DCMFNet: Deep Cross-Modal Fusion Network for Referring Image Segmentation with Iterative Gated Fusion

Zhen Huang*
Mingcheng Xue*
huangzhen@mail.dlut.edu.cn
xmc_andy@163.com
Dalian University of Technology
Dalian, Liaoning, China

Yu Liu
yuliu@dlut.edu.cn
Dalian University of Technology
Dalian, China

Kaiping Xu
xkp13@tsinghua.org.cn
Dalian University of Technology
Dalian, Liaoning, China

Jiangquan Li
ljqwangyibiu@163.com
Dalian University of Technology
Dalian, Liaoning, China

Chenyang Yu
yu736314362@gmail.com
Dalian University of Technology
Dalian, Liaoning, China

ABSTRACT

Cross-modal fusion aims to establish a consistent correspondence between arbitrary modalities. Due to the inherent differences between these modalities, accurately modeling their correspondence is a challenging task. Referring image segmentation (RIS) is a fundamental cross-modal task that intends to segment a desired object from an image based on a given natural language expression. In this paper, we propose an efficient algorithm called the Deep Cross-Modal Fusion Network (DCMFNet) to address this challenge. The proposed algorithm leverages the contextual information from linguistic context to guide the modeling of the visual context, gradually highlighting the referent in the image. The network architecture employs an innovative fusion strategy known as Iterative Gated Fusion (IGF) to capture the consistency relationship between multi-modal features. IGF iteratively adjusts the relative importance of features at each level based on high-level semantics, emphasizing the shared information while suppressing the irrelevant parts. Specifically, IGF consists of cascaded fusion units and gating units. The fusion units integrate high-level semantics with the features from the previous layer to enhance the representation. The gating units perceive the discrepancy between the enhanced features and the original representation, and selectively weight and integrate the important features for further refinement. Through multi-layer iterative optimization, IGF gradually establishes a fine-grained correspondence between arbitrary modalities. Extensive experimental results on the Referring Image Segmentation task demonstrate the effectiveness and utility of the proposed method.

*Both authors contributed equally to this research.

Graphics Interface 2024, June 03–06, 2024, Barrington, Halifax

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Human-centered computing → Text input**; **Text input**; • **Computing methodologies → Computer vision**;

KEYWORDS

referring image segementation, novel fusion strategy, cross-modal fusion, vision and language, context modeling

ACM Reference Format:
Zhen Huang, Mingcheng Xue, Yu Liu, Kaiping Xu, Jiangquan Li, and Chenyang Yu. 2024. DCMFNet: Deep Cross-Modal Fusion Network for Referring Image Segmentation with Iterative Gated Fusion. In *Proceedings of Graphics Interface 2024*. ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Referring image segmentation (RIS) is a challenging multimodal task involving computer vision and natural language processing. This task requires a comprehensive understanding and accurate modeling of the correspondence between vision and language to correctly segment the particular object described by a natural language expression in the image. Unlike the traditional semantic segmentation problem [3, 5, 38, 44] that aims to classify each pixel into predefined labels, referring image segmentation is not confined to predefined categories and makes the pixel-level prediction based on categories contained in natural language expressions. Similar to many interesting scene understanding problems that combine visual and linguistic data for reasoning such as vision-language navigation [50], visual question answering [1, 14, 61], cross-modal retrieval [7, 36, 40], etc., the RIS problem shows the potential way to use language to guide an intelligent body to interact with the environment, which has a wide range of application scenarios such as interactive image editing [8], language-driven human-computer interaction [47], etc. The RIS task has gained wider scholarly attention in recent years, and some existing works [10, 12, 19, 54] have achieved excellent performance. It is worth noting that there are two major challenges to further address in this task, one of which is how to establish a more consistent visual-linguistic correspondence so that the referent can be accurately identified in complex visual and linguistic scenarios, and the other is how to capture more

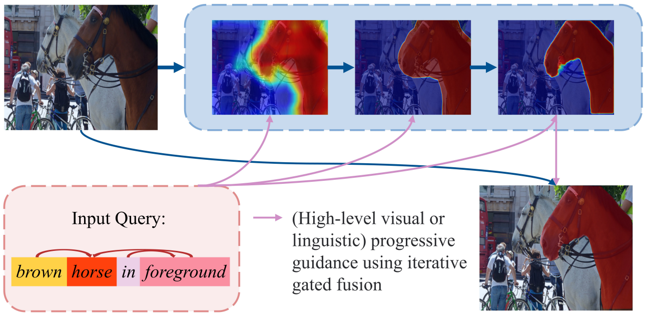

Figure 1: Illustration of our proposed deep cross-modal fusion network for referring image segmentation. Given an input image and a natural language expression, the proposed model first exploits linguistic context to contiguously guide visual context modeling to build the consistent correspondence between vision and language, which progressively highlights the referent. Then the model refines the prediction mask of the referent by utilizing high-level visual features to guide the integration of low-level visual features. Noteworthy, the fusion processes of multi-modal and multi-level visual features are both done by the proposed iterative gated fusion.

valid visual information related to the referent to refine the final prediction mask.

Benefiting from the development of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), we can gain a deeper understanding of vision and language and promote building more consistent relationships between modalities with the powerful learning capabilities of these networks. Prior methods to solve the RIS task [17, 29, 33, 42] concatenate visual features, spatial coordinates, and linguistic features to obtain the multimodal feature representation and then rely on the deep learning models to learn a particular correspondence between co-embedded elements. The main limitation of this approach is that it can only model the shallow interaction between vision and language and cannot accurately predict the segmentation mask. Recently, some referring segmentation works [10, 19, 21, 22, 39, 57] utilize the attention mechanisms to model the visual-linguistic interaction. These approaches can roughly divide into two types, i.e., modular attention networks [22, 39, 57] and attention-reasoning structure networks (e.g., attention-graph structured reasoning [13, 19], attention-multimodal tree reasoning [21], and attention-based cross-modal transformer [10]). These attention-based methods deepen the interaction between vision and language and achieve remarkable results. However, the single-layer attention fusion strategy adopted by most of them still has limitations in modeling multimodal feature interactions, that is, lack of deep interaction and underutilization of language guidance. Specifically, these attention fusion strategies use a single-layer structure to model the correspondence between modalities by computing an attention map to update the modalities. But such single-layer fusion strategies may lead to inaccurate relationship modeling and make it difficult to fully utilize language features that provide crucial prompts for deep interaction. Furthermore, previous referring segmentation works have rarely explored the linguistic context to guide visual context modeling in the encoder, which ignores the potential of encoders to align multimodal features. We believe that high-level visual and linguistic features with sufficient semantic information can interact to form a common semantic space that can be learned by encoders in favor of highlighting the spatial regions associated with the referent.

Refining the prediction mask is another major challenge for the referring image segmentation task. After constructing a consistent correspondence between vision and language, the model typically generates a fine-grained multimodal feature that implicitly highlights the spatial region where the referent resides. In order to obtain a more accurate prediction mask, it is necessary to supplement the visual information related to the referent. A popular approach in this field is to integrate multi-level visual features. Some previous works (e.g., CMSA [57], CMPC [19]) individually and repeatedly process visual features at different levels and then use gating mechanisms to aggregate the multi-level visual information, which seriously increases the computational cost. Some recent works (e.g., BUSNet [54], BRINet [18], LSCM [21]) adopt bidirectional (i.e., bottom-up and top-down) pathways to fuse multi-level visual features using attention or gating mechanisms, which are also prone to the redundant computation. Unlike these methods, we introduce a novel iterative gated fusion to integrate multi-level visual context on a simple single path, which reduces redundant calculations efficiently and is proven utility.

