SAM as the Guide: Mastering Pseudo-Label Refinement in Semi-Supervised Referring Expression Segmentation

Danni Yang [*1] **Jiayi Ji** [*1] **Yiwei Ma** [1] **Tianyu Guo** [1] **Haowei Wang** [1 2] **Xiaoshuai Sun** [1] **Rongrong Ji** [1]

Abstract

In this paper, we introduce SemiRES, a semi-supervised framework that effectively leverages a combination of labeled and unlabeled data to perform RES. A significant hurdle in applying semi-supervised techniques to RES is the prevalence of noisy pseudo-labels, particularly at the boundaries of objects. SemiRES incorporates the Segment Anything Model (SAM), renowned for its precise boundary demarcation, to improve the accuracy of these pseudo-labels. Within SemiRES, we offer two alternative matching strategies: IoU-based Optimal Matching (IOM) and Composite Parts Integration (CPI). These strategies are designed to extract the most accurate masks from SAM's output, thus guiding the training of the student model with enhanced precision. In instances where a precise mask cannot be matched from the available candidates, we develop the Pixel-Wise Adjustment (PWA) strategy, guiding the student model's training directly by the pseudo-labels. Extensive experiments on three RES benchmarks—RefCOCO, RefCOCO+, and G-Ref reveal its superior performance compared to fully supervised methods. Remarkably, with only 1% labeled data, our SemiRES outperforms the supervised baseline by a large margin, e.g. +18.64% gains on RefCOCO val set. The project code is available at https://github.com/nini0919/SemiRES.

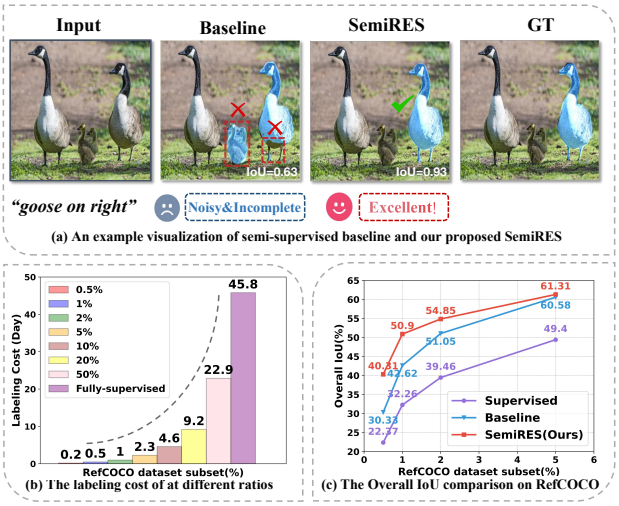

Figure 1. (a) A large number of noisy and incomplete cases exist in pseudo-labels. The proposed SemiRES can address this issue. (b) Analysis shows labeling a small portion of RefCOCO data can greatly reduce costs. (c) Our method substantially improves performance, even with a small number of annotated samples.

1. Introduction

Referring Expression Segmentation (RES) has attracted considerable attention from the fields of vision and language research (Hu et al., 2016; Chen et al., 2019b; Huang et al., 2020; Liu et al., 2019; 2023b). Unlike common visual grounding tasks such as phrase localization (Bajaj et al., 2019; Chen et al., 2017; Dogan et al., 2019; Plummer et al., 2015; 2017) and referring expression comprehension (Yang et al., 2020; Yu et al., 2018; Deng et al., 2021; Kamath et al., 2021; Huang et al., 2021), RES requires precise pixel-level segmentation of an object within an image, as directed by a referring expression, which goes beyond simple bounding box identification.

Despite advancements, the high demand for labeled data presents a significant barrier to the deployment of RES, particularly in domains where labeling is prohibitively expensive, such as medical imaging and autonomous driving. The labor intensity of the task is underscored by findings

[*]Equal contribution [1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Xiamen University, 361005, P.R. China. [2]Youtu Lab, Tencent, Shanghai, China. Correspondence to: Xiaoshuai Sun <xssun@xmu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

from (Kim et al., 2023a), indicating that an average mask segmentation requires approximately 79.1 seconds. This is further exemplified by the extensive annotation efforts needed for benchmark datasets like RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016), and G-Ref (Mao et al., 2016; Nagaraja et al., 2016), which consist of tens of thousands of labeled instances and necessitate substantial time investments, as illustrated in Fig. 1 (b). The cost and time implications, alongside the potential for inaccuracies in manual labeling, present considerable challenges for the scalability and reliability of RES models, highlighting the urgent need for more efficient methodologies.

To address the challenges outlined above, we pioneer a semi-supervised learning framework tailored for RES, utilizing a combination of a small subset of image-text pairs with segmentation annotations and a large corpus of unannotated pairs to train the model. This approach has been widely validated across the fields of computer vision (Sohn et al., 2020; Olsson et al., 2021; Chen et al., 2023), natural language processing (Miyato et al., 2016; Cheng & Cheng, 2019; Zou & Caragea, 2023), and vision-language research (Yang et al., 2023; Sun et al., 2023; Jin et al., 2023). Yet, its application in RES has not yet been explored. We establish a baseline for semi-supervised RES, encompassing a comprehensive pipeline that goes beyond data augmentation (Zhang et al., 2017; Hendrycks et al., 2019; Cubuk et al., 2019; Zhao et al., 2023a) and Exponential Moving Average (Kingma & Ba, 2014; He et al., 2020; Grill et al., 2020; Tarvainen & Valpola, 2017) training mechanisms. However, this baseline encounters a substantial challenge: pseudo-labels are significantly noisy, particularly at the edges of instances, which can trap the model in suboptimal performance, as depicted in Fig. 1 (a). The crux of semi-supervised learning lies in refining these pseudo-labels to enhance their quality. Previous methods have addressed this by employing confidence-based pseudo-label filtering strategies (Sohn et al., 2020) or auxiliary correcting networks (Kwon & Kwak, 2022; Kim et al., 2023a). While intuitively appealing, relying solely on confidence for filtering may lead to the under-utilization of unlabeled data and lack flexibility in handling diverse noise in pseudo-labels.

To tackle the aforementioned issues, we introduce a novel semi-supervised RES framework named SemiRES. The motivation behind SemiRES is to leverage the robust segmentation capabilities of SAM (Kirillov et al., 2023) to rectify pseudo-labels, especially around the edges of instances. Specifically, we employ SAM to extract multi-scale masks from original images to build a proposal library. The central concept of SemiRES is to retrieve one or multiple proposals from this library to reconstruct pseudo-labels. To achieve this, we propose two alternative strategies: IoU-based Optimal Matching (IOM) and Composite Parts Integration (CPI). The first assumes that the proposal library contains masks closely approximating the target instance, thus utilizing IoU to directly identify and replace the pseudo-label with the most corresponding mask from the library. The second strategy moves away from this assumption, instead using the pseudo-label to select different part-specific proposals from the library to assemble a complete mask. In cases where a suitable replacement cannot be retrieved from the proposal library, we default to optimizing the student model using the pseudo-label itself. To enhance training in such scenarios, we have devised a Pixel-Wise Adjustment (PWA) strategy that adjusts the final loss on a per-pixel basis according to the confidence levels on the pseudo-label.

To further qualitatively validate the effectiveness of our proposed SemiRES, we conduct extensive experiments on three RES benchmark datasets—RefCOCO, RefCOCO+, and G-Ref. Our experiments show that SemiRES notably surpasses both supervised and semi-supervised baselines in all settings, for example, gaining +18.64% and +8.28% on 1% labeled RefCOCO as shown in Fig. 1 (c), which highlights its significant real-world application potential.

To sum up, the contributions of this paper are three-fold:

- We first present SemiRES, a semi-supervised framework tailored for RES that efficiently trains models using a minimal amount of labeled data, thus reducing dependence on expensive pixel-level annotations.
- We introduce two alternative strategies, IoU-based Optimal Matching (IOM) and Composite Parts Integration (CPI), that leverage the SAM's edge-segmentation proficiency to produce superior-quality pseudo-labels.
- Our SemiRES framework achieves notable performance improvements on three benchmark datasets RefCOCO, RefCOCO+, and G-Ref, demonstrating significant gains in model accuracy while concurrently cutting down on labeling costs.

