SLViT: Scale-Wise Language-Guided Vision Transformer for Referring Image Segmentation

Shuyi Ouyang[1] , **Hongyi Wang**[1] , **Shiao Xie**[1] , **Ziwei Niu**[1] , **Ruofeng Tong**[1,3] , **Yen-Wei Chen**[2*] and **Lanfen Lin**[1*]
[1]Zhejiang University
[2]Ritsumeikan University
[3]Zhejiang Lab

Abstract

Referring image segmentation aims to segment an object out of an image via a specific language expression. The main concept is establishing global visual-linguistic relationships to locate the object and identify boundaries using details of the image. Recently, various Transformer-based techniques have been proposed to efficiently leverage long-range cross-modal dependencies, enhancing performance for referring segmentation. However, existing methods consider visual feature extraction and cross-modal fusion separately, resulting in insufficient visual-linguistic alignment in semantic space. In addition, they employ sequential structures and hence lack multi-scale information interaction. To address these limitations, we propose a Scale-Wise Language-Guided Vision Transformer (SLViT) with two appealing designs: (1) Language-Guided Multi-Scale Fusion Attention, a novel attention mechanism module for extracting rich local visual information and modeling global visual-linguistic relationships in an integrated manner. (2) An Uncertain Region Cross-Scale Enhancement module that can identify regions of high uncertainty using linguistic features and refine them via aggregated multi-scale features. We have evaluated our method on three benchmark datasets. The experimental results demonstrate that SLViT surpasses state-of-the-art methods with lower computational cost. The code is publicly available at: https://github.com/NaturalKnight/SLViT.

1 Introduction

Referring segmentation refers to the task of segmenting an object based on a given text description that may contain information about the target's action, category, color, position in the image, etc [Cheng *et al.*, 2014; Hu *et al.*, 2016]. It has a promising application prospects in many fields, such as language-based man-machine interaction. Unlike the conventional semantic and instance segmentation tasks, referring segmentation task requires precise perception of the locations of different objects in an image, making global visual-linguistic relationships modeling indispensable. Moreover, effective edge detection of the target objects requires details of the image, necessitating high-quality local visual features.

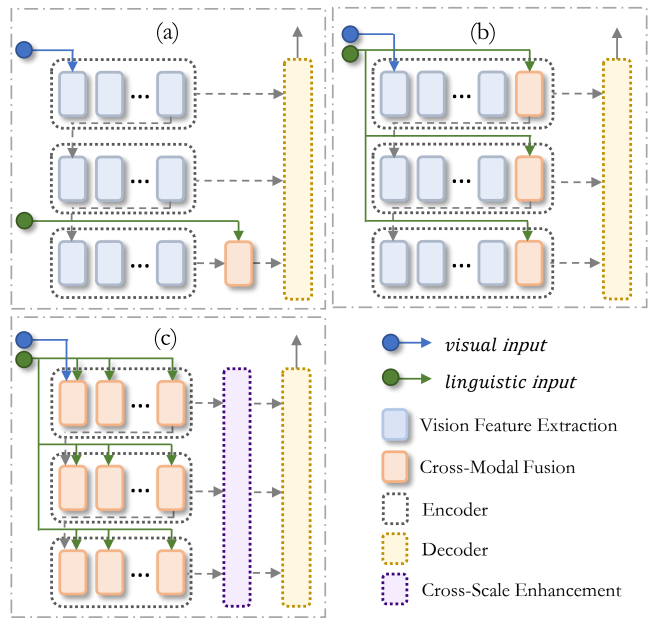

Figure 1: Comparison of existing Transformer-based architectures ((a) and (b)) for referring segmentation with our SLViT (c).

In contrast to linear fusion methods [Hu *et al.*, 2016; Liu *et al.*, 2017] adopting Fully Convolutional Networks (FCN) for feature learning and prediction, various attention mechanisms [Shi *et al.*, 2018; Ye *et al.*, 2019] have been proposed to learn rich visual-linguistic information. Transformers can naturally model long-distance dependencies via attention mechanisms, which are well suited to cross-modal fusion and hence an appropriate choice for referring segmentation task. Therefore, several Vision Transformer (ViT) [Dosovitskiy *et al.*, 2020] methods have been put forth which significantly improve performance for this task. Figure 1(a) illustrates a Transformer-based architecture for referring image segmentation, i.e., VLT [Ding *et al.*, 2021], which fuses vision and language features after vision feature extraction

*Corresponding Authors: Lanfen Lin (llf@zju.edu.cn), Yen-Wei Chen (chen@is.ritsumei.ac.jp).

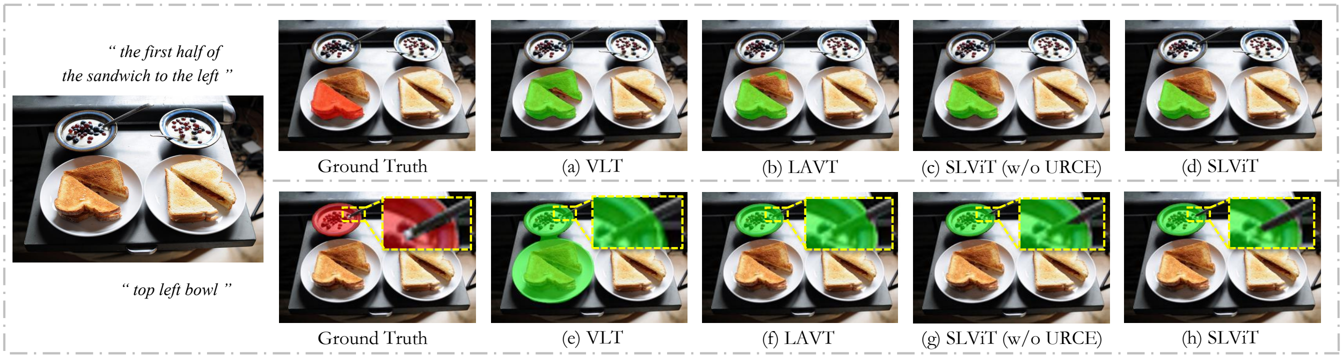

Figure 2: Qualitative results of different approaches. The prediction masks for the image with the referring text *"the first half of the sandwich to the left"* and *"top left bowl"* are shown in (a)-(d) and (e)-(h), respectively.

