Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 15, No. 2, 2026

Multi-scale Feature Extraction and Dynamic Aggregation for Referring Image Segmentation

Xinzhuo Gao, Quange Tan, Siyu Meng, Mingjie Wang and Rong Wang *

College of Information and Cyber Security, People's Public Security University of China, Beijing, China
*** Corresponding author:** Rong Wang (Email: dbdxwangrong@163.com)

Abstract: Image referring segmentation aims to accurately locate and segment the corresponding target regions in an image based on textual descriptions. Existing methods mainly face three major challenges: Firstly, there is insufficient information interaction in the extraction process of multi-scale visual features, making it dif-ficult to balance local details and global semantics. Secondly, cross-modal semantic alignment lacks a bidirectional guidance mechanism, resulting in weak semantic consistency between image and text features. Thirdly, the feature aggregation strategy lacks dynamic adaptability and cannot flexibly adjust feature weights according to scene complexity, leading to insufficient feature discriminability. To address the above issues, a referring image segmentation method based on a chained collaborative mechanism for multi-scale feature extraction and dynamic aggregation is proposed. Firstly, a multi-scale adaptive feature fusion module is constructed. The image features after patch embedding are split into four channel sub-spaces through group convolution. After extracting multi-scale features via adaptive max-pooling, the gating network dynamically adjusts the fusion weights to realize the adaptive interaction between local details and global semantics, thus improving the diversity of feature expression. Secondly, a dual-attention modal alignment module is designed. Channel-first convolutional attention is introduced into image fea-tures, which adjusts the weights of different channels and spatial positions in the feature map under the guidance of text features to strengthen the expression of key regions. Efficient channel attention is in-troduced into text features, which takes image features as prior and enhances the semantic relevance of text through channel interaction, improving the accuracy and efficiency of feature alignment. Finally, a dy-namic feature aggregation module is proposed to explore the non-linear dependencies between channels, generate fine-grained feature weights, and dynamically adjust features of each scale through competitive weight allocation, so as to realize the complementary aggregation of global and local features and improve the discriminability of features. Experimental results demonstrate that the proposed method achieves In-tersection over Union (IoU) scores of 74.34% and 66.59% on testA and testB of the RefCOCO dataset, 68.99% and 53.87% on testA and testB of the RefCOCO+ dataset, and 61.65% on the test set of the G-Ref dataset, respectively, verifying the effectiveness of the proposed method.

Keywords: Referring Image Segmentation; Multi-scale Feature; Feature Extraction; Attention Mechanism; Dynamic Aggregation.

1. Introduction

Referring Image Segmentation (RIS), a core task at the intersection of computer vision and natural language processing, aims to precisely localize target objects in an image based on a textual description. Its essence lies in establishing an accurate mapping between visual features and textual semantics. Unlike traditional instance segmentation, which focuses on identifying and segmenting all objects in an image, RIS achieves precise localization and segmentation of a specific target by integrating object appearance features with the semantic information of the text, coupled with contextual reasoning. This linguistic query-based approach enables direct and accurate pointing to specific regions and objects, facilitating higher-level semantic understanding and interaction. This task holds significant practical value in scenarios such as intelligent interaction, autonomous driving, and image editing. For instance, in an intelligent shopping guide system, a user's natural language command like "pick the snack in the blue packaging on the shelf" allows the system to quickly locate and segment the target product. In autonomous driving, given a description like "identify the pedestrian squatting on the right side of the road," the model can perform real-time alignment of visual targets and textual semantics. Consequently, RIS has garnered widespread attention in the field of computer vision.

Although existing RIS methods have evolved in feature extraction and modal interaction strategies, they still face the following challenges: First, insufficient information interaction in multi-scale visual features. Early methods, such as the initial approach proposed by Hu et al.[1], relied on simple matching of image features with bag-of-words encoded text, depending on a single-pathway feature extraction that struggles to balance local details and global semantic information. Second, a lack of bidirectional guidance mechanisms in cross-modal semantic alignment. While many methods, like the progressive framework by Huang et al. [2], improve accuracy through multi-stage alignment, the role of text features remains fixed, lacking bidirectional feedback. Third, inadequate dynamic adaptability in feature aggregation strategies. Existing models often employ fixed weight allocation strategies. For example, Li et al.[3]utilized Transformers for cross-modal fusion but adopted concatenation and fixed-weight aggregation for multi-scale visual features, failing to effectively preserve detailed information.

Examining the field's development trend, image segmentation tasks have progressively evolved from single-modality independent modeling to multi-modal deep fusion. Methods like CLIP[4] achieve cross-modal association between vision and language through contrastive learning, while SAM [5] supports multi-category target segmentation

with its versatile segmentation capabilities. These approaches provide crucial insights for RIS: constructing dynamic, bidirectional cross-modal interaction mechanisms is the core direction for overcoming current bottlenecks. However, RIS has its unique characteristics—namely, the ambiguity of linguistic descriptions, the diversity of visual targets, and the complexity of semantic mapping between modalities. These require models not only to extract high-quality multi-scale features but also to achieve precise vision-language alignment and dynamic aggregation. Existing methods have limitations in addressing these challenges: early single-modality modeling approaches, such as the two-stream network by Mao et al.[6], suffer from performance degradation in complex description scenarios due to the lack of dynamic cross-modal interaction. Transformer-based models like SLViT by Ouyang et al.[7], although enhancing fusion through multi-scale linguistic guidance, exhibit insufficient dynamic feature aggregation and are susceptible to redundant information interference. These issues render current models less robust in practical applications, failing to meet real-world demands.

To address these issues, this paper proposes a novel Referring Image Segmentation method based on multi-scale feature extraction and dynamic feature aggregation, building upon the ETRIS[8] model. Our approach forms a chain-like collaborative framework consisting of three modules: a Multi-scale Adaptive Feature Fusion Module, a Dual-Attention Modality Alignment Module, and a Feature Dynamic Aggregation Module. The Multi-scale Adaptive Feature Fusion Module achieves adaptive fusion of multi-scale features, providing a rich visual foundation for precise segmentation. The Dual-Attention Modality Alignment Module enables bidirectional semantic alignment between image and text, ensuring cross-modal consistency. The Feature Dynamic Aggregation Module implements dynamic weight allocation for features, optimizing the discriminative power of the final representations. These three components form a closed-loop optimization, enhancing multi-scale information representation, improving cross-modal feature interaction efficiency, and optimizing dynamic feature weight allocation. The main contributions are as follows:

(1) We construct a Multi-scale Adaptive Feature Fusion Module (MAFF). This module splits patch-embedded image features into four channel subspaces via group convolution, employs adaptive max-pooling and upsampling to extract contextual information at different scales while preserving high-frequency details. A lightweight gating network is introduced to generate spatial attention maps based on shallow features, dynamically adjusting fusion weights for cross-scale features. This achieves adaptive fusion of shallow detail features and deep semantic features, effectively overcoming the limitations of fixed weights in traditional multi-scale fusion and enhancing the diversity of feature representations.

