

Contrastive Grouping with Transformer for Referring Image Segmentation

Jiajin Tang¹ Ge Zheng¹ Cheng Shi¹ Sibe Yang^{1,2†}

¹School of Information Science and Technology, ShanghaiTech University

²Shanghai Engineering Research Center of Intelligent Vision and Imaging

{tangjj, zhengge, shicheng2022, yangsb}@shanghaitech.edu.cn

Abstract

Referring image segmentation aims to segment the target referent in an image conditioning on a natural language expression. Existing one-stage methods employ per-pixel classification frameworks, which attempt straightforwardly to align vision and language at the pixel level, thus failing to capture critical object-level information. In this paper, we propose a mask classification framework, Contrastive Grouping with Transformer network (CGFormer), which explicitly captures object-level information via token-based querying and grouping strategy. Specifically, CGFormer first introduces learnable query tokens to represent objects and then alternately queries linguistic features and groups visual features into the query tokens for object-aware cross-modal reasoning. In addition, CGFormer achieves cross-level interaction by jointly updating the query tokens and decoding masks in every two consecutive layers. Finally, CGFormer cooperates contrastive learning to the grouping strategy to identify the token and its mask corresponding to the referent. Experimental results demonstrate that CGFormer outperforms state-of-the-art methods in both segmentation and generalization settings consistently and significantly. Code is available at <https://github.com/Toneyaya/CGFormer>.

1. Introduction

Referring Image Segmentation (RIS) aims to segment the target referent in an image given a natural language expression [12, 17, 64]. It attracts increasing attention in the research community and is expected to show its potential in real applications, such as human-robot interaction via natural language [50] and image editing [3]. Compared to classical image segmentation that classifies pixels or masks into a closed set of fixed categories, RIS requires locating referent at pixel level according to the free-form natural languages with open-world vocabularies. It faces chal-

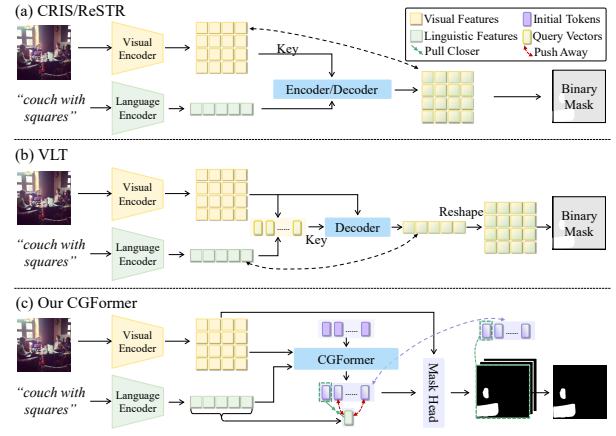


Figure 1. Comparison of Transformer-based RIS methods and our CGFormer. (a) CRIS [51] and ReSTR [25] fuse relevant linguistic features into visual features, while (b) VLT [12] generates query vectors to query visual features for segmentation. In contrast, (c) our CGFormer introduces learnable query tokens and explicitly groups visual features into tokens conditioning on language. Co-operating the grouping strategy with contrastive learning, it identifies the token and its mask corresponding to the referent.

lenges of comprehensively understanding vision and language modalities and aligning them at pixel level.

Existing works mainly follow the segmentation framework of *per-pixel classification* [17, 34] integrated with multi-modal fusion to address the challenges. They introduce various fusion methods [19, 20, 26, 32, 46, 63] to obtain vision-language feature maps and predict the segmentation results based on the feature maps. Recently, the improvement of vision-language fusion for RIS mainly lies in utilizing the Transformer [12, 25, 51, 62]. LAVT [62] integrates fusion module into the Transformer-based visual encoder. CRIS [51] and ReSTR [25] fuse linguistic features into each feature of the visual feature maps, as shown in Figure 1a. In contrast, VLT [12] integrates the relevant visual features into language-conditional query vectors via transformer decoder, as shown in Figure 1b.

Although these methods have improved the segmenta-

[†]Sibe Yang is the corresponding author.

tion accuracy, they still face several intrinsic limitations. First, the works [25, 51, 62] based on pixel-level fusion only model the pixel-level dependencies for each visual feature, which fails to capture the crucial object/region-level information. Therefore, they cannot accurately ground expressions that require efficient cross-modal reasoning on objects. Second, although VLT [12]’s query vectors contain object-level information after querying, it directly weights and reshapes different tokens into one multi-modal feature map for decoding the final segmentation mask. Therefore, it loses the image’s crucial spatial priors (relative spatial arrangement among pixels) in the reshaping process. More importantly, it does not model the inherent differences between query vectors, resulting in that even though different query vectors comprehend expressions in their own way, they still focus on similar regions but fail to focus on different regions and model their relations.

In this paper, we aim to propose a simple and effective framework to address these limitations. Instead of using *per-pixel classification* framework, we adopt an end-to-end *mask classification* framework [1, 16] (see Figure 1c) to explicitly capture object-level information and decode segmentation masks for both the referent and other disturbing objects/stuffs. Therefore, we can simplify RIS task by finding the corresponding mask for the expression. Note that our framework differs from two-stage RIS methods [53, 64] which require explicitly detecting the objects first and then predicting the mask in the detected bounding boxes.

