ReSTR: Convolution-free Referring Image Segmentation Using Transformers

Namyup Kim[1] Dongwon Kim[1] Cuiling Lan[2] Wenjun Zeng[3] Suha Kwak[1]
[1]POSTECH [2]Microsoft Research Asia [3]EIT Institute for Advanced Study
http://cvlab.postech.ac.kr/research/restr/

Abstract

Referring image segmentation is an advanced semantic segmentation task where target is not a predefined class but is described in natural language. Most of existing methods for this task rely heavily on convolutional neural networks, which however have trouble capturing long-range dependencies between entities in the language expression and are not flexible enough for modeling interactions between the two different modalities. To address these issues, we present the first convolution-free model for referring image segmentation using transformers, dubbed ReSTR. Since it extracts features of both modalities through transformer encoders, it can capture long-range dependencies between entities within each modality. Also, ReSTR fuses features of the two modalities by a self-attention encoder, which enables flexible and adaptive interactions between the two modalities in the fusion process. The fused features are fed to a segmentation module, which works adaptively according to the image and language expression in hand. ReSTR is evaluated and compared with previous work on all public benchmarks, where it outperforms all existing models.

1. Introduction

Throughout the recent years, there have been witnessed remarkable advances in semantic segmentation in terms of both efficacy and efficiency [4, 5, 15, 28, 33, 51, 52]. However, its application to real-world downstream tasks is still limited. Since the task is designed to deal with only a predefined set of classes (*e.g.*, "car", "person"), semantic segmentation models are hard to address undefined classes and specific entities of user interest (*e.g.*, "a red Ferrari", "a man wearing a blue hat").

Referring image segmentation [12] has been studied to resolve this limitation by segmenting an image region corresponding to a natural language expression given as query. As this task is no longer restricted by predefined classes, it enables a large variety of applications such as human-robot interaction and interactive photo editing. Referring image segmentation is however more challenging than semantic segmentation since it demands to comprehend individual entities and their relations expressed in the language expression (*e.g.*, "a car behind the taxi next to the building"), and to fully exploit such structured and relational information in the segmentation process. For this reason, models for the task should be capable of capturing interactions between semantic entities in both modalities as well as joint reasoning over the two different modalities.

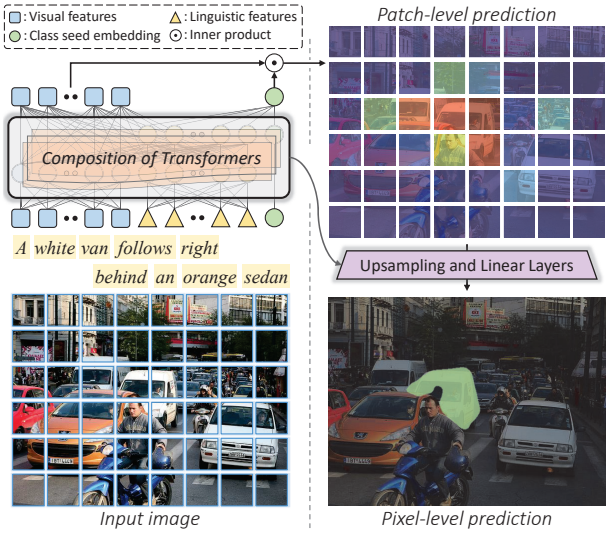

Figure 1. Our convolution-free architecture for Referring image Segmentation using TRansformer (ReSTR) takes a set of non-overlapped image patches and that of word embeddings, and captures intra- and inter-modality interactions by transformers. Then, ReSTR takes a class seed embedding to produce an adaptive classifier which examines whether each image patch contains a part of target entity. Finally, a series of upsampling and linear layers computes a pixel-level prediction in a coarse-to-fine manner.

Existing methods for referring image segmentation [3, 11, 12, 13, 14, 16, 22, 25, 31, 37, 46] have adopted convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to extract visual and linguistic features, respectively. In general, these features are integrated into a multimodal feature map through convolution layers applied to a concatenation of the two features, so-called concatenation-convolution operation. On top of the multimodal feature map, recent methods [11, 13, 14, 16, 46] further employ attention mechanisms [40, 43] so that the fea-

ture map effectively captures interactions between semantic entities. The final multimodal features are then fed as input to a segmentation module.

Although these methods have shown remarkable results on the challenging task, they share the following limitations. First, they have trouble handling long-range interactions between semantic entities within each modality. Referring image segmentation requires to capture such interactions since language expressions often involve complicated relations between entities to precisely indicate target region. In this aspect, both of CNNs and RNNs are limited due to the locality of their basic building blocks. Second, existing models have difficulty in modeling sophisticated interactions between the two modalities. They aggregate visual and linguistic features through the concatenation-convolution operation, which is a fixed and handcrafted way of feature fusion and thus could not be sufficiently flexible and effective to handle a large variety of referring image segmentation scenarios.

To overcome the aforementioned limitations, we propose the first convolution-free model for Referring image Segmentation using TRansformers, dubbed ReSTR. Its overall pipeline is illustrated briefly in Fig. 1. First of all, ReSTR extracts visual and linguistic features through transformer encoders [40]. The two encoders, namely *vision encoder* and *language encoder*, take a set of non-overlapped image patches and that of word embeddings as input, respectively, and extract their features while considering their long-range interactions within each modality. By using transformers for both modalities, we take advantage of capturing global context from the beginning of feature extraction and unifying network topology for the two modalities [32].

