

# HARIS: Human-Like Attention for Reference Image Segmentation

1<sup>st</sup> Mengxi Zhang

School of Electrical and Information Engineering  
Tianjin University  
Tianjin, China.  
mengxizhang@tju.edu.cn

2<sup>nd</sup> Heqing Lian

Xiao Ying AI Lab  
Xiao Ying Company  
Beijing, China  
lianheqing@xiaoyingai.com

3<sup>rd</sup> Yiming Liu

Xiao Ying AI Lab  
Xiao Ying Company  
Beijing, China  
liuyiming@xiaoyingai.com

4<sup>th</sup> Jie Chen\*

School of Electronics Engineering and Computer Science  
Peking University  
Shenzhen, China  
chenj@pcl.ac.cn

**Abstract**—Referring image segmentation (RIS) aims to locate the particular region corresponding to the language expression. Existing methods incorporate features from different modalities in a *bottom-up* manner. This design may get some unnecessary image-text pairs, which leads to an inaccurate segmentation mask. In this paper, we propose a referring image segmentation method called HARIS, which introduces the Human-Like Attention mechanism and uses the parameter-efficient fine-tuning (PEFT) framework. To be specific, the Human-Like Attention gets a *feedback* signal from multi-modal features, which makes the network center on the specific objects and discard the irrelevant image-text pairs. Besides, we introduce the PEFT framework to preserve the zero-shot ability of pre-trained encoders. Extensive experiments on three widely used RIS benchmarks and the PhraseCut dataset demonstrate that our method achieves state-of-the-art performance and great zero-shot ability.

**Index Terms**—Referring Image Segmentation, Vision-Language Understanding, Multi-modal Learning.

## I. INTRODUCTION

Referring image segmentation [1] (RIS) is a typical multi-modal task, which aims to locate the particular region corresponding to the language expression. Different from conventional image segmentation that generates masks of some fixed categories, the target of RIS is to locate the referents according to the free-form language expressions. Therefore, the main challenge for RIS is to get a comprehensive vision-language understanding and effectively fuse features from different modalities.

Existing RIS methods design different methods to fuse visual and linguistic features. Recently, some methods [2]–[4] introduce the attention mechanism for vision-language fusion and achieve significant improvement. CRIS [2] adopts the cross-attention mechanism to get multi-modal features. ReLA [4] designs the region-based attention to explicitly model relationships between image regions and each word. However, these methods only explore the vision-language

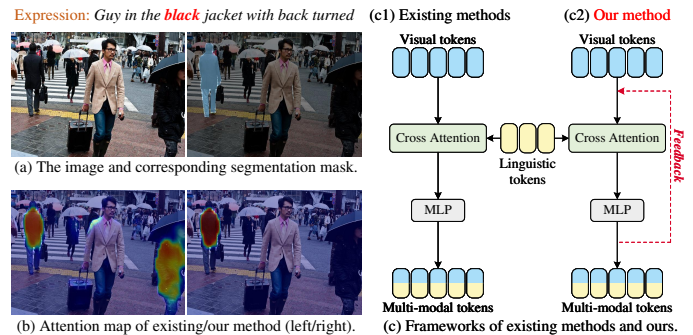


Fig. 1: The attention maps and frameworks of existing methods and our method. Existing methods fuse features from different modalities based on the cross-attention mechanism, which follows a *bottom-up* manner, as shown in the left part of (c). As a result, the attention map generated by these methods may contain some irrelevant image-text pairs (the left part of (b)). For example, the correct region for ‘black’ is the left part of the image. However, existing methods also take the right part as relevant regions of ‘black’. Different from existing methods, our method (the right part of (c)) introduces a feedback signal, which comes from modulated multi-modal tokens. Therefore, our method gets an accurate region for the word ‘black’ (the right part of (b)).

relationships in a *bottom-up* manner, which may lead to some unnecessary vision-language pairs. As shown in Fig. 1, only the guy on the left accords with the expression ‘Guy in the black jacket with his back turned’. However, conventional fusion methods take both the right and left image regions as the relevant objects for ‘black’ due to the bottom-up mechanism.

