Referring Image Segmentation Using Text Supervision

Fang Liu[1,2,*], Yuhao Liu[2,*], Yuqiu Kong[1,†], Ke Xu[2,†], Lihe Zhang[1] ,
Baocai Yin[1], Gerhard Hancke[2] , Rynson Lau[2,†]
[1] Dalian University of Technology, [2] City University of Hong Kong
{fawnliu2333, yuhaoLiu7456, kkangwing}@gmail.com, {yqkong, zhanglihe, ybc}@dlut.edu.cn,
{gp.hancke, rynson.lau}@cityu.edu.hk

Abstract

Existing Referring Image Segmentation (RIS) methods typically require expensive pixel-level or box-level annotations for supervision. In this paper, we observe that the referring texts used in RIS already provide sufficient information to localize the target object. Hence, we propose a novel weakly-supervised RIS framework to formulate the target localization problem as a classification process to differentiate between positive and negative text expressions. While the referring text expressions for an image are used as positive expressions, the referring text expressions from other images can be used as negative expressions for this image. Our framework has three main novelties. First, we propose a bilateral prompt method to facilitate the classification process, by harmonizing the domain discrepancy between visual and linguistic features. Second, we propose a calibration method to reduce noisy background information and improve the correctness of the response maps for target object localization. Third, we propose a positive response map selection strategy to generate high-quality pseudo-labels from the enhanced response maps, for training a segmentation network for RIS inference. For evaluation, we propose a new metric to measure localization accuracy. Experiments on four benchmarks show that our framework achieves promising performances to existing fully-supervised RIS methods while outperforming state-of-the-art weakly-supervised methods adapted from related areas. Code is available at https://github.com/fawnliu/TRIS.

1. Introduction

Referring Image Segmentation (RIS) aims to segment a target object from an input image according to the input linguistic query. It has many applications such as text-based image editing [8, 10], human-computer interaction [58, 64], E-commercial search engine [76, 71].

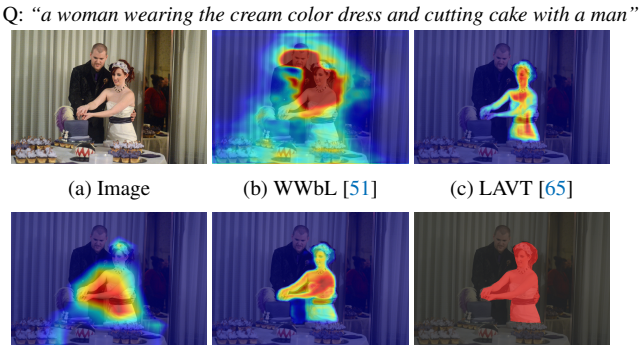


Figure 1. Given an input image and a query expression (a), our method leverages a text-to-image response optimization process to locate the target object in the first step (d), and then produce the final map using a segmentation network in the second step (e). Our method outperforms the weakly-supervised method [51] (b) and achieves competitive result compared to the state-of-the-art fully-supervised RIS method [65] (c). Note that our method is trained using only text expressions, which are already available.

Hu *et al*. [22] propose the first RIS method, which separately extracts visual and linguistic features and then concatenates them to predict the segmentation map. Subsequently, many methods are proposed to fuse multi-modal features for RIS in different ways, *e.g*., with unidirectional fusion [66, 37] and bidirectional fusion [23, 15] using CNNs, and long-range cross-modal modeling [65, 28] using transformers. However, almost all existing RIS methods rely on the fully-supervised learning scheme to produce accurate results, which requires time-consuming and labour-intensive pixel-level annotations. (For example, skilled annotators may need an average of 79s [44] to label a polygon-based instance mask for an image in MS COCO [33].) Recently, Feng *et al*. [16] propose to use bounding box annotations for weakly-supervised RIS, but it is still costly for labeling large-scale RIS datasets (on average 7s to annotate a single bounding box). Hence, there is still a need to design RIS models with cheaper supervision signals.

Besides using bounding boxes, there are also other kinds

*Joint first authors. †Joint corresponding authors.

of weak supervision signals studied in other computer vision tasks, including scribbles [70, 69], points [18] and class labels [57, 47]. However, scribbles and points still require and are sensitive to human involvement. Although class labels work well for other visual tasks, *e.g.*, salient object detection [57], they are unsuitable for RIS due to their lack of instance information.

We observe that the input referring texts used in the RIS task have already provided distinctive descriptions, which describe one or several properties, of the target object that can be used to locate the object. Inspired by this, we propose a novel weakly-supervised RIS framework, which learns to locate the target objects by learning how to classify positive and negative texts, where the positive texts are the referring texts that are used to describe the corresponding objects of the input images while the negative texts are the referring texts that describe objects from other images. Our framework has three main technical novelties. First, to facilitate classification process, we propose a bilateral prompt method to harmonize the modal discrepancy between visual and linguistic features. Second, we propose a calibration method that consists of a foreground enhancement process and a background suppression process to enhance the derived response maps for target object localization. Third, we propose a novel positive response map selection strategy to derive high-quality pseudo-labels from the enhanced response maps, which are used to train a segmentation network for RIS inference. In addition, we propose a new metric to evaluate localization accuracy, which can reduce in-box errors of the existing pointing-game accuracy metric.

As shown in Fig. 1, our method can locate the target object and produce the segmentation map accurately. Even though it is trained using only text expressions, it produces a comparable result to the fully-supervised method [65]. In summary, this paper has the following main contributions:

- We propose a novel weakly-supervised RIS framework that is supervised only by the readily available referring texts and does not require any extra annotations.
- Our framework has three main technical novelties. First, we propose a bilateral prompt method to harmonize the visual and linguistic modal discrepancy. Second, we propose a calibration method to improve the correctness of the response maps for localization. Third, we propose a response map selection strategy to generate high-quality pseudo labels for the segmentation of the target objects.
- We propose a new metric for the evaluation of localization accuracy. Extensive experiments on four benchmarks show that our framework can produce promising results compared to previous fully-supervised RIS methods and outperforms existing weakly-supervised baselines adapted from related tasks.