through encoders. The architecture shown in Fig.1(b) is employed in EFN [Feng *et al.*, 2021] and LAVT [Yang *et al.*, 2022]. This architecture includes a fusion module at the end of each stage to fuse extracted visual features with linguistic modality information. The existing Transformer-based designs take advantage of long-range dependencies and hierarchical structure to enhance performance, although Transformer designs can be further improved. Specifically, visual feature extraction and cross-modal fusion are considered into two independent steps in the existing works, which leaves room for improvement in visual-linguistic alignment in semantic space. Additionally, current approaches adopt sequential structures that result in single-scale representations at each level, despite the fact that multi-scale feature interaction has been shown to be more beneficial for capturing the core semantic information.

By revisiting previous successful works and analyzing requirements of the referring segmentation task, we argue a effective method for such task should have the following characteristics: (i) A robust fusion encoder network to capture local visual and global visual-linguistic information. To accurately pinpoint the target instance with varying characteristics, both rich local visual features and positional global cross-modal relationships are crucial. (ii) Multi-scale information interaction to capture cross-scale dependencies and address complex scale differences. For dense prediction tasks like referring segmentation, the incorporation of complementing information from multiple scales is helpful.

Therefore, taking the aforementioned analysis into account, we propose a novel referring image segmentation architecture (in Figure 1(c)), namely Scale-Wise Language-Guided Vision Transformer (SLViT). In SLViT, we propose an integrated vision-language encoder network design, with a novel attention mechanism called Language-Guided Multi-Scale Fusion Attention (LMFA) to comprehensively extract multi-scale local visual features and model global cross-modal relationships. It improves visual-linguistic alignment in semantic space in a lightweight manner. As shown in Figure 2(a)-(c), LMFA significantly improves performance in locating objects. Considering the spatial correlation between patches at different scales through downsampling, it is beneficial to perform interactions at different scales of the same region for feature refinement. Furthermore, there are regions where the semantic information is temporarily uncertain, making targeted cross-scale enhancement needed. We design a cross-scale feature fusing module named Uncertain Region Cross-Scale Enhancement (URCE) to identify regions of high uncertainty, represented by variance of cross-modal attention scores between scales, and then refine features of the regions using complementary information from multiple scales. The accuracy of boundary identification is improved by URCE, which is qualitatively shown in Figure 2(e)-(h).

In summary, our contributions are three-folded:

1. We propose Language-Guided Multi-Scale Fusion Attention (LMFA) module in our integrated vision-language encoder with the ability of integrated local visual feature extraction and global cross-modal relationships modeling in referring segmentation. LMFA improves visual-linguistic alignment in semantic space.
2. We design a multi-scale feature fusion module named Uncertain Region Cross-Scale Enhancement (URCE). URCE uses the variance of cross-modal correlations between adjacent stages to identify regions of high uncertainty and refines features of these regions with complementing information from multiple stages, which helps in identifying satisfactory boundaries.
3. Based on the aforementioned moudules, we design a novel framework named SLViT for referring segmentation task. We conducted thorough experiments on SLViT with three benchmark datasets, and the experimental results show that SLViT outperforms current state-of-the-art methods with lower computational cost.

2 Related Works

Referring segmentation. For referring segmentation, the early methods [Hu *et al.*, 2016; Liu *et al.*, 2017; Li *et al.*, 2018] directly concatenate visual and linguistic features, adopting FCN for cross-modal feature learning and prediction, lacking attention to the relationship between modalities. Differently, numerous attention-based fusion methods have been proposed for this task. Vision-guided linguistic attention [Shi *et al.*, 2018] and Cross-Modal Self-Attention module [Ye *et al.*, 2019] are proposed to learn visual content correspond-