Next, a self-attention encoder aggregates the visual and linguistic features into a patch-wise multimodal features. This multimodal fusion encoder enables sophisticated and flexible interactions between features of the two modalities thanks to its self-attention layers. Moreover, the fusion encoder takes a class seed embedding as another input. The class seed embedding is transformed adaptively by the fusion encoder to a classifier for the target entity described in the language expression.

Finally, the outputs of the multimodal fusion encoder, *i.e.*, the patch-wise multimodal features and the adaptive classifier, are fed as input to the segmentation decoder. The decoder computes the final segmentation map in a coarse-to-fine manner. The adaptive classifier is first applied to each multimodal feature as a classifier to examine whether each image patch contains a part of target entity. The coarse, patch-level prediction is then converted into a pixel-level segmentation map by a series of upsampling and linear layers. Thanks to the powerful transformer encoders, this simple and efficient decoder is able to produce accurate segmentation results, achieving the state of the art on four public benchmarks for referring image segmentation.

In summary, the contribution of this work is three-fold:

- Our network is the first convolution-free architecture for referring image segmentation. It captures long-range interactions between vision and language modalities and unifies the network topology for the two different modalities by transformers.
- To encode the fine comprehension of the two modalities, we carefully design the multimodal fusion encoder with the class seed embedding which is transformed to an adaptive classifier for referring image segmentation.
- ReSTR achieves the state of the art on four public benchmarks without bells and whistles.

2. Related Work

2.1. Semantic Segmentation

Semantic segmentation has been significantly improved with the emergence of deep neural networks. Based on a Fully Convolutional Network (FCN) [28] for pixel-level prediction on an end-to-end framework, many approaches are proposed to overcome the several limitations of the network. Since FCN predicts a coarse output mask, the early approaches [1, 4, 26, 54] focus on performing high-resolution predictions. The former studies propose methods to extend the receptive field of CNN by dilated convolutions [5, 47] and to capture multiscale contexts by a feature pyramid pooling scheme [5, 44, 52]. The several approaches propose encoder-decoder structures [6, 23, 33, 36, 49] to model coarse-to-fine framework by multi-level feature fusion. Recently, semantic segmentation has been studied to capture contextual information by attention mechanism [15, 50, 53].

However, the mentioned methods have used variants of FCN architecture that limit local context encoding by convolutional layers. Moreover, since this task is defined to predict segmentation masks within only a predefined set of classes, semantic segmentation models have limitations to apply to real applications.

2.2. Referring Image Segmentation

In contrast to predefined pixel-level classification as semantic segmentation, referring image segmentation aims at grouping the pixels as mask corresponded to a given natural language expression. The pioneering work [12] proposes extracting visual and linguistic features from CNN and RNN, respectively, and generating multimodal features by concatenating tiled linguistic features and visual feature maps. Based on this framework, the early approaches suggest the methods to perform high-resolution prediction by ConvLSTM [25] and the encoder-decoder architecture by intermediate connections [22]. Follow-up studies propose

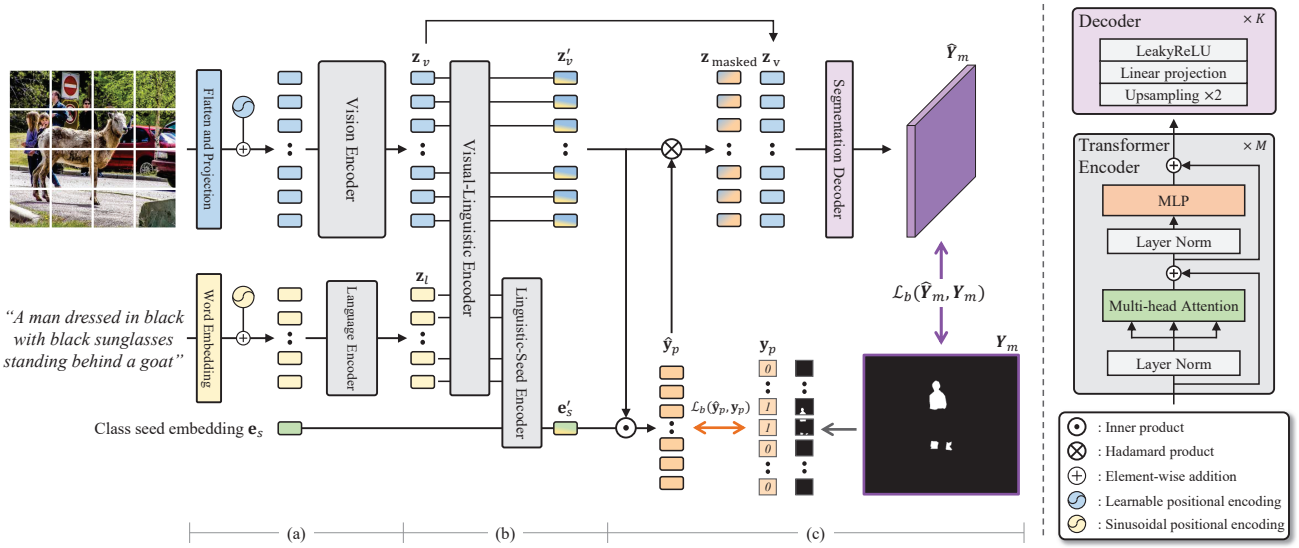

Figure 2. (*Left*) Overall architecture of ReSTR. (a) The feature extractors for the two modalities are composed on transformer encoders, independently. (b) The multimodal fusion encoder consists of the two transformer encoders: *the visual-linguistic encoder* and *the linguistic-seed encoder*. (c) The coarse-to-fine segmentation decoder transforms a patch-level prediction to a pixel-level prediction. (*Right*) Transformer encoder used in all encoders and the composition of the coarse-to-fine segmentation decoder.

an attention mechanism to fuse the visual and linguistic features and multi-level feature aggregation to produce high-resolution segmentation maps [3, 11, 13, 46]. Recent studies [8, 19] suggest a multimodal fusion encoder using transformers [40] to capture long-range interactions between visual and linguistic features.