Specifically, we propose a Contrastive Grouping with Transformer (CGFormer) network consisting of the *Group Transformer* and *Consecutive Decoder* modules. The *Group Transformer* aims to capture object-level information and achieve object-aware cross-modal reasoning. The success of applying query tokens in object detection [1, 69] and instance segmentation [8, 16, 54] could be a potential solution. However, it is non-trivial to apply them to RIS. Without the annotation supervision of other mentioned objects other than the referent, it is hard to make tokens pay attention to different objects and distinguish the token corresponding to the referent from other tokens. Therefore, although we also specify query tokens as object-level information representations, we explicitly group the visual feature map’s visual features into query tokens to ensure that different tokens focus on different visual regions without overlaps. Besides, we can further cooperate contrastive learning with the grouping strategy to make the referent token attend to the referent-relevant information while forcing other tokens to focus on different objects and background regions, as shown in Figure 1c. In addition, we alternately query the linguistic features and group the visual features into the query tokens for cross-modal reasoning.

Furthermore, integrating and utilizing multi-level feature maps are crucial for accurate segmentation. Previous

works [32, 40, 63] fuse visual and linguistic features at multiple levels in parallel and later integrate them via ConvLSTM [47] or FPNs [30]. However, their fusion modules are solely responsible for cross-modal alignment at each level, which fails to perform joint reasoning for multiple levels. Therefore, we propose a *Consecutive Decoder* that jointly updates query tokens and decodes masks in every two consecutive layers to achieve cross-level reasoning.

To evaluate the effectiveness of CGFormer, we conduct experiments on three standard benchmarks, *i.e.*, RefCOCO series datasets [39, 41, 65]. In addition, unlike semantic segmentation, RIS is not limited by the close-set classification but to open-vocabulary alignment. It is necessary to evaluate the generalization ability of RIS models. Therefore, we introduce new subsets of training sets on the three datasets to ensure the categories of referents in the test set are not seen in the training stage, inspired by the zero-shot visual grounding [44] and open-set object detection [67].

In summary, our main contributions are as follows,

- We propose a Group Transformer cooperated with contrastive learning to achieve object-aware cross-modal reasoning by explicitly grouping visual features into different regions and modeling their dependencies conditioning on linguistic features.
- We propose a Consecutive Decoder to achieve cross-level reasoning and segmentation by jointly performing the cross-modal inference and mask decoding in every two consecutive layers in the decoder.
- We are the first to introduce an end-to-end mask classification framework, the Contrastive Grouping with Transformer (CGFormer), for referring image segmentation. Experimental results demonstrate that our CGFormer outperforms all state-of-the-art methods on all three benchmarks consistently.
- We introduce new splits on datasets for evaluating generalization for referring image segmentation models. CGFormer shows stronger generalizability compared to state-of-the-art methods thanks to object-aware cross-modal reasoning via contrastive learning.

2. Related Work

Referring Image Segmentation (RIS) aims to segment objects from images according to natural language expressions. The pioneering work [17] uses the concatenation operation to fuse the linguistic and visual features. Some following works [7, 17, 18, 22, 23, 26, 37] extract textual features for the expressions at the sentence level, while other works [2, 14, 32, 40] employ word vectors as textual representations. Considering that natural language naturally contains structured information [45, 59] that can be exploited to align with visual constituents, some methods explicitly decompose expressions into different components [20, 53,

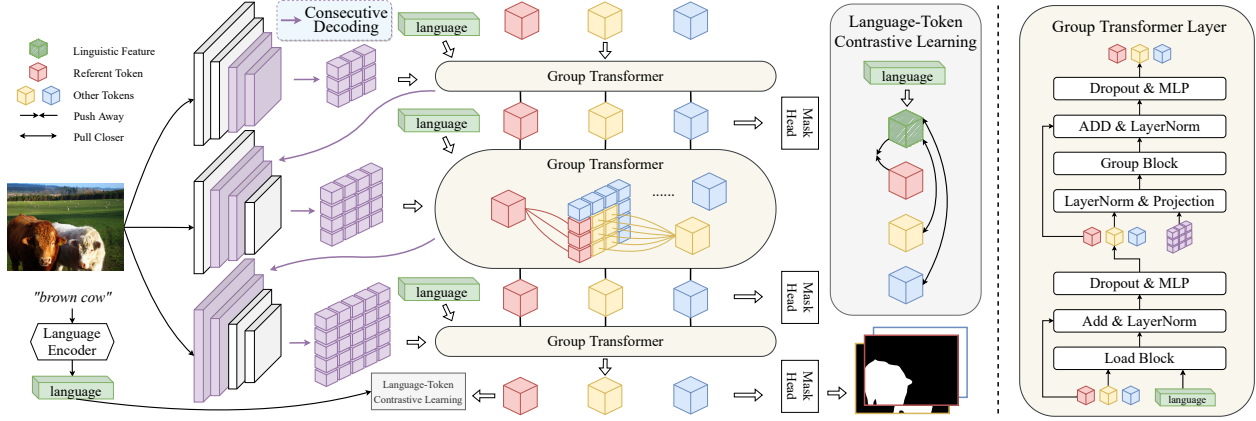


Figure 2. Overall framework of the proposed CGFormer. We first extract visual and linguistic features and then feed them to Group Transformer to integrate multi-modal features to object-level tokens in a Consecutive Decoding way. Next, we distinguish the referent token from others via Contrastive Learning between language-token pairs and decode segmentation results for tokens via mask head.

[58, 61] or apply soft component division via the attention mechanisms [12, 15, 19, 46, 57, 63, 64]. The composed components are then aligned with visual constituents via the well-designed module networks [53, 56, 60, 61, 64] or attention mechanism [29, 46, 63] and interact with each other via graph convolution networks [19, 20] or the transformer [12].

