Learning from Box Annotations for Referring Image Segmentation

Guang Feng, Lihe Zhang, Zhiwei Hu, Huchuan Lu

***Abstract*—Referring image segmentation (RIS) has obtained an impressive achievement by Fully Convolutional Networks (FCNs). However, previous RIS methods require a large number of pixel-level annotations. In this paper, we present a weakly-supervised RIS method by using bounding box annotations. In the first stage, we introduce an adversarial boundary loss to extract the object contour from the bounding box, which is then used to select appropriate region proposals for pseudo ground-truth generation. In the second stage, we design a co-training strategy to purify the pseudo labels. Specifically, we train two networks and interactively guide them to pick clean labels for each other's networks, which can weaken the effect of noisy labels on model training. Experiment results on four benchmark datasets demonstrate that the proposed method can produce high-quality masks with a speed of 63 FPS.**

***Index Terms*—Weakly-supervised referring image segmentation, bounding box annotation, adversarial boundary loss, co-training strategy.**

I. Introduction

Referring image segmentation (RIS) is a cross-modal task that combines vision and language. It aims to segment the visual region (object or stuff) related to language description. RIS segments the image into two categories: foreground and background, and the definition of the category is only related to the referring expression, not the predefined semantic categories. Therefore, RIS not only needs a deep understanding of the semantic context of images, but also considers the mutual embedding between visual and linguistic information.

Benefit from the development of deep learning technology, the performance of RIS has achieved great progress [1]–[8]. However, these works are designed in a fully supervised manner, the annotation work of the segmentation mask is time-consuming and laborious. Recently, some weakly-supervised image segmentation methods [9]–[14] have been proposed to alleviate the need for a large quantity of pixel-level class labels. Among them, weak label can be image-level tag [9], [10], [12], bounding box [11], [13], scribble [14], etc. But in the field of RIS, weakly-supervised learning is still a blank. Considering the particularity of this task, the image-level tag usually does not accurately point to the target region (e.g., as shown in Fig. 1, the foreground region is the bird on the left, but the image-level label possibly describes all birds). While both bounding box or scribble can indicate the location information of the foreground. In this work, we deeply investigate the usage of bounding box to train the RIS network.

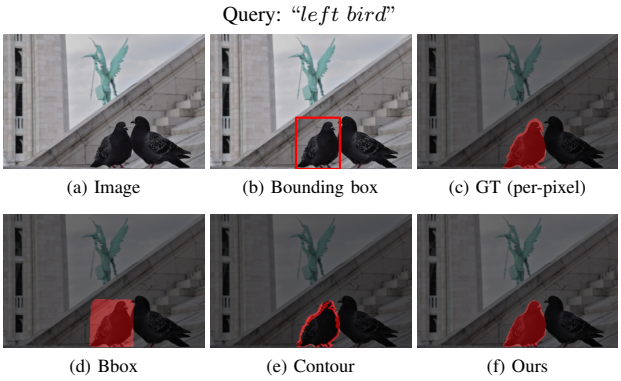

Fig. 1: Weakly supervised segmentation with the box-level annotations. (a) Input image. (b) Bounding box. (c) Pixel-wise ground-truth. (d) Result of using the cross entropy loss on the box-level annotations. (e) Result of using the adversarial boundary loss. (f) Our final result.

There have been some works [11], [13], [15]–[18] dedicated to using bounding box annotation to realize weakly-supervised image segmentation. Among them, some methods such as BoxSup [15] and Box2Seg [18] mainly rely on region proposals to generate pseudo labels, while other methods [13], [17] apply global constraints to define the loss function and directly use them to train the network. However, these proposal-based methods cannot capture the general shape of object region, which makes them only be able to ensure that the picked proposals belong to the bounding box as much as possible. Thus, due to the diversity of object semantics and background appearances as well as the varied shape and size of proposals, it is extremely difficult to select the proposals most similar to the ground truth. Besides, although global constraint based models achieve the end-to-end training, they lack an explicit description of the object contour, which eventually leads to ambiguous prediction on the boundary pixels.

In this work, we present a novel box-supervised method for referring image segmentation. It is well known that good proposals can provide important object-level prior knowledge for RIS, but bounding box annotations cannot directly perceive the shape of the foreground, it is not enough to accurately select the proposals that meet the requirements. A reasonable assumption is *if the contour of the foreground can be inferred from the bounding box, we can use it as a powerful prior*

This work was supported by the National Key R&D Program of China #2018AAA0102000, the National Natural Science Foundation of China #61876202, the Liaoning Natural Science Foundation #2021-KF-12-10, and the Fundamental Research Funds for the Central Universities #DUT20ZD212.

G. Feng, L. Zhang, Z. Hu and H. Lu are with School of Information and Communication Engineering, Dalian University of Technology, Dalian, China (e-mail: fengguang.gg@gmail.com, zhanglihe@dlut.edu.cn, hzw950822@mail.dlut.edu.cn, lhchuan@dlut.edu.cn).

to filter region proposals. With this in mind, we propose an adversarial boundary loss. It includes two terms: one is used to promote the main body of the object to have higher activation, while the other is used to reversely inhibit high activation region. Under their confrontation, the main parts of the object are suppressed, while the contour of the object is highlighted. We utilize the learned contour to pick out only a very few proposals and generate pseudo ground-truth.

The pseudo labels contain noisy information, which can worsen the generalization of the model. To alleviate this problem, we design a learning paradigm based on the co-training strategy to filter out the noisy labels. Specifically, we use the cross-entropy loss as the basis to determine the confidence of the pseudo labels, and then select the pixels with small loss for back-propagation. Meanwhile, we use two networks to mutually guide the back-propagation of their peer networks, which avoids the sample selection bias that may appear in the self-training of single network.

Our main contributions are as follows:

- We design an adversarial boundary loss to capture the contour of the foreground, which is trained based on bounding box annotations. The learned contour is used to filter proposals to obtain accurate pseudo labels.
- We introduce a co-training strategy to purify pseudo labels, which promotes two networks to mutually guide their peer networks, thereby reducing the influence of the wrongly labeled pixels in back-propagation.
- Extensive experiments on UNC, UNC+, Google-Ref, and ReferIt demonstrate that the proposed method is effective and achieves comparable accuracy than fully supervised counterparts. And the model runs at about 63 FPS without any post-processing.

II. Related Work

A. Fully Supervised RIS

RIS is a cross-modal image segmentation task, and its segmentation result is the visual region referred to the referring expression. Compared with traditional image segmentation tasks, RIS is more challenging because it needs to realize the mutual matching between various types of linguistic and visual information, such as entities, attributes and relationships. Early methods [1], [2], [19], [20] directly concatenate linguistic and visual features to realize the feature fusion between different modalities. But these works lack deep interaction between multi-modal features. Later, some works [3], [4], [6], [7], [21]–[26] utilize attention mechanisms to implement the context modeling of language and vision. Shi et al. [3] adopt query attention to extract key-words corresponding to each visual region. This method considers the relationship between multi-modal features for the first time. Later, CMSA [4], [27] models the dependency relationship between all of the word-vision mixed features in a fully connected way. BRINet [6], [28] learns a bi-directional cross-modal attention module to realize the mutual guidance of language and visual features. CMPC [22] first perceives all the entities in the image based on the entity and attribute words, and then model the relationships of all entities using relational words. LSCM [7] builds a word graph based on dependency parsing tree to construct the cross-modal context. CEFNet [29] employs a co-attention mechanism to realize the parallel update of language and visual features, which promotes the consistency of language and vision in the semantic space. And this method adopts the encoder fusion strategy for the first time, thus realizing the gradual guidance of language to vision. In addition, Lang2Seg [30] uses the consistency constraints of language and visual to promote the fusion between multi-modal features. STEP [31] uses multiple ConvLSTM to densely integrate five levels of cross-modal features. Although the above fully supervised RIS has achieved competitive performance, it is expensive and labor-intensive to annotate the pixel-level masks. In contrast, the proposed method only relies on bounding box annotations for training.