Unlike the existing work, we propose a new convolution-free architecture to encode contextual information at every stage of our model and efficiently transform a patch-level prediction to a high-resolution segmentation map in a coarse-to-fine manner.

2.3. Vision Transformer

From the introduction of transformers by [40] as a self-attention module for NLP, many approaches adopt this module in computer vision tasks for the advantages of this module including long-range dependencies, dynamic kernel depended on input, and less visual inductive bias than CNNs. Several studies employ transformers for an attention module in/on CNNs as a CNN-transformer hybrid network [2, 35, 38, 41, 43, 51, 55]. Recent approaches replace CNNs with transformers as a convolution-free architecture in image classification [9, 17, 27, 42], object detection [27, 42], semantic segmentation [27, 39, 42, 55] and multimodal learning [32]. In particular, transformers are deployed to semantic segmentation tasks to overcome the inherent limitation of FCN-like architecture. For example, Zheng *et al.* [55] utilize transformer backbone as a global context feature extractor and then convolutional layers for a decoder in the hybrid manner. Strudel *et al.* [39] propose a convolution-free architecture for semantic segmentation by self-attention with visual features and a set of learnable classes embeddings. Inspired by the paradigm, we adopt transformers for referring image segmentation for the above advantages and use an adaptive classifier as an extension of the learnable class queries used in semantic segmentation transformers [39, 41].

3. Proposed Method

This section elaborates on ReSTR, our convolutional-free transformer network for referring image segmentation. Its detailed architecture is illustrated in Fig. 2. To capture long-range interactions for each modality, ReSTR first extracts visual and linguistic features by transformer encoders [40] independently (Sec. 3.1). Then, it forwards visual and linguistic features in parallel to a multimodal fusion encoder to capture fine relations across these two modalities (Sec. 3.2). Finally, an efficient decoder for a coarse-to-fine segmentation converts patch-level prediction into high-resolution pixel-level prediction (Sec. 3.3).

3.1. Visual and Linguistic Feature Extraction

To extract visual and linguistic features, we choose transformers [9] for both modalities. A transformer encoder is M sequential transformers, each of which consists of Multi-headed Self-Attention (MSA), Layer Normalization (LN), and Multilayer Perceptron (MLP) blocks:

$$\bar{\mathbf{z}}_{i+1} = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_i)) + \mathbf{z}_i, \tag{1}$$

$$\mathbf{z}_{i+1} = \mathrm{MLP}(\mathrm{LN}(\bar{\mathbf{z}}_{i+1})) + \bar{\mathbf{z}}_{i+1}, \tag{2}$$

where $\mathbf{z}_i \in \mathbb{R}^{N \times D}$ denotes input features of the i-th layer of the transformer encoder, N is the input size of each modality, and D is the channel dimension of the features. LN is

applied to the output of the transformer encoder. MSA is composed of k Self-Attention (SA) operations on queries $\mathbf{q} \in \mathbb{R}^{N \times D_h}$, keys $\mathbf{k} \in \mathbb{R}^{N \times D_h}$, and values $\mathbf{v} \in \mathbb{R}^{N \times D_h}$, which are obtained by linear projections of input features $\mathbf{z}$, independently:

$$\mathrm{MSA}(\mathbf{z}) = [\mathrm{SA}_1(\mathbf{z}), \mathrm{SA}_1(\mathbf{z}), \cdots, \mathrm{SA}_k(\mathbf{z})]\mathbf{W}_{\mathrm{MSA}}, \quad (3)$$

$$\mathrm{SA}(\mathbf{z}) = A\mathbf{v}, \quad (4)$$

$$A = \mathrm{softmax}(\mathbf{q}\mathbf{k}^\top / \sqrt{D_h}), \quad (5)$$

where $A \in \mathbb{R}^{N \times N}$ is dot-product attention, $[\cdot, \cdot]$ denotes concatenation, and $\mathbf{W}_{\mathrm{MSA}} \in \mathbb{R}^{kD_h \times D}$ is a linear projection. D_h is set to D/k following [9]. The transformer encoder, which is composed of transformers, is denoted by $\mathrm{Transformers}(\cdot)$.

Vision encoder. An input image $\mathbf{x}^v \in \mathbb{R}^{H \times W \times C_v}$ is transformed to a set of patch embeddings $\mathbf{x}^p \in \mathbb{R}^{N_v \times D_v}$ by splitting the input image into the patches without overlapping and mapping them with a linear projection. Let $N_v = HW/P^2$ be the number of patches, P be the patch size, and D_v be the projected channel dimension. We add learnable 1D positional encoding $\mathbf{E}^v_{\mathrm{pos}} \in \mathbb{R}^{N_v \times D_v}$ to the patch embeddings to obtain input to the vision encoder, $\mathbf{z}^v_0 = \mathbf{x}^p + \mathbf{E}^v_{\mathrm{pos}}$. We feed $\mathbf{z}^v_0$ into the vision encoder to produce the patch-wise visual features $\mathbf{z}_v \in \mathbb{R}^{N_v \times D_v}$:

$$\mathbf{z}_v = \mathrm{Transformers}(\mathbf{z}^v_0; \boldsymbol{\theta}_v), \quad (6)$$

where $\boldsymbol{\theta}_v$ are the parameters of the vision encoder.