(2) We design a Dual-Attention Modality Alignment Module (DAMA). For image features, Channel Prior Convolutional Attention (CPCA) is introduced, using the textual description as guidance. Attention masks are generated via channel-spatial convolutional attention to weaken background noise and highlight key regions corresponding to the text. Concurrently, for text features, Efficient Channel Attention (ECA) is introduced, using image features as prior knowledge. Text features are calibrated through a channel interaction mechanism to suppress descriptions irrelevant to the image, enhancing semantic relevance. This effectively mitigates semantic偏差 caused by unidirectional alignment and improves cross-modal semantic consistency.

(3) We propose a Feature Dynamic Aggregation Module (FDA). This module achieves dynamic allocation of attention weights through the synergy of dual activation functions. The Sigmoid function is utilized to model non-linear channel-wise dependencies, generating fine-grained feature weights to enhance local key features. This is combined with the Softmax function for competitively allocating weights, dynamically adjusting features across scales and suppressing redundant information. By fusing features from different scales, complementary aggregation of global and local features and preservation of detailed information are achieved, effectively enhancing feature discriminability and enabling the model to adaptively adjust feature importance based on scene complexity.

The remainder of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 elaborates on the proposed Referring Image Segmentation method based on multi-scale feature extraction and dynamic feature aggregation. Section 4 presents experimental evaluations and comparisons with existing methods. Finally, Section 5 summarizes the work of this paper.

2. Related Work

The development of image referring segmentation has consistently centered on the advancement of feature extraction capabilities and the improvement of cross-modal fusion mechanisms, with these two aspects mutually reinforcing one another: advancements in feature extraction technology provide richer visual representations for modal fusion, while the optimization of fusion mechanisms drives feature extraction to evolve in directions better aligned with semantic requirements. The evolutionary trajectory of existing methods can be traced from the perspectives of feature extraction and cross-modal fusion.

2.1. Feature Extraction Methods

Early feature extraction methods were constrained by single-modality modeling approaches, predominantly employing fixed architectures for visual feature encoding. Although the two-stream network proposed by Mao et al. [6] distinguished between image and text feature extraction pathways, the visual branch still relied on a single CNN structure, lacking the capability to model multi-scale contexts and struggling to capture complete object contours in occlusion scenarios. When Rohrbach et al.[9] proposed an alignment strategy based on reconstruction loss, their adopted visual feature extraction network failed to account for variations in object scale, exhibiting limited perception of image details and difficulty handling ambiguous scenes.

With the development of Feature Pyramid Networks (FPN), multi-scale feature extraction gradually became mainstream. Liu et al.[10] introduced the FPN architecture to obtain visual features at different levels, fusing shallow details with deep semantics through bottom-up and top-down pathways. However, their feature fusion weights were assigned with fixed ratios, unable to dynamically adjust scale attention based on textual descriptions. Yang et al. [11] further refined multi-scale feature hierarchies in the LAVT model by interacting textual semantics with features of different

resolutions, yet feature extraction still relied on fixed hierarchical outputs from pre-trained CNNs, lacking adaptive adjustment mechanisms tailored for referring tasks, resulting in insufficient feature responses in semantically relevant regions.

The introduction of Transformer architecture brought a new paradigm to feature extraction. Li et al. [3] embedded image patches and encoded global features through Transformers, but their multi-scale processing was achieved merely through simple downsampling, neglecting the semantic correlations among features at different scales. The ReSTR model proposed by Kim et al. [12] achieved end-to-end feature extraction entirely based on Transformers, yet the abandonment of convolutional operations led to blurred local texture features, performing poorly in fine-grained segmentation tasks. The SLViT model by Ouyang et al.[7] attempted to adjust feature receptive fields through multi-scale convolutional kernels, but the convolution parameters remained fixed, unable to dynamically focus on critical scales according to textual descriptions, leaving limitations in the scene adaptability of feature extraction.

2.2. Cross-Modal Fusion Methods

In the early fusion stage, the core approach centered on independent single-modality feature modeling, with fusion strategies being relatively coarse. Hu et al. [1] made the first attempt to match text features encoded by a bag-of-words model with image region features using cosine similarity. Here, the text served merely as a static label and did not participate in the dynamic modulation of visual features, resulting in poor segmentation performance in scenarios with polysemous words. Mao et al.[6] concatenated CNN visual features and RNN text features through fully connected layers. This early fusion method failed to capture deep correlations between modalities, leading to a significant drop in accuracy when the text contained attribute descriptions. Rohrbach et al.[9] proposed an alignment strategy based on reconstruction loss, forcing textual phrases to reconstruct image region features. While this enhanced modality association, it did not resolve semantic ambiguity issues in complex description scenarios.

The multi-modal interaction phase began to explore bidirectional fusion mechanisms, yet limitations of unidirectional dependency persisted. The progressive framework proposed by Huang et al.[2] filtered candidate regions using textual keywords before performing fine-grained feature alignment. Although this established a clear guiding role for text over vision, it lacked reverse calibration of textual semantics by visual features, making it prone to localization bias in ambiguous description scenarios. Feng et al.[13] designed a fusion encoder network that progressively refined multi-modal features using linguistic features and embedded attention mechanisms to promote cross-modal consistency. However, their attention computation was confined to local feature interactions, lacking global semantic guidance. Jing et al. [14] proposed a two-stage "localization-segmentation" framework, which first narrowed the search range through cross-modal feature alignment and then finely segmented candidate regions, improving efficiency and accuracy. Nevertheless, information loss occurred during feature transmission between the two stages, leading to suboptimal performance on small object segmentation. Feng et al.[15] designed a bidirectional relation reasoning network that modeled inter-modal dependencies through vision-guided language attention and language-guided vision attention, enhancing semantic correlation consistency. However, their relation reasoning was based solely on pairwise feature interactions, making it difficult to capture hierarchical relationships among multiple objects in complex scenes and thus prone to localization shift when segmenting group targets.

The stage of Transformers and pre-trained models achieved global-scale cross-modal associations, but limitations in fusion dynamics remain. Wang et al. [16] constructed a cross-modal framework based on the CLIP pre-trained model, leveraging contrastive learning to enhance vision-text semantic associations. By freezing the pre-trained encoder and training only a lightweight segmentation head, they achieved efficient transfer. However, the adaptability of pre-trained features to downstream tasks was insufficient, resulting in subpar performance on fine-grained segmentation tasks. Zhang et al. [17] designed a mutual perception attention module enabling vision and text features to calibrate each other during fusion. Combined with the Segment Anything Model, this enhanced the model's robustness in segmenting unknown targets. Nevertheless, weight allocation during multi-scale feature aggregation remained fixed, providing inadequate suppression of redundant information.

