Beyond One-to-One: Rethinking the Referring Image Segmentation

Yutao Hu[1]*, Qixiong Wang[2]*, Wenqi Shao[2], Enze Xie[3],
Zhenguo Li[3], Jungong Han[4], Ping Luo[1,2]†
[1]The University of Hong Kong [2]Shanghai AI Laboratory
[3]Huawei Noah's Ark Lab [4]The University of Sheffield

Abstract

Referring image segmentation aims to segment the target object referred by a natural language expression. However, previous methods rely on the strong assumption that one sentence must describe one target in the image, which is often not the case in real-world applications. As a result, such methods fail when the expressions refer to either no objects or multiple objects. In this paper, we address this issue from two perspectives. First, we propose a Dual Multi-Modal Interaction (DMMI) Network, which contains two decoder branches and enables information flow in two directions. In the text-to-image decoder, text embedding is utilized to query the visual feature and localize the corresponding target. Meanwhile, the image-to-text decoder is implemented to reconstruct the erased entity-phrase conditioned on the visual feature. In this way, visual features are encouraged to contain the critical semantic information about target entity, which supports the accurate segmentation in the text-to-image decoder in turn. Secondly, we collect a new challenging but realistic dataset called Ref-ZOM, which includes image-text pairs under different settings. Extensive experiments demonstrate our method achieves state-of-the-art performance on different datasets, and the Ref-ZOM-trained model performs well on various types of text inputs. Codes and datasets are available at `https://github.com/toggle1995/RIS-DMMI`.

1. Introduction

Referring image segmentation aims to segment the target object described by a given natural language expression. Compared to the traditional semantic segmentation task [32, 41, 3], referring image segmentation is no longer restricted by the predefined classes and could segment specific individuals selectively according to the description of text, which has large potential value for various applications such as human-robot interaction [43] and image editing [2]. Despite the recent progress, there are still several important challenges that need to be addressed in order to make this technology more applicable in real-world scenarios.

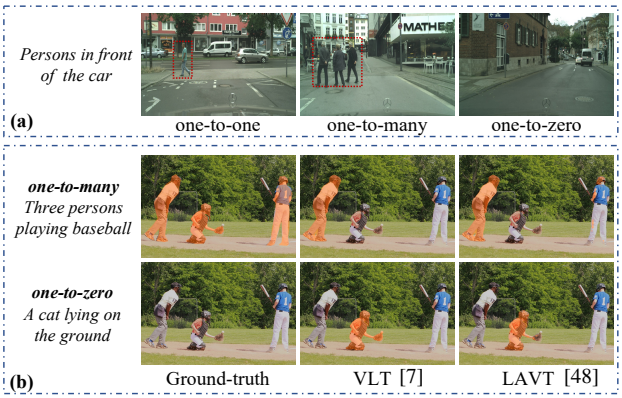

Figure 1: (a) Taking the autonomous driving as an example, text expression may refer to varying number of targets, depending on the specific real-world scenario. (b) When the sentences refer to multiple or no targets, existing methods cannot realize accurate segmentation.

In referring image segmentation, most previous methods only concentrate on the one-to-one setting, where each sentence only indicates one target in the image. However, as shown in Fig. 1(a), one-to-many and one-to-zero settings, where the sentence indicates many or no targets in the image, respectively, are also common and critical in the real-world applications. Unfortunately, previous methods tend to struggle when confronting one-to-many and one-to-zero samples. As illustrated in Fig. 1(b), the recent SOTA method, LAVT [48], only localizes one person in the image when given the description "Three persons playing baseball". As for one-to-zero input, previous methods still segment one target even if it is completely irrelevant to the given text. Therefore, it is imperative to enable the model to adapt to various types of text inputs.

*Equal contribution.
†Corresponding author.

We attribute this problem to two main factors. First, although existing methods design various ingenious modules to align multi-modal features, most of them only supervise the pixel matching of the segmentation map, which cannot ensure the significant semantic clues from the text are fully incorporated into the visual stream. As a result, visual features lack the comprehensive understanding of the entity being referred to in the expression, which limits the capacity when the model confronts various types of text inputs. Second, all popular datasets [21, 38, 36] for referring image segmentation are established under the one-to-one assumption. In the training, the model is enforced to localize one entity that is most related to the text. As a result, the model trained on these datasets is prone to overfitting and only remembers to segment the object with the largest response, which leads to the failure when segmenting one-to-many and one-to-zero samples.

To address the aforementioned issues, this paper proposes a Dual Multi-Modal Interaction Network (DMMI) to achieve robust segmentation when given various types of text expressions, and establishes a new comprehensive dataset Ref-ZOM (**Z**ero/**O**ne/**M**any). In the DMMI network, we address the referring segmentation task in a dual manner, which not only incorporates the text information into visual features but also enables the information flow from visual stream to the linguistic one. As illustrated in Fig. 2, the whole framework contains two decoder branches. On the one hand, in the text-to-image decoder, linguistic information is involved into the visual features to segment the corresponding target. On the other hand, we randomly erase the entity-phrase in the original sentence and extract the incomplete linguistic feature. Then, in the image-to-text decoder, given the incomplete text embedding, we utilize the Context Clue Recovery (CCR) module to reconstruct the missing information conditioned on the visual features. Meanwhile, multi-modal contrastive learning is also deployed to assist the reconstruction. By doing so, the visual feature is encouraged to fully incorporate the semantic clues about target entity, which promotes the multi-modal feature interaction and leads to more accurate segmentation maps. Additionally, to facilitate the two decoder parts, we design a Multi-scale Bi-direction Attention (MBA) module to align the multi-modal information in the encoder. Beyond the interaction between single-pixel and single-word [48], the MBA module enables the multi-modal interaction in the local region with various sizes, leading to a more comprehensive understanding of multi-modal features.