To address the limitations of the above methods, we propose a deep cross-modal fusion network (DCMFNet) to build more consistent corresponding relations between modalities and thus improve segmentation performances. Figure 1 shows an example that illustrates the deep cross-modal fusion network, where language continuously guides visual context modeling in the encoder and high-level visual features guide low-level visual features to supplement detail information, all of which are fused with iterative gated fusion, resulting in the referent being gradually highlighted and refined. The proposed iterative gated fusion (IGF) strategy employs a multi-layer structure to deepen the interaction between modalities, with a bidirectional fusion unit and an adaptive gating unit (ASGate) embedded in each layer to dynamically reconcile the relative strength of features in each spatial region according to high-level semantics so as to highlight the referent and suppress the others. Specifically, the gating module adaptively selects spatial regions of high-level semantic concern, and the fusion unit builds long-range dependencies between modalities to weight features. Through intra-layer and inter-layer iterative optimization, IGF gradually builds consistent correspondence between modalities, which helps the DCMFNet generate the accurate segmentation mask.

Our main contributions can be summarized as follows:

- We propose a deep cross-modal fusion network (DCMFNet) for the referring image segmentation task. DCMFNet fully exploits the potential of high-level semantic guidance and encoders to build consistent correspondence between modalities, thereby improving segmentation performances.
- We propose a novel fusion strategy called Iterative Gated Fusion (IGF), which can deeply fuse multi-modal and multi-level contextual information.

- The proposed method outperforms many previous state-of-the-art methods on multiple referring image segmentation datasets. Extensive experiment results demonstrate the effectiveness and utility of the proposed method.

2 RELATED WORK

Referring image segmentation aims to segment the specific object corresponding to a natural language expression in an image. Different from traditional unimodal semantic segmentation [3, 5, 38, 44] and instance segmentation [35], the key for referring image segmentation is to learn a particular correspondence between modalities by building the deep interaction between vision and language.

Pioneering referring image segmentation methods [18]– [21] usually adopt a concatenation-convolution scheme, which typically relies on deep learning models to learn the particular correspondence between vision and language. LSTM-CNN [18] directly concatenates the visual feature map with spatial coordinates and linguistic features and then feeds the com- bined features into a fully convolutional network (FCN) [1] to generate the segmentation mask. Later, recurrent interaction fusion strategy is introduced in [19]–[21], RMI [20] mim- ics human decision-making process to progressively perform cross-modal interaction in a word-reading order. DMN [21] exploits the recursive nature of language to integrate visual feature maps in multiple steps. In RRN [19], the concatenated multimodal features top-down integrate multi-level features to refine the segmentation mask.

Recently, some works consider attention mechanisms to learn the correspondence between vision and language. These attention-based methods can be roughly divided into two types: modular attention networks and attention-reasoning structure networks. The former updates the corresponding modalities by calculating the correlation between multiple modalities, and the latter uses the reasoning structure combined with attention to aggregate the global context. For example, CMSA [57] introduces the self-attention mechanism to capture long-term dependencies for adaptively focusing on informative words and important regions. ESE-FN [48] learns modal and channel-wise Expansion-Squeeze-Excitation (ESE) attentions for attentively fusing the multi-modal features in the modal and channel-wise ways. MMSA [16] designs multiperspective and hierarchical fusion modules to perform mutual attention fusion. KWA [45] adopts a vision-guided linguistic attention mode to learn the importance of words to each spatial region. BRINet [18] explores the bidirectional guidance between visual and linguistic features. Cross-image attention is introduced in [22] for enhancing visual cues. EFN [12] transforms the vision encoder into a multimodal feature learning network. In addition, since the graph and tree structures can represent data relationships, some works introduce them to perform multimodal reasoning combined with attention. CMPC [19] constructs a multimodal graph and utilizes the graph convolution to reason among vertexes for highlighting the referent. Language graph or tree structures parsed from the expression [55] are introduced in NMTree [34], LSCM [21], and BUSNet [54], they take the dependencies among words as prior knowledge to restrict the communication among word nodes for modeling valid multimodal context. More recently, VLT [10] and ReSTR [25] introduce the cross-modal transformer to build the deep interaction between multimodal features at the decoding stage, achieving state-of-the-art performances. Recent work SAM [27] excels at producing high-quality masks by leveraging diverse prompts like points or boxes. Unlike traditional SAM models that require large-scale training, our proposed approach enables precise mask generation with small-scale training. This achievement is attributed to our innovative modeling strategy, i.e., globally, we explore using language to guide visual encoding in the encoder and using the high-level feature to guide low-level feature integration in the decoder, and locally, we establish the deep interaction between guidance and guided features by embedding the iterative gated fusion module in the network.

In this paper, we exploit the potential of high-level semantic guidance and encoders to establish consistent correspondences between modalities. Furthermore, we also propose a novel and practical iterative gated fusion module capable of deeply integrating multi-modal and multi-level contextual information.

3 METHOD

The overall architecture of the proposed network is illustrated in Figure 2. Given an input image and a natural language expression, we first extract linguistic features from the text encoder and then embed linguistic features into different stages of the vision encoder via the proposed iterative gated fusion (IGF) module to guide the visual context modeling. Each iteration of the gated fusion module generates a finer feature that more precisely highlights the referent. The feature output by the IGF module will be fed into the next visual encoding stage for the encoder to learn the correspondence of multimodal contexts. To clarify the boundary of the referent and generate an accurate mask, we supplement the visual detail information to the spatial regions where the referent resides. We first extract the multi-scale contextual information via the Atrous Spatial Pyramid Pooling (ASPP) module [6] and then utilize the generated high-level semantic features to guide the integration of low-level visual features via the IGF module. In the following sections, we elaborate on the design of the iterative gated fusion module in Section 3.1, language-guided visual encoding in Section 3.2, and decoder in Section 3.3.

3.1 Iterative Gated Fusion

The iterative gated fusion (IGF) module is a simple but effective deep fusion module, which progressively deepens the interaction between the guidance features and guided features within multiple iteration steps. In this work, the guidance features refer to the features connected by the pink line in Figure 2 (e.g., linguistic feature L , high-level visual feature M_{dec}), and the guided features refer to the features connected by the blue line (e.g., low-level visual feature V_1 and high-level visual feature V_3, V_4).

The details of the iterative gated fusion module are depicted in Figure 3. Given the guidance feature X and guided feature $Y \in \mathbb{R}^{C_v \times H \times W}$, we first resize them to keep the spatial size consistent and then feed the feature maps into the 1×1 convolution layer respectively to obtain the initial inputs of the iterative process $x \in \mathbb{R}^{C_l' \times H \times W}$ and $G_0 \in \mathbb{R}^{C_v' \times H \times W}$, where C_l', C_v', H, and W denote the channel numbers, height, and width of the initial inputs, respectively. Then, the input features x and G_0 are passed into the

