Neurocomputing 583 (2024) 127599

Contents lists available at ScienceDirect

Neurocomputing

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

Survey Paper

A survey of methods for addressing the challenges of referring image segmentation

Lixia Ji [a,b], Yunlong Du [a], Yiping Dang [a], Wenzhao Gao [d], Han Zhang [a,c,*]

[a] School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou, Henan 450002, China
[b] College of Computer Science, Sichuan University, Chengdu, Sichuan 610041, China
[c] Intelligent Policing Key Laboratory of Sichuan Province, Sichuan Police College, Luzhou, Sichuan 646000, China
[d] Intelligent Systems Research Centre, School of Computing, Engineering & Intelligent Systems, Ulster University, Magee campus, Northern Ireland, BT487JL, United Kingdom

ARTICLE INFO

Communicated by W. Wang

Keywords:
Referring image segmentation
Multi-modal learning
Computer vision
Natural language processing

ABSTRACT

Referring image segmentation is guided by natural language descriptions to separate the target objects in an image. This task is different from semantic segmentation and instance segmentation in that it involves unique challenges such as multimodal information fusion, variability of natural language expressions, and model robustness. In recent years, the emergence of deep learning techniques has led to innovative ideas and methods for solving these problems. We systematically analyze the main challenges of referring image segmentation and summarize the existing solutions. These include strategies such as multimodal fusion, expression query, multimodal pre-training, and robustness. In addition, we provide an overview of several datasets commonly used in referring image segmentation and analyze the performance of various representative approaches in comparison to different datasets, visual backbone models and threshold settings. Our focus also extends to the challenges and future developments in the field of referring image segmentation. Our survey paper will provide a comprehensive technical reference for future researchers.

1. Introduction

Referring Image Segmentation (RIS) is a multimodal task. This research intersects the domains of Computer Vision [1] and Natural Language Processing [2]. RIS presents a greater challenge than other multimodal tasks, such as Visual Question Answering [3] and Visual Dialog [4], as it necessitates effective coordination and reasoning between language and vision to accurately segment the target region within an image.

Closely related to referring image segmentation are the fields of referring expression comprehension (REC) [5] and video segmentation. All of these studies involve using natural language expressions to locate or segment target objects within an image or video. Qiao et al. [6] provide a comprehensive survey and comparison of existing methods addressing the problem of REC. Their review categorized these methods according to their mechanism for encoding the visual and textual modalities and highlighted the developmental trends and some unresolved issues in the field of REC by revealing the connections between new methods. In contrast, Zhou et al. [7] provide an extensive review of work related to video segmentation, covering all aspects from task definition to categorization, from algorithms to datasets, and from unresolved problems to future research directions. Their review primarily focuses on two main branches of video segmentation: Video object segmentation and Video semantic segmentation. Unlike video segmentation, RIS requires the simultaneous execution of both tasks involved in video segmentation. This is because RIS not only needs to segment the target objects within the image, but also ensure that the segmentation results align with the semantics of the expression description. This necessitates the model's full comprehension of the linguistic expression's meaning, the image content, and the correspondence between them. To understand the progress in the field of RIS, we searched on the Web of Science using the keyword "referring image segmentation". The search results, categorized by year and shown in Fig. 1, indicating a gradual increase in research activities in the field of RIS in recent years. Through an extensive review of the literature, we have obtained a comprehensive understanding of the current state and trends of research in this field. However, there are still significant challenges in improving the performance of image segmentation. The following are the current major challenges in referring image segmentation:

- Multimodal Fusion: Referring image segmentation necessitates the concurrent handling of both textual and visual modalities.

* Corresponding author at: School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou, Henan 450002, China.
E-mail address: zhang_han@zzu.edu.cn (H. Zhang).

https://doi.org/10.1016/j.neucom.2024.127599
Received 29 November 2023; Received in revised form 14 March 2024; Accepted 19 March 2024
Available online 24 March 2024

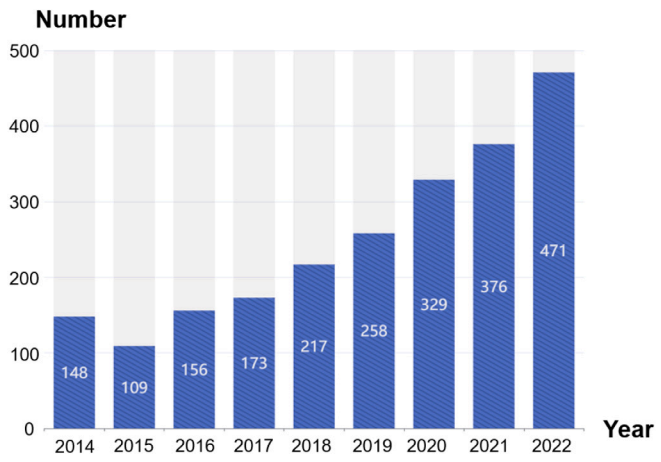


Fig. 1. Number of publications related to "referring image segmentation" retrieved from Web of Science during 2014–2022.

Visual information is high-dimensional and spatial, encapsulating the shape, color, location, and other attributes of the target object within the image. Conversely, linguistic information is low-dimensional and sequential, detailing the properties, relationships, and categories of the target objects within the image. Therefore, it is crucial to account for the disparities and correlations between these two modalities during fusion. The challenges lie in effectively mining the correlations between multi-scale image features [8] and text, and integrating the fused features of the modalities at various stages. This is essential to maintain the consistency of the multimodal data across different shapes of the target objects, thereby enhancing the effectiveness and accuracy of segmentation [9].

- Diversity of Expression: Despite the abundance of research on multimodal fusion in recent years, the enhancement in method performance is not significant. This is primarily due to previous methods not fully accounting for the fact that a single target can be described in various expressions, with these expressions potentially encompassing different attributes, relationships, locations, and other information. This necessitates an improvement in the model's adaptability, enabling it to cope with diverse referring expressions.
- Robustness: Currently, most studies operate under an idealized assumption that each expression corresponds to a unique target object within the image. However, in practical applications, there may be multiple targets that match the description or there may be no targets that match the description at all. Therefore, it is crucial to equip the model with the ability to handle these multimodal inconsistencies and provide reasonably reliable results.

These challenges represent the core difficulties in the field of referring image segmentation and are crucial directions for future development. However, review [10] focuses on encoding and decoding methods, without systematically summarizing and analyzing how to tackle these challenges. This lack of comprehensive analysis has motivated us to write this paper.

This review will focus on the methodologies addressing the aforementioned challenges, analyzing the diverse strategies and techniques employed by existing methods to tackle the core problem of referring image segmentation. We will also explore their respective strengths, weaknesses, and potential directions for future improvements. Through this work, we aim to provide a comprehensive and in-depth overview of the field, offering valuable references for researchers in this domain. Our major contributions include:

- In contrast to previous related reviews [6,7,10], we focus on the challenges in the task of referring image segmentation. Classifying different challenges according to their approaches to tackling them helps to understand the design principles and optimization goals of various methods more clearly.
- By conducting experiments on three benchmark datasets as well as a new dataset, gRefCOCO. The performance results of these methods are analyzed to mine the correlation between the methods, which provides valuable guidance for future research.
- Finally, we discuss unsolved problems, highlight model generalization and complexity issues and outline future scope.

2. Method

As shown in Fig. 2, we collect and organize the related literature on these challenges, categorized into the following areas: multimodal fusion, expression query, multimodal pre-training, and robustness.

2.1. Multimodal fusion

Multimodal fusion [11] is a pivotal technique in addressing the task of referring image segmentation. It effectively establishes the relationship between visual and linguistic modalities, thereby bridging the modality gap. To underscore the essence of multimodal fusion and the strengths of various methods, we categorize existing methods based on the information fusion modes into CNN-LSTM, Attention Mechanism, Multi-level Feature Fusion and Auto-regressive Vertex Generation. Subsequently, we analyze the key techniques and advantages of each method, as well as their performance in the task of referring image segmentation. This aids readers in better understanding and comparing the similarities and differences among these methods.

2.1.1. CNN-LSTM

Motivation Intuitively, to perform RIS, visual and linguistic features need be independently extracted from encoder networks. Subsequently, these features are fused together for prediction using a cross-modal decoder. Under the framework of CNN-LSTM, the Convolutional Neural Network (CNN) [12] processes the input image to generate rich image representations. The Long Short-Term Memory (LSTM) network [13] encodes the expression for feature extraction.

Methods Long et al. [14] propose a Fully Convolutional Network (FCN) that assigns each pixel in the image to a category. As shown in Fig. 3, a segmentation mask can be generated directly from the input image without the need for a region proposal. Hu et al. [15] introduce a method for segmentation from natural language expression, initially proposing the use of the CNN-LSTM framework to extract visual features from image and linguistic features from natural language expression, thereby addressing the task of referring image segmentation. Liu et al. [16] design a recursive multimodal interaction model capable of encoding sequential interactions between individual words, visual information, and spatial information, thereby capturing the complex relationship between text and image. RMI [16] uses a multimodal LSTM network for linguistic representation, which can encompass visual features and capture spatial variations of multimodal information, thereby generating coarse localization masks. These masks are then progressively refined by a unidirectional LSTM network. Li et al. [17] propose a Recurrent Refinement Network (RRN) to address the lack of multi-scale semantics in image representations. The RRN can efficiently utilize the inherent feature pyramid of the convolutional neural network to capture semantics at different scales. The RRN can match each word with each pixel in the image to generate an initial segmentation mask, which is then refined and iteratively optimized using a recursive optimization module, to obtain a high-quality pixel segmentation mask.

