the hyper-parameters for training the teacher and the student models as follows. For the teacher training, as the target attention score plays the most critical role in grounding target objects, we set δ_1 = 2 and δ_2 = δ_3 = 1 in Equation (5) to calculate the triad-level attention score. For the knowledge distillation, we utilize φ = 2 to transfer the target prediction information from the teacher to the student. For the student model training, we adopt different values of η in Equation (8) to validate the effect of target-guided knowledge distillation for the student training and model performance improvement, and we elaborate on the results in the hyper-parameter analysis section. To update the student parameters, we set the decay weight γ in EMA to 0.9997.

We conduct all experiments on an NVIDIA GeForce GTX 3090Ti GPU. We adopt the Adam optimizer with an initial learning rate of 1.26e-5 to train the teacher and student models. The total iteration number for each training phase is up to 1500,000, and the training phase is around 10 hours on a single GPU.

C. Comparison with State-of-the-Art Models

In order to validate the performance of the proposed architecture, we compare the grounding accuracy acquired by our approach with the SOTA weakly supervised REG models, including VC [30], ARN [8], KPRN [9], IGN[31], EARN [32], DTMR [10], and Cycle-Free approach [33]. The listed approaches train their model by utilizing the ground truth bounding boxes, and some listed models employ different settings to achieve their best grounding scores on the benchmark datasets. For example, ARN adopts $L_{lan} + L_{att}$ setting to acquire the best accuracy on RefCOCO and RefCOCOg and utilize $L_{lan} + L_{adp} + L_{att}$ to achieve better results on RefCOCO+. In contrast, KPRN achieves the best

TABLE I

PERFORMANCE (ACC%) COMPARISON WITH STATE-OF-THE-ART APPROACHES ON REFCOCO, REFCOCO+, AND REFCOCOG. THE BEST GROUNDING ACCURACY ON EACH SPLIT IS IN BOLD.

Approaches	Settings	RefCOCO			RefCOCO+			RefCOCOg
		val	testA	testB	val	testA	testB	val
VC [30]	w/o reg	-	13.59	21.65	-	18.79	24.14	25.14
VC [30]	GT	-	17.34	20.98	-	23.24	24.91	33.79
VC [30]	w/o α	-	33.29	30.13	-	34.60	31.58	30.26
ARN [8]	L_{adp}+L_{att}	33.07	36.43	29.09	33.53	36.40	29.23	33.19
ARN [8]	L_{lan}+L_{adp}	33.60	35.65	31.48	34.40	35.54	32.60	34.50
ARN [8]	L_{lan}+L_{att}	38.05	35.27	36.47	34.51	34.40	36.12	39.62
ARN [8]	L_{lan}+L_{adp}+L_{att}	34.26	36.01	33.07	34.53	36.01	33.75	34.66
KPRN [9]	hard	35.04	34.74	36.53	35.10	32.75	36.76	35.44
KPRN [9]	hard+attr	34.93	33.76	36.98	35.31	33.46	37.27	38.37
KPRN [9]	soft	34.43	33.82	35.45	35.96	35.24	36.96	33.56
KPRN [9]	soft+attr	36.34	35.28	37.72	37.16	36.06	39.29	36.65
IGN[31]	Base	31.05	34.39	28.16	31.13	34.44	29.59	32.17
IGN[31]	CCL	34.78	37.64	32.59	34.29	36.91	33.56	34.92
EARN [32]	L_{lan}+L_{adp}	35.31	37.07	32.66	35.50	37.39	33.65	38.99
EARN [32]	L_{lan}+L_{att}	34.93	33.76	36.98	35.31	33.46	37.27	38.37
EARN [32]	L_{lan}+L_{adp}+L_{att}	38.08	38.25	38.59	37.54	37.58	37.92	45.33
DTMR [10]	GT	39.21	41.14	37.72	39.18	**40.01**	38.08	43.24
Cycle-Free [33]	GT	39.58	**41.46**	37.96	39.20	39.63	37.59	-
Proposed	teacher	38.96	39.37	39.41	39.67	39.98	39.93	47.71
Proposed	student(η = 2.0)	**39.70**	39.92	**39.63**	**40.20**	39.94	**40.27**	**47.99**

grounding accuracy on RefCOCO and RefCOCO+ with the soft+attr setting, and obtains the highest grounding score on RefCOCOg with hard+attr. We summarize the comparison results in Table 1.