Language encoder. We transform a natural language expression to a set of word embeddings $\mathbf{x}^l \in \mathbb{R}^{N_l \times C_l}$, where N_l is the maximum length of sentence and C_l is the dimension of the word embeddings. We add a sinusoidal 1D positional encoding $\mathbf{e}^l_{\mathrm{pos}} \in \mathbb{R}^{N_l \times C_l}$ to the word embeddings, as $\mathbf{z}^l_0 = \mathbf{x}^l + \mathbf{e}^l_{\mathrm{pos}}$. Linguistic features $\mathbf{z}_l \in \mathbb{R}^{N_l \times D_l}$ are generated by feeding $\mathbf{z}^l_0$ into the language encoder which consists of transformers.

3.2. Multimodal Fusion Encoder

The multimodal fusion encoder consists of two transformer encoders, namely *visual-linguistic encoder* and *linguistic-seed encoder* as shown in Fig. 2 (b). In detail, we use the visual features $\mathbf{z}_v$, the linguistic features $\mathbf{z}_l$ and a class seed embedding $\mathbf{e}_s \in \mathbb{R}^{1 \times D}$ as input for the multimodal fusion encoder. $\mathbf{e}_s$ is the trainable parameters, initialized randomly. We first normalize $\mathbf{z}_v$ and $\mathbf{z}_l$ and feed each of them into a different linear layer to adjust their channel dimension to be the same as D. Then, the visual-linguistic encoder takes the visual and linguistic features as inputs to produce patch-wise multimodal features $\mathbf{z}'_v \in \mathbb{R}^{N_v \times D}$:

$$[\mathbf{z}'_v, \mathbf{z}'_l] = \mathrm{Transformers}([\mathbf{z}_v, \mathbf{z}_l]; \boldsymbol{\theta}_{vl}), \quad (7)$$

where $\boldsymbol{\theta}_{vl}$ are the parameters of the visual-linguistic encoder, $\mathbf{z}'_l \in \mathbb{R}^{N_l \times D}$ denotes visual-attended linguistic features. Since the visual and linguistic features are fed into the visual-linguistic encoder in parallel, we obtain the patch-wise multimodal features by fine and flexible interactions between the visual and linguistic features.

Then, we feed the class seed embedding $\mathbf{e}_s$ and the visual-attended linguistic features $\mathbf{z}'_l$ into the linguistic-seed encoder:

$$\mathbf{e}'_s = \mathrm{Transformers}([\mathbf{z}'_l, \mathbf{e}_s]; \boldsymbol{\theta}_{ls}), \quad (8)$$

where $\boldsymbol{\theta}_{ls}$ are the parameters of transformers for the linguistic-seed encoder, and $\mathbf{e}'_s \in \mathbb{R}^{1 \times D}$ is an adaptive classifier. Since a single fixed classifier is not sufficient for referring segmentation where a target mask varies by a language expression in hand, $\mathbf{e}'_s$ acts as an adaptive classifier that examines if each patch contains a part of a target entity.

The multimodal fusion encoder is designed to produce the adaptive classifier that satisfies the following two requirements demanded in referring image segmentation. First, since referring image segmentation aims to segment a region corresponding to a language expression, the adaptive classifier should comprehend fine relations of the language expression. Moreover, since an input image has regions irrelevant to the language expression (*e.g.*, background), the class seed embedding directly attending to the visual information can lead to an adaptive classifier corrupted by the irrelevant regions. Nevertheless, since the appearance of the target entity described in a language expression can differ by images, it is beneficial to produce the adaptive classifier using the visual-attended linguistic features.

Therefore, we build the multimodal fusion encoder using these two transformer encoders alternatively to generate the adaptive classifier that meets the aforementioned conditions. We empirically verify the superiority of our multimodal fusion encoder in Sec. 4.3.

3.3. Coarse-to-Fine Segmentation Decoder

A patch-level prediction $\hat{\mathbf{y}}_p \in \mathbb{R}^{N_v \times 1}$ is calculated by an inner product between the patch-wise multimodal features $\mathbf{z}'_v$ and the adaptive classifier $\mathbf{e}'_s$:

$$\hat{\mathbf{y}}_p = \sigma\left(\frac{\mathbf{z}'_v \mathbf{e}'^\top_s}{\sqrt{D}}\right), \quad (9)$$

where σ is the sigmoid function and $\sqrt{D}$ is a normalization factor [40].

We suggest an efficient segmentation decoder to compensate for the low-resolution patch-level prediction (*e.g.*, $N_v = H/P \times W/P$). First, the decoder produces masked multimodal features $\mathbf{z}_{\mathrm{masked}} \in \mathbb{R}^{N_v \times D}$:

$$\mathbf{z}_{\mathrm{masked}} = \mathbf{z}'_v \otimes \hat{\mathbf{y}}_p, \quad (10)$$

where $\otimes$ denotes Hadamard product operation over the channel dimension D. Then, before forwarding to the segmentation decoder, we concatenate the patch-wise visual features and the masked multimodal features as $[\mathbf{z}_v, \mathbf{z}_{\text{masked}}] \in \mathbb{R}^{N_v \times 2D}$ to guide the segmentation decoder through visual semantics. The segmentation decoder is composed of K sequential blocks, each of which consists of upsampling with factor 2, linear projection with channel reduction by $1/2$ of input dimension, and activation function, where $K = \log P$ and P is the patch size. Finally, the output features of the decoder is fed into a linear projection and reshaped to generate a pixel-level prediction $\hat{Y}_m \in \mathbb{R}^{H \times W \times 1}$. At inference time, we only use the pixel-level prediction $\hat{Y}_m$ as the final prediction.