2. Related Works

Referring Image Segmentation detects and segments target objects according to the input text expressions. Hu *et al*. [22] propose the first RIS method with a CNN model. Many follow-up methods are then proposed, mainly focusing on addressing the linguistic and visual feature fusion problem. Some methods adopt recurrent refinement [7, 30, 35] or dynamic filters [42, 9, 26] to fuse visual and linguistic features. To capture long-range dependencies between two modalities, several methods explore attention mechanisms [66, 23, 15, 37, 52, 40, 67, 60, 24, 63] or transformer-based architectures [14, 29, 28, 65, 59, 75, 36].

However, all these methods are fully-supervised. Training them requires pixel-level annotations, which are time-consuming and labor-intensive to obtain. Recently, Feng *et al*. [16] propose the first weakly-supervised RIS method, which is based on bounding box annotations. Although bounding boxes are cheaper to annotate, they still require a significant amount of effort on a large-scale dataset. In this paper, we propose a weakly-supervised method that uses only the available text expressions for training.

Weakly-Supervised Learning tries to reduce labelling efforts with different kinds of weak labels, *e.g.*, bounding boxes [11, 31], scribbles [70, 69], points [18], class labels [72, 46, 57, 3, 61, 54, 53], and text expressions [51, 12, 19, 38, 56, 25, 2, 5], for various computer vision tasks.

Among these proposed weak labels, image-level class labels are very popular, as they provide high-level semantic information but are relatively cheap to label. The above methods propose different class activation map generation processes to locate the objects, *e.g.*, global average pooling [72], global max pooling [46], global smoothing pooling [57], and normalized global weighted pooling [3]. Xie *et al*. [61] propose to further use background class labels to improve the quality of maps. While also using text input, class labeling does not work in our task due to the lack of instance-level information.

Language expressions are commonly used as weak supervision signals in the weakly-supervised visual grounding (WSG) task (*i.e.*, detecting the target objects with bounding boxes according to the input expressions). However, most WSG methods [12, 19, 38, 56, 25, 55] resort to additional pre-trained object detectors to produce a set of proposals and then select the one with the highest confidence score by matching visual features with phrase features. To avoid the dependence on pre-trained detectors, Arbelle *et al*. [5] propose a self-supervised method to train the network by randomly blending the images according to the expressions. Shaharabany *et al*. [51] compute the relevancy heatmaps and treat them as GTs to supervise the generated response maps. Akbari *et al*. [2] propose a posterior probability-based multi-modal loss to guide the network to predict cor-

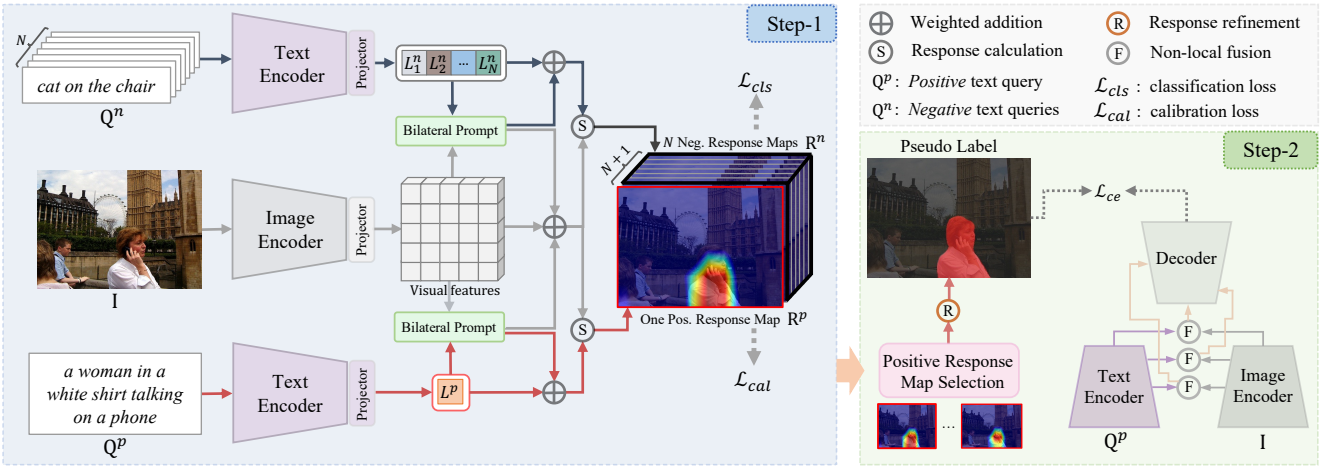


Figure 2. Our weakly-supervised RIS framework has two steps. **Step-1**: It learns to classify input text expressions, and uses the positive text expression to help localize the target object to produce response maps. **Step-2**: It feeds the pseudo-labels derived from the response maps to train a segmentation network for RIS inference. We propose a bilateral prompt method to harmonize the multi-modal discrepancy and a calibration method to enhance the response maps in Step-1. We propose a positive response map selection strategy in Step-2 to help select the best response maps as pseudo-labels.

relation scores for the related image-sentence pairs.

While these methods are close in spirit to our work, they only detect the target objects with bounding boxes. Our experiments demonstrate that it is non-trivial to obtain pixel-wise segmentation maps from their bounding-box results. In contrast, our method learns to detect the target objects in a pixel-wise manner via a text-to-image optimization process, which allows accurate localization of the target.

3. Our Framework

The main objective of the weakly-supervised RIS task is to establish a pixel-wise association between the visual content and the input referring expressions without using pixel-level annotations. We note that the input referring expressions inherently possess discriminative information pertaining to the localization of the target objects or regions. In light of this observation, our framework is designed to learn to classify positive and negative expressions of each input image, through which it also learns to localize the target object in the image as described by the positive expressions. The positive expressions are the referring text expressions that are used to describe the target object of the input image, while the negative expressions are the referring text expressions taken from other images. Our classification process involves the modelling of text-to-image responses, which can learn to associate the visual contents in the input image with the positive expressions.

Fig. 2 shows our weakly-supervised RIS framework, which has two steps. The first step models the text-to-image response of the classification process to locate the target object as specified by the text expressions. We propose a bilateral prompt method in this step to harmonize the discrepancy between the visual and linguistic features, and a calibration method to enhance the completeness and correctness of the positive response maps. The second step leverages the response maps from the first step to produce pseudo labels. These pseudo labels are then used to train a segmentation network for RIS inference. A positive response map selection strategy is proposed to select the best response maps for pseudo-label generation.