Performance and limitations We summarize the performance of all discussed methods in Table 2. CNN-LSTM is a simple and effective

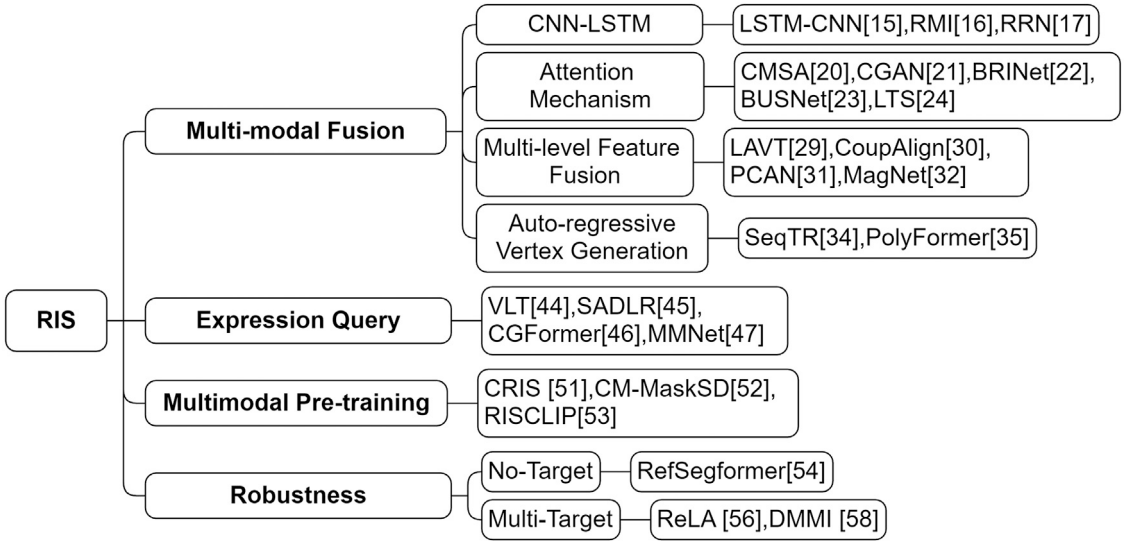


Fig. 2. Overview of existing methods of referring image segmentation.

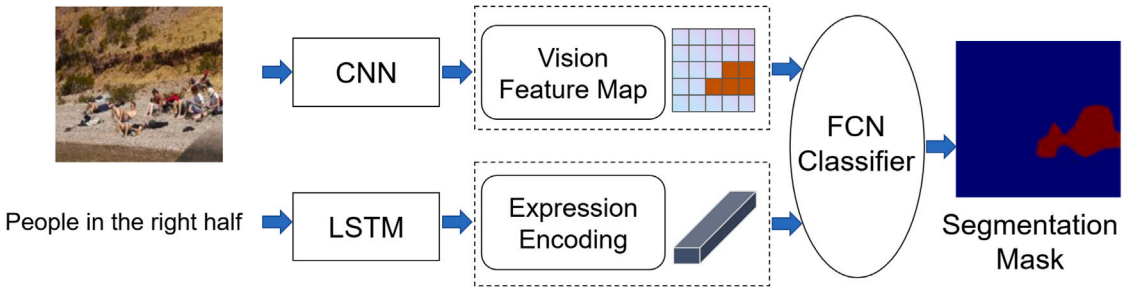


Fig. 3. The process of inference using the CNN-LSTM framework.

multimodal fusion method. However, it cannot capture complex linguistic and image features due to the limitation of the local receptive field of CNN. It can be seen that CNN-LSTM methods (e.g., LSTM-CNN [15], RMI [16], and RRN [17]) perform well when dealing with simple RefCOCO datasets, but their performance deteriorates when dealing with more complex RefCOCO+ and G-Ref datasets. The limitations of these methods are that they focus only on the vector representation of a single modality, ignore the modality gap, and do not adequately consider complex linguistic representations and interactions between images. These issues are addressed in later studies of other multimodal fusion methods and significantly improve the performance and generalization ability of the model.

2.1.2. *Attention mechanism*

Motivation Referring image segmentation involves the interaction and fusion of information from both visual and linguistic modalities. However, the CNN-LSTM framework is unable to capture complex linguistic and image features due to the limitation of the local receptive field of CNN. The attention mechanism [18,19] is an effective technique for multimodal information processing. It establishes elemental connections between visual and textual information, associating each pixel or region in the image with each word in the text. This allows the model to more accurately capture the correspondence between image and text, thereby realizing semantically rich visual and textual fusion representations, and enhancing the segmentation performance of the model. This section will introduce some methods for referring image segmentation that are based on the attention mechanism.

Methods Ye et al. [20] proposed a cross-modal self-attention (CMSA) module to efficiently capture long-distance dependencies between linguistic and visual features. The module can adaptively attend to keywords in the expression and key regions in the input image. By computing mutual attention on linguistic and visual features, it achieves global dependency modeling between linguistic and visual features. However, this simple inter-modal interaction lacks attention and differentiation between different targets. To address this, Luo et al. [21] proposed a cascade-grouped attention network (CGAN), which consists of cascade-grouped attention (CGA) and instance-level attention loss (ILA). This method can perform layer-by-layer inference on images, effectively distinguishing between different instances, recognizing objects matching text descriptions, and enhancing the correlation between text and images. The ILA loss is then embedded into each layer of the CGA to directly guide the attention-learning process, improving the consistency between the text and visual instances, enabling the model to better understand the description and map it to the correct visual instances.

Both [20,21] are language-guided unidirectional reasoning processes that ignore the problem of modality gaps. To address this issue, Hu et al. [22] proposed a bidirectional relationship inferring network (BRINet) to model cross-modal information dependencies. The network can utilize a visual-guided language attention module to learn the adaptive linguistic context corresponding to each visual region, thus effectively filtering out expression-irrelevant regions. The bidirectional cross-modal attention module (BCAM) is then co-constructed with the language-guided visual attention module to learn the relationship between multimodal features and enhance the semantic matching between the target object and the expression. Inspired by BRINet [22], Yang et al. [23] propose a novel bottom-up shift and bidirectional attention refinement module, which work together to accomplish visual inference and accurate segmentation in a single stage. The bottom-up process explicitly aligns the text with the visual area to identify

all entities mentioned in the expression. Subsequently, the bidirectional attention refinement module improves the segmentation results by integrating multi-level features through bidirectional attentional propagation to capture the visual details associated with the referenced objects.

The methods of BRINet [22] and BUSNet [23] focus on designing an implicit and recursive feature interaction mechanism to fuse visual-linguistic features without explicitly modeling the localization information of the referring instances. Jing et al. [24] propose a locate-then-segment (LTS) method. This method decomposes the task of referring image segmentation into two stages: first locate, then segment, rather than directly generating the final segmentation mask. The location stage uses a cross-modal interaction module to fuse linguistic and visual features to obtain cross-modal representation and combines location prior information to generate the detection box of the target object. A lightweight segmentation network is used in the segmentation stage to generate detailed segmentation masks based on the detection box and cross-modal features obtained in the location stage.

Performance and limitations The attention mechanism reduces computational resource consumption by focusing on the important parts of the input and can utilize global context information to enhance segmentation accuracy. Currently, methods based on the attention mechanism model the multilevel relationship between images and language, thereby improving the model's sensitivity to complex natural language expressions. By applying the attention mechanism on top of the CNN-LSTM framework, both CGAN [21] and BRINet [22] outperform traditional CNN-LSTM methods on the RefCOCO and RefCOCO+ datasets. Experimental results from BRINet further demonstrate the importance of enforcing attention across multiple modalities, and the need to consider the effects of both textual guidance and image guidance. Additionally, ablation studies with CGAN have revealed the effectiveness of overlaying multiple different attention modules. However, as datasets do not provide corresponding annotations, complex loss functions need to be designed to adjust the correct allocation of attention.

2.1.3. Multi-level feature fusion

Motivation Visual Transformer (ViT) [25] have been effectively applied in various visual tasks. However, ViT requires the computation of self-attention weights [26] across the entire image, which leads to high computational complexity. Furthermore, visual transformers only consider single-scale images, which restricts their adaptability to the segmentation of targets of varying sizes. The Swin Transformer [27] is a hierarchical visual transformer model. It accomplishes multi-level feature [28] extraction by partitioning the image into windows of varying sizes and applying self-attention mechanisms within each window. Multilevel feature fusion methods, based on the Swin Transformer, enable the fusion of image features and textual information across multiple scales. This allows the model to adapt to images of different resolutions, showing higher sensitivity when dealing with small targets and details. Additionally, it reduces the computational complexity of the model, enhancing its efficiency in scenarios that require high accuracy and real-time processing.

Methods Yang et al. [29] propose a language-aware visual transformer (LAVT) network. Using hierarchical levels of image feature structure, this network integrates visual and linguistic characteristics. The LAVT network employs a multi-level design within the Swin Transformer backbone, forming a hierarchical approach to the visual structure of language perception. Each stage consists of a Transformer encoding layer, a multimodal fusion unit, and a learnable gating unit. The multimodal fusion unit uses a pixel-word attention mechanism to assess the interplay between individual pixel locations and corresponding linguistic attributes. The gating unit can adaptively control the influence of linguistic features on visual features. LAVT provides a robust benchmark for future referring image segmentation tasks.