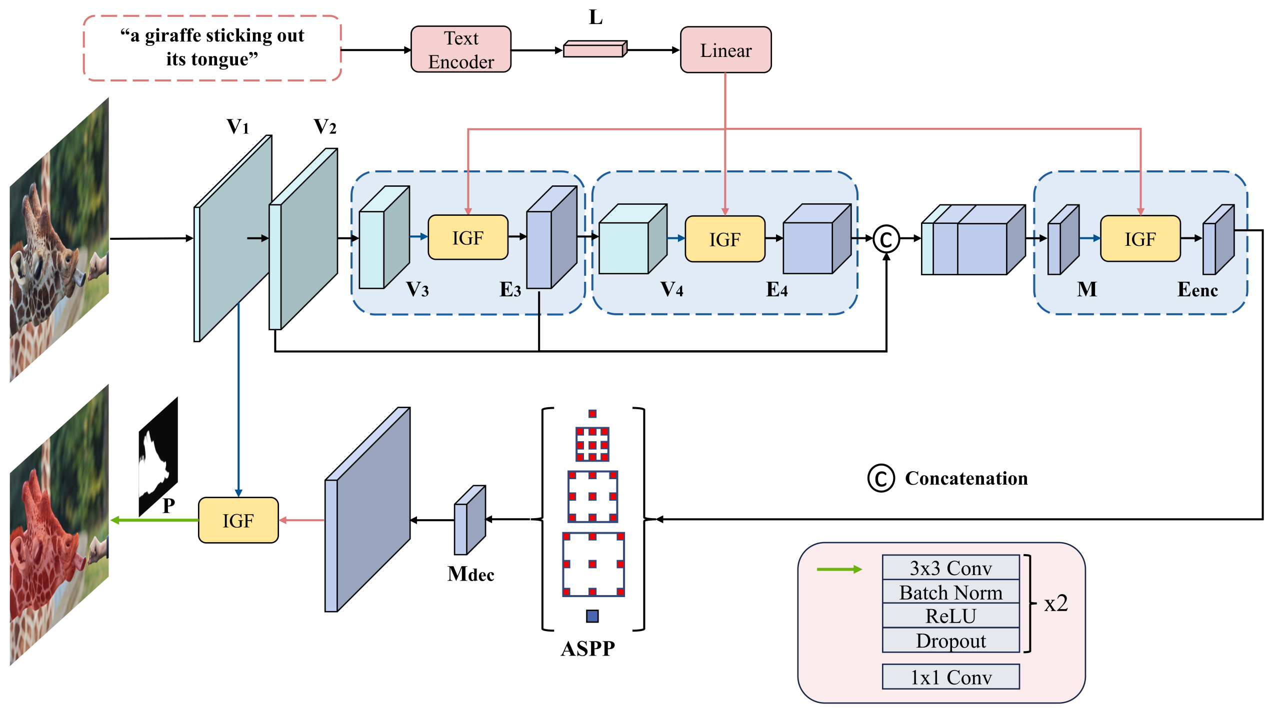

Figure 2: An overview of our approach. The proposed method consists of two stages including language-guided visual encoding and multi-level visual feature fusion. In the first stage, we use a text encoder to extract the language features L and then selectively embed language features into different stages of visual encoding to guide visual context modeling and exploit the encoder to learn multi-modal feature representations. In the first stage, the network models the consistent correspondence between vision and language and generates the fine features E_{enc} that highlight the spatial region where the referent resides. To obtain a more accurate prediction mask, we feed E_{enc} into ASPP [6] to extract multi-scale information and supplement the visual detail information to the referent based on the indication of the highlighted spatial region (i.e., using the high-level visual feature M_{dec} to guide the integration of the low-level visual feature V_1). The constructed network adopts a novel and practical iterative gated fusion (IGF) to unify the multi-modal and multi-level visual feature fusion.

IGF layers cascaded in depth (denoted $\text{IGF}^{(1)}, \text{IGF}^{(2)}, ..., \text{IGF}^{(L)}$) to perform deep interaction. The interaction at the t-th time step occurs between x and G_{t-1}, given by:

$$F_t, G_t = \text{IGF}^{(t)}(x, G_{t-1}), \tag{1}$$

where $F_t \in \mathbb{R}^{C'_v \times H \times W}$ and $G_t \in \mathbb{R}^{C'_v \times H \times W}$ are hidden states updated by the fusion unit and gated unit respectively with taking the current input x and previous hidden state G_{t-1} as inputs.

Fusion Unit. Each IGF layer first uses a multi-head bilinear fusion [2] to associate the current input x and the previous hidden state G_{t-1}, which models the long-range dependencies between modalities to enhance the response of spatial regions related to guidance features and weaken the others. The structure of the fusion unit as shown in Figure 4 The update process of the fusion unit is formulated as follows:

$$F_t = \tau(\sum_{i=1}^{5}(\tau(W_1 x) \odot \tau(W_2 G_{t-1}))), \tag{2}$$

where $W_1 \in \mathbb{R}^{C'_v \times C'_l}$ and $W_2 \in \mathbb{R}^{C'_v \times C'_v}$ are weight matrices for linear transformation, $\odot$ denotes the element-wise multiplication, $\tau(\cdot)$ denotes the tanh function, Σ denotes integrating the multi-head output features, i.e., stacking and summing along the channel dimension.

The guided features can deeply fuse with guidance features within multiple rounds of fusion processing. However, the fusion process may introduce noise (e.g., unimportant semantic information and spatial details irrelevant to the referent). Therefore, we propose an adaptive selection gate (ASGate), which can dynamically select useful information and suppress interference caused by noise to achieve the signal response shift toward the referent-related regions.

Gating Unit. Figure 5 shows the structure of the proposed Adaptive Selection Gate (ASGate). The gating unit takes the fused feature F_t generated by the current fusion unit and the hidden feature G_{t-1} generated by the previous gating unit as input. In the gating unit, the input features are first integrated and fed into a convolutional layer with a sigmoid function to form a learnable referent-aware weight matrix to weighted the fused feature F_t. The aware matrix assigns higher weights to spatial regions with high-level semantic interest and low weights to noisy features. Then, the gating unit perceives the difference between the feature F_t and the weighted

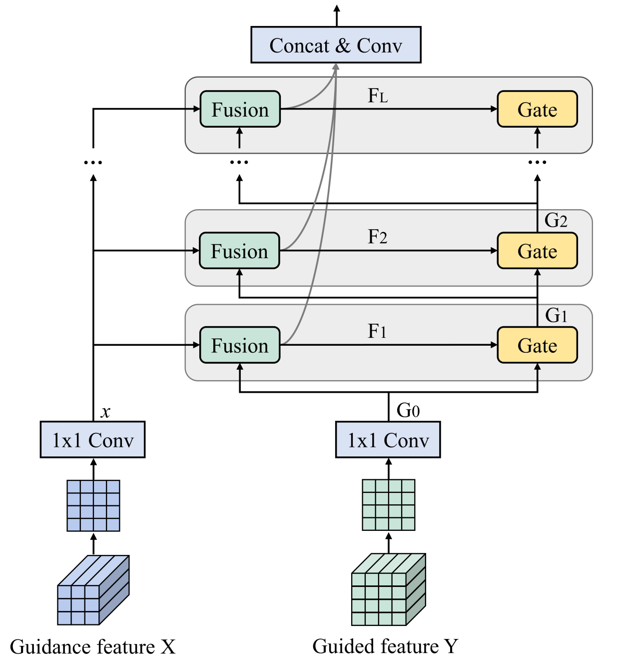

Figure 3: Illustration of the proposed Iterative Gated Fusion (IGF) layers. IGF takes the guidance feature X and guided feature Y as inputs, first adjusts their spatial size and channel number, then deepens the interaction between input features through the fusion-gate scheme within L times optimization, and finally aggregates the output features of each layer as the output of IGF. In the multi-layer structure, the guidance feature X (e.g., linguistic features, high-level visual features) can continuously guide the guided features Y (e.g., low-level visual features) to build a more consistent correspondence between modalities.