B. Box-Supervised Image Segmentation

A few works [11], [13], [15]–[18], [32] attempt to use box-level annotations to train semantic segmentation or instance segmentation networks. These weakly-supervised methods can be roughly divided into two categories. One is to use some unsupervised methods to generate pixel-wise pseudo labels, and then use them to train the segmentation model. The other is to utilize box annotations to derive the global constraints and establish a new loss function directly. For example, BoxSup [15] employs a recursive training process, which first generates a set of candidate segments, and then uses these candidates to train the semantic segmentation network. Subsequently, the learned semantic features are used to pick out better candidates. SDI [16] also adopts an iterative training procedure to refine the segmentation mask. Song et al. [11] use unsupervised dense CRF to generate proposals, and then a filling rate guided adaptive loss is proposed to train the network. The above methods all rely on the bounding box as a priori to generate pixel-level pseudo labels. When the semantics in the bounding box is complex, it is difficult for these methods to guarantee the overlap rate of the proposal and the object mask. Recently, Kervadec et al. [13] propose to use the global constraints derived from the bounding box annotations to establish a loss function, which no longer depends on the pixel-wise mask loss. BBTP [17] converts the bounding box supervised instance segmentation into a multiple instance learning (MIL) task. BBTP generates the positive and negative bags according to the tightness prior for MIL. All these methods do not explicitly consider the boundary information of object. In contrast, our method defines an adversarial boundary loss to predict a coarse contour of object, and then combines bounding box and the predicted contour as a priori to filter the candidate proposals, which makes the proposals having high overlap with the object be selected.

C. Learning with Noisy Labels

Few works [11], [18] regard box-supervised segmentation as a noisy label learning problem. Song et al. [11] first use Dense-CRF [33] as an assistant to count the filling rate of each class of samples, and then a filling-rate guided loss is used to train the model. Furthermore, Kulharia et al. [18] presents

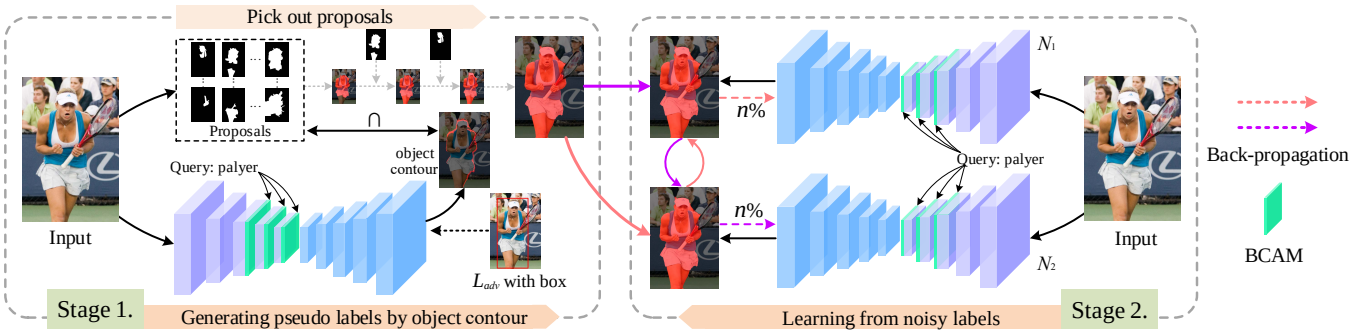

Fig. 2: Overview of the proposed method. In the first stage, we apply the adversarial boundary loss ($\mathcal{L}_{adv} = \mathcal{L}_{tight} + \lambda \cdot \mathcal{L}_{zero}$) to generate the contour of object, and then utilize it to select appropriate proposals to construct pseudo labels. In the second stage, we design two networks to control the back-propagation of each pixel position of their peer networks, which can filter out the wrong pixel-level labels in the pseudo ground truth. ∩: intersection, BCAM: bi-directional cross-modal attention module.

a soft constraint to control the filtering rate of each sample. These methods are similar to the self-training approach [34]. They rely on the model itself to pick out clean labels, but they also inherited the same inferiority of accumulated error caused by the sample selection bias. Different from them, we adopt two networks to interactively determine useful labels for their peer networks. These two networks provide different views, which can filter different types of noise and avoid the rapid accumulation of errors.

III. Proposed Method

A. Overview

Our pipeline is built on the bi-directional relationship inferring network (BRINet) [6]. In BRINet, the core is the bi-directional cross-modal attention module (BCAM), which can effectively model the relationship between features of different modalities. Specifically, BCAM consists of a vision-guided linguistic attention module (VLAM) and a language-guided visual attention module (LVAM). VlAM aims to learn the adaptive language context corresponding to each visual region. LVAM uses the learned adaptive language context to guide the update of each visual region. Through their joint action, BCAM realizes the mutual embedding between cross-modal features. But different from BRINet [6], we adopt the encoder fusion strategy, that is, the BCAM is inserted into the ResNet101 [35] backbone in a residual manner, as shown in Fig. 2. This operation can realize the deep interweaving between multi-modal features. The decoder adopts the FPN [36] structure. The proposed box-supervised method is mainly divided into two stages. In the first stage, we present an adversarial boundary loss to capture the object contour, and then combine it and bounding box to select appropriate proposals. These picked proposals are fused as pseudo labels for the subsequent training. In the second stage, to alleviate the effect of the noisy pseudo labels, we train two new segmentation networks simultaneously and they guide each other to block the back-propagation of false labels.

B. Pseudo Label Generation

For weakly referring image segmentation, a general idea is to utilize unsupervised region proposal methods (e.g. MCG [37], GOP [38], etc.) to generate a set of segments, and then use the box-level annotation as a priori to select the qualified candidate mask. But the bounding box can only provide the location information of the object, it cannot describe its shape. Therefore, at the boundary of object, the matching rate between the object mask and the selected proposals (segments) cannot be guaranteed. The wrong boundary information provided by these proposals will affect the training of the network. To this end, we try to capture the contour of object and use it as a priori to pick out the proposals.

