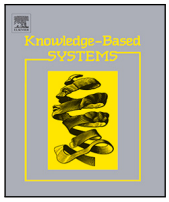
Knowledge-Based Systems 286 (2024) 111437

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Adaptive knowledge distillation and integration for weakly supervised referring expression comprehension

Jinpeng Mi [a,*], Stefan Wermter [b], Jianwei Zhang [c]

[a] Institute of Machine Intelligence, University of Shanghai for Science and Technology, China
[b] Knowledge Technology, Department of Informatics, University of Hamburg, Germany
[c] Technical Aspects of Multimodal Systems, Department of Informatics, University of Hamburg, Germany

ARTICLE INFO

Keywords:
Referring expression comprehension
Adaptive knowledge distillation
Dynamic knowledge integration
Target- and interaction-aware knowledge

ABSTRACT

Weakly supervised referring expression comprehension (REC) aims to ground target objects in images according to given referring expressions, while the mappings between image regions and referring expressions are unavailable during the model training phase. Existing models typically reconstruct the multimodal relationships to ground targets by utilizing off-the-shelf information, and ignore to further exploit helpful knowledge to enhance the model performance. To address this issue, we propose an adaptive knowledge distillation architecture to enrich the predominant pattern of weakly supervised REC and transfer the target-aware and interaction-aware knowledge from a pre-trained teacher grounder to enhance the grounding performance of the student model. Specifically, in order to encourage the teacher to impart more reliable knowledge, we present a Knowledge Confidence-Based Adaptive Temperature (KCAT) learning approach to learn optimal temperatures to transfer the target-aware and interaction-aware knowledge with higher prediction confidence. Moreover, to urge the student to absorb more helpful knowledge, we introduce a Student Competency-Based Adaptive Weight (SCAW) learning strategy to dynamically integrate the distilled target-aware and interaction-aware knowledge to enhance the student's grounding certainty. We conduct extensive experiments on three benchmark datasets, RefCOCO, RefCOCO+, and RefCOCOg, to validate the proposed approach. Experimental results demonstrate that our approach achieves superior performance over state-of-the-art methods with the aid of adaptive knowledge distillation and integration. The code and trained models are available at: https://github.com/dami23/WREC_AdaptiveKD.

1. Introduction

Referring expression comprehension (REC) locates target objects in images according to the given referring expressions by comprehensively understanding the context in images and expressions. REC bridges object detection and natural language understanding, and can be applied to multiple tasks, such as image retrieval [1,2], visual question answering [3,4], visual language navigation [5,6], and human–robot interaction [7–9].

Benefiting from the accessibility of a large volume of manually annotated datasets, existing supervised-based approaches achieve promising grounding accuracy. The dataset collection requires detailed annotations of bounding boxes, referring expressions, and the corresponding mapping between each bounding box and referring expressions. However, building such datasets is time consuming and laborious. In order to reduce the labor intensity of manual annotations and expand the applications of REC in practical scenarios, plenty of weakly supervised methods [10–13] have been proposed.

In contrast to fully supervised REC, which learns the image region-expression alignment by utilizing the annotated mapping between image regions and referring expressions, weakly supervised REC does not require cross-modal mapping annotations during the model training stage, and the unavailability of the cross-modal mapping relationships between regions and expressions poses challenges for weakly supervised REC. Thus, to locate target regions, the core of weakly supervised REC lies in exploiting valuable information from visual images and textual expressions to facilitate the cross-modal mapping reconstruction. Existing models propose various strategies to build the corresponding relationships. For instance, Liu et al. [10] introduce an Adaptive Reconstruction Network (ARN) to reconstruct the mapping by combining the subjection, location, and context features in visual images and textual expressions. Liu et al. [11] present a Knowledge-guided Pairwise Reconstruction Network (KPRN) to learn the correspondence between the detected image regions and expressions. Sun et al. [12] propose a

* Corresponding author.
E-mail addresses: jinpeng.mi@uni-hamburg.de (J. Mi), stefan.wermter@uni-hamburg.de (S. Wermter), jianwei.zhang@uni-hamburg.de (J. Zhang).

https://doi.org/10.1016/j.knosys.2024.111437
Received 25 April 2023; Received in revised form 17 January 2024; Accepted 22 January 2024
Available online 23 January 2024

Discriminative Triad Matching and Reconstruction (DTMR) framework to learn the relationship between image regions and expressions by utilizing the parsed linguistic structures. Most surprisingly, the heuristic rule-based approach achieves state-of-the-art (SOTA) performance on the weakly supervised REC task. Zhang et al. [13] develop counterfactual transformation schemes to optimize the visual region and textual expression alignments. However, these methods mainly utilize off-the-shelf information to reconstruct the relationships, such as semantics in deep visual features [10,11], linguistic structure [12], and ignore to further explore helpful knowledge to boost grounding performance.

Thus, one natural question posed on weakly supervised REC could be: is there a more effective strategy to exploit and utilize reliable and helpful information to learn a better grounding model? Inspired by the salient attribute of knowledge distillation [14], we employ knowledge distillation as a unique scheme to enrich the predominant pattern of weakly supervised REC and facilitate the model's grounding performance. In other words, we leverage the knowledge transferred from a pre-trained teacher model as high-quality pseudo-ground-truth labels to guide the training process of the student model and bolster its grounding performance.

Moreover, according to the styles of referring expressions in the benchmark datasets RefCOCO [15], RefCOCO+ [15], and RefCOCOg [16], the description information of target objects comprises the target attributes, the interaction information between the image regions, and the combination of the attribute description and interaction information. Namely, the target-aware and interaction-aware information in expressions play critical roles in disambiguating target objects. In addition, existing methods demonstrate that transferring knowledge from multiple teachers [17] or multi-level knowledge from a teacher model [18] significantly improves the performance of the student model. Inspired by these observations, we explore and make full use of the target-aware and interaction-aware prediction information learned by the teacher model and transfer them from the teacher to boost the grounding performance of the student.

Unlike the existing knowledge distillation approaches that utilize an identical temperature hyper-parameter during the knowledge distillation process, we propose a Knowledge Confidence-Based Adaptive Temperature (KCAT) learning approach that learns optimal temperatures to urge the teacher model to transfer more reliable knowledge to the student. Moreover, because the target-aware and interaction-aware knowledge have complementary semantic strengths for grounding target objects, we distill them to promote student learning from multiple perspectives of the teacher model. On the other hand, to avert introducing information redundancy and to encourage the student to absorb more helpful knowledge, we further present a Student Competency-Based Adaptive Weight (SCAW) learning approach to dynamically integrate the distilled target-aware and interaction-aware knowledge to facilitate the student's prediction certainty.

We conduct extensive experiments and ablation studies on benchmark datasets RefCOCO [15], RefCOCO+ [15], and RefCOCOg [16] to evaluate our proposed framework. The proposed approach significantly improves the grounding performance of weakly supervised REC and outperforms SOTA models on the validation and testB sets of RefCOCO, RefCOCO+, and RefCOCOg.

In summary, the main contributions of this paper are summarized as follows:

- We propose to enhance the predominant grounding pattern of weakly supervised REC with adaptive knowledge distillation, and employ the transferred knowledge as the high-quality pseudo-ground-truth labels to boost the grounding performance of the student model.
- We present an adaptive temperature learning approach to learn optimal temperatures according to the confidence of the distilled knowledge, and introduce an adaptive weight learning strategy to dynamically assign weights to fuse the target-aware and interaction-aware knowledge based on the student prediction certainty.
- We conduct extensive experiments and ablation studies on the benchmark datasets, and our proposed approach outperforms SOTA grounding performance on several splits of the benchmarks.

2. Related work

2.1. Supervised referring expression comprehension

Supervised REC aims to locate target objects in images by jointly understanding the semantics of the images and the given referring expressions, where the relationship annotations between region proposals and referring expressions are available during the training phase. The pioneering methods of REC [15,16] directly locate target objects by calculating the matching score between the visual features of image regions and the text representations of the referring expressions. On this basis, Hu et al. [19] and Yu et al. [20] propose modular frameworks to improve the grounding performance. In order to better capture the crucial relationship information between visual regions and textual expressions, graph neural network-based methods [21–23] represent the images and referring expressions in the form of graphs, and ground target objects by aligning the generated multimodal graphs. According to the target object grounding pattern, these approaches can be sorted as the two-stage paradigm. Specifically, these models first adopt a pre-trained object detection model, such as Faster R-CNN [24], to detect and extract the visual features of the candidates, and then ground target objects by calculating the matching score between the visual features and the expression text representations. To relieve the burden of object detection and improve the inference speed of the two-stage approaches, one-stage models are considered an alternative orientation.