3.1. Text-to-Image Response Modeling

We first explain how we model text-to-image responses, which are used to localize the target objects as specified by the expressions in the query classification process.

Given an input image $\mathbf{I} \in \mathbb{R}^{H_I \times W_I \times 3}$ and a text expression query $\mathbf{Q} \in \mathbb{R}^{T}$ with T words, we first extract the visual features $\mathbf{V}_e \in \mathbb{R}^{H \times W \times C_v}$ and text features $\mathbf{L}_e \in \mathbb{R}^{1 \times C_l}$ by an image encoder E_v and a text encoder E_l [49], where $H = H_I/s$ and $W = W_I/s$. C_v and C_l denote the numbers of channels of visual and text features, respectively. s is the ratio of down-sampling. We then use a projector layer to transform $\mathbf{V}$ and $\mathbf{L}$ to a unified hidden dimension C_d, *i.e.*, $\mathbf{V} \in \mathbb{R}^{H \times W \times C_d}$ and $\mathbf{L} \in \mathbb{R}^{1 \times C_d}$. The L_2 channel-wise normalization is used to regularize the output of the projection layer. To harmonize the domain discrepancy, we propose a bilateral prompt method (Sec. 3.2) to update the visual and text features as:

$$\hat{\mathbf{V}} = \mathbf{V} + \alpha \mathbf{V}^{'}; \quad \hat{\mathbf{L}} = \mathbf{L} + \beta \mathbf{L}^{'}, \\ \text{with} \quad \mathbf{V}^{'}, \mathbf{L}^{'} = \mathit{\Gamma}_{\text{Prompt}}\left(\mathbf{V}, \mathbf{L}\right), \tag{1}$$

where $\hat{\mathbf{V}}$ and $\hat{\mathbf{L}}$ denote the updated visual and text features, and $\mathbf{V}^{'}$ and $\mathbf{L}^{'}$ are the residual enrichment prompted from

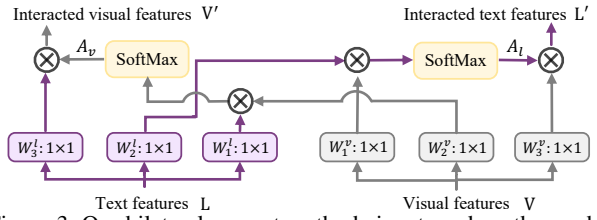


Figure 3. Our bilateral prompt method aims to reduce the modality differences between visual and linguistic features. The text features $\mathbf{L}$ are enhanced by the visual features $\mathbf{V}$ via the attention map A_l, which updates $\mathbf{L}$ to $\mathbf{L}^{'}$ for classification. The visual features $\mathbf{V}$ are enhanced by the text features $\mathbf{L}$ via the attention map A_v to update $\mathbf{V}$ to $\mathbf{V}^{'}$ for localization.

$\mathbf{L}$ and $\mathbf{V}$. α and β are weights to control the proportion of bilateral prompting. $\mathcal{I}_{\text{Prompt}}$ is our bilateral prompt method.

We have now aligned $\hat{\mathbf{V}}$ and $\hat{\mathbf{L}}$ explicitly. After reshaping, the response between i-th flattened pixel of $\hat{\mathbf{V}}$ and j-th query of $\hat{\mathbf{L}}$ can then be calculated as:

$$\mathbf{R}_{i,j} = \sum_{\vartheta=0}^{C_d} \hat{\mathbf{V}}_{i,\vartheta} \cdot \hat{\mathbf{L}}_{j,\vartheta}, \tag{2}$$

where ϑ and $\cdot$ represent channel index and element-wise multiplication, respectively. We employ a learnable temperature parameter τ [49] to scale and constrain the range of $\mathbf{R}_{i,j}$. In this way, we establish the relationship of each pixel to the whole referring expression $\mathbf{Q}$, and different pixels will echo responses of different degrees. A higher response indicates that the pixel is more likely to belong to the target object described by the positive expression.

3.2. Bilateral Prompt

To facilitate learning knowledge from the pretrained model [49], we propose a bilateral prompt method to harmonize the domain discrepancy between visual and text features. Unlike previous prompt methods that either refine visual and text features separately [17], or use the visual features to refine the text features [50, 74, 73], our bilateral prompt method enhances the features of one modality with those of the other modality. Enriching the text features with the visual features facilitates the classification process, while enriching the visual features with the text features helps localize the target objects.

Fig. 3 shows our bilateral prompt method. Given input features $\mathbf{V} \in \mathbb{R}^{H\times W\times C_d}$ and $\mathbf{L} \in \mathbb{R}^{1\times C_d}$, we first compute two attention maps as:

$$\mathbf{A}_l = \text{SoftMax}\left((\mathbf{V}W_1^v) \otimes (\mathbf{L}W_2^l)^\top / \sqrt{C_d}\right), \tag{3}$$

$$\mathbf{A}_v = \text{SoftMax}\left((\mathbf{L}W_1^l) \otimes (\mathbf{V}W_2^v)^\top / \sqrt{C_d}\right), \tag{4}$$

where $\mathbf{A}_l \in \mathbb{R}^{1\times HW}$ and $\mathbf{A}_v \in \mathbb{R}^{HW\times 1}$ denote the affinity propagation of visual-to-text features and text-to-visual features, respectively. $W_*^v \in \mathbb{R}^{C_d\times C_d}$ and $W_*^l \in \mathbb{R}^{C_d\times C_d}$ are the learnable parameters for $\mathbf{V}$ and $\mathbf{L}$. $\otimes$ denotes matrix multiplication. The bilateral prompt is then formulated as:

$$\mathbf{L}^{'} = \mathbf{A}_l^\top \otimes (\mathbf{V}W_3^v), \quad \mathbf{V}^{'} = Re(\mathbf{A}_v^\top \otimes (\mathbf{L}W_3^l)), \tag{5}$$

where Re is a shape transform function, $\mathbf{L}^{'} \in \mathbb{R}^{1\times C_d}$ and $\mathbf{V}^{'} \in \mathbb{R}^{H\times W\times C_d}$ denote visual-enhanced linguistic features and language-guided visual features, respectively. In this way, the learned bilateral residuals (*i.e.*, $\mathbf{L}^{'}$ and $\mathbf{V}^{'}$) can be used to update both visual and linguistic features to effectively harmonize their discrepancy, according to Eq. (1).