In this paper, we propose a novel RIS method that uses the Human-Like Attention mechanism based on the parameter-efficient fine-tuning framework. First, we design the Human-Like Attention to avoid unnecessary vision-language pairs. Unlike conventional bottom-up fusion methods, we introduce an

\*: Corresponding Author

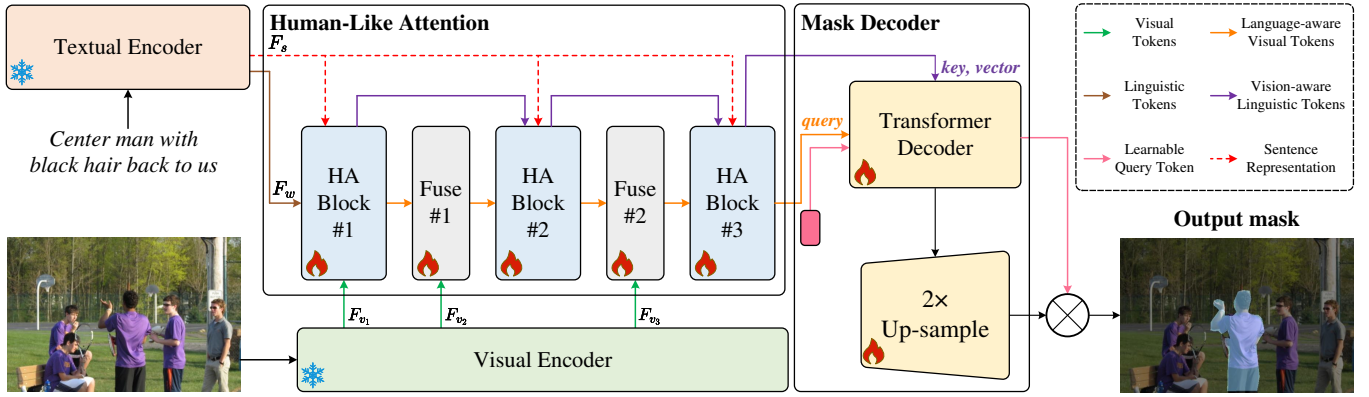


Fig. 2: The overview of HARIS. The input image is fed into the image encoder and outputs visual features ( $F_{v_1}, F_{v_2}, F_{v_3}$ ) from different layers. Correspondingly, we send the language expression to the text encoder and get linguistic features  $F_w$  and sentence representation  $F_s$ . Then, these features are sent into the Human-Like Attention blocks to get multi-modal features. Besides, we use a hierarchical architecture to use both semantic and grained visual features. Last, we send the multi-modal features and a learnable query token to the Transformer Decoder and get the final segmentation mask.

additional *feedback* signal derived from multi-modal features. This signal is integrated with the visual tokens, serving as the secondary input for the attention block. This innovative design is inspired by bionics. Specifically, in human’s vision and thought, the cognition of things usually iteratively advances, using new information to refine existing knowledge. Our Human-Like Attention mechanism pushes the network to re-visit feature fusion across varied modalities. Thus, our method makes the network center on the specific object and neglects the irrelevant ones, as depicted in Fig. 1, where the right part about ‘black’ is discarded. Second, we leverage the parameter-efficient fine-tuning (PEFT) framework to preserve the zero-shot ability of the visual and textual encoder. In the training process, we only set the parameters of the fusion block and mask decoder learnable, which saves computational resources and avoids catastrophic forgetting. To be summarized, our main contributions are summarized as follows:

- We design a novel RIS method called HARIS, which leverages the PEFT framework to preserve the great generation ability for encoders of different modalities.
- We propose Human-Like Attention to reduce the interference from unnecessary vision-language pairs, which makes the network focus on the referred object.
- Our approach exhibits state-of-the-art performance on three widely used RIS datasets, *i.e.*, RefCOCO, RefCOCO+, and G-Ref. Additionally, our method achieves excellent zero-shot ability on the PhraseCut dataset.

## II. RELATED WORK

### A. Referring Image Segmentation

Prior RIS approaches [5], [6] concatenate visual and linguistic features before sending them to the convolution neural networks (CNN) to generate multi-modal features. However, these methods are sub-optimal for multi-modal fusion due to the limitations of CNN modeling. Since the attention mechanism demonstrates promising performance in diverse

domains, certain methods [3], [4], [7] exploit the attention mechanism in the field of RIS. ReSTR [7] leverages the Transformer architecture to capture long-range dependencies of different modalities. More recently, ReLA [4] proposes the region-based attention mechanism to model the region-region and region-language dependencies explicitly. However, these methods only use *bottom-up attention*, which may lead to some irrelevant vision-language pairs. To address this issue, we propose the Human-Like Attention mechanism to make the model rethink the relationships between visual and linguistic features, similar to human cognitive processes.