Recently, the research interest has shifted toward developing a better framework for vision-language fusion. LAVT [62] adopts Swin Transformer [33] as the visual encoder and integrates vision-language fusion modules at the last four encoding layers in the visual encoder. Alternatively, ReSTR [25] and CRIS [51], which first encode vision and language with a dual encoder and then fuse visual and linguistic features by resorting to a multi-modal transformer encoder or cross-modal decoder. Unlike existing one-stage RIS studies that are based on per-pixel classification framework, we convert the pixel-level alignment to the mask-level by selecting the mask corresponding to the expression.

Semantic and Instance Segmentation. Semantic segmentation aims to segment regions according to visual semantics by labeling every pixels [34]. Mainstream methods adopt the segmentation framework of per-pixel classification. Specifically, FCNs [34] adopt a stack of convolutional blocks to classify pixels. Further, ASPP [4, 5] and GCN [42] are applied to improve FCNs with larger receptive fields. Transformer-based models [48, 68] further capture long-range dependencies. Unlike semantic segmentation, instance segmentation requires predicting both the masks and categories at the instance level [28]. To achieve this goal, mainstream methods adopt the segmentation framework of mask classification. Specifically, Mask R-CNN [16] employs a two-stage framework, which first generates a set of proposals and then predicts the masks and categories for proposals. Moreover, DETR [1] adopts an end-to-end seg-

mentation framework that uses a large number of learnable query tokens to represent instances and predicts the mask and category of the instance based on each corresponding token. Recently, MaskFormer [9] expands the DETR and can be applied to both semantic and instance segmentation.

To address RIS, we further exploit the advantages of per-pixel and mask classification by using the hard assignment to ensure that each pixel can only be grouped into one query token and avoid the overlap between the tokens’ masks.

Hard Assignment is a reparameterization method to solve the problem of non-differentiable argmax operation. In recent works, Gumbel-Softmax [21, 38], a hard assignment method, has been applied to semantic segmentation to group pixels with similar semantics [55, 66]. For example, GroupViT [55] initializes several sets of queries representing multiple level semantics and uses Gumbel-Softmax to group pixels from low level to high level. Similar to GroupViT, K-means Mask Transformer [66] generates several queries as clustering centers and clusters pixels with similar semantics via Gumbel-Softmax.

However, bottom-up clustering based on semantics is not applicable in RIS because RIS requires distinguishing the target region and disturbing regions with similar semantics. Therefore, instead of utilizing Gumbel-Softmax to group pixels from bottom to up, we use it to compare between our specialized tokens, which can avoid the target region and the disturbing regions being grouped together.

3. Method

The framework of our proposed CGFormer is shown in Figure 2. First, we adopt the visual encoder and language encoder to encode images and referring expressions (see Section 3.1). Second, we achieve object-aware cross-modal reasoning via the proposed Group Transformer (see Section 3.2). Next, we implement the cross-level reasoning and

segmentation via the proposed Consecutive Decoder (see Section 3.3). Finally, we apply contrastive learning to distinguish the referent token from other tokens and obtain the mask corresponding to the referent token as the segmentation result (see Section 3.4).

3.1. Visual Encoder and Language Encoder

Visual Encoder. Following the previous work [62], we employ Swin Transformer [33] as the visual encoder for fair comparison. For an input image $I \in \mathbb{R}^{H \times W \times 3}$ with the size of $H \times W$, we extract its visual feature maps at four stage $i \in \{1, 2, 3, 4\}$, which we denote it as $V = \{V_i\}_{i=1}^4$, $V_i \in \mathbb{R}^{H_i \times W_i \times C_i^v}$. Here, each stage corresponds to an encoding block of Swin Transformer, and H_i , W_i and C_i^v denote the height, width and channel dimension of V_i .

Language Encoder. We adopt BERT [11] as the language encoder following the previous work [62]. Given an expression contains L words, we extract its linguistic feature and denote it as $e \in \mathbb{R}^{C^l}$, where C^l is the channel dimension. In addition, we obtain the word representations by removing the last pooling layer, which is denoted as $F \in \mathbb{R}^{L \times C^l}$.

3.2. Group Transformer

We propose the Group Transformer to achieve object-aware cross-modal inference. Group Transformer uses query tokens to represent object-level information and updates query tokens by alternately querying the linguistic features and grouping visual features. Specifically, we first initialize a set of learnable tokens representing the different objects/regions (see Section 3.2.1). Then we query the linguistic and visual features for tokens via our proposed novel Group Transformer layer (see Section 3.2.2). After alternative reasoning, tokens capture the rich object characteristics relevant to the referring expressions. Note that we use multiple Group Transformer layers in different stages of the decoder, which will be introduced in Section 3.3.1.

3.2.1 Definition of Query Token

Inspired by the apply of query tokens [1, 9, 54] in object detection and semantic segmentation, we randomly initialize N learnable tokens $T \in \mathbb{R}^{N \times C^t}$ where C^t is the channel dimension of tokens, to represent referent and other disturbing objects/stuffs.

Next, we feed these tokens T to Group Transformer layers to capture the object-level information conditioning on the expression and update the features for tokens. For simplicity of demonstration, we use $T_{i-1} \in \mathbb{R}^{N \times C^t}$ to represent the features of tokens output from the $(i-1)$ -th layer and input them to the i -th layer of the Group Transformer.

3.2.2 Group Transformer Layer

The right part of Figure 2 illustrates the architecture of a single Group Transformer layer. The Load block and Group

block are two core blocks to achieve object-aware cross-modal reasoning. In addition, we follow standard transformer [49] to employ LayerNorm for feature normalization and MLP with activation function for nonlinear mapping. Specifically, the Load block preloads the linguistic information each token should focus on at the current layer. The Group block, the critical component of Group Transformer, performs the cross-modal interaction and groups the visual features into query tokens to ensure that different tokens focus on different visual regions without overlap.