For patch-level classification, we generate patch-level labels by splitting the ground-truth label $Y_m \in \mathbb{R}^{H \times W \times 1}$ into a set of patch labels whose number of patches is the same as the patch-level prediction $\hat{\mathbf{y}}_p \in \mathbb{R}^{N_v \times 1}$ by following criteria:

$$\mathbf{y}_p^i = \begin{cases} 1, & \text{if } h(p_{ij}) > \tau \\ 0, & \text{otherwise} \end{cases}, \tag{11}$$

where $\mathbf{y}_p^i$ denotes the patch-level labels of the i-th patch p_i, j is the number of pixels in a patch, $h(\cdot)$ indicates the average pooling over spatial dimension, and τ is a thresholding hyperparameter.

The network is trained by the binary cross-entropy loss $\mathcal{L}_b(\hat{Y}, Y)$ on the patch-level prediction $\hat{\mathbf{y}}_p$ and the pixel-level prediction $\hat{Y}_m$:

$$\mathcal{L}(\hat{\mathbf{y}}_p, \mathbf{y}_p, \hat{Y}_m, Y_m) = \lambda \mathcal{L}_b(\hat{\mathbf{y}}_p, \mathbf{y}_p) + \mathcal{L}_b(\hat{Y}_m, Y_m), \tag{12}$$

where λ is a balancing hyperparameter.

4. Experiments

4.1. Experimental Setting

Datasets. We conduct experiments on four datasets, ReferIt [20], UNC [48], UNC+ [48], and Gref [30], which are widely used in referring image segmentation task. ReferIt [20] contains 19,894 images with 130,525 language expressions for 96,654 masks which are collected from IAPR TC-12 [10]. UNC, UNC+, and Gref are collected from COCO [24] dataset. UNC and UNC+ consist of 19,994 images with 142,209 language expressions for 50,000 masks and 19,992 images with 141,564 language expressions for 49,856 masks, respectively. The difference between UNC and UNC+ is that UNC+ does not contain the words that indicate location properties (*e.g.*, left, top, front) in expressions and contains the only appearance expressions. Gref contains 25,711 images with 104,560 language expressions for 54,822 objects.

Implementation details. We use ViT-B-16 [9] pretrained on ImageNet-21K [7] for the vision encoder which has 12 layers, 16 patch size, 768 channel dimensions, 12 heads of MSA, and 3,072 dimensions of channel expansion in MLP. We use pretrained GloVe [34] embeddings for language expressions. The language encoder consists of 6 transformer layers, and has 300 channel dimensions as GloVe embeddings, 12 heads of MSA and 3,072 dimensions of channel expansion in MLP. The maximum length of a language expression N_l is set to 20 following previous work. The multimodal fusion encoder consists of the same transformer as the vision encoder. The number of layers of the segmentation decoder is 4 since the patch size is 16. In all experiments, the models are optimized by AdamW [29] with weight decay of $5e-4$; the initial learning rate is $1e-5$ and decreases with polynomial decay [4]. We set a batch size of 8 and train for 400,000 iterations with warm-up period for 40,000 iterations to reach the initial learning rate. We resize input images to 480×480. We set τ in Eq. (11) and λ in Eq. (12) to 0.8 and 0.1 for all experiments, respectively.

Evaluation protocol. Following previous work [12, 25], we adopt the cumulative Intersection-over-Union (IoU) metric, where total intersections are divided by the total unions over all test samples. Then, we evaluate the accuracy at the $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ IoU thresholds.

4.2. Comparisons with the State of the Art

We compare ReSTR with other referring image segmentation models on four benchmarks. As summarized in Table 1, ReSTR achieves outstanding performance without inefficient postprocessing (*e.g.*, DenseCRF [21]) compared with the previous arts on all public benchmarks except for UNC+ *testB* set. Following [25], we discuss the relationship between language expression length and performance as summarized in Table 2. The results demonstrate the ReSTR clearly outperforms previous methods on most groups of expression length except for the 1-5 length group on Gref *val* set. Moreover, the performance of ACM using an attention mechanism for long-range interactions between the two modalities drops 13.71%p from 1-5 to 11-20 length group on the Gref *val* set, while that of ReSTR drops by 6.81%p. It demonstrates that our method is better to capture the long-range interactions between the two modalities compared to previous methods. Note that the recent methods [8, 18, 45] use a visual backbone pretrained on COCO object detection dataset and evaluate their models on only three benchmarks based on COCO dataset. In contrast, our visual backbone is pretrained for ImageNet classification, and ReSTR is evaluated on all benchmarks.

4.3. Analysis of Variants of Fusion Encoder

To verify our design choice for the multimodal fusion encoder, we investigate variants of the fusion encoder. We use 4 transformer layers, denoted as $\{f_1, f_2, f_3, f_4\}$, in all variants of the encoder.

