CRIS: CLIP-Driven Referring Image Segmentation

Zhaoqing Wang[1,2*]   Yu Lu[3*]   Qiang Li[4*]   Xunqiang Tao[2]   Yandong Guo[2]
Mingming Gong[5]   Tongliang Liu[1]

[1]University of Sydney; [2]OPPO Research Institute; [3]Beijing University of Posts and Telecommunications
[4]Kuaishou Technology; [5]University of Melbourne

{derrickwang005,taoxunqiang}@gmail.com; aniki@bupt.edu.cn liqiang03@kuaishou.com
yandong.guo@live.com; mingming.gong@unimelb.edu.au; tongliang.liu@sydney.edu.au

Abstract

Referring image segmentation aims to segment a referent via a natural linguistic expression. Due to the distinct data properties between text and image, it is challenging for a network to well align text and pixel-level features. Existing approaches use pretrained models to facilitate learning, yet separately transfer the language/vision knowledge from pretrained models, ignoring the multi-modal corresponding information. Inspired by the recent advance in Contrastive Language-Image Pretraining (CLIP), in this paper, we propose an end-to-end CLIP-Driven Referring Image Segmentation framework (CRIS). To transfer the multi-modal knowledge effectively, CRIS resorts to vision-language decoding and contrastive learning for achieving the text-to-pixel alignment. More specifically, we design a vision-language decoder to propagate fine-grained semantic information from textual representations to each pixel-level activation, which promotes consistency between the two modalities. In addition, we present text-to-pixel contrastive learning to explicitly enforce the text feature similar to the related pixel-level features and dissimilar to the irrelevances. The experimental results on three benchmark datasets demonstrate that our proposed framework significantly outperforms the state-of-the-art performance without any post-processing.

1. Introduction

Referring image segmentation [14, 49, 50] is a fundamental and challenging task at the intersection of vision and language understanding, which can be potentially used in a wide range of applications, including interactive image editing and human-object interaction. Unlike semantic and instance segmentation [8, 10, 12, 47], which requires segmenting the visual entities belonging to a pre-determined set of categories, referring image segmentation is not limited to
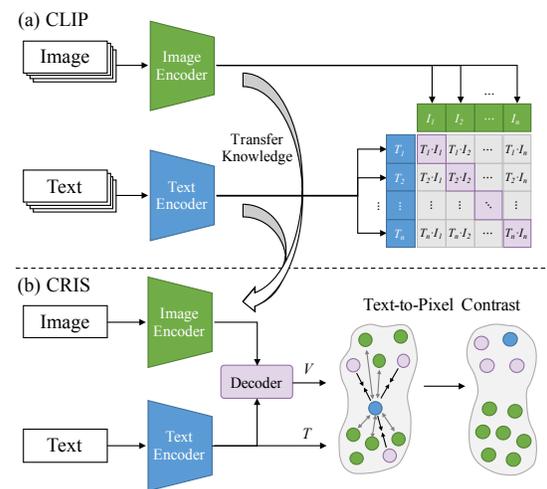

Figure 1. **An illustration of our main idea.** (a) CLIP [39] jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of image I and text T, which can capture the multi-modal corresponding information. (b) To transfer this knowledge of the CLIP model from image level to pixel level, we propose a CLIP-Driven Referring Image Segmentation (CRIS) framework. Firstly, we design a vision-language decoder to propagate fine-grained semantic information from textual features to pixel-level visual features. Secondly, we combine all pixel-level visual features V with the global textual feature T and adopt contrastive learning to pull text and related pixel-wise features closer and push other irrelevances away.

indicating specific categories but finding a particular region according to the input language expression.