In the Ref-ZOM, we establish a comprehensive and challenging dataset to promote the referring image segmentation when given various types of text inputs. On the one hand, compared to the existing widely-used datasets [21, 38, 36], the text expressions are more complex in Ref-ZOM. It is not limited to the one-to-one assumption, and instead, the expression can refer to multiple or no targets within the image. Additionally, the language style in our Ref-ZOM is much more flowery than the short phrases found in [21]. On the other hand, Ref-ZOM also surpasses most mainstream datasets in terms of size, containing 55078 images and 74942 annotated objects.

We conduct extensive experiments on three popular datasets [21, 38, 36] and our DMMI achieves state-of-the-art results. Meanwhile, we reproduce some representative methods on our newly established Ref-ZOM dataset, where DMMI network consistently outperforms existing methods and exhibits strong ability in handling one-to-zero and one-to-many text inputs. Moreover, the Ref-ZOM-trained network performs remarkable generalization capacity when being transferred to different datasets without fine-tuning, highlighting its potential for real-world applications.

The main contributions of this paper are summarized as follows:

- We find the deficiency of referring image segmentation when meeting the one-to-many and one-to-zero text inputs, which strongly limits the application value in real-world scenarios.
- We propose a Dual Multi-Modal Interaction (DMMI) Network to enable the information flow in two directions. Besides the generation of segmentation map, DMMI utilizes the image-to-text decoder to reconstruct the erased entity-phrase, which facilitates the comprehensive understanding of the text expression.
- We collect a new challenging dataset, termed as Ref-ZOM, in which the text inputs are not limited to the one-to-one setting. The proposed dataset provides a new perspective and benchmark for future research.
- Extensive experimental results show the proposed DMMI network achieves new state-of-the-art results on three popular benchmarks, and exhibits superior capacity in handling various types of text inputs on the newly collected Ref-ZOM.

2. Related Work

2.1. Referring Image Segmentation

Referring image segmentation is first introduced by [15]. Early approaches [15, 27, 29, 37] generally employ Convolutional Neural Networks (CNNs) [3, 40, 13] and Recurrent Neural Networks (RNNs) [14, 17] to extract relevant visual and linguistic features. After feature extraction, the concatenation-convolution operation is employed to fuse multi-modal features. However, it fails to exploit the inherent interaction between image and text. To overcome this shortcoming, some approaches [16, 18, 47] establish relation-aware reasoning based on the multi-modal graph.

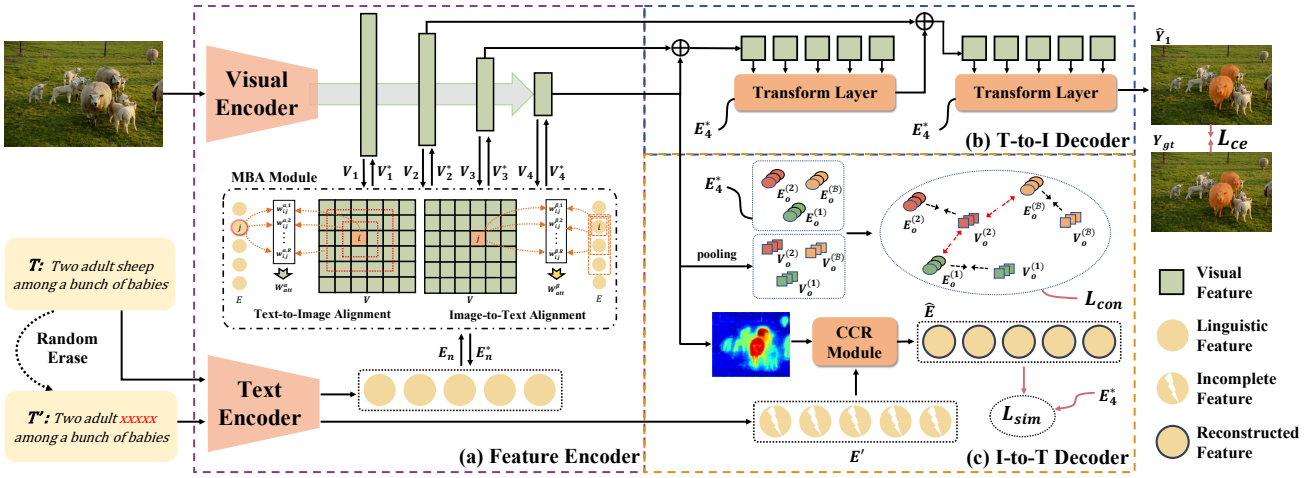

Figure 2: The whole framework of the proposed Dual Multi-Modal Interaction (DMMI) Network. **(a)** The feature encoder, in which the visual encoder and text encoder are utilized to extract visual and linguistic feature, respectively. Meanwhile, MBA module is employed to perform multi-modal feature interaction. Notably, $w_{i,j}$ denotes the j-th point in the i-th row of attention weight. **(b)** Text-to-image decoder, in which the text embedding is utilized to query the visual feature to generate the prediction map. **(c)** Image-to-text decoder, in which CCR module is utilized to reconstruct the erased linguistic information condition on the visual feature. $\mathcal{L}_{con}$ and $\mathcal{L}_{sim}$ are implemented to assist the reconstruction.