However, the multimodal fusion part of LAVT only performs pixel-word matching. To enrich the multimodal fusion information, Zhang et al. [30] propose a hierarchical visual-semantic alignment method called Coup Align. This method achieves more accurate localization and segmentation through a combination of sentence-mask alignment and word-pixel alignment and enforces object-mask constraints. During the fusion stage, word-pixel alignment modules are used to merge linguistic and visual features. During the decoder stage. The sentence-mask alignment module assigns weights to the masks using sentence embeddings to pinpoint the referenced object.

LAVT [29] and Coup Align [30] do not fully utilize the spatial information of the image to enhance the representation of visual features. To address this problem, Chen et al. [31] design a position-aware contrastive alignment network (PCAN). This network utilizes a location-aware module to extract the location information of all objects related to the expression as a priori knowledge. It also utilizes a contrastive language understanding module to effectively guide the interaction between vision and language to enhance the multimodal alignment effect by comparing the features of referring objects and neighboring objects.

PCAN [31] can enhance the spatial perception of the model through location encoding and location attention, but it still requires explicit alignment information for supervised learning, which may increase the cost of data annotation and the complexity of the model. How can the model itself achieve explicit alignment? Current approaches primarily focus on designing and improving multimodal alignment modules. Current methodologies lack direct supervisory signals for detailed visual contexts, resulting in a disconnect between the linguistic elements and their visual counterparts. The Mask Grounding [32] uses BERT [33] to process the referring expressions and randomly replaces some word tokens with special mask tokens. The fusion stage is comparable to the multi-level process of LAVT. It fuses image text features and uses segmentation masks as supervisory information to improve the alignment between linguistic and visual features.

Performance and limitations In referring image segmentation tasks, the Swin Transformer typically outperforms ViT as it can better handle targets and details at different scales, and more effectively match expressions with visual objects. Based on the experimental results in Table 2, we observe that most state-of-the-art methods opt for the Swin Transformer as the visual encoder, further demonstrating the advantages of multilevel feature fusion methods in handling complex visual tasks. As shown in Table 4, we can see that there is a significant difference in the performance of different methods under different thresholds on the RefCOCO validation set. When the Intersection over Union (IoU) threshold is 0.9, the performance of all models dramatically decreases. This is because the higher the threshold, the higher the fine-grained requirement for segmentation. The LAVT, by considering the multi-level feature information of the image, can better handle fine-grained visual information and still achieve the highest results. In addition, for conventional CNN-LSTM and attention mechanism methods, there are large differences in their results between different thresholds. For LAVT, however, the differences between each threshold are relatively small. However, as multimodal fusion methods continue to evolve, we see an increase in method design complexity and model size, making inter-modal fusion and target segmentation increasingly complex. This complexity will increase the computational complexity and detection time of the model. As shown in Table 5, multilevel visual feature methods achieve good performance but at the cost of having the highest computational complexity and longer runtimes.

2.1.4. Auto-regressive vertex generation

Motivation Although multilevel feature fusion results in smoother target segmentation edges, the shape and size of the segmented target may vary significantly, and there may be occlusions. This can lead to inaccurate or incomplete segmentation results. The decoders of LAVT [29] and Coup Align [30] can only classify the target pixels based

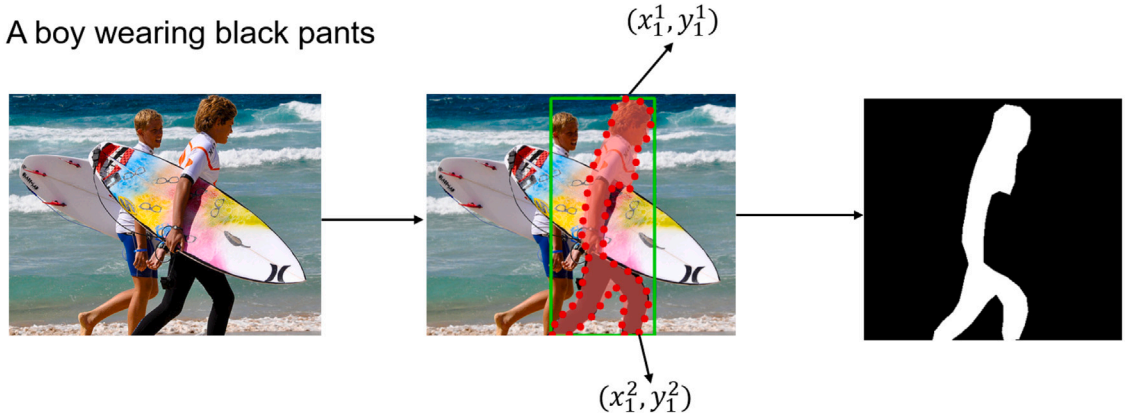


Fig. 4. The illustration of segmentation mask generation from polygon vertex sequences.

on the results of the encoder fusion, ignoring the sequence information in the linguistic query. However, the auto-regressive vertex generation approach enables the decoder to generate coordinate sequences. It can perform autoregressive vertex generation directly in coordinate space without relying on fixed-size feature maps or predefined anchor frames, thus significantly reducing coding redundancy and uncertainty. Furthermore, it can utilize sequence information from expression to guide the order and location of vertex generation, thereby improving the accuracy and completeness of segmentation.

Methods Zhu et al. [34] design a simple and universal network, SeqTR, which transforms the problem of referring image segmentation into a coordinate point prediction problem. Based on image and text inputs, SeqTR uniformly samples N points clockwise on the mask contour to generate a sequence of coordinates from a binary mask, and then quantizes these points into positive integers. This allows SeqTR to predict the coordinate markings of the mask using the standard Transformer architecture and the cross-entropy loss function, eliminating the need for an additional convolutional mask decoder.

However, SeqTR [34] can only generate a single rough segmentation polygon mask and is unable to efficiently extract the contours of objects with complex shapes. Liu et al. [35] propose a sequence-to-sequence approach, PolyFormer, as shown in Fig. 4. This method transforms the segmentation problem into a sequential polygon generation problem. It generates a sequence of polygon vertices in an autoregressive manner based on the input image blocks and textual query tokens. PolyFormer excels in extracting the contours of objects with complex shapes and occlusions. This is due to its novel polygon encoding method that represents polygon vertices in polar coordinates relative to the center of the image block, thus directly predicting precise floating-point coordinates. This approach eliminates the need for coordinate quantization and does not rely on fixed-size feature maps or predefined anchor frames, significantly reducing coding redundancy and uncertainty. Additionally, PolyFormer introduces a polygon ordering mechanism based on occlusion relationships, allowing the generated polygons to be automatically optimized in occlusion order. This mechanism enhances the quality and accuracy of contours.

Performance and limitations Auto-regressive Vertex generation employs a sequence-to-sequence framework that naturally integrates multimodal features into input sequences and outputs multitask predictions. This approach designs an innovative regression-based decoder capable of generating continuous 2D coordinates without the need for quantization errors. Notably, this sequence-to-sequence framework can handle referring expression comprehension and referring image segmentation tasks in a unified manner. Both SeqTR and PolyFormer methods demonstrate exceptional performance on both REC and RIS tasks. These methods enhance the flexibility and accuracy of segmentation mask generation. This provides a fresh perspective for understanding and addressing this problem, as well as new possibilities and directions for future research.

2.2. Expression query

Motivation The diversity of referring expressions presents a challenge for cross-modal fusion, especially when understanding complex expressions that involve relationships between multiple objects, or when dealing with rare or ambiguous expressions. Expression query methods optimize segmentation results by retrieving target regions from images based on different queries. As shown in Fig. 5, the input expression "People on the sofa looking at phone", although all the people in the image satisfy the condition "on the sofa", linguistic self-attention struggles to distinguish the importance between woman and man. Previous methods mainly focused on the relationship between objects, but ignored the complexity of referring expressions. To improve the consistency of queries, expression query methods generate multiple queries, such as "Man looking at phone" and "woman looking at camera", each corresponds to a unique interpretation of the referring expression. This can better distinguish the target, thereby improving the accuracy of target recognition.

Methods In previous work on Vision Transformer [36–38], the queries are typically a pre-defined set of vectors that have been trained, with each vector dedicated to predicting a specific target. These queries primarily focus on all objects in the image rather than the target demonstrated by the referring expression. However, in referring image segmentation tasks, the target can be any part of the image, and its attributes can vary greatly. Therefore, a fixed query vector may not accurately represent the attributes of the target. To address this issue, methods for referring image segmentation tasks should be capable of adaptively generating query vectors based on given images and referring expressions. For referring expressions, words vary in importance. Several research works have tackled this problem by quantifying the significance of each word. For instance, Ding and Luo et al. [39,40] assign weights to each word in an expression, while [41,42] define a set of clusters, such as location, attribute, and entity.