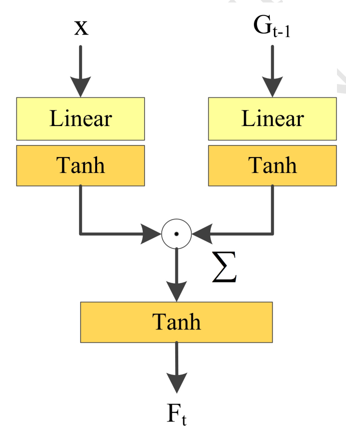

Figure 4: The structure of the fusion unit.

feature f_t and aggregates the difference region through the concatenation and convolution operations followed by a non-linear activation function. Finally, the generated feature G_t will be used

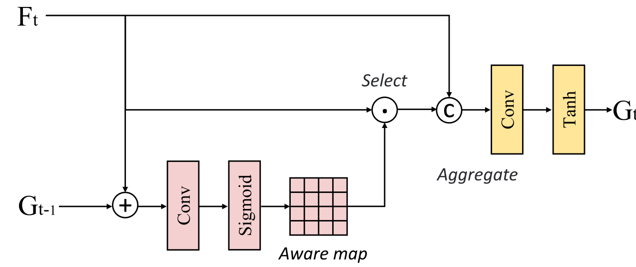

Figure 5: Illustration of the proposed Adaptive Selection Gate (ASGate). ASGate is an implementation of the gating unit in iterative gating fusion, which can perceive, select, and aggregate the important features corresponding to the spatial regions concerned by guidance features.

as the input feature for the next round of optimization. The gating process of the adaptive selection gate to generate the hidden feature G_t can be formulated as follows:

$$\begin{aligned} z_t &= \sigma(W_z(F_t + G_{t-1}) + b_z), \\ f_t &= F_t \odot z_t, \\ G_t &= \tanh(W_g([F_t; f_t]) + b_g), \end{aligned} \tag{3}$$

where $\sigma(\cdot)$ denotes the sigmoid function. [;] denotes the concatenation operation. $W_z \in \mathbb{R}^{C'_v \times C'_v}$ and $W_g \in \mathbb{R}^{(C'_v + C'_v) \times C'_v}$ represent the learnable parameters of the 3×3 convolution operations. b_z and b_g are biases.

Multi-layer Progressive Interaction. The IGF module adopts a multi-layer progressive interaction strategy. The guidance features deeply fuse with guided features within L times fusing and gating iterations. The IGF module integrates the fused features $F^{(L)} = [F_1, F_2, ..., F_L]$ generated by L layers as the output. The process of generating the output feature of the IGF module can be formulated as follows:

$$IGF^* = \mathrm{Conv}([F_1; F_2; ...; F_L]), \tag{4}$$

where $IGF^* \in \mathbb{R}^{C_v \times H \times W}$ is the output feature of IGF layers, its shape is the same as the guided feature Y. $\mathrm{Conv}(\cdot)$ denotes the 3×3 convolution operation.

3.2 Language-Guided Visual Encoding

In this section, we elaborate on the design of the vision encoder in DCMFNet. To efficiently utilize the encoder to model valid multi-modal context, we mainly consider the following points: (1) Differential processing of high-level and low-level visual features: During the visual encoding, high-level visual features contain rich semantic information, suitable for interacting with linguistic features, while low-level visual features own amounts of spatial detail information that are suitable for refining the identified referent. (2) Multiple-step guided encoding: To model the deep interaction between vision and language, we selectively embed IGF layers into different stages of the vision encoder, realizing the contiguous guidance of language to visual encoding from local to the whole.

As shown in Figure 2, given a natural language expression, we first employ a language encoder to extract the linguistic feature $L \in \mathbb{R}^{C_L}$ and then apply a linear layer to map it to $L \in \mathbb{R}^{C_l}$. For

Table 1: Comparison with the state-of-the-art methods on four datasets using overall IoU as metric. "-" denotes no data available. DCRF denotes DenseCRF [28] post-processing.

	Vision encoder	UNC			UNC+			G-ref	ReferIt
		val	testA	testB	val	testA	testB	val	test
RMI [33]	ResNet-101	44.33	44.74	44.63	29.91	30.37	29.43	34.40	57.34
RRN+DCRF [29]	ResNet-101	55.33	57.26	53.95	39.75	42.15	36.11	36.45	63.63
MAttNet [58]	Res101-MRCN	56.51	62.37	51.70	46.67	52.39	40.08	-	-
NMTree [34]	Res101-MRCN	56.59	63.02	52.06	47.40	53.01	41.56	-	-
CMSA+DCRF [57]	ResNet-101	58.32	60.61	55.09	43.76	47.60	37.89	39.98	63.80
STEP [4]	ResNet-101	60.04	63.46	57.97	48.19	52.33	40.41	46.40	64.13
CGAN [39]	ResNet-101	59.25	62.37	53.94	46.16	51.37	38.24	46.54	-
BRINet+DCRF [18]	ResNet-101	61.35	63.37	59.57	48.57	52.87	42.13	48.04	63.46
LSCM+DCRF [21]	ResNet-101	61.47	64.99	59.55	49.34	53.12	43.50	48.05	66.57
CMPC+DCRF [19]	ResNet-101	61.36	64.53	59.64	49.56	53.44	43.23	49.05	65.53
SANet [31]	ResNet-101	61.84	64.95	57.43	50.38	55.36	42.74	44.53	65.88
TV-Net [22]	ResNet-101	61.87	65.61	60.10	50.30	54.43	43.52	49.92	65.38
BUSNet [54]	ResNet-101	62.56	65.61	60.38	50.98	56.14	43.51	49.98	-
EFN [12]	ResNet-101	62.76	65.69	59.67	51.50	55.24	43.01	51.93	66.70
DCMFNet-Res101 (Ours)	ResNet-101	**65.84**	**69.34**	**63.09**	**54.78**	**60.03**	**49.30**	**51.99**	**66.74**
ReSTR [25]	Transformer	67.22	69.30	64.45	55.78	60.44	48.27	54.48	70.18
DCMFNet-Trans (Ours)	Transformer	**71.00**	**73.49**	**67.17**	**60.55**	**66.34**	**52.18**	**57.79**	**68.36**

an input image, we first employ a four-stage vision encoder to extract the visual features $\{V_1, V_2, V_3\}$ of the first three stages, where $V_i \in \mathbb{R}^{C_v^i \times H_i' \times W_i'}$, $i \in \{1, 2, 3\}$, with C_v^i, H_i', and W_i' being the channel number, height, and width of the visual feature map at i-th stage, respectively. Then, we insert the IGF layers to perform the deep interaction between the high-level visual feature V_3 and linguistic feature L to obtain the input feature E_3 of the next encoding stage. Next, in the fourth encoding stage, we again insert the IGF layers to realize the second guidance of language to visual encoding, obtaining the visual feature E_4. To further align linguistic features with multi-level visual features, we incorporate the encoding features V_2, E_3, and E_4 to form a multi-level visual feature M and fuse it with L via the IGF layers forming the deeply fused feature E_{enc}. Finally, the low-level visual feature V_1 and high-level visual feature E_{enc} are input to the decoder for predicting the mask P. The process by which linguistic features guide visual encoding can be formulated as follows:

$$\begin{aligned} E_3 &= \mathrm{IGF}(L, V_3), \\ E_4 &= \mathrm{IGF}(L, \mathrm{Encoder}(E_3)), \\ M &= \mathrm{ConvBN}([V_2 \downarrow; E_3 \downarrow; E_4]), \\ E_{enc} &= \mathrm{IGF}(L, M), \\ P &= \mathrm{Decoder}(V_1, E_{enc}), \end{aligned} \tag{5}$$

where ConvBN denotes the 3×3 convolutional layer attached with a batch normalization layer. $\downarrow$ denotes the downsampling operation. E_3, E_4, and E_{enc} represent the language-guided visual encoding features.