2. Related Work

2.1. Referring Expression Segmentation

Referring Expression Segmentation (Chen et al., 2019b; Huang et al., 2020; Liu et al., 2023b;a; Hu et al., 2023; Kim et al., 2023b; Yang et al., 2022b) is a multimodal task involving both image segmentation and natural language understanding. It aims to identify specific target regions within an image according to natural language expressions. In recent years, RES methods have made significant progress, with Transformer-based backbones emerging as the predominant choice for this task. Additionally, RES methods can be categorized into two main types: one-stage and two-stage approaches. One-stage methods (Chen et al., 2019a; Hu et al., 2020; Hui et al., 2020) usually use end-to-end networks for

prediction, while two-stage methods (Yu et al., 2018; Liu et al., 2022) employ an instance segmentation network to generate a set of instance proposals before selecting the target instance. Recently, GRES (Liu et al., 2023a)introduces multi-target and no-target expressions, extending the classic RES to refer to an arbitrary number of target objects.

2.2. Semi-Supervised Semantic Segmentation

Semi-supervised learning (French et al., 2019; Xu et al., 2021; Wu et al., 2021; Olsson et al., 2021; Jiang et al., 2022; Yang et al., 2023), using a small amount of labeled data alongside a larger pool of unlabeled data for training, has widespread applications in computer vision and natural language processing. In recent years, semi-supervised semantic segmentation has developed rapidly, with many new research works emerging (Ouali et al., 2020; Chen et al., 2021; Wang et al., 2022; Yang et al., 2022a). Building upon FixMatch (Sohn et al., 2020), PseudoSeg (Zou et al., 2020) extends weak consistency to strong consistency in segmentation scenarios and incorporates a calibration module for pseudo-label refinement. ReCo (Liu et al., 2021a) samples classes prone to confusion across all categories to assist the segmentation network for better representations. AugSeg (Zhao et al., 2023b), a simple yet effective method, primarily enhances the performance of semi-supervised semantic segmentation through data augmentation.

2.3. Segment Anything Model

The recent Segment Anything Model (SAM) (Kirillov et al., 2023) has made significant advancements in pushing the boundaries of segmentation, greatly boosting the development of foundational models for computer vision. Trained on SA-1B of over 1 billion masks, it aims to segment any object in any given image without requiring any additional task-specific adaptation. SAM is proved to be capable of solving various tasks, such as medical image analysis (Ma & Wang, 2023; Shi et al., 2023), adversarial attacks (Guan et al., 2023; Zhang et al., 2023a), image inpainting (Yu et al., 2023), image editing (Xie et al., 2023), image captioning (Wang et al., 2023). Recently, several works (Zhang et al., 2023b; Liu et al., 2023c) have incorporated SAM for one-shot learning, contributing significantly to SAM's multifaceted development. In this paper, we investigate how to leverage SAM to enhance pseudo-labels and improve the performance of semi-supervised learning.

3. Method

3.1. Task Definition

Before diving into SemiRES, we begin with illustrating the task definition of semi-supervised referring expression segmentation (RES). We usually have a small labeled dataset $\mathcal{D}_l = \left\{\left(\left(\mathcal{I}_i^l, \mathcal{T}_i^l\right), Y_i^l\right)\right\}_{i=1}^{N^l}$ and a much larger unlabeled dataset $\mathcal{D}_u = \left\{\left(\left(\mathcal{I}_i^u, \mathcal{T}_i^u\right), \varnothing\right)\right\}_{i=1}^{N^u}$, where $\mathcal{I}_i^l, \mathcal{I}_i^u$ denote the i-th labeled and unlabeled image, respectively; $\mathcal{T}_i^l$ and $\mathcal{T}_i^u$ are the corresponding language expressions; N^l and N^u are the number of labeled and unlabeled data, with $N^l \ll N^u$. It is crucial to note that the unlabeled set $\mathcal{D}_u$ lacks ground truth mask labels, utilizing language expressions solely as input. Our primary aim is to leverage this small labeled set alongside a large unlabeled set to achieve competitive performance in the RES task.

3.2. Semi-Supervised Baseline

We introduce a semi-supervised baseline for RES based on a teacher-student network structure. This approach unfolds in two stages:

Stage1: Burn-In Stage. In the semi-supervised teacher-student framework, achieving proper parameter initialization is crucial for accelerating the convergence of training during the mutual learning stage (Liu et al., 2021b). During the Burn-In stage, we train the pre-trained model using only labeled data. The optimization objective is defined as follows:

$$\mathcal{L}_{sup} = \frac{1}{H \times W} \sum_{j=1}^{H \times W} \mathcal{L}_{BCE}\left(M_{i,j}^l, Y_{i,j}^l\right), \quad (1)$$

where $M_{i,j}^l$ denotes the prediction mask of Burn-In model for the j-th pixel of i-th labeled image, $Y_{i,j}^l$ denotes the corresponding ground truth, $\mathcal{L}_{BCE}$ denotes binary cross entropy loss (Csiszár, 2008).

Stage2: Mutual-Learning Stage. After the Burn-In stage, we use the trained weights θ to initialize both the teacher and student models. This process is defined as follows:

$$\theta_t \leftarrow \theta, \theta_s \leftarrow \theta, \quad (2)$$

where $\theta_t, \theta_s, \theta$ denote the parameters of the teacher, student and Burn-In model, respectively.

During the mutual learning stage, the teacher generates pseudo-labels for unlabeled data to supervise the training of the student, which is defined as follows:

$$\mathcal{L}_{unsup} = \frac{1}{H \times W} \sum_{j=1}^{H \times W} \mathcal{L}_{BCE}\left(M_{i,j}^u, \hat{M}_{i,j}^u\right), \quad (3)$$

where $M_{i,j}^u$ and $\hat{M}_{i,j}^u$ denote the predicted mask for j-th pixel of i-th unlabeled image by student and teacher, respectively.

Simultaneously, the student continues to train on a small subset of labeled data, jointly optimizing with these two components of loss function, which is defined as follows:

$$\mathcal{L} = \lambda_{sup}\mathcal{L}_{sup} + \lambda_{unsup}\mathcal{L}_{unsup}, \quad (4)$$

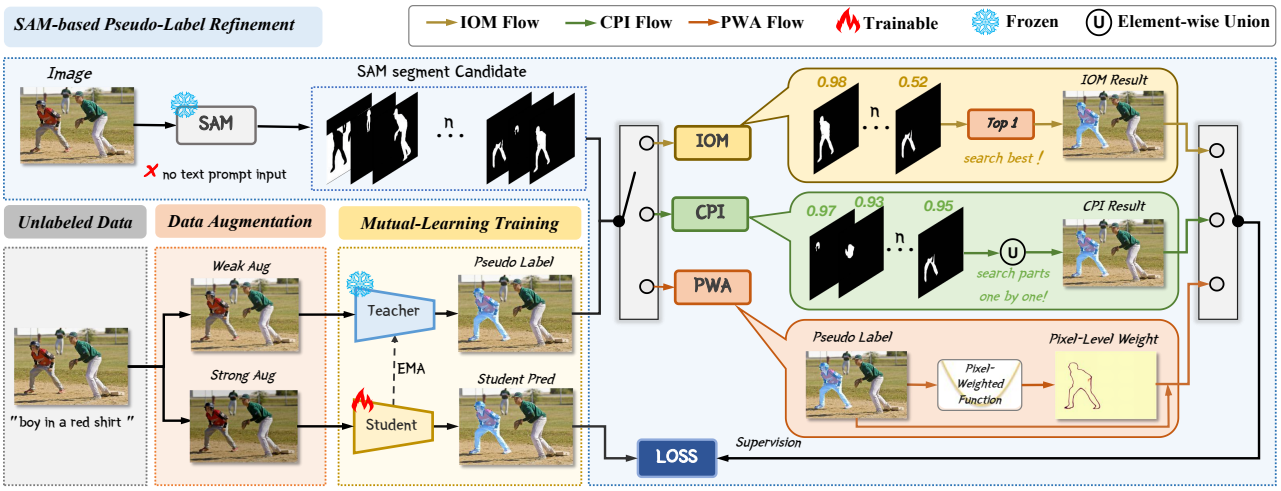

Figure 2. An overview of the proposed SemiRES, featuring a teacher-student network with data augmentation and mutual learning. It includes SAM-based pseudo-label refinement using IOM or CPI strategies, and PWA supervision when matches are not found.

where λ_{sup} and λ_{unsup} is the hyperparameter of supervised loss $\mathcal{L}_{sup}$ and unsupervised loss $\mathcal{L}_{unsup}$.