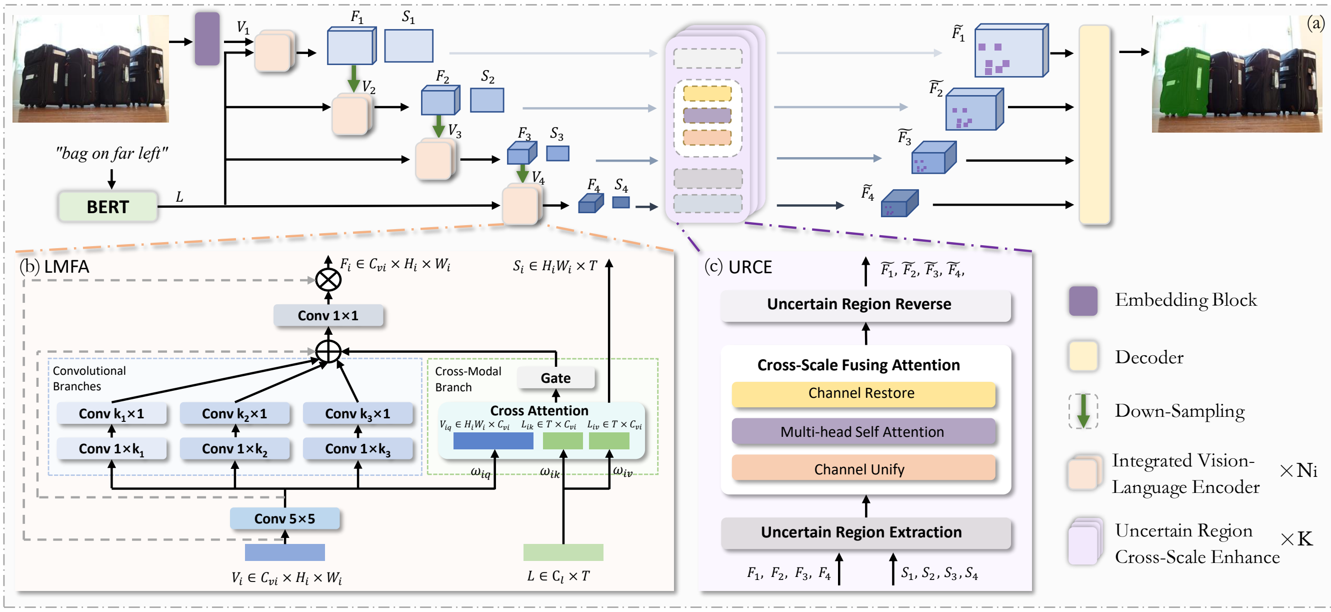

Figure 3: An illustration of SLViT. First, the input image and referring expression pass the embedding block and language encoder BERT respectively to get visual feature V_1 and linguistic feature L, which are sent to Integrated Vision-Language Encoder. Encoders learn useful cross-modal features $F_i, i \in \{1,2,3,4\}$ and records the correlation maps $S_i, i \in \{1,2,3,4\}$ between the two modalities, in which Language-Guided Multi-Scale Fusion Attention (LMFA) captures local visual details and global visual-linguistic cues. The Uncertain Region Cross-Scale Enhancement (URCE) then identifies uncertain regions in the image and enhances patches in $F_i, i \in \{1,2,3,4\}$ corresponding to them, which interacts among different scales. Finally, the reinforced features $\tilde{F}_i, i \in \{1,2,3,4\}$ are sent to the decoder block for final segmentation.

ing to verbal expression. [Hu *et al.*, 2020] capture the mutual guidance between two modalities by a bi-directional relationship inferring network. [Yu *et al.*, 2018] and [Huang *et al.*, 2020] use knowledge about sentence structure to capture attributes in cross-modal features, while [Hui *et al.*, 2020] exploits syntactic structures between words to guide cross-modal context aggregation. Recently, Transformer-based methods have improved the ability to model long-distance cross-modal dependencies and made significant progress in referring segmentation. A VLT framework with encoder-decoder is proposed in [Ding *et al.*, 2021] using attention mechanism to enhance global context information. In EFN [Feng *et al.*, 2021], a collaborative attention mechanism is presented to gradually refine multi-modal features using language and promote cross-modal expression. LAVT [Yang *et al.*, 2022] achieves the early fusion of linguistic and visual features between encoders of ViT.

Vision transformer and multi-scale architecture. Transformer models have been widely used for several computer vision tasks. The ViT model applies self-attention in shallow layers enhancing performance for vision tasks. Since the computational complexity of self-attention is a quadratic polynomial in the number of tokens, it is difficult to directly apply it to a large number of tokens. Therefore, in order to improve the performance of fine-grained tasks such as segmentation, various attention mechanisms have been developed in some recent works [Liu *et al.*, 2021b; Ren *et al.*, 2022; Wang *et al.*, 2021; Guo *et al.*, 2022] to reduce computational cost while retaining valuable visual information. Various studies [Zheng *et al.*, 2021; Gu *et al.*, 2022] using Transformer with multi-scale designs for segmentation tasks have been presented using knowledge between different scales. For referring segmentation task, many Transformer-based methods [Ding *et al.*, 2021; Feng *et al.*, 2021; Yang *et al.*, 2022] exploiting sequential structures have been proposed. These methods leverage sequential structures and lack sufficient multi-scale interaction.

3 Scale-Wise Language-Guided Vision Transformer

3.1 Overview

The proposed SLViT learns local visual and global visual-linguistic cues within scales in an integrated way as well as modeling inter-scale dependencies of uncertain regions. The structure of SLViT is shown in Figure 3.

In a hierarchical manner, we propose the ***[integrated vision-language encoder] - [cross-scale enhancement] - decoder*** framework. The encoder (Sec.3.2) includes a novel lightweight attention module (Sec.3.3) that uses simultaneous multi-scale convolutional operations and gated cross-modal attention to capture local visual features and global visual-linguistic correlations. We also propose to use variance between adjacent stages of cross-modal correlation to assess the uncertainty of regions in the image. For regions of high uncertainty, we design a novel cross-scale feature fusion module (Sec.3.4) to automatically refine mutual regions at different scales via the complementary information among them. Finally, the enhanced representations are sent to the decoder block (Sec.3.5) for the final prediction. In the following subsections, we describe each components of SLViT in detail.

3.2 Integrated Vision-Language Encoder

In order to improve the alignment of visual-linguistic features in semantic space, we propose an integrated vision-language encoder to capture visual and cross-modal features in an integrated way. The block structure of our encoder follows the design of ViT [Dosovitskiy *et al.*, 2020] but we design a novel attention mechanism (Sec.3.3) replacing the conventional self-attention mechanism.