3.3 Decoder

Multi-Level Feature Fusion. It has become a common idea in RIS that multi-level feature fusion can improve segmentation results. Unlike previous works, our whole process does not have repeated treatment and only uses the IGF module to complete the multimodal and multi-scale feature aggregation. At the decoding stage, we first take the visual feature E_{enc} from the encoder fed into the ASPP module [6] to capture multi-scale context, forming the feature M_{dec}. Then the high-level guidance feature M_{dec} is resized to the spatial size of the low-level visual feature V_1 and deeply interacts with V_1 via IGF layers to supplement spatial details of the concerned regions. The process of multi-level feature fusion can be formulated as follows:

$$E_{dec} = \mathrm{IGF}(M_{dec} \uparrow, V_1), \tag{6}$$

where $\uparrow$ denotes the upsampling operation.

Segmentation. After obtaining the fine feature E_{dec}, we need a segmentation structure to transform the feature E_{dec} into a segmentation mask. Following [30], we adopt a hierarchical segmentation structure to process the multimodal feature. The segmentation structure consists of two stacked 3×3 convolution and one 1×1 convolution for classifying pixels (represented by the green line in Figure 2). Each 3×3 convolutional layer attaches with batch normalization and ReLU activation, and a Dropout layer.

4 EXPERIMENTS

4.1 Experimental Setup

Datasets. To verify the effectiveness of our proposed method, we conduct extensive experiments on four benchmark datasets for RIS: RefCOCO (UNC) [59], RefCOCO+ (UNC+) [59], G-Ref [41], and ReferIt [24].

The RefCOCO and RefCOCO+ (i.e., UNC and UNC+) datasets are collected from the MS COCO dataset [32]. The RefCOCO dataset contains 19,994 images with 142,209 language expressions for 50,000 objects, and the RefCOCO+ dataset contains 141,564 expressions referring to 49,856 objects in 19,992 images. In the RefCOCO dataset, each image has multiple objects with the same category, and referring expressions are not restricted. The language expressions in the RefCOCO+ dataset contain more appearance information without location information.

The G-Ref dataset is collected from the MS COCO dataset via Amazon's Mechanical Turk and includes 104,560 expressions referring to 54,822 objects in 26,711 images. Compared to the other datasets, G-Ref has longer language descriptions.

The ReferIt dataset is built upon the IAPR TC-12 [11] and comprises 19,894 images with 130,525 language expressions for 96,654 segmented image regions. Its annotations contain not only objects but also stuff (e.g., ground and water).

Implementation Details. Our network is trained using Adam [26] optimizer with an initial learning rate of $1.82e^{-05}$ for the vision encoder and $1.82e^{-04}$ for the rest modules. For feature dimensions, considering the GPU memory limits, we set $C'_v = C_v/2$, $C'_l = C_l = 256$ for the first two IGF layers, $C'_v = C_v = C'_l = 256$ for the third IGF layers, $C'_v = C_v = C'_l = 128$ for the last IGF layers, and $C_L = 768$. We adopt the language representation model [53] as our language encoder. The deeplab ResNet-101 [15] and vision transformer [37] are used as the visual backbone. During the end-to-end training, we set the maximum length of language expression to 15 for RefCOCO, RefCOCO+, ReferIt, and 20 for G-Ref and resize the input image to 480×480. We train our network for 15 epochs with a batch size of 6 on an NVIDIA RTX3090 GPU and use the sigmoid cross-entropy loss as the loss function to guide the network training.

Evaluation Metrics. Following the setup of previous works [10, 22], we use two metrics for the experimental evaluation: Overall Intersection-over-Union (Overall IoU) and Prec@X. The Overall IoU metric calculates the ratio of the total intersection areas and the total union areas between the predicted mask and the ground-truth mask for all test samples, which reflects the overall performance of the proposed methods. The Prec@X metric calculates the percentage of test samples whose IoU exceeds the threshold X, which shows the precision distribution of predicted masks in detail.

4.2 Comparison with State-of-the-art Approaches

To make comparisons as fair as possible, reducing the impact of the vision encoder with different capacities on performances, we compare DCMFNet-Res101 with EFN [12], DCMFNet-Trans with ReSTR [25], respectively. The results in Table 1 show that our method outperforms many other methods on multiple datasets. In particular, DCMFNet-Res101 achieves the average IoU gains of 3.38%, 4.79% on the UNC, UNC+ datasets over EFN. More remarkably, our DCMFNet-Trans achieves the average IoU gains of 3.56%, 4.86%, 3.31% on the UNC, UNC+, G-Ref datasets over ReSTR. We attribute these performance gains to our modeling scheme. Locally, we build the deep interaction between guidance and guided features by embedding the IGF module in the network. Globally, we explore using language to guide visual encoding in the encoder and using the high-level feature to guide low-level feature integration in the decoder.

In addition, we compare DCMFNet-Dark53 with VLT [10], and LTS [23] on three sets using different metrics. Since VLT and LTS methods are implemented by TensorFlow, and the proposed DCMFNet is implemented by PyTorch, we reproduced the DarkNet-53[43] backbone with PyTorch. Although the performance of our reproduced DarkNet-53 is not as good as that used by VLT and LTS, it can be seen from the results in Table 3 that the DCMFNet-Dark53 with a lower-performance backbone has achieved the higher performance than VLT and LTS. Darknet-53 is better than ResNet-101, and it has a similar performance to ResNet-152 [43]. In Table 1, DCMFNet-Res101 with ResNet-101 achieved comparable performance to VLT with DarkNet-53. These results demonstrate the effectiveness of the proposed method.

4.3 Ablation Studies

We conduct extensive experiments on the validation set of the RefCOCO dataset to investigate why DCMFNet is effective.

Language-Guided Visual Encoding. Our language-guided visual encoding divides into two stages. In the first stage (denoted LGVE-1), we selectively embed the IGF layers into the encoding layers of the vision encoder to achieve local guidance of language to visual encoding, formed the visual feature E_i is fed into the next encoding layer for the encoder to learn the multimodal feature representation. In the second stage (denoted LGVE-2), we again embed the IGF layers for further cross-modal alignment to generate the feature E_{enc} for decoding.

In rows 1 to 7 of Table 2, we carefully control variables to verify the effectiveness of language-guided visual encoding. Firstly, we evaluate the first stage of language-guided visual encoding (corresponding to rows 1 to 5 of Table 2, LGVE-1). We can see that the Overall IoU and Prec@X performances of the model with LGVE-1 (rows 3 to 5 of Table 2) are higher than that of the model without co-embedding visual and linguistic features during the visual encoding (row 1 and row 2 of Table 2), which indicates the co-embedding of visual and linguistic elements at the visual encoding stage is beneficial for the network to learn higher-quality multimodal representation.

In addition, we selectively embed IGF layers in different visual encoding layers to further explore the co-embedding of visual and linguistic features at the visual encoding stage. From the results of Table 2, we can see that the model with the IGF layers embedded in the last two visual encoding layers (row 4 of Table 2, Figure 2, Full model) achieves the best metric performances, while the model with the IGF layers embedded in the highest layer (row 3 of Table 2) drops by 2.61% average Prec@X performance and 1.3% Overall IoU performance compared to the full model. Similarly, the model with IGF layers embedded in the last three encoding layers (row 5 of Table 2) drops by the 1.35% average Prec@X performance and 1.27% Overall IoU performance. We consider the reasons for such results are that only co-embeds language features in the highest layer cannot fully exploit the learning ability of the vision encoder, while co-embeds language features at the third from last encoding layer may affect the discrimination of the referent in subsequent visual coding processes due to the visual feature V_2 lacking sufficient semantic information and containing lots of spatial detail information irrelevant to the referent, the visual feature V_3 from the penultimate encoding layer owning more semantic information can interact with language features more effectively, enabling the encoder to learn more accurate multimodal representation.

Table 2: Ablation studies of deep cross-modal fusion network on the RefCOCO val set.