Recently, due to the breakthrough of Transformer in computer vision community [8, 46, 12, 1], Transformer-based backbones have become dominant in referring image segmentation for both visual and linguistic feature extraction [31, 6, 19, 20]. Meanwhile, the self-attention mechanism [42] in the Transformer also inspired numerous studies that employ cross-attention blocks for better cross-model alignment. For instance, VLT [7] utilizes the cross-attention module to generate the query vectors by comprehensively understanding the multi-model features, which are then used to query the given image through the Transformer decoder. LAVT [48] finds that early fusion of multi-modal features via cross-attention module brings better cross-modal alignments. Moreover, CRIS [44] utilizes the Transformer block to transfer the strong ability of image-text alignment from the pre-trained CLIP model [39].

However, most previous methods only supervise the visual prediction and cannot ensure the semantic clues in the text expressions have been incorporated into the visual features. As a result, these methods tend to struggle when handling the text expressions that refer to either no objects or multiple objects. In this work, we establish a dual network and emphasize that the information flow from image to text is beneficial for comprehensive understanding of text expression. Furthermore, we collect a new dataset called RefZOM, which contains various types of text inputs and compensates for the limitations of existing benchmarks.

2.2. Visual-Language Understanding

Video-Language understanding has received rapidly growing attention in recent years and plays an important role in various tasks such as video-retrieval [10], image-text matching [24] and visual question answering [53, 25]. In these tasks, effective multi-modal interaction and comprehensive understanding of both visual and linguistic features are critical in achieving great performance. Some previous works employ masked word prediction (MWP) to achieve this goal, where a proportion of words in a sentence are randomly masked, and the masked words are predicted under the condition of visual inputs [23, 11, 54]. Most MWP methods directly predict the value of the token. In our work, instead of predicting the single token, we reconstruct the holistic representation of text embedding and measure the global similarity, leading to the comprehensive understanding of the entire sentence.

Moreover, recently popular vision-language pre-training models [39, 26, 49, 50] have demonstrated the remarkable ability of contrastive learning in cross-modal representation learning. Motivated by their success, we incorporate the contrastive loss in our image-to-text decoder to facilitate text reconstruction. The experimental results reflect that the two components are highly complementary and effectively enhance the semantic clues in visual features.

3. Method

The Dual Multi-Modal Interaction (DMMI) network adopts the encoder-decoder paradigm, which is illustrated in Fig. 2. In the encoder part, the visual encoder and text encoder are utilized to extract visual and linguistic features, respectively. During this process, the Multi-scale Bi-direction attention (MBA) module is employed to perform cross-modal interaction. After feature extraction, the two

modalities are delivered to the decoder part. In the text-to-image decoder, the text embedding is utilized to query the visual feature and generate the segmentation mask. While in the image-to-text decoder, we employ the Context Clue Recovery (CCR) module to reconstruct the erased information of target entity conditioned on the visual features. Meanwhile, the contrastive loss is utilized to promote the learning of CCR module. We elaborate each component of the DMMI network in detail in the following sections.

3.1. Feature Encoder

Given the text expression T, we randomly mask the entity-phrase via TextBlob Tool [33] and generate its corresponding counterpart T'. Then, we feed both T and T' into the text encoder to generate the linguistic features $E = \{e_l\}_{l=1}^{L}$ and $E' = \{e'_l\}_{l=1}^{L} \in \mathbb{R}^{C_t \times L}$, where C_t and L indicate the number of channels and the length of the sentence. For the input image X, we utilize the visual encoder to extract the multi-level visual features $V_n \in \mathbb{R}^{C_n \times H_n \times W_n}$. Here, C_n, H_n and W_n denote the number of channels, height and width, and n indicates features in the n-th stage. During the feature extraction, MBA module is hierarchically applied to perform cross-modal feature interaction.

3.1.1 Hierarchical Structure

As illustrated in Fig. 2 (a), the visual encoder is implemented as a hierarchical structure with four stages, which is conducted with the MBA module alternately. For the shallow layer feature V_1 extracted from the first stage of visual encoder, we deliver it to MBA module with linguistic feature E_1 and obtain V_1^* and E_1^*. Then, V_1^* is sent back to the visual encoder, based on which V_2 is extracted through the next stage. Meanwhile, E_1^* is also noted as the E_2 that will be utilized in the next MBA module. Similarly, V_2 and E_2 are fed to MBA module again, and the generated V_2^* will be delivered to the next part of visual encoder. By doing so, the visual and linguistic features are jointly refined, achieving cross-modal alignment in both text-to-image and image-to-text directions.