Most one-stage methods directly fuse the referring expression text representations with the extracted visual features to locate target objects. For instance, Yang et al. [25] directly integrate the text representations into the object detection model YOLO v3 [26] to predict the bounding boxes of the target objects. The newly proposed one-stage models introduce multiple multimodal feature fusion tactics to improve the target grounding accuracy. For example, text semantics-aware approaches [27–29] emphasize the role of text semantics to obtain unique visual representations for candidate regions, Luo et al. [30] reduce the difference between the two modes of text and vision, Sun et al. [31] strengthen the reasoning clues of target objects by leveraging attention-based multimodal data fusion, and Huang et al. [32] propose Landmark Feature Convolution to describe the relationship between referring expressions and target objects. Although the one-stage models avoid the dependence on the pre-trained candidate region generation models, they require predefined anchors and are vulnerable to the inconsistency of modal information.

With the prosperity of pre-trained models in multimodal tasks, researchers investigate to improve the performance of one-stage models by introducing BERT [33] into REC, such as VL-BERT [34], TransVG [35], MDETR [36], Word2Pix [37], and OFA [38]. These methods utilize Transformer [39] to jointly learn contextualized representations for image regions and expressions, and they aim to facilitate REC by learning the generalizable representations from large-scale data. Albeit these models achieve promising results on the benchmark datasets, they demand enormous computational power and a longer training time to complete their training.

2.2. Weakly supervised referring expression comprehension

Weakly supervised REC methods train models without the mapping annotations between region proposals and corresponding referring expressions. In this mode, Rohrbach et al. [40] first propose to ground target objects by reconstructing queries via the calculated attention scores between image regions and queries. Subsequently, Niu et al. [41] take advantage of the reciprocity between the candidate region and

the referring expression to model their relationship, and Liu et al. [10] integrate the matching score between each region proposal and referring expression with the contextualized feature of the region proposal to rebuild the cross-modal mappings. Based on [10], Liu et al. [42] propose to filter unrelated candidate proposals via an entity enhancement strategy to improve the grounding accuracy. Additionally, Liu et al. [11] build the mappings by leveraging prior knowledge acquired from pre-trained Faster RCNN [24] and ground target objects through the learned pairwise matching score, and Sun et al. [12] locate target objects via triad-level matching learning and reconstruction.

Unlike these reconstruction-based methods, Zhang et al. [13] utilize counterfactual results to facilitate the alignment between visual features and textual representations. Sun et al. [43] present a Cycle-free model and develop a region describer to generate a textual description for each region proposal, and grounds targets via the semantic similarity between the acquired region descriptions and the referring expressions. Jin et al. [44] introduce an anchor-based contrastive learning scheme to align the regions and expressions, and the proposed approach is employed as a teacher model to generate pseudo-labels to improve the performance of common REC models.

In contrast to the methods mentioned above, we focus on devising a novel framework that can enrich the prevalent pattern of the weakly supervised REC with knowledge distillation and aim to explore and transfer helpful knowledge to achieve a better grounding model.

2.3. Knowledge distillation

Knowledge distillation, initially introduced in [14], transfers knowledge from a pre-trained teacher model to a student model to promote the student model's performance. In recent years, plenty of knowledge distillation methods have been proposed to transfer different kinds of knowledge, including output probability [14,45,46], intermediate layer representations [47–50], inter-class correlation [51,52], and knowledge learned in earlier training epochs [53–55]. Studies on output probability distillation aim to optimize the student training phase with learned prediction logits, intermediate layer-based methods transfer features from the teacher to enhance the student, inter-class correlation-related approaches distill the relationship learned by the teacher to the student, and models distilled knowledge from earlier training epochs investigate to obtain richer supervision information from the teacher. These approaches focus on distilling one type of knowledge from a single pre-trained teacher. In contrast, we attempt to transfer multiple kinds of knowledge with complementary properties to enhance the performance of the student, and we also investigate to transfer knowledge with higher confidence by learning dynamic temperatures.

To facilitate student learning from multiple perspectives of the teacher, some methods distill multi-level knowledge or transfer knowledge from multiple teachers [17,18,56,57]. Instead of equally utilizing ensemble knowledge or distilling the average of the ensemble knowledge from the teacher, various strategies have been introduced to address the importance of multi-level knowledge for training students. For instance, Du et al. [58] distill the ensemble knowledge through a multi-objective optimization strategy, Liu et al. [59] calculate the fusing weight via latent representation, Kwon et al. [60] employ the entropy of the teacher's labels to obtain the fusing weight, and Li et al. [61] emphasize the intra-class variance retained by the teacher model to enhance the performance of the student.

The most relevant contribution of our work is the development of adaptive weight learning strategies for distilling ensemble knowledge from multiple teachers. Unlike the existing adaptive and dynamic knowledge distillation approaches, we aim to distill knowledge with higher confidence to the student via adaptive temperature and encourage the student to digest more reliable knowledge based on the student's competency.

3. Problem formulation and teacher grounder

3.1. Problem formulation

Given an image I with M regions of interest (RoIs) $O = \{o_i\}_{i=1}^{M}$ and a referring expression E, weakly supervised REC aims to locate the target region $o^* \in O$ by learning the relationship between E and all region candidates o_i. Specifically, we aim to learn a model to ground o^* via reconstructing the mapping between o_i and E, and select the region candidate with the maximum matching score as the target object o^*:

$$o^* = \arg\max_{o_i \in R} \mathrm{G}(o_i, E), \tag{1}$$

where $\mathrm{G}(\cdot,\cdot)$ denotes the mapping reconstruction operation.

3.2. Textual and visual representation encoding

Textual Feature Encoding. To extract the textual representations, we first employ the natural language parsing method introduced in DTMR [12] to parse referring expressions. DTMR parses expressions into discriminative triads, where each triad comprises the discriminative description information for the target object, the related subject, and the relationship between the target object and the related subject. Formally, each expression E is parsed into multiple triads, and each discriminative triad includes a target unit u_t, a subject unit u_s, and a relationship unit u_r. We then adopt GloVe [62] to extract textual embeddings e_t, e_s, $e_r \in \mathbb{R}^{1\times 300}$ for u_t, u_s, and u_r respectively.

Visual Feature Encoding. For the given images, we adopt Faster R-CNN [24] to detect region candidates o_i for each image I, and utilize ResNet-101 [63] to extract a visual feature $f_v^i \in \mathbb{R}^{7\times 7\times 2048}$ from the 4-th layer of the ResNet-101 with RoI pooling [24] as the visual feature representation of each o_i.

To explore the spatial relation between the region candidates, following [15], we utilize a 5-dimensional vector $f_{sp}^i = [\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w\cdot h}{W\cdot H}]$ to encode the absolute location feature for each o_i, where x_{tl}, y_{tl}, x_{br}, y_{br} represent the top left and bottom right positions, respectively. w and h are the width and height of the region candidate, W and H represent the width and height of the image I, and $\frac{w\cdot h}{W\cdot H}$ is the relative size of o_i in image I. The spatial vector f_{sp}^i is then projected to a vector $\bar{f}_{sp}^i \in \mathbb{R}^{1\times 512}$, which is combined with f_v^i as the input for the reconstruction of the correlated referring expressions.

3.3. Teacher grounder

In order to transfer the target-aware and interaction-aware knowledge, we pre-train a teacher grounder at first. For each region proposal pair (o_i, o_j), we learn the matching score between the region candidate o_i and the parsed target element u_t through a pairwise attention module

$$s_i^t = \mathrm{Softmax}(w_2^t \Psi(w_1^t \Psi(f_v^i \oplus e_t) + b_1^t) + b_2^t), \tag{2}$$

where w_1^t, b_1^t, w_2^t, b_2^t are learnable parameters, Ψ represents the ReLU activation function, and $\oplus$ denotes the concatenation operation. The hierarchical pairwise attention learns the matching score using the syntax structure of the expressions and thus can further promote target grounding.

Similarly, we model the matching score between (o_i, o_j) and the subject element u_s and the discriminative relationship element u_r via

$$\begin{aligned} s_j^s &= \mathrm{Softmax}(w_2^s \Psi(w_1^s \Psi(f_v^j \oplus e_s) + b_1^s) + b_2^s), \\ s_{i,j}^r &= \mathrm{Softmax}(w_2^r \Psi(w_1^r \Psi(\bar{f}_v \oplus e_r) + b_1^r) + b_2^r), \end{aligned} \tag{3}$$

where f_v^j represents the visual feature of the related subject o_j, $\bar{f}_v = f_v^i \oplus \bar{f}_{sp}^i \oplus f_v^j \oplus \bar{f}_{sp}^j$ and $\bar{f}_{sp}^i$ is the projected spatial vector of o_j.