Since the image and language modality maintain different properties, it is difficult to explicitly align textual features with pixel-level activations. Benefiting from the powerful capacity of the deep neural network, early approaches [14, 22, 25, 33] concatenate textual features with each visual activation directly, and use these combined features to generate the segmentation mask. Subsequently, to

---

*Equal contribution

Language: *"a blond haired , blue eyed young boy in a blue jacket"*   Language: *"a zebra ahead of the other zebra"*

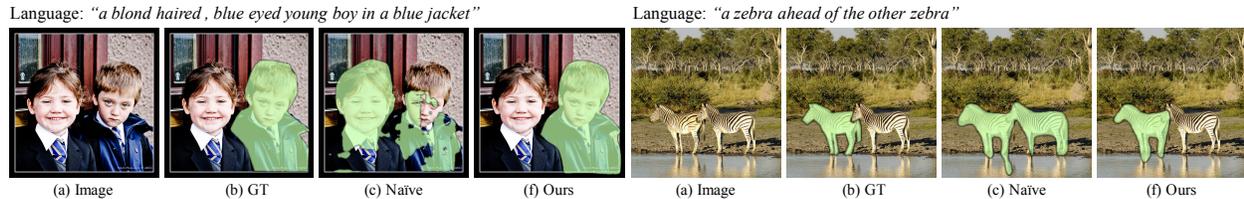| (a) Image | (b) GT | (c) Naïve | (f) Ours | (a) Image | (b) GT | (c) Naïve | (f) Ours |

Figure 2. **Comparison between the direct fine-tuning and our proposed methods.** "Naive" denotes the direct fine-tuning mentioned in section 4. Compared with the direct fine-tuning, our method can not only leverage the powerful cross-modal matching capability of the CLIP, but also learn fine-grained visual representations.

address the lack of adequate interaction between two modalities, a series of methods [4,16,17,42,49] adopt the language-vision attention mechanism to better learn cross-modal features.

Existing methods [4, 16, 17, 42, 49] leverage external knowledge to facilitate learning in common, while they mainly utilize a single-modal pretraining (*e.g.*, the pretrained image or text encoder), which is short of multi-modal correspondence information. By resorting to language supervision from large-scale unlabeled data, vision-language pretraining [34, 39, 46] is able to learn ample multi-modal representations. Recently, the remarkable success of the CLIP [39] has shown its capability of learning SOTA image-level visual concepts from 400 million image-text pairs, which assists many multi-modal tasks achieve significant improvements, including image-text retrieval [39], video-text retrieval [6, 31]. However, as shown in Figure 2, the direct usage of the CLIP can be sub-optimal for pixel-level prediction tasks, *e.g.*, referring image segmentation, duo to the discrepancy between image-level and pixel-level prediction. The former focuses on the global information of an input image, while the latter needs to learn fine-grained visual representations for each spatial activation.

In this paper, we explore leveraging the powerful knowledge of the CLIP model for referring image segmentation, in order to enhance the ability of cross-modal matching. Considering the characteristics of referring image segmentation, we propose an effective and flexible framework named CLIP-Driven Referring Image Segmentation (CRIS), which can transfer ample multi-modal corresponding knowledge of the CLIP for achieving text-to-pixel alignment. Firstly, we propose a visual-language decoder that captures long-range dependencies of pixel-level features through the self-attention operation and adaptively propagate fine-structured textual features into pixel-level features through the cross-attention operation. Secondly, we introduce the text-to-pixel contrastive learning, which can align linguistic features and the corresponding pixel-level features, meanwhile distinguishing irrelevant pixel-level features in the multi-modal embedding space. Based on this scheme, the model can explicitly learn fine-grained visual concepts by interwinding the linguistic and pixel-level visual features.

Our main contributions are summarized as follow:

- We propose a CLIP-Driven Referring Image Segmentation framework (CRIS) to transfer the knowledge of the CLIP model for achieving text-to-pixel alignment.

- We take fully advantage of this multi-modal knowledge with two innovative designs, *i.e.*, the vision-language decoder and text-to-pixel contrastive learning.

- The experimental results on three challenging benchmarks significantly outperform previous state-of-the-art methods by large margins (*e.g.*, $+4.89\,IoU$ on Ref-COCO, $+8.88\,IoU$ on RefCOCO+, $+5.47\,IoU$ on G-Ref).