To maintain the stability of pseudo-labels, we forego gradient backpropagation for updating the teacher model's parameters. Instead, we employ the Exponential Moving Average (EMA) method to create an aggregated model reflecting both the current and previous states. EMA's effectiveness has been substantiated in numerous studies (Kingma & Ba, 2014; Ioffe & Szegedy, 2015; He et al., 2020; Grill et al., 2020; Tarvainen & Valpola, 2017). The use of EMA not only improves the teacher model's accuracy but also its stability, making it a valuable tool during the mutual learning stage, which is formulated as follows:

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s, \tag{5}$$

where α is the decay coefficient of EMA, typically set within the small range of 0.9 to 0.999.

3.3. The Proposed SemiRES

3.3.1. Overview

The overview of our proposed SemiRES is depicted in Fig. 2. SemiRES inherits the semi-supervised framework introduced in Sec. 3.2 and proposes new strategies to address the challenges of noisy pseudo-labels encountered by regular semi-supervised frameworks, which limit the extraction of knowledge from unlabeled data. The core idea is to exploit the powerful edge segmentation capabilities of SAM. The central question of our research is how to utilize these masks to refine the noisy pseudo-labels. In this paper, we propose two alternative matching strategies, IoU-based Optimal Matching (IOM) and Composite Parts Integration (CPI), to select masks that contribute to the final pseudo-labels, as detailed in Sec. 3.3.2. Moreover, when segments generated by SAM cannot be matched with pseudo-labels, we introduce a Pixel-Wise Weighted Adjustment (PWA) scheme to focus the model on more reliable pixels, thereby improving performance, as outlined in Sec. 3.3.3.

3.3.2. SAM-based Pseudo-Label Refinement

Despite SAM's powerful segmentation capabilities, effectively harnessing these for pseudo-label refinement is an area ripe for investigation. We have formulated two strategies for matching SAM-generated segments with the original pseudo-labels to improve their accuracy. Before deploying these strategies, we utilize SAM's "Segment Everything" feature to create an extensive proposal library of multi-scale candidate segments for our dataset offline, eliminating the need for specific prompts. To optimize storage space, we implement the Run Length Encoding (RLE) algorithm[1]. Considering that SAM is capable of producing hundreds to thousands of intricate segments per image, adopting efficient storage solutions is crucial. Importantly, while the RLE algorithm achieves high compression rates, it also preserves the precision of the candidate masks.

IoU-based Optimal Matching (IOM). To achieve our goal, we initially consider a more straightforward approach, premised on the robust multi-scale segmentation ability of the Segment Anything Model (SAM). We hypothesize that the proposal library, constructed as previously mentioned, likely contains a close approximation of the ideal target seg-

[1] https://en.wikipedia.org/wiki/Run-length_encoding

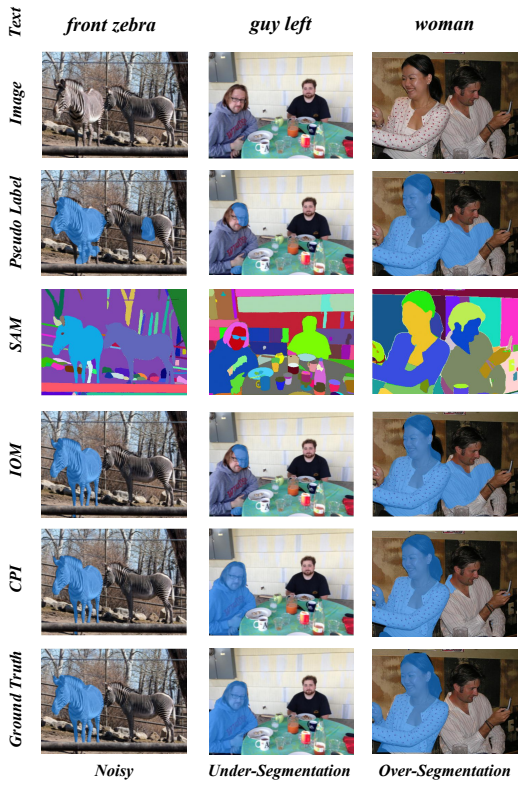

Figure 3. Visualization of the principles behind IOM and CPI addressing pseudo-label issues in different cases.

mentation. Therefore, our task simplifies devising a method to retrieve this optimal mask from the library.

Our method involves an IoU-based selection process, which computes the similarity between the pseudo-labels and each segment generated by SAM. We aim to identify the segment with the highest similarity score, ensuring it aligns closely with the overall target mask. The similarity measure is calculated using the Intersection over Union (IoU) metric, a standard in object detection and segmentation tasks that quantifies the extent of overlap between two areas. By selecting the segment with the top-1 IoU score from the candidate pool, we can effectively align our model's output with the most accurate representation of the intended segmentation, as detailed below:

$$s^k = \frac{\sum_{j=1}^{H\times W}\left(\hat{M}_{i,j}^u \cap \widetilde{M}_{i,j,k}^u\right)}{\sum_{j=1}^{H\times W}\left(\hat{M}_{i,j}^u \cup \widetilde{M}_{i,j,k}^u\right)}, \tag{6}$$

where $\hat{M}_{i,j}^u$ and $\widetilde{M}_{i,j,k}^u$ denote the pseudo-labels and k-th segment mask generated by SAM for j-th pixel of i-th unlabeled image $\mathcal{I}_i^u$, respectively. When the score s^k exceeds a certain threshold IoU_{rate}, the matched mask will replace the pseudo-label. The detailed matching schemes can be found in Algorithm 1, particularly within lines 4 to 9.

Composite Parts Integration (CPI). In our exploration of semi-supervised referring expression segmentation, we recognize that while the IoU-based Optimal Matching (IOM) strategy is generally straightforward and effective, it may falter in certain scenarios. One such instance occurs when the proposal library lacks an ideal target segmentation, rendering even the most sophisticated matching algorithm incapable of finding an appropriate guide mask. Another instance is when the disparity between the pseudo-labels and the desired segmentation is too substantial to allow for effective correction. We have noted that the original pseudo-labels generated by the teacher model can suffer from either under-segmentation or over-segmentation of the target instances, as depicted in Fig. 3. These inaccuracies diminish the quality of the pseudo-labels, providing erroneous guidance to the student model and impeding its learning.

Under-segmentation is characterized by incomplete coverage of the target instance, missing activation for certain region pixels. To address this, we aim to identify larger regions within the proposal library to rectify the pseudo-labels. Our selection is based on the overlap ratio with the pseudo-labels, calculated as follows:

$$s_1^k = \frac{\sum_{j=1}^{H\times W}\left(\hat{M}_{i,j}^u \cap \widetilde{M}_{i,j,k}^u\right)}{\sum_{j=1}^{H\times W}\left(\hat{M}_{i,j}^u\right) + \epsilon}, \tag{7}$$

where ϵ is the smoothing factor to prevent a denominator of zero. When the overlap ratio s_1^k exceeds a predefined threshold $inter_1$, the k-th segment $\widetilde{M}_{i,:,k}^u$ generated by SAM is selected and subsequently merged to replace the pseudo-labels. This method is referred to as Composite Parts Integration for Under-segmentation (CPI-U). Conversely, over-segmentation introduces erroneous regions into the segmentation. To mitigate this, we seek to leverage SAM's segmentation to filter out the extraneous noise. The selection is based on the overlap ratio with the candidate mask, computed as:

$$s_2^k = \frac{\sum_{j=1}^{H\times W}\left(\hat{M}_{i,j}^u \cap \widetilde{M}_{i,j,k}^u\right)}{\sum_{j=1}^{H\times W}\left(\widetilde{M}_{i,j,k}^u\right)}. \tag{8}$$

Likewise, when the ratio s_2^k is above the set threshold $inter_2$, the segment $\widetilde{M}_{i,:,k}^u$ generated by SAM is chosen and integrated to refine the pseudo-labels. This approach is termed Composite Parts Integration for Over-segmentation (CPI-O). When both conditions are met, we form the overarching CPI strategy. The detailed matching schemes can be found in Algorithm 1, particularly within lines 10 to 15.