Inspired by [43], Ding et al. [44] propose the vision-language transformer (VLT). To enhance the model's diverse understanding of expressions, VLT dynamically generates multiple sets of queries, thereby enhancing the model's understanding of expressions and facilitating effective interaction between visual and linguistic information. Additionally, VLT's query balancing module adaptively selects the most appropriate features for fusion, thereby improving the accuracy of visual localization. VLT flattens the channel dimensions of visual features to generate a unidirectional inference process with attention weights between text and vision. Yang et al. [45] iteratively refine multimodal features through a semantics-aware dynamic convolution module. This module can generate a kernel based on the input feature vectors and use it to convolve the multimodal features. The feature vectors generated by these kernels, referred to as "queries", can be seen as concepts highlighted in the multimodal feature map. In each iteration,

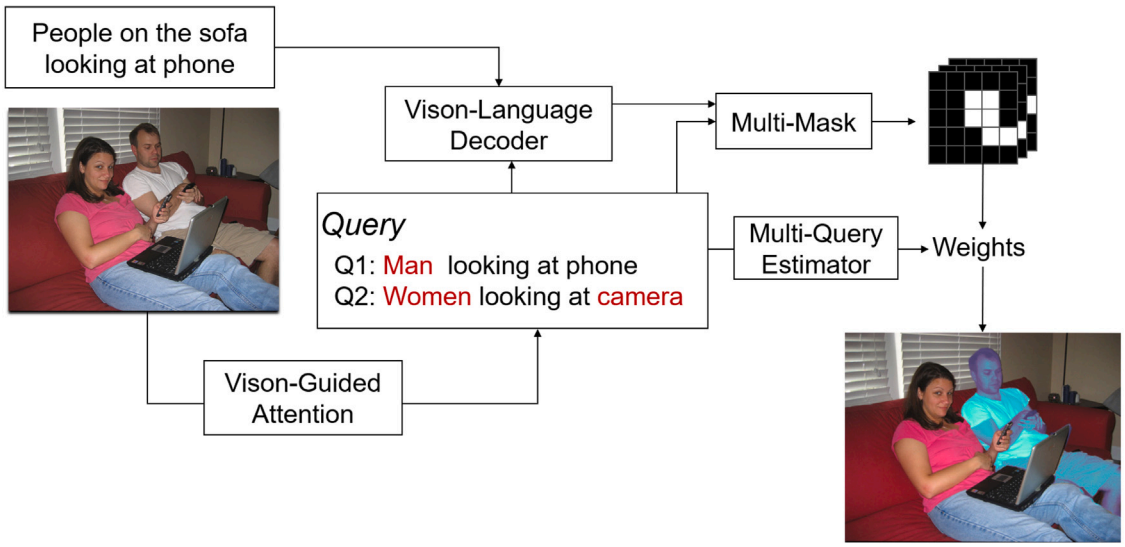


Fig. 5. The expression query method matches a target by generating multiple queries.

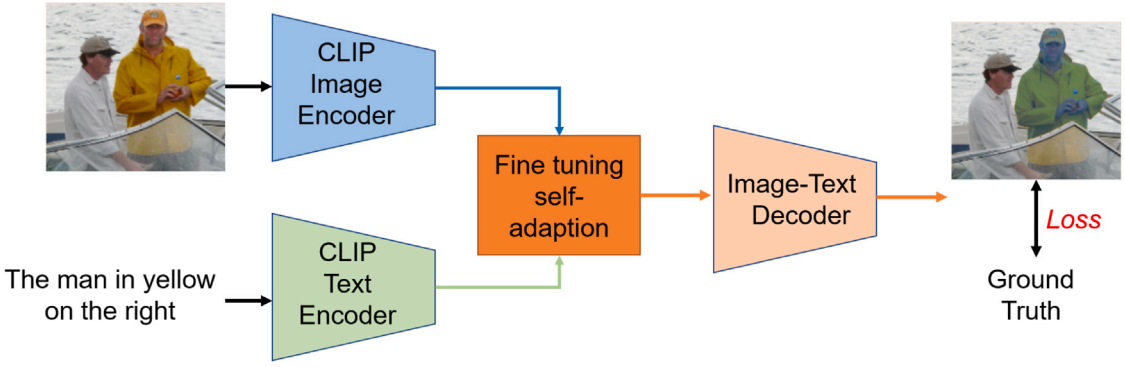


Fig. 6. The segmentation process of referring image segmentation method based on CLIP.

the module updates the queries based on the segmentation results of the previous iteration, thereby enhancing the multimodal features related to the target while suppressing features unrelated to the target.

To enhance the semantic information of the query, Tang et al. [46] design a contrastive group with a transformer network (CGFormer). This network introduces learnable query sequences to represent object-level information. It then updates the query sequences with linguistic features and categorizes visual features, associating the visual features with the corresponding query sequences for cross-modal inference with object perception.

Yan et al. [47] propose a multi-mask network(MMNet). Initially, it fuses global textual and visual features and generates multiple queries representing different aspects of referring expressions. These queries and global multimodal features are input into a visual-language decoder to obtain decoded features. These features, along with the queries, are then input into a multi-mask projector to generate corresponding segmentation masks. Finally, a weighted sum is computed based on these masks and their weights to produce the final segmentation result. This approach reduces the annotation cost and model complexity by automatically generating masks for supervised learning.

Performance and limitations While the multilevel visual feature method considers the multilevel visual information of an image, it overlooks the diversity of referring expressions. This could lead to an inability to accurately determine the target indicated by the expression, even if we can detect all objects in the image. To address this issue, VLT [44] successfully applies learnable queries to the RIS task, which we refer to as the expression query method. Notably, the performance of methods such as SADLR [45], CGFormer [46], and MMNet [47] on benchmark datasets is comparable to that of the multilevel feature fusion methods (e.g., LAVT and PACN), indicating that the expression query method can effectively handle complex referring expressions. As shown in Table 5, compared to the multilevel feature fusion method (LAVT), the expression query method (VLT), as a lighter and more efficient approach, can handle complex referring expressions while also reducing the computational complexity of the model.

As shown in Table 6, the number of queries significantly affects the model performance. We used the slightly more challenging dataset RefCOCO+ for the ablation study. VLT uses ResNet50 as the visual backbone, while MMNet uses ViT-L. When the quantity of queries escalates from 1 to 16, the performance of the VIT method improves by about 5%. Similarly, the MMNet method improves its performance by about 3% when the quantity of queries escalates from 1 to 16. This suggests that multiple queries indicative of diverse aspects of information, aiding model in obtaining robust multimodal features. In general, the model performance improves with the increase in the number of queries, but this escalation does not consistently lead to superior results. As the query count rises, the performance gradually stabilizes and may even decrease. Too many queries may interfere with the selection of the final result.

2.3. Multimodal pre-training

Motivation Referring image segmentation requires effective fusion of visual and linguistic cues to match and segment a target object in an

image. However, the model's ability to generalize can be hindered by the limited training data and the lack of a unified learning framework. To address these issues, multimodal pre-training models (as shown in Fig. 6) can be constructed to enhance the interaction between image and language features, thereby improving segmentation performance. These models, trained on large datasets to learn shared representations of visual and linguistic elements, providing a unified knowledge representation for downstream multimodal tasks. The multimodal pre-training approach is a promising research direction as it can effectively handle sparse and unbalanced data and improve the generalization ability of the models.

Methods In recent years, the development of jointly pretrained visual and linguistic models [48,49] has provided unified knowledge representations for downstream tasks. CLIP [50] is a model for learning multimodal representations from matching relations between images and language descriptions. With its powerful multimodal information and migration capabilities, CLIP has achieved state-of-the-art performance in several multimodal tasks.

Several methods for referring image segmentation based on CLIP have been proposed, each leveraging and optimizing the multimodal features of CLIP from various angles. The CRIS [51] method employs a CLIP pre-training model to extract and fuse image and language features, enhancing the consistency between these two modalities. It propagates text features to visual features at each pixel level, thereby fostering modality consistency. This method uses a text-pixel contrast learning strategy, which enforces text features to resemble relevant pixel-level features and differ from irrelevant ones, thereby adaptively generating segmentation masks. To achieve efficient fine-grained feature alignment, Wang et al. [52] propose a method called cross-modality masked self-distillation (CM-MaskSD) for referring image segmentation. This method introduces two symmetrical masked self-distillation branches based on multi-modal segmentation, aiming for a more refined alignment of visual and textual features. During the training process, visual and textual dual-guided masked self-diffusion is jointly utilized to bridge the modality gap, achieving fine-grained feature alignment between modalities.

However, both approaches optimize cross-modal feature fusion only with the rich multimodal pre-training knowledge of CLIP, but this knowledge is not well adapted to the tasks of RIS. To tackle this issue, Kim et al. [53] proposed a novel framework for adapting frozen CLIP feature residuals to RIS via a fusion adapter and a backbone adapter. The fusion adapter is a multimodal adaptive mechanism that enhances the interaction and alignment between image and language features. The backbone adapter is a unimodal self-attention mechanism that can infuse new knowledge beneficial for referring image segmentation, including spatial and edge information. Both adapters are used to fine-tune the adaptation of RIS, without disrupting the pre-training knowledge of CLIP.

Performance and limitations As shown in Table 2, the methods CRIS [51], CM-MaskSD [52], and RISCLIP [53] utilize large-scale multimodal pre-training models for feature extraction and processing, all achieving performance exceeding 70%. This demonstrates the advantages of multimodal pre-training models in handling complex multimodal information. Multimodal pre-training encoder CLIP-B exhibits excellent performance compared to current advanced visual encoders. Notably, RISCLIP [53] successfully adapts the rich visual information from the pre-training visual encoder to the RIS task, achieving the best results. This proves that multimodal pre-training visual encoders can enhance the model's generalizability and adaptability. However, while methods based on pre-trained models have clear advantages, they require fine-tuning for each dataset, which increases the model complexity and computational cost.