For reconstructing the triad elements via the extracted visual representations, we first compute the weighted sum of the regions' visual features and the associated triad element matching scores as follows:

$$h_t = \sum_{i=1}^{N} s_i^t f_v^i, \qquad h_s = \sum_{j=1}^{N} s_j^s f_v^j, \qquad h_r = \sum_{i,j=1}^{N} s_{i,j}^r \bar{f}_v. \tag{4}$$

We then utilize a Multilayer Perceptron (MLP) to learn the linguistic embeddings responding to the parsed triad elements via

$$\bar{e}_t = \mathrm{MLP}(h_t), \qquad \bar{e}_s = \mathrm{MLP}(h_t), \qquad \bar{e}_r = \mathrm{MLP}(h_r). \tag{5}$$

To ground the target o^*, we first calculate the triad-level attention score $Att_{triad}(o_i)$ and select the region candidate with the maximum $Att_{triad}(o_i)$ as the target o^*. The grounding process can be formulated as

$$\begin{aligned} &Att_{triad}(o_i) = \gamma_1 s_i^t + \gamma_2 s_j^s + \gamma_3 s_{i,j}^r, \\ &o^* = \arg\max_{o_i \in R} Att_{triad}(o_i), \end{aligned} \tag{6}$$

where γ_1, γ_2 and γ_3 are the weighting hyper-parameters to imbalance the impact of each attention score to the overall matching score. $\gamma_1 s_i^t$ is the target prediction score and directly contributes to locating the target objects. $(\gamma_2 s_j^s + \gamma_3 s_{i,j}^r)$ denotes the interaction prediction information between image regions described in expressions and is employed as an important auxiliary information of $\gamma_1 s_i^t$ to disambiguate the target regions.

We train the teacher model by minimizing the mean squared error between the parsed triad element textual representations and the reconstructed triad element embeddings, and the final reconstruction loss L_{WREC} is defined as

$$\begin{aligned} &L_t = \left\| e_t - \bar{e}_t \right\|_2^2, \qquad L_s = \left\| e_s - \bar{e}_s \right\|_2^2, \qquad L_r = \left\| e_r - \bar{e}_r \right\|_2^2, \\ &L_{WREC} = \mu_1 L_t + \mu_2 L_s + \mu_3 L_r, \end{aligned} \tag{7}$$

where μ_1, μ_2, μ_3 are the weights wo balance the impact of L_t, L_s, L_t.

4. Adaptive knowledge distillation and integration

The proposed adaptive knowledge distillation and integration approach for weakly supervised REC aims to build a novel training strategy, which regards the target- and interaction-related prediction information learned by the pre-trained teacher grounder as high-quality pseudo-labels to guide the training process of the student, and to further boost the student's grounding performance. In order to encourage the teacher to transfer more reliable knowledge to the student, we propose a Knowledge Confidence-Based Adaptive Temperature (KCAT) learning module to adaptively learn optimal temperatures during the knowledge distillation. To encourage the student to absorb more helpful knowledge, we introduce a Student Competency-Based Adaptive Weight (SCAW) learning module to dynamically integrate the target-aware and the interaction-aware knowledge to enhance the student's prediction certainty. The proposed architecture diagram is shown in Fig. 1.

4.1. Overview of knowledge distillation

For transferring the teacher's knowledge to the student model, we employ the knowledge distillation method introduced in [14] that distills the softened logits prediction knowledge from the teacher with Kullback–Leibler (KL) divergence loss. We denote $P = \{p_i\}_{i=1}^{M}$ as the output logits of the models, where M is the number of the RoIs of given images. The softened prediction information is acquired by function $\Phi(p_i) = \mathrm{Softmax}(p_i)$. For the output logit vector P^T of the teacher, and the student logit vector P^S, the objective of knowledge distillation is given by

$$L_{KD} = \mathrm{KL}(\Phi(P^T/\tau) \parallel \Phi(P^S/\tau)), \tag{8}$$

where τ denotes the temperature hyper-parameter.

4.2. Target-aware knowledge distillation

The target-aware prediction information $\gamma_1 s_i^t$ directly contributes to grounding the target regions. Thus, we distill the target-aware prediction knowledge learned by the teacher to facilitate the target-related prediction information learning during student training. Specifically, we first reactivate the target-aware prediction knowledge $K_{Tar}^T = \gamma_1 s_i^t$ learned by the teacher model and distill the softened target-aware prediction knowledge via

$$L_{KD}^{Tar} = \mathrm{KL}(\Phi(K_{Tar}^T/\tau_{Tar}) \parallel \Phi(Att_{triad}^S(o_i)/\tau)), \tag{9}$$

where

$$Att_{triad}^S(o_i) = \gamma_1 s_i^{t,S} + \gamma_2 s_j^{s,S} + \gamma_3 s_{i,j}^{r,S} \tag{10}$$

is the triad-level reconstruction score learned by the student model.

The overall loss of the target-aware knowledge distillation is formulated as

$$\begin{aligned} &L_{WREC}^S = \sum_{A \in \{t,s,r\}} \left\| e_A - \bar{e}_A^S \right\|_2^2, \\ &L_{Tar} = L_{WREC}^S + \lambda L_{KD}^{Tar}, \end{aligned} \tag{11}$$

where $\bar{e}_A^S, A \in \{t, s, r\}$ denotes the triad element embeddings reconstructed by the student model, and λ represents the trade-off parameter to balance the importance of the knowledge distillation loss and the triad-level reconstruction loss for the student training.

4.3. Interaction-aware knowledge distillation

In order to ground target objects according to the given referring expressions, the interaction-related prediction information between region candidates described in the expressions also plays a key role in obtaining discriminative cues for grounding targets. Thus, we also re-attend and transfer the interaction-aware prediction information learned by the teacher to enrich the grounding proofs for the student model. Concretely, we utilize the interaction-aware prediction knowledge $K_{Inter}^T = \gamma_2 s_j^s + \gamma_3 s_{i,j}^r$ and distill the softened interaction-aware knowledge by

$$L_{KD}^{Inter} = \mathrm{KL}(\Phi(K_{Inter}^T/\tau_{Inter}) \parallel \Phi(Att_{triad}^S(o_i)/\tau)). \tag{12}$$

The loss of the interaction-aware knowledge distillation is defined as

$$L_{Inter} = L_{WREC}^S + \lambda L_{KD}^{Inter}. \tag{13}$$

4.4. Knowledge confidence-based adaptive temperature learning

During the knowledge distillation, temperature τ balances the ground truth label knowledge and the softened prediction knowledge learned by the teacher model. A fixed temperature is not necessarily the optimal value for distilling knowledge in the whole training process, and it may also impede the helpful knowledge distillation from the teacher to the student. On the other hand, if the teacher obtains unreliable grounding predictions on some region-expression pairs, directly distilling the knowledge with low confidence will hinder the grounding performance of the student. Moreover, Liu et al. [64] demonstrate that temperature scaling on the teacher can bring about more calibrated predictions. To achieve a more effective knowledge distillation, we propose a Knowledge Confidence-based Adaptive Temperature (KCAT) learning approach to dynamically learn optimal temperatures for distilling more reliable knowledge from the teacher model.

According to [60], the entropy of the teacher output logit vector can be deemed a confidence indicator. In other words, a higher entropy implies that the teacher has more confidence in specific samples, while a lower entropy indicates a higher prediction uncertainty of the teacher. Thus, we utilize the entropy of the logits of the knowledge learned by the teacher as a proxy to assess the confidence of each kind of