3.1.2 Multi-scale Bi-direction attention Module

The MBA module jointly refines the visual feature V and linguistic feature E to achieve text-to-image and image-to-text alignment. To simplify the notation, here we drop the subscript of features from different stages. Inspired by the success of self-attention [42], most recent works utilize the cross-attention operation to perform the cross-modal feature interaction. During this process, the visual feature V is first flattened to $\mathbb{R}^{C \times N}$, where $N = W \times H$. Then, the feature interaction is formulated as:

$$V^* = \text{softmax}(\frac{(W_q V)^T (W_k E)}{\sqrt{\hat{C}}})\,(W_v E)^T \tag{1}$$

where W_q, W_k and W_v are three transform functions unifying the number of channels to $\hat{C}$. However, Eq. 1 only establishes the relationship between a single pixel and a single word. In fact, beyond the single point representation, local visual regions and text sequences also store critical information for the comprehensive understanding of multimodal features. Following this idea, we design two alignment strategies in MBA to capture the relationship between visual features and text sequences in different local regions.

Text-to-Image Alignment. To fully leverage the structure information in various regions, we compute the affinities coefficients $W_{att}^{\alpha,r}$ between each token and different local regions Ω_r^α, in which r indicates different spatial sizes and its value ranges from 1 to R. Ω_r^α will slide across the whole spatial plane of the visual feature. Then, given region $\Omega_r^\alpha(i)$ centered at the position i, the i-th row weight $w_i^{\alpha,r}$ in attention matrix $W_{att}^{\alpha,r} \in \mathbb{R}^{N \times L}$ is calculated as:

$$w_i^{\alpha,r} = \text{softmax}(\sum_{m \in \Omega_r^\alpha(i)} \frac{(W_q^\alpha V^m)^T (W_k^\alpha E)}{\sqrt{\hat{C}}}) \tag{2}$$

where $w_i^{\alpha,r} \in \mathbb{R}^{1 \times L}$, m enumerates all spatial positions in $\Omega_r^\alpha(i)$ and $V^m \in \mathbb{R}^{\hat{C} \times 1}$ denotes one specific feature vector in $\Omega_r^\alpha(i)$. Then, for all Ω_r^α, the final affinities coefficient is calculated as:

$$W_{att}^\alpha = \sum_{r=1}^{R} \lambda_r^\alpha W_{att}^{\alpha,r} \tag{3}$$

where λ_r^α is a learnable parameter reflecting the importance of regions in different sizes. Finally, after the process of transform function W_v^α, the linguistic information is incorporated into the visual feature:

$$V^* = W_{att}^\alpha (W_v^\alpha E)^T \tag{4}$$

Image-to-Text Alignment. In human perception, to fully comprehend the language expression, we will associate the context information rather than understanding each word separately. Therefore, for each visual pixel, we also establish the connection with various text sequences Ω_r^β, where r indicates different lengths of the sequence and Ω_r^β slides across the whole sentence. For text sequence $\Omega_r^\beta(i)$ starting at position i, we calculate the i-th row weight $w_i^{\beta,r}$ in affinity coefficients $W_{att}^{\beta,r} \in \mathbb{R}^{L \times N}$ as:

$$w_i^{\beta,r} = \text{softmax}(\sum_{m \in \Omega_r^\beta(i)} \frac{(W_q^\beta E^m)^T (W_k^\beta V)}{\sqrt{\hat{C}}}) \tag{5}$$

where $w_i^{\beta,r} \in \mathbb{R}^{1 \times N}$, m enumerates all tokens in $\Omega_r^\beta(i)$ and $E^m \in \mathbb{R}^{\hat{C} \times 1}$ represents one specific feature vector in $\Omega_r^\beta(i)$. Then, similar to Eq. 3, we average the $W_{att}^{\beta,r}$ through

a set of learnable parameters λ_r^β to obtain the W_{att}^β. Afterwards, the visual information is involved to generate the refined text embedding as follows:

$$E^* = W_{att}^\beta (W_v^\beta V)^T \tag{6}$$

3.2. Text-to-Image Decoder

The whole structure of text-to-image decoder is depicted in Fig. 2(b). As advocated in [41, 3], we implement skip-connections between the encoder and decoder to introduce the spatial information stored in the shallow layers. Specifically, the text-to-image decoder can be described as:

$$\begin{cases} Y_4 = V_4^* \\ Y_n = \psi\left(\phi\left(Y_{n+1}, V_n^*\right), E_4^*\right) \quad n = 3, 2 \end{cases} \tag{7}$$

in which $\psi\left(\cdot\right)$ indicates the Transformer decoder layer. $\phi\left(\cdot\right)$ consists of two 3×3 convolutions followed by batch normalization and the ReLU function, in which features from the shallow parts of the encoder are aggregated with the decoder feature. Then, a series of convolution operations are applied on Y_2 to produce two class score maps $\hat{Y}_1$, which is considered as the final visual prediction of DMMI network. Finally, we calculate the binary cross-entropy loss for $\hat{Y}_1$ with Y_{gt}, which is denoted as $\mathcal{L}_{ce}$.

3.3. Image-to-Text Decoder

3.3.1 Context Clue Recovery Module

Besides the text-to-image decoder, DMMI network promotes the referring segmentation in a dual manner and facilitates the information flow from visual to text, which is illustrated in Fig. 2(c). For the incomplete linguistic feature $E' = \{e_l'\}_{l=1}^L$, we utilize CCR module to reconstruct its masked information under the guidance of visual feature V_g^*. To support the precise reconstruction, the visual feature is encouraged to contain essential semantic clues stored in the $E = \{e_l\}_{l=1}^L$, which boosts the sufficient multi-modal interaction in the encoder part and support the accurate segmentation in the text-to-image decoder.