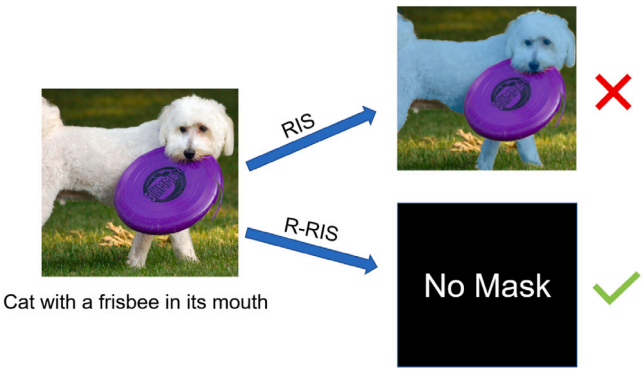


Fig. 7. Correct output of the robustness model.

2.4. Robustness

Motivation Most existing referring image segmentation methods are rely on the perfect presumption that there is always a target in the image that matches the description. However, in reality, the description may not consistently correspond with the visual content. In such cases, the robust model can provide accurate feedback, rather than erroneously segmenting irrelevant objects. Therefore, we expect referring image segmentation methods to make correct judgments when dealing with situations with no targets or multiple targets. In the following, we will explore and evaluate some robust referring image segmentation methods, which will contribute to the advancement of this field in practical applications.

2.4.1. No-target

Methods As shown in Fig. 7, the expression "Cat with a frisbee in its mouth" does not correspond to any target in the image and should be judged as "no mask". However, previous methods incorrectly segmented the dog because they were trained based on the assumption of matching expressions with targets, without considering the mismatch situations that might exist in actual applications. To solve this problem, Wu et al. [54] propose a Transformer-based model that uses a multi-head cross-attention mechanism [55] and a token-based visual-language fusion module. By associating image and text features through learnable tokens, it can be easily extended to robust referring image segmentation tasks (R-RIS). They constructed the R-RIS dataset, adding misleading negative expressions, meaning the object described does not exist in the image. At the same time, they proposed new evaluation criteria to evaluate the performance of the model in RIS and no-target situations, comprehensively assessing the robustness of the model.

2.4.2. Multi-target

Methods In response to the issue of multiple targets in an image, Liu et al. [56] propose generalized referring expression segmentation (GRES) to handle multi-target scenarios. They extended the segmentation of a single image to a related set of images to accommodate complex environments. GRES can handle any number of targets in an expression. They constructed a dataset named gRefCOCO to cover scenarios with multiple targets, no target and a single target. They proposed a baseline model called ReLA that adaptively segments images into regions containing sub-instance cues and models dependencies between regions and between regions and language. ReLA consists of a mask extractor, a region encoder and a region selector. The mask extractor uses R-CNN [57] to retrieve masks and features for each instance in the image. The region encoder uses a self-attention mechanism to contextually encode each region and computes a similarity score between each region and the expression. The region selector uses a variable clustering layer to cluster regions into different categories

Table 1
The key features of the benchmark datasets.

Dataset	Number of images	Number of expressions	Number of objects	Object categories
ReferItGame	19,894	130,525	96,654	238
RefCOCO	19,994	142,209	50,000	80
RefCOCO+	19,992	141,564	49,856	80
G-Ref	26,711	104,560	54,822	80

Table 2
Comparison with existing image segmentation methods on RefCOCO, RefCOCO + and G - Ref datasets.

Method	Visual encoder	Categories	RefCOCO(easy)			RefCOCO+(medium)			G-Ref (hard)	
			val	testA	testB	val	testA	testB	val(U)	test(U)
Hu [15]	ResNet-101	CNN-LSTM	–	–	–	–	–	–	28.14	–
RMI [16]	ResNet-101	CNN-LSTM	45.18	45.69	45.57	30.48	25.64	29.50	34.52	–
RRN [17]	ResNet-101	CNN-LSTM	55.33	57.26	53.95	39.75	42.15	36.11	–	–
CMSA [18]	ResNet-101	Attention mechanism	58.32	60.61	55.09	43.76	47.60	37.89	39.98	–
BRINet [22]	ResNet-101	Attention mechanism	61.35	63.37	59.57	48.57	52.87	42.12	–	–
BUSNet [23]	ResNet-101	Attention mechanism	63.27	66.41	61.39	51.76	56.87	44.13	–	–
CGAN [21]	DarkNet-53	Attention mechanism	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69
LTS [24]	DarkNet-53	Attention mechanism	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
LAVT [29]	Swin-Base	Multi-level visual features	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09
PCAN [31]	Swin-Base	Multi-level visual features	73.71	76.26	70.47	64.01	70.01	54.81	64.43	65.68
CoupAlign [30]	Swin-Base	Multi-level visual features	74.70	77.76	70.58	62.92	68.34	56.69	62.84	62.22
MagNet [32]	Swin-Base	Multi-level visual features	75.24	78.24	71.05	66.16	71.32	58.14	65.36	66.03
SeqTR [34]	DarkNet-53	Auto-regressive vertex generation	71.70	73.31	69.82	63.04	66.73	58.97	64.69	65.74
Poly Former [35]	Swin-Base	Auto-regressive vertex generation	74.82	76.64	71.06	67.64	72.89	59.33	67.76	69.05
VLT [44]	DarkNet-53	Expression query	67.52	70.47	65.24	56.30	60.98	50.08	54.96	57.73
SADLR [45]	Swin-Base	Expression query	74.24	76.25	70.06	64.28	69.09	55.19	63.60	63.56
CGFormer [46]	Swin-Base	Expression query	74.75	77.30	70.64	64.54	71.00	57.14	64.68	65.09
MMNet [47]	ViT-L	Expression query	75.01	77.81	71.59	68.44	72.81	59.86	66.52	67.28
CRIS [51]	CLIP-L	Multi-modal pretraining	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36
CM-MaskSD [52]	CLIP-L	Multi-modal pretraining	74.89	77.54	71.28	67.47	71.80	59.91	66.53	66.63
RISCLIP [53]	CLIP-L	Multi-modal pretraining	76.92	80.99	73.04	69.33	74.56	61.87	69.20	70.19
RefSegformer [54]	Swin-Base	Robustness(No-target)	73.22	75.64	70.09	63.50	68.69	55.44	62.56	63.07
ReLA [56]	Swin-Base	Robustness(Multi-target)	73.82	76.48	70.18	66.04	71.02	57.65	65.00	65.97
DMMI [58]	Swin-Base	Robustness(Multi-target)	74.13	77.13	70.16	63.98	69.73	57.03	63.46	64.19

based on the similarity score and outputs the segmentation probability for each category.

Similarly, Hu et al. [58] design a dual multimodal interaction network (DMMI), which enables bidirectional information flow interaction between text and image through two decoder branches. This allows the visual features to reflect the key semantic details regarding the target entity, thereby providing the text-to-image decoder with more accurate segmentation support.

Performance and limitations The methods RefSegformer [54], ReLA [56], and DMMI [58] take into account scenarios with no-target and multi-targets, demonstrating excellent performance on the benchmark datasets. As shown in Table 3, significant findings were made on the gRefCOCO dataset. While baseline models such as LTS [24], VLT [44] and CRIS [51] perform well on standard RIS tasks, their performance needs improvement when dealing with complex scenarios involving multi-target recognition and no-target. The oIoU evaluation metric fails to distinguish between scenarios with and without targets, whereas gIoU more accurately reflects model performance. Notably, the ReLA [56] model outperforms other models in all tests, demonstrating robustness in handling multi-target and no-target scenarios. However, despite the ReLA model surpassing other baseline models in performance, its accuracy still needs improvement. Currently, the performance of the ReLA model has not reached a level suitable for practical applications. Future RIS methods need to consider how to enhance model robustness to achieve better performance.

3. Datasets and evaluation

In this section, we review the benchmark datasets and the new dataset. The key features of the benchmark datasets are summarized in Table 1. Under the unified evaluation metrics, we conduct performance experiments on datasets.

3.1. Benchmark datasets

ReferItGame [59]: In 2014, Kazemzadeh et al. construct the first large-scale referring expression dataset, ReferItGame, containing 130,525 referring expressions in real-life scenes. The dataset covers 96,654 target objects in 19,894 images. However, images in this dataset often contain only a single object of the given category, leading to simpler descriptions that emphasize contextual information while overlooking object details.

RefCOCO and RefCOCO+ [60]: These datasets, similar to ReferItGame, were collected using a similar method and further expanded on the MSCOCO images [61]. The main difference between RefCOCO and RefCOCO+ is that in RefCOCO+, the use of location words is prohibited, emphasizing descriptions of the appearance features of the target objects. The RefCOCO dataset includes 50,000 objects and 142,209 referring expressions in 19,994 images, while the RefCOCO+ dataset includes 49,856 objects and 141,564 referring expressions in 19,992 images. Both datasets are divided into training, validation, test set A, and test set B. Test set A contains multiple target categories, while test set B contains multiple instances of the same target category.