As shown in Figure 3(a), our encoder has a pyramid structure, which contains 4 stages with decreasing spatial resolutions. There are N_i blocks in our encoder for i-th stage. Given input of a pair of an image and a referring expression, our model outputs a segmentation mask for the specified instance. We extract language features via a language encoder BERT [Devlin *et al.*, 2018]. The language feature provided into encoders are denoted as $L \in \mathbb{R}^{C_l \times T}$, where C_l is the number of channel, T is the number of words. The given image passes through an embedding block to obtain initial vision input $V_1 \in \mathbb{R}^{C_{v1} \times H_1 \times W_1}$ of the encoder, where C_{v1} is the number of channels, H_1 and W_1 are height and width of the feature maps in first stage. Each stage contains a down-sampling block and a stack of integrated vision-language encoders. The down-sampling block consists of a convolution with stride of 2 and kernel size of 3×3, followed by a batch normalization layer. For each stage, the stack of integrated cross-modal feature maps F_i can be represented as:

$$F_i = \begin{cases} Ilve(V_1, L), & i = 1 \\ Ilve(Down(F_{i-1}), L), & i = 2, 3, 4 \end{cases} \tag{1}$$

where function $Down(\cdot)$ indicates the down-sampling block, function $Ilve(\cdot)$ indicates blocks in our encoder to catch integrated cross-modal features, i indexes the stage. We obtain the visual inputs of stages $2, 3, 4$ through $V_i = Down(F_{i-1})$.

3.3 Language-Guided Multi-Scale Fusion Attention

As depicted in Figure 3(b), our proposed attention mechanism, namely Language-Guided Multi-Scale Fusion Attention (LMFA), contains four parts: a convolution operation to capture preliminary local feature, a multi-scale convolutional activation to aggregate multi-scale local visual features, a gated cross-modal activation to aggregate global visual-linguistic relationships, and a 1×1 convolution operation to model relationships between branches. In i-th stage, given the visual input $V_i \in \mathbb{R}^{C_{vi} \times H_i \times W_i}$ and the linguistic input $L \in \mathbb{R}^{C_l \times T}$, we obtain the preliminary local visual feature map V_i^{Local} employing a 5×5 convolution operation.

Multi-scale convolutional activation. There are three concurrent convolutional branches with different kernel sizes to capture local features of different receptive fields, which has spatial inductive-bias in modelling rich local visual information. Multi-scale convolutional activation $Att_i^{conv} \in \mathbb{R}^{C_{vi} \times H_i \times W_i}$ can be obtained using the following equation:

$$Att_i^{conv} = \sum_{t=1}^{3} Conv_R^t(Conv_C^t(V_i^{Local})), \tag{2}$$

where t indexes the convolutional branch, $Conv_R^t$ indicates a $1 \times k_t$ convolution function for horizontal linear features, $Conv_C^t$ indicates a $k_t \times 1$ convolution function for vertical linear features. The strip-like convolution kernels aims in obtaining detailed local visual information with low cost.

Gated cross-modal activation. We utilize a gated cross-modal attention to model global visual-linguistic relationships. The steps to get gated cross-modal activation $Att_i^{cross} \in \mathbb{R}^{C_{vi} \times H_i \times W_i}$ are described as follows:

$$V_{iq} = flatten(\omega_{iq}(V_i)), \tag{3}$$

$$L_{ik}, L_{iv} = \omega_{ik}(L), \omega_{iv}(L), \tag{4}$$

$$S_i = V_{iq}^T L_{ik}, \tag{5}$$

$$Att_i^{cross} = Gate(unflatten(softmax(\frac{S_i}{\sqrt{C_l}})L_{iv}^T)), \tag{6}$$

where ω_{iq}, ω_{ik}, ω_{iv} are projection functions, $Gate(\cdot)$ indicates a 1×1 convolution and a GELU function, $flatten(\cdot)$ means unrolling the two spatial dimensions into one dimension in row-major, and $unflatten(\cdot)$ indicates the opposite operation. Here, $S_i \in \mathbb{R}^{H_i W_i \times T}$ is the attention scores between the V_{iq} and L_{ik}, which represents the degree of correlation between two modalities. In the last block of each stage, S_i is provided to URCE. ω_{iq} is implemented as a 1×1 convolution followed by instance normalization with C_{vi} number of output channels. Each of ω_{ik} and ω_{iv} is implemented as a 1×1 convolution with C_{vi} number of output channels.

Integrated attention. We apply a convolution to coordinate convolutional branches and the cross-modal branch obtaining integrated attention weights and reweight the input V_i of LMFA. We obtain the integrated cross-modal feature map $F_i \in \mathbb{R}^{C_{vi} \times H_i \times W_i}$ using the following equation:

$$F_i = Conv_{1\times1}(Att_i^{conv} + Att_i^{cross} + V_i^{Local}) \odot V_i, \tag{7}$$

where $\odot$ is element-wise matrix multiplication operation, and $Conv_{1\times1}$ indicates a 1×1 convolution function to model relationships between branches.

3.4 Uncertain Region Cross-Scale Enhancement

Multi-scale information is crucial for capturing boundary details. To optimize spatial correspondence and minimize redundancy, we propose an Uncertain Region Cross-Scale Enhancement (URCE) module. URCE targets high-uncertainty regions and facilitates interaction across scales within our hierarchical model. Refer to Figure 3(c) for the URCE pipeline.