Although the methods in the aforementioned feature extraction and cross-modal fusion stages have made significant progress in multi-scale modeling and modal interaction, continuously advancing the performance of image referring segmentation, three limitations still persist: First, insufficient information interaction in multi-scale visual features. Whether based on CNN feature pyramids or Transformer patch embedding, existing methods struggle to achieve dynamic correlation and complementarity across scales while simultaneously preserving local details and global semantics. Second, the lack of bidirectional guidance mechanisms in cross-modal semantic alignment. Most methods still primarily employ text-guided vision, lacking reverse calibration and filtering of textual semantics by visual features, leading to decreased semantic consistency in ambiguous description scenarios. Third, inadequate dynamic adaptability in feature aggregation strategies. Current models predominantly adopt fixed weight allocation approaches, unable to dynamically adjust the contribution of features across different scales and modalities based on the correlation strength between textual descriptions and image content, resulting in limited capability to suppress redundant information. To address these issues, this paper proposes an image referring segmentation method based on multi-scale feature extraction and dynamic aggregation, achieving collaborative optimization through the construction of a Multi-scale Adaptive Feature Fusion module, a Dual-Attention Modality Alignment module, and a Feature Dynamic Aggregation module.

3. Methodology

The image referring segmentation method based on multi-scale feature extraction and dynamic feature aggregation proposed in this paper mainly consists of three modules: Multi-scale Adaptive Feature Fusion, Dual-Attention Modality Alignment, and Feature Dynamic Aggregation, as illustrated in Fig 1. First, in the Multi-scale Adaptive Feature Fusion module, multi-scale features of the patch-embedded image are extracted through a multi-branch parallel structure. A lightweight gating network is then employed to perform

weighted fusion of features at different scales, achieving full utilization of multi-scale features and enhancing the diversity of feature representations. Second, in the Dual-Attention Modality Alignment module, the CPCA attention mechanism is utilized to adjust the importance of different channels and spatial positions in the image features containing multi-scale information, thereby enhancing the representational capacity of key image regions. Simultaneously, the ECA attention mechanism is applied to weight the individual channels of the token-embedded text features, enriching the semantic information of the textual features and improving the model's ability to comprehend the correlation between images and text. Finally, in the Feature Dynamic Aggregation module, multiple aggregation units are employed to construct a multi-scale receptive field. Combined with dual activation functions, fine-grained feature weights are generated to dynamically adjust features across different scales. This enables complementary aggregation of global and local features, enhancing the discriminative power of the final features.

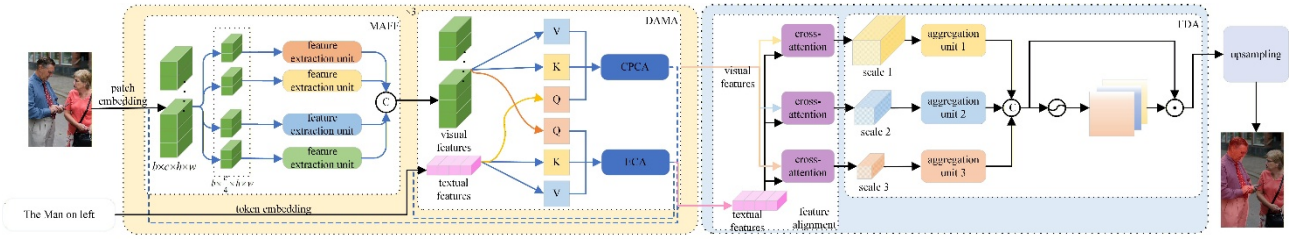


Fig 1. A Framework of the proposed method

3.1. Multi-scale Adaptive Feature Fusion Module

Conventional feature extraction modules often rely on fixed-structure convolutional kernels or single-scale pooling operations to process feature maps, which makes it challenging to effectively capture multi-scale contextual information characterized by diverse distributions within an image. To address this limitation, this paper constructs a multi-scale adaptive feature fusion module, which innovatively introduces a dynamic weight adjustment mechanism, as illustrated in Fig 2. First, group convolution is employed to partition the patch-embedded image features into multi-scale channel subspaces along the channel dimension. Subsequently, adaptive pooling and upsampling operations are utilized to extract contextual information across different scales. Finally, the dynamic weight adjustment mechanism facilitates adaptive interaction between local details and global semantics, thereby preserving spatial information while enhancing the model's ability to capture fine-grained features and significantly improving the diversity of feature representations.

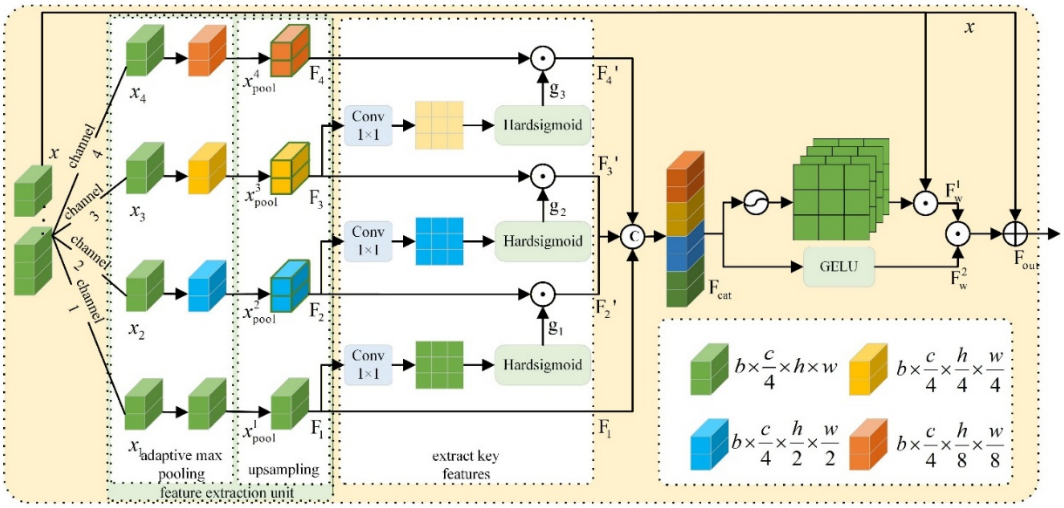


Fig 2. Structure of multi-scale adaptive feature fusion module

This module employs group convolution to evenly split the patch-embedded image features into four groups along the channel dimension. Through distinct feature extraction units, adaptive max-pooling is applied to extract contextual information at different scales, as formulated in Equation (1), yielding $x_{\text{pool}}^{i}, i=1,2,3,4$.