1) Object contour prediction: We design an adversarial boundary loss function, which can enforce the network to learn a rough contour of object under the supervision of bounding box. First, the tightness prior assume that the target region is sufficiently close to the sides of its box annotation [39]. Inspired by this property, it can infer that at least one pixel in each row or column of the bounding box mask belongs to the foreground. For the binary classification (foreground/background) task of RIS, we define $C \in [0,1]^{H \times W}$ as the prediction map of the network in the first stage, where 1 and 0 represent the contour and non-contour, respectively. Let C^i_{row} and C^j_{col} represent the i^{th} row and the j^{th} column of the prediction map, respectively. We define $P_{row}(i) = \max(C^i_{row})$ and $P_{col}(j) = \max(C^j_{col})$. $P_{row}(i)$ and $P_{col}(j)$ compute the maximum value along the i^{th} row and j^{th} column of the prediction map C, respectively. When the box is across the i^{th} row or j^{th} column, we think that this row or column passes the foreground region. The prediction value of $P_{row}(i)$ or $P_{col}(j)$ should be close to $\mathbf{1}$. Otherwise, $P_{row}(i)$ and $P_{col}(j)$ should be close to $\mathbf{0}$. According to this prerequisite, we define the

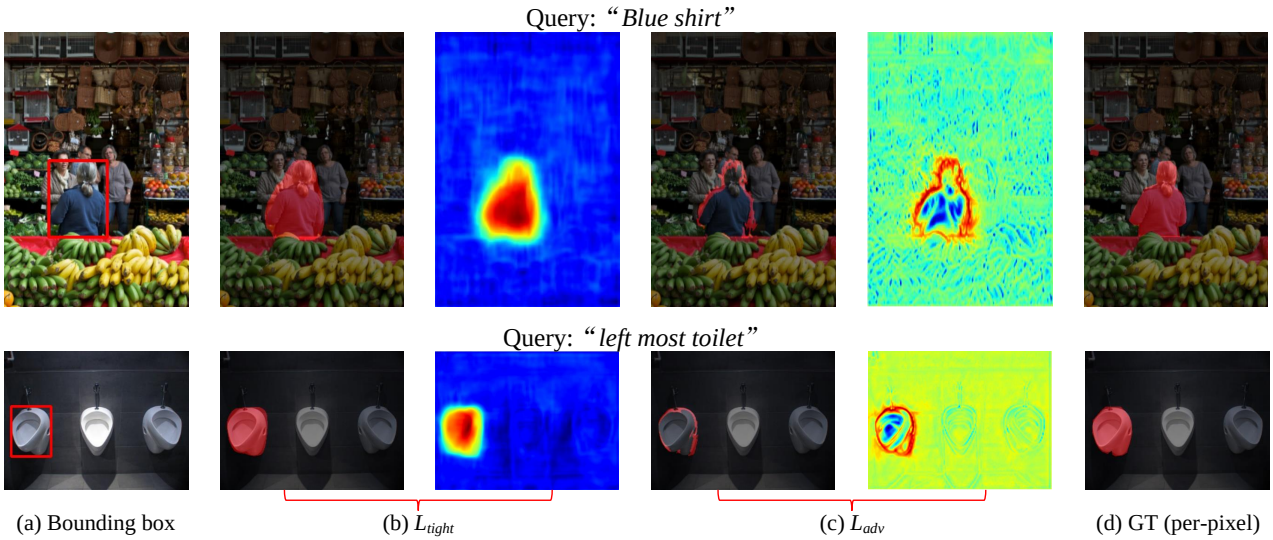

Fig. 3: Visual examples based on different loss functions.

loss function as:

$$\mathcal{L}_{tight} = \sum_{C^i_{row}, C^j_{col} \cap \mathcal{B} \notin \varnothing} -[\log(P_{row}(i)) + \log(P_{col}(j))] + \sum_{C^i_{row}, C^j_{col} \cap \mathcal{B} \in \varnothing} -[\log(1 - P_{row}(i)) + \log(1 - P_{col}(j))], \tag{1}$$

where $\mathcal{B}$ represents the rectangular region surrounded by the bounding box. $\mathcal{L}_{tight}$ can directly supervise the segmentation network based on bounding box annotations. In Eq. 1, the row-wise constraints, i.e., $\sum_{C^i_{row} \bigcap \mathcal{B} \notin \varnothing} -\log(\max(C^i_{row}))$ and $\sum_{C^i_{row} \bigcap \mathcal{B} \in \varnothing} -\log(1 - \max(C^i_{row}))$, ensure that the maximum activation region locates these rows which intersect the bounding box (BB). Meanwhile, the column-wise constraints will restrict that the maximum activation region locates these columns which intersect with the BB. Under their mutual actions, $\mathcal{L}_{tight}$ finally drives the network to predict the foreground within the BB. Then we define a zero constraint as:

$$\mathcal{L}_{zero} = \sum_{i=1}^{H} \sum_{j=1}^{W} -\log(1 - C_{i,j}), \tag{2}$$

where H and W represent the height and width of the input image, respectively. The purpose of $\mathcal{L}_{zero}$ is to make the prediction map tend to be all 0. We combine the two loss functions to form the adversarial boundary loss:

$$\mathcal{L}_{adv} = \mathcal{L}_{tight} + \lambda \cdot \mathcal{L}_{zero}, \tag{3}$$

where the balance parameter λ is empirically set to 0.05. This is because that a larger balance parameter will make the prediction map completely close to 0. Besides, $\mathcal{L}_{tight}$ is a global constraint and $\mathcal{L}_{zero}$ is a local constraint. Their gradients are in different orders of magnitude.

Fig. 3 shows some visual examples based on different loss functions. The training results based on $\mathcal{L}_{tight}$ are shown in Fig. 3 (b). From the heat map, it can be found that $\mathcal{L}_{tight}$ drives the network to activate the main body of the object, and the non-core region of the object presents a lower activation value. While the high-activation region presents larger gradient in $\mathcal{L}_{zero}$. Therefore, with their interaction, the high-activation region is easily suppressed and the region with relatively low activation is highlighted. Moreover, from Fig. 3 (b), it can be seen that the segmentation result based on $\mathcal{L}_{tight}$ is often larger than the object itself. Because these falsely detected parts do not belong to the foreground, they are easy to be suppressed during the back propagation of $\mathcal{L}_{zero}$. Therefore, by balancing $\mathcal{L}_{zero}$ and $\mathcal{L}_{tight}$ with a suitable parameter λ, the learned prediction map presents a hollow form (as shown in Fig. 3 (c)). We can finally learn a acceptable contour.

2) Proposal Selection: We first adopt a unsupervised proposal method [38] to generate some region proposals. Then, the predicted contour mask is utilized to select only a few suitable proposals, so that their fusion result can cover the object as precisely as possible. Formally, we define the following objective function:

$$\arg\max_{y} \{\mathcal{C} \cap b(\bigcup_{p_i \in \mathcal{P}} y_i \cdot p_i)\} \quad s.t. \quad y_i = 0 \quad or \quad 1, \mathcal{P} \subseteq \text{box}, \tag{4}$$

where $\mathcal{C}$ denotes a set of the pixels at the contour, which is obtained by binarizing the prediction map C. $\mathcal{P}$ denotes the set of proposals. b represents the boundary extraction operator. To extract the boundary of the binary mask constructed by the selected proposals, we first use dilation and erosion operations to process the fused mask respectively, and then the boundary map can be achieved by subtracting the eroded mask from the dilated mask. If the proposal p_i is selected, we set $y_i = 1$, otherwise $y_i = 0$. The union of all chosen proposals is taken as the pseudo ground-truth mask. Eq. (4) means that the boundary of the pseudo label has the largest intersection with the contour of object. This NP-hard problem can be solved by a greedy algorithm described in Alg. 1. Our greedy algorithm starts from a seed proposal sp, and its boundary has the largest intersection with the object contour $\mathcal{C}$. Next, it utilizes the sp as the basis to add (remove) the proposal to the set $\mathcal{S}$ until

Algorithm 1: Optimization Process of Eq. 4.