Uncertain region extraction. Considering computational cost and efficiency, we perform cross-scale enhancement only for the regions with the highest uncertainty.

Visual-linguistic correlation of each patch in i-th stage is indicated by $R_i \in H_{vi}W_{vi}$, which is obtained by $R_i = \sum_1^{C_l} S_i$. Here, S_i is the cross-modal attention score map from LMFA in i-th stage. The variation of visual-linguistic

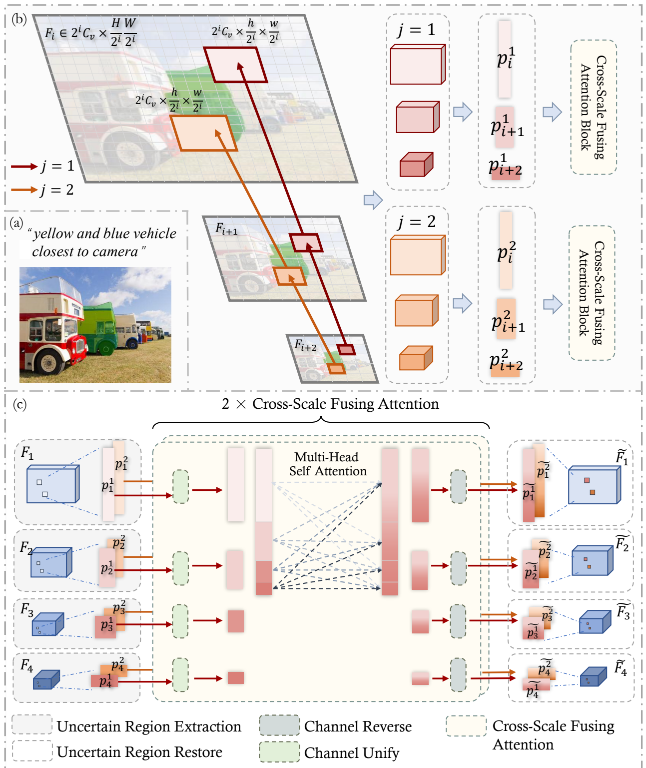

Figure 4: An illustration of the Uncertain Region Cross-Scale Enhancement.

correlation R_i at each coordinate between adjacent stages is used to represent the uncertainty of the corresponding region in the image. We select K most uncertain regions to utilize cross-scale enhancement. The steps are described as follows:

$$map_U = \sum_{i=2}^{4} |Down(R_i) - Down(R_{i-1})|, \tag{8}$$

$$Index = TopK(map_U), \tag{9}$$

where $map_U \in H_{v4}W_{v4}$ means uncertainty map of each coordinate, which is obtained by the correlation difference between adjacent stages. Here, $Down(\cdot)$ indicates function to downsample R_i into size of feature maps of 4-th stage, and $TopK(\cdot)$ indicates the function to find the array index of uncertain regions with largest K values in the array map_U.

The K uncertain regions marked by $Index$, which correspond to patches in different scales, are first rearranged into 2D tensor. Let p_i^j denotes the patch of j-th uncertain region in i-th stage. As an example of $K = 2$, Figure 4(b) illustrates the process of finding the target patches in three successive feature maps F_i, F_{i+1}, F_{i+2} from i-th to (i+2)-th stage with sample inputs (in Figure 4(a)).

Cross-scale fusing attention. We model cross-scale clues for uncertain patches correlated spatially, as shown in Figure 4(c). For j-th uncertain region, patches from multiple stages are passed through *Channel Unify* to have the same channel dimension C_f. Then we concatenate them to obtain cross-scale feature P_{cross}^j corresponding to j-th uncertain region, which can be described as follows:

$$Concat[\omega_p(p_1^j), \omega_p(p_2^j), \omega_p(p_3^j), \omega_p(p_4^j)] \rightarrow P_{cross}^j, \tag{10}$$

where $p_i^j \in 2^i C_v \times \frac{h}{2^i} \frac{w}{2^i}$ is the feature map of j-th uncertain patch in i-th stage, h and w are size of patches in first stage, and ω_p indicates functions to unify the channel to C_f and to rearrange the feature map into 2D tensor. Then we model cross-scale dependencies as follows:

$$\tilde{P_{cross}^j} = MSA(LN(P_{cross}^j)) + P_{cross}^j, \tag{11}$$

where $LN(\cdot)$ indicates the layer normalization operator, and $MSA(\cdot)$ indicates multi-head self-attention. Then we reverse the enhanced sequence back to patches according to the order of concatenation:

$$\omega_p^{reverse}(Split(\tilde{P_{cross}^j})) \rightarrow \tilde{p_1^j}, \tilde{p_2^j}, \tilde{p_3^j}, \tilde{p_4^j}, \tag{12}$$

where $\omega_p^{reverse}$ indicates *Channel Reverse*. Here, $\omega_p^{reverse}$ and $Split(\cdot)$ are inverse operations of previous operations ω_p and $Concat[\cdot]$, respectively. Then we employ *Uncertain Region Restore* that replaces p_i^j with $\tilde{p_i^j}$ to obtain final cross-modal feature maps $\tilde{F_i}$ for each stage.

3.5 Decoder and Segmentation

After cross-scale enhancement, a decoder network is employed to capture high-level semantics. We aggregate features $\tilde{F_1}, \tilde{F_2}, \tilde{F_3}, \tilde{F_4}$ from URCE and use a lightweight Hamburger [Geng *et al.*, 2021] to further model the global context. We obtain the final prediction results by the following equation:

$$Out = Seg(Ham(Concat[\tilde{F_1}, \tilde{F_2}, \tilde{F_3}, \tilde{F_4}])), \tag{13}$$

where F_i is the cross-modal feature maps from stages, $Ham(\cdot)$ indicates a Hamburger function, and $Seg(\cdot)$ indicates a 1×1 convolution and an upsampling function for final prediction.