### B. Attention Mechanism

Recently, the attention mechanism has achieved great success in various vision-language tasks. Pioneer methods [8], [9] use the Transformer-decoder-based architecture to align visual and linguistic features. Later, GLIP [10] introduces a bidirectional attention mechanism to align these features more precisely. Recently, the multi-modal large language models (MLLM) have demonstrated great performance on various multi-modal tasks. Typically, the architectures of MLLM are a stack of Transformers. In this paper, we exploit the practicality of top-down attention in RIS, which may inspire other vision-language tasks.

## III. METHOD

The overall framework of HARIS is shown in Fig. 2. We first extract visual features ( $F_{v_1}, F_{v_2}, F_{v_3}$ ) and linguistic features ( $F_w, F_s$ ) from frozen image encoder and text encoder, respectively. Specifically,  $F_{v_1}/F_{v_2}/F_{v_3} \in \mathbb{R}^{H \times W \times C}$  symbolize the visual features from shallow/middle/deep layers, where  $H$  and  $W$  denote height and width of the feature map,  $C$  is the number of channels. For the convenience of feature fusion, we flatten these visual features into a sequence, forming them into the shape of  $L_v \times C$ ,  $L_v = H \times W$ .  $F_w \in \mathbb{R}^{L_t \times C_t}$  denotes feature for each word, and  $F_s \in \mathbb{R}^{1 \times C_t}$  is the representation

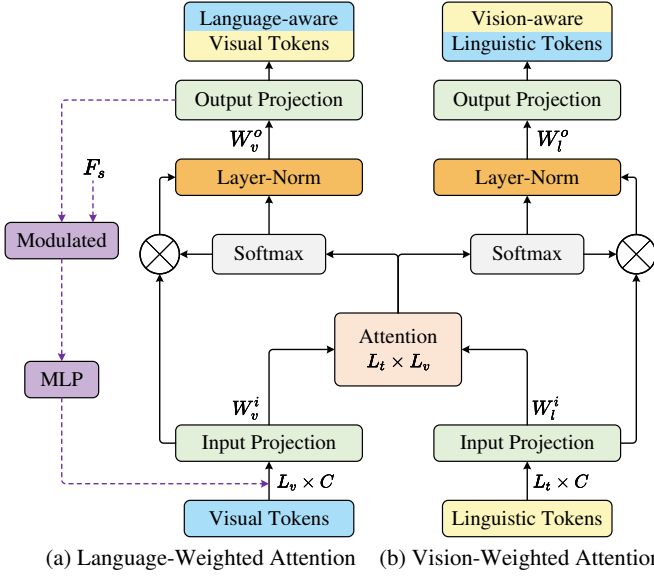


Fig. 3: The architecture of Human-Like Attention block. This block consists of two branches: Language-Weighted Attention and Vision-Weighted Attention. In the Language-Weighted Attention, we introduce the feedback signal, which is modulated by  $F_s$  and fed into the MLP layer. Then, the feedback signal together with visual tokens acts as the second-round inputs for Language-Weighted Attention.

of the whole sentence. The sentence length is denoted as  $L_t$ , while the channel number of linguistic features is represented by  $C_t$ .

Then, we send the visual features and linguistic features into the Human-Like Attention block to obtain the multi-modal features. In particular, we utilize both semantic and grained information by a hierarchical architecture. Finally, the Mask Decoder takes the multi-modal features to get the segmentation mask. The details of each module are described in the following section.

#### A. Human-Like Attention Block

Previous methods fuse visual features based on the bottom-up attention mechanism. Although these methods show satisfactory results, they may produce some irrelevant image-text pairs, as shown in Fig. 1. To address this issue, we propose the integration of Human-Like Attention block, leveraging an additional feedback signal to eliminate redundant image-text pairs. The architecture of the Human-Like Attention block is depicted in Fig. 3.

First, we model the relationships between visual tokens and linguistic tokens as follows,

$$\begin{aligned} E_v &= F_v W_v^i, E_l = F_l W_l^i, \\ A &= \text{Softmax}\left(\frac{E_v E_l^T}{\sqrt{C}}\right), \end{aligned} \quad (1)$$

where  $W_v^i \in C \times C$  and  $W_l^i \in C_t \times C$  are two learnable matrices to transform tokens of  $F_v$  and  $F_l$  from different

modalities into the same feature dimension.  $A \in L_v \times L_t$  is the attention matrix.  $\frac{1}{\sqrt{C}}$  denotes the scale factor.