Specifically, given the visual feature V_g^*, we employ a Transformer decoder layer $\mathcal{D}(E', V_g^*)$ to recover the missed information in the $E' = \{e_l'\}_{l=1}^L$, where visual feature V_g^* is employed to query the E'. Notably, we extract V_g^* from middle part of the text-to-image decoder, which contains both spatial and semantic information. The output of $\mathcal{D}(E', V_g^*)$ is considered as the reconstructed text embedding, which is denoted as $\hat{E} = \{\hat{e}_l\}_{l=1}^L$.

To enforce the CCR module to precisely recover the missing information, we measure the similarity distance between the reconstructed embedding $\hat{E}$ and E_4^*, and calculate $\mathcal{L}_{sim}$ as:

$$\mathcal{L}_{\text{sim}} = \delta * \left(1 - \cos\left[\text{Detach}\left(E_4^*\right), \hat{E}\right]\right) \tag{8}$$

Here, δ is an indicator that will be set to 0 if this sample is a one-to-zero case, where the text input is unrelated to the corresponding image, making it impossible to reconstruct linguistic information. Additionally, $\text{Detach}(E_4^*)$ refers to stopping the gradient flow of E_4^* in Eq. 8, which prevents E_4^* from being misled by $\hat{E}$ in the optimization.

3.3.2 Multi-modal Contrastive Learning

We calculate the contrastive loss to reduce the distance between visual feature and its corresponding linguistic one, which is helpful in reconstructing the text embedding from the visual representation. Specifically, we aggregate features from different parts of text-to-image decoder to generate $\tilde{V}_d^*$. Then, for visual feature $\tilde{V}_d^* \in \mathbb{R}^{\mathcal{B}\times N\times C}$ and linguistic feature $\tilde{E}_4^* \in \mathbb{R}^{\mathcal{B}\times L\times C}$ in a batch, we pool them into V_o and $E_o \in \mathbb{R}^{\mathcal{B}\times C}$. Afterwards, the contrastive loss is computed as:

$$\mathcal{L}_{con} = \mathcal{L}_{I\to T} + \mathcal{L}_{T\to I} \tag{9}$$

where $\mathcal{L}_{I\to T}$ and $\mathcal{L}_{T\to I}$ denote image-to-text and text-to-image contrastive loss respectively:

$$\mathcal{L}_{I\to T} = -\frac{1}{\mathcal{B}}\sum_{i=1}^{\mathcal{B}} \delta^{(i)} * \log \frac{\exp\left(V_o^{(i)} \cdot E_o^{(i)}/\tau\right)}{\sum_{j=1}^{\mathcal{B}} \exp\left(V_o^{(i)} \cdot E_o^{(j)}/\tau\right)} \tag{10}$$

$$\mathcal{L}_{T\to I} = -\frac{1}{\mathcal{B}}\sum_{i=1}^{\mathcal{B}} \delta^{(i)} * \log \frac{\exp\left(E_o^{(i)} \cdot V_o^{(i)}/\tau\right)}{\sum_{j=1}^{\mathcal{B}} \exp\left(E_o^{(i)} \cdot V_o^{(j)}/\tau\right)} \tag{11}$$

where $V_o^{(i)} \in \mathbb{R}^C$ and $E_o^{(i)} \in \mathbb{R}^C$ denote i^{th} sample in a batch, $\mathcal{B}$ indicates the batch size. Meanwhile, δ is the one-to-zero indicator, τ is the temperature hyper-parameter that scales the logits. Finally, the total loss is combined as the summation of $\mathcal{L}_{ce}$, $\mathcal{L}_{sim}$ and $\mathcal{L}_{con}$ over the batch.

4. Ref-ZOM Dataset

We collect Ref-ZOM to address the limitations of mainstream datasets [21, 38, 36] that only contain one-to-one samples. Following previous works [21, 38, 36], images in Ref-ZOM are selected from COCO dataset [28]. Generally, Ref-ZOM contains 55078 images and 74942 annotated objects, in which 43,749 images and 58356 objects are utilized in training, and 11329 images and 16,586 objects are employed in testing. Notably, Ref-ZOM is the first dataset that contains one-to-zero, one-to-one, and one-to-many samples simultaneously. It is worthwhile to mention that although the VGPHRASECUT dataset [45] includes some one-to-many samples, it lacks one-to-zero cases, which makes it less applicable than Ref-ZOM. Due to the space limitation, we only illustrate a selection of representative samples from Ref-ZOM in Fig. 3. More detailed information can be found in the supplementary materials.

One-to-many. We collect one-to-many samples in three different ways, as illustrated in the first row of Fig. 3 from left to right. (1) We manually create some image-text pairs

Figure 3: Selected samples from our newly collected Ref-ZOM datasets. From top to down are image-text pairs under one-to-many, one-to-zero and one-to-one settings.

based on the expressions from COCO_Caption and annotate the corresponding target in a two-player game [21, 52]. Specifically, given an image with caption expressions and annotations, the first annotator selects and modifies the sentence to describe the masked objects. Then, only given the image, the second annotator is asked to select the targets according to the text expression from the first one. The image-text pair will be collected only when the second annotator selects the targets correctly. (2) Based on the existing one-to-one referring image segmentation dataset, we combine the text expression describing different targets in one image to compose the one-to-many expressions. (3) We utilize the category information with the prompt template to compose some text samples. Generally, Ref-ZOM contains 41842 annotated objects under one-to-many settings.