3.3. Localization via Text-to-Image Optimization

Classification. We formulate a classification process for the network to learn to select the positive expressions from a set of positive and negative expressions, with which we optimize the text-to-image response maps to localize the target object. We implement the classification process in a contrastive learning manner [20]. In our problem, we use the input image $\mathbf{I}$, a positive text expression $\mathbf{Q}^p \in \mathbb{R}^T$ as a positive sample (or the anchor), and randomly sample N negative text expressions $\mathbf{Q}^n \in \mathbb{R}^{N\times T}$ (*i.e.*, text expressions of other images) from the whole dataset as negative samples. We compute the response maps $\mathbf{R}^a \in \mathbb{R}^{HW\times(1+N)}$ (which include $\mathbf{R}^p \in \mathbb{R}^{HW\times 1}$ for the positive expression $\mathbf{Q}^p$, and $\mathbf{R}^n \in \mathbb{R}^{HW\times N}$ for the negative expressions $\mathbf{Q}^n$) of image $\mathbf{I}$, as shown in Fig. 2. We then compute the image-level score y_j for each text query $\mathbf{Q}_j$ as:

$$y_j = \max_i \mathbf{R}_{i,j}^a + \frac{1}{HW}\sum_{i=1}^{HW} \mathbf{R}_{i,j}^a + \psi(\mathbf{R}_{\cdot,j}^a), \tag{6}$$

where $\psi(\mathbf{R}_{\cdot,j}^a)$ is a regularization term proposed by [32]. We use it to re-balance our negative and positive text queries. $\mathbf{y}_j$ ranges from 0 to 1, and the larger it is, the better the current query $\mathbf{Q}_j$ matches with the input image.

The classification process is then supervised by:

$$\mathcal{L}_{cls}(y,z) = -\frac{1}{N+1}\sum_{j=1}^{1+N} z_j \log\left(\frac{1}{1+e^{-y_j}}\right) + (1-z_j)\log\left(\frac{e^{-y_j}}{1+e^{-y_j}}\right), \tag{7}$$

where $\mathbf{z} \in \mathbb{R}^{1\times(1+N)}$ is an easy to obtain supervision signal, with 1 indicating a positive query and 0 a negative one.

Calibration. Given the coarse response maps derived from the classification process, we propose a calibration method to enhance the correctness of the positive response map $\mathbf{R}^p$ by contrasting the target object with other objects of the same image (*i.e.*, considering them as background noise).

Specifically, we multiply the input image $\mathbf{I}$ with $\mathbf{R}^p$ to obtain the target object and use it as the anchor. We use the positive query $\mathbf{Q}^p$ as the positive sample, and additionally sample K queries that describe other objects of the same image as negative samples (*i.e.*, $\mathbf{F}_k^n, k \in 1, 2, \ldots, K$). The calibration process can then be formulated as:

$$\mathcal{L}_{cal} = -(\log S(\mathbf{I}, \mathbf{R}^p, \mathbf{Q}^p) + \sum_{k=1}^{K} \log\left(1 - S(\mathbf{I}, \mathbf{R}^p, \mathbf{F}_k^n)\right)), \tag{8}$$

where $S(\cdot, \cdot, \cdot)$ is a similarity function to measure the matching scores of the target object and a query, as:

$$S(\mathbf{I}, \mathbf{R}^p, \mathbf{Q}) = \varphi(E_v(\mathbf{I} \odot up(\mathbf{R}^p)), E_l(\mathbf{Q})), \tag{9}$$

where E_v and E_l are CLIP [49] visual and text encoders. $\odot$ and $up()$ are Hadamard product and an up-sampling function, respectively. $\varphi(\cdot, \cdot)$ computes cosine similarity score.

Eq. (8) calibrates the positive response map in two ways. The first term helps learn more compact foreground regions by encouraging higher correlation of the instance-wise target object to the positive query. The second term suppresses noisy information of other background objects by decreasing the matching scores between foreground object regions and those negative queries.

3.4. Pseudo Labels Generation and Segmentation

Positive Response Map Selection (PRMS). There are usually several text expressions available that refer to the same target object in one image, but describing different properties of the object. Although these expressions are discriminative enough for neural networks to localize the target object, the corresponding response maps may not be exactly the same. Hence, we select the response map of the best quality by computing the cumulative similarity scores.

Specifically, we first compute the text-to-image response $\mathbf{R}_t^p$ by Eq. (2) for each $\mathbf{Q}_m^p$ of M positive queries. We then compute the similarity between its masked target object with those of all positive queries, and sum up all the object-to-text similarity scores to reflect the accuracy of the current response map as:

$$CS_t = \sum_{m=1}^{M} S(\mathbf{I}, \mathbf{R}_t^p, \mathbf{Q}_m^p),\ t \in 1, 2, \ldots, M. \tag{10}$$

We select the response map $\mathbf{R}_t^p$ with the maximum cumulative score CS_t as the response map of the target object.

Segmentation. We use [1] to refine our response maps and conduct thresholding on the responses to obtain the pseudo labels for training the segmentation network for RIS inference. The segmentation network has an image encoder, a text encoder, a multi-modality fusion module, and a decoder. We use the same encoders as in Eq. (9), and the decoder is symmetric to the encoder. Since the segmentation network only needs to predict the segmentation map, we use the non-local module [65] to enrich the visual features with the text features in the last three encoder layers, and send them to the decoder. We train the segmentation network with standard cross-entropy loss ($\mathcal{L}_{ce}$).