4 Experiments

4.1 Dataset and Evaluation

We perform experiments on three widely used benchmark datasets for referring image segmentation, including RefCOCO [Yu *et al.*, 2016], RefCOCO+ [Yu *et al.*, 2016], and G-Ref [Mao *et al.*, 2016; Nagaraja *et al.*, 2016]. They have 19,994, 19,992, and 26,711 images respectively, containing 50,000, 49,856, and 54,822 references and 142,209, 141,564, and 104,560 reference expressions.

Following previous works [Wang *et al.*, 2022; Yang *et al.*, 2022], we evaluate our proposed method with overall intersection-over-union (oIoU), mean intersection-over-union (mIoU), and precision at various thresholds. The oIoU is the ratio between the total intersection area and the total union areas. Precision refers to the proportion of test samples with IoU values higher than the threshold.

Method	Language Model	RefCOCO			RefCOCO+			G-Ref		
		val	test A	test B	val	test A	test B	val(U)	test(U)	val(G)
MAttNet [Yu *et al.*, 2018]	Bi-LSTM	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61	-
CMSA [Ye *et al.*, 2019]	None	58.32	60.61	55.09	43.76	47.60	37.89	-	-	39.98
CAC [Chen *et al.*, 2019b]	Bi-LSTM	58.90	61.77	53.81	-	-	-	46.37	46.95	44.32
STEP [Chen *et al.*, 2019a]	Bi-LSTM	60.04	63.46	57.97	48.19	52.33	40.41	-	-	46.40
BRINet [Hu *et al.*, 2020]	LSTM	60.98	62.99	59.21	48.17	52.32	42.11	-	-	48.04
LSCM [Hui *et al.*, 2020]	LSTM	61.47	64.99	59.55	49.34	42.12	43.50	-	-	48.05
CMPC+ [Liu *et al.*, 2021a]	LSTM	62.47	65.08	60.82	50.25	54.04	43.47	-	-	49.89
MCN [Luo *et al.*, 2020b]	Bi-GRU	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	-
EFN [Feng *et al.*, 2021]	Bi-GRU	62.76	65.69	59.67	51.50	55.24	43.01	-	-	51.93
BUSNet [Yang *et al.*, 2021]	Self-Att	63.27	66.41	61.39	51.76	56.87	44.13	-	-	50.56
CGAN [Luo *et al.*, 2020a]	Bi-GRU	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	46.54
ISFP [Liu *et al.*, 2022]	Bi-GRU	65.19	68.45	62.73	52.70	56.77	46.39	52.67	53.00	50.08
LTS [Jing *et al.*, 2021]	Bi-GRU	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25	-
VLT [Ding *et al.*, 2021]	Bi-GRU	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	49.76
ReSTR [Kim *et al.*, 2022]	Transformer	67.22	69.30	64.45	55.78	60.44	48.27	54.48	-	-
CRIS [Wang *et al.*, 2022]	Transformer	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36	-
LAVT [Yang *et al.*, 2022]	BERT	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	60.50
Ours (w/o URCE)	BERT	73.34	75.98	70.21	63.72	68.81	55.72	62.74	63.23	60.55
Ours	BERT	**74.02**	**76.91**	**70.62**	**64.07**	**69.28**	**56.14**	**62.75**	**63.57**	**60.94**

Table 1: Comparison with state-of-the-art methods in terms of overall IoU on three benchmark datasets. U: The UMD partition. G: The Google partition. Language model shows the the main learnable function that transforms word embeddings before multi-modal feature fusion.

4.2 Implementation Details

We conduct experiments using PyTorch library and use BERT implementation from HuggingFace's Transformer library [Wolf *et al.*, 2020]. Convolutions in LMFA's convolutional branches and our decoder are initialized with weights pre-trained on ImageNet-22K from the SegNeXt [Guo *et al.*, 2022]. Language encoder of our model is initialized using official pre-trained weights of BERT with 12 layers and hidden size 768. In convolutional branches of LMFA, we use $k_1 = 7, k_2 = 11, k_3 = 21$ kernel sizes for our convolutions. The rest of weights in our model are randomly initialized.

Following, we use AdamW optimizer with weight decay 0.01. The learning rate is initialed as $3e$-5 and scheduled by polynomial learning rate decay with a power of 0.9. All the models are trained for 60 epochs with a batch size of 16. Each reference has 2-3 sentences on average, and we randomly sample one referring expression per object in a epoch. Image size is adjusted to 480×480 without data augmentation.

4.3 Comparison with the State-of-the-Arts

We compare the performance of our proposed method with state-of-the-art methods on three widely-used datasets using the oIoU metric. Experimental results are reported in Table 1 and the best results are highlighted in bold. As shown, our SLViT without Uncertain Region Cross-Scale Enhancement (w/o URCE) outperforms all other methods. This has an improvement of 1.58 oIoU over the Val Split set of the RefCOCO+ dataset, while on average an improvement of 0.83 oIoU across all 9 validation sets of the three datasets. It indicates the efficacy of integrated vision-language encoder with LMFA to improve visual-linguistic alignment in semantic space. It is helpful in capturing detailed local visual features and modeling global visual-linguistic relationships in an integrated manner. Additionally, an increment of oIoU (+0.93 at most) is achieved when utilizing URCE, which achieves the new SOTA on these datasets. Furthermore, our proposed SLViT uses significantly less GFLOPs and parameters than previously proposed Transformer-based approaches while achieving a higher oIoU score, as shown in Figure 5(b).