For the grounding accuracy gain, our proposed architecture outperforms the accuracy on several splits compared with the SOTA approaches. Concretely, the teacher model outperforms the SOTA approaches on the testB set of RefCOCO, val and testB splits of RefCOCO+, and val set of RefCOCOg. While the student model improves the accuracy by 0.12%, 1.0%, and 2.66% on the validation sets of the datasets compared with the SOTA methods, Cycle-Free [33] and EARN [32], respectively. Moreover, the student model boosts the grounding accuracy 1.67% and 2.19% on testB of RefCOCO and RefCOCO+, against Cycle-Free [33] and DTMR [10]. While the results on testA of RefCOCO and RefCOCO+ are lower than Cycle-Free [33] and DTMR [10].

Some qualitative visualization results on RefCOCO, RefCOCO+, and RefCOCOg are shown in Figure 2, where the referring expressions are positioned under the corresponding images, and the ground truths and the grounded target regions are denoted as solid green and red rectangles, respectively. For the correctly grounded examples, the model can precisely parse the expressions into triad elements and ground the target objects according to the aggregated matching scores. For the incorrectly grounded referring expressions, the model fails to predict the correct targets based on the parsed discriminative triads and corresponding visual features. For instance, for the failure sample "man looking at phone", our model can not distinguish the target "man" with the minor visual difference between region candidates. In addition, for the last failure case "the mid sized vase", the parsed triad is "vase-self-vase" which includes the target element "vase", the discriminative relationship "self" and the subject element "vase", so the model fails to locate the correct "vase" in the corresponding image.

D. Hyper-parameter Analysis

In this section, we analyze the performance of the proposed architecture by setting different hyper-parameter η in Equation (8). In the knowledge distillation experiments, we set the hyper-parameter η over an extensive range from 0.001 to 10, and the careful adjustments of η bring some extra performance improvement. The grounding results achieved by the teacher and the student with different η are listed in Table 2.

As observed from Table 2, the results show that the student performance is sensitive to the choice of η. Specifically, compared with the accuracy acquired by the teacher, the smaller $\eta \in \{0.001, 0.005\}$ damage the performance on RefCOCO and RefCOCOg, and also weaken the performance on RefCOCO+. With the increase of η, the grounding accuracy achieved by the student on the datasets is improved. When η = 0.05, the student achieves the best grounding accuracy on the testA split of RefCOCO and RefCOCO+. And $\eta \in \{1.0, 1.5, 2.0, 3.0, 5.0\}$, the acquired accuracy outperforms the SOTA approaches on several splits. Specifically, the model with η = 1.0 attains the best grounding score on the testB set of RefCOCO+. When setting η = 2.0, the student acquires the best accuracy on val and testB sets of RefCOCO, val