Methods	DCRF	ReferIt	UNC			UNC+			Gref
		test	*val*	*testA*	*testB*	*val*	*testA*	*testB*	*val*
LSTM-CNN [12]		48.03	-	-	-	-	-	-	28.14
RMI [25]	✓	58.73	45.18	45.69	45.57	29.86	30.48	29.50	34.52
DMN [31]		52.81	49.78	54.83	45.13	38.88	44.22	32.29	36.76
RRN [22]	✓	63.63	55.33	57.26	53.95	39.75	42.15	36.11	36.45
CMSA [46]	✓	63.80	58.32	60.61	55.09	43.76	47.60	37.89	39.98
STEP [3]		64.13	60.04	63.46	57.97	48.19	52.33	40.41	46.40
BRINet [13]	✓	63.46	61.35	63.37	59.57	48.57	52.87	42.13	48.04
LSCM [16]	✓	66.57	61.47	64.99	59.55	49.34	53.12	43.50	48.05
CMPC [14]	✓	65.53	61.36	64.54	59.64	49.56	53.44	43.23	49.05
ACM [11]		66.70	62.76	65.69	59.67	51.50	55.24	43.01	51.93
BUSNet [45]	✓	-	63.27	66.41	61.39	51.76	56.87	44.13	50.56
LTS [18]		-	65.43	67.76	63.08	54.21	58.32	48.02	54.40
VLT [8]		-	65.65	68.29	62.73	55.50	59.20	**49.36**	52.99
ReSTR (Ours)		**70.18**	**67.22**	**69.30**	**64.45**	**55.78**	**60.44**	48.27	**54.48**

Table 1. Quantitative results on four datasets in IoU (%). DCRF denotes using post-procession by DenseCRF [21]. The best results are in bold, while second-best ones are underlined.

	Length	1-5	6-7	8-10	11-20
Gref	R+RMI [25]	35.34	31.76	30.66	30.56
	BRINet [13]	51.93	47.55	46.33	46.49
	ACM [11]	**59.92**	52.94	49.56	46.21
	ReSTR (Ours)	58.72	**53.47**	**53.96**	**51.91**
	Length	1-2	3	4-5	6-20
UNC	R+RMI [25]	44.51	41.86	35.05	25.95
	BRINet [13]	65.99	64.83	56.97	45.65
	ACM [11]	68.73	65.58	57.32	45.90
	ReSTR (Ours)	**72.38**	**69.46**	**61.19**	**50.21**
	Length	1-2	3	4-5	6-20
UNC+	R+RMI [25]	35.72	25.41	21.73	14.37
	BRINet [13]	59.12	46.89	40.57	31.32
	ACM [11]	61.62	52.18	43.46	31.52
	ReSTR (Ours)	**65.72**	**54.81**	**47.65**	**37.02**
	Length	1	2	3-4	5-20
ReferIt	R+RMI [25]	68.11	52.73	45.69	34.53
	BRINet [13]	75.28	62.62	56.14	44.40
	ACM [11]	78.19	66.63	60.30	46.18
	ReSTR (Ours)	**80.82**	**69.78**	**63.66**	**50.73**

Table 2. Performance according to variants of referring length on Gref, UNC, UNC+ and ReferIt in IoU (%). The best results are in bold, while second-best ones are underlined.

First, as illustrated in Fig. 3(a), we present a variant of the fusion encoder which takes all features simultaneously as inputs, denoted as Vanilla Multimodal Encoder (VME). Since all inputs are given in parallel, VME can learn the fine relations between all features. However, the adaptive classifier can be undesirably biased to the visual features by the imbalance of the length of features between visual and linguistic features ($N_v \gg N_l$). As shown in Table 3(a), we measure attention scores of the visual and linguistic features to the class seed embedding. In detail, we split the attentions of the class seed embedding $\mathbf{a} \in \mathbb{R}^{1\times(N_v+N_l+1)}$ in the attention matrix A in Eq. (5) into the visual and linguistic attentions $\mathbf{a}_v \in \mathbb{R}^{1\times N_v}$ and $\mathbf{a}_l \in \mathbb{R}^{1\times N_l}$, respectively. Then, we sum each modality attention across feature dimension to obtain a_v^i and a_l^i as the attention score of VME of i-th transformer layer. Finally, we average the attention score of each layer over the dataset. The results demonstrate that the attentions for the class seed embedding are biased to the visual features. We hypothesize that the bias of attentions results from the imbalance of the length of features between the visual and linguistic features, which is $N_v : N_l = 900 : 20$ in our experiments, that leads to the adaptive classifier capturing less fine relations of a language expression.

Layer	a_v	a_l
f_1	82.4	16.8
f_2	98.9	1.0
f_3	98.7	1.2
f_4	98.1	1.7

(a)

Encoder	# params	MACs	IoU
VME	28.35M	31.36G	51.27
IME	28.35M	15.96G	45.89
CME	28.35M	15.96G	**52.81**
CME†	14.18M	15.96G	52.79

(b)

Table 3. (a) Averaged attention score (%) of the visual and linguistic features to the class seed embedding at each transformer layer of VME on Gref *train* set. (b) Performance of the variants of the multimodal fusion encoder on Gref *val* set in IoU (%). † denotes the fusion encoder with weight sharing. The best results are in bold, while second-best ones are underlined.

To resolve this problem, we consider disconnecting interactions between the visual features and the class seed embedding as illustrated in Fig. 3(b), denoted as Independent Multimodal Encoder (IME). In other words, the class seed embedding interacts with only the linguistic features. Therefore, IME restricts the class seed embedding from being adaptively transformed to an adaptive classifier with the visual information.