Load Block is expected to preload what linguistic information the query tokens should focus on at the current layer. The load layer is implemented by a classical cross-attention block [49], accepting the input tokens $T_{i-1} \in \mathbb{R}^{N \times C^t}$ as the query and the word vectors $F \in \mathbb{R}^{L \times C^l}$ extracted by the text encoder (see Section 3.1) as the key and value. Concretely, the computation of the Load block is as follows:

$$\begin{aligned} T_i^q &= T_{i-1} W_q, F^k = F W_k, F^v = F W_v, \\ T_i^l &= (\text{softmax}(\frac{T_i^q (F^k)^\top}{\sqrt{C^l}}) F^v) W_c, \end{aligned} \quad (1)$$

where $W_q, W_c \in \mathbb{R}^{C^t \times C^t}$ and $W_k, W_v \in \mathbb{R}^{C^l \times C^l}$ are learnable projection matrices. And T_i^q, F^k and F^v are query, key, and value in the cross attention, respectively. We end up with linguistic-enhanced representations for tokens, $T_i^l \in \mathbb{R}^{N \times C^t}$, and feed them into the Group block to query and group the relevant visual features for tokens.

Group Block interacts between vision and language and groups visual features from the feature map into linguistic-enhanced query tokens T_i^l . We denote the feature map as $D_i \in \mathbb{R}^{H_i \times W_i \times C_i^v}$, which is fused from the feature maps in two consecutive layers in Consecutive Decoder (refer to Section 3.3 for details). Firstly, we project T_i^l and D_i into a common feature space:

$$T_i' = T_i^l W_t, D_i' = \text{flatten}(D_i) W_d, \quad (2)$$

where $W_t \in \mathbb{R}^{C^t \times C^t}$ and $W_d \in \mathbb{R}^{C_i^v \times C^t}$ are learnable projection matrices, and flatten operation flattens the feature map D_i into the visual feature with $H_i W_i$ vectors. Then, we calculate the similarities $S_{pixel} \in \mathbb{R}^{N \times H_i W_i}$ between every pairwise features of tokens T_i' and features D_i' :

$$S_{pixel} = \text{norm}(T_i') \text{norm}(D_i')^\top, \quad (3)$$

where norm means L2 normalization for vectors.

Next, based on the similarities S_{pixel} , we group the features in D_i' and correspond the groups to tokens T_i' . However, the grouping operation with straightforward hard assignment is non-differentiable. Therefore, we adopt a learnable Gumbel-softmax [21, 38] to hard assign the features in D_i' to the tokens T_i' and generate the mask $S_{mask} \in$

$\mathbb{R}^{N \times H_i W_i}$ of the grouping. The computation is as follows:

$$\begin{aligned} S_{gumbel} &= \text{softmax}((S_{pixel} + G)/\tau), \\ S_{onehot} &= \text{onehot}(\text{argmax}_N(S_{gumbel})), \\ S_{mask} &= (S_{onehot})^\top - \text{sg}(S_{gumbel}) + S_{gumbel}, \end{aligned} \quad (4)$$

where $G \in \mathbb{R}^{N \times H_i W_i}$ samples from the Gumbel(0, 1) distribution, τ is the learnable significance coefficient to assist in finding a more suitable assign boundary, sg is the stop gradient operator. Here, argmax_N means selecting the corresponding token of T'_i with the highest similarity for each feature in D'_i , and the onehot operation transforms the token indexes into $H_i W_i$ one-hot vectors $S_{onehot} \in \mathbb{R}^{H_i W_i \times N}$. The mask $S_{mask} \in \mathbb{R}^{N \times H_i W_i}$ indicates the grouping from the features D'_i to the tokens T'_i .

Finally, we integrate the features D'_i to update tokens T'_i based on the mask S_{mask} , which is computed as follows:

$$T_i = \text{MLP}(S_{mask} D'_i) + T'_i, \quad (5)$$

where MLP is the multilayer perceptron. And T_i are the updated features of tokens via the Group block, which capture the rich object/region characteristics relevant to the linguistic features.

3.3. Consecutive Decoder

We further perform cross-level reasoning via the proposed Consecutive Decoder. Figure 3 shows the architecture of our Consecutive Decoder and its comparison to the parallel cross-modal fusion. Previous works [2, 18–20, 32, 37, 40, 63] model the vision-language interaction at multiple levels in parallel and late integrate multi-level results. The sole interaction at a single level fails to perform joint interaction across various levels. In contrast, the Consecutive Decoder achieves cross-level reasoning by jointly updating the query tokens in every two consecutive decoder layers, and the two-level cross-modal information will be consecutively propagated in multiple levels from bottom to up.

Specifically, the Consecutive Decoder contains three stages. At each decoding stage, it first fuses feature maps at two levels as the input of the Group Transformer layer to update tokens (see Section 3.3.1) and then decodes the corresponding mask of each token through the Mask Head (see Section 3.3.2).