4. Experiments

4.1. Experiment Setups

Datasets. We conduct experiments on four benchmarks: ReferIt [27], RefCOCO [68], RefCOCO+ [68] and RefCOCOg [41]. ReferIt has $19,894$ images with $13,0525$ annotated referring expressions, which are usually shorter and more succinct. The other three datasets are all based on MSCOCO [33], and each of them contains (images, annotated expressions) as: (19,994, 142,209), (19,992, 141,564) and (26,711, 104,560). While the expressions of RefCOCO focus more on the position property of objects, those of RefCOCO+ focus more on appearance. Compared to these two datasets, RefCOCOg is more challenging as their expressions are usually longer and more complex. This dataset has two partitions, *i.e.*, the Google [41] and UMD [43] partitions. Both are used in our experiments.

Implementation Details. We use ResNet-50 [21] as our default image encoder, and CLIP [49] to initialize the image and text encoders. Both α and β used in our bilateral prompt are set to 0.1. We set the numbers of negative samples in the classification loss and calibration method to $N = 47$ and $K = 6$, respectively. For images that do not have enough negative queries (of only one object with expressions), we randomly sample queries from the rest of dataset to supplement K to 6. The number of hidden dimensions is $C_d = 1024$, and the down-sampling ratio is $s = 32$. The loss functions used in Step-1 is $\mathcal{L} = \lambda * \mathcal{L}_{cls} + \mathcal{L}_{cal}$. λ is used to ensure the numerical and gradient equivalence during training, and we empirically set λ to 5.

We implement our framework on PyTorch [45] and train it for 15 epochs with a batch size of 48 on an NVIDIA RTX3090 GPU with 24GB of memory. During training, we resize the input images to 320×320 and set the maximum length of each referring expression to 20. The network is optimized using the AdamW optimizer [39] with a weight decay of $1e^{-2}$ and an initial learning rate of $5e^{-5}$ with polynomial learning rate decay. The training settings are the same for both steps. During inference, we feed the input image and query text to Step-2 of the segmentation network to produce the segmentation map.

Evaluation Metrics. Following [14, 66], we adopt the mask intersection-over-union (IoU) and Prec@0.5 (P@0.5)

Metric	Method	*Sup.*	Backbone	ReferIt	RefCOCO			RefCOCO+			RefCOCOg		
				test	val	testA	testB	val	testA	testB	val (G)	val (U)	test (U)
IoU	LAVT [65]	$\mathcal{F}$	Swin-Base	-	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	60.50
	CTS [16]	$\mathcal{B}$+$\mathcal{T}$	ResNet-101	62.44	58.01	60.52	55.48	47.12	50.86	40.26	46.03	-	-
	AMR† [48]	$\mathcal{T}$	ResNet-50	18.98	14.12	11.69	17.47	14.13	11.47	18.13	15.83	15.46	15.59
	GroupViT† [62]	$\mathcal{T}$	GroupViT	21.73	18.03	18.13	19.33	18.15	17.65	19.53	19.97	19.80	20.09
	CLIP-ES† [34]	$\mathcal{T}$	ViT-Base	18.67	13.79	15.23	12.87	14.57	16.01	13.53	14.16	13.93	14.09
	GbS† [5]	$\mathcal{T}$	VGG16	14.21	14.59	14.60	14.97	14.49	14.49	15.77	14.21	13.75	14.20
	WWbL† [51]	$\mathcal{T}$	VGG16	27.68	18.26	17.37	19.90	19.85	18.70	21.64	21.84	21.75	21.82
	Ours (Step-1)	$\mathcal{T}$	ResNet-50	33.33	25.11	26.47	23.80	22.31	21.61	22.86	26.93	26.62	27.27
	Ours (Step-2)	$\mathcal{T}$	ResNet-50	**44.57**	**31.17**	**32.43**	**29.56**	**30.90**	**30.42**	**30.80**	**36.00**	**36.19**	**36.23**
PointM	AMR† [48]	$\mathcal{T}$	ResNet-50	7.12	15.55	5.52	28.91	16.33	5.90	30.27	25.51	24.96	26.14
	GroupViT† [62]	$\mathcal{T}$	GroupViT	27.44	25.01	26.30	24.42	25.92	26.06	26.12	30.02	30.90	30.98
	CLIP-ES† [34]	$\mathcal{T}$	ViT-Base	52.90	41.33	50.61	30.34	46.55	56.20	33.32	49.08	46.22	45.75
	GbS† [5]	$\mathcal{T}$	VGG16	30.30	21.58	19.52	25.95	20.95	18.34	25.96	24.64	24.60	25.38
	WWbL† [51]	$\mathcal{T}$	VGG16	42.84	31.28	31.15	30.79	34.47	33.30	36.10	29.32	32.13	31.37
	Ours (Step-1)	$\mathcal{T}$	ResNet-50	61.70	51.92	60.88	43.02	40.85	40.94	41.13	52.48	51.98	53.29
	Ours (Step-2)	$\mathcal{T}$	ResNet-50	**67.00**	**54.72**	**65.64**	**43.40**	**53.72**	**61.30**	**45.24**	**58.01**	**58.84**	**58.70**

Table 1. Quantitative comparison of different methods using the *IoU* and *PointM* metrics on four RIS benchmarks. *Sup.* denotes the supervision type ($\mathcal{F}$: full supervision, $\mathcal{B}$: box-level labels, $\mathcal{T}$: text description labels). (G) and (U) denote the Google and UMD dataset partitions of RefCOCOg. † indicates the methods adapted from other tasks. "-" denotes unavailable values.

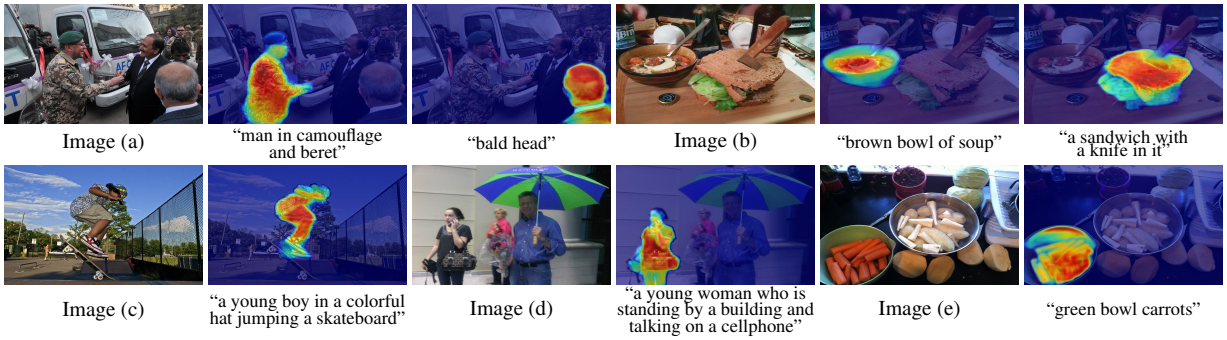


Figure 4. Qualitative results of the proposed RIS method.