3.3.3. Pixel-Wise Weighted Adjustment

Despite the effectiveness of our two strategies for refining pseudo-labels, there are cases where the scores do not ex-

Algorithm 1 Pseudo code for our proposed SemiRES

Input: Teacher's predicted pseudo mask $\hat{M}_i^u$ for i-th unlabeled image $\mathcal{I}_i^u$, the multi-scale mask $\{\widetilde{M}_{i,:,k}^u\}_{k=1}^{N_i}$ generated by SAM, the number of masks N_i generated by SAM for $\mathcal{I}_i^u$, selected strategy S, current top-1 score s_{top}
Output: Enhanced pseudo mask $\ddot{M}_i^u$

```
Initialize M̈_i^u ← ∅, s_top ← 0;
for k in 1...N_i do
    Get k-th SAM's segment M̃_{i,:,k}^u for image I_i^u;
    if S == "IOM" then
        Compute the score s^k by Eq.(6);
        if s^k > IoU_rate and s^k > s_top then
            M̈_i^u ← M̃_{i,:,k}^u , s_top ← s;
        end if
    end if
    if S == "CPI" then
        Compute the score s_1^k, s_2^k by Eq.(7) and Eq.(8);
        if s_1^k > inter_1 or s_2^k > inter_2 then
            M̈_i^u ← M̈_i^u ∪ M̃_{i,:,k}^u;
        end if
    end if
end for
if M̈_i^u == ∅ then
    Replace the enhanced pseudo-label M̈_i^u with the
    teacher's predicted pseudo-labels: M̈_i^u ← M̂_i^u;
end if
```

ceed a certain threshold, indicating a mismatch between SAM-generated segments and the current pseudo-labels. In such situations, as inspired by previous work (Yang et al., 2023), we implement the Pixel-Wise Weighted Adjustment (PWA). PWA's core objective is to assign weights to pixels based on their confidence levels. High-confidence pixels, with scores near 0 or 1, indicate certainty in foreground or background prediction and are given higher weights. In contrast, pixels with scores around 0.5, often associated with noise or ambiguity, receive lower weights to reduce their influence on training. The mapping function Ψ for translating pixel confidence into weights is defined as:

$$\Psi(\hat{M}_{i,j}^u) = \gamma - \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\hat{M}_{i,j}^u - \mu)^2}{2\sigma^2}\right), \quad (9)$$

where γ, σ^2, μ are hyperparameters, which are set to 1.3, 0.1, and 0.5 respectively.

Therefore, the loss for i-th unlabeled image $\mathcal{I}_i^u$ is defined as follows:

$$\mathcal{L}_{unsup} = \frac{1}{H \times W} \sum_{j=1}^{H \times W} \Psi(\hat{M}_{i,j}^u) * L_{BCE}\left(M_{i,j}^u, \hat{M}_{i,j}^u\right). \quad (10)$$

4. Experiment

4.1. Datasets

We verify the effectiveness of our proposed method on three standard RES benchmark datasets, RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016), and G-Ref (Mao et al., 2016; Nagaraja et al., 2016). Images in these datasets are collected from the MS-COCO dataset (Lin et al., 2014) and are attached with one or more short captions.

RefCOCO & RefCOCO+ contains 19,994, 19,992 images, with 50,000, 49,856 annotated objects and 142,209, 141,564 annotated expressions, respectively. RefCOCO and RefCOCO+ are split into four parts, *i.e.*, train, val, testA and testB. The expressions of RefCOCO are mainly about absolute position, while the ones of RefCOCO+ includes more information related to attributes.

G-Ref contains 26,711 images, with 54,822 annotated objects and 104,560 annotated expressions. In contrast, G-Ref contains more intricate expressions, with an average length of 8.4 words, making the dataset more challenging. Moreover, the G-Ref dataset is split into two distinct partitions, one maintained by UMD and the other by Google, and we present results for UMD split.

4.2. Implementation Details

Our experimental setup uses LAVT (Yang et al., 2022b) as the baseline for the RES network, employing the same Swin Transformer (Liu et al., 2021c) and BERT (Devlin et al., 2018) backbones for visual and linguistic modalities, respectively. We implement our SemiRES model in PyTorch (Paszke et al., 2019), training it on 4 RTX3090 GPUs with 3 labeled and 3 unlabeled samples per GPU. Optimization is done using the AdamW optimizer, with an initial learning rate of 5×10^{-5} and weight decay of 10^{-2}. Data augmentation includes RandomColorJitter and RandomGaussianBlur. We set the EMA rate at 0.996 and use pre-trained weights of the ViT-Huge version for SAM in generating multi-scale masks.

We use the overall Intersection-over-Union (oIoU) metric (Ding et al., 2021; Yang et al., 2022b; Liu et al., 2023b), a standard in RES, to measure the overlap ratio between predicted masks and ground truth.

4.3. Experimental Results

4.3.1. Comparison with supervised model and baseline

In Tab. 1, we conduct experiments on RefCOCO, RefCOCO+ and G-Ref under the setting of 0.5%, 1%, 2% and 5% labeled data. From the results, it can be observed that the performance of the supervised model dramatically drops when lacking sufficient labeled data. For instance,

Table 1. **Comparison of supervised, baseline and our proposed SemiRES on RefCOCO, RefCOCO+ and G-Ref.** For all approaches, we use LAVT (Yang et al., 2022b) as the RES model. "Supervised" denotes the fully-supervised training with only labeled data. "Baseline" denotes the plain semi-supervised training with data augmentation using both labeled and unlabeled data.

Methods	***RefCOCO***											
	0.5%			1%			2%			5%		
	val	testA	testB	val	testA	testB	val	testA	testB	val	testA	testB
Supervised	22.37	25.35	19.28	32.26	35.71	28.02	39.46	42.50	35.26	49.40	53.72	44.87
Baseline	30.33	35.18	25.74	42.62	48.86	37.43	51.05	54.75	46.34	60.58	64.98	54.85
SemiRES	40.31	46.48	34.88	50.90	57.54	44.48	54.85	60.39	48.52	61.31	66.64	55.94

Methods	***RefCOCO+***											
	0.5%			1%			2%			5%		
	val	testA	testB	val	testA	testB	val	testA	testB	val	testA	testB
Supervised	20.83	24.53	16.25	24.76	29.11	20.29	28.88	32.41	24.49	37.60	42.32	32.39
Baseline	25.89	30.42	20.23	30.91	35.83	24.98	36.98	41.44	30.63	46.29	52.46	38.61
SemiRES	31.99	38.06	25.92	36.49	42.86	28.58	40.41	46.84	33.30	47.00	54.42	38.74

Methods	***G-Ref***							
	0.5%		1%		2%		5%	
	val(U)	test(U)	val(U)	test(U)	val(U)	test(U)	val(U)	test(U)
Supervised	18.33	18.69	24.31	24.72	28.23	29.86	37.25	38.62
Baseline	26.02	27.62	30.91	31.51	37.07	38.55	46.67	48.39
SemiRES	31.81	33.40	34.76	36.18	42.15	43.49	47.61	50.11

with 0.5% labeled data, the overall IoU on RefCOCO val set is only 22.37%. We also compare the plain semi-supervised baseline, as mentioned in Sec. 3.2, which surpasses the supervised method in all settings, *i.e.*, exhibiting a +7.96% improvement on RefCOCO val set, with 0.5% labeled data. Most importantly, our proposed SemiRES achieves state-of-the-art performance compared to baseline. In comparison to the supervised model, SemiRES gains +17.94%, +18.64%, +15.39%, and +11.91% on RefCOCO val set under the setting of 0.5%, 1%, 2%, and 5% labeled data, respectively.

Table 2. Ablation study of SAM Matching Strategy.

SAM Matching Strategy	val	testA	testB
baseline	42.62	48.86	37.43
IOM	49.66	55.26	44.09
CPI-O	45.23	52.53	38.30
CPI-U	**50.90**	**57.54**	44.48
CPI	50.37	56.88	**44.52**

4.3.2. Ablation Study

To validate the effectiveness of components in SemiRES, we conduct the ablation study on the 1% labeled data and the remaining 99% unlabeled data on RefCOCO.

The comparison of different matching strategies. In Tab. 2, we evaluate our two proposed matching strategies: IOM and CPI. Both strategies significantly surpass the baseline, demonstrating their effectiveness. IOM, which matches the top-1 IoU mask from SAM, attains 49.66% oIoU on the RefCOCO val set. This result not only indicates the simplicity and efficacy of IOM but also corroborates SAM's exceptional segmentation ability, capable of producing ideal masks in most instances. However, IOM is slightly less effective compared to CPI, particularly when pseudo-labels segment only a small part of the target.