3.3.1 Consecutive Decoding

We intersperse multi-scale and cross-modal reasoning at each decoding layer. Specifically, for the i -th decoder layer, we first adopt a convolutional module to fuse the visual feature map V_i at the current layer i and the multi-modal feature map D_{i-1} output from the previous layer $i - 1$ of the Consecutive Decoder to generate the feature map $D_i \in \mathbb{R}^{H_i \times W_i \times C_i^v}$. Then, we update the query tokens

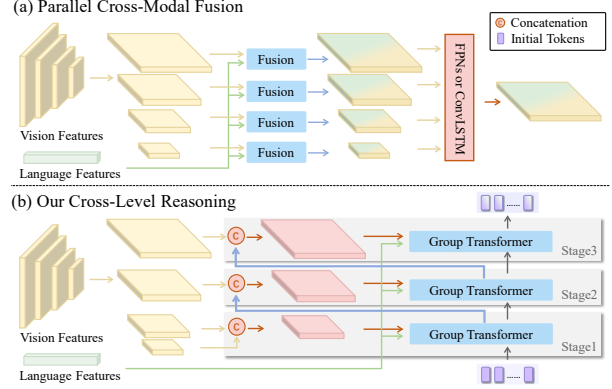


Figure 3. Comparison of (a) parallel cross-modal fusion and (b) our cross-level reasoning via the Consecutive Decoder.

T_{i-1} output from the previous Consecutive Decoder layer by querying them on the feature map D_i via the Group Transformer layer. The calculation is as follows:

$$\begin{aligned} D_i &= \text{Conv}([V_i; \text{Up}(D_{i-1})]), i \in \{2, 3, 4\} \\ T_i &= \text{GroupTransformerLayer}(T_{i-1}, D_i), \end{aligned} \quad (6)$$

where Conv is the convolution layer, Up refers to up-sampling D_{i-1} to the scale of V_i , $[\cdot]$ denotes concatenation along the channel dimension.

Particularly, for the first decoder layer ($i = 1$), we skip the fusion and cross-modal interaction and let $D_1 = V_1$ and $T_1 = T$, where T are the initialized tokens defined in Section 3.2.1.

3.3.2 Mask Head

For i -th decoder layer, our mask head takes updated tokens T_i and visual feature map D_i as inputs and output the segmentation probabilities $Z_i \in \mathbb{R}^{N \times H_i \times W_i}$ for tokens via dynamic convolutions [6]. For n -th token with feature $T_i^{(n)}$, we first project it to convolution kernels $W_i^{(n)}$ and then predict the segmentation probabilities $Z_i^{(n)} \in \mathbb{R}^{H_i \times W_i}$ based on the kernels, which is computed as follows,

$$\begin{aligned} W_i^{(n)} &= \text{MLP}(T_i^{(n)}), \\ Z_i^{(n)} &= \text{Sigmoid}(\text{Conv}_{W_i^{(n)}}(D_i)), \end{aligned} \quad (7)$$

where the superscript (n) denotes the features, kernels, and predicted probabilities corresponding to the n -th token, and the $\text{Conv}_{W_i^{(n)}}$ means the convolution layer with the convolution kernels $W_i^{(n)}$.

3.4. Contrastive Learning

We use contrastive learning to distinguish the referent token from other tokens by maximizing the similarity between the referent token and the expression and minimizing the similarities between negative pairs. For simplicity

	Method	RefCOCO			RefCOCO+			G-Ref			ReferIt
		val	test A	test B	val	test A	test B	val-U	test-U	val-G	test
mIoU	DMN [40]	49.78	54.83	45.13	38.88	44.22	32.29	-	-	36.76	52.81
	MCN [37]	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	-	-
	CGAN [36]	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	46.54	-
	LTS [23]	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25	-	-
	VLT [12]	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	49.76	-
	CRIS [51]	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36	-	-
	Our CGFormer	76.93	78.70	73.32	68.56	73.76	61.72	67.57	67.83	65.79	66.42
oIoU	RRN [26]	55.33	57.26	53.93	39.75	42.15	36.11	-	-	36.45	63.63
	MAttNet [64]	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61	-	-
	CMSA [63]	58.32	60.61	55.09	43.76	47.60	37.89	-	-	39.98	63.80
	CMPC [19]	61.36	64.53	59.64	49.56	53.44	43.23	-	-	49.05	65.53
	LSCM [20]	61.47	64.99	59.55	49.34	53.12	43.50	-	-	48.05	66.57
	CEFNet [14]	62.76	65.69	59.67	51.50	55.24	43.01	51.93	-	-	66.70
	BUSNet [61]	63.27	66.41	61.39	51.76	56.87	44.13	-	-	50.56	-
	ReSTR [25]	67.22	69.30	64.45	55.78	60.44	48.27	54.48	-	-	-
	LAVT [62]	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	60.50	-
	Our CGFormer	74.75	77.30	70.64	64.54	71.00	57.14	64.68	65.09	62.51	73.36

Table 1. Comparison with state-of-the-art models in referring image segmentation on RefCOCO, RefCOCO+, G-Ref, and ReferIt datasets.

of demonstration, we suppose the first token represents the referent and other tokens represent non-target objects/stuff. The contrastive loss between the tokens with features $T_4 \in \mathbb{R}^{N \times C^t}$ output from last decoder layer and the expression $e \in \mathbb{R}^{C^l}$ is computed as follows,

$$\mathcal{L}_{cl} = -\log\left(\frac{\exp(s(T_4^{(1)}, e))}{\sum_{n=2}^N \exp(s(T_4^{(n)}, e))}\right), \quad (8)$$

where $s(\cdot, \cdot)$ is used to compute the similarity.

In addition, we combine the dice loss [27] and binary cross-entropy loss as the segmentation loss, \mathcal{L}_{seg} . And we use the segmentation loss on multiple levels to supervise the learning of masks Z_i . The total loss \mathcal{L} is the sum of contrastive loss \mathcal{L}_{cl} and segmentation loss \mathcal{L}_{seg} :

$$\mathcal{L} = \mathcal{L}_{cl} + \mathcal{L}_{seg}. \quad (9)$$

During the inference, we predict the mask based on the referent token’s segmentation probabilities of the last decoder layer, *i.e.*, $Z_4^{(1)}$.