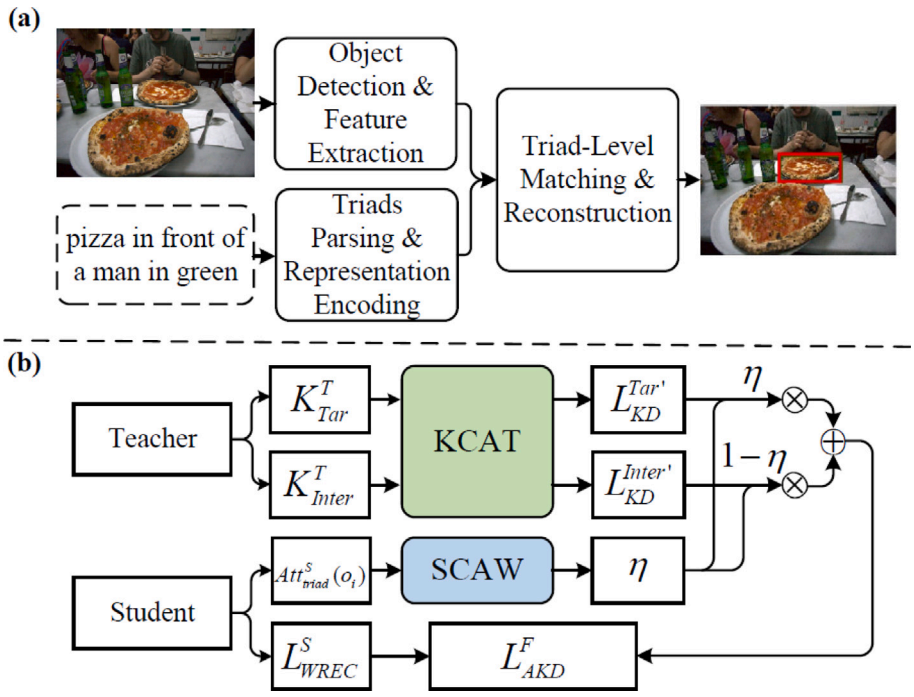


Fig. 1. Architecture diagram of the proposed adaptive knowledge distillation and integration for weakly supervised REC. (a) Teacher Grounder, which locates target objects in images by triad-level matching and reconstruction. (b) Training of the student model with adaptive knowledge distillation and integration. The proposed framework comprises a Knowledge Confidence-Based Adaptive Temperature (KCAT) learning module and a Student Competency-Based Adaptive Weight (SCAW) learning module. KCAT learns optimal temperatures to encourage the teacher to impart more reliable knowledge to the student model. SCAW urges the student to absorb more helpful knowledge by dynamically fusing the target-aware knowledge and the interaction-aware knowledge to boost the prediction certainty of the student. K^T_{Tar} and K^T_{Inter} denote the target-aware knowledge and the interaction-aware knowledge learned by the teacher grounder. $L^{Tar'}_{KD}$ and $L^{Inter'}_{KD}$ represent the target-aware knowledge distillation loss and the interaction-aware knowledge distillation loss with adaptive temperatures. $Att^S_{triad}(o_i)$ is the triad-level attention score learned by the student model, and L^S_{WREC} indicates the triad-level reconstruction loss. η is the adaptive weight learned by SCAW to balance the contribution of $L^{Tar'}_{KD}$ and $L^{Inter'}_{KD}$ for the student model training. L^F_{AKD} is the final adaptive knowledge distillation loss for training the student model.

knowledge. Specifically, we learn the adaptive temperatures for the target-aware and the interaction-aware knowledge distillation by

$$C_{Tar} = \text{Entropy}(K^T_{Tar}) = -\sum_{i=1}^{M} k^T_{Tar,i} log(k^T_{Tar,i}),$$
$$\tau'_{Tar} = \text{Sigmoid}(\varphi(C_{Tar})), \tag{14}$$

and

$$C_{Inter} = \text{Entropy}(K^T_{Inter}) = -\sum_{i=1}^{M} k^T_{Inter,i} log(k^T_{Inter,i}),$$
$$\tau'_{Inter} = \text{Sigmoid}(\varphi(C_{Inter})), \tag{15}$$

where φ denotes MLP employed to learn the dynamic temperatures.

Accordingly, the target-aware knowledge distillation loss (Eq. (9)) and the interaction-aware knowledge distillation loss (Eq. (12)) with the learned adaptive temperatures τ'_{Tar} and τ'_{Inter} are reformulated to

$$L^{Tar'}_{KD} = \text{KL}(\Phi(K^T_{Tar}/\tau'_{Tar}) \parallel \Phi(Att^S_{triad}(o_i)/\tau)), \tag{16}$$

and

$$L^{Inter'}_{KD} = \text{KL}(\Phi(K^T_{Inter}/\tau'_{Inter}) \parallel \Phi(Att^S_{triad}(o_i)/\tau)). \tag{17}$$

4.5. Student competency-based adaptive weight learning

In order to guarantee that the student learns knowledge from multiple perspectives, we transfer the target-aware and interaction-aware knowledge to promote the grounding performance of the student. Because of the different patterns used to describe specific targets in referring expressions, simply combining the target-aware and interaction-aware knowledge will introduce information redundancy and may discourage the student from learning more helpful knowledge and may bring information redundancy during knowledge distillation. Additionally, it is unnecessary to distill the knowledge mastered by the student with a high prediction certainty. Hence, we present a Student Competency-Based Weight (SCAW) learning module to dynamically adjust the weight to integrate the target-aware and interaction-aware knowledge, so that the student can effectively inherit the knowledge from the teacher.

Better student competency indicates that the student acquires higher prediction certainty on specific samples. Motivated by the uncertainty sampling policy in Active Learning [65], we employ entropy to measure the uncertainty of the instances and informative samples. Lower entropy of specific instances denotes higher prediction uncertainty on the current training samples. Thus, we utilize the student prediction uncertainty as the proxy to represent the competency of the student, and adopt the entropy of the student output logit vector to learn the adaptive weight for transferring the target-aware and interaction-aware knowledge. Concretely, we learn the competency-based adaptive weight via

$$V = \text{Entropy}(\Phi(Att^S_{triad}(o_i))),$$
$$\eta = -V / \sum_{i=1}^{M} att_{i,c} log(att_{i,c}), \tag{18}$$

where $att_{i,c}$ denotes the column vector of $Att^S_{triad}(o_i)$, and c is the dimension of $Att^S_{triad}(o_i)$.

We transfer the target-aware and interaction-aware knowledge with the learned adaptive weight η by

$$L^{TI}_{KD} = \eta L^{Tar}_{KD} + (1-\eta) L^{Inter}_{KD},$$
$$L^F_{KD} = L^S_{WREC} + \lambda L^{TI}_{KD}. \tag{19}$$

Finally, we integrate the dynamic temperature learned by KCAT with the adaptive weight obtained by SCAW to distill more reliable

Algorithm 1: Adaptive Knowledge Distillation and Integration for Weakly Supervised REC

Input: image I and expression E;
pre-trained teacher grounder T;
total iteration number $N = 150{,}000$;
current iteration i and training step $k = i/10{,}000$;
temperature hyper-parameter $\tau = 1$;
trade-off hyper-parameter λ;
EMA decay weight $\delta = 0.9997$.
Output: student model S with parameters θ_S.

while $i < N$ **do**
  for *I, E in dataloader* **do**
    $\gamma_1 s_i^t, \gamma_2 s_j^s + \gamma_3 s_{i,j}^r = T(I, E)$;
    $Att_{triad}^S(o_i) = S(I, E)$;
    $\tau'_{Tar} \leftarrow \text{Entropy}(\gamma_1 s_i^t)$; // Learn adaptive temperature for the target-aware knowledge distillation using Eq. 14
    $\tau'_{Inter} \leftarrow \text{Entropy}(\gamma_2 s_j^s + \gamma_3 s_{i,j}^r)$; // Learn adaptive temperature for the interaction-aware knowledge distillation using Eq. 15
    $L_{KD}^{Tar'} = \text{KL}(\Phi(K_{Tar}^T/\tau'_{Tar}) \parallel \Phi(s/\tau))$; // Compute target-aware knowledge distillation loss with τ'_{Tar} using Eq. 16
    $L_{KD}^{Inter'} = \text{KL}(\Phi(K_{Inter}^T/\tau'_{Inter}) \parallel \Phi(s/\tau))$; // Compute interaction-aware distillation loss with τ'_{Inter} using Eq. 17
    $\eta \leftarrow \text{Entropy}(Att_{triad}^S(o_i))$; // Learn competency-based adaptive weight using Eq. 18
    $L_{KD}^{TI} = \eta L_{KD}^{Tar'} + (1\text{-}\eta) L_{KD}^{Inter'}$;
    $L_{AKD}^F = L_{WREC}^S + \lambda L_{KD}^{TI}$. // Compute final loss using Eq. 20
  end
  $\theta_S^k \leftarrow \delta\theta_S^{k-1} + (1-\delta)\theta_S^k$ // Update student parameters using EMA
end

knowledge from the teacher to boost the grounding certainty of the student. The final training loss L_{AKD}^F of the proposed approach is formulated as

$$\begin{aligned} L_{AKD}^{TI} &= \eta L_{KD}^{Tar'} + (1-\eta) L_{KD}^{Inter'}, \\ L_{AKD}^F &= L_{WREC}^S + \lambda L_{AKD}^{TI}. \end{aligned} \tag{20}$$

Moreover, inspired by Mean Teacher [66] and its application in REC [67], we employ EMA [66] to update the parameters of the student and improve its training efficiency. The proposed adaptive knowledge distillation approach is summarized in Algorithm 1.

5. Experiments

5.1. Datasets and metric

We train and validate the proposed framework on RefCOCO [15], RefCOCO+ [15], and RefCOCOg [16]. The images of the three datasets originate from the MSCOCO dataset [68].

RefCOCO includes 19,994 images with 142,210 expressions for 50,000 referents. We adopt the UNC split introduced by [15], which divides RefCOCO into training, validation, testA, and testB sets. The training set comprises 120,624 expressions for 42,404 objects in 16,994 images, and the validation includes 10,834 expressions for 3811 objects in 1500 images. In comparison, testA has 5657 expressions for 1975 objects in 750 person-centric images, and testB possesses 5095 object-centric expressions for 1810 objects in 750 images.