G-Ref [62,63]: This dataset features longer and more complex expressions, often entire expressions rather than phrases. It includes 104,560 expressions for 54,822 objects in 26,711 images. Each object has an average of 1.91 expressions, and each image has an average of 3.91 expressions. The average length of expressions in RefCOCO is 3.61 words, RefCOCO+ is 3.53 words, and RefCOCOg is 8.43 words.

3.2. Robustness datasets

While benchmark datasets excel in multiple aspects, there is still scope for improvement when dealing with complex real-world scenarios. These datasets often lack detailed annotations of the model's inference process, so performance evaluation primarily relies on the

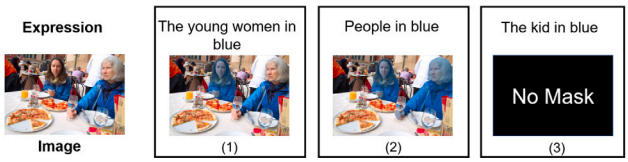


Fig. 8. Robust scene dataset.

Table 3
Performance of baseline models on the gRefCOCO dataset.

Model	val		testA		testB	
	oIoU	oIoU	oIoU	gIoU	oIoU	gIoU
LTS [24]	52.30	52.70	61.87	62.64	49.96	50.42
VLT [44]	52.51	52.00	53.95	63.20	50.52	50.88
CRIS [51]	55.34	56.27	63.82	63.42	51.04	51.79
LAVT [29]	57.64	58.40	65.32	65.90	55.04	55.83
ReLA [56]	62.42	63.60	69.26	70.03	59.88	61.02

final prediction results. Moreover, enhancing the robustness of models using benchmark datasets presents certain challenges. Below are some designs of robust datasets:

Robust RefCOCO dataset [54]: This dataset extends RefCOCO and introduces negative text inputs to test the model's ability to recognize objects not present in the image. Negative text inputs are generated in two ways: using words associated with categories that are not present in the image or using words that do not match the categories present in the image. The dataset consists of 19,994 images and 183,773 expressions, of which 41,563 are negative text inputs.

gRefCOCO [56]: This dataset aims to enhance the robustness of models when dealing with multi-target, no-target, and single-target scenarios. It extends the type and number of referring expressions in the RefCOCO dataset, enabling it to describe any number and type of target objects in an image. As shown in Fig. 8, this dataset not only covers single-target expressions but also includes a variety of expressions such as multi-target and no-target, significantly different from benchmark datasets. The gRefCOCO dataset is closer to real-world applications and poses greater challenges to the model.

3.3. Evaluation

An evaluation metric is a measure used to quantify the consistency between model outputs and actual outputs. In the task of referring image segmentation, it is common to compare the similarity between the segmentation mask produced by the model and the true mask. This comparison is typically done using a metric called Intersection over Union (IoU). As shown in Eq. (1), G and P denote the real bounding box and the predicted bounding box respectively, $I_G \cap I_P$ is the area of the intersection region of the real and predicted boxes, and $I_G \cup I_P$ is the area of the concatenation region of the real and predicted bounding boxes and the ratio of the two is the IoU. The prediction result of the IoU is larger than the set threshold value (generally set to 0.5), the prediction will be regarded as valid.

$$IoU_{G,P} = \frac{I_G \cap I_P}{I_G \cup I_P} \tag{1}$$

Settings and Metrics: We use PyTorch [64] to implement these methods. There are various visual encoders in the methods, including ResNet-101 [65], DarkNet-53 [66], Swin-Base, ViT-L, and CLIP-L. Among them, ResNet-101 and ResNet-53 and DarkNet-53 are convolutional neural network (CNN)-based visual encoders, which were pre-trained on PASCAL VOC [67] and MSCOCO training sets. Swin-Base, ViT-L and CLIP-L are Transformer-based visual encoders. We use two common metrics to evaluate the performance of different methods:

Table 4
Performance of some methods on the RefCOCO validation set for different value thresholds.

Model	Categories	prec@0.5	prec@0.7	prec@0.9	oIoU
RMI [16]	CNN-LSTM	42.99	22.75	2.23	45.18
RRN [17]	CNN-LSTM	47.59	26.53	3.17	46.95
CMSA [20]	Attention mechanism	66.44	50.77	10.96	58.32
LTS [24]	Attention mechanism	75.16	60.74	14.41	65.43
LAVT [29]	Multi-level visual features	84.46	75.28	34.30	72.73
PCAN [31]	Multi-level visual features	82.53	73.87	33.37	69.51

Table 5
FLOPs and Runtimes on the RefCOCO validation set for three representative methods.

Model	Categories	FLOPs	Times
LTS [24]	Attention mechanism	133.3G	41.0 ms
VLT [44]	Expression query	142.6G	41.9 ms
LAVT [29]	Multi-level features fusion	197.4G	45.6 ms

- Overall Intersection over Union (oIoU): This is the proportion of the intersection region to the union region for all test samples. Each sample contains a set of text expressions and an image. This metric is used to measure the accuracy of target localization.
- Precision: For each category, the precision is calculated as the ratio of the count of correctly predicted pixels for that category to the total number of pixels predicted for that category. We examine precision at different thresholds. Specifically, we measure the percentage of test samples where the IoU score is above a specific threshold $X \in \{0.5, 0.7, 0.9\}$.

It is worth noting that cIoU favors larger objects [68,69]. These metrics help us gain a deeper comprehension of the effectiveness of different methods in target detection and localization tasks. gIoU: The Generalized Intersection over Union (gIoU) is a metric proposed in [56] to address the issue of models being more likely to achieve higher oIoU scores for multi-target samples with larger foreground areas. Similar to the mean IoU, gIoU treats objects of all sizes equally. Moreover, gIoU incorporates a special treatment for samples without targets. For no-target samples, the IoU values of true positive no-target samples are regarded as 1, while IoU values of false negative samples are treated as 0. This ensures the completeness and consistency of computation.

As shown in Table 2, we evaluated the corresponding methods on the benchmark datasets (RefCOCO, RefCOCO+ and G-Ref). Here, due to page limitations, we only present the results for oIoU.

The three datasets represent three layers of complexity: easy, medium, and hard. Among them, RefCOCO is the simplest dataset with shorter expressions, larger target areas and less background interference. RefCOCO+ is of medium difficulty with longer expressions, smaller target areas, and more background interference. G-Ref is the most challenging dataset with expressions involving more complex semantics, finer granularity in the target areas, and severe background interference.

As shown in Table 4, it is evident that multilevel visual fusion methods have significant advantages. This insight serves as an important reference for future research. Specifically, when designing new models and methods, the incorporation of multi-scale feature information should be considered to enhance the model performance. Additionally, this suggests that when evaluating model performance, different thresholds should be taken into account to provide a comprehensive assessment of the model performance.

As shown in Table 6, the number of queries setting directly affects the performance of the model. In practice, we may need to conduct several experiments to find the optimal number of queries. At the same time, these results also reveal the importance of the query generation module, which is able to generate multiple queries representing different information, thus improving the performance of the model. In

Table 6
Performance gain by different Query Number on the test set of RefCOCO+.

Model	N_q	prec@0.5	prec@0.7	prec@0.9	oIoU
VLT [44]	1	50.17	34.75	4.66	44.83
	2	52.85	39.66	8.30	47.07
	4	53.06	40.38	8.92	46.79
	8	55.57	44.24	12.62	49.04
	16	55.84	41.68	10.76	49.36
	32	55.57	44.43	12.50	49.27
MMNet [47]	1	76.62	66.26	13.84	65.42
	2	77.53	68.13	14.39	65.80
	4	78.08	68.54	14.29	66.13
	8	78.69	68.05	14.83	66.85
	16	79.86	69.03	14.90	67.59
	32	79.45	68.71	14.91	67.26

future research, we can further explore the design and optimization of the query generation module with the aim of achieving better results on the referring image segmentation.

4. Discussion and future scope

In this section, we will summarize the commonalities among the various methods listed, discuss the shortcomings of current referring image segmentation methods and provide direction for the future development of this field.

4.1. Summary and analysis of existing methods

Multimodal fusion methods are central to accomplishing multimodal tasks. Within the domain of referring image segmentation, fusion strategies typically involve capturing visual and linguistic features from encoder networks then making predictions through a cross-modal decoder. Whether it is CNN-LSTM, attention mechanisms, multilevel feature fusion, or auto-regressive vertex generation methods, they all strive to effectively extract and fuse features of visual and linguistic modalities. Multilevel feature fusion methods co-embed multilayer visual and linguistic characteristics in the encoding process, fully leveraging the potential of Transformers to handle complex visual information and obtain rich multimodal features. However, this method significantly increases computational demands due to the consideration of multilevel visual feature information. Auto-regressive vertex generation methods aim to reduce coding redundancy, making the model more lightweight by transforming the segmentation problem into a problem of sequentially generating polygon vertices.

Multimodal tasks are still in their infancy when it comes to processing complex natural languages, and expression query methods aid models in better handling complex and diverse expressions, promoting improved segmentation using matched fusion information. However, bridging the gap between modalities is challenging when relying solely on a single-module approach, thus necessitating a combination of powerful multimodal fusion processing methods for enhanced results. This limitation also suggests that modules addressing different modalities can be stacked upon each other to bolster the performance of the model.

The multimodal pre-training approach leverages the joint representation between visual concepts and linguistic semantics, and through efficient fine-tuning techniques, it transfers the multimodal fusion information from the multimodal pre-trained model to the referring image segmentation. However, we have noticed that there is still limited research on referring image segmentation methods that improve model generalization, such as those for few-shot and zero-shot scenarios. This is an important direction for our future research.