CPI-O, tailored for over-segmentation scenarios, effectively eliminates noise by filtering out excessively segmented small regions. Nonetheless, its performance increment is less pronounced (45.23% vs. 42.62%) when the noisy region enlarges and starts matching with new noisy areas. In contrast, CPI-U, designed to tackle under-segmentation, emerges as the most performant strategy (50.90% vs. 42.62%). This superior performance of CPI-U can be attributed to its efficient resolution of the common under-segmentation problem in pseudo-labels. When CPI-O and CPI-U are combined into a CPI strategy, there is a minor decrease in performance, likely due to CPI-U inadvertently introducing noise in the pseudo-label refinement process.

The impact of the matching threshold. In the proposed SemiRES, we use specific thresholds to optimize the matching rates for the IOM and CPI strategies. Our ablation study, shown in Tab. 3, reveals that IOM achieves its best performance with an IoU_{rate} of 0.5. For CPI, a higher performance is observed when $inter_1$ and $inter_2$ are both set

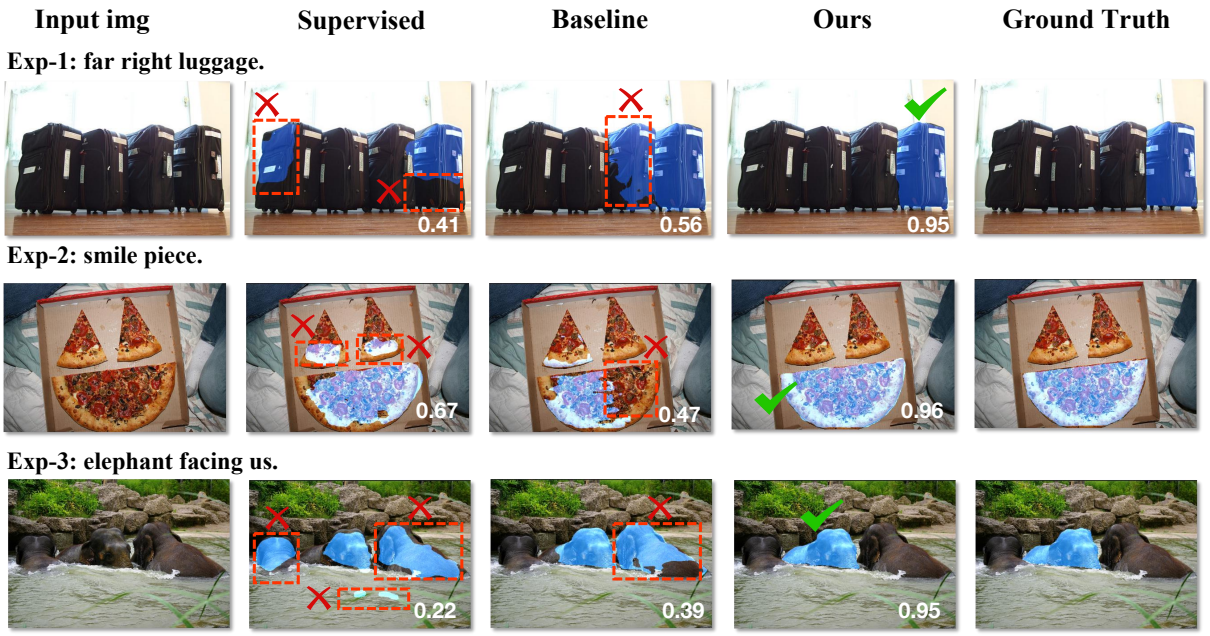

Figure 4. **Qualitative analysis for SemiRES, supervised model, semi-supervised baseline and ground truth.** The white number in the bottom right corner represents the IoU value between predicted image and ground truth. The object enclosed by the red dashed box represents incorrect segmentation. Here we use supervised and semi-supervised model trained on 1% labeled data for visualization.

Table 3. Ablation study of different thresholds for matching.

$inter_1$	$inter_2$	IoU_{rate}	val	testA	testB
-	-	0.7	48.24	55.34	41.83
-	-	0.6	49.36	56.35	42.94
-	-	0.5	49.66	55.26	44.09
0.7	0.7	-	50.37	56.88	**44.52**
0.6	0.6	-	50.18	56.97	44.51
0.5	0.5	-	49.38	56.40	43.07
0.7	-	-	**50.90**	57.54	44.48
0.6	-	-	50.13	**57.66**	43.71
0.5	-	-	50.21	57.03	44.17

to 0.7. This setting is crucial as lower thresholds in CPI could include noise, especially in the CPI-U variant, where $inter_1 = 0.7$ yields the best results. By default, unless specified otherwise, IoU_{rate}, $inter_1$, and $inter_2$ are set to 0.5, 0.7, and 0.7, respectively.

Comparison with filtering-based method. In our analysis, as presented in Tab. 4, we compare SemiRES with a confidence filtering method that eliminates the lowest 5% of pseudo-labels based on confidence scores. SemiRES demonstrates superior performance, achieving 50.90% versus 42.99% on the 1% RefCOCO validation set. This result suggests that the filtering-based method is overly rigid, leading to suboptimal use of pseudo-labels.

Effectiveness of different Components. We present our

Table 4. Comparison of SemiRES and filtering-based method.

Semi-Supervised Settings	val	testA	testB
Supervised	32.26	35.71	28.02
baseline	42.62	48.86	37.43
+confidence filtering	42.99	48.78	35.96
+SemiRES	**50.90**	**57.54**	**44.48**

Table 5. Ablation study on various components: MLT (mutual learning training), DA (data augmentation), PWA (pixel-wise adjustment), and Refine (SAM matching refinement).

MLT	DA	PWA	Refine	val	testA	testB
✗	✗	✗	✗	32.26	35.71	28.02
✓	✗	✗	✗	40.84	46.46	35.22
✓	✓	✗	✗	42.62	48.86	37.43
✓	✓	✓	✗	43.09	49.72	37.10
✓	✓	✗	✓	49.96	57.12	43.86
✓	✓	✓	✓	**50.90**	**57.54**	**44.48**

experimental analysis in Tab. 5, systematically ablated to evaluate each component of SemiRES. The first row uses only 1% labeled data for supervised training. The second row adds the remaining 99% as unlabeled data for basic semi-supervised learning. The third and fourth rows include data augmentation and the PWA module, demonstrating their effectiveness with incremental improvements (40.84% vs. 42.62% vs. 43.09%). The fifth row shows a significant

gain from using SAM for pseudo-label refinement, with a 7.34% increase (49.96% vs. 42.62%). Finally, the sixth row indicates that PWA continues to improve performance even when SAM candidates do not match the pseudo-labels, highlighting the synergistic effect of these modules.

4.4. Qualitative Analysis

We showcase qualitative results in Fig. 4, comparing SemiRES with a supervised model, a semi-supervised baseline, and ground truth. Impressively, SemiRES corrects errors from both the supervised model and the semi-supervised baseline. For instance, in the first example, while the supervised and baseline models fail to interpret "far right" correctly, leading to inaccurate identification of the luggage, SemiRES precisely localizes the target. In the second example, SemiRES effectively understands "smile" and accurately segments the correct pizza. In a more complex third scenario with several elephants, SemiRES successfully identifies the elephant facing towards us, demonstrating its advanced understanding.

5. Conclusion

In this work, we present a novel semi-supervised framework, namely SemiRES, to address the challenge of costly annotations in RES. SemiRES incorporates two innovative matching strategies that leverage the robust segmentation capabilities of SAM to refine the quality of pseudo-labels. In situations where SAM is unable to rectify the pseudo-labels, we employ the Pixel-Wise Adjustment (PWA) strategy, which utilizes the original pseudo-labels for efficient training directly. Our extensive experiments demonstrate that SemiRES achieves competitive results on three RES benchmark datasets, underscoring its viability and effectiveness for real-world applications.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

This work was supported by National Key R&D Program of China (No.2023YFB4502804), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62072389), the National Natural Science Fund for Young Scholars of China (No. 62302411), China Postdoctoral Science Foundation (No. 2023M732948), the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001), and partially sponsored by CCF-NetEase ThunderFire Innovation Research Funding (NO. CCF-Netease 202301).

References

Bajaj, M., Wang, L., and Sigal, L. G3raphground: Graph-based language grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4281–4290, 2019.

Chen, D.-J., Jia, S., Lo, Y.-C., Chen, H.-T., and Liu, T.-L. See-through-text grouping for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7454–7463, 2019a.

Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., and Savvides, M. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023.

Chen, K., Kovvuri, R., and Nevatia, R. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 824–832, 2017.

Chen, X., Yuan, Y., Zeng, G., and Wang, J. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2622, 2021.

Chen, Y.-W., Tsai, Y.-H., Wang, T., Lin, Y.-Y., and Yang, M.-H. Referring expression object segmentation with caption-aware consistency. *arXiv preprint arXiv:1910.04748*, 2019b.

Cheng, Y. and Cheng, Y. Semi-supervised learning for neural machine translation. *Joint training for neural machine translation*, pp. 25–40, 2019.

Csiszár, I. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 113–123, 2019.

Deng, J., Yang, Z., Chen, T., Zhou, W., and Li, H. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1769–1779, 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ding, H., Liu, C., Wang, S., and Jiang, X. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16321–16330, 2021.

Dogan, P., Sigal, L., and Gross, M. Neural sequential phrase grounding (seqground). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4175–4184, 2019.

French, G., Laine, S., Aila, T., Mackiewicz, M., and Finlayson, G. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Guan, Z., Hu, M., Zhou, Z., Zhang, J., Li, S., and Liu, N. Badsam: Exploring security vulnerabilities of sam via backdoor attacks. *arXiv preprint arXiv:2305.03289*, 2023.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Hu, R., Rohrbach, M., and Darrell, T. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 108–124. Springer, 2016.

Hu, Y., Wang, Q., Shao, W., Xie, E., Li, Z., Han, J., and Luo, P. Beyond one-to-one: Rethinking the referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4067–4077, 2023.

Hu, Z., Feng, G., Sun, J., Zhang, L., and Lu, H. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4424–4433, 2020.

Huang, B., Lian, D., Luo, W., and Gao, S. Look before you leap: Learning landmark features for one-stage visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16888–16897, 2021.

Huang, S., Hui, T., Liu, S., Li, G., Wei, Y., Han, J., Liu, L., and Li, B. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10488–10497, 2020.

Hui, T., Liu, S., Huang, S., Li, G., Yu, S., Zhang, F., and Han, J. Linguistic structure guided context modeling for referring image segmentation. In *European Conference on Computer Vision*, pp. 59–75. Springer, 2020.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Jiang, Y., Li, X., Chen, Y., He, Y., Xu, Q., Yang, Z., Cao, X., and Huang, Q. Maxmatch: Semi-supervised learning with worst-case consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5970–5987, 2022.

Jin, L., Luo, G., Zhou, Y., Sun, X., Jiang, G., Shu, A., and Ji, R. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2681–2690, 2023.

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Kim, B., Jeong, J., Han, D., and Hwang, S. J. The devil is in the points: Weakly semi-supervised instance segmentation via point-guided mask representation. *arXiv preprint arXiv:2303.15062*, 2023a.

Kim, D., Kim, N., Lan, C., and Kwak, S. Shatter and gather: Learning referring image segmentation with text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15547–15557, 2023b.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Kwon, D. and Kwak, S. Semi-supervised semantic segmentation with error localization network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9957–9967, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, C., Jiang, X., and Ding, H. Instance-specific feature propagation for referring segmentation. *IEEE Transactions on Multimedia*, 2022.

Liu, C., Ding, H., and Jiang, X. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23592–23601, 2023a.

Liu, D., Zhang, H., Wu, F., and Zha, Z.-J. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4673–4682, 2019.

Liu, J., Ding, H., Cai, Z., Zhang, Y., Satzoda, R. K., Mahadevan, V., and Manmatha, R. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18653–18663, 2023b.

Liu, S., Zhi, S., Johns, E., and Davison, A. J. Bootstrapping semantic segmentation with regional contrast. *arXiv preprint arXiv:2104.04465*, 2021a.

Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., and Shen, C. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023c.

Liu, Y.-C., Ma, C.-Y., He, Z., Kuo, C.-W., Chen, K., Zhang, P., Wu, B., Kira, Z., and Vajda, P. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021b.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021c.

Ma, J. and Wang, B. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.

Miyato, T., Dai, A. M., and Goodfellow, I. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.

Nagaraja, V. K., Morariu, V. I., and Davis, L. S. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 792–807. Springer, 2016.

Olsson, V., Tranheden, W., Pinto, J., and Svensson, L. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1369–1378, 2021.

Ouali, Y., Hudelot, C., and Tami, M. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12674–12684, 2020.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Plummer, B. A., Mallya, A., Cervantes, C. M., Hockenmaier, J., and Lazebnik, S. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE international conference on computer vision*, pp. 1928–1937, 2017.

Shi, P., Qiu, J., Abaxi, S. M. D., Wei, H., Lo, F. P.-W., and Yuan, W. Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics*, 13(11): 1947, 2023.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Sun, J., Luo, G., Zhou, Y., Sun, X., Jiang, G., Wang, Z., and Ji, R. Refteacher: A strong baseline for semi-supervised referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19144–19154, 2023.

Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Wang, T., Zhang, J., Fei, J., Ge, Y., Zheng, H., Tang, Y., Li, Z., Gao, M., Zhao, S., Shan, Y., et al. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*, 2023.

Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., and Le, X. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4248–4257, 2022.

Wu, J., Fan, H., Zhang, X., Lin, S., and Li, Z. Semi-supervised semantic segmentation via entropy minimization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2021.

Xie, D., Wang, R., Ma, J., Chen, C., Lu, H., Yang, D., Shi, F., and Lin, X. everything: A text-guided generative system for images editing. *arXiv preprint arXiv:2304.14006*, 2023.

Xu, Y., Shang, L., Ye, J., Qian, Q., Li, Y.-F., Sun, B., Li, H., and Jin, R. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pp. 11525–11536. PMLR, 2021.

Yang, D., Ji, J., Sun, X., Wang, H., Li, Y., Ma, Y., and Ji, R. Semi-supervised panoptic narrative grounding. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 7164–7174, 2023.

Yang, L., Qi, L., Feng, L., Zhang, W., and Shi, Y. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. *arXiv preprint arXiv:2208.09910*, 2022a.

Yang, Z., Chen, T., Wang, L., and Luo, J. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pp. 387–404. Springer, 2020.

Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., and Torr, P. H. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18155–18165, 2022b.

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016.

Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., and Berg, T. L. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1307–1315, 2018.

Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., and Chen, Z. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

Zhang, C., Zhang, C., Kang, T., Kim, D., Bae, S.-H., and Kweon, I. S. Attack-sam: Towards evaluating adversarial robustness of segment anything model. *arXiv preprint arXiv:2305.00866*, 2023a.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., and Li, H. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023b.

Zhao, Z., Yang, L., Long, S., Pi, J., Zhou, L., and Wang, J. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11350–11359, 2023a.

Zhao, Z., Yang, L., Long, S., Pi, J., Zhou, L., and Wang, J. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *CVPR*, 2023b.

Zou, H. P. and Caragea, C. Jointmatch: A unified approach for diverse and collaborative pseudo-labeling to semi-supervised text classification. *arXiv preprint arXiv:2310.14583*, 2023.

Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., and Lee, Y. J. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.

Zou, Y., Zhang, Z., Zhang, H., Li, C.-L., Bian, X., Huang, J.-B., and Pfister, T. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020.

A. The Labeling Cost at Different Ratios

In semi-supervised learning, the more labeled data there is, the higher the cost of annotation. The three RES benchmark datasets RefCOCO, RefCOCO+, G-Ref contain 50000, 49856 and 54822 annotated objects, respectively. Following the label budget calculation in (Kim et al., 2023a) , manually labeling a mask for one instance takes approximately 79.1 seconds. Therefore, it can be observed that annotating masks for the entire dataset is a very time-consuming task. Based on this, we calculate the required annotation time for labeling 0.5%, 1%, 2%, 5%, 10%, 20%, and 50% of the data on the RefCOCO as follows:

- 0.5%: 50000 × 0.005 × 79.1 / 60 / 60 / 24 = 0.2 day
- 1%: 50000 × 0.01 × 79.1 / 60 / 60 / 24 = 0.5 day
- 2%: 50000 × 0.02 × 79.1 / 60 / 60 / 24 = 1.0 day
- 5%: 50000 × 0.05 × 79.1 / 60 / 60 / 24 = 2.3 day
- 10%: 50000 × 0.1 × 79.1 / 60 / 60 / 24 = 4.6 day
- 20%: 50000 × 0.2 × 79.1 / 60 / 60 / 24 = 9.2 day
- 50%: 50000 × 0.5 × 79.1 / 60 / 60 / 24 = 22.9 day
- Fully-supervised: 50000 × 79.1 / 60 / 60 / 24 = 45.8 day

B. The Impact of the Proportion of Labeled Data

We further conduct the experiment using our SemiRES framework with a larger proportion of labeled data under the settings of 10%, 20%, 30%, 40%, and 50% of labeled data, as shown in Tab. 6. We observe that when the labeled data approaches 30%, the performance of our method nearly matches that of fully supervised models. This validates that our proposed SemiRES maintains great segmentation performance with a significant reduction in annotation costs.