	E_i (LGVE-1)	E_{enc} (LGVE-2)	E_{dec} (MLFF)	Prec@0.5	Prec@0.6	Prec@0.7	Prec@0.8	Prec@0.9	Overall IoU
1	-	-	-	79.32	74.28	65.30	41.39	16.13	66.92
2	-	✓	✓	81.19	77.30	71.43	59.38	30.58	69.19
3	E_4	✓	✓	81.60	77.97	72.16	60.48	31.47	69.70
4	E_3, E_4	✓	✓	**83.81**	**80.30**	**74.78**	**63.83**	**34.02**	**71.00**
5	E_2, E_3, E_4	✓	✓	82.80	79.27	73.45	62.18	32.29	69.73
6	E_3, E_4	-	✓	83.40	80.09	74.59	62.94	33.71	70.74
7	E_3, E_4	✓	-	83.14	79.35	71.85	54.26	20.89	68.66
Ablation studies of Multi-Level Feature Fusion (MLFF):									
8	DCMFNet w/o MLFF			83.14	79.35	71.85	54.26	20.89	68.66
9	with Concat Fusion			83.39	79.72	73.32	60.04	26.91	70.23
10	with Gated Fusion [57]			83.25	79.68	73.57	61.10	28.60	70.40
11	with Gated Bi-directional Fusion [18]			83.63	79.92	74.24	62.03	29.06	70.55
12	with Iterative Gated Fusion (Ours)			**83.81**	**80.30**	**74.78**	**63.83**	**34.02**	**71.00**

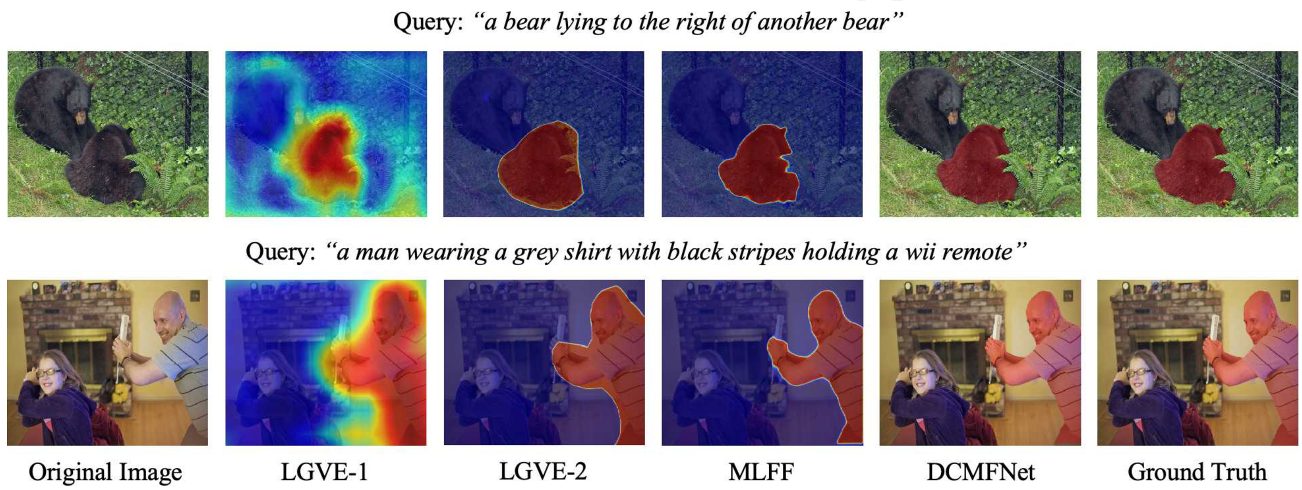

Figure 6: Attention results from different stages of DCMFNet. Note: LGVE-1 and LGVE-2 denote the first and second stages of language-guided visual encoding, respectively, and described in Section 4.3, while MLFF denotes Multi-Level Feature Fusion, described in Section 3.3.

Table 3: Comparison between our method, VLT [10], and LTS [23] using DarkNet-53 [43] as the vision encoder on three sets.

Method	UNC testA	UNC+ testA	G-ref val
LTS [23]	67.76	58.32	-
VLT [10]	68.29	59.20	49.76
DCMFNet-Dark53 (Ours)	68.38	59.24	51.50

Method	Prec@0.5	Prec@0.9	IoU
LTS [23]	78.47	12.92	67.76
DCMFNet-Dark53 (Ours)	82.15	28.18	68.38

Further, we evaluate the second stage of language-guided visual encoding (LGVE-2). We remove the IGF module that aligns language and multi-level visual features (row 6 of Table 2) from the full model, that is, removing the third IGF module in the encoder in Figure 2, resulting in a drop of 0.4% average Prec@X performance and 0.24% Overall IoU performance. Such results are reasonable because the multi-level visual feature M formed by the integration of V_2, E_3, E_4, the introduction of feature V_2 brings more inconsistency information (e.g., irrelevant spatial detail information), while the inconsistency can be reduced by further aligning the language and visual feature.

The above results demonstrate the effectiveness of language-guided visual encoding and IGF layers integrating multimodal features.

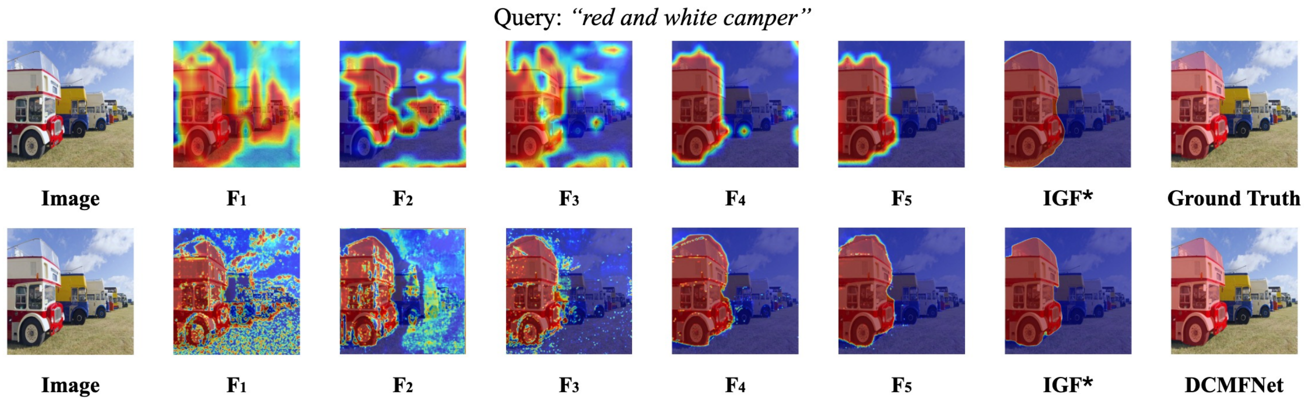

Figure 7: Attention results from different layers of IGF. The attention results of the IGF layers in the first row are from LGVE-2, and those in the second row are from MLFF. Note: F_i is the output feature of the i-th layer of IGF, and IGF^* is the final output feature of the IGF module.

Table 4: Ablation studies of depth L of IGF layers.

L	Overall IoU	Prec@0.5	Prec@0.7	Prec@0.9
1	69.91	82.67	72.99	31.08
2	70.37	83.34	74.66	33.19
3	70.48	83.35	74.01	33.21
4	70.68	83.58	**74.86**	33.39
5	**71.00**	**83.81**	74.78	**34.02**
6	70.36	83.35	74.00	33.01
7	70.33	83.64	74.54	33.14

Table 5: Ablation studies of the output feature IGF^* of IGF layers

Settings	Overall IoU	Prec@0.5	Prec@0.7
Conv(G_5)	70.15	83.39	73.88
Conv(F_5)	70.25	83.50	74.31
Conv($[F_1; F_2; F_3; F_4; G_5]$)	70.30	83.60	74.55
Conv($[F_1; F_2; F_3; F_4; F_5]$)	**71.00**	**83.81**	**74.78**

Table 6: Ablation studies of the gating unit of IGF layers.