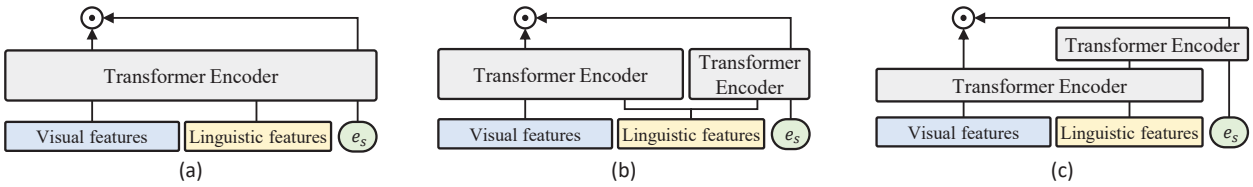

Figure 3. The variants of the multimodal fusion encoder based on transformer architecture. (a) Self-attention fusion encoder on all sequences in parallel. (b) Independent fusion encoder between the visual features and the class seed embedding. (c) Indirect conjugating fusion encoder between the visual features and the class seed embedding.

Encoder *# layers*	Encoder *w share*	Decoder	Prec@0.5	Prec@0.6	Prec@0.7	Prec@0.8	Prec@0.9	IoU
2			52.60	45.59	36.59	23.54	5.23	48.12
		✓	52.86	36.61	38.93	26.37	7.90	48.43
4			61.77	55.86	46.86	30.88	8.18	52.81
		✓	64.91	59.94	51.73	37.70	12.23	**54.48**
	✓	✓	64.27	59.01	50.70	35.85	11.46	54.07
6			63.36	57.88	48.75	33.46	8.75	52.84
	✓		63.05	57.32	48.19	32.47	8.47	52.59

Table 4. Performance for ablation study of ReSTR on Gref *val* set. *# layers* denotes the number of transformer layers in the multimodal fusion encoder. *w share* denotes weight sharing of the multimodal fusion encoder.

To this end, we propose a structure that indirectly conjugates the class seed embedding and the visual features with the linguistic features as medium, denoted as indirect Conjugating Multimodal Encoder (CME) as illustrated in Fig. 3(c). As mentioned in Sec. 3.2, the design aims to avoid interaction between the irrelevant visual features and the class seed embedding by indirectly interactions via the linguistic features. Furthermore, CME produces the adaptive classifier for the target entity described in the language expression by fine interactions between the linguistic features and the class seed embedding.

As summarized in Table 3(b), we compare the three variants of the multimodal fusion encoder on performance, computational cost (MACs), and the number for parameters (# params) of these encoders without the segmentation decoder. These results demonstrate the superiority of CME over the other variants of the fusion encoder in performance and efficiency. In addition, we also experiment CME with weight sharing (CME†) between transformer layers of the visual-linguistic encoder and between those of the linguistic-seed encoder. The result shows CME† is still better performance with lower parameters and computational cost than the other variants.

4.4. In-depth Analysis of ReSTR

We investigate our framework on the *val* set of Gref dataset which contains the longer and more complicated language expressions than the others.

Effect of the number of transformer layers in the multimodal fusion encoder. We study the impact of the number of transformer layers in the multimodal fusion encoder by varying the number of transformers to $\{2, 4, 6\}$. Since the multimodal fusion encoder is composed of two transformer encoders, the encoder always has an even number of the transformer layers. As summarized in Table 4, the performance is fairly increased until using 4 transformer layers and marginally increased using 6 transformer layers.

Effect of the segmentation decoder. We investigate the contribution of the segmentation decoder. As summarized in Table 4, the decoder improves IoU by 1.67%p when used along with the 4 transformer layers in the fusion encoder. However, when with the fusion encoder with 2 transformer layers, the improvement made by the segmentation decoder is only 0.31%p. When coupled with the shallow fusion encoder that produces relatively larger potion of false patch-level predictions, the effect of the segmentation decoder is marginal since it is trained to refine the mask of the positive patches. The results demonstrate that the decoder is specialized to refine a patch-level prediction to a fine pixel-level prediction. Note that the analysis of the segmentation decoder is examined except for the fusion encoder with 6 transformer layers due to the memory shortage.

Effect of weight sharing. In Table 4, we also present the performance of the model with weight sharing. Using weight sharing, the number of parameters remains the same regardless of the number of transformer layers that the multimodal fusion encoder contains. The results show that the performance degradation incurred by weight sharing is marginal. It demonstrates that ReSTR could be used in an efficient manner with little loss of performance using weight sharing.