Metric	MG [2]	GbS†[5]	WWbL†[51]	Ours$^*_{s_1}$	Ours$_{s_1}$	Ours$_{s_2}$
Acc$_{box}$	15.15	12.67	24.02	38.14	39.90	**50.79**
PointIt	47.52	48.12	57.04	70.10	72.56	**74.94**

Table 2. Quantitative comparison of our method and WSGs on the ReferIt *test* set. * denotes using ImageNet [13] weights to initialize the image encoder. s_1 and s_2 denote Step-1 and Step-2, respectively. Refer to the Supplemental for more comparisons.

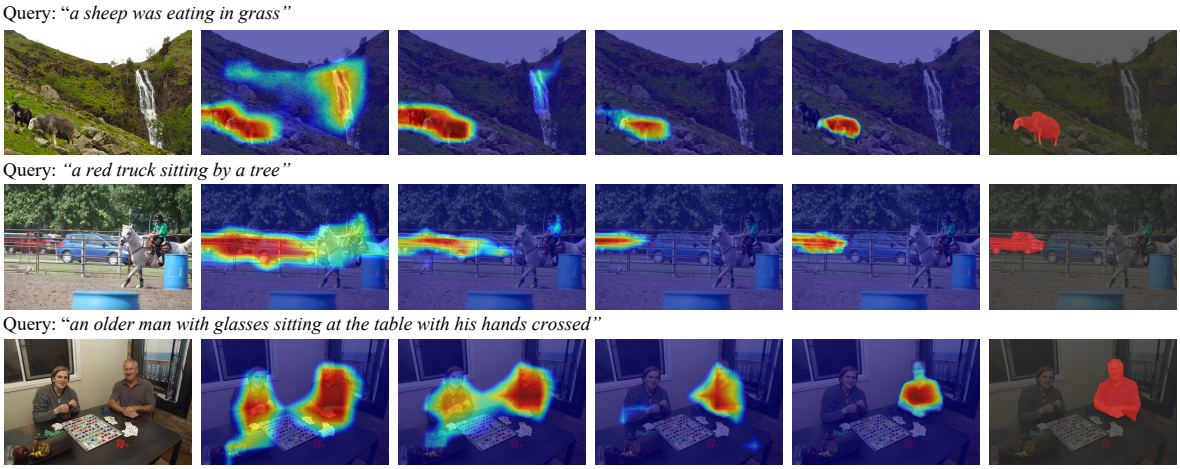
metrics to evaluate the segmentation accuracy. Following [2, 51], we also use the pointing-game (PointIt) and box accuracy (Acc$_{box}$) to measure the localization performance.

However, we note that PointIt tends to yield inaccurate localization scores in our task, as PointIt may count the hit that falls into the box but out of ground truth mask as corrected. Hence, we propose a new metric (PointM), which is formulated as: $PointM = \frac{\#Hits}{\#Hits+\#Misses}$, where *#Hits* and *#Misses* are the numbers of Hits and Misses. If the maximum point of the response map falls within the ground truth mask regions, it is counted as a Hit. Otherwise, it is considered as a Miss.

4.2. Comparison to State-of-the-arts

We validate the effectiveness of our framework by comparing it with fully supervised RIS method (LAVT [65]), weakly-supervised RIS method (CTS [16]), and other related weakly-supervised methods (MG [2], AMR [48], GroupViT [62], CLIP-ES [34], GbS [5] and WWbL [51]).

Table 1 reports the quantitative comparisons of the segmentation and localization accuracy of different methods on four benchmarks. Our approach demonstrates a significant performance improvement compared to existing methods [48, 62, 34] that typically generate pseudo-labels using only class labels. This is because they lack the capability to discriminate and reason the relations between the instance-level objects in the image, making their adaption to RIS more challenging. Our framework also exhibits supe-



(a) Image (b) $\mathcal{L}_{cls}$ (c) $\mathcal{L}_{cls}$+BP (d) $\mathcal{L}_{cls}$+BP+ $\mathcal{L}_{cal}$ (e) $\mathcal{L}_{cls}$+BP+$\mathcal{L}_{cal}$+Step2 (f) GT

Figure 5. Visualization of the ablation studies to demonstrate the effectiveness of each component. BP is the bilateral prompt.

rior performance in comparison to text-based weakly supervised methods [5, 51]. For instance, it outperforms WWbL with IoU gains of 16.56 and 14.16 on the ReferIt test and RefCOCOg val(G) sets. This is due to three reasons. First, the formulations of target object localization are different. WWbL uses the output relevance maps of [6] (which mainly model semantic information of class labels) as GTs to learn response maps. It may be incorrect if [6] fails to localize the target object. In contrast, we locate target objects via the response map generated from text-to-image optimization process, which makes our model more sensitive to information from both modalities. Second, our bilateral prompt generates language-guided visual features and image-enhanced text features for better feature alignment. Third, we leverage instance-wise negative samples to suppress noisy background regions, which calibrates the target's position, while WWbL cannot reduce the background noise. Although CTS achieves better performance than ours, it segments the target object using an auxiliary bounding-box supervision and an offline mask proposal network [4]. In contrast, our approach does not learn from manually annotated masks or bounding boxes. The comparison demonstrates that it is possible to train a RIS model only using texts.