Module	Params	Overall IoU	Prec@0.5
ConvLSTMCell [46]	4.72M	70.17	82.94
ConvGRUCell [9]	3.54M	70.61	83.45
ASGate (Ours)	**1.77M**	**71.00**	**83.81**

Multi-Level Feature Fusion. In this ablation study, we further evaluate the multi-level feature fusion (denoted MLFF) component. We first remove the IGF layers in the decoder, that is, removing the last IGF layers in Figure 2, resulting in a drop of 5.45% average Prec@X performance and 2.34% Overall IoU performance (row 7 of Table 2) compared to the full model. Then, we carefully compare the ablation results of rows 2 to 6 of Table 2 with those of row 7 and find that integrating low-level visual features can significantly improve the performances of Prec@0.8 and Prec@0.9 with high thresholds, meaning the proportion of predicted masks that are highly consistent with the ground-truth increases. In addition, we also compare IGF with the previous multi-level feature fusion modules (Concat Fusion, Gated Fusion [57] and Gated Bi-directional Fusion [18]), and the results of Table 2 shows that MLFF with iterative gated fusion achieves better performances. Such results demonstrate the effectiveness of MLFF and IGF layers in integrating multi-level features.

Depth of IGF. We further examine the effect of the depth of IGF layers on the network. From the results in Table 4, we observe that the IGF with multiple layers achieves better Overall IoU and Prec@X performances than the IGF with a single layer, which verifies the effectiveness of the multi-step progressive interaction strategy adopted by the IGF layers. Moreover, we also observe that with the increase of the depth L of IGF layers, the performances of the network first steadily increase and achieve the best at $L = 5$ and then decrease at $L = 6$. We consider the reason for such results is that the deeper IGF layers enable the deep interaction between guidance features and guided features as well as make the difficulty of network optimization. Similar observations are also reported by [60].

Settings for the output feature of IGF. Table 5 shows different settings of the output feature of IGF layers. We can see that incorporating features from multiple layers achieves better performance than setting only using the last layer. The setting formed by integrating the output features of each fusion unit achieves the best performance, so we use the setting as the default.

Adaptive Selection Gate. Table 6 shows the performance of using different gated mechanisms as the gating unit of the IGF layers. We can see that compared with ConvLSTMCell [46] and ConvGRUCell [9], the proposed ASGate achieves the better performances under different metrics with fewer parameters. These results are reasonable since ASGate is more concise and is more targeted. ASGate perceives the differences between the enhanced visual features and the original visual representation, adaptively

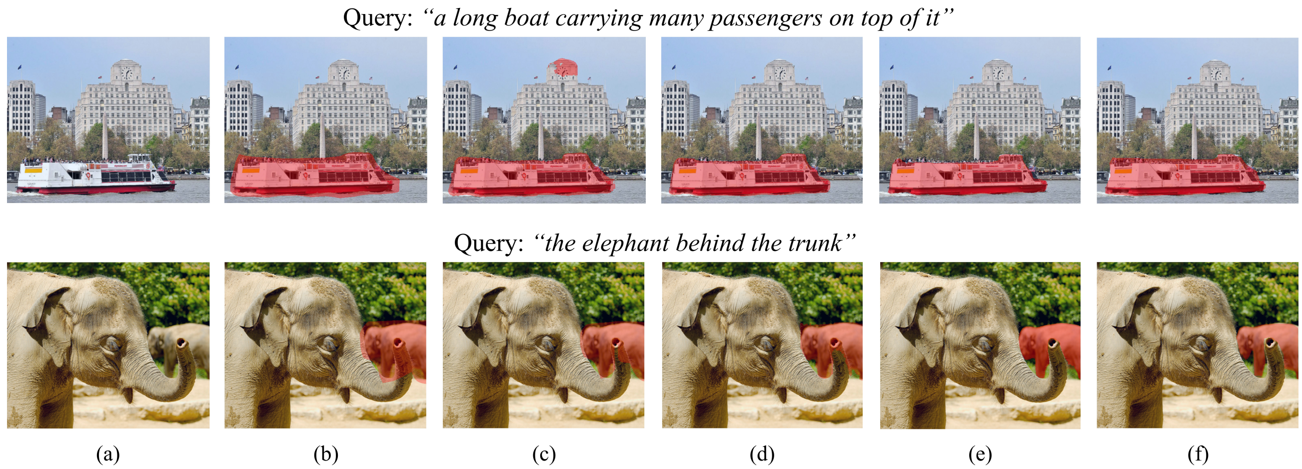

Figure 8: Prediction results with different stages removed. (a) Original Image. (b) DCMFNet without MLFF. (c) DCMFNet without LGVE-1. (d) DCMFNet without LGVE-2. (e) DCMFNet. (f) Ground Truth.

selects spatial regions of high-level semantic interest and aggregates them with the enhanced visual features in preparation for enhancement and suppression of the fusion unit.

Qualitative results. We first visualize the attention results from different stages of DCMFNet in Figure 6. The visualization shows that DCMFNet gradually highlights the referent from LGVE-1 to LGVE-2 and clarifies the boundaries of the referent at MLFF, finally generating the predicted mask close to the ground truth. We also visualize the attention results from different layers of IGF in Figure 7, respectively. It can be seen that with the depth L increasing, the signal response gradually shifts toward spatial regions related to the referent. In addition, we visualize the predicted results with different components removed in Figure 8. These qualitive results demonstrate the effectiveness of the proposed method.

Limitation. The proposed network uses language to continuously guides visual context modeling, which relies on the accuracy of semantic information extracted by the text encoder. Recent studies [20, 49, 51, 52, 56] have shown that transformer and bert based text encoders show great potential in enhancing the RIS task performance. In addition, the paper explores the potential ability of the vision encoder to align vision and language, leaving unexplored the potential of the language encoder to align visual and linguistic features, and we will explore this possibility in the future.

5 CONCLUSION

In this paper, we introduce a deep cross-modal fusion network (DCMFNet) for the referring image segmentation task. DCMFNet achieves cross-modal and multi-level feature alignment to segment the referent from an image by embedding the core component Iterative Gated Fusion (IGF) layers multiple times in the encoder and decoder. The proposed method outperforms many previous state-of-the-art methods on multiple benchmark datasets.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[2] Hedi Ben-younes, Rémi Cadène, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2631–2639. https://doi.org/10.1109/ICCV.2017.285

[3] Yihong Cao, Hui Zhang, Xiao Lu, Yurong Chen, Zheng Xiao, and Yaonan Wang. 2023. Adaptive Refining-Aggregation-Separation Framework for Unsupervised Domain Adaptation Semantic Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* (2023), 1–1. https://doi.org/10.1109/TCSVT.2023.3243402

[4] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. 2019. See-Through-Text Grouping for Referring Image Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 7453–7462. https://doi.org/10.1109/ICCV.2019.00755

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 4 (2018), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *CoRR* abs/1706.05587 (2017). arXiv:1706.05587 http://arxiv.org/abs/1706.05587

[7] Qingrong Cheng, Zhenshan Tan, Keyu Wen, Cheng Chen, and Xiaodong Gu. 2022. Semantic pre-alignment and ranking learning with unified framework for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* (2022).