4. Experiments

4.1. Datasets and Implementation Details

Datasets. We conduct experiments on four common benchmark datasets, RefCOCO [65], RefCOCO+ [65], G-Ref [39, 41], and ReferIt [24]. The images of the first three datasets are all based on MSCOCO [31], but are annotated with different settings. RefCOCO has a short average description length of 3.5 words, RefCOCO+ is limited to not describing absolute locations of referents, and G-Ref has a longer word count per expression (8.4 words). We follow previous works [32, 61] to split the RefCOCO and RefCOCO+ into training, validation, testA and testB. For the

G-Ref, we apply both partitions of UMD and Google for the evaluation. In addition, the ReferIt dataset, which is also the main benchmark for referring image segmentation, includes 19,894 images sourced from the IAPR TC-12 [13].

Implementation Details. Following [62], our visual encoder is pre-trained on ImageNet22K [10], text encoder is initialized with the weights from HuggingFace [52], and image size is 480×480 . The hyperparameters, C_i^v , C^l and C^t are $1024/2^{i-1}$, 768 and 512, respectively. We adopt AdamW [35] as the optimizer with initialized learning rate $1e-4$ and train the model for 50 epochs with batch size 64. All experiments are conducted on NVIDIA Tesla A40 GPUs. Following [62], we adopt overall IoU (oIoU), mean IoU (mIoU), and precision at the 0.5, 0.7, and 0.9 thresholds of IoU as our main evaluation metrics.

Implementation for Generalization. We introduce new splits on RefCOCO series datasets to validate the generalization, inspired by [44]. Specifically, we split them according to the splits of seen and unseen classes on MSCOCO of open-vocabulary detection [67]. Image-text pairs in the original training sets whose referent categories belong to the seen classes are selected as the new training sets. Likewise, the image-text pairs of test sets are also split into seen and unseen subsets according to whether the categories of the referents belong to seen classes. And the categories of referents in the unseen splits are not seen in the training stage. We consider that CRIS [51] employs CLIP [43] as the encoder network for transferring the knowledge of CLIP to achieve text-to-pixel alignment. Therefore, for the generalization experiments, we also take the text encoder of CLIP as the language encoder for our CGFormer and LAVT [62] for a fair comparison. For all methods, we train 50 epochs using the official code and select best-performing models

Dataset	Method	val		test	
		seen	unseen	seen	unseen
RefCOCO	CRIS [51]	68.66	52.77	52.77	52.66
	LAVT [62]	73.05	61.35	72.31	57.66
	Ours	75.52	63.17	74.63	59.03
RefCOCO+	CRIS [51]	61.49	48.08	60.46	45.26
	LAVT [62]	61.17	41.49	60.97	38.67
	Ours	67.44	51.24	66.35	48.11
G-Ref(U)	CRIS [51]	58.64	42.63	59.68	38.88
	LAVT [62]	60.16	42.33	60.37	41.38
	Ours	65.60	46.11	65.67	42.31
G-Ref(G)	CRIS [51]	42.36	32.84	\	
	LAVT [62]	57.33	40.43		
	Ours	62.85	45.05		

Table 2. Comparison for generalization setting on the validation and test sets of RefCOCO, RefCOCO+ and G-Ref datasets using mean IoU(%). (U): UMD partition. (G): Google partition.

on the validation set for comparison. We use mIoU as the evaluation metric to eliminate the influence of categories because referents with different categories differ in size.

4.2. Comparison with State-of-the-Art Methods

As shown in Table 1 and Table 2, we compare CGFormer with state-of-the-art methods [12, 25, 51, 62] on the four benchmarks and validate its generalization ability on our split datasets. CGFormer outperforms state-of-the-art methods on all the splits on the three datasets consistently.

Comparison on Referring Image Segmentation. Table 1 illustrates the comparison on common splits. Our CGFormer improves the average oIoU by 1.78%, 2.35%, 2.82% and 6.66% on RefCOCO, RefCOCO+, G-Ref, and ReferIt datasets respectively, compared to the previous best-performing methods [14, 62]. This demonstrates that our object-aware reasoning and joint decoding not only achieve a better understanding of the location and appearance information in RefCOCO but also adapt to the various forms of expressions in RefCOCO+ and G-Ref.

Besides, the following three comparisons show the effectiveness of our CGFormer from different perspectives: (1) CRIS [51] is the recently proposed CLIP-based pixel-level contrastive learning method. Compared to CRIS, CGFormer achieves clear performance improvements of 6.40%, 6.67% and 7.59% on the three RefCOCO series datasets, which indicates that our mask-level contrastive framework is more capable than the pixel-level alignment. (2) Compared to other methods that capture object-level information, such as MAttNet [64] and BUSNet [61], CGFormer significantly surpasses them by 13.96% and 15.58% in terms of average oIoU on RefCOCO and RefCOCO+ datasets, respectively. These results imply that our end-to-end token-based object information capturing is more sim-

	Method	P@0.5	P@0.7	P@0.9	oIoU
1	baseline	75.31	61.48	16.85	65.70
2	1+one token	77.28	64.94	19.47	66.39
3	1+N tokens	77.70	65.12	19.44	66.46
4	3+grouping	83.94	72.09	23.43	70.81
5	4+hard assignment	84.59	74.92	33.75	72.44
6	5+multi-scale	85.80	76.31	35.35	73.28
7	5+CD (ours full)	87.23	78.69	38.77	74.75
8	VLT(Swin-B+BERT)*	83.24	72.81	24.64	70.89
9	w/o cos	85.64	76.23	33.96	73.37
10	w/o learnable τ	86.14	76.99	36.48	73.50