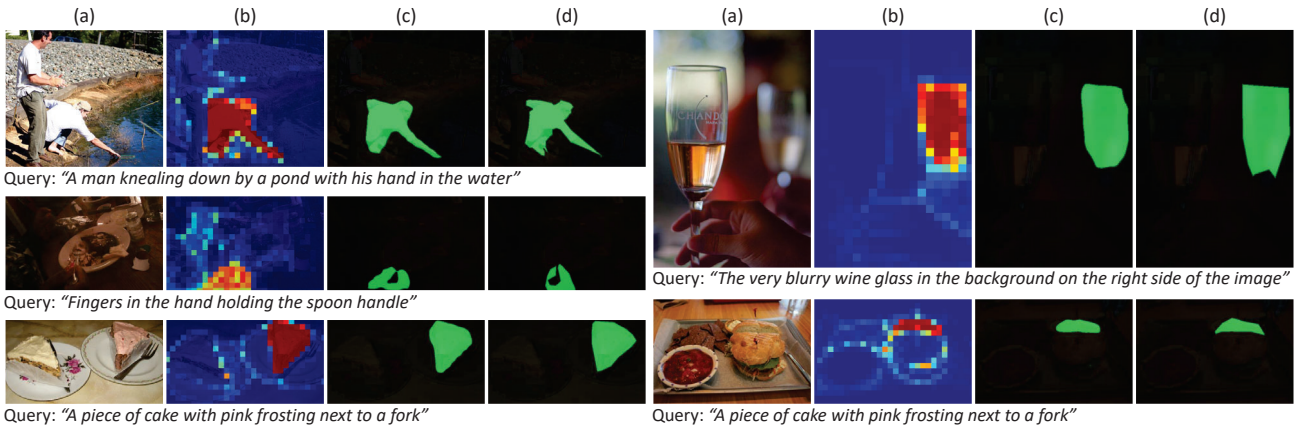

Figure 4. Qualitative results of ReSTR on Gref *val* set. (a) Input image. (b) Patch-level prediction. (c) ReSTR. (d) Ground truth.

Figure 5. Visualization examples of ReSTR according to different language expression queries for an image on Gref *val* set.

Methods	DCRF	# params	MACs	IoU
BRINet [13]	✓	241.18M	367.63G	48.04
LSCM [16]	✓	127.91M	130.45G	48.05
CMPC [14]	✓	118.66M	126.66G	49.05
ACM [11]	✗	232.78M	124.68G	51.93
ReSTR (CME)	✗	122.87M	52.29G	**54.48**
ReSTR (CME†)	✗	108.70M	52.29G	54.07

Table 5. Comparison of computations and performance with recent methods. Both are evaluated on Gref *val* set in IoU (%). † denotes the multimodal fusion encoder with weight sharing and MACs is computed with an input image of 320×320.

Qualitative analysis. As illustrated on Fig. 4, the patch-level predictions of ReSTR are roughly localized on the target patches and the boundaries of relational objects. Then, the patch-level predictions are transformed to fine pixel-level predictions by the segmentation decoder in a coarse-to-fine manner. Moreover, in Fig. 5, we provide the visualization examples of the predictions when varying language expressions are given as queries. These visualizations show that ReSTR is able to predict the segmentation masks corresponding to different language expressions on an image.

Computation cost analysis. In Table 5, we present the number of parameters and MACs of ReSTR and recent studies whose codes are publicly available. ReSTR achieves the best accuracy with the least computation since it employs the efficient segmentation decoder. Also, the size of the visual feature used in previous work is 4 times bigger than ours.

5. Conclusion

We have proposed ReSTR, the first convolution-free model for referring image segmentation. ReSTR adopts transformers for both visual and linguistic modalities to capture global context from feature extraction. It also includes the multimodal fusion encoder composed of transformers to encode fine and flexible interactions between these features of the two modalities. Also, the multimodal fusion encoder computes an adaptive classifier for patch-level classification. Furthermore, we have proposed a segmentation decoder to refine the patch-level predictions to the pixel-level prediction in a coarse-to-fine manner. ReSTR outperformed the existing referring image segmentation techniques on all public benchmarks. The fact that computational cost quadratically increases as patch size decreases is the potential limitation of our work. Since the performance of the dense prediction tasks heavily depends on the patch size when using the visual transformer [39], it introduces an undesirable trade-off between performance and computational cost. To alleviate this, integrating linear-complexity transformer architectures would be a promising research direction, which we leave for future work.

Acknowledgement. We thank Manjin Kim and Sehyun Hwang for fruitful discussions. This work was supported by MSRA Collaborative Research Program, and the NRF grant and the IITP grant funded by Ministry of Science and ICT, Korea (NRF-2021R1A2C3012728, IITP-2020-0-00842, No.2019-0-01906 Artificial Intelligence Graduate School Program–POSTECH).

References

[1] Gedas Bertasius, Lorenzo Torresani, Stella X. Yu, and Jianbo Shi. Convolutional random walk networks for semantic image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

[3] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[8] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021.

[9] 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. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[10] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villasenor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer vision and image understanding*, 2010.

[11] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[12] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Proc. European Conference on Computer Vision (ECCV)*, 2016.

[13] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

[16] 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 *Proc. European Conference on Computer Vision (ECCV)*, 2020.

[17] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proc. International Conference on Machine Learning (ICML)*, 2021.

[18] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[19] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021.

[20] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[21] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2011.

[22] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] 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 *Proc. European Conference on Computer Vision (ECCV)*, 2014.

[25] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

[26] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

[27] 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 *Proc. IEEE International Conference on Computer Vision*

(ICCV), 2021.

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *Proc. International Conference on Learning Representations (ICLR)*, 2019.

[30] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[32] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2021.

[33] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

[34] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[35] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *Proc. Neural Information Processing Systems (NeurIPS)*, 2019.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[37] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[38] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[39] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2017.

[41] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[42] 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. *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021.

[43] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[44] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[45] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[46] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[47] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *Proc. International Conference on Learning Representations (ICLR)*, 2016.

[48] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Proc. European Conference on Computer Vision (ECCV)*, 2016.

[49] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2011.

[50] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding. Acfnet: Attentional class feature network for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 6798–6807, 2019.

[51] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[52] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[53] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[54] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip Torr. Conditional random fields as recurrent neural networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

[55] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.