Then we get the multi-modal features in a bidirectional way, which is formulated by,

$$\begin{aligned} F_{l2v} &= \text{LayerNorm}(A E_l + E_v) W_v^o, \\ F_{v2l} &= \text{LayerNorm}(A^T E_v + E_l) W_l^o, \end{aligned} \quad (2)$$

where  $\text{LayerNorm}(\cdot)$  denotes the layer-normalization layer.  $W_v^o$  and  $W_l^o$  denote weights of linear layers for mapping.  $F_{l2v}$  and  $F_{v2l}$  symbolize language-aware visual tokens and vision-aware linguistic tokens, which are the outputs of Language-Weighted Attention and Vision-Weighted Attention, respectively. After we get these multi-modal features, we design an extra branch to get the feedback signal. This branch can be viewed as the human's thinking process, where they use new knowledge to refine existing knowledge. Such a design makes the model center on the referring objects accurately. The feedback branch is shown by the purple dash lines in Fig. 3.

Specifically, we first use the whole sentence representation  $F_s^T$  to get modulated features  $\bar{F}_{l2v}$ . The mathematical process is shown below.

$$\bar{F}_{l2v} = \text{Softmax}\left(\frac{F_{l2v} F_s^T}{\sqrt{C}}\right) F_s + F_{l2v}. \quad (3)$$

Then, the feedback signal is obtained by a Multi-Layer Perceptron (MLP). We add the feedback signal to the visual tokens as the second-round inputs of Language-Weighted Attention. Finally, we run the feed-forward path of Language-Weighted Attention again and get the final language-aware visual tokens.

#### B. Hierarchical Design

A high-quality referring image segmentation mask requires both global and local information, which serves for precise location and accurate details. To this end, we introduce the hierarchical design to utilize both semantic and grained visual features:  $F_{v1}$ ,  $F_{v2}$ , and  $F_{v3}$ . Specifically, we fuse visual features from different layers after each Human-Like Attention block. Take Fuse #1 as the example, the mathematical process is formulated as follows,

$$\hat{F}_{v2} = \text{CBA}(\text{Concat}(\bar{F}_{l2v}, F_{v2})), \quad (4)$$

where  $\text{CBA}(\cdot)$  is the sequential operation consisting of convolutional layer with the kernel size of  $3 \times 3$  and stride 1.  $\hat{F}_{v2}$  is the input for next Human-Like Attention Block #2.

#### C. Mask Decoder

To fully utilize the multi-modal tokens from dual attention branches, we build the Mask Decoder based on the standard Transformer Decoder architecture [11], [12]. In particular, we concatenate a learnable query token  $M$  with vision-aware linguistic tokens as the query of the Transformer. Correspondingly, we set the language-aware visual tokens as the

TABLE I: IoU comparisons with previous state-of-the-art methods. U: UMD split; G: Google split.

Methods	RefCOCO			RefCOCO+			G-Ref		
	val	testA	testB	val	testA	testB	val (U)	test (U)	val (G)
ReSTR (CVPR 2022)	67.22	69.30	64.45	55.78	60.44	48.27	-	-	54.48
CRIS (CVPR 2022)	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36	-
ETRIS (ICCV 2023)	70.51	73.51	66.63	60.10	66.89	50.17	59.82	59.91	57.88
RefTR (NIPS 2021)	70.56	73.49	66.57	61.08	64.69	52.73	58.73	58.51	-
LAVT (CVPR 2022)	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	60.50
VLT (TPAMI 2022)	72.96	75.96	69.60	63.53	68.43	56.92	63.49	66.22	62.80
MCRES (CVPR 2023)	74.92	76.98	70.84	64.32	69.68	56.64	63.51	64.90	61.63
<b>HARIS (Ours)</b>	<b>76.17</b>	<b>78.45</b>	<b>71.37</b>	<b>66.01</b>	<b>71.92</b>	<b>57.02</b>	<b>65.05</b>	<b>66.60</b>	<b>64.55</b>

key and value. In the training process, the learnable query token assembles the linguistic information and interacts with the visual information, thus getting a comprehensive vision-language understanding. This process is formulated as follows.

$$F^m = \text{TD}(\text{Concat}(M, F_{l2v}), F_{v2l}, F_{v2l}), \quad (5)$$

where  $\text{TD}(q, k, v)$  represents a standard Transformer Decoder with the input of query/key/vector ( $q/k/v$ ).  $M$  symbolizes the learnable query token.  $\text{Concat}(\cdot, \cdot)$  represents the concatenation operation.