Table 3. Ablation study on the validation set of RefCOCO. CD: Consecutive Decoder. cos: cosine similarity operation. τ : learnable parameter in Gumble Softmax. Results with * refer to [62].

ple and effective. (3) Moreover, we improve the performance of VLT [12] by 10.76%, 13.33%, and 12.88% on the three RefCOCO series datasets, respectively. VLT also adopts query tokens to model the object-level information, however, it does not consider the inherent differences between tokens. The large gains show the superiority of our grouping strategy cooperated with contrastive learning to distinguish tokens.

Comparison on Generalization. We compare CGFormer with LAVT [62] and CRIS [51] as LAVT is the current best-performing model in referring image segmentation and CRIS transfers the knowledge of the strong generalizable CLIP model [43]. As is shown in Table 2, CGFormer outperforms LAVT and CRIS for both seen and unseen splits on all three datasets consistently. On the RefCOCO+ dataset that relies on understanding object-level attributes rather than location information, our performance exceeds LAVT by 5.8% and 9.6% in terms of average mIoU on seen and unseen splits, respectively. In addition, our CGFormer performs significantly better than other methods on the G-ref dataset with more complex languages. The mIoU of our CGFormer outperforms CIRS by 20.49% and 12.21% on seen and unseen splits of G-Ref(G), respectively.

4.3. Ablation Study

The results of the ablation study are shown in Table 3. **Baseline and Grouping Strategy.** (1) The baseline extracts the visual feature map V_4 of the visual encoder and predicts the segmentation result from the map via dynamic convolutions with the learned kernels from linguistic feature e . (2) We improve the baseline by generating a linguistic-conditioned query token and querying relevant visual features on V_4 to capture object-level information, which slightly improves the baseline by 0.69%. (3) We further extend one token to multiple tokens, but the two models have similar performance. The results suggest that simply adding tokens cannot boost performance, as these tokens

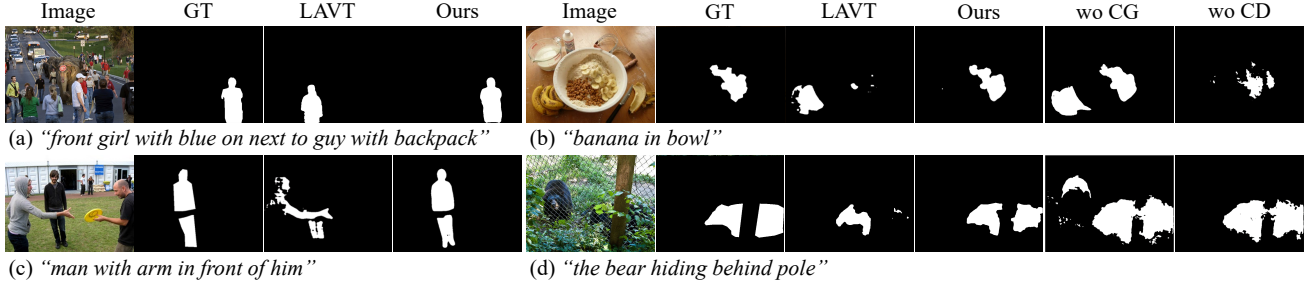


Figure 4. Visualization results of our CGFormer, its variants, and LAVT [62]. CG: Contrastive Grouping. CD: Consecutive Decoder.

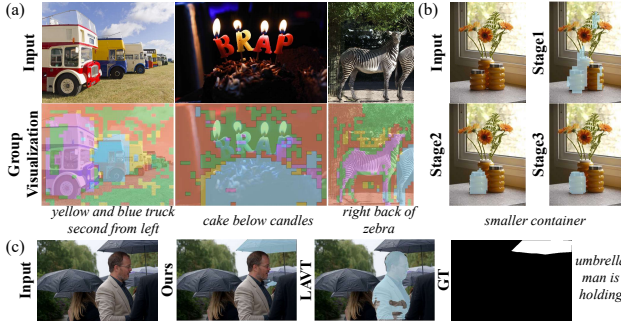


Figure 5. Visualization of grouping results for (a) different tokens (in different colors), (b) the referent token in three stages and (c) segmentation results of unseen objects.

are likely to focus on similar information rather than distinct regions. (4) Our grouping strategy cooperated with contrastive loss to make tokens can focus on different regions and let them distinguishable, which delivers a 4.35% improvement. (5) We further use the hard assignment with learnable Gumbel Softmax to obtain a more refined grouping that achieves an improvement of 1.63%.

Multi-Scale Decoding. We extend the single-scale model (row 5) to multi-scale one (row 6) by first parallel updating tokens at multiple levels and then integrating these tokens to predict the segmentation mask over the multi-scale visual feature map fused by FPNs [30]. The 0.84% improvement of oIoU shows the effectiveness of multi-scale features.

We further connect grouping layers by consecutive decoding, which is applied in our final model (row 7). The 1.47% improved oIoU of the model using a consecutive decoder (row 7) over the model using parallel querying (row 6) suggests that our joint querying and decoding in the decoder is a more desirable solution than aggregating different levels of information in parallel.