Input: a set of the pixels of the contour $\mathcal{C}$, a set of proposals $\mathcal{P}$, i^{th} proposal p_i, number of all proposals T;
Definition: b: boundary extraction operator, sp: seed proposal, $\mathcal{S}$: a set of selected proposals;

$sp = p_1$;
for i=2, 3 ,..., T **do**
  if $|\mathcal{C} \cap b(p_i)| > |\mathcal{C} \cap b(sp)|$ and $p_i \subseteq$ box **then**
    $sp = p_i$;
  end
end
$\mathcal{S} \leftarrow sp$;
for i=1, 2 ,..., T **do**
  if $|\mathcal{C} \cap b(p_i \cup sp)| > |\mathcal{C} \cap b(sp)|$ and $p_i \subseteq$ box **then**
    $\mathcal{S} \leftarrow \mathcal{S} \cup p_i$;
  end
end
for i=1, 2 ,..., T **do**
  if $|\mathcal{C} \cap b(\mathcal{S} \setminus p_i)| > |\mathcal{C} \cap b(\mathcal{S})|$ and $p_i \subseteq$ box **then**
    $\mathcal{S} \leftarrow \mathcal{S} \setminus p_i$;
  end
end
Output: $\mathcal{S}$.

the boundary of the set $\mathcal{S}$ has the largest intersection with the object contour.

3) Difference between adversarial boundary loss ($\mathcal{L}_{adv}$) and MIL loss: The MIL Loss [17] is built on a local ROI region, which does not capture the information outside this region. In contrast, the $\mathcal{L}_{adv}$ is built on the full-resolution prediction map, which is a global constraint and avoids the process of sampling positive and negative bags. It is more flexible and lightweight. Moreover, MIL is directly used to train the segmentation network, while $\mathcal{L}_{adv}$ uses $\mathcal{L}_{tight}$ and $\mathcal{L}_{zero}$ together to enforce the network to capture the contour of the object. Then we can use the learned contour to derive a more accurate pseudo ground truth (PGT).

C. Learning from Noisy Labels

The pseudo label inevitably includes noise. In order to decrease its negative effect on network training, we attempt to select the high-confidence pixel-level labels to participate in the supervision. Generally speaking, people are more likely to catch out others' error or bug but ignore their own's as a result of personal bias. Therefore, the self-training in a single network easily transfers the biased selection from the network back to itself, thereby accelerating the accumulation of errors. To conquer this problem, we build two networks to interactively determine useful labels for their peer networks. The co-training between the two networks can provide diverse and complementary information about the samples and filter out different noise. Our method is presented in Alg. 2. First, for two networks N_1 and N_2, we calculate and sort the cross-entropy loss of each pixel in a batch based on the pseudo label M. We consider that the pixel with smaller loss

Algorithm 2: Co-training algorithm.

Input: image I_t, language expression L_t, retention ratio $R(t)\%$, iteration t, pseudo labels M, networks N_1 and N_2;
Definition: $\mathcal{L}(\cdot,\cdot)$: cross-entropy loss; argsort: Return the pixel index in ascending order of the loss;

for t=1, 2 ,..., T **do**
  idx$_1$ = argsort$\big(\mathcal{L}(N_1(I_t, L_t),\ M)\big)$ [1: top $R(t)\%$];
  idx$_2$ = argsort$\big(\mathcal{L}(N_2(I_t, L_t),\ M)\big)$ [1: top $R(t)\%$];
  Update N_1 with M and idx$_2$;
  Update N_2 with M and idx$_1$;
end
Output: Updated N_1 and N_2.

is more likely to be correctly labeled. Then we record the index of the top $R(t)\%$ smallest-loss pixels for each network. Finally, the index is used to control the back-propagation of the corresponding position in its peer network. In general, this strategy employs two models with different initialization parameters, which provide different views (two independent sets of features) for image referring segmentation. Therefore, their mutual supervision can deal with heavier noise.

Retention Ratio: The deep model first learns the simple and general patterns of the real data, and then gradually fit the noise [40]. Thus we can use a larger retention ratio $R(t)\%$ at the beginning of training to better fit the real data. Specifically, we set $R(t) = 1 - \eta \cdot (t/T)$, where t denotes the iterations. The maximum iterations T is set to 50,000. The noise level η is empirically set to 0.1.

IV. Experiments

In this section, we first introduce the datasets and evaluation metrics. Next, the implementation details of our method are described. Lastly, the performance comparison and ablation study are presented. And we strictly follow the experimental setup of previous methods for training and testing to ensure the fairness.

A. Datasets and Evaluation Metrics

To evaluate the effectiveness of the proposed method, we conduct extensive experiments on four public referring image segmentation (RIS) datasets: UNC [42], UNC+ [42], Google-Ref [43] and ReferIt [44].

1) UNC: The UNC consists of 19,994 images with 142,209 language expressions for 50,000 objects regions. All its images and expressions build upon the MS COCO dataset using a two-player game [44]. Usually, these selected images contain two or more objects of the same category.

2) UNC+: Similar to the UNC dataset, UNC+ is also collected from the MS COCO dataset. UNC+ contains 141,564 referring expressions for 49,856 objects in 19,992 images. But different from UNC, its language expression does not contain words that indicate location information. Therefore, this expression puts forward higher requirements for the cross-modal fusion model.

TABLE I: Quantitative evaluation of different methods on four datasets. -: no data available. D: DenseCRF [33] post-processing.

	Methods	Resolution	Backbone	ReferIt	UNC			UNC+			G-Ref	FPS↑
				test	val	testA	testB	val	testA	testB	val	
Fully	LSTM-CNN$_{16}$ [1]	320×320	ResNet101	48.03	-	-	-	-	-	-	28.14	18
	RMI+D$_{17}$ [2]	320×320	ResNet101	58.73	45.18	45.69	45.57	29.86	30.48	29.50	34.52	14
	DMN$_{18}$ [20]	320×320	ResNet101	52.81	49.78	54.83	45.13	38.88	44.22	32.29	36.76	-
	KWA$_{18}$ [3]	320×320	ResNet101	59.19	-	-	-	-	-	-	36.92	-
	RRN+D$_{18}$ [19]	320×320	ResNet101	63.63	55.33	57.26	53.95	39.75	42.15	36.11	36.45	24
	MAttNet$_{18}$ [41]	∼1000×600	ResNet101	-	56.51	62.37	51.70	46.67	52.39	40.08	-	-
	lang2seg$_{19}$ [30]	320×320	ResNet101	-	58.90	61.77	53.81	-	-	-	-	-
	CGAN$_{20}$ [8]	416×416	DarkNet53	-	59.25	62.37	53.94	46.16	51.37	38.24	46.54	-
	CMSA+D$_{21}$ [27]	320×320	ResNet101	63.99	58.92	62.01	56.03	43.87	47.79	38.33	42.84	13
	BRINet+D$_{21}$ [6]	320×320	ResNet101	63.46	61.35	63.37	59.57	48.57	52.87	42.13	48.04	9
	CMPC+D$_{21}$ [22]	320×320	ResNet101	65.53	61.36	64.54	59.64	49.56	53.44	43.23	49.05	17
Weakly	Cross Entropy	320×320	ResNet101	49.68	42.01	42.10	42.79	36.10	37.41	33.82	36.62	63
	Ours	320×320	ResNet101	62.44	58.01	60.52	55.48	47.12	50.86	40.26	46.03	63
	Ours+D	320×320	ResNet101	62.83	58.83	61.31	56.01	47.85	51.54	40.73	46.42	-

(a) Image (b) cut off guy (c) black and white sweater in corner (d) Image (e) the blue cup on the top right (f) red sauce bowl

(g) Image (h) the chocolate covered donut in the middle (i) donut on left in front (j) bottom row of donuts second from left (k) Image (l) the woman with black hair and a black shirt

(m) Image (n) ceiling (o) Image (p) water below boat (q) Image (r) lower right corner

Fig. 4: Visual examples of referring image segmentation by our method.