$$x_{\text{pool}}^{i}=\begin{cases} x_i, & i=1 \\ AdaptiveMaxPool\left(x_i, \dfrac{\text{H}}{2^{i-1}}, \dfrac{\text{W}}{2^{i-1}}\right), & i>1 \end{cases} \tag{1}$$

Here, *AdaptiveMaxPool* denotes the adaptive max-pooling operation. Upsampling is applied to x_{pool}^{i} as shown in Equation (2), yielding the multi-scale image features $\text{F}_i, i=1,2,3,4$.

$$\text{F}_i=\begin{cases} x_{\text{pool}}^{i}, & i=1 \\ Interpolate\left(x_{\text{pool}}^{i}\right), & i>1 \end{cases} \tag{2}$$

Here, *Interpolate* represents the nearest neighbor interpolation upsampling operation. The features $\text{F}_1, \text{F}_2, \text{F}_3$ are passed through a 1×1 convolution to generate an attention map in the spatial dimension. This attention map is then

linearly processed using the *Hardsigmoid* function to produce corresponding weights g_1, g_2, g_3, which dynamically adjust the importance of different spatial positions within the features, as formulated in Equation (3).

$$g_i = Hardsigmoid\left(\mathrm{Conv}_{1\times1}\left(F_i\right)\right), i = 1,2,3 \tag{3}$$

Here, *Hardsigmoid* function is designed using a piecewise linear approximation to prevent gradients from approaching zero when input values deviate significantly from zero, as expressed in Equation (4), thereby effectively mitigating the vanishing gradient problem.

$$Hardsigmoid(x) = \max(0, \min(1, \frac{x}{6} + \frac{1}{2})) \tag{4}$$

Here, x denotes the input to the function. When $x \in [-3,3]$, the function is linear with slope $\frac{1}{6}$ and intercept $\frac{1}{2}$; when $x < -3$, the output is 0; and when $x > 3$, the output is 1.

To achieve adaptive fusion of shallow detail features and deep semantic features, with a focus on better capturing fine-grained details, the weights g_1, g_2, g_3 obtained from formula (3) are multiplied with the features F_2, F_3, F_4, thereby obtaining enhanced multi-granularity features $F_s^{'}$ as shown in formula (5).

$$F_s^{'} = g_{s-1} \cdot F_s, s = 2,3,4 \tag{5}$$

The weighted features $F_s^{'}$ are concatenated with the features F_1 along the channel dimension to obtain richer feature representations, as shown in Equation (6).

$$F_{cat} = Cat\left(\left[F_1, F_2^{'}, F_3^{'}, F_4^{'}\right]\right) \tag{6}$$

Here, *Cat* denotes the feature concatenation process. A dual-branch structure is employed to enhance the concatenated features F_{cat}, as described in Equations (7) and (8):

$$F_w^1 = Softmax\left(F_{cat}\right) \cdot x \tag{7}$$

$$F_w^2 = GELU\left(\mathrm{Conv}\left(F_{cat}\right)\right) \tag{8}$$

The result obtained by multiplying F_w^1 and F_w^2 is then added via a residual connection to the original input features x, yielding the enhanced output features F_{out}, as shown in Equation (9).

$$F_{out} = F_w^1 \cdot F_w^2 + x \tag{9}$$

3.2. Dual-Attention Modality Alignment Module

To better understand the correspondence between images and texts, effectively capture and utilize key information, and improve the accuracy of image-text alignment, this paper designs a dual-attention modality alignment module. It innovatively introduces bidirectional attention guidance to dynamically perceive the semantic differences between image and text features, thereby enhancing the model's ability to focus on key features in both modalities and improving cross-modal semantic consistency.

The structure of the dual-attention modality alignment module is shown in Fig 3. First, the image features and text features are respectively passed through independent linear layers to adjust them to the same dimension. Next, the linearly transformed image features and text features are fed into a cross-attention module. Here, the text features serve as the Query, while the image features act as the Key and Value. Attention weights are generated by computing similarity scores, enabling the alignment and weighting of text features with respect to image features. Simultaneously, the image features are used as Query to interact with the text features, capturing textual semantic information and achieving bidirectional information fusion. Finally, the image features aligned by cross-attention are fed into CPCA attention [19]. After passing through channel sub-attention to obtain refined channel-wise features, spatial sub-attention is applied via convolution to generate spatial attention weights. The refined channel-wise features are then multiplied by the spatial attention weights, enhancing the image features sequentially from channel and spatial dimensions. Meanwhile, the text features aligned by cross-attention are fed into ECA attention [20]. Global average pooling is used to extract global channel information, followed by one-dimensional convolution to capture channel dependencies and generate channel attention weights. After normalization via Sigmoid, these weights are multiplied element-wise with the text features along the channel dimension, thereby enhancing text information from the channel perspective and reducing interference from redundant information.