The robustness approach enhances the multimodal fusion feature by adding more constraints, thereby improving the model's capability to discern the mismatch problem between the image and the referring expression. Although current methods may still have limitations in addressing the issue of image-text mismatch, there is a need to enhance the model's robustness.However, we observe that there are still constraints in the development of robustness today, primarily because the training datasets for research on these types of methods are limited. Consequently, we propose that future research methodologies should incorporate more datasets to enhance the model's robustness further.

4.2. Model generalization

In the field of referring image segmentation, we currently encounter an important problem, the majority of existing datasets mainly contain relatively simple attribute and relationship descriptions. However, existing models may struggle to handle new combinations of concepts, which poses a challenge for model generalization.

To tackle this issue, meta-learning approaches [70] propose the construction of synthetic training groups and a variety of synthetic testing groups to deal with different extents of fresh combinations. Furthermore, few-shot approaches [71,72] learn segmentation masks directly from image-level referring expressions, eliminating the need for pixel-level annotations. Additionally, some methods [73] achieve zero-shot referring image segmentation by making use of the cross-modal understanding from CLIP. However, the lack of fine-grained annotations in datasets makes it challenging for models to learn sufficient information for accurate predictions. Therefore, enhancing the generalizability of models is a crucial issue that needs addressing. Future research should explore the effective utilization of unlabeled data to improve the generalization performance of models.

4.3. Model complexity

In terms of model design, to enhance performance, existing multimodal fusion methods employ more complex attention mechanisms and compute multi-scale image information. However, these methods face challenges due to significant computational complexity and memory usage. To resolve this matter, future work should explore new algorithms and architectures to optimize the computational efficiency of Transformers. This could include developing more efficient self-attention mechanisms and innovative model compression techniques to reduce the demand on computational resources while maintaining model performance.

In the field of video segmentation, Ding et al. [74] extend the Transformer architecture by adding limited memory, enabling efficient querying of an entire video using referring expressions. This work designs memory that can persistently store global video content, dynamically collect local temporal context and segmentation history. The model can fully and flexibly understand the expression generates an adaptive query through the local–global context and the specific content of each frame. This vector is utilized to query the matching frame to generate a mask. In addition, the memory is processed with linear time complexity and constant size memory for video, avoiding expensive computational costs. Xu et al. [75] investigate the problem of efficient tuning and propose a adapter. This adapter injects information on specific tasks into pre-trained models to facilitate cross-modal knowledge interaction. They developed the lightweight decoder which achieves comparable or superior performance to challenging benchmark tests while only updating 1.61% to 3.38% of the backbone parameters. These studies provide solution ideas to reduce the complexity and computational cost of models while maintaining performance.

4.4. Future scope

The future development of referring image segmentation will involve advancements in modularization of method categories and enrichment of datasets. These factors will collectively drive the performance enhancement of RIS, enabling it to better handle complex real-world problems.

Table 7
Performance comparison of different methods interactions on RefCOCO validation set.

Model	prec@0.5	prec@0.7	prec@0.9	oIoU
VLT [44]	83.80	74.10	25.60	70.31
VLT [44]+CGAN [21]	80.33	69.94	22.80	66.41
VLT [44]+SADLR [45]	85.00	76.17	27.81	71.64
LAVT [29]	85.49	76.44	35.16	73.12
LAVT [29]+CGAN [21]	85.14	75.98	34.35	72.56
LAVT [29]+SADLR [45]	86.90	78.76	37.36	74.24

4.4.1. Modular approach design

In terms of model design, we could potentially achieve superior performance by simultaneously combining the multilevel visual information of multi-scale images with expressive query methods, generating multiple queries based on the fused information of images and text. As shown in Table 7, we observe a comparison of the performance of three interacting approaches: expressive query, attention mechanism, and multilevel feature fusion. Notably, representative expressive query and multilevel feature fusion methods, such as VLT [44] and LAVT [29], demonstrate significant performance improvements when combined with CGAN [21] and SADLR [45]. These results reveal an important phenomenon: different methods are not independent but can interact to produce a cumulative effect, thereby enhancing the model performance. For instance, LAVT can achieve further performance improvement by integrating the fine-grained attention design of CGAN with the semantic query of SADLR. This provides us with a crucial insight: when designing models, we should consider integrating different methods to achieve better results. This fusion and interaction of methods open up new possibilities and directions for the future development of referring image segmentation.

4.4.2. Abundant datasets

Current benchmark datasets range from RefCOCO to RefCOCO+ to G-Ref. These datasets serve as crucial tools for testing and enhancing the performance of referring image segmentation. The gRefCOCO dataset, which addresses arbitrary targets, is an extension of RefCOCO. Therefore, future datasets can build upon existing benchmark datasets, incorporating improvements to enhance performance based on specific challenges that need addressing. We anticipate the emergence of more datasets in the future to cater to diverse needs and challenges, providing a broader scope and more possibilities for the development of referring image segmentation.

5. Conclusion

This paper provides a comprehensive and in-depth examination of the field of referring image segmentation, systematically categorizing and analyzing the multiple challenges currently faced. We delve into existing solutions such as multimodal fusion, expression query, multimodal pre-training, and robustness methods, detailing the representative models and their performance in each category. Through experiments and thorough analysis of widely used datasets, we further explore the effectiveness of these methods and compare the performance of different methods under various settings, providing a comprehensive evaluation of the results. Despite the progress made in referring image segmentation, our findings reveal that there are still some unresolved issues. Considering the generalizability and complexity of the model, we contemplate the future scope of this field.

CRediT authorship contribution statement

Lixia Ji: Writing – review & editing, Conceptualization. **Yunlong Du:** Writing – review & editing, Writing – original draft, Investigation, Conceptualization. **Yiping Dang:** Writing – original draft, Resources, Investigation. **Wenzhao Gao:** Data curation. **Han Zhang:** Writing – review & editing, Supervision, Funding acquisition, Formal analysis.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the Henan Provincial Major Science and Technology Special Project, China [grant numbers 231100210200]; Intelligent Policing Key Laboratory of Sichuan Province, China [grant numbers ZNJW2024KFQN005]; the Henan Provincial Key Scientific Research Project, China [grant numbers No. 24A520047]. Henan Provincial Key R&D and Promotion Special Program (Science and Technology Tackling), China [grant numbers 232102210128].

References

[1] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, et al., Deep learning for computer vision: A brief review, Comput. Intell. Neurosci. 2018 (2018).
[2] K. Chowdhary, K. Chowdhary, Natural language processing, in: Fundamentals of Artificial Intelligence, Springer, 2020, pp. 603–649.
[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, Vqa: Visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433.
[4] S. Kottur, J.M. Moura, D. Parikh, D. Batra, M. Rohrbach, Visual coreference resolution in visual dialog using neural module networks, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 153–169.
[5] Z. Chen, P. Wang, L. Ma, K.-Y.K. Wong, Q. Wu, Cops-ref: A new dataset and task on compositional referring expression comprehension, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 10083–10092, http://dx.doi.org/10.1109/CVPR42600.2020.01010.
[6] Y. Qiao, C. Deng, Q. Wu, Referring expression comprehension: A survey of methods and datasets, IEEE Trans. Multimed. 23 (2020) 4426–4440.
[7] T. Zhou, F. Porikli, D.J. Crandall, L. Van Gool, W. Wang, A survey on deep learning technique for video segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (6) (2022) 7099–7122.
[8] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2021) 3523–3542.
[9] H. Li, Y. Chen, Q. Zhang, D. Zhao, Bifnet: Bidirectional fusion network for road segmentation, IEEE Trans. Cybern. 52 (9) (2021) 8617–8628.
[10] Z.Y. Qiu Shuang, W. Shikui, A survey of referring image segmentation, J. Signal Process. 38 (1003-0530(2022)06-1144-11) (2022) 1144, http://dx.doi.org/10.16798/j.issn.1003-0530.2022.06.002.
[11] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.
[12] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[13] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, J. Schmidhuber, LSTM: A search space odyssey, IEEE Trans. Neural Netw. Learn. Syst. 28 (10) (2016) 2222–2232.
[14] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 3431–3440, http://dx.doi.org/10.1109/CVPR.2015.7298965.
[15] R. Hu, M. Rohrbach, T. 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.
[16] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, A. Yuille, Recurrent multimodal interaction for referring image segmentation, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 1280–1289, http://dx.doi.org/10.1109/ICCV.2017.143.
[17] R. Li, K. Li, Y.-C. Kuo, M. Shu, X. Qi, X. Shen, J. Jia, Referring image segmentation via recurrent refinement networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5745–5753, http://dx.doi.org/10.1109/CVPR.2018.00602.

[18] W. Zhang, S. Tang, Y. Cao, S. Pu, F. Wu, Y. Zhuang, Frame augmented alternating attention network for video question answering, IEEE Trans. Multimed. 22 (4) (2020) 1032–1041, http://dx.doi.org/10.1109/TMM.2019.2935678.

[19] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, M. Tan, Visual grounding via accumulated attention, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7746–7755, http://dx.doi.org/10.1109/CVPR.2018.00808.

[20] L. Ye, M. Rochan, Z. Liu, Y. Wang, Cross-modal self-attention network for referring image segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 10494–10503, http://dx.doi.org/10.1109/CVPR.2019.01075.