In addition, we also report the comparisons of the PointIt and Acc_{box} accuracy with WSGs in Table 2. On the ReferIt *test* set, our framework in Step-1 brings PointIt improvement of 25.04, 24.44, and 15.52 compared to MG, GbS, and WWbL, respectively. Even using pretrained ImageNet [13] weights to initialize the image encoder, our approach still outperforms WWbL by large margins (Acc_{box}: 14.12; PointIt: 13.06). This demonstrates the effectiveness of our framework in locating and segmenting target objects. Some visual examples are shown in Fig. 4, where we can see that our method can localize the target objects in challenging scenes (*e.g.*, long and complex sentence in Image (d)) without extra annotations for training.

$\mathcal{L}_{cls}$	BP		$\mathcal{L}_{cal}$		IoU	PointIt	P@0.5	PointM
	P_{t2v}	P_{v2t}	P_{term}	N_{term}				
✓					18.94	51.42	3.87	36.93
✓	✓				19.70	53.47	4.21	39.14
✓	✓	✓			20.43	56.49	4.47	40.92
✓	✓	✓	✓		24.18	64.03	8.33	49.85
✓	✓	✓		✓	22.43	60.81	6.44	45.58
✓	✓	✓	✓	✓	**27.81**	**66.99**	**12.75**	**53.69**

Table 3. Ablation studies of different components on the RefCOCOg (U) *train* set. P_{t2v} and P_{v2t} denote the two variants detached from the proposed bilateral prompt (BP), which indicate the unilateral prompt from textual to visual features only and the opposite direction. P_{term} and N_{term} are the positive enhancement and negative suppression processes.

4.3. Ablation Studies and Analyses

We conduct ablation studies on the text-to-image response map prediction, and report the quantitative results on RefCOCOg (U) train set in Table 3. First, we remove the proposed bilateral prompt and calibration method as our baseline (1-st row), and it can already obtain an initial localization (*e.g.*, PointM: 36.93). To validate the bilateral prompt, we transform it into two unilateral prompts (*i.e.*, P_{t2v} and P_{v2t}) that update only one-way information (*i.e.*, from text to visual or the opposite direction). We can see that the bilateral prompt works better than the unilateral prompts and significantly enhances the localization than the baseline (*e.g.*, PointM increased from 36.93 $\rightarrow$ 40.92). In addition, we split the calibration method into two processes, *i.e.*, the positive foreground enhancement process (P_{term}) and negative background suppression process (N_{term}). Both improve performance, and when they are combined, performance is further boosted.

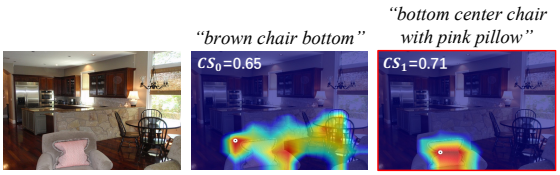


Figure 6. Qualitative comparison of with and without PRMS on RefCOCO+ set. The cumulative score of the response map for each query is placed in the upper left corner of its results. $\odot$ indicates the maximum response point. We highlight the best result with red boxes.

IoU	RefCOCO	RefCOCO+	RefCOCOg	
			Google	UMD
w/o PRMS	25.11	22.31	26.93	26.62
w PRMS	**25.90**	**24.48**	**27.33**	**27.06**

Table 4. Quantitative comparisons of our method without (w/o) and with (w) the PRMS on different *val* sets.

We visualize the effect of each component in Fig. 5. Although the model has higher responses at the target object when only Eq. (7) is used, it lacks instance information and suffers from background noise. The bilateral prompt enhances the localization of the referred target (see the shades of color in (c)), and attenuates the false responses (*e.g.*, the waterfall in case-1). The calibration method can further enhance the correctness by suppressing irrelevant noisy backgrounds (*e.g.*, the black goat in case-1 and the woman in case-3) and maintaining the completeness of the target object. In (e), Step-2 further improves the performance.

Positive Response Map Selection (PRMS). We conduct the add-on experiment on three benchmark* val sets to verify the effectiveness of the positive response map selection strategy. As shown in Table 4, PRMS generally boosts the performance on all three datasets. In particular, it brings around 10% IoU improvement on the challenging RefCOCO+ dataset. In Fig. 6, we also show visual comparisons of two cases between the RIS performance with and without PRMS. From the comparison, we can see that PRMS is able to select the best response map out of a collection of response maps that corresponds to the queries describing the same object with different properties. This facilitates the training of the segmentation network in Step-2.

Numbers of Negative Queries. The comparison results are shown in Table 5. We can see that when we do not use the negative text descriptions for classification, the performance is extremely low. As the number of negative samples gradually increases, the discriminative power of our model improves and reaches to the peak at $N+1=48$. When continuing to increase the N to 60, the performance starts to decline instead, due to the huge imbalance of positive and negative queries. We also investigate the effects of different K in our calibration method, and we can observe a similar trend (the performance is best when K=6).

*If the target has only one text expression, *e.g.*, ReferIt dataset, this strategy will degenerate to an identical mapping.

	$N+1$							K		
	1	2	6	12	24	48	60	3	6	12
IoU	11.16	19.91	22.19	24.05	26.19	**27.81**	27.61	27.05	**27.81**	27.63
PointIt	16.73	43.35	58.18	64.73	66.45	**66.99**	66.24	65.81	**66.99**	66.82
PointM	8.01	32.96	43.71	49.92	52.46	**53.69**	53.05	52.94	**53.69**	53.16

Table 5. Comparisons of numbers of negative queries N for Eq. (7) and K for Eq. (8) on the RefCOCOg (U) *train* set.

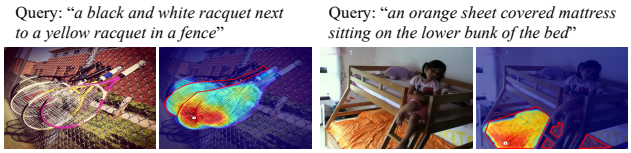


Figure 7. Two failure cases. Both images are from the RefCOCOg (U) *val* set, and the results are from Step-2. The GT objects are indicated with red contours for visualization.

5. Conclusion

In this paper, we have presented a novel RIS framework that uses only the available text descriptions as a supervision signal for training. Our work has three main technical contributions. First, we have proposed a bilateral prompt method to help harmonize the discrepancy between the visual and linguistic features. Second, we have proposed a calibration method to help reduce background noise to improve the quality of the response maps. Third, we have proposed a positive response map selection strategy to help obtain high-quality pseudo labels for training a segmentation network for RIS inference. To reduce the in-box error of the existing PointIt metric, we have proposed a new metric (PointM) for a more accurate localization evaluation. Extensive results demonstrate the effectiveness of our method using only text descriptions as the supervision signal.