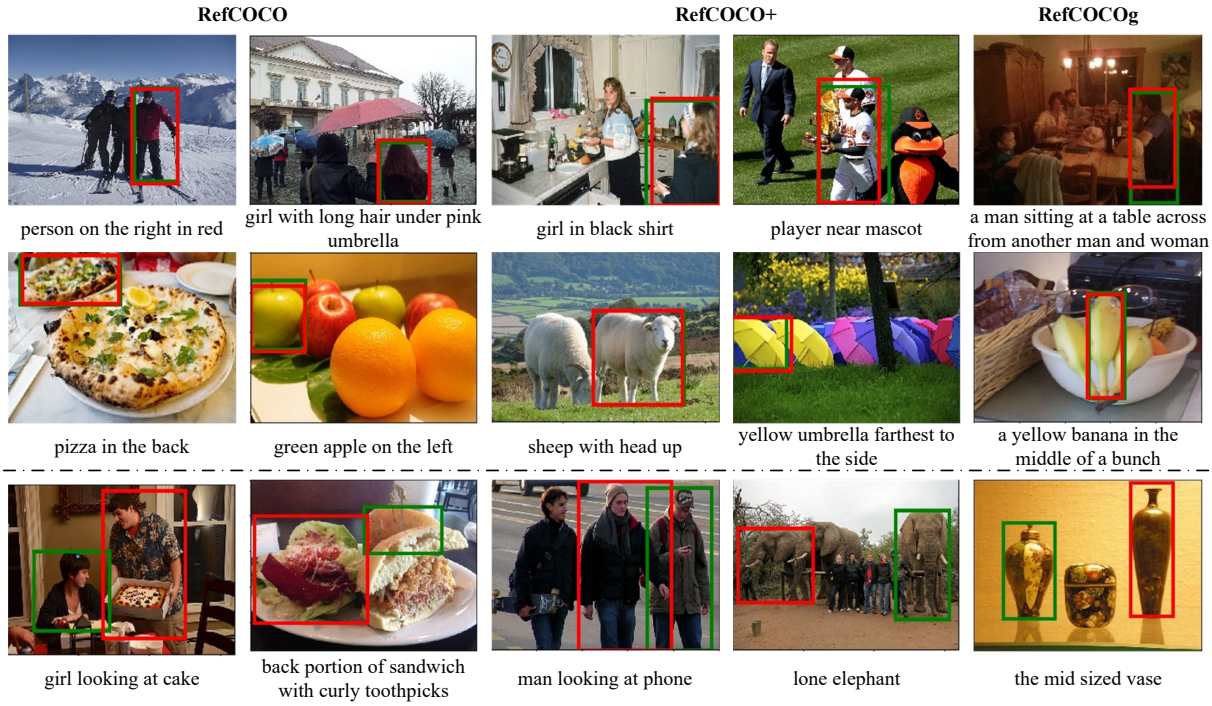

Fig. 2. Qualitative visualization results of the proposed approach on RefCOCO, RefCOCO+, and ReCOCOg datasets. Some incorrect grounding examples are listed under the dashed line. The green boxes represent the ground truths and the red boxes denote the grounded targets by our approach.

TABLE II

GROUNDING ACCURACY OF THE ACQUIRED BY THE STUDENT MODEL ON REFCOCO, REFCOCO+, AND REFCOCOG WHEN THE VALUE OF η IN EQUATION (8) VARIES FROM 0.001 TO 10.0. THE BEST RESULTS ARE IN BOLD.

η	RefCOCO			RefCOCO+			RefCOCOg
	val	testA	testB	val	testA	testB	val
teacher	38.96	39.37	39.41	39.67	39.98	39.93	47.70
0.001	37.53	38.50	38.10	39.68	39.85	40.29	46.73
0.005	38.20	39.08	38.51	39.64	40.10	39.97	47.59
0.01	38.50	39.33	38.98	39.58	39.84	40.07	47.86
0.05	39.39	**40.20**	39.12	39.60	**40.19**	40.38	47.80
0.1	39.37	40.16	39.63	39.70	39.85	40.11	47.78
0.5	39.53	39.79	39.21	39.80	39.56	40.27	47.69
1.0	39.46	39.83	39.12	39.82	40.10	**40.60**	47.83
1.5	39.57	39.99	39.51	40.01	39.92	40.46	47.84
2.0	**39.70**	39.92	**39.63**	**40.20**	39.94	40.27	**47.99**
2.5	39.42	39.67	39.21	40.04	40.01	39.93	47.77
3.0	39.62	39.61	39.41	40.10	39.96	40.29	47.80
5.0	39.57	39.67	39.37	40.04	39.87	40.15	47.72
8.0	39.26	39.49	39.23	39.98	39.94	39.76	47.58
10.0	39.23	39.28	39.20	39.74	39.89	39.72	47.36

set of RefCOCO+ and RefCOCOg, and surpasses the SOTA models on five splits of the benchmarks. We thus select the grounding results acquired by the student model with η = 2.0 to compare with the SOTA methods. While we set η to bigger values, such as 8.0 and 10.0, the performance of the student decreases.