[21] G. Luo, Y. Zhou, R. Ji, X. Sun, J. Su, C.-W. Lin, Q. Tian, Cascade grouped attention network for referring expression segmentation, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1274–1282.

[22] Z. Hu, G. Feng, J. Sun, L. Zhang, H. Lu, Bi-directional relationship inferring network for referring image segmentation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 4423–4432, http://dx.doi.org/10.1109/CVPR42600.2020.00448.

[23] S. Yang, M. Xia, G. Li, H.-Y. Zhou, Y. Yu, Bottom-up shift and reasoning for referring image segmentation, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 11261–11270, http://dx.doi.org/10.1109/CVPR46437.2021.01111.

[24] Y. Jing, T. Kong, W. Wang, L. Wang, L. Li, T. Tan, Locate then segment: A strong pipeline for referring image segmentation, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 9853–9862, http://dx.doi.org/10.1109/CVPR46437.2021.00973.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[26] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, J. Zhong, Attention is all you need in speech separation, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2021, pp. 21–25, http://dx.doi.org/10.1109/ICASSP39728.2021.9413901.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 9992–10002, http://dx.doi.org/10.1109/ICCV48922.2021.00986.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 936–944, http://dx.doi.org/10.1109/CVPR.2017.106.

[29] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, P.H. Torr, LAVT: Language-aware vision transformer for referring image segmentation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 18134–18144, http://dx.doi.org/10.1109/CVPR52688.2022.01762.

[30] Z. Zhang, Y. Zhu, J. Liu, X. Liang, W. Ke, Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation, Adv. Neural Inf. Process. Syst. 35 (2022) 14729–14742.

[31] B. Chen, Z. Hu, Z. Ji, J. Bai, W. Zuo, Position-aware contrastive alignment for referring image segmentation, 2022, ArXiv abs/2212.13419.

[32] Y.X. Chng, H. Zheng, Y. Han, X. Qiu, G. Huang, Mask grounding for referring image segmentation, 2023, ArXiv abs/2312.12198.

[33] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, 2019.

[34] C. Zhu, Y. Zhou, Y. Shen, G. Luo, X. Pan, M. Lin, C. Chen, L. Cao, X. Sun, R. Ji, Seqtr: A simple yet universal network for visual grounding, in: European Conference on Computer Vision, Springer, 2022, pp. 598–615.

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

[36] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H. Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.

[37] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.

[38] B. Cheng, A. Schwing, A. Kirillov, Per-pixel classification is not all you need for semantic segmentation, Adv. Neural Inf. Process. Syst. 34 (2021) 17864–17875.

[39] H. Ding, H. Zhang, J. Liu, J. Li, Z. Feng, X. Jiang, Interaction via bi-directional graph of semantic region affinity for scene parsing, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15848–15858.

[40] G. Luo, Y. Zhou, X. Sun, L. Cao, C. Wu, C. Deng, R. Ji, Multi-task collaborative network for joint referring expression comprehension and segmentation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 10031–10040, http://dx.doi.org/10.1109/CVPR42600.2020.01005.

[41] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, T.L. Berg, MAttNet: Modular attention network for referring expression comprehension, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1307–1315, http://dx.doi.org/10.1109/CVPR.2018.00142.

[42] S. Huang, T. Hui, S. Liu, G. Li, Y. Wei, J. Han, L. Liu, B. Li, Referring image segmentation via cross-modal progressive comprehension, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 10485–10494, http://dx.doi.org/10.1109/CVPR42600.2020.01050.

[43] M. Mitchell, K. Van Deemter, E. Reiter, Generating expressions that refer to visible objects, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 1174–1184.

[44] H. Ding, C. Liu, S. Wang, X. Jiang, Vision-language transformer and query generation for referring segmentation, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 16301–16310, http://dx.doi.org/10.1109/ICCV48922.2021.01601.

[45] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, P.H.S. Torr, Semantics-aware dynamic localization and refinement for referring image segmentation, in: AAAI Conference on Artificial Intelligence, 2023.

[46] J. Tang, G. Zheng, C. Shi, S. Yang, Contrastive grouping with transformer for referring image segmentation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 23570–23580, http://dx.doi.org/10.1109/CVPR52729.2023.02257.

[47] Y. Yan, X. He, W. Wan, J. Liu, MMNet: Multi-mask network for referring image segmentation, 2023, arXiv preprint arXiv:2305.14969.

[48] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Adv. Neural Inf. Process. Syst. 32 (2019).

[49] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, Vl-bert: Pre-training of generic visual-linguistic representations, 2019, arXiv preprint arXiv:1908.08530.

[50] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[51] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, T. Liu, CRIS: CLIP-driven referring image segmentation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 11676–11685, http://dx.doi.org/10.1109/CVPR52688.2022.01139.

[52] W. Wang, X. He, Y. Zhang, L. Guo, J. Shen, J. Li, J. Liu, CM-MaskSD: Cross-modality masked self-distillation for referring image segmentation, IEEE Trans. Multimed. (2024) 1–11, http://dx.doi.org/10.1109/TMM.2024.3358085.

[53] S. Kim, M. Kang, J. Park, RISCLIP: Referring image segmentation framework using CLIP, 2023, arXiv preprint arXiv:2306.08498.

[54] J. Wu, X. Li, X. Li, H. Ding, Y. Tong, D. Tao, Towards robust referring image segmentation, 2022, arXiv preprint arXiv:2209.09554.

[55] H. Ding, C. Liu, S. Wang, X. Jiang, VLT: Vision-language transformer and query generation for referring segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (6) (2023) 7900–7916, http://dx.doi.org/10.1109/TPAMI.2022.3217852.

[56] C. Liu, H. Ding, X. Jiang, GRES: Generalized referring expression segmentation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 23592–23601, http://dx.doi.org/10.1109/CVPR52729.2023.02259.

[57] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 2980–2988, http://dx.doi.org/10.1109/ICCV.2017.322.

[58] Y. Hu, Q. Wang, W. Shao, E. Xie, Z. Li, J. Han, P. Luo, Beyond one-to-one: Rethinking the referring image segmentation, in: 2023 IEEE/CVF International Conference on Computer Vision, ICCV, 2023, pp. 4044–4054, http://dx.doi.org/10.1109/ICCV51070.2023.00376.

[59] S. Kazemzadeh, V. Ordonez, M. Matten, T. 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.

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

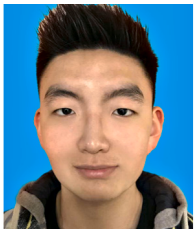
[61] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[62] J. Mao, J. Huang, A. Toshev, O. Camburu, A.L. Yuille, K. Murphy, Generation and comprehension of unambiguous object descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 11–20.
[63] V.K. Nagaraja, V.I. Morariu, L.S. Davis, 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, Springer, 2016, pp. 792–807.
[64] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).
[65] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[66] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, 2018, arXiv preprint arXiv:1804.02767.
[67] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, T. Darrell, Natural language object retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4555–4564.
[68] C. Wu, Z. Lin, S. Cohen, T. Bui, S. 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.
[69] Y. Wu, Z. Zhang, C. Xie, F. Zhu, R. Zhao, Advancing referring expression segmentation beyond single image, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2628–2638.
[70] L. Xu, M.H. Huang, X. Shang, Z. Yuan, Y. Sun, J. Liu, Meta compositional referring expression segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19478–19487.
[71] R. Strudel, I. Laptev, C. Schmid, Weakly-supervised segmentation of referring expressions, 2022, arXiv preprint arXiv:2205.04725.
[72] H. Li, M. Sun, J. Xiao, E.G. Lim, Y. Zhao, Fully and weakly supervised referring expression segmentation with end-to-end learning, IEEE Trans. Circuits Syst. Video Technol. (2023).
[73] S. Yu, P.H. Seo, J. Son, Zero-shot referring image segmentation with global-local context features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19456–19465.
[74] C. Liang, W. Wang, T. Zhou, J. Miao, Y. Luo, Y. Yang, Local-global context aware transformer for language-guided video segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (8) (2023) 10055–10069, http://dx.doi.org/10.1109/TPAMI.2023.3262578.
[75] Z. Xu, Z. Chen, Y. Zhang, Y. Song, X. Wan, G. Li, Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 17503–17512.

Lixia Ji was born in 1979. She is currently an Associate Professor of computer science and applications with Zhengzhou University, and is also the Deputy Director of Zhengzhou Key Laboratory of Blockchain and Data Intelligence. Her current research interests include knowledge mining, multi-modal learning, and data intelligence.

Yunlong Du was born in 2000. He received the B.E. degree in software engineering from Henan University in 2022. Currently, he is working on his master's degree at Zhengzhou University. His current research interests include image segmentation and multi-modal learning.

Yiping Dang was born in 2001. She received the B.E. degree in software engineering from Wuhan University of Science and Technology in 2022. Currently, she is working on her master's degree at Zhengzhou University. Her current research interests include named entity recognition and artificial intelligence.



Wenzhao Gao was born in 2001 and he is an undergraduate student of Artificial Intelligence at the Ulster University.

Han Zhang was born in 1985. She received the Ph.D. degree in Computer Science from Information Engineering University in 2020. She is a Lecturer at Zhengzhou University and a Visiting Scholar at the University of Ulster in the UK. Her main research interest is natural language processing and information security.