RefCOCO+ contains 19,992 images with 141,564 expressions for 49,856 referents. The split is the same as RefCOCO. The training partition contains 120,191 expressions for 42,278 objects in 16,992 images, and the validation partition includes 10,758 expressions for 3,805 objects in 1500 images. Similar to the split in RefCOCO, testA has 5726 expressions for 1975 objects in 750 images, and testB comprises 4889 expressions for 1798 objects in 750 images. Compared to RefCOCO, the expressions in RefCOCO+ pay more attention to describing the attribute differences between objects.

RefCOCOg comprises 95,010 expressions for 25,799 images with 49,822 referents, and the average length of RefCOCOg referring expressions is longer than those of RefCOCO and RefCOCO+ expressions. We utilize the Google splits [16], which includes train and validation partitions. The training split contains 85,474 expressions for 44,820 objects in 21,149 images, and the validation split includes 9536 expressions for 5000 objects in 4650 images.

Evaluation Metric. We utilize the Intersection over Union (IoU) score widely employed in existing methods [12,15,19] to validate the performance of our proposed approach. We calculate the IoU score between the predicted image region and the ground truth. If the IoU score exceeds 0.5, we select the predicted region as the correct grounding.

5.2. Implementation details

We set the hyper-parameters for the experiments as follows. For the teacher model training, we select $\gamma_1 = 2$ and $\gamma_2 = \gamma_3 = 1$ in Eq. (6) to acquire the triad-level attention weight, as the target-aware attention score directly serves to ground the targets. We set $\mu_1 = \mu_2 = \mu_3 = 1$ in Eq. (7) to learn the reconstruction loss. During the knowledge distillation and the student training, we adopt the same values of γ_1, γ_2, and γ_3 within the teacher model. In addition, we use $\delta = 0.9997$ as the EMA decay weight to update the student model parameters.

We adopt the Adam optimizer with an initial learning rate 1.26e−5 to train our model. We employ an iteration-based learning rate schedule to decay the learning rate by 0.1 every 30,000 iterations. We train and evaluate the models on one single NVIDIA RTX A6000 through a total of 150,000 iterations.

5.3. Comparison with state-of-the-art

In order to demonstrate the effectiveness of the proposed framework, we compare the grounding accuracy with SOTA weakly supervised REC methods, including VC [41], ARN [10], KPRN [11], IGN [13], EARN [42], DTMR [12], and Cycle-Free [43]. The comparison results are reported in Table 1. All the listed approaches adopt the ground truth bounding box to train their models. Apart from the ground truth training, some listed models adopt different settings to acquire their best grounding accuracy on the datasets. For example, KPRN [11] utilizes $soft + attr$ setting to achieve the best performance on RefCOCO and RefCOCO+, and employs $hard + attr$ to obtain the highest grounding accuracy on RefCOCOg.

We select the teacher model, SCAW, and SCAW+KCAT with weight coefficient $\lambda = 1$ to compare with the SOTA approaches. As can be observed from Table 1, our proposed approach achieves new SOTA performance on several splits. Compared to the SOTA method Cycle-Free [43], the grounding accuracy acquired by SCAW is 1.47% lower on testA of RefCOCO. In contrast, the accuracy on testB split of RefCOCO surpasses Cycle-Free by 1.73%, and the results on testA, testB, and val of RefCOCO+ exceeds Cycle-Free by 0.93%, 0.59%, and 2.34%, respectively. Compared to the results acquired by the SCAW+KCAT, the accuracy on testB splits RefCOCO and RefCOCO+ outperforms Cycle-Free by 3.28% and 3.50%. In addition, compared with the best result of RefCOCOg [42], the grounding accuracy acquired by our approach outperforms EARN by 2.63%. Besides, SCAW and SCAW+KCAT outperform SOTA grounding accuracy on several splits. These acquired results indicate the effectiveness of our proposed approach.

Table 1
Performance (Acc%) comparison with state-of-the-art approaches on RefCOCO, RefCOCO+, and RefCOCOg. The best grounding results are in bold.

Approaches	Settings	RefCOCO			RefCOCO+			RefCOCOg
		val	testA	testB	val	testA	testB	val
VC [41]	w/o reg	–	13.59	21.65	–	18.79	24.14	25.14
	–	–	17.34	20.98	–	23.24	24.91	33.79
	w/o α	–	33.29	30.13	–	34.60	31.58	30.26
ARN [10]	L_{adp}+L_{att}	33.07	36.43	29.09	33.53	36.40	29.23	33.19
	L_{lan}+L_{adp}	33.60	35.65	31.48	34.40	35.54	32.60	34.50
	L_{lan}+L_{att}	38.05	35.27	36.47	34.51	34.40	36.12	39.62
	L_{lan}+L_{adp}+L_{att}	34.26	36.01	33.07	34.53	36.01	33.75	34.66
KPRN [11]	hard	35.04	34.74	36.53	35.10	32.75	36.76	35.44
	hard+attr	34.93	33.76	36.98	35.31	33.46	37.27	38.37
	soft	34.43	33.82	35.45	35.96	35.24	36.96	33.56
	soft+attr	36.34	35.28	37.72	37.16	36.06	39.29	36.65
IGN [13]	Base	31.05	34.39	28.16	31.13	34.44	29.59	32.17
	CCL	34.78	37.64	32.59	34.29	36.91	33.56	34.92
EARN [42]	L_{lan}+L_{adp}	35.31	37.07	32.66	35.50	37.39	33.65	38.99
	L_{lan}+L_{att}	34.93	33.76	36.98	35.31	33.46	37.27	38.37
	L_{lan}+L_{adp}+L_{att}	38.08	38.25	38.59	37.54	37.58	37.92	45.33
DTMR [12]	–	39.21	41.14	37.72	39.18	40.01	38.08	43.24
Cycle-Free [43]	–	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
	SCAW	39.71	39.99	39.69	**40.13**	**40.22**	39.93	47.75
	SCAW+KCAT	**39.84**	39.90	**40.24**	40.05	40.12	**41.09**	**47.96**

Table 2
Ablation studies on RefCOCO, RefCOCO+, and RefCOCOg with different settings to validate KCAT.

Temperature	Settings	RefCOCO				RefCOCO+				RefCOCOg
		val	testA	testB	Avg	val	testA	testB	Avg	val
–	Teacher	38.96	39.37	39.41	39.25	39.67	39.98	39.93	39.86	47.71
Fixed	Target-aware	39.51	39.83	39.43	39.59	39.83	40.19	40.29	40.10	47.69
	Interaction-aware	39.10	39.54	39.45	39.36	39.38	39.87	40.23	39.83	46.95
	SCAW	39.71	**39.99**	39.69	39.80	**40.13**	**40.22**	39.93	40.09	47.75
Adaptive	Target-aware	39.80	39.88	40.07	39.91	39.98	39.84	40.63	40.15	47.88
	Interaction-aware	38.88	39.81	39.61	39.43	39.46	39.61	40.15	39.74	47.05
	SCAW+KCAT	**39.84**	39.90	**40.24**	**39.99**	40.05	40.12	**41.09**	**40.42**	**47.96**

5.4. Ablation study

To evaluate the performance of each module, we conduct extensive ablation experiments on the three benchmark datasets.

5.4.1. Knowledge confidence-based adaptive temperature

We first evaluate the effects of KCAT by utilizing multiple settings, including the knowledge distillation with fixed temperature $\tau = 1$ to transfer the target-aware knowledge, the interaction-aware knowledge, and the adaptively fused knowledge acquired by SCAW via Eqs. (11), (13), and (19), respectively. We then substitute the fixed τ with the adaptive temperatures obtained by the KCAT module to demonstrate the benefits of KCAT for transferring the target-aware and interaction-aware knowledge, which utilizes Eqs. (16), (17), and (20) to train the models. For a fair comparison, we set $\lambda = 1$ in these experiments.

The comparison results are listed in Table 2. At first glance, the results acquired by the teacher model and the other settings demonstrate the effects on the knowledge distillation for the model grounding performance. From the comparison of the target-aware knowledge distillation with fixed temperature and adaptive temperature, we can find that KCAT improves the grounding accuracy on the splits of the three datasets except for testA split of RefCOCO+, and the average accuracy on three datasets is improved by 0.32%, 0.05%, and 0.19% respectively. By comparing the accuracy obtained by SCAW and SCAW+KCAT, we can observe that KCAT obviously enhances the grounding accuracy on the testB splits of RefCOCO and RefCOCO+, the accuracy acquired by SCAW+KCAT surpasses 0.55% and 1.16%, respectively. These comparison results demonstrate that KCAT promotes transferring knowledge with higher confidence from the teacher to the student model and further boosts the student's grounding performance.

Table 3
Grounding results on RefCOCO, RefCOCO+, and RefCOCOg with fixed distilling temperatures τ' to verify the effectiveness of KCAT.