Others. (1) We further validate the necessity of the proposed contrastive grouping by comparing our CGFormer (row 7) with VLT [12] (row 8) using the same visual backbone and text encoder. We significantly outperforms VLT by 3.99%, 5.88%, 14.13% and 3.86% in terms of P@0.5, P@0.7, P@0.9 and oIoU, respectively. (2) We replace the cosine similarity with dot produce (row 9) or fix the param-

eter τ in the Gumble softmax to 0.1 (row 10), which results in a reduction of about 1.3% in oIoU.

4.4. Visualization

Figure 4 visualizes segmentation results. The expression in (a) refers to a girl in a complex scenario with a crowd of several dozen people. For (b), CGFormer accurately recognizes the challenging visual concept “*sliced banana*” and distinguishes it from a similar object with the same visual concept. The (c) demonstrates that our joint grouping and decoding comprehensively understand the expression “*hands*” and “*man*”. The (d) illustrates that CGFormer entirely segments the “*bear*” even though it is shaded by “*pole*” thanks to our object-aware reasoning.

Variants Results of our models without the Contrastive Grouping and Consecutive Decoder are shown in the two rightmost columns of Figure 4, respectively. Contrastive Grouping captures object-level information to distinguish between similar objects, and the Consecutive Decoder help obtain more precise segmentation results.

Qualitative results of grouping are shown in Figure 5 (a) and (b), which demonstrates: (1) Both the referent token and others represent certain meaningful objects/regions. (2) Tokens can partition the objects of the same categories (e.g., the different trucks) and different categories (e.g., the cake and candles). (3) The grouping can be more precise in multiple stages (see results for referent token in b).

5. Conclusion

This paper proposes a novel Contrastive Grouping with Transformer network (CGFormer) for referring image segmentation, which achieves object-aware cross-modal and cross-level reasoning. The experimental results demonstrate the superiority and the generalization ability of the proposed CGFormer compared to state-of-the-art methods.

Acknowledgment: this work was supported by the National Natural Science Foundation of China (No.62206174), the Shanghai Pujiang Program (No.21PJ1410900), and the Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).

References

- [1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 3, 4
- [2] 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. 2, 5
- [3] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8721–8729, 2018. 1
- [4] 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*, 40(4):834–848, 2017. 3
- [5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3
- [6] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11039, 2020. 5
- [7] 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. 2
- [8] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 2
- [9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 3, 4
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6
- [11] 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. 4
- [12] 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. 1, 2, 3, 6, 7, 8
- [13] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villaseñor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer vision and image understanding*, 114(4):419–428, 2010. 6
- [14] 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. 2, 6, 7
- [15] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 3
- [16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 3
- [17] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016. 1, 2
- [18] 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. 2, 5
- [19] Shaoifei 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. 1, 3, 5, 6
- [20] Tianrui Hui, Si Liu, Shaoifei 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. 1, 2, 3, 5, 6
- [21] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 3, 4
- [22] Yang Jiao, Zequn Jie, Weixin Luo, Jingjing Chen, Yu-Gang Jiang, Xiaolin Wei, and Lin Ma. Two-stage visual cues enhancement network for referring image segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1331–1340, 2021. 2
- [23] 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. 2, 6
- [24] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 6
- [25] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring im-

- age segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18145–18154, 2022. 1, 2, 3, 6, 7
- [26] 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. 1, 2, 6
- [27] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019. 6
- [28] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2359–2367, 2017. 3
- [29] Liang Lin, Pengxiang Yan, Xiaoqian Xu, Sibe Yang, Kun Zeng, and Guanbin Li. Structured attention network for referring image segmentation. *IEEE Transactions on Multimedia*, 24:1922–1932, 2021. 3
- [30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2, 8
- [31] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 6
- [32] 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. 1, 2, 5, 6
- [33] 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. 3, 4
- [34] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 3
- [35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [36] 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. 6
- [37] 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. 2, 5, 6
- [38] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 3, 4
- [39] 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. 2, 6
- [40] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *European Conference on Computer Vision*, pages 630–645, 2018. 2, 5, 6
- [41] 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. 2, 6
- [42] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters—improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017. 3
- [43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6, 7
- [44] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4694–4703, 2019. 2, 6
- [45] Cheng Shi and Sibe Yang. Spatial and visual perspective-taking via view rotation and relation reasoning for embodied reference understanding. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 201–218. Springer, 2022. 2
- [46] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *European Conference on Computer Vision*, pages 38–54, 2018. 1, 3
- [47] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 2
- [48] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 3
- [49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4
- [50] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation.

- In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6629–6638, 2019. 1
- [51] 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. 1, 2, 3, 6, 7
- [52] 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. 6
- [53] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020. 2, 3
- [54] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *European Conference on Computer Vision*, pages 553–569. Springer, 2022. 2, 4
- [55] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 3
- [56] Sibe Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4145–4154, 2019. 3
- [57] Sibe Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4644–4653, 2019. 3
- [58] Sibe Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9952–9961, 2020. 2
- [59] Sibe Yang, Guanbin Li, and Yizhou Yu. Propagating over phrase relations for one-stage visual grounding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 589–605. Springer, 2020. 2
- [60] Sibe Yang, Guanbin Li, and Yizhou Yu. Relationship-embedded representation learning for grounding referring expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2765–2779, 2020. 3
- [61] Sibe 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. 2, 3, 6, 7
- [62] 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. 1, 2, 3, 4, 6, 7, 8
- [63] 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. 1, 2, 3, 5, 6
- [64] 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. 1, 2, 3, 6, 7
- [65] 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. 2, 6
- [66] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hatwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. *arXiv preprint arXiv:2207.04044*, 2022. 3
- [67] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2, 6
- [68] 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. 3
- [69] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2