We believe the leading cause is the different expression styles in the datasets for defining the target objects. The expressions in RefCOCO+ abandon the absolute position information to describe objects, so the accuracy on RefCOCO+ is relatively less sensitive to the varies of η. In contrast, the RefCOCO expressions combine attribute, position information, and appearance to portray objects. Thus, the performance on RefCOCO varies obviously. The expressions in RefCOCOg concentrate on the relation descriptions among the candidate targets and their adjacent image regions. Consequently, the larger η values weaken the grounding accuracy on RefCOCOg.

V. CONCLUSION

In this paper, we propose a novel architecture that introduces knowledge distillation to address weakly supervised REG. Motivated by the salient attribute of knowledge distillation and the role of expression components for grounding the target objects, we leverage knowledge distillation as a unique scheme to exploit and transfer valuable information to acquire a better grounding model. Specifically, we reactivate the target-related prediction information learned by the pre-trained teacher model and adopt the knowledge distillation as a unique channel to guide the training process of the student model and boost its grounding performance. Extensive experiments on the public benchmark datasets demonstrate the substantial benefits of the proposed architecture.

VI. ACKNOWLEDGEMENT

This work is partly funded by the German Research Foundation (DFG) and National Science Foundation (NSFC) in the project Crossmodal Learning under contract Sonderforschungsbereich Transregio 169, the National Natural Science Foundation under contract No. 92048205, and the DAAD German Academic Exchange Service under the CASY project.

REFERENCES

[1] E. Krahmer and K. Van Deemter, "Computational generation of referring expressions: A survey," *Computational Linguistics*, vol. 38, no. 1, pp. 173–218, 2012.

[2] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, X. Wang, and M. Zhou, "Visual question generation as dual task of visual question answering," in *CVPR*, 2018, pp. 6116–6124.

[3] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *ECCV*, 2016, pp. 241–257.

[4] M. Shridhar, D. Mittal, and D. Hsu, "Ingress: Interactive visual grounding of referring expressions," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 217–232, 2020.

[5] J. Mi, J. Lyu, S. Tang, Q. Li, and J. Zhang, "Interactive natural language grounding via referring expression comprehension and scene graph parsing," *Frontiers in Neurorobotics*, vol. 14, p. 43, 2020.

[6] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, "Grounding of textual phrases in images by reconstruction," in *ECCV*, 2016, pp. 817–834.

[7] F. Xiao, L. Sigal, and Y. Jae Lee, "Weakly-supervised visual grounding of phrases with linguistic structures," in *CVPR*, 2017, pp. 5945–5954.

[8] X. Liu, L. Li, S. Wang, Z.-J. Zha, D. Meng, and Q. Huang, "Adaptive reconstruction network for weakly supervised referring expression grounding," in *ICCV*, 2019, pp. 2611–2620.

[9] X. Liu, L. Li, S. Wang, Z.-J. Zha, L. Su, and Q. Huang, "Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding," in *ACM MM*, 2019, pp. 539–547.

[10] M. Sun, J. Xiao, E. G. Lim, S. Liu, and J. Y. Goulermas, "Discriminative triad matching and reconstruction for weakly referring expression grounding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4189–4195, 2021.

[11] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[12] H. Qiu, H. Li, Q. Wu, F. Meng, H. Shi, T. Zhao, and K. N. Ngan, "Language-aware fine-grained object representation for referring expression comprehension," in *ACM MM*, 2020, pp. 4171–4180.