3) Google-Ref: The G-Ref dataset is still derived from MS COCO. The annotations are based on Mechanical Turk instead of using a two-player game. The expression of G-ref is usually long, and the average length of sentences is 8.43 words. In G-Ref, it contains 26,711 images with 104,560 expressions for 54,822 segmented object regions.

4) ReferIt: ReferIt is collected from the IAPR TC-12 [45]. It has 19,894 natural images with 130,525 referring expressions for 96,654 visual regions. Note that, the annotation of ReferIt contains objects and stuff. The expressions are shorter and more succinct than the other datasets.

5) Evaluation Metrics: Following the setup of former works [6], [7], [22], we utilize Overall Intersection-over-Union (Overall IoU) and Prec@X to evaluate the proposed method. The Overall IoU metric represents the ratio between the total intersection regions and the total union regions of the predicted segmentation mask and the ground truth for all the test images. The Prec@X measures the percentage of the IoU score of the prediction mask exceeding the threshold X, where $\mathrm{X} \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

B. Implementation Details

Our framework is built on the public pytorch toolbox and is trained on an Nvidia RTX 3090 GPU. In the process of training, all the networks (i.e., for object contour prediction and noisy label learning) use the BRINet of encoder fusion. And they all use the SGD optimizer with an initial learning rate of 0.002 and divided by 10 after 50,000 iterations, and the maximum iterations are set to 90,000. The weight decay and batch size are 0.0005 and 16, respectively. Follow the previous works [6], [7], all input images are resized to 320×320. The maximum length of each referring expression is set to 20. Besides, when training G-ref, we use the UNC model as a pre-training model to avoid over-fitting. We use the network N_1 of stage 2 in the testing stage (as shown in Fig.2), and

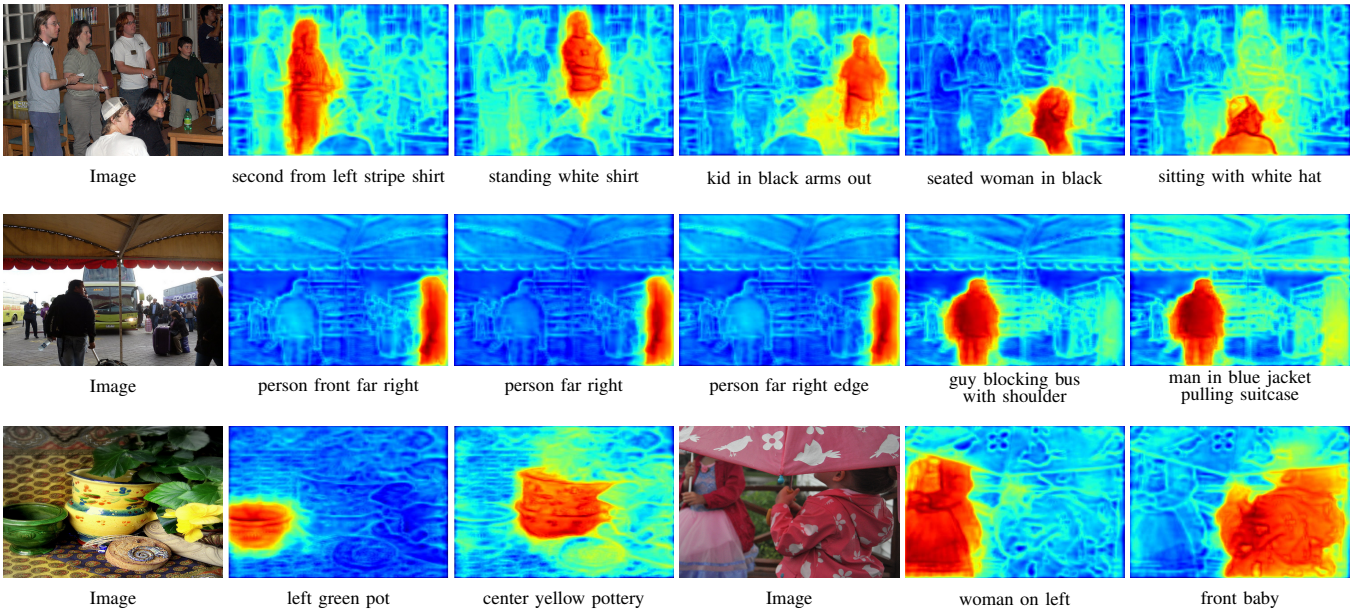

Fig. 5: Segmentation heatmaps of referring image segmentation by our method.

TABLE II: IoU for different length referring expressions on Google-Ref, UNC, UNC+ and ReferItGame.

	Length	1-5	6-7	8-10	11-20
G-Ref	R+LSTM [2]	32.29	28.27	27.33	26.61
	R+RMI [2]	35.34	31.76	30.66	30.56
	BRINet [6]	51.93	47.55	46.33	46.49
	Ours	51.67	46.90	44.54	41.59

	Length	1-2	3	4-5	6-20
UNC	R+LSTM [2]	43.66	40.60	33.98	24.91
	R+RMI [2]	44.51	41.86	35.05	25.95
	BRINet [6]	65.99	64.83	56.97	45.65
	Ours	63.12	60.65	53.79	42.75

	Length	1-2	3	4-5	6-20
UNC+	R+LSTM [2]	34.40	24.04	19.31	12.30
	R+RMI [2]	35.72	25.41	21.73	14.37
	BRINet [6]	59.12	46.89	40.57	31.32
	Ours	57.14	46.79	39.11	28.91

	Length	1	2	3-4	5-20
ReferIt	R+LSTM [2]	67.64	52.26	44.87	33.81
	R+RMI [2]	68.11	52.73	45.69	34.53
	BRINet [6]	75.28	62.62	56.14	44.40
	Ours	73.19	62.87	56.16	44.10

the proposed method runs at about 63 FPS with 320×320 input resolution. For the box-level ground truth, since UNC, UNC+ and G-Ref are collected by MS COCO, they already contain bounding box annotations. For ReferIt, we utilize the rectangular box that tightly encloses their foreground mask as the bounding box annotations.

C. Performance Comparison

This is the first work of box-level supervised referring image segmentation (RIS). To verify the effectiveness of the proposed algorithm, we compare it with some fully-supervised methods, such as LSTM-CNN [1], RMI [2], DMN [20], KWA [3], RRN [19], MAttNet [41], lang2seg [30], CGAN [8], BRINet [6], CMPC [22], and CMSA [27].