[8] Yu Cheng, Zhe Gan, Yitong Li, Jingjing Liu, and Jianfeng Gao. 2020. Sequential Attention GAN for Interactive Image Editing. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 4383–4391. https://doi.org/10.1145/3394171.3413551

[9] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1724–1734. https://doi.org/10.3115/v1/d14-1179

[10] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. 2021. Vision-Language Transformer and Query Generation for Referring Segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 16301–16310. https://doi.org/10.1109/ICCV48922.2021.01601

[11] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villasenor, and Michael Grubinger. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer vision and image understanding* 114, 4 (2010), 419–428.

[12] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. 2021. Encoder fusion network with co-attention embedding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15506–15515.

[13] Junhao Feng, Guohua Wang, Changmeng Zheng, Yi Cai, Ze Fu, Yaowei Wang, Xiao-Yong Wei, and Qing Li. 2023. Towards Bridged Vision and Language: Learning Cross-modal Knowledge Representation for Relation Extraction. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[16] Lijun He, Ziqing Wang, Liejun Wang, and Fan Li. 2023. Multimodal Mutual Attention-Based Sentiment Analysis Framework Adapted to Complicated Contexts. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[17] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. 2016. Segmentation from Natural Language Expressions. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 9905)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer, 108–124. https://doi.org/10.1007/978-3-319-46448-0_7

[18] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. 2020. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4424–4433.

[19] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. 2020. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10488–10497.

[20] Ziling Huang and Shin'ichi Satoh. 2023. Referring Image Segmentation via Joint Mask Contextual Embedding Learning and Progressive Alignment Network. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 7753–7762.

[21] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. 2020. Linguistic Structure Guided Context Modeling for Referring Image Segmentation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X (Lecture Notes in Computer Science, Vol. 12355)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 59–75. https://doi.org/10.1007/978-3-030-58607-2_4

[22] Yang Jiao, Zequn Jie, Weixin Luo, Jingjing Chen, Yu-Gang Jiang, Xiaolin Wei, and Lin Ma. 2021. Two-stage Visual Cues Enhancement Network for Referring Image Segmentation. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 1331–1340. https://doi.org/10.1145/3474085.3475222

[23] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. 2021. Locate then segment: A strong pipeline for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9858–9867.

[24] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 787–798. https://doi.org/10.3115/v1/d14-1086

[25] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. 2022. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18145–18154.

[26] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023).

[28] Philipp Krähenbühl and Vladlen Koltun. 2011. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger (Eds.). 109–117. https://proceedings.neurips.cc/paper/2011/hash/beda24c1e1b46055dff2c39c98fd6fc1-Abstract.html

[29] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. 2018. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5745–5753.

[30] Liang Lin, Pengxiang Yan, Xiaoqian Xu, Sibei Yang, Kun Zeng, and Guanbin Li. 2021. Structured Attention Network for Referring Image Segmentation. *IEEE Transactions on Multimedia* (2021).

[31] Liang Lin, Pengxiang Yan, Xiaoqian Xu, Sibei Yang, Kun Zeng, and Guanbin Li. 2022. Structured Attention Network for Referring Image Segmentation. *IEEE Trans. Multim.* 24 (2022), 1922–1932. https://doi.org/10.1109/TMM.2021.3074008

[32] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 8693)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[33] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan L. Yuille. 2017. Recurrent Multimodal Interaction for Referring Image Segmentation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 1280–1289. https://doi.org/10.1109/ICCV.2017.143

[34] Daqing Liu, Hanwang Zhang, Zheng-Jun Zha, and Feng Wu. 2019. Learning to Assemble Neural Module Tree Networks for Visual Grounding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 4672–4681. https://doi.org/10.1109/ICCV.2019.00477

[35] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. 2018. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8759–8768.

[36] Zejun Liu, Fanglin Chen, Jun Xu, Wenjie Pei, and Guangming Lu. 2022. Image-Text Retrieval with Cross-Modal Semantic Importance Consistency. *IEEE Transactions on Circuits and Systems for Video Technology* (2022), 1–1. https://doi.org/10.1109/TCSVT.2022.3220297

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

[39] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. 2020. Cascade Grouped Attention Network for Referring Expression Segmentation. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 1274–1282. https://doi.org/10.1145/3394171.3414006

[40] Lei Ma, Hongliang Li, Fanman Meng, Qingbo Wu, and King Ngi Ngan. 2017. Learning Efficient Binary Codes From High-Level Feature Representations for Multilabel Image Retrieval. *IEEE Transactions on Multimedia* 19, 11 (2017), 2545–2560. https://doi.org/10.1109/TMM.2017.2703089

[41] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 11–20. https://doi.org/10.1109/CVPR.2016.9

[42] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. 2018. Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 630–645.

[43] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[44] Hao Sheng, Ruixuan Cong, Da Yang, Rongshan Chen, Sizhe Wang, and Zhenglong Cui. 2022. UrbanLF: A Comprehensive Light Field Dataset for Semantic Segmentation of Urban Scenes. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 11 (2022), 7880–7893. https://doi.org/10.1109/TCSVT.2022.3187664

[45] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. 2018. Key-Word-Aware Network for Referring Expression Image Segmentation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI (Lecture Notes in Computer Science, Vol. 11210)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer, 38–54. https://doi.org/10.1007/978-3-030-01231-1_3

[46] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna

Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 802–810. https://proceedings.neurips.cc/paper/2015/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html

[47] Mohit Shridhar and David Hsu. 2018. Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction. In *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, Hadas Kress-Gazit, Siddhartha S. Srinivasa, Tom Howard, and Nikolay Atanasov (Eds.). https://doi.org/10.15607/RSS.2018.XIV.028

[48] Xiangbo Shu, Jiawen Yang, Rui Yan, and Yan Song. 2022. Expansion-squeeze-excitation fusion network for elderly activity recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 8 (2022), 5281–5292.

[49] Wei Su, Peihan Miao, Huanzhang Dou, Gaoang Wang, Liang Qiao, Zheyang Li, and Xi Li. 2023. Language adaptive weight generation for multi-task visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10857–10866.

[50] Liuyi Wang, Zongtao He, Ronghao Dang, Huiyi Chen, Chengju Liu, and Qijun Chen. 2022. RES-StS: Referring Expression Speaker via Self-training with Scorer for Goal-Oriented Vision-Language Navigation. *IEEE Transactions on Circuits and Systems for Video Technology* (2022), 1–1. https://doi.org/10.1109/TCSVT.2022.3233554

[51] Wenxuan Wang, Xingjian He, Yisi Zhang, Longteng Guo, Jiachen Shen, Jiangyun Li, and Jing Liu. 2024. CM-MaskSD: Cross-Modality Masked Self-Distillation for Referring Image Segmentation. *IEEE Transactions on Multimedia* (2024).

[52] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2022. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11686–11695.

[53] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[54] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. 2021. Bottom-up shift and reasoning for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11266–11275.

[55] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A Fast and Accurate One-Stage Approach to Visual Grounding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 4682–4692. https://doi.org/10.1109/ICCV.2019.00478

[56] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18155–18165.

[57] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10502–10511.

[58] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1307–1315.

[59] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling Context in Referring Expressions. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 9906)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer, 69–85. https://doi.org/10.1007/978-3-319-46475-6_5

[60] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6281–6290.

[61] Lichen Zhao, Daigang Cai, Jing Zhang, Lu Sheng, Dong Xu, Rui Zheng, Yinjie Zhao, Lipeng Wang, and Xibo Fan. 2022. Towards Explainable 3D Grounded Visual Question Answering: A New Benchmark and Strong Baseline. *IEEE Transactions on Circuits and Systems for Video Technology* (2022), 1–1. https://doi.org/10.1109/TCSVT.2022.3229081