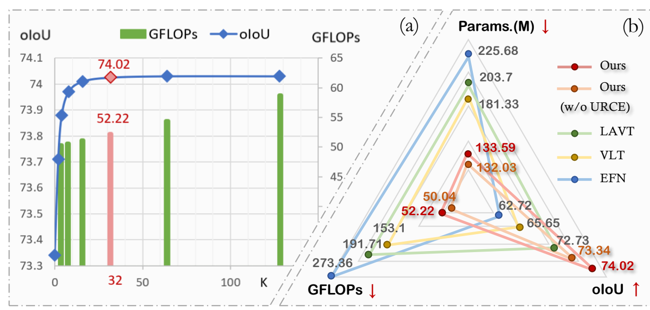

Figure 5: (a) Ablation study using different numbers of uncertain regions. (b) Comparison of our SLViT (w/o URCE), SLViT with Transformer-based methods EFN, VLT and LAVT on Params, GFLOPs and oIoU.

4.4 Ablation Study

Ablation on LMFA design. We have conducted an ablation study on LMFA design on the RefCOCO validation set. Results are shown in Table 2(a)(b). k_tB indicates the convolutional branch containing a $1 \times k_t$ convolution and a $k_t \times 1$ convolution. Gate represents a 1×1 convolution and a GELU function in the cross-modal branch, enhancing the network's adaptive ability. We employed a single convolution branch in

				P@0.5	P@0.7	P@0.9	Mean IoU	Overall IoU
(a) Comparison of kernel sizes with a single convolutional branch								
small		5 B		81.03	70.36	25.88	71.09	69.91
		7 B		81.80	71.28	26.33	71.83	70.08
medium		**11 B**		82.17	71.79	27.96	72.63	71.55
		15 B		82.23	71.75	27.92	72.60	71.56
large		**21 B**		82.68	72.96	29.47	73.02	71.79
		23 B		82.59	72.78	29.36	72.99	71.77
(b) Ablation on design choices								
7B	11B	21B	Gate					
✓	✓			83.26	73.47	29.97	73.44	72.26
	✓	✓		84.31	73.51	30.11	73.89	72.53
✓		✓		84.55	73.64	30.37	74.08	72.68
✓	✓	✓		85.16	74.12	31.00	75.07	73.31
✓	✓	✓	✓	**86.74**	**75.84**	**35.10**	**75.96**	**74.02**
(c) Effectiveness of URCE								
SLViT(w/o URCE)				85.23	74.57	31.36	75.29	73.34
SLViT(w/ URCE)				**86.74**	**75.84**	**35.10**	**75.96**	**74.02**
(d) URCE on various stages								
S1	S2	S3	S4					
		✓	✓	85.66	74.59	30.73	74.49	72.97
✓	✓	✓		86.42	75.51	**35.17**	75.50	73.76
	✓	✓	✓	86.48	75.65	33.79	75.45	73.66
✓	✓	✓	✓	**86.74**	**75.84**	35.10	**75.96**	**74.02**
(e) Features used for final prediction								
F_2, F_3, F_4				84.87	74.13	30.39	75.09	73.27
F_1, F_2, F_3, F_4				85.23	74.57	31.36	75.29	73.34
$\tilde{F}_2, \tilde{F}_4, \tilde{F}_4$				86.12	75.23	34.79	75.57	73.86
$F_1, \tilde{F}_2, \tilde{F}_4, \tilde{F}_4$				85.52	74.43	33.11	75.45	73.64
$\tilde{F}_1, \tilde{F}_2, \tilde{F}_3, \tilde{F}_4$				**86.74**	**75.84**	**35.10**	**75.96**	**74.02**

Table 2: Ablation studies on the RefCOCO validation set.

LMFA to evaluate the impact of various convolution kernel sizes k_t. In Table 2(a), sizes 7 and 21 show superior performance among those with comparable computational costs. Observing Table 2(b), it follows that each part contributes to the final performance.

Number of uncertain regions to enhance. We explore the number of uncertain regions K to utilize cross-scale enhancement. In Figure 5(a), when increasing the number of selected uncertain regions, the oIoU metrics increase sharply at the beginning and then tends to stabilize. The increase in the value of K is linearly correlated with the increasement of computing cost. We choose $K = 32$ as the default setting.

Effectiveness of URCE. To verify the performance of URCE, we have compared SLViT to SLViT (w/o URCE) in Table 2(c). It shows this ablation leads to a drop of 0.68 and 0.67 absolute point in overall IoU and mean IoU respectively, and a drop of an average of 2.18 points in precision across the three thresholds. In addition, *Ours* and *Ours (w/o URCE)* in Figure 5(b) also show that URCE improves performance with a slight increase in parameters and GFLOPs.

Ablation of URCE on various stages. Given integrated cross-modal features from different stages, URCE forms corresponding regions of them into a sequence for joint refinement in single forward pass. Si means the integrated cross-modal feature F_i from i-th stage used as input for URCE. Multiple input sequences are compared in the Table 2(d). The performance boost shows the benefit of multi-scale feature interaction and detailed context in global reasoning.