1) Quantitative Evaluation: The results of different methods are reported in Tab. I. Among them, 'Fully' indicates the fully-supervised method, and 'Weakly' denotes the weakly-supervised method. 'Cross Entropy' represents the prediction results of training on bounding box annotations with the cross entropy loss. It can be seen that the performance of the proposed method is close to that of these fully supervised methods. However, our method only requires box-level annotations, which achieves a trade-off between labeling efficiency and model performance. We analyze the relationship between the language length and the segmentation performance in Tab. II, which indicates that our method is on a par with the fully-supervised models. Meanwhile, we report the running speed of different methods in Tab. I. The proposed method is highly efficient with a speed of 63 FPS, which is averagely 3 times faster than other competitors.

2) Qualitative Evaluation: We give some representative segmentation results in Fig. 4. They contain a variety of scenarios, including objects touching the image boundary (b, c), complex semantics (e∼j), long and difficult sentences (h, j, l), referring expression without location information (b, f, l, n) and stuff region (n∼r). These results show that our method can effectively segment the foreground even in complex scenes. Besides, Fig. 5 visualize some heatmaps, which are obtained by averaging the feature of the last convolutional layer in the decoder. The results indicate that our weakly-supervised method can well perceive the contour of the foreground region.

TABLE III: Ablation study on the UNC val, testA and testB datasets. PGT: pseudo ground truth; Self-T: self-training; Co-T: co-training; Fully: full supervision training; P@X: Prec@X.

	$\mathcal{L}_{ce}$	$\mathcal{L}_{tight}$	PGT$_{adv}$	Self-T	Co-T(N_1)	N_1+N_1	PGT$_{tight}$	Fully	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	O-IoU
val	✓								43.13	25.85	9.38	2.39	0.14	42.01
		✓							65.95	52.56	34.62	14.90	1.38	52.54
		✓	✓						68.95	61.44	48.62	30.34	7.64	56.41
		✓	✓	✓					68.07	60.10	47.20	28.48	7.50	56.67
		✓	✓		✓				70.33	62.76	49.34	29.92	7.79	58.01
		✓	✓			✓			71.76	64.22	50.82	31.12	8.57	59.33
		✓					✓		68.81	58.51	43.73	24.59	6.30	55.10
								✓	71.90	66.59	59.15	44.91	17.86	61.39
testA	✓								41.59	17.54	3.66	0.39	0.00	42.10
		✓							70.02	55.26	33.76	11.37	0.35	54.20
		✓	✓						72.76	65.00	51.19	29.15	5.14	58.48
		✓	✓	✓					71.84	64.04	49.80	28.02	5.56	58.78
		✓	✓		✓				74.49	66.36	52.55	30.37	6.29	60.52
		✓	✓			✓			75.59	68.45	53.77	31.06	6.79	61.63
		✓					✓		72.51	61.46	44.88	22.73	3.25	56.73
								✓	75.13	70.57	62.88	48.35	16.07	63.98
testB	✓								48.13	32.91	16.76	4.02	0.51	42.79
		✓							63.08	52.93	37.94	20.12	2.51	51.47
		✓	✓						65.63	57.59	47.62	32.33	10.89	54.48
		✓	✓	✓					63.83	55.68	45.53	30.81	10.32	54.04
		✓	✓		✓				66.69	57.96	46.50	31.60	10.15	55.48
		✓	✓			✓			68.18	60.16	49.26	33.54	11.50	57.02
		✓					✓		64.53	56.00	45.16	29.68	9.64	53.53
								✓	66.99	61.69	54.88	43.08	19.41	58.14

TABLE IV: Relationship between hyperparameter x and segmentation results.

	x = 0.5	x = 0.75	x = 1.00	x = 1.25	x = 1.5
UNC val	57.71	57.76	58.01	57.88	57.61
UNC testA	60.19	60.26	60.52	60.73	60.12
UNC testB	55.07	55.60	55.48	55.15	55.60

D. Ablation Study

We compare a series of ablation studies on the UNC dataset to further verify our main contributions.

1) Comparison of $\mathcal{L}_{tight}$ and $\mathcal{L}_{ce}$: $\mathcal{L}_{tight}$ is a relaxed constraint based on tightness prior, which provides necessary supervision information for foreground segmentation. $\mathcal{L}_{ce}$ pixel by pixel computes and accumulates the prediction error. Hence, a large number of wrong labels in the bounding box mask make the network memorize these inaccurate patterns gradually, thereby worsening the performance of the network. We evaluate the two baselines in Tab. III, from which we can see that the results of $\mathcal{L}_{tight}$ are significantly better than those of $\mathcal{L}_{ce}$.

2) Effectiveness of Adversarial Boundary Loss: We use PGT$_{adv}$ to represent the pseudo ground truth derived from the object contour learned by adversarial boundary loss ($\mathcal{L}_{adv}$). Compared with the baseline $\mathcal{L}_{tight}$, the PGT$_{adv}$ brings 7.37%, 7.90%, 5.85% improvement in terms of Overall IoU on the UNC-val, UNC-testA and UNC-testB, respectively. $\mathcal{L}_{tight}$ mainly considers the global perception of visual region, it cannot capture the pixel-level details. However, the pseudo ground truth derived from the adversarial boundary loss supplements clear local information about the boundary. By combining both global and local contexts, our approach achieves significant performance improvements. Some visual results in Fig. 6 also verify this inference. Besides, to further demonstrate the effectiveness of the adversarial boundary loss $\mathcal{L}_{adv}$, we also compare the results of using the object contour (from $\mathcal{L}_{adv}$) and only using the prediction map of $\mathcal{L}_{tight}$ to filter the proposals. The prediction results of $\mathcal{L}_{adv}$ and $\mathcal{L}_{tight}$ are shown in Fig. 3. The method of using $\mathcal{L}_{tight}$ to select the proposals is similar to Eq. 4:

$$\arg\max_{y}\{\mathrm{IoU}(\mathcal{C}, \bigcup_{p_i\in\mathcal{P}} y_i \cdot p_i)\} \;\; s.t. \;\; y_i = 0 \;\; or \;\; 1, \mathcal{P} \subseteq \mathrm{box}, \tag{5}$$

where $\mathcal{C}$ represents a set of pixels at the foreground of the prediction map. $\mathcal{P}$ denotes the set of proposals. The main difference between Eq. 4 and Eq. 5 is that Eq. 4 calculates the intersection between the boundary of the selected proposals and the predicted contour, while Eq. 5 calculates the IoU between the selected proposals and the predicted mask. This is because calculating the intersection between proposals and prediction mask here will make the foreground of the pseudo ground truth tend to infinity. The experimental results are shown in the Tab. III. We can find that the training result of the pseudo labels derived from the object contour (PGT$_{adv}$) is better than that of the pseudo labels derived from $\mathcal{L}_{tight}$ (PGT$_{tight}$). Especially on prec@0.7, prec@0.8, and prec@0.9, PGT$_{adv}$ achieves significant performance improvement.