One-to-zero. We carefully select 11937 images from COCO dataset [28], which are not included in [21, 38, 36]. Next, we randomly pair each image with a text expression taken from either the COCO captions or the text pools in [21, 38, 36]. Finally, we conduct a thorough double-checking process to verify that the selected text expressions are unrelated to the corresponding images.

One-to-one. First, we randomly select some samples from existing datasets [21, 38, 36]. Meanwhile, we manually create some new samples based on the category information with the prompt template, which is similar with the third strategy in the creation of one-to-many samples. In total, there are 42421 one-to-one objects in the Ref-ZOM.

5. Experiments

5.1. Implementation Details

We evaluate the performance of DMMI with two different visual encoders, ResNet-101 and Swin-Transformer-Base (Swin-B), which are initialized with classification weights pre-trained on ImageNet-1K and ImageNet-22K [5], respectively. Our text encoder is the base_BERT model with 12 layers and the hidden size of 768, which is initialized with the official pre-trained weights [6]. In the training, we utilize AdamW as the optimizer with a weight decay of 0.01. Moreover, the initial learning rate is set to 5e-5 with a polynomial learning rate decay policy. The images are resized to 448 $\times$ 448 and the maximum sentence length is set to 20. Additionally, we only randomly erase one phrase in each iteration for each sentence.

In DMMI network, the values of r for Ω_r^α and Ω_r^β are set to [1, 2, 3]. Specifically, Ω_1^α, Ω_2^α, and Ω_3^α correspond to spatial regions of size 1×1, 3×3, and 5×5, respectively. In addition, when $r = 1, 2, 3$, Ω_r^β corresponds to text sequences with 1, 2, and 3 tokens, respectively. Furthermore, all Transformer layers in the decoder are set with 8 heads and temperature τ in $\mathcal{L}_{con}$ equals to 0.05.

5.2. Datasets and Metrics

In addition to our newly collected Ref-ZOM, we evaluate our method on three mainstream referring image segmentation datasets, RefCOCO[21], RefCOCO+[21] and G-Ref[38, 36]. Notably, G-Ref has two different partitions, which are established by UMD [38] and Google, respectively [36]. We evaluate our method on both of them.

In the test, for one-to-one and one-to-many samples, we employ the overall intersection-over-union (oIoU), the mean intersection-over-union (mIoU), and prec@X to evaluate the quality of segmentation masks. The oIoU measures the ratio between the total intersection area and the total union area added from all test samples, while the mIoU averages the IoU score of each sample across the whole test set. Prec@X measures the percentage of test images with an IoU score higher than the threshold $X \in \{0.5, 0.7, 0.9\}$. As for the one-to-zero samples, since there is no target included in the image, IoU-based metrics are not applicable. Thus, we utilize image-level accuracy (Acc) to evaluate the performance. For each one-to-zero sample, its Acc value is 1 only when all points in the prediction mask are classified as the background. Otherwise, the Acc value is 0. We average the Acc value across the whole test set.

5.3. Comparison with State-of-the Arts

In Table 1, we compare the proposed DMMI network with the state-of-the-art methods on RefCOCO, RefCOCO+, and G-Ref in terms of the oIoU metric. The table is divided into two parts according to their visual encoder. The first part reports the performance of methods equipped with CNNs as the visual encoder, while the second part presents the methods using Transformer-based structure or pre-trained backbones beyond ImageNet as the visual encoder. Generally speaking, DMMI delivers the best performance in two conditions. Here, taking the second part as the example for analysis. On the RefCOCO dataset, we sur-

Table 1: Comparison with state-of-the-art methods in terms of oIoU(%) on three datasets. In G-Ref, U and G denote the UMD and Google partition, respectively. The best results are in bold.

Method	Visual Encoder	RefCOCO			RefCOCO+			G-Ref		
		val	test A	test B	val	test A	test B	val (U)	test (U)	val (G)
RRN [16]	ResNet-101	55.33	57.26	53.93	39.75	41.25	36.11	–	–	36.45
MAttNet [51]	ResNet-101	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61	–
CAC [4]	ResNet-101	58.90	61.77	53.81	–	–	–	46.37	46.95	44.32
LSCM [18]	ResNet-101	61.47	64.99	59.55	49.34	53.12	43.50	–	–	48.05
CMPC+ [30]	ResNet-101	62.47	65.08	60.82	50.25	54.04	43.47	–	–	49.89
MCN [35]	DarkNet-53	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	–
EFN [9]	ResNet-101	62.76	65.69	59.67	51.50	55.24	43.01	–	–	51.93
BUSNet [47]	ResNet-101	63.27	66.41	61.39	51.76	56.87	44.13	–	–	50.56
CGAN [34]	ResNet-101	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	46.54
LTS [19]	DarkNet-53	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25	–
VLT [7]	DarkNet-53	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	49.76
DMMI (Ours)	ResNet-101	**68.56**	**71.25**	**63.16**	**57.90**	**62.31**	**50.27**	**59.02**	**59.24**	**55.13**
ReSTR [22]	ViT-B	67.22	69.30	64.45	55.78	60.44	48.27	–	–	54.48
CRIS [44]	CLIP-R101	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36	–
LAVT [48]	Swin-B	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	60.50
DMMI (Ours)	Swin-B	**74.13**	**77.13**	**70.16**	**63.98**	**69.73**	**57.03**	**63.46**	**64.19**	**61.98**

Table 2: Comparisons with some representative methods on the newly collected Ref-ZOM dataset.