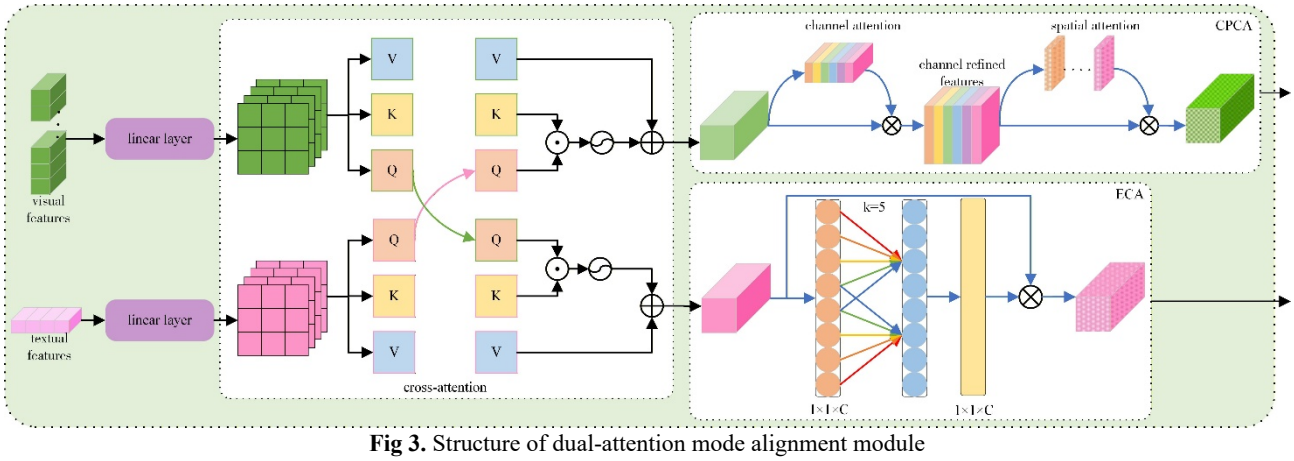


Fig 3. Structure of dual-attention mode alignment module

3.3. Feature Dynamic Aggregation Module

Traditional methods often employ static weighting strategies when dealing with multi-scale feature fusion, which leads to the loss of local detailed information and the fragmentation of global context. To better preserve local details, aggregate contextual information, enhance the discriminability of multi-scale features, and achieve complementary fusion of cross-scale information, this paper constructs a Feature Dynamic Aggregation Module. Its core innovation lies in proposing a dynamic aggregation strategy based on competitive allocation. As illustrated in Fig 4, taking the image features processed by the dual-attention modality alignment module as input, the Sigmoid function is first used to establish non-linear relationships across channels and generate fine-grained weights. Subsequently, the Softmax function is applied to perform competitive allocation of these weights, dynamically adjusting the importance of features at different scales. Finally, global and local features are complementarily aggregated to enhance the discriminability of the features.

First, global average pooling is applied to the input image features $f_i, i=1,2,3$ at each scale using the aggregation unit to extract global information. Subsequently, a dilated convolution operation is performed on the image features to obtain the enhanced features f_a^i, as shown in Equation (10):

$$f_a^i = SE_i\left(GAP\left(f_i\right)\right), i=1,2,3 \tag{10}$$

Here, SE_i represents the dilated convolution module corresponding to features at different scales. Then, the enhanced features from the three scales are concatenated into a feature vector. Weights for each scale are generated through the *Sigmoid* and *Softmax* functions, as shown in Equation (11). These weights are then multiplied with their corresponding features to obtain the weighted multi-scale features f' as illustrated in Equation (12):

$$w = Softmax\left(Sigmoid\left(Cat\left(\left[f_a^1, f_a^2, f_a^3\right]\right)\right)\right) \tag{11}$$

$$f' = w_1 \cdot f_1 + w_2 \cdot f_2 + w_3 \cdot f_3 \tag{12}$$

Here, w_i represents the dynamic weights, and $f' \in \mathrm{R}^{3\times1\times c}$ is a tensor containing the weights for the three scales. Finally, the weighted and fused features f' are passed through the *Sigmoid* function and a convolution operation to generate weights. These weights are then multiplied with f' to obtain the final output f_{out} of the feature dynamic aggregation module, as shown in Equation (13):

$$f_{\text{out}} = \mathrm{Conv}\left(Sigmoid\left(f'\right)\right) \cdot f' \tag{13}$$

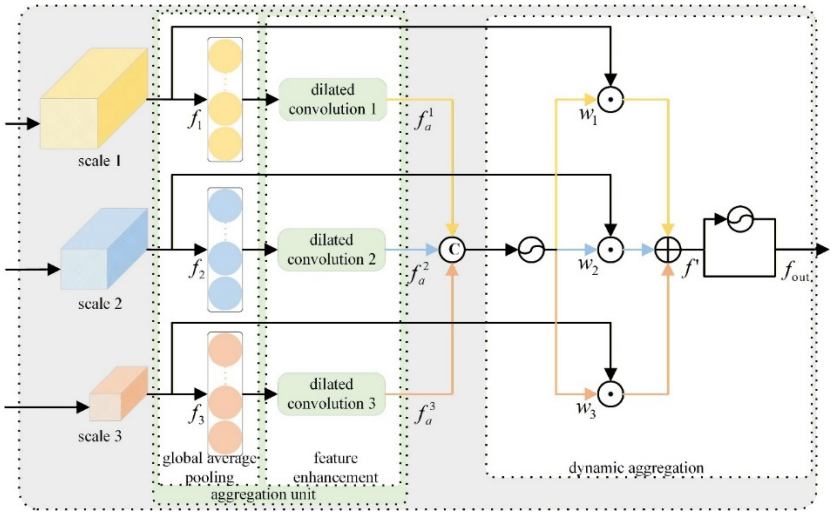


Fig 4. Structure of feature dynamic aggregation module

3.4. Loss Function

This paper designs a joint loss function that integrates the optimization characteristics of binary cross-entropy loss and Dice loss. An equal-proportion linear weighted fusion strategy is adopted to construct the total loss function $\mathrm{L_T}$, aiming to minimize the error between global prediction accuracy and local boundary localization, as shown in Equation (14).

$$\mathrm{L_T} = \lambda_1 \mathrm{L_{BCE}} + \lambda_2 \mathrm{L_{Dice}} \tag{14}$$

Here, λ_1 and λ_2 are both set to 1, ensuring a balanced propagation of gradient magnitudes between the two loss functions while taking into account the complexity of hyperparameter tuning.

The binary cross-entropy loss function calculates the binary cross-entropy between the predicted values and the ground truth labels, enabling the model to adjust its weights based on the proximity of predictions to the true values. This drives the model to optimize overall classification accuracy, as shown in Equation (15). The Dice loss function, on the other hand, measures the spatial overlap between the predicted segmentation region and the ground truth region. It strengthens the model's ability to capture subtle differences at the segmentation boundaries, thereby improving prediction accuracy for contour regions, as shown in Equation (16).

$$\mathrm{L_{BCE}} = -\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\left(y_i \log\left(p_i\right) + \left(1-y_i\right)\log\left(1-p_i\right)\right) \tag{15}$$

$$\mathrm{L_{Dice}} = 1 - \frac{2\sum_{i=1}^{\mathrm{N}} y_i p_i}{\sum_{i=1}^{\mathrm{N}} y_i + \sum_{i=1}^{\mathrm{N}} p_i} \tag{16}$$

Here, N denotes the number of pixels in the image, y_i represents the ground truth label of the i-th pixel, and p_i indicates the predicted value of the i-th pixel.

4. Experiments

4.1. Dataset and Evaluation Metrics

4.1.1. Dataset

The experiments in this paper are conducted on three commonly used referring image segmentation datasets: RefCOCO [22], RefCOCO+ [22], and G-Ref [23]. The details of these datasets are shown in Table 1.

