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

Shuyi Ouyang\textsuperscript{1}, Hongyi Wang\textsuperscript{1}, Shiao Xie\textsuperscript{1}, Ziwei Niu\textsuperscript{1}, Ruofeng Tong\textsuperscript{1,3}, Yen-Wei Chen\textsuperscript{2\ast} and Lanfen Lin\textsuperscript{1\ast}

\textsuperscript{1}Zhejiang University
\textsuperscript{2}Ritsumeikan University
\textsuperscript{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.

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.
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 an 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 modules, 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 \( V_i \) 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.

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 of edge 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 \textit{integrated vision-language encoder} - \textit{cross-scale enhancement} - \textit{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. We extract language features via a language encoder to capture linguistic relationships, and a 1 × 1 convolution operation to obtain preliminary local visual feature, a multi-scale convolutional activation to aggregate multi-scale local visual features, which has spatial inductive-bias in modelling rich 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 \( \text{Att}^\text{cross} \in \mathbb{R}^{C_v \times H_i \times W_i} \) are described as follows:

\[
\text{Att}^\text{cross} = \text{Gate}(\text{unflatten}(\text{softmax}(\frac{S_i}{\sqrt{C_v}})T^T_{i,v})),
\]

where \( \omega_{iq}, \omega_{ik}, \omega_{iv} \) are projection functions, \( \text{Gate}(\cdot) \) indicates a 1×1 convolution and a GELU function, \( \text{flatten}(\cdot) \) means unrolling the two spatial dimensions into one dimension in row-major, and \( \text{unflatten}(\cdot) \) indicates the opposite operation. Here, \( S_i \in \mathbb{R}^{H_i \times W_i \times T} \) 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. \( \omega_{iq} \) is implemented as a 1×1 convolution followed by instance normalization with \( C_v \) number of output channels. Each of \( \omega_{ik} \) and \( \omega_{iv} \) is implemented as a 1×1 convolution with \( C_v \) 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 \( F_i \in \mathbb{R}^{C_v \times H_i \times W_i} \) using the following equation:

\[
F_i = \text{Conv}^\text{conv}_1(\text{Att}^\text{conv}_i + \text{Att}^\text{cross} + V_i^{\text{Local}}) \odot V_i,
\]

where \( \odot \) is element-wise matrix multiplication operation, and \( \text{Conv}^\text{conv}_1 \) indicates a 1×1 convolution function to model relationships between branches.

### 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 \( V_i \in \mathbb{R}^{C_v \times H_i \times W_i} \) and the linguistic input \( L \in \mathbb{R}^{C_l \times T} \), we obtain the preliminary local visual feature map \( V_i^{\text{Local}} \) 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 \( \text{Att}^\text{conv}_i \in \mathbb{R}^{C_v \times H_i \times W_i} \) can be obtained using the following equation:

\[
\text{Att}^\text{conv}_i = \sum_{t=1}^3 \text{Conv}^t_R(\text{Conv}^t_C(V_i^{\text{Local}})),
\]
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:

$$\text{map}_U = \sum_{i=2}^{4} |\text{Down}(R_i) - \text{Down}(R_{i-1})|,$$

$$\text{Index} = \text{TopK}(\text{map}_U),$$

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

The $K$ uncertain regions marked by $\text{Index}$, which correspond to patches in different scales, are first rearranged into 2D tensor. Let $p_j^i$ 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 $F_1, F_{i+1}, F_{i+2}$ 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 $P_{\text{cross}}^j$ corresponding to $j$-th uncertain region, which can be described as follows:

$$\text{Concat} [\omega_{p_4}, \omega_{p_3}, \omega_{p_2}, \omega_{p_1}] \rightarrow P_{\text{cross}}^j,$$

where $p_i^j \in 2^i C_f \times \frac{h}{2^i} \frac{w}{2^i}$ is the feature map of $j$-th uncertain patch in $i$-th stage, $h$ and $w$ are size of patches in first stage, and $\omega_i$ 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:

$$P_{\text{cross}}^j = \text{MSA}(\text{LN}(P_{\text{cross}}^j)) + P_{\text{cross}}^j,$$

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

$$\omega_{p_4}^{\text{reverse}}(\text{Split}(P_{\text{cross}}^j)) \rightarrow p_4^j, p_3^j, p_2^j, p_1^j,$$

where $\omega_{p_4}^{\text{reverse}}$ indicates Channel Reverse. Here, $\omega_{p_4}^{\text{reverse}}$ and $\text{Split}(\cdot)$ are inverse operations of previous operations $\omega_{p_4}$ and $\text{Concat}(\cdot)$, 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:

$$\text{Out} = \text{Seg}(\text{Ham}([\text{Concat}(\tilde{F}_1, \tilde{F}_2, \tilde{F}_3, \tilde{F}_4)]),$$

where $F_i$ is the cross-modal feature maps from stages, $\text{Ham}(\cdot)$ indicates a Hamburger function, and $\text{Seg}(\cdot)$ indicates a $1 \times 1$ 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.
### 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 initialized 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 $480 \times 480$ 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).

### 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 $1 \times 1$ convolution and a GELU function in the cross-modal branch, enhancing the network’s adaptive ability. We employed a single convolution branch in a sequential manner. The results show that the adaptive ability of LMFA is improved by using the proposed ablation study.

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

Table 1: Comparison of 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.
Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)

<p>| (a) Comparison of kernel sizes with a single convolutional branch |</p>
<table>
<thead>
<tr>
<th>P@0.5</th>
<th>P@0.7</th>
<th>P@0.9</th>
<th>Mean IoU</th>
<th>Overall IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td>5 B</td>
<td>81.03</td>
<td>70.36</td>
<td>25.88</td>
</tr>
<tr>
<td></td>
<td>7 B</td>
<td>81.80</td>
<td>71.28</td>
<td>26.33</td>
</tr>
<tr>
<td>medium</td>
<td>11 B</td>
<td>82.17</td>
<td>71.79</td>
<td>27.96</td>
</tr>
<tr>
<td></td>
<td>15 B</td>
<td>82.23</td>
<td>71.75</td>
<td>27.92</td>
</tr>
<tr>
<td>large</td>
<td>21 B</td>
<td>82.68</td>
<td>72.96</td>
<td>29.47</td>
</tr>
<tr>
<td></td>
<td>23 B</td>
<td>82.59</td>
<td>72.78</td>
<td>29.36</td>
</tr>
</tbody>
</table>

| (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 | 72.97 |
| 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 |
| F₁, F₂, F₃ | F₁, F₂, F₃, F₄ |
| 84.87 | 74.13 | 30.39 | 75.09 | 73.27 |
| 85.23 | 74.57 | 31.36 | 75.29 | 73.34 |
| 86.12 | 75.23 | 34.79 | 75.57 | 73.86 |
| 85.52 | 74.43 | 33.11 | 75.45 | 73.64 |
| 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 $k_i$. 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 IoU metric increases sharply at the beginning and then tends to stabilize. The increase in the value of $K$ is linearly correlated with the increase in 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. $S_i$ 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.

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 $F_1, F_2, F_3, F_4$, 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 different stages in LAVT, SLViT (w/o URCE) and SLViT, respectively.

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

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.

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
[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