Method	oIoU	mIoU	Acc
MCN [35]	55.03	54.70	75.81
CMPC [16]	56.19	55.72	77.01
VLT [7]	60.21	60.43	79.26
LAVT [48]	64.45	64.78	83.11
DMMI (Ours)	**68.77**	**68.21**	**87.02**

pass the second-best method by 1.4%, 1.31%, and 1.37% on val, testA, and testB subsets, respectively. On RefCOCO+ dataset, our DMMI network achieves a significant gain over the SOTA method, with increases of 1.84%, 1.35%, and 1.93% on the val, testA, and testB subsets, respectively. On the UMD partition of G-Ref dataset, 2.22% and 2.1% oIoU improvements are obtained, while a 1.48% increase is also observed on the Google partition. Such improvements are consistent in the first part of Table 1.

In addition, we reproduce some representative methods on the newly collected Ref-ZOM dataset and evaluate our DMMI against these methods. The performance comparison is presented in Table 2. Here, our DMMI is equipped with Swin-B as the visual encoder. As shown, our method achieves the best performance in handling the one-to-many and one-to-zero settings. To be more specific, DMMI outperforms the second-best method by 4.32% and 3.43% in terms of oIoU and mIoU. Moreover, in terms of the metric for one-to-zero samples, DMMI surpasses the secondary method by 3.91% in Acc results.

5.4. Ablation Study

In this part, we perform several ablation studies to evaluate the effectiveness of the key components in our DMMI network on both G-Ref_(U) and Ref-ZOM datasets. The results are listed in Table 3.

Effect of Image-to-Text Decoder. The first three rows in Table 3 verify the effectiveness of the image-to-text decoder. First, we remove the whole image-to-text decoder and report the performance in the first row of Table 3, where a 1.7% performance degradation could be observed on G-Ref. This reflects the image-to-text decoder contributes a lot to producing accurate segmentation result. Next, we add the similarity loss $\mathcal{L}_{sim}$ and report the results in the second row of Table 3. We can find $\mathcal{L}_{sim}$ brings significant improvements. Especially, on the Ref-ZOM, the accuracy improves by 1.48% and 1.64% in terms of oIoU and Acc. Meanwhile, 0.71% oIoU gain is also found on G-Ref when the network is equipped with $\mathcal{L}_{sim}$. This demonstrates that through the reconstruction of incomplete text embedding in the training, DMMI is learned to fully incorporate the semantic clues about the entity targets into the visual features, which brings superior performance when facing various types of text inputs. Additionally, we verify the effectiveness of $\mathcal{L}_{con}$ in the third row. Compared to the baseline in the first row, performance goes up by 0.59% and 0.82% on the Ref-ZOM in terms of oIoU and Acc. This suggests $\mathcal{L}_{con}$ also contributes to the comprehensive understanding of target entity by pairing corresponding multi-modal features. Moreover, in the seventh row, we can find the best performance is achieved when the network equipped with $\mathcal{L}_{sim}$ and $\mathcal{L}_{con}$ simultaneously, reflecting the contrastive learning and text reconstruction are highly complementary.

Effect of MBA module. In the fourth to sixth row of Table 3, we conduct experiments to investigate the effectiveness of MBA module. On the one hand, as shown in the fifth and seventh row, if we prohibit the multi-modal interaction between various local regions, and only imple-

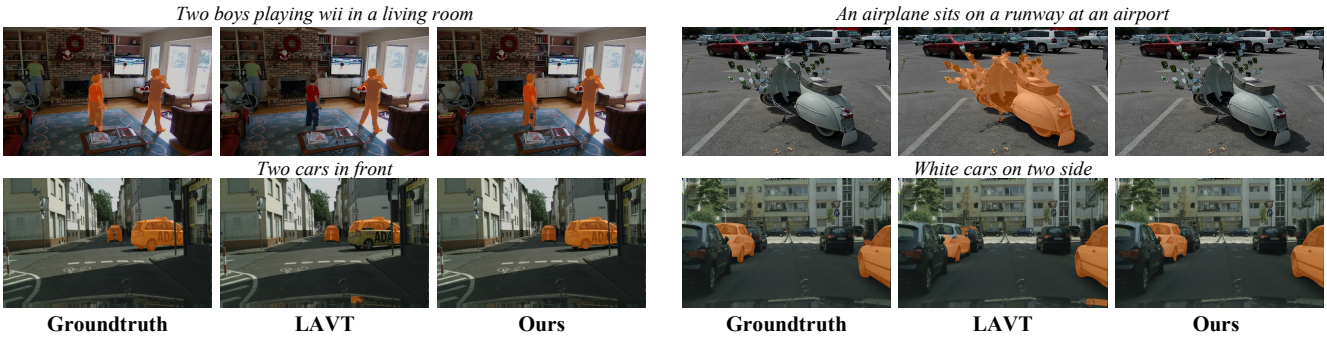

Figure 4: Comparisons of segmentation maps generated by LAVT and our DMMI network.