C. The Impact on Different RES Frameworks

To assess the generalizability of SemiRES, we integrate it with the region-based GRES baseline ReLA (Liu et al., 2023a). As Tab. 7 shows, SemiRES significantly enhances performance compared to the "Supervised" approach (43.57% vs. 31.54%), demonstrating its robust generalization capabilities across different RES frameworks.

D. The Working Principles of IOM and CPI

To provide a clearer understanding of the working principles underlying the two SAM-based Pseudo-Label Refinement modules we have introduced, namely the IOM and CPI method, as depicted in Fig. 5. Specifically, we select a sample to elucidate how IOM and CPI operate in matching pseudo-labels. Specifically, IOM selects the best-matching segment proposal that meets the criteria from the candidate pool within SAM and uses it as the refined pseudo-label. IOM algorithm selects the segment with index 7 as the optimal match to refine the pseudo-label. On the other hand, CPI leverages the multi-scale characteristics obtained from SAM segmentation, systematically selecting qualifying partial segments, and eventually merging them together to form the final result. For CPI, it selects all segments that meet the conditions, such as indices 3, 4, 5, 6, and 7. In the end, it takes the union of these segments as the final refined pseudo-label. Observably, IOM and CPI, while both aimed at refining pseudo-labels within the SAM differ significantly in their operational approach and specialization. IOM is tailored for simpler, less noisy label refinement tasks, while CPI is designed to address more complex segmentation errors, together providing a comprehensive solution for enhancing the accuracy of pseudo-labels in segmentation tasks.

Table 6. The impact of the proportion of labeled data.

Setting	Labeled Fraction	val	testA	testB
SemiRES	10%	63.60	68.43	60.20
	20%	65.97	69.15	62.08
	30%	68.54	70.62	63.90
	40%	68.76	71.69	65.23
	50%	69.45	72.94	65.68
Fully-Supervised	-	72.73	75.82	68.79

Table 7. Results of different state-of-the-art (SOTA) RES approaches equipped with the SemiRES strategy.

RES Baseline	Settings	val	testA	testB
LAVT (Yang et al., 2022b)	Supervised	32.26	35.71	28.02
	Baseline	42.62	48.86	37.43
	SemiRES	**50.90**	**57.54**	**44.48**
GRES (Liu et al., 2023a)	Supervised	31.54	35.93	25.37
	Baseline	38.65	43.53	34.63
	SemiRES	**43.57**	**49.05**	**39.05**

E. Comparasion with other class-agnostic proposal network

To validate the tight combination between our SemiRES and SAM, we replace SAM with SEEM (Zou et al., 2024) for generating class-agnostic mask proposals and applied our matching strategy to refine pseudo-labels. As shown in Tab. 8, the results maintained under the same configurations as detailed in our paper, indicated that the performance did not match that achieved using SAM for proposal extraction. Our analysis suggests that this discrepancy stems from the designed CPI algorithm capitalizing on SAM's strong seg-

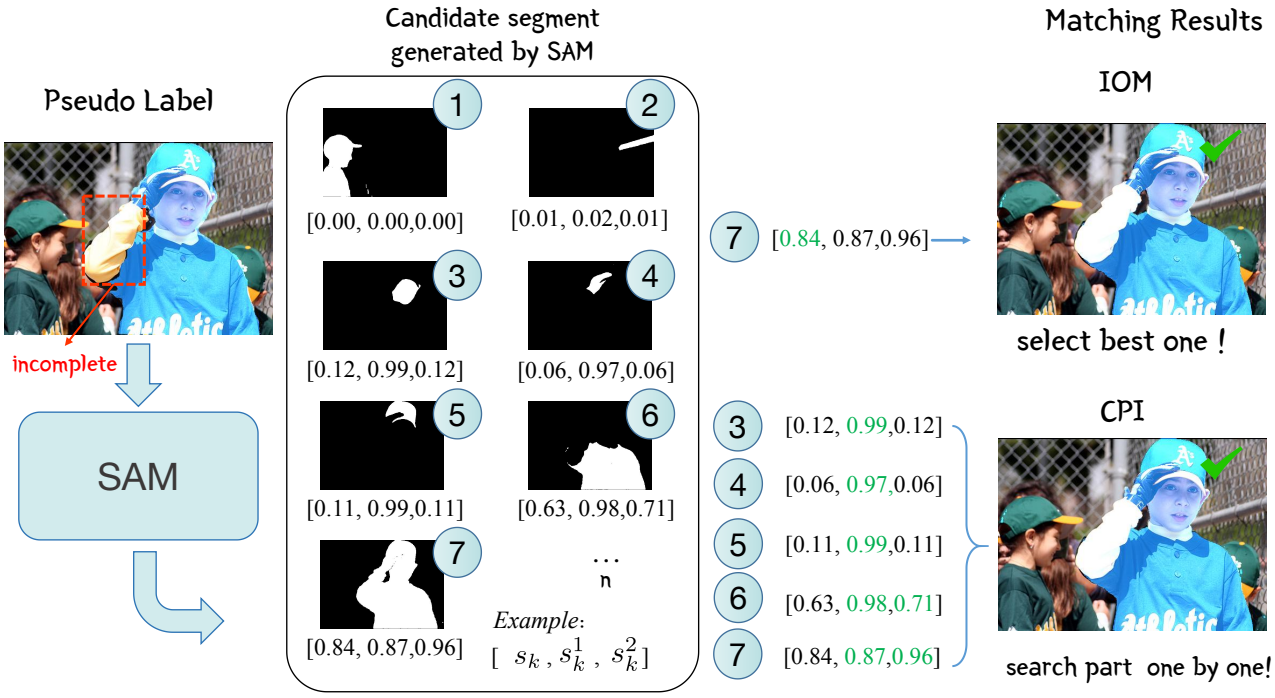

Figure 5. Demonstration of how IOM and CPI operate in matching pseudo-labels. The caption for this image is "kid looking at you".

Table 8. Comparison with other class-agnostic proposal network.

Semi-Supervised Settings	val	testA	testB
Supervised	32.26	35.71	28.02
SemiRES(+SEEM)	47.61	56.28	43.62
SemiRES(+SAM)	**50.90**	**57.54**	**44.48**

Table 9. The potential of SemiRES for detecting small objects.

Setting	Method	val	testA	testB
Top 5% small objects	Supervised	11.37	13.84	11.75
	SemiRES	20.45	27.21	17.04
Top 10% small objects	Supervised	13.70	14.88	12.59
	SemiRES	24.98	29.76	18.05

mentation capability and its ability to extract multi-scale proposal masks, which facilitated effective pseudo-label optimization, a feat not typically achievable with general class-agnostic proposal networks.

F. The potential of SemiRES for detecting small objects

To explore the potential of our proposed SemiRES method for detecting small objects, we curated samples from the RefCOCO training, validation, and test sets, organizing them by the size of the ground truth masks from smallest to largest, and specifically selected the top 5% and 10% of samples featuring small objects. As shown in Tab. 9, experiments conducted on these subsets serve to validate the efficacy of our SemiRES approach in comparison to traditional supervised methods when focusing on small objects. Our findings reveal that our method significantly outperforms the supervised approach in this area. This improvement is attributed to the SAM's ability to utilize a multi-scale library of offline-generated masks, which includes candidates for small objects, allowing for the refinement of pseudo-labels.