Nonetheless, our approach does have limitations. If a scene has similar semantics in the foreground and background, it can be distracted by other objects of the same category, and thus produce false localization, *e.g.*, Fig. 7(a). It may also have difficulties in handling occluded target objects, *e.g.*, Fig. 7(b). A possible solution may be to incorporate structured linguistic features with visual features to enhance the model's reasoning abilities.

Acknowledgements: This work was supported by the National Key R&D Program of China #2018AAA0102003, the Fundamental Research Funds for the Central Universities DUT22JC06, and the National Natural Science Foundation of China #62172073, #62006037, #U19B2039 and #62276046.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019. 5

[2] Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. In *CVPR*, 2019. 2, 6

[3] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, 2020. 2

[4] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *CVPR*, 2014. 7

[5] Assaf Arbelle, Sivan Doveh, Amit Alfassy, Joseph Shtok, Guy Lev, Eli Schwartz, Hilde Kuehne, Hila Barak Levi, Prasanna Sattigeri, Rameswar Panda, et al. Detector-free weakly supervised grounding by separation. In *ICCV*, 2021. 2, 6, 7

[6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021. 7

[7] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *ICCV*, 2019. 2

[8] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-based image editing with recurrent attentive models. In *CVPR*, 2018. 1

[9] Yi-Wen Chen, Yi-Hsuan Tsai, Tiantian Wang, Yen-Yu Lin, and Ming-Hsuan Yang. Referring expression object segmentation with caption-aware consistency. In *BMVC*, 2019. 2

[10] Myungsub Choi. Referring object manipulation of natural images using conditional classifier-free guidance. In *ECCV*, 2022. 1

[11] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015. 2

[12] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *CVPR*, 2019. 2

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6, 7

[14] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021. 2, 5

[15] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, 2021. 1, 2

[16] Guang Feng, Lihe Zhang, Zhiwei Hu, and Huchuan Lu. Learning from box annotations for referring image segmentation. *TNNLS*, 2022. 1, 2, 6

[17] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 4

[18] Shuyong Gao, Wei Zhang, Yan Wang, Qianyu Guo, Chenglong Zhang, Yangji He, and Wenqiang Zhang. Weakly-supervised salient object detection using point supervison. In *AAAI*, 2022. 2

[19] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *ECCV*, 2020. 2

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 4

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[22] Ronghang Hu, Marcus Rohrbach, Trevor Darrell, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 1, 2

[23] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *CVPR*, 2020. 1, 2

[24] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*, 2020. 2

[25] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. In *CVPR*, 2022. 2

[26] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*, 2021. 2

[27] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 5

[28] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *CVPR*, 2022. 1, 2

[29] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. In *NeurIPS*, 2021. 2

[30] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *CVPR*, 2018. 2

[31] Zijian Liang, Pengjie Wang, Ke Xu, Pingping Zhang, and Rynson WH Lau. Weakly-supervised salient object detection on light fields. *IEEE TIP*, 2022. 2

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 4

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 5

[34] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. *arXiv preprint arXiv:2212.09506*, 2022. 6

[35] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, 2017. 2

[36] Fang Liu, Yuqiu Kong, Lihe Zhang, Guang Feng, and Baocai Yin. Local-global coordination with transformers for referring image segmentation. *Neurocomputing*, 2023. 2

[37] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *TPAMI*, 2021. 1, 2

[38] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *CVPR*, 2021. 2

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[40] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, 2020. 2

[41] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 5

[42] Tuay Edgar Margffoy, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *ECCV*, 2018. 2

[43] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 5

[44] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *ICCV*, 2017. 1

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[46] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR*, 2015. 2

[47] Yongri Piao, Jian Wang, Miao Zhang, and Huchuan Lu. Mfnet: Multi-filter directive network for weakly supervised salient object detection. In *ICCV*, 2021. 2

[48] Jie Qin, Jie Wu, Xuefeng Xiao, Lujun Li, and Xingang Wang. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In *AAAI*, 2022. 6

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 4, 5

[50] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022. 4

[51] Tal Shaharabany, Yoad Tewel, and Lior Wolf. What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. In *NeurIPS*, 2022. 1, 2, 6, 7

[52] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *ECCV*, 2018. 2

[53] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Weakly-supervised salient instance detection. *arXiv preprint arXiv:2009.13898*, 2020. 2

[54] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to detect instance-level salient objects using complementary image labels. *IJCV*, 2022. 2

[55] Josiah Wang and Lucia Specia. Phrase localization without paired training examples. In *ICCV*, 2019. 2

[56] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *CVPR*, 2021. 2

[57] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 2

[58] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*, 2019. 1

[59] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, 2022. 2

[60] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *CVPR*, 2020. 2

[61] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: cross language image matching for weakly supervised semantic segmentation. In *CVPR*, 2022. 2

[62] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, 2022. 6

[63] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *CVPR*, 2021. 2

[64] Xin Yang, Ke Xu, Shaozhe Chen, Shengfeng He, Baocai Yin Yin, and Rynson Lau. Active matting. In *NeurIPS*, 2018. 1

[65] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 1, 2, 5, 6

[66] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, 2019. 1, 2, 5

[67] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 2

[68] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 5

[69] Siyue Yu, Bingfeng Zhang, Jimin Xiao, and Eng Gee Lim. Structure-consistent weakly supervised salient object detection with local saliency coherence. In *AAAI*, 2021. 2

[70] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *CVPR*, 2020. 2

[71] Xiaoyang Zheng, Zilong Wang, Ke Xu, Sen-Yuan Li, Tao Zhuang, Qingwen Liu, and Xiaoyi Zeng. Make: Vision-language pre-training based product retrieval in taobao search. In *The ACM Web Conference (Industry Track)*, 2023. 1

[72] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2

[73] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 4

[74] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 4

[75] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *ECCV*, 2022. 2

[76] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *CVPR*, 2021. 1