3) Comparison of Self-training (Self-T) and Co-training (Co-T): From Tab. III, it can be seen that Self-T only brings a slight performance improvement. In contrast, Co-T achieves

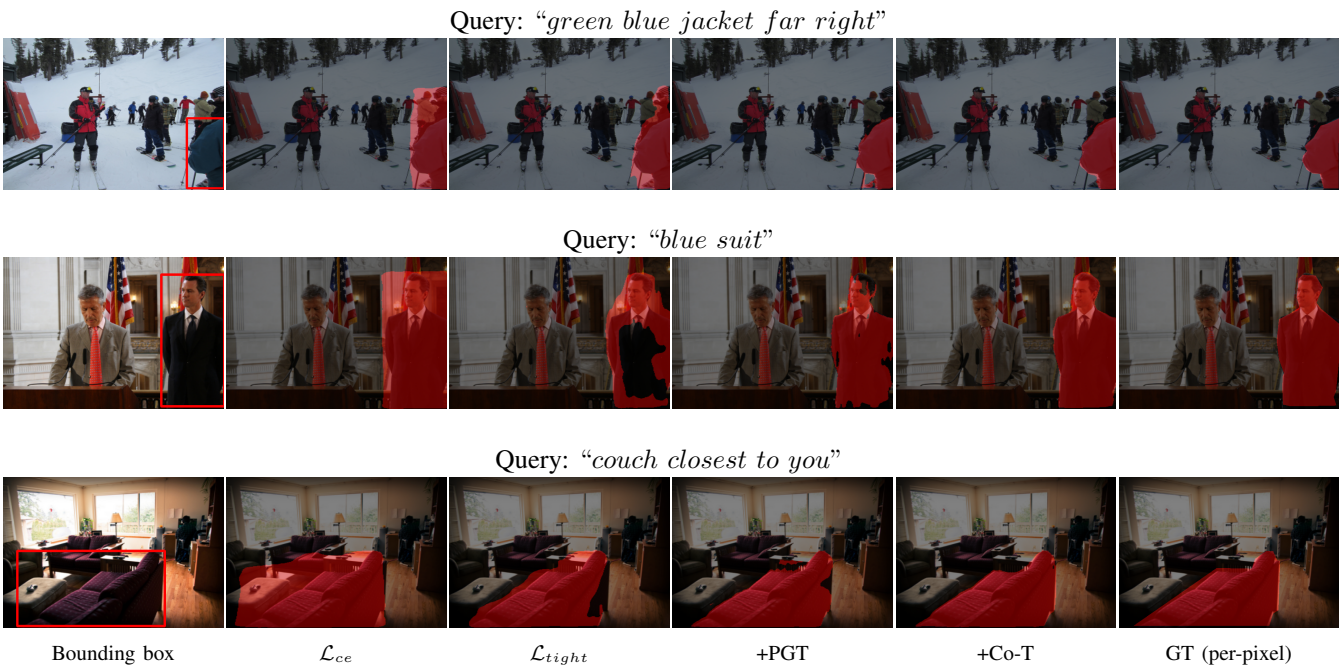

Fig. 6: Visual examples of the proposed modules.

Fig. 7: Visual examples of the failure cases.

the gains (Overall IoU) of 2.84%, 3.49%, 1.84% on the UNC-val, UNC-testA and UNC-testB, respectively. This indicates that the designed co-training strategy can effectively block the interference of noisy labels.

4) Results of N_1 + N_2: We compute the mean of the predicted maps output by network N_1 and network N_2, and the results (N_1+N_2) are shown in Tab. III. The results show that the average prediction maps of N_1 and N_2 show higher segmentation accuracy. However, the use of two networks will lead to the increase of computational cost. In the future, we can consider using different prediction strategies in different situations to achieve a dynamic balance between performance and computational cost.

5) Comparison with fully supervised training: In order to fairly reflect the performance gap between our proposed weakly-supervised model and the fully-supervised network. We train the same modified BRINet in a fully supervised manner. The results are shown in Tab. III. The weakly supervised model achieves about 3% performance drop.

6) Parameter sensitivity: The retention ratio $R(t)\%$ and maximum iterations T in Sec. III-C plays an important role. To investigate the effectiveness of these parameters, we compare the performance with a wider range of variable changes. Let n and m denote noise level and maximum iterations, respectively. The Overall IoU under different hyperparameters on the UNC val are: $\mathrm{n}\underline{0.05}\mathrm{m}\underline{50\mathrm{K}}$: 58.08%; $\mathrm{n}\underline{0.15}\mathrm{m}\underline{50\mathrm{K}}$: 57.32%; $\mathrm{n}\underline{0.10}\mathrm{m}\underline{40\mathrm{K}}$: 57.81%; $\mathrm{n}\underline{0.10}\mathrm{m}\underline{50\mathrm{K}}$: 58.01%; $\mathrm{n}\underline{0.10}\mathrm{m}\underline{60\mathrm{K}}$: 57.93%. These results show that our co-training strategy is not sensitive to the parameters in a great range.

In addition, the retention rate is a dynamic parameter. In this paper, a linear function is used to represent the variation law of this parameter. To verify the rationality of the linear function, we further define the calculation formula of the retention rate as: $R(t) = 1 - \eta \cdot (t/T)^{x}$, where T = 50000, η = 0.1. Obviously, when the hyperparameter x is equal to 1, the function degenerates to linear linear. When x is not equal to 1, it is nonlinear. The relationship between the value of x and the segmentation results (Overall IoU) is shown in Tab. IV. We can find that the linear function obtains the most stable

performance.

7) Effect of the boundary box without strict annotation on the results: We expand the width and height of the box annotation to 1.1 times the original one. On the UNC train, the Overall IoU of the pseudo label generated by the original box and the expanded box are 63.15% and 63.00%, respectively. Furthermore, We expand the box annotations to 1.2 times, the Overall IoU of the pseudo label on the UNC train is 60.55%. These results show that the change of bounding box within a certain range do not seriously affect the performance of the generated pseudo-annotations.

8) Analysis of the ratio of object to box area: We supplement the analysis about the relationship between the ratio of the object to the box area and the quality (Overall IoU) of the recovered pseudo label. The results on the UNC train are: [0, 0.5]: 51.98%; (0.5, 0.75]: 67.45%; (0.75, 1]: 73.92%.

E. Failure Cases

We also present some interesting failure cases in Fig. 7. In the first two rows, we can find that the query only contains some implicit appearance information, such as the number '1' on the sweatshirt and the word 'relax' on the bottle. Our method fails to capture this information well. This is because this paper mainly focuses on weakly-supervised segmentation. In the future, we can design more effective cross-modal fusion modules to alleviate this problem. In the third row, the referred object in the image is very blurry, which makes it difficult for our method to obtain an accurate object boundary. A reasonable boundary-aware mechanism may alleviate it. Besides, some unclear expressions may also lead to incorrect segmentation (the fourth row).

V. Conclusion

In this paper, we present a two-stage training method for referring image segmentation with box-level weak supervision. In the first stage, we design an adversarial boundary loss to capture the contour of the foreground region, and then utilize the contour as a priori to pick out suitable region proposals to construct the pseudo ground truth. In the second stage, we introduce a co-training strategy to promote two network to mutually purify pseudo labels for their peers. This mechanism can filter different types of noise and avoid the rapid accumulation of errors. Extensive experiments on four datasets demonstrate the effectiveness of the proposed method in accuracy and speed. In the future, we can consider applying the proposed weakly supervision algorithm to other image segmentation tasks, such as instance segmentation, saliency detection. Theoretically, the proposed method can be adopted to solve the corresponding problem only by constructing the bounding box of each instance or salient object.