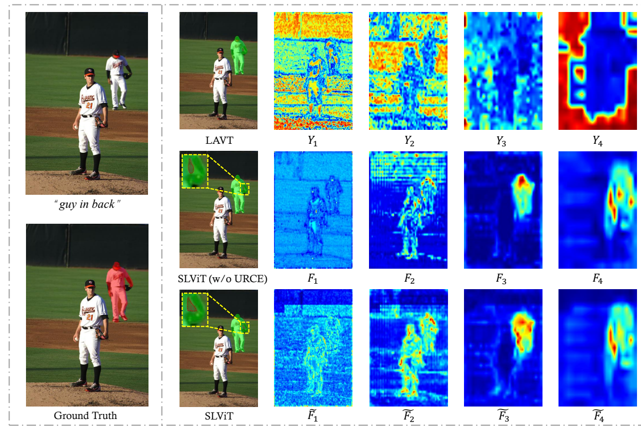

Figure 6: Visualization of the feature maps from different stages in LAVT, SLViT (w/o URCE) and SLViT, respectively.

Ablation of features used for final prediction. We conduct several experiments to assess the influence of the decoder with various sequences as input. As shown in Table 2(e), features $\tilde{F}_1, \tilde{F}_2, \tilde{F}_3, \tilde{F}_4$, which are enhanced by URCE, are the best choices to be sent to the decoder network.

4.5 Interpretation of SLViT

In Figure 6, we visualize the feature maps from LAVT [Yang *et al.*, 2022], our SLViT (w/o URCE) and SLViT. Comparing F_1, F_2, F_3, F_4 from our SLViT (w/o URCE) to Y_1, Y_2, Y_3, Y_4 from LAVT's various stages, the feature maps F_1 to F_4 gradually focus on the target instance and show more visual-linguistic alignment in semantic space as stages go deeper. It indicates that in LMFA, multi-scale convolutional activation learns valuable local visual information, and gated cross-modal activation is useful in identifying the relative position of the target object. We interpret the role of URCE by comparing the segmentation results and the feature maps F_i, $\tilde{F}_i$. Impressively in F_2 and $\tilde{F}_2$, the fence as the background in $\tilde{F}_2$ has been eliminated, which indicates that URCE is able to filter out irrelevant objects. As F_3 and $\tilde{F}_3$ are being observed, the edge of the target object in $\tilde{F}_3$ is more concerned. Therefore, URCE can eliminate interference items and improve the accuracy of boundary prediction by cross-scale enhancement of uncertain regions.

5 Conclusion

In this paper, we propose a novel Transformer-based framework named SLViT for referring image segmentation. SLViT captures rich local visual features and models global visual-linguistic relationships in an integrated manner at each stage. The proposed network design interacts cross-modal features of uncertain regions between different scales with spatial correspondence. Experiments show that SLViT outperforms existing methods on three benchmark datasets with lower computational cost.

Acknowledgments

This work was supported in part by the National Key Research and Development Project (No. 2022YFC2504605), Zhejiang Provincial Natural Science Foundation of China (No. LZ22F020012), Major Technological Innovation Project of Hangzhou (No. 2022AIZD0147), Major Scientific Research Project of Zhejiang Lab (No. 2020ND8AD01).

References

[Chen *et al.*, 2019a] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7454–7463, 2019.

[Chen *et al.*, 2019b] 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.

[Cheng *et al.*, 2014] Ming-Ming Cheng, Shuai Zheng, Wen-Yan Lin, Vibhav Vineet, Paul Sturgess, Nigel Crook, Niloy J Mitra, and Philip Torr. Imagespirit: Verbal guided image parsing. *ACM Transactions on Graphics (ToG)*, 34(1):1–11, 2014.

[Devlin *et al.*, 2018] 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.

[Ding *et al.*, 2021] 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.

[Dosovitskiy *et al.*, 2020] 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.

[Feng *et al.*, 2021] 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.

[Geng *et al.*, 2021] Zhengyang Geng, Meng-Hao Guo, Hongxu Chen, Xia Li, Ke Wei, and Zhouchen Lin. Is attention better than matrix decomposition? *arXiv preprint arXiv:2109.04553*, 2021.

[Gu *et al.*, 2022] Jiaqi Gu, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, and David Z Pan. Multi-scale high-resolution vision transformer for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12094–12103, 2022.

[Guo *et al.*, 2022] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv preprint arXiv:2209.08575*, 2022.

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

[Hu *et al.*, 2020] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4424–4433, 2020.

[Huang *et al.*, 2020] 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.

[Hui *et al.*, 2020] 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 *European Conference on Computer Vision*, pages 59–75. Springer, 2020.

[Jing *et al.*, 2021] 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.

[Kim *et al.*, 2022] 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.

[Li *et al.*, 2018] 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 Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018.

[Liu *et al.*, 2017] 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 International Conference on Computer Vision*, pages 1271–1280, 2017.

[Liu *et al.*, 2021a] 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*, 2021.

[Liu *et al.*, 2021b] 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.

[Liu *et al.*, 2022] Chang Liu, Xudong Jiang, and Henghui Ding. Instance-specific feature propagation for referring segmentation. *IEEE Transactions on Multimedia*, 2022.

[Luo *et al.*, 2020a] 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.

[Luo *et al.*, 2020b] 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.

[Mao *et al.*, 2016] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

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

[Ren *et al.*, 2022] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted self-attention via multi-scale token aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10853–10862, 2022.

[Shi *et al.*, 2018] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–54, 2018.

[Wang *et al.*, 2021] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[Wang *et al.*, 2022] 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.

[Wolf *et al.*, 2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[Yang *et al.*, 2021] 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.

[Yang *et al.*, 2022] 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.

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

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

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

[Zheng *et al.*, 2021] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.