After that, we up-sample  $F^m$  by two sequential blocks, each of which contains a convolution layer, a batch-normalization layer, and an up-sample operation. Finally, we multiply the up-sampled features with the learnable query token  $M$  and get the final segmentation mask.

#### D. Loss

We adopt the linear combination of focal loss [13]  $\mathcal{L}_f$  and dice loss [14]  $\mathcal{L}_d$  as the optimizing target. The overall loss  $\mathcal{L}$  is calculated as follows.

$$\mathcal{L} = \mathcal{L}_f + \mathcal{L}_d. \quad (6)$$

## IV. EXPERIMENTS

### A. Datasets and Metrics

We evaluate our method on three typical RIS datasets, *i.e.*, RefCOCO & RefCOCO+ [15], and G-Ref [16]. These three datasets, sourced from MSCOCO [17], are annotated with different linguistic styles. The average expression length of RefCOCO/RefCOCO+ is approximately 3.61/3.53 words, respectively. Notably, RefCOCO+ differs from RefCOCO by excluding expressions related to absolute positions, such as 'left/right'. G-Ref has a longer average expression length of 8.4 words. Similar to prior studies, we evaluate RefCOCO and RefCOCO+ across validation/testA/testB splits. For G-Ref, we employ both partitions of UMD and Google.

As for the evaluation metrics, we adopt two widely used metrics, *i.e.*, mask Intersection-over-Union (IoU) score and precision with thresholds ( $\text{Pr}@X$ ). The IoU score reflects the quality of predicted masks, while  $\text{Pr}@X$  is the proportion of predicted masks that achieve an IoU score exceeding a specific threshold  $X \in \{70, 80, 90\}$ .

### B. Implementation Details

We utilize a visual encoder pre-trained by [18] and a text encoder pre-trained by [19]. During the training process, we set the parameters of these encoders frozen. The maximum word length is tailored to each dataset: for RefCOCO and RefCOCO+, it is set at 17 words; while for G-Ref, the limit is extended to 25 words. We train the network with the Adam optimizer for 50 epochs, where the initial learning rate is  $1e^{-4}$  and decreases by a factor of 0.1 at the 30th epoch. All experiments are conducted on 4 Nvidia V100 with a batch size of 16.

### C. Main Results

We compare our proposed method, HARIS, with a series of previous state-of-the-art (SOTA) methods on three widely-used datasets, *i.e.*, RefCOCO, RefCOCO+, and G-Ref. According to Tab. I, our approach exceeds others on each split of all datasets even with frozen visual and text encoders.

On the RefCOCO dataset, the IoU performance of our proposed method surpasses the second-best SOTA method, *i.e.*, MCRES [20]. To be specific, our method achieves 1.25%, 1.47%, and 0.53% IoU gain on the val, testA, and testB split, respectively. The significant improvement reveals the effectiveness of our proposed HARIS.

Besides, our HARIS also achieves new SOTA performance by 1.69%/2.24%/0.10% on three splits, which verifies that our method is also competitive for expressions without absolute positioning, such as 'left', 'right'.

Finally, on the most challenging G-Ref dataset where the length and style are various, our method improves over other SOTA methods by 1.54%, 0.38%, and 1.75% IoU on the val (U), test (U), and val (G) split, respectively. Based on these significant improvements, we assert that our method gets a more holistic image-text understanding ability and thus gets a high-quality segmentation mask.

### D. Ablation Study

We conduct ablation experiments to investigate the specific contributions of each component of HARIS on the RefCOCO val split.. The first experiment (# 1) removes the hierarchical structure (HS) and only utilizes the features from the last layer of the visual encoder. In the second experiment (# 2), the decoder is replaced by point-wise multiplying multi-modal features and globally averaged language features, which is similar to previous works [2]. In the third experiment (# 3), we

TABLE II: The ablation study of each component in our HARIS.

Settings	Methods	Pr@70	Pr@80	Pr@90	IoU
# 1	w/o HS	74.87	65.42	38.16	71.75
# 2	r. DE	79.26	69.98	41.15	74.98
# 3	r. CA	77.45	67.67	41.08	74.72
# 4	<b>Ours</b>	<b>80.35</b>	<b>71.79</b>	<b>42.72</b>	<b>76.17</b>

TABLE III: The ablation study of each component in our HARIS.