References

[1] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Eur. Conf. Comput. Vis.* Springer, 2016, pp. 108–124.

[2] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille, "Recurrent multimodal interaction for referring image segmentation," in *Int. Conf. Comput. Vis.*, 2017, pp. 1271–1280.

[3] H. Shi, H. Li, F. Meng, and Q. Wu, "Key-word-aware network for referring expression image segmentation," in *Eur. Conf. Comput. Vis.*, 2018, pp. 38–54.

[4] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 10 502–10 511.

[5] X. Liu, L. Li, S. Wang, Z.-J. Zha, D. Meng, and Q. Huang, "Transferrable referring expression grounding with concept transfer and context inheritance," in *ACM Int. Conf. Multimedia*, 2020, pp. 3938–3946.

[6] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu, "Bi-directional relationship inferring network for referring image segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 4424–4433.

[7] T. Hui, S. Liu, S. Huang, G. Li, S. Yu, F. Zhang, and J. Han, "Linguistic structure guided context modeling for referring image segmentation," in *Eur. Conf. Comput. Vis.* Springer, 2020, pp. 59–75.

[8] G. Luo, Y. Zhou, R. Ji, X. Sun, J. Su, C.-W. Lin, and Q. Tian, "Cascade grouped attention network for referring expression segmentation," in *ACM Int. Conf. Multimedia*, 2020, pp. 1274–1282.

[9] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 136–145.

[10] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3791–3800.

[11] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3136–3145.

[12] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 12 275–12 284.

[13] H. Kervadec, J. Dolz, S. Wang, E. Granger, and I. B. Ayed, "Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 365–381.

[14] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 12 546–12 555.

[15] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Int. Conf. Comput. Vis.*, 2015, pp. 1635–1643.

[16] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 876–885.

[17] C.-C. Hsu, K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, and Y.-Y. Chuang, "Weakly supervised instance segmentation using the bounding box tightness prior," *Adv. Neural Inform. Process. Syst.*, vol. 32, pp. 6586–6597, 2019.

[18] V. Kulharia, S. Chandra, A. Agrawal, P. Torr, and A. Tyagi, "Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation," in *Eur. Conf. Comput. Vis.* Springer, 2020, pp. 290–308.

[19] R. Li, K. Li, Y.-C. Kuo, M. Shu, X. Qi, X. Shen, and J. Jia, "Referring image segmentation via recurrent refinement networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 5745–5753.

[20] E. Margffoy-Tuay, J. C. Pérez, E. Botero, and P. Arbeláez, "Dynamic multimodal instance segmentation guided by natural language queries," in *Eur. Conf. Comput. Vis.*, 2018, pp. 630–645.

[21] 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 Int. Conf. Multimedia*, 2019, pp. 539–547.

[22] S. Huang, T. Hui, S. Liu, G. Li, Y. Wei, J. Han, L. Liu, and B. Li, "Referring image segmentation via cross-modal progressive comprehension," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 10 488–10 497.

[23] Z. Liu, J. Deng, L. Li, S. Cai, Q. Xu, S. Wang, and Q. Huang, "Ir-gan: Image manipulation with linguistic instruction by increment reasoning," in *ACM Int. Conf. Multimedia*, 2020, pp. 322–330.

[24] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vision-language transformer and query generation for referring segmentation," in *Int. Conf. Comput. Vis.*, 2021, pp. 16 321–16 330.

[25] S. Yang, M. Xia, G. Li, H.-Y. Zhou, and Y. Yu, "Bottom-up shift and reasoning for referring image segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 11 266–11 275.

[26] Y. Jing, T. Kong, W. Wang, L. Wang, L. Li, and T. Tan, "Locate then segment: A strong pipeline for referring image segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 9858–9867.

[27] L. Ye, M. Rochan, Z. Liu, X. Zhang, and Y. Wang, "Referring segmentation in images and videos with cross-modal self-attention network," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.
[28] G. Feng, Z. Hu, L. Zhang, J. Sun, and H. Lu, "Bidirectional relationship inferring network for referring image localization and segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, 2021.
[29] G. Feng, Z. Hu, L. Zhang, and H. Lu, "Encoder fusion network with co-attention embedding for referring image segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 15 506–15 515.
[30] Y.-W. Chen, Y.-H. Tsai, T. Wang, Y.-Y. Lin, and M.-H. Yang, "Referring expression object segmentation with caption-aware consistency," in *Brit. Mach. Vis. Conf.*, 2019.
[31] D.-J. Chen, S. Jia, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "See-through-text grouping for referring image segmentation," in *Int. Conf. Comput. Vis.*, 2019, pp. 7454–7463.
[32] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Int. Conf. Comput. Vis.*, 2015, pp. 1742–1750.
[33] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Adv. Neural Inform. Process. Syst.*, 2011, pp. 109–117.
[34] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, pp. 542–542, 2009.
[35] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.
[36] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2117–2125.
[37] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 328–335.
[38] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *Eur. Conf. Comput. Vis.* Springer, 2014, pp. 725–739.
[39] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box prior," in *Int. Conf. Comput. Vis.* IEEE, 2009, pp. 277–284.
[40] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2017, pp. 233–242.
[41] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 1307–1315.
[42] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Eur. Conf. Comput. Vis.* Springer, 2016, pp. 69–85.
[43] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 11–20.
[44] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," 2014, pp. 787–798.
[45] H. J. Escalante, C. A. Hernández, J. A. Gonzalez, A. López-López, M. Montes, E. F. Morales, L. E. Sucar, L. Villaseñor, and M. Grubinger, "The segmented and annotated iapr tc-12 benchmark," *Comput. Vis. Image Und.*, vol. 114, no. 4, pp. 419–428, 2010.

Guang Feng received his B.E. degree in electronic information engineering from Qingdao University, Qingdao, China and M.E. degree in signal and information processing from the University of Jinan, Jinan, China in 2015 and 2018, respectively. He is currently a Ph.D. candidate in the School of Information and Communication Engineering, Dalian University of Technology, Dalian, China. His research interests include saliency detection and referring expression comprehension.

Lihe Zhang received the M.S. degree and the Ph.D. degree in Signal and Information Processing from Harbin Engineering University (HEU), Harbin, China, in 2001 and from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2004, respectively. He is currently a Full Professor with the School of Information and Communication Engineering, Dalian University of Technology (DUT). His current research interests include computer vision and pattern recognition.

Zhiwei Hu received the M.Eng. degree in Electronics and Communication Engineering from Dalian University of Technology (DUT) in 2021. His research interests include computer vision and Natural Language Processing.

Huchuan Lu received the Ph.D. degree in System Engineering and the M.S. degree in Signal and Information Processing from Dalian University of Technology (DUT), Dalian, China, in 2008 and 1998, respectively. He joined the faculty in 1998 and currently is a Full Professor of the School of Information and Communication Engineering, DUT. His research interests include computer vision and pattern recognition with focus on visual tracking, saliency detection, and segmentation. He is an Associate Editor of the IEEE Transactions on Cybernetics and IEEE Transactions on Circuits, Systems for Video Technology.