Table 1. Referring image segmentation dataset

Dataset	Number of images	Referring expressions	Object instances	Average query length
RefCOCO	19994	142210	50000	3.61
RefCOCO+	19992	141564	49856	3.53
G-Ref	25799	95010	49822	8.43

4.1.2. Evaluation Metrics

This paper adopts the commonly used evaluation metrics for referring image segmentation: Intersection over Union (IoU) and Precision. IoU is a key metric for measuring the degree of overlap between the predicted result and the ground truth region, representing the ratio of the intersection area to the union area of the predicted region and the ground truth region, as shown in Equation (17):

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{17}$$

Here, A represents the predicted segmentation region, and B represents the ground truth segmentation region. $|A \cap B|$ denotes the intersection area of the predicted and ground truth regions, while $|A \cup B|$ denotes their union area. Precision, on the other hand, calculates the prediction accuracy of the model under different IoU thresholds, as shown in Equation (18):

$$Pr@t = \frac{s'}{s} \tag{18}$$

Here, t denotes the IoU threshold. In this paper, t is set to 0.5, 0.6, 0.7, 0.8, and 0.9. s' represents the number of predicted samples with IoU greater than or equal to t, and s is the total number of samples.

4.2. Experimental Setup

This experiment was conducted on the Ubuntu 22.04 LTS operating system, using an NVIDIA GeForce RTX 4090 graphics card with 24GB of video memory and 64GB of system RAM. The model was optimized using the Adam optimizer, with an initial learning rate set to 0.0001. The training epoch was set to 50, and the batch size was 32. The loss function employed was a combination of binary cross-entropy loss and Dice loss.

4.3. Comparative Experiments

To validate the effectiveness of the proposed method, comparative experiments were conducted between our approach and existing referring image segmentation methods on three datasets: RefCOCO, RefCOCO+, and G-Ref. Several representative methods in the current field were selected as baselines for comparison. All methods were tested under identical experimental conditions, including the same datasets, data preprocessing pipelines, training parameter settings, and evaluation metrics, ensuring a fair and accurate comparison of performance differences. The comparative results are shown in Table 2.

Table 2. Comparation of different methods

Method	RefCOCO			RefCOCO+			G-Ref		
	val	testA	testB	val	testA	testB	val(u)	test(u)	val(g)
RMI[24]	45.18	45.69	45.57	29.86	30.48	29.50	—	—	—
DMN[25]	49.78	54.83	45.13	38.88	44.22	32.29	—	—	36.76
RRN[26]	55.33	57.26	53.95	39,75	42.15	35.11	—	—	36.45
MAttNet[27]	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61	—
NMTree[28]	56.59	63.02	52.06	47.40	53.01	41.56	46.59	47.88	—
CMSA[29]	58.32	60.61	55.09	43.76	47.60	37.89	—	—	39.98
CAC[30]	58.90	61.77	53.81	—	—	—	46.37	46.95	44.32
BRINet[15]	61.35	63.37	59.57	48.57	52.87	42.13	—	—	48.04
EFN[2]	62.76	65.69	59.67	51.50	55.24	43.01	—	—	—
BUSNet[31]	63.27	66.41	61.39	51.76	56.87	44.13	—	—	—
CGAN[32]	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	—
CRIS[16]	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36	—
ETRIS [8]	70.93	73.80	66.48	62.44	68.52	53.83	60.42	60.39	55.28
Ours	71.05	74.34	66.59	62.84	68.99	53.87	61.10	61.65	58.40

The proposed method outperforms the compared methods across all three datasets (RefCOCO, RefCOCO+, and G-Ref). The RMI [24] method emphasizes multi-modal interactive fusion but does not sufficiently extract multi-scale features or

mine semantic information. In contrast, our proposed Multi-scale Adaptive Feature Fusion module accurately captures key information, enhances the ability to capture image details, and provides rich features for segmentation, leading to significant performance improvements. Specifically, on RefCOCO testA, our method exceeds RMI by 28.63%, and on RefCOCO+ testA, it achieves a 38.51% improvement. Although CMSA[29] achieves feature interaction through cross-modal self-attention, its unidirectional mapping from vision to language struggles to fully capture bidirectional semantic correlations. Our designed Dual-Attention Modality Alignment module enhances image-text correlation and achieves precise image-text matching, resulting in a 13.71% improvement on RefCOCO testA and surpassing CMSA by 21.39% on RefCOCO+ testA. The CRIS[16] method employs a fixed-weight aggregation strategy for multi-modal feature fusion. In contrast, our proposed Feature Dynamic Aggregation module dynamically adjusts feature attention based on different input characteristics, enhancing feature representation capability. This leads to a 1.14% improvement on RefCOCO testA and surpasses CRIS by 0.91% on RefCOCO+ testA.

As a recent method in the field, ETRIS [8] relies on a cross-modal Transformer architecture for single-dimensional semantic interaction. Its feature aggregation adopts a fixed-weight strategy, which struggles to adapt to the differentiated requirements of detail features and semantic information for multi-scale targets. Furthermore, its cross-modal alignment process lacks dual constraints from channel and spatial dimensions, making it prone to alignment drift in complex semantic scenarios. Compared to ETRIS, our method designs the MAFF module to provide multi-scale feature support spanning from shallow details to deep semantics. The DAMA module leverages channel-spatial dual-dimension attention mechanisms to achieve precise alignment of key semantics between text and images. The FDA module dynamically assigns weights based on the aligned features, overcoming the limitations of fixed weights. Through the chain collaboration of the MAFF, DAMA, and FDA modules, our method fundamentally transcends the reliance of existing methods on single interaction modes and static aggregation strategies, achieving adaptive matching to the characteristics of different datasets.

4.4. Ablation Experiment Results and Case Analysis

4.4.1. Ablation Experiment Results

To validate the effectiveness of the three modules—Multi-scale Adaptive Feature Fusion (MAFF), Dual-attention Mechanism for Multimodal Alignment (DMAM), and Feature Dynamic Aggregation (FDA)—ablation experiments were conducted on the RefCOCO, RefCOCO+, and G-Ref datasets. First, experiments were performed on a baseline without any of the modules. Subsequently, the MAFF, DMAM, and FDA modules were added incrementally. Next, pairwise combinations of the modules were tested, including MAFF with DMAM, MAFF with FDA, and DMAM with FDA. Finally, all three modules were integrated simultaneously into the network. Following these steps, ablation experiments were carried out to comprehensively evaluate the contribution and effectiveness of each module by analyzing various metrics on the test sets.

The ablation experiment results on the RefCOCO dataset are shown in Table 3.

Adding only DMAM improved the IoU to 74.33%, making it the most effective single module. This indicates that DMAM effectively captures key associations between images and text through its channel-spatial attention mechanism, suppressing redundant information and significantly enhancing target localization accuracy.