[13] R. Hu, A. Rohrbach, T. Darrell, and K. Saenko, "Language-conditioned graph networks for relational reasoning," in *ICCV*, 2019, pp. 10 294–10 303.

[14] M. Feng, Z. Li, Q. Li, L. Zhang, X. Zhang, G. Zhu, H. Zhang, Y. Wang, and A. Mian, "Free-form description guided 3d visual graph network for object grounding in point cloud," in *ICCV*, 2021, pp. 3722–3731.

[15] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *ECCV*, 2016, pp. 69–85.

[16] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *CVPR*, 2016, pp. 11–20.

[17] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling relationships in referential expressions with compositional modular networks," in *CVPR*, 2017, pp. 1115–1124.

[18] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *CVPR*, 2018, pp. 1307–1315.

[19] S. Yang, G. Li, and Y. Yu, "Dynamic graph attention for referring expression comprehension," in *ICCV*, 2019, pp. 4644–4653.

[20] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. v. d. Hengel, "Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks," in *CVPR*, 2019, pp. 1960–1968.

[21] C. Jing, Y. Wu, M. Pei, Y. Hu, Y. Jia, and Q. Wu, "Visual-semantic graph matching for visual grounding," in *ACM MM*, 2020, pp. 4041–4050.

[22] S. Chen and B. Li, "Multi-modal dynamic graph transformer for visual grounding," in *CVPR*, 2022, pp. 15 534–15 543.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, vol. 28, 2015.

[24] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo, "A fast and accurate one-stage approach to visual grounding," in *ICCV*, 2019, pp. 4683–4693.

[25] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[26] Y. Liao, S. Liu, G. Li, F. Wang, Y. Chen, C. Qian, and B. Li, "A real-time cross-modality correlation filtering method for referring expression comprehension," in *CVPR*, 2020, pp. 10 880–10 889.

[27] J. Ye, X. Lin, L. He, D. Li, and Q. Chen, "One-stage visual grounding via semantic-aware feature filter," in *ACM MM*, 2021, pp. 1702–1711.

[28] G. Luo, Y. Zhou, X. Sun, L. Cao, C. Wu, C. Deng, and R. Ji, "Multi-task collaborative network for joint referring expression comprehension and segmentation," in *CVPR*, 2020, pp. 10 034–10 043.

[29] M. Sun, W. Suo, P. Wang, Y. Zhang, and Q. Wu, "A proposal-free one-stage framework for referring expression comprehension and generation via dense cross-attention," *IEEE Transactions on Multimedia*, 2022.

[30] Y. Niu, H. Zhang, Z. Lu, and S.-F. Chang, "Variational context: Exploiting visual and textual context for grounding referring expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 347–359, 2019.

[31] Z. Zhang, Z. Zhao, Z. Lin, X. He *et al.*, "Counterfactual contrastive learning for weakly-supervised vision-language grounding," *NeurIPS*, vol. 33, pp. 18 123–18 134, 2020.

[32] X. Liu, L. Li, S. Wang, Z.-J. Zha, Z. Li, Q. Tian, and Q. Huang, "Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[33] M. Sun, J. Xiao, E. G. Lim, and Y. Zhao, "Cycle-free weakly referring expression grounding with self-paced learning," *IEEE Transactions on Multimedia*, 2021.

[34] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *ICLR*, 2015.

[35] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *CVPR*, 2019, pp. 3967–3976.

[36] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *ICCV*, 2019, pp. 1365–1374.

[37] L. Zhang and K. Ma, "Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors," in *ICLR*, 2021.

[38] J. Fan and Z. Zhang, "Memory-based cross-image contexts for weakly supervised semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[39] H.-R. Wei, S. Huang, R. Wang, X. Dai, and J. Chen, "Online distilling from checkpoints for neural machine translation," in *NAACL-HLT*, 2019, pp. 1932–1941.

[40] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[42] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NeurIPS*, vol. 30, 2017.