G. Quantitative Statistics for Pseudo-label Refinement

Due to the presence of noise and incompleteness in pseudo-labels, our proposed two SAM-based Pseudo-Label Refinement modules effectively enhance the quality of pseudo-labels, facilitating the mutual learning process between teacher and student models. To further analyze the roles of the designed IOM and CPI, along with their variants, we numerically compute their frequency for positive, negative corrections and no correction of pseudo-labels, as illustrated in Fig. 6. Here we use the ground truth of all unlabeled data in three RES datasets to validate the correction performance of our methods. Positive correction indicates that, after

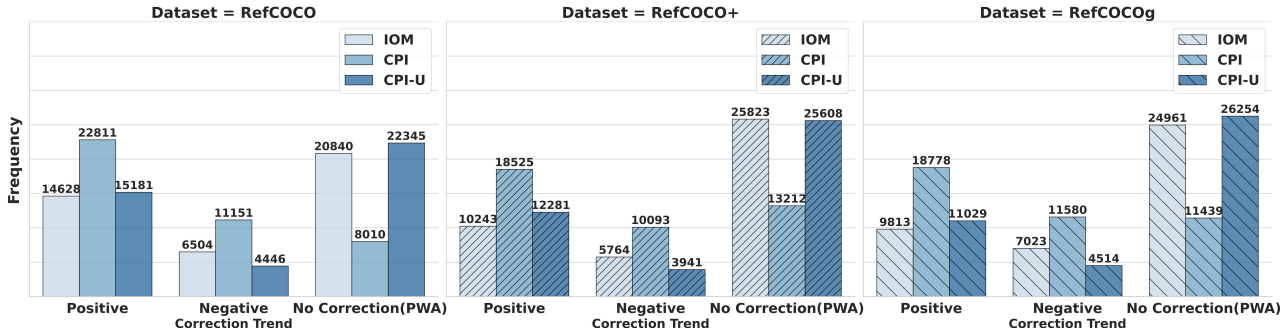

Figure 6. Quantitative statistics of positive, negative corrections and no corrections(PWA) for IOM, CPI and variants.

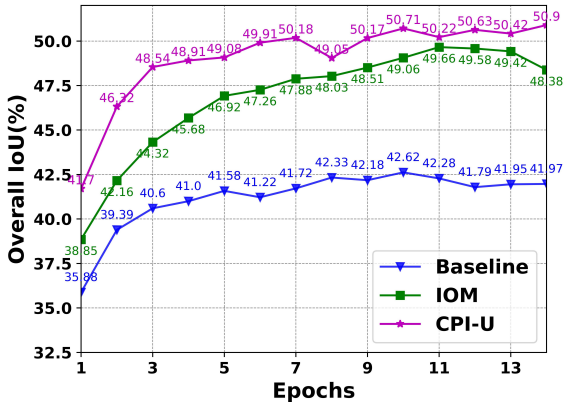

Figure 7. Training curves of two matching strategies of SemiRES and semi-supervised baselines.

the matching refinement, the corrected pseudo-label has a higher IoU with the ground truth than the original pseudo-label, thereby improving the quality of the pseudo-label. Negative correction, on the other hand, is the opposite; the quality of the updated pseudo-label is worse, exacerbating the misguidance in student learning. No correction indicates that the matching algorithms do not find a suitable correction for the pseudo label, thus the Pixel-Wise Adjustment (PWA) strategy mentioned in this paper is used for adjustment. The histogram results show that, among the three datasets for RES, the CPI method performs best in positive correction, while its variant CPI-U has the least occurrence of negative correction. And CPI-U exhibits faster convergence as shown in Fig. 7.

H. Additional Visualizations

We present more comparative visualizations of our proposed SemiRES with both supervised and baseline models, alongside the ground truth, as illustrated in Fig. 8. Through these extensive examples, it becomes apparent that our proposed SemiRES excels in understanding semantic attributes such as absolute positional terms, spatial relations, and colors. Moreover, it demonstrates a strong capability in comprehending the semantics of complex sentences. Furthermore, it adeptly handles intricate details, showcasing robust proficiency in managing nuanced information. These visualizations underscore the effectiveness of our proposed SemiRES in correcting noisy and incomplete pseudo-labels, significantly enhancing performance, especially in scenarios with limited data.

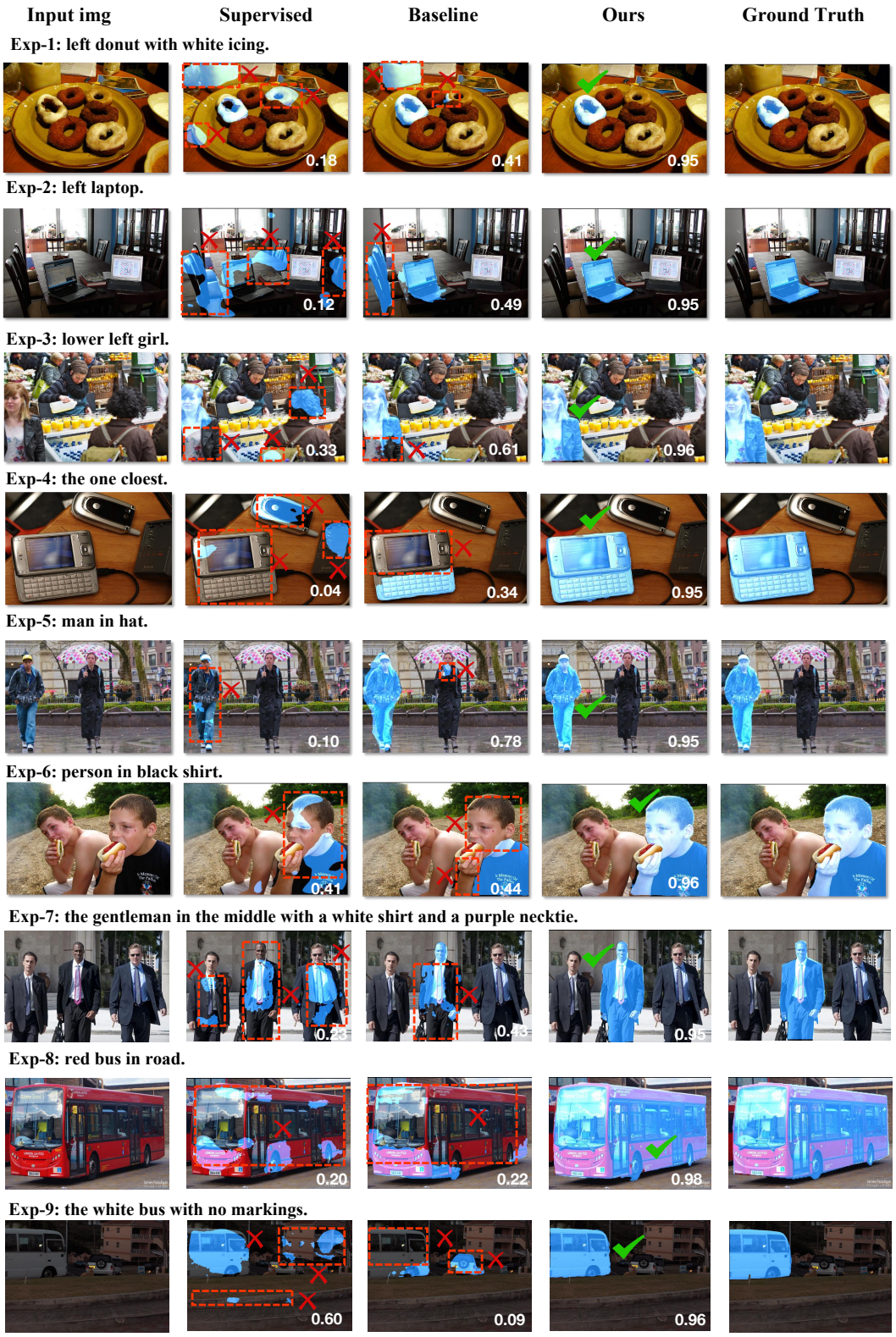
Input img
Supervised
Baseline
Ours
Ground Truth
Exp-1: left donut with white icing.
0.18
0.41
0.95
Exp-2: left laptop.
0.12
0.49
0.95
Exp-3: lower left girl.
0.33
0.61
0.96
Exp-4: the one cloest.
0.04
0.34
0.95
Exp-5: man in hat.
0.10
0.78
0.95
Exp-6: person in black shirt.
0.41
0.44
0.96
Exp-7: the gentleman in the middle with a white shirt and a purple necktie.
0.95
Exp-8: red bus in road.
0.20
0.22
0.98
Exp-9: the white bus with no markings.
0.60
0.09
0.96

Figure 8. More visualization results of our proposed SemiRES, compared with the supervised and baseline model. The red dashed bounding boxes denote regions where our model has made accurate predictions, while other models have made inaccurate predictions.