Furthermore, when DMAM and FDA were combined, the IoU further increased to 74.48%, suggesting that the cross-modal alignment capability of DMAM and the dual-activation weighting of FDA form a chain optimization effect, complementing each other during feature aggregation. The combination of MAFF and FDA resulted in an IoU improvement to 74.23%, demonstrating that multi-scale features generated by MAFF, when processed through FDA's dual-weight mechanism, effectively integrate shallow details and deep semantics, preserving edge information while enhancing category discriminability. However, when MAFF and DMAM were combined, the IoU dropped to 73.81%, as the extensive detail features introduced by MAFF interfered with DMAM's attention mechanism in focusing on text-guided regions, leading to reduced localization accuracy. When all three modules were used together, conflicts arose between DMAM's channel-spatial attention and FDA's dual-activation weights during feature aggregation, excessively suppressing or reinforcing certain key features, thereby weakening the synergistic effect among modules.

Table 3. Ablation experiments on RefCOCO dataset

MAFF	DAMA	FDA	IoU	Pr@50	Pr@60	Pr@70	Pr@80	Pr@90
			73.80	87.48	84.46	77.30	59.15	17.02
✓			73.87	87.70	84.07	77.64	59.54	**17.55**
	✓		74.33	88.37	84.64	77.78	59.73	17.43
		✓	73.90	87.61	84.78	78.06	59.20	16.51
✓	✓		73.81	87.56	84.27	77.64	59.59	16.10
✓		✓	74.23	88.14	84.90	**78.15**	**60.60**	17.36
	✓	✓	**74.48**	**88.88**	85.15	77.92	59.29	16.39
✓	✓	✓	74.34	88.49	**85.24**	77.58	59.64	16.63

The ablation experiment results on the RefCOCO+ dataset are shown in Table 4.

Adding only FDA improved the IoU to 69.71%, making it the most effective single module. This suggests that FDA, through dynamic weight allocation, effectively integrates bimodal features in complex backgrounds and multi-scale target scenarios, significantly optimizing segmentation results.

However, when MAFF and FDA were combined, the IoU dropped to 69.14%, indicating that multi-scale features, without modal alignment, interfered with FDA's dynamic weight allocation, reducing its feature aggregation efficiency. The combination of DMAM and FDA led to a significant IoU drop to 68.19%, demonstrating that DMAM's channel-spatial attention mask and FDA's dual-activation weights caused redundant suppression, weakening the expression of key features.

Table 4. Ablation experiments on RefCOCO+ dataset

MAFF	DAMA	FDA	IoU	Pr@50	Pr@60	Pr@70	Pr@80	Pr@90
			68.52	81.61	77.93	71.06	53.58	13.85
✓			68.97	81.68	78.01	70.99	53.88	14.69
	✓		68.69	81.28	78.31	71.69	53.79	14.79
		✓	69.71	82.73	79.25	72.51	55.27	15.30
✓	✓		68.86	81.49	78.47	71.34	54.61	16.26
✓		✓	69.14	82.01	79.29	72.49	54.99	14.97
	✓	✓	68.19	80.86	77.38	70.66	53.34	14.34
✓	✓	✓	68.99	81.84	78.57	72.04	54.03	15.11

This phenomenon was most pronounced in the RefCOCO+ dataset, as its textual descriptions rely more on appearance features, and DMAM's cross-modal alignment mechanism, without dynamic aggregation support, failed to effectively suppress background interference, leading to attention imbalance. When all three modules were used together, the IoU was 68.99%, only slightly above the baseline. In the RefCOCO+ dataset, FDA's feature aggregation dominance was more evident, while the introduction of MAFF and DMAM led to feature regulation conflicts, limiting overall performance improvement.

The ablation experiment results on the G-Ref dataset are shown in Table 5.

Adding only DMAM improved the IoU to 61.80%, making it the most effective single module. This indicates that DMAM's channel-spatial attention mechanism effectively handles the extensive positional relationships and composite semantic information in the dataset's textual descriptions, accurately aligning key text-image features, suppressing background interference, and highlighting target regions, significantly improving localization accuracy in complex semantic scenarios.

Table 5. Ablation experiments on G-Ref dataset

MAFF	DAMA	FDA	IoU	Pr@50	Pr@60	Pr@70	Pr@80	Pr@90
			60.39	69.63	64.22	55.38	40.01	13.76
✓			61.12	70.83	65.19	56.92	41.16	13.31
	✓		61.80	71.84	66.51	58.58	42.67	14.56
		✓	61.15	70.87	65.90	57.94	42.61	13.76
✓	✓		61.45	71.50	65.94	58.12	42.76	13.96
✓		✓	60.99	70.59	65.85	57.65	42.02	13.71
	✓	✓	60.84	70.62	65.13	56.71	42.01	13.84
✓	✓	✓	61.65	71.43	66.43	58.05	42.78	13.98

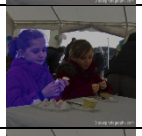
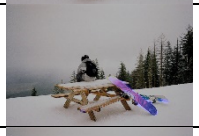
However, when MAFF and DMAM were combined, the IoU dropped to 61.45%, as multi-scale features introduced redundant information without semantic alignment, interfering with DMAM's focus on text-guided regions. The combination of FDA and DMAM resulted in an IoU drop to 60.84%, as FDA's dynamic weight allocation relies on aligned features, and DMAM's attention mask, without dynamic aggregation support, struggled to optimize feature complementarity, leading to performance degradation. When all three modules were used together, the IoU reached 61.65%, outperforming the pairwise combinations. This suggests that the multi-scale features provided by MAFF offer a richer input foundation for FDA's dynamic aggregation, while DMAM's cross-modal alignment mechanism provides semantic priors for FDA's weight allocation. Their synergy mitigates feature regulation conflicts, achieving a certain degree of performance recovery.

4.4.2. Visualization of Segmentation Results

To intuitively demonstrate the effectiveness of the MAFF, DMAM, and FDA modules in referring image segmentation, this paper visualizes the segmentation results of the baseline model and its combinations with different modules (baseline+MAFF, baseline+DMAM, baseline+FDA, and the integration of all three modules), as shown in Table 6. The baseline model produces relatively coarse segmentation results with unclear object boundaries and exhibits significant mis-segmentations in complex backgrounds or small object regions. With the addition of the MAFF module alone, the model enhances its ability to capture fine-grained details by effectively extracting multi-scale features, resulting in clearer segmentation boundaries and notably improved performance in scenarios with complex backgrounds and multi-scale objects. Incorporating only the DMAM module enables more precise alignment between image and text features, allowing the model to better understand the semantic correspondence between textual descriptions and visual content, thereby further refining segmentation results and reducing mis-segmented regions. When the FDA module is added independently, the model leverages dynamic weight generation and cross-scale feature interaction to more effectively integrate feature information across different scales, enhancing the accuracy and robustness of

segmentation—particularly for objects with intricate textures and complex shapes. Ultimately, the simultaneous integration of all three modules yields the best segmentation performance. The synergistic effect of these modules significantly advances referring image segmentation, enabling the model to achieve more precise results across diverse scenarios, with clear object boundaries, substantially reduced mis-segmentations, and improved adaptability to objects of varying scales and complex backgrounds.