τ'	RefCOCO			RefCOCO+			RefCOCOg
	val	testA	testB	val	testA	testB	val
0.1	**39.97**	40.11	40.06	39.54	39.43	39.54	47.69
0.5	39.87	**40.20**	**40.61**	40.05	39.91	40.91	47.92
1	39.71	39.99	39.69	**40.13**	**40.22**	39.93	47.75
5	38.96	39.08	39.39	39.72	39.63	39.99	47.36
10	37.66	36.63	39.78	39.17	38.11	40.25	46.21
SCAW+KCAT	39.84	39.90	40.24	40.05	40.12	**41.09**	**47.96**

To further validate the benefits of KCAT, we conduct experiments with fixed temperatures for distilling the knowledge from the teacher. Concretely, we utilize fixed temperature $\tau = 1$ for the student, and adjust the temperatures for transferring the target-aware and the interaction-aware knowledge from 0.1 to 10 to verify the gain of KCAT, i.e., the temperatures in Eqs. (16) and (17) are set to $\tau'_{Tar} = \tau'_{Inter} = \tau' \in \{0.1, 0.5, 1, 5, 10\}$. In these experiments, we set $\lambda = 1$ and summarize the obtained grounding results in Table 3.

As observed from Table 3, the models with $\tau' = 0.1$ and $\tau' = 0.5$ acquire the best accuracy on RefCOCO val, and RefCOCO testA and testB, respectively. However, the grounding results on RefCOCO+ and RefCOCOg are lower than that achieved by SCAW+KCAT. When τ' is set to 1, the model obtains the best accuracy on val and testA of

Table 4
Ablation studies with different η in Eq. (19) to evaluate the performance of SCAW.

η	RefCOCO			RefCOCO+			RefCOCOg
	val	testA	testB	val	testA	testB	val
0	39.10	39.54	39.45	39.38	39.87	40.23	46.95
0.3	39.26	39.56	39.35	39.85	39.49	40.48	47.24
0.5	39.07	39.61	39.55	39.66	39.96	40.42	47.64
0.8	39.32	39.81	39.49	39.89	39.85	40.05	47.66
1.0	39.51	39.83	39.43	39.83	40.19	40.29	47.69
Sum	39.44	39.37	39.21	40.00	40.05	**40.46**	47.46
SCAW	**39.71**	**39.99**	**39.69**	**40.13**	**40.22**	39.93	**47.75**

RefCOCO, but the accuracy on other splits is inferior to that obtained by SCAW+KCAT. The models with $\tau' \in \{5, 10\}$, the grounding accuracy decreases dramatically. SCAW+KCAT acquires better accuracy on each split of the datasets, and it demonstrates that KCAT encourages the teacher to transfer more reliable knowledge and impels the student to learn more helpful knowledge from the teacher.

5.4.2. Student competency-based adaptive weight

We evaluate the benefits of SCAW by setting multiple values for η in Eq. (19). In these experiments, we employ the same fixed temperature $\tau = \tau'_{Tar} = \tau'_{Inter} = 1$ to train the teacher and the student models to exclude the impact of KCAT. Specifically, we set $\eta \in \{0, 0.3, 0.5, 0.8, 1.0\}$ and $\lambda = 1$ to train the models. Note that $\eta = 1.0$ and $\eta = 0$ denote two special variants of SCAW, i.e., the target-aware knowledge distillation and the interaction-aware knowledge distillation. And $\eta = 0.5$ represents the equal contribution of the target-aware and the interaction-aware knowledge during the knowledge distillation phase. Additionally, we sum the target-aware and interaction-aware loss by removing the trade-off parameter η in Eq. (19) and denote the variation as "Sum" in Table 4.

As can be observed from Table 4, SCAW acquires the best grounding accuracy on six splits, except testB of RefCOCO+. By comparing the results listed in Line 1 and Line 6, transferring the knowledge by directly summing the target-aware and the interaction-aware knowledge will decrease the grounding accuracy on RefCOCO and RefCOCOg, whereas the "Sum" strategy improves the grounding performance on val and testA of RefCOCO+. The primary reason is that the target description pattern in the datasets contributes to the different accuracy gains. The expressions in RefCOCO+ adopt appearance discrimination and object interaction to define targets, while RefCOCO describes target objects by utilizing the object attribute and absolute location.

From the comparison between the results of several variations with different values of η and SCAW, the obtained results confirm the effectiveness of the dynamic weight for transferring the target-aware and interaction-aware knowledge. Transferring the knowledge according to the student's competency ensures the student absorbs more reliable knowledge, further boosts the student's prediction certainty and avoids bringing information redundancy during the knowledge distillation process. Moreover, SCAW outperforms the SOTA methods on six splits of the benchmark datasets. Due to that, we also select SCAW to compare with the SOTA in Table 1.

5.5. Comparison with online distillation

The proposed scheme transfers knowledge in an offline manner, where the knowledge is distilled from a pre-trained teacher model. This offline manner requires more training time and decreases knowledge distilling efficiency. In order to validate the effectiveness of the proposed approach, we also conduct experiments that transfer the target-aware and interaction-aware knowledge in an online distillation, i.e. directly transfer the multiple knowledge during the model training. In these online distillation implementations, we keep the same settings in KCAT and SCAW with the knowledge inherited from the pre-trained teacher. We summarize the obtained results in Table 5.

From Table 5, it can be observed that the model learned from the pre-trained teacher acquires better grounding accuracy on five splits of the benchmarks. In comparison, the online distilled model with SCAW surpasses the model learned from the pre-trained teacher on testA of RefCOCO by 0.21%, and the online model with SCAW+KCAT outperforms the offline model on val of RefCOCOg by 0.30%, while the average accuracy on RefCOCO acquired by the models learned from the pre-trained teacher is higher than the online one. Moreover, except for the teacher model pre-training, the training route of the online distillation model spends about 10 h on one single A6000 GPU. It is almost identical to the training duration of the student model learned from the pre-trained teacher. In a nutshell, the proposed knowledge transferring strategy demands a pre-trained teacher, but it acquires better grounding accuracy than the online distillation scheme.

5.6. Hyper-parameter analysis

In this section, we evaluate the performance of the adaptive knowledge distillation branch via setting the hyper-parameter λ in Eq. (20), which is adopted as a trade-off between the triad-level reconstruction loss and the knowledge distillation loss. In order to better indicate the effects of λ, we summarize the acquired results in Table 6, where λ varies from 0.01 to 10. As shown in Table 6, when λ is set to 1, our proposed method achieves the best average grounding accuracy on the benchmark datasets and outperforms SOTA models on six splits. Due to that, we select the results obtained by setting $\lambda = 1$ to compare with SOTA methods in Table 1.

As can be observed from Table 6, compared with the results obtained by the teacher grounder, $\lambda \in \{0.3, 0.5, 0.8, 1, 3, 5\}$ prompts the accuracy improvements on all splits of RefCOCO, RefCOCO+, and RefCOCOg, whereas $\lambda \in \{0.01, 0.05, 0.1, 8, 10\}$ decreases the grounding accuracy on some splits. The primary reason is that the different patterns of expressions in the datasets result in different performance. The expressions in RefCOCOg pay more attention to the relations among the target objects and their neighboring regions. As a result, larger λ values worsen the grounding accuracy. In contrast, the expressions in RefCOCO combine object attributes and location descriptions to define target candidates, so the best results are achieved when $\lambda \in \{0.3, 0.5, 0.8, 1, 3, 5\}$. The expressions in RefCOCO+ utilize more appearance-related phrases to depict objects rather than location descriptions, $\lambda \in \{8, 10\}$ enhances the grounding accuracy on val splits, but deteriorates the accuracy on testA and testB.

In addition, as can be observed from Table 6, the results on RefCOCO are sensitive to λ, where the performance fluctuation is more apparent than on RefCOCO+ and RefCOCOg. For RefCOCOg, the volatility contributed by our adaptive knowledge distillation framework is relatively tiny. Specifically, the margin value between the best and worst accuracy is 0.73%, while the margin values of RefCOCO val, testA, and testB are 1.90%, 1.73%, and 2.03%, respectively. These results demonstrate that the adaptive knowledge distillation promotes transferring knowledge with higher quality and confidence, and the transferred knowledge can boost the student's grounding performance.

5.7. Qualitative results

We present some qualitative visualization results on RefCOCO, RefCOCO+, and RefCOCOg in Fig. 2. The referring expressions are positioned under the related images, the grounded target objects, and the ground truth regions are denoted as solid red and green bounding boxes. The grounding results in the first row show examples comprising multiple objects of the same category. The second row shows some complex samples with long referring expressions. These results demonstrate that our approach can help ground targets in the hard samples

Table 5
Performance (Acc%) comparison with online distillation scheme. The best grounding results are in bold.