2. Related Work

Vision-Language Pretraining. Vision-Language pretraining has made rapid progress in recent years and achieved impressive performance on various multi-modal downstream tasks. By resorting to semantic supervision from large-scale image data, several approaches [34, 39, 46] were proposed to learn visual representations from text representations. MIL-NCE [34] mainly explored leveraging noisy large-scale Howto100M [35] instructional videos to learn a better video encoder via an end-to-end manner. SimVLM [46] reduced the training complexity by leveraging large-scale weak supervision, and adopted a single prefix language modeling objective in an end-to-end manner. Benefit from the large-scale image and text pairs collected from the Internet, a recent approach, *i.e.*, Contrastive Language-Image Pretraining (CLIP) [39], achieved the notable success of aligning two modalities representations in the embedding space. CLIP adopted contrastive learning with high-capacity language models and visual feature encoders to capture compelling visual concepts for zero-shot image classification. More recently, a series of works [6, 31, 38, 43] were proposed to transfer the knowledge of CLIP models to downstream tasks and achieved promising results, such as video caption, video-text retrieval, and image generation. Different from these works, we transfer these image-level visual concepts to referring image segmentation for leveraging multi-modal corresponding information.

Contrastive Learning Date back to [9], these approaches learned representations by contrasting positive pairs against negative pairs. Several approaches [2, 3, 11, 23, 48] were proposed to treat each image as a class and use contrastive loss-based instance discrimination for representation learning. Recently, VADeR and DenseCL [37, 45] proposed to explore pixel-level contrastive learning to fill the gap between self-supervised representation learning and dense prediction tasks. Besides, CLIP [39] proposed a promising alternative that directly learns transferable visual concepts from large-scale collected image-text pairs by using cross-modal contrastive loss. In this paper, we propose a CLIP-Driven Referring Image Segmentation (CRIS) framework to transfer the knowledge of the CLIP model to referring image segmentation in an end-to-end manner.

Referring Image Segmentation Referring image segmentation is to segment a target region (*e.g.*, object or stuff) in an image by understanding a given natural linguistic expression, which was first introduced by [14]. Early works [22, 25, 33] first extracted visual and linguistic features by CNN and LSTM, respectively, and directly concatenated two modalities to obtain final segmentation results by a FCN [28]. In [50], they proposed a two-stage method that first extracted instances using Mask R-CNN [12], and then adopted linguistic features to choose the target from those instances. Besides, MCN [30] designed a framework achieving impressive results. They learned to optimize two related tasks, *i.e.*, referring expression comprehension and segmentation, simultaneously.

As the attention mechanism arouses more and more interests, a series of works are proposed to adopt the attention mechanism. It is powerful to extract the visual contents corresponding to the language expression. [42] used the vision-guided linguistic attention to aggregate the linguistic context of each visual region adaptively. [49] designed a Cross-Modal Self-Attention (CSMA) module to focus on informative words in the sentence and crucial regions in the image. [15] proposed a bi-directional relationship inferring network that adopted a language-guided visual and vision-guided linguistic attention module to capture the mutual guidance between two modalities. Besides, LTS [18] designs a strong pipeline that decouples the task into a "Locate-Then-Segment" scheme by introducing the position prior. EFNet [7] designs a co-attention mechanism to use language to refine the multi-modal features progressively, which can promote the consistent of the cross-modal information representation. More recently, VLT [5] employs transformer to build a network with an encoder-decoder attention mechanism for enhancing the global context information. Different from previous methods, we aim to leverage the knowledge of the CLIP, in order to improving the compatibility of multi-modal information and boost the ability of cross-modal matching.

3. Methodology

As illustrated in Figure 3, we introduce how the proposed CRIS framework transfers the knowledge of CLIP to referring image segmentation to achieve text-to-pixel alignment by leveraging multi-modal corresponding information. Firstly, we use a ResNet [13] and a Transformer [44] to extract image and text features respectively, which are further fused to obtain the simple multi-modal features. Secondly, these features and text features are fed into the vision-language decoder to propagate fine-grained semantic information from textual representations to pixel-level visual activations. Finally, we use two projectors to produce the final prediction mask, and adopt the text-to-pixel contrastive loss to explicitly align the text features with the relevant pixel-level visual features.

3.1. Image & Text Feature Extraction

As illustrated in Figure 3, the input of our framework consists of an image I and a referring expression T.

Image Encoder. For an input image $I \in \mathbb{R}^{H \times W \times 3}$, we utilize multiple visual features from the 2th-4th stages of the ResNet, which are defined as $F_{v2} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_2}$, $F_{v3} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_3}$, and $F_{v4} \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_4}$, respectively. Note that C is the feature dimension, H and W are the height and width of the original image.