Settings	Methods	Pr@70	Pr@80	Pr@90	IoU
# 5	w/o FB	77.41	68.74	40.62	75.06
# 6	w/o VW	78.02	68.43	41.36	74.89
# 7	w/o LW	75.96	65.95	38.74	72.86
# 4	<b>Ours</b>	<b>80.35</b>	<b>71.79</b>	<b>42.72</b>	<b>76.17</b>

replace our Human-Like Attention Block with a conventional cross-attention block.

1) *Effectiveness of Human-Like Attention*: To alleviate irrelevant image-text pairs, we design the Human-Like Attention mechanism. As illustrated in Tab. II, the replacement of the conventional cross-attention (CA) block brings 1.45% descent (# 3), which indicates the effectiveness of the feedback signal. To get a more comprehensive understanding of Human-Like Attention, we conduct a series of experiments to evaluate each design in Human-Like Attention. First, we remove the feedback (FB) branch and only preserve the bidirectional attention mechanism. As shown in Tab. II, this strategy (# 5) results in 1.11% IoU drop. This is because the feedback branch reduces some unnecessary image-text pairs. Then, we explore the impacts of Vision-Weighted (VW) Attention (# 6) and Language-Weighted (LW) Attention (# 7). According to Tab. II, Vision-Weighted (VW) Attention and Language-Weighted (LW) Attention bring 1.28% and 3.31% IoU improvement, respectively. These improvements indicate that Language-Weighted Attention is more important than Vision-Weighted Attention.

2) *Effectiveness of PEFT Framework*: To demonstrate the superiority of the PEFT framework, we compare our method with others in terms of zero-shot ability. Specifically, we use the test split of PhraseCut [21] to emphasize the zero-shot feature. PhraseCut contains 1287 categories, much more abundant than the 80 categories in COCO. We report the zero-shot performance of our method and others in Tab. IV.

TABLE IV: Zero-shot performance of different methods.

Training Set	IoU results on PhraseCut		
	CRIS	LAVT	<b>Ours</b>
RefCOCO	15.53	16.68	<b>21.62</b>
RefCOCO+	16.30	16.64	<b>21.30</b>
G-Ref	16.24	16.05	<b>22.93</b>

As illustrated in Tab. IV, the zero-shot ability of our method significantly exceeds previous methods. This property benefits from the PEFT framework, which preserves the zero-shot ability of these encoders. In contrast, other methods set these

encoders trainable to get a great performance on specific datasets. However, the number of these specific datasets is insufficient for the convergence of encoders and thus destroys the zero-shot ability.

## V. CONCLUSION

In this paper, we propose a novel referring image segmentation method called HARIS, which introduces the Human-Like Attention mechanism based on the parameter-efficient fine-tuning framework. Specifically, we introduce an extra feedback branch from multi-modal features and feed it into the attention block as the second-round input. This design alleviates some irrelevant image-text pairs. Besides, we introduce the parameter-efficient fine-tuning framework into RIS, which preserves the zero-shot ability of the pre-trained encoders. Extensive experiments on three widely used benchmarks and PhraseCut demonstrate that HARIS achieves new state-of-the-art performance and great zero-shot ability.

## REFERENCES

- [1] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell, "Segmentation from natural language expressions," in *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 108–124.
- [2] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu, "Cris: Clip-driven referring image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11686–11695.
- [3] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang, "Vlt: Vision-language transformer and query generation for referring segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [4] Chang Liu, Henghui Ding, and Xudong Jiang, "Gres: Generalized referring expression segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23592–23601.
- [5] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1307–1315.
- [6] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha, "Learning to assemble neural module tree networks for visual grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4673–4682.
- [7] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak, "Restr: Convolution-free referring image segmentation using transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18145–18154.
- [8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12888–12900.
- [9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.
- [10] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiyu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al., "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10965–10975.
- [11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.
- [12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1290–1299.

- [13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [14] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.
- [15] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg, “Referitgame: Referring to objects in photographs of natural scenes,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 787–798.
- [16] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy, “Generation and comprehension of unambiguous object descriptions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 11–20.
- [17] 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 *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
- [18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al., “Segment anything,” *arXiv preprint arXiv:2304.02643*, 2023.
- [19] 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 *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [20] Li Xu, Mark He Huang, Xindi Shang, Zehuan Yuan, Ying Sun, and Jun Liu, “Meta compositional referring expression segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19478–19487.
- [21] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji, “Phrasecut: Language-based image segmentation in the wild,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10216–10225.