Scheme	Settings	RefCOCO			RefCOCO+			RefCOCOg
		val	testA	testB	val	testA	testB	val
Proposed	SCAW	39.71	39.99	39.69	**40.13**	**40.22**	39.93	47.75
	SCAW+KCAT	**39.84**	39.90	**40.24**	40.05	40.12	**41.09**	47.96
Online	SCAW	39.33	**40.20**	39.06	39.74	39.77	39.95	47.61
	SCAW+KCAT	39.51	40.14	39.35	40.10	39.42	40.70	**48.26**

Fig. 2. Qualitative results acquired by our proposed model on RefCOCO, RefCOCO+, and RefCOCOg. The referring expressions are positioned under the corresponding images. The solid green bounding boxes indicate the ground truth, and the red boxes are the predicted targets. We also attach some incorrect grounding samples under the dotted line.

Table 6
Grounding accuracy on RefCOCO, RefCOCO+, and RefCOCOg when the value of λ in Eq. (20) varies from 0.01 to 10.

λ	RefCOCO			RefCOCO+			RefCOCOg
	val	testA	testB	val	testA	testB	val
Teacher	38.96	39.37	39.41	39.67	39.98	39.93	47.71
0.01	37.95	38.59	38.21	39.55	39.75	39.54	47.34
0.05	38.60	39.51	38.25	39.54	39.77	40.54	48.09
0.1	39.16	39.44	38.37	39.65	40.15	39.68	47.99
0.3	39.41	39.74	39.61	39.86	**40.27**	40.34	48.04
0.5	39.41	39.97	39.27	39.78	39.82	40.11	47.93
0.8	**39.85**	**40.30**	40.10	40.02	39.70	**41.11**	47.86
1.0	39.84	39.90	**40.24**	40.05	40.12	41.09	47.96
3.0	39.73	39.97	39.51	39.95	39.91	40.48	47.73
5.0	39.70	39.92	39.43	40.12	39.82	40.21	**48.07**
8.0	39.73	39.58	39.18	40.07	39.77	39.72	47.66
10.0	39.49	39.12	38.59	**40.15**	39.68	39.62	47.59

that comprise complex expressions and multiple objects of the same category.

For comparison, we also list some incorrect grounding examples under the dotted line. One type of incorrect grounding is caused by triads erroneously parsed from the expressions. For instance, for the case "number 14" and the related image, the expression is parsed into a target element "number", a subject element "14", and a discriminative relationship element "self". The wrongly parsed triad brings challenges to ground the target region. Besides incorrect expression parsing, ambiguous expressions dramatically affect the model performance. For example, the expression "left player in back" is not enough to define the target region of the players in the left rear. Finally, the sparsity of training data in the datasets and object occlusion could lead to incorrect groundings, which also pose challenges in other vision tasks.

6. Conclusion

In this paper, we propose an adaptive knowledge distillation and integration architecture to enrich the dominant pattern of weakly

supervised REC. Specifically, the proposed adaptive knowledge distillation architecture integrates a Knowledge Confidence-Based Adaptive Temperature (KCAT) learning approach with a Student Competency-Based Adaptive Weight (SCAW) learning strategy to boost the model grounding performance. KCAT learns adaptive distilling temperatures according to knowledge confidence to transfer more reliable knowledge from the teacher to the student model. SCAW enhances the student's prediction certainty by learning dynamic weight to integrate the target-aware and interaction-aware knowledge based on the student's competency. The experimental results achieved on three benchmark datasets demonstrate that our approach outperforms SOTA models under fair comparison, and extensive ablation experiments indicate the superiority of the knowledge confidence-based adaptive temperature learning and the student competency-aware interrelated knowledge dynamic integration.

CRediT authorship contribution statement

Jinpeng Mi: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Stefan Wermter:** Writing – review & editing. **Jianwei Zhang:** Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets that have been used are public, and we published the code on Github.

Acknowledgment

We would like to thank Hugo Carneiro for his helpful advices to revise the manuscript. This work is partly funded by the German Research Foundation (DFG) and National Science Foundation (NSFC), China in the project Crossmodal Learning under contract Sonderforschungsbereich Transregio 169, and the DAAD German Academic Exchange Service under the CASY project.

References

[1] A. Gordo, J. Almazán, J. Revaud, D. Larlus, Deep image retrieval: Learning global representations for image search, in: Proceedings of European Conference on Computer Vision, ECCV, 2016, pp. 241–257.

[2] C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 43 (4) (2020) 1445–1451, http://dx.doi.org/10.1109/TPAMI.2020.2975798.

[3] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, X. Wang, M. Zhou, Visual question generation as dual task of visual question answering, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 6116–6124.

[4] L. Liu, M. Wang, X. He, L. Qing, H. Chen, Fact-based visual question answering via dual-process system, Knowl.-Based Syst. 237 (2022) 107650, http://dx.doi.org/10.1016/j.knosys.2021.107650.

[5] Y. Qi, Q. Wu, P. Anderson, X. Wang, W.Y. Wang, C. Shen, A.v.d. Hengel, REVERIE: Remote embodied visual referring expression in real indoor environments, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 9982–9991.

[6] A. Ku, P. Anderson, R. Patel, E. Ie, J. Baldridge, Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding, in: Proceedings of Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 4392–4412.

[7] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, J. Tan, Interactively picking real-world objects with unconstrained spoken language instructions, in: IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2018, pp. 3774–3781.

[8] J. Mi, J. Lyu, S. Tang, Q. Li, J. Zhang, Interactive natural language grounding via referring expression comprehension and scene graph parsing, Front. Neurorobotics 14 (2020) 43, http://dx.doi.org/10.3389/fnbot.2020.00043.

[9] M. Shridhar, D. Mittal, D. Hsu, INGRESS: Interactive visual grounding of referring expressions, Int. J. Robot. Res. 39 (2–3) (2020) 217–232, http://dx.doi.org/10.1177/0278364919897133.

[10] X. Liu, L. Li, S. Wang, Z.-J. Zha, D. Meng, Q. Huang, Adaptive reconstruction network for weakly supervised referring expression grounding, in: Proceedings of IEEE/CVF Conference on Computer Vision, ICCV, 2019, pp. 2611–2620.

[11] X. Liu, L. Li, S. Wang, Z.-J. Zha, L. Su, Q. Huang, Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding, in: Proceedings of ACM International Conference on Multimedia, ACM MM, 2019, pp. 539–547.

[12] M. Sun, J. Xiao, E.G. Lim, S. Liu, J.Y. Goulermas, Discriminative triad matching and reconstruction for weakly referring expression grounding, IEEE Trans. Pattern Anal. Mach. Intell. 43 (11) (2021) 4189–4195, http://dx.doi.org/10.1109/TPAMI.2021.3058684.

[13] Z. Zhang, Z. Zhao, Z. Lin, X. He, et al., Counterfactual contrastive learning for weakly-supervised vision-language grounding, in: Conference on Neural Information Processing Systems, Vol. 33, NeurIPS, 2020, pp. 18123–18134.

[14] G. Hinton, O. Vinyals, J. Dean, et al., Distilling the knowledge in a neural network, 2015, arXiv preprint:1503.02531 2(7).

[15] L. Yu, P. Poirson, S. Yang, A.C. Berg, T.L. Berg, Modeling context in referring expressions, in: Proceedings of European Conference on Computer Vision, ECCV, 2016, pp. 69–85.

[16] J. Mao, J. Huang, A. Toshev, O. Camburu, A.L. Yuille, K. Murphy, Generation and comprehension of unambiguous object descriptions, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 11–20.

[17] K.-P. Huang, T.-h. Feng, Y.-K. Fu, T.-Y. Hsu, P.-C. Yen, W.-C. Tseng, K.-W. Chang, H.-y. Lee, Ensemble knowledge distillation of self-supervised speech models, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2023, pp. 1–5.

[18] T. Ma, W. Tian, Y. Xie, Multi-level knowledge distillation for low-resolution object detection and facial expression recognition, Knowl.-Based Syst. 240 (2022) 108136, http://dx.doi.org/10.1016/j.knosys.2022.108136.

[19] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, K. Saenko, Modeling relationships in referential expressions with compositional modular networks, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 1115–1124.

[20] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, T.L. Berg, MAttNet: Modular attention network for referring expression comprehension, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 1307–1315.

[21] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, A.v.d. Hengel, Neighbourhood Watch: Referring expression comprehension via language-guided graph attention networks, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 1960–1968.

[22] C. Jing, Y. Wu, M. Pei, Y. Hu, Y. Jia, Q. Wu, Visual-semantic graph matching for visual grounding, in: Proceedings of ACM International Conference on Multimedia, ACM MM, 2020, pp. 4041–4050.

[23] S. Chen, B. Li, Multi-modal dynamic graph transformer for visual grounding, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 15534–15543.

[24] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: Conference on Neural Information Processing Systems, Vol. 28, NeurIPS, 2015.