Text Encoder. For an input expression $T \in \mathbb{R}^L$, we adopt a Transformer [44] modified by [40] to extract text features $F_t \in \mathbb{R}^{L \times C}$. The Transformer operates on a lower-cased byte pair encoding (BPE) representation of the text with a 49,152 vocab size [41], and the text sequence is bracketed with [SOS] and [EOS] tokens. The activations of the highest layer of the transformer at the [EOS] token are further transformed as the global textual representation $F_s \in \mathbb{R}^{C'}$. Note that C and C' are the feature dimension, L is the length of the referring expression.

Cross-modal Neck. Given multiple visual features and the global textual representation F_s, we obtain the simple multi-modal feature $F_{m4} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$ by fusing F_{v4} with F_s:

$$F_{m4} = Up(\sigma(F_{v4}W_{v4}) \cdot \sigma(F_s W_s)), \qquad (1)$$

where $Up(\cdot)$ denotes $2\times$ upsampling, $\cdot$ denotes the element-wise multiplication, σ denotes ReLU, W_{v4} and W_s are two learnable matrices to transform the visual and textual representations into the same feature dimension. Then, the multi-modal features F_{m2} and F_{m3} are obtained by:

$$\begin{aligned} F_{m3} &= [\sigma(F_{m4}W_{m4}), \ \sigma(F_{v3}W_{v3})], \\ F_{m2} &= [\sigma(F_{m3}W_{m3}), \ \sigma(F'_{v2}W_{v2})], \ F'_{v2} = Avg(F_{v2}), \end{aligned}$$
$$(2)$$

where $Avg(\cdot)$ denotes a kernel size of 2×2 average pooling with 2 strides, respectively. $[,]$ is the concatenation operation. Subsequently, we concatenate three multi-modal features and
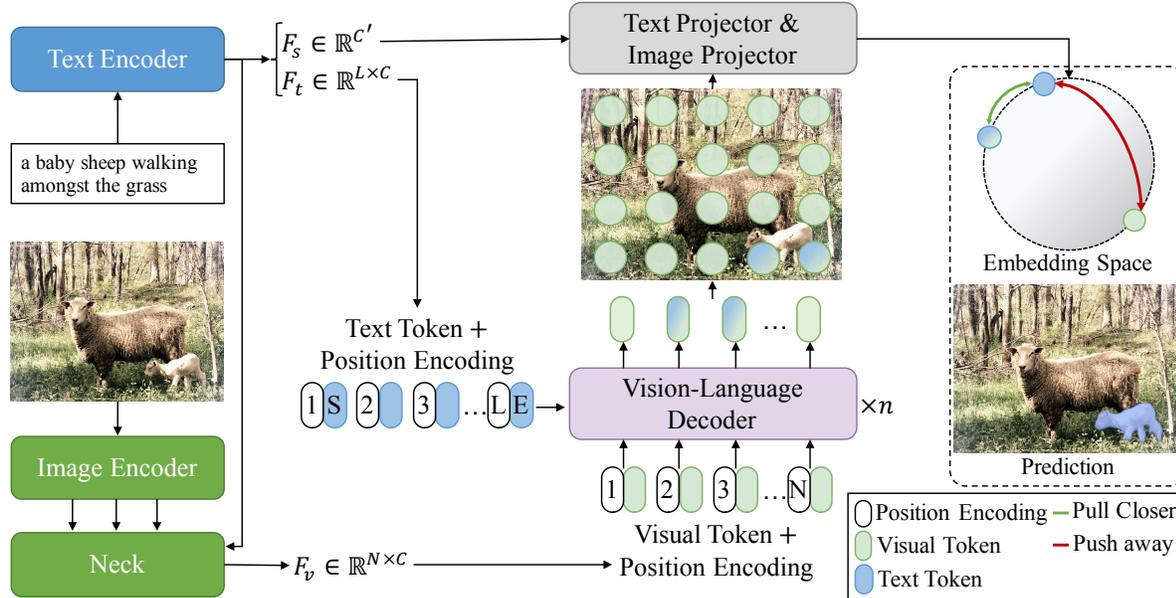
Figure 3. **The overview of the proposed CLIP-Driven Referring Image Segmentation (CRIS) framework.** CRIS mainly consists of a text encoder, an image encoder, a cross-modal neck, a vision-language decoder, and two projectors. The vision-language decoder is used to adaptively propagate semantic information from textual features to visual features. The text-to-pixel contrastive learning is used to explicitly learn fine-grained multi-modal corresponding information by interwinding the text features and pixel-level visual features.