Table 3: Ablation study of different components in DMMI network on G-Ref and Ref-ZOM datasets. Notably, "Bi-D" indicates the bi-direction operation in MBA module and I2T denotes the "Image-to-Text".

	MBA		I2T Decoder		G-Ref	Ref-ZOM	
	Bi-D	MS	$\mathcal{L}_{sim}$	$\mathcal{L}_{con}$	oIoU	oIoU	Acc
1	✓	✓			61.76	65.77	83.91
2	✓	✓	✓		62.47	67.25	85.55
3	✓	✓		✓	62.13	66.36	84.73
4			✓	✓	62.05	67.13	85.09
5	✓		✓	✓	62.20	67.31	85.82
6		✓	✓	✓	62.48	67.52	86.11
7	✓	✓	✓	✓	**63.46**	**68.77**	**87.02**

ment the interaction between single-word and single-pixel, the segmentation results drop significantly. Specifically, 1.46% and 1.2% degradation are observed on Ref-ZOM, which demonstrates the interaction in a large region benefits the comprehensive understanding of multi-modal features. On the other hand, we forbid the bi-direction mechanism in MBA module by removing the image-to-text alignment and only retaining the text-to-image one. The results are listed in the sixth row in Table 3, in which the performance drops a lot compared to the whole network. This reflects that mutually refining the multi-modal features in the interaction contributes to producing the accurate segmentation map.

5.5. Visualization

In this section, we visualize some segmentation maps generated from DMMI and LAVT [48] to further demonstrate the superiority of our method.

Zero-shot to Ref-ZOM. We first visualize some segmentation maps when the model is trained on the G-Ref and transferred to Ref-ZOM under the zero-shot condition. The results are illustrated in the first row of Fig. 4. Since G-Ref only contains one-to-one samples, it is challenging to directly utilize the G-Ref-trained model to address the one-to-many and one-to-zero cases. As shown in the first sample, our DMMI network could precisely localize two boys and distinguish the women in the background. However, LAVT could only localize one boy with the largest size in the image. As for the second sample in the first row, DMMI also handles the one-to-zero case successfully.

Zero-shot to Cityscapes. To further verify the generalization ability of DMMI, we directly transfer the Ref-ZOM-trained networks to the Cityscapes dataset and give the model some expressions as the text input. The training images in Ref-ZOM all come from the COCO dataset, where the image style is quite different from that in Cityscapes. Thus, it is challenging to produce satisfactory performance when the model is transferred to Cityscapes without fine-tuning. As shown in the second row of Fig. 4, DMMI presents the satisfactory performance. Specifically, when we give the text "White cars on two side", DMMI could precisely localize the corresponding targets while LAVT segment many irrelevant cars and fails to produce accurate segmentation map, demonstrating the great generalization ability of our method.

6. Conclusion

In this paper, we point out the limitations of existing referring image segmentation methods in handling expressions that refer to either no objects or multiple objects. To solve this problem, we propose a Dual Multi-Modal Interaction (DMMI) Network and establish the Ref-ZOM dataset. In the DMMI network, besides the visual prediction, we reconstruct the erased entity-phrase based on the visual features, which promotes the multi-modal interaction. Meanwhile, in the newly collected Ref-ZOM, we include image-text pairs under one-to-zero and one-to-many settings, making it more comprehensive than previous datasets. Experimental results show that the proposed method outperforms the existing method by a large margin, and Ref-ZOM dataset endows the network with remarkable generalization ability in understanding various text expressions. We hope our work provides a new perspective for future research.

Acknowledgement

This paper is partially supported by the National Key R&D Program of China No.2022ZD0161000 and the General Research Fund of Hong Kong No.17200622.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[2] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8721–8729, 2018.

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[4] Yi-Wen Chen, Yi-Hsuan Tsai, Tiantian Wang, Yen-Yu Lin, and Ming-Hsuan Yang. Referring expression object segmentation with caption-aware consistency. *arXiv preprint arXiv:1910.04748*, 2019.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

[10] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 214–229. Springer, 2020.

[11] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022.

[12] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, 2022.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Proceedings of the European Conference on Computer Vision*, pages 108–124. Springer, 2016.

[16] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10488–10497, 2020.

[17] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[18] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 59–75. Springer, 2020.

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

[20] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[21] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, 2014.

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

[23] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[24] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4654–4662, 2019.

[25] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10313–10322, 2019.

[26] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

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

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[29] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1271–1280, 2017.

[30] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4761–4775, 2021.

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[32] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[33] Steven Loria et al. textblob documentation. *Release 0.15*, 2(8), 2018.

[34] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1274–1282, 2020.

[35] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10034–10043, 2020.

[36] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2016.

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

[38] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Proceedings of the European Conference on Computer Vision*, pages 792–807. Springer, 2016.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[40] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[43] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6629–6638, 2019.

[44] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11686–11695, 2022.

[45] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020.

[46] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

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

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

[49] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv preprint arXiv:2209.09407*, 2022.

[50] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[51] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[52] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[53] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019.

[54] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8746–8755, 2020.