Table 6. Ablation study segmentation results

Text input	Left picture guy in green	Left zebra	Kid without bangs in her face	Big bear	A yellow and red snowboard
Original image					
baseline					
+MAFF					
+DMAM					
+FDA					
+MAFF+DMAM +FDA					

Furthermore, the combination of the MAFF and DMAM modules improves the IoU to 61.45%, while the integration of all three modules further raises the IoU to 61.65%. This result indicates that in the G-Ref dataset, DMAM plays a pivotal role in handling complex linguistic expressions, and although the synergy among the other modules contributes to marginal performance gains, the overall improvement is limited. This is attributable to the inherently more complex linguistic expressions in the G-Ref dataset, where DMAM's semantic alignment capability assumes a dominant role, rendering the optimization effects of the other modules relatively minor.

5. Summary

This paper proposes a referring image segmentation method based on a chain-like synergistic mechanism for multi-scale feature extraction and dynamic aggregation. By constructing a collaborative framework comprising a Multi-scale Adaptive Feature Fusion module (MAFF), a Dual-attention Mechanism for Multimodal Alignment module (DMAM), and a Feature Dynamic Aggregation module (FDA), the proposed method enhances multi-scale feature representation, dynamically optimizes feature weight allocation, and improves cross-modal feature interaction efficiency, thereby achieving more precise image segmentation. Experimental results and visual analysis of segmentation outcomes on three mainstream datasets—RefCOCO, RefCOCO+, and G-Ref—demonstrate that the proposed method effectively captures fine-grained details, strengthens the alignment between image and text features, and addresses issues such as insufficient multi-scale adaptive feature fusion, unidirectional cross-modal feature alignment, and static feature aggregation, thus validating its effectiveness. Future work will focus on optimizing the segmentation model to achieve higher accuracy and enable broader application to other vision tasks.

Acknowledgments

The authors gratefully acknowledge the financial support from the National Natural Science Foundation of China (Grant No. 62076246) and the Fundamental Research Funds for the Central Universities (Grant No. 2024JKF12).

References

[1] Hu R, Rohrbach M, Darrell T. Segmentation from natural language expressions[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 108-124.

[2] Huang S, Hui T, Liu S, et al. Referring image segmentation via cross-modal progressive comprehension[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10488-10497.

[3] Li M, Sigal L. Referring transformer: A one-step approach to multi-task visual grounding[J]. Advances in neural information processing systems, 2021, 34: 19652-19664.

[4] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]// International conference on machine learning. PmLR, 2021: 8748-8763.

[5] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]// Proceedings of the IEEE/CVF international conference on computer vision. 2023: 4015-4026.

[6] Mao J, Huang J, Toshev A, et al. Generation and comprehension of unambiguous object descriptions[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 11-20.

[7] Ouyang S, Wang H, Xie S, et al. SLViT: Scale-Wise Language-Guided Vision Transformer for Referring Image Segmentation [C]// IJCAI. 2023: 1294-1302.

[8] Xu Z, Chen Z, Zhang Y, et al. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 17503-17512.

[9] Rohrbach A, Rohrbach M, Hu R, et al. Grounding of textual phrases in images by reconstruction[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 817-834.

[10] Liu S, Hui T, Huang S, et al. Cross-modal progressive comprehension for referring segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 4761-4775.

[11] Yang Z, Wang J, Tang Y, et al. Lavt: Language-aware vision transformer for referring image segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 18155-18165.

[12] Kim N, Kim D, Lan C, et al. Restr: Convolution-free refer-ring image segmentation using transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 18145-18154.

[13] Feng G, Hu Z, Zhang L, et al. Encoder fusion network with co-attention embedding for referring image segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15506-15515.

[14] Jing Y, Kong T, Wang W, et al. Locate then segment: A strong pipeline for referring image segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9858-9867.

[15] Feng G, Hu Z, Zhang L, et al. Bidirectional relationship inferring network for referring image localization and segmentation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 34(5): 2246-2258.

[16] Wang Z, Lu Y, Li Q, et al. Cris: Clip-driven referring image segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11686-11695.

[17] Zhang M, Liu Y, Yin X, et al. MARIS: Referring Image Segmentation via Mutual-Aware Attention Features[J]. arXiv preprint arXiv:2311.15727, 2023.

[18] Sun L, Dong J, Tang J, et al. Spatially-adaptive feature modulation for efficient image super-resolution[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 13190-13199.

[19] Huang H, Chen Z, Zou Y, et al. Channel prior convolutional attention for medical image segmentation. arXiv 2023[J]. arXiv preprint arXiv:2306.05196.

[20] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural net-works[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11534-11542.

[21] Xie L, Li C, Wang Z, et al. Shisrcnet: Super-resolution and classification network for low-resolution breast cancer histopathology image[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2023: 23-32.

[22] Kazemzadeh S, Ordonez V, Matten M, et al. Referitgame: Referring to objects in photographs of natural scenes[C]// Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 787-798.

[23] Yu L, Poirson P, Yang S, et al. Modeling context in referring expressions[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer International Publishing, 2016: 69-85.

[24] Liu C, Lin Z, Shen X, et al. Recurrent multimodal interaction for referring image segmentation[C]//Proceedings of the IEEE international conference on computer vision. 2017: 1271-1280.

[25] Margffoy-Tuay E, Pérez J C, Botero E, et al. Dynamic multimodal instance segmentation guided by natural language queries[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 630-645.

[26] Li R, Li K, Kuo Y C, et al. Referring image segmentation via recurrent refinement networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5745-5753.

[27] Yu L, Lin Z, Shen X, et al. Mattnet: Modular attention network for referring expression comprehension[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1307-1315.

[28] Liu D, Zhang H, Wu F, et al. Learning to assemble neural module tree networks for visual grounding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4673-4682.

[29] Ye L, Rochan M, Liu Z, et al. Cross-modal self-attention network for referring image segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10502-10511.

[30] Chen Y W, Tsai Y H, Wang T, et al. Referring expression object segmentation with caption-aware consistency[J]. arXiv preprint arXiv:1910.04748, 2019.

[31] Yang S, Xia M, Li G, et al. Bottom-up shift and reasoning for referring image segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 11266-11275.

[32] Luo G, Zhou Y, Ji R, et al. Cascade grouped attention net-work for referring expression segmentation[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1274-1282.