[25] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, J. Luo, A fast and accurate one-stage approach to visual grounding, in: Proceedings of IEEE/CVF Conference on Computer Vision, ICCV, 2019, pp. 4683–4693.

[26] J. Redmon, A. Farhadi, YOLOv3: An incremental improvement, 2018, arXiv preprint arXiv:1804.02767.

[27] Y. Liao, S. Liu, G. Li, F. Wang, Y. Chen, C. Qian, B. Li, A real-time cross-modality correlation filtering method for referring expression comprehension, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 10880–10889.

[28] H. Qiu, H. Li, Q. Wu, F. Meng, H. Shi, T. Zhao, K.N. Ngan, Language-aware fine-grained object representation for referring expression comprehension, in: Proceedings of ACM International Conference on Multimedia, ACM MM, 2020, pp. 4171–4180.

[29] J. Ye, X. Lin, L. He, D. Li, Q. Chen, One-stage visual grounding via semantic-aware feature filter, in: Proceedings of ACM International Conference on Multimedia, ACM MM, 2021, pp. 1702–1711.

[30] G. Luo, Y. Zhou, X. Sun, L. Cao, C. Wu, C. Deng, R. Ji, Multi-task collaborative network for joint referring expression comprehension and segmentation, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 10034–10043.

[31] M. Sun, W. Suo, P. Wang, Y. Zhang, Q. Wu, A proposal-free one-stage framework for referring expression comprehension and generation via dense cross-attention, IEEE Trans. Multimed. 1 (2022) 1–13, http://dx.doi.org/10.1109/TMM.2022.3147385.

[32] B. Huang, D. Lian, W. Luo, S. Gao, Look Before You Leap: Learning landmark features for one-stage visual grounding, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 16888–16897.
[33] J.D.M.-W.C. Kenton, L.K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, NACCAL-HLT, 2019, pp. 4171–4186.
[34] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, VL-BERT: Pre-training of generic visual-linguistic representations, in: The International Conference on Learning Representations, ICLR, 2020.
[35] J. Deng, Z. Yang, T. Chen, W. Zhou, H. Li, TransVG: End-to-end visual grounding with transformers, in: Proceedings of IEEE/CVF Conference on Computer Vision, ICCV, 2021, pp. 1769–1779.
[36] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, N. Carion, MDETR-modulated detection for end-to-end multi-modal understanding, in: Proceedings of IEEE/CVF Conference on Computer Vision, ICCV, 2021, pp. 1780–1790.
[37] H. Zhao, J.T. Zhou, Y.-S. Ong, Word2Pix: Word to pixel cross-attention transformer in visual grounding, IEEE Trans. Neural Netw. Learn. Syst. 1 (2022) 1–11, http://dx.doi.org/10.1109/TNNLS.2022.3183827.
[38] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, H. Yang, OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, in: International Conference on Machine Learning, ICML, 2022, pp. 23318–23340.
[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: Conference on Neural Information Processing Systems, Vol. 30, NeurIPS, 2017.
[40] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, B. Schiele, Grounding of textual phrases in images by reconstruction, in: Proceedings of European Conference on Computer Vision, ECCV, 2016, pp. 817–834.
[41] Y. Niu, H. Zhang, Z. Lu, S.-F. Chang, Variational Context: Exploiting visual and textual context for grounding referring expressions, IEEE Trans. Pattern Anal. Mach. Intell. 43 (1) (2019) 347–359, http://dx.doi.org/10.1109/TPAMI.2019.2926266.
[42] X. Liu, L. Li, S. Wang, Z.-J. Zha, Z. Li, Q. Tian, Q. Huang, Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding, IEEE Trans. Pattern Anal. Mach. Intell. 45 (3) (2023) 3003–3018, http://dx.doi.org/10.1109/TPAMI.2022.3186410.
[43] M. Sun, J. Xiao, E.G. Lim, Y. Zhao, Cycle-free weakly referring expression grounding with self-paced learning, IEEE Trans. Multimed. 25 (2023) 1611–1621, http://dx.doi.org/10.1109/TMM.2021.3139467.
[44] L. Jin, G. Luo, Y. Zhou, X. Sun, G. Jiang, A. Shu, R. Ji, RefCLIP: A universal teacher for weakly supervised referring expression comprehension, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2681–2690.
[45] B. Zhao, Q. Cui, R. Song, Y. Qiu, J. Liang, Decoupled knowledge distillation, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 11953–11962.
[46] Z. Long, F. Ma, B. Sun, M. Tan, S. Li, Diversified branch fusion for self-knowledge distillation, Inf. Fusion 90 (2023) 12–22, http://dx.doi.org/10.1016/j.inffus.2022.09.007.
[47] S. Ahn, S.X. Hu, A. Damianou, N.D. Lawrence, Z. Dai, Variational information distillation for knowledge transfer, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 9163–9171.
[48] K. Shuang, Q. Yang, J. Loo, R. Li, M. Gu, Feature distillation network for aspect-based sentiment analysis, Inf. Fusion 61 (2020) 13–23, http://dx.doi.org/10.1016/j.inffus.2020.03.003.
[49] H. Yang, G. Jeon, K. Liu, Y. Liu, X. Yang, Feature similarity rank-based information distillation network for lightweight image superresolution, Knowl.-Based Syst. 266 (2023) 110437, http://dx.doi.org/10.1016/j.knosys.2023.110437.
[50] C. Li, G. Cheng, G. Wang, P. Zhou, J. Han, Instance-aware distillation for efficient object detection in remote sensing images, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–11, http://dx.doi.org/10.1109/TGRS.2023.3238801.
[51] W. Park, D. Kim, Y. Lu, M. Cho, Relational knowledge distillation, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 3967–3976.
[52] C. Tan, J. Liu, X. Zhang, Improving knowledge distillation via an expressive teacher, Knowl.-Based Syst. 218 (2021) 106837, http://dx.doi.org/10.1016/j.knosys.2021.106837.
[53] C. Yang, L. Xie, C. Su, A.L. Yuille, Snapshot Distillation: Teacher-student optimization in one generation, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 2859–2868.
[54] H. Zhao, X. Sun, J. Dong, Z. Dong, Q. Li, Knowledge distillation via instance-level sequence learning, Knowl.-Based Syst. 233 (2021) 107519, http://dx.doi.org/10.1016/j.knosys.2021.107519.
[55] Q. Li, Q. Hu, S. Qi, Y. Qi, D. Wu, Y. Lin, J.S. Dong, Stochastic ghost batch for self-distillation with dynamic soft label, Knowl.-Based Syst. 241 (2022) 107936, http://dx.doi.org/10.1016/j.knosys.2021.107936.
[56] X. Lan, X. Zhu, S. Gong, Knowledge distillation by on-the-fly native ensemble, in: Conference on Neural Information Processing Systems, Vol. 2, NeurIPS, 2018.
[57] N. Dvornik, C. Schmid, J. Mairal, Diversity with Cooperation: Ensemble methods for few-shot classification, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 3723–3731.
[58] S. Du, S. You, X. Li, J. Wu, F. Wang, C. Qian, C. Zhang, Agree to Disagree: Adaptive ensemble knowledge distillation in gradient space, in: Conference on Neural Information Processing Systems, Vol. 33, NeurIPS, 2020, pp. 12345–12355.
[59] Y. Liu, W. Zhang, J. Wang, Adaptive multi-teacher multi-level knowledge distillation, Neurocomputing 415 (2020) 106–113, http://dx.doi.org/10.1016/j.neucom.2020.07.048.
[60] K. Kwon, H. Na, H. Lee, N.S. Kim, Adaptive knowledge distillation based on entropy, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2020, pp. 7409–7413.
[61] C. Li, G. Cheng, J. Han, Boosting knowledge distillation via intra-class logit distribution smoothing, IEEE Trans. Circuits Syst. Video Technol. 1 (2023) 1–12, http://dx.doi.org/10.1109/TCSVT.2023.3327113.
[62] J. Pennington, R. Socher, C.D. Manning, GloVe: Global vectors for word representation, in: Proceedings of Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.
[63] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.
[64] Z. Zhang, M. Sabuncu, Self-distillation as instance-specific label smoothing, in: Conference on Neural Information Processing Systems, Vol. 33, NeurIPS, 2020, pp. 2184–2195.
[65] B. Settles, Active learning literature survey, Comput. Sci. Tech. Rep. (2009) 1–47.
[66] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Conference on Neural Information Processing Systems, Vol. 30, NeurIPS, 2017.
[67] G. Luo, Y. Zhou, J. Sun, S. Huang, X. Sun, Q. Ye, Y. Wu, R. Ji, What goes beyond multi-modal fusion in one-stage referring expression comprehension: An empirical study, 2022, arXiv preprint:2204.07913.
[68] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, in: Proceedings of European Conference on Computer Vision, ECCV, 2014, pp. 740–755.