use a 1×1 convolution layer to aggregate them:

$$F_m = Conv([F_{m_2}, F_{m_3}, F_{m_4}]), \qquad (3)$$

where $F_m \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$. Finally, we concatenate a 2D spatial coordinate feature $F_{coord} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 2}$ with F_m and fuse that by a 3×3 convolution [27]. The visual feature $F_v \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$ is calculated as follow,

$$F_v = Conv([F_m, F_{coord}]). \qquad (4)$$

As shown in figure 3, we flatten the spatial domain of F_v into a sequence, forming the visual feature $F_v \in \mathbb{R}^{N \times C}$, $N = \frac{H}{16} \times \frac{W}{16}$, which is utilized in the following process.

3.2. Vision-Language Decoder

We design a vision-language decoder to adaptively propagate fine-grained semantic information from textual features to visual features. As shown in Figure 3, the decoder module takes textual features $F_t \in \mathbb{R}^{L \times C}$ and pixel-level visual features $F_v \in \mathbb{R}^{N \times C}$ as inputs, which can provide ample textual information corresponding to visual features. To capture positional information, the fixed sine spatial positional encodings are added to F_v [1] and F_t [44], respectively. The vision-language decoder composed of n layers is applied to generate a sequence of evolved multi-modal features $F_c \in \mathbb{R}^{N \times C}$. Following the standard architecture of the transformer [44], each layer consists of a multi-head self-attention layer, a multi-head cross-attention layer, and a feed-forward network. In one decoder layer, F_v is first

sent into the multi-head self-attention layer to capture global contextual information:

$$F_v' = MHSA(LN(F_v)) + F_v, \qquad (5)$$

where F_v' is the evolved visual feature, $MHSA(\cdot)$ and $LN(\cdot)$ denote the multi-head self-attention layer and Layer Normalization [21], respectively. The multi-head self-attention mechanism is composed of three point-wise linear layers mapping F_v to intermediate representations, queries $Q \in \mathbb{R}^{N \times d_q}$, keys $K \in \mathbb{R}^{N \times d_k}$ and values $V \in \mathbb{R}^{N \times d_v}$. Multi-head self-attention is then calculated as follows,

$$MHSA(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \qquad (6)$$

After that, the multi-head cross-attention layer is adopted to propagate fine-grained semantic information into the evolved visual features, where one point-wise linear layer maps F_v' to Q, and the other two linear layers map F_t to K and V. To obtain the multi-modal feature F_c, the output query Q is further computed by a MLP block of two layers with Layer Normalization and residual connections:

$$\begin{aligned} F_c' &= MHCA(LN(F_v'), F_t) + F_v', \\ F_c &= MLP(LN(F_c')) + F_c', \end{aligned} \qquad (7)$$

where $MHCA(\cdot)$ denotes the multi-head cross-attention layer, and F_c' is the intermediate features. The evolved multi-modal feature F_c is utilized for the final segmentation mask. Note that the hyper-parameter n is discussed in the following experiment section.

3.3. Text-to-Pixel Contrastive Learning

Although the CLIP [39] learns powerful image-level visual concepts by aligning the textual representation with the image-level representation, this type of knowledge is sub-optimal for referring image segmentation, due to the lack of more fine-grained visual concepts.

To tackle this issue, we design a text-to-pixel contrastive loss, which explicitly aligns the textual features with the corresponding pixel-level visual features. As illustrated in Figure 3, image and text projector are adopted to transform F_c and F_s as follow:

$$
\begin{aligned}
z_v &= F_c' W_v + b_v, \ F_c' = Up(F_c), \\
z_t &= F_s W_t + b_t,
\end{aligned}
\tag{8}
$$

where $z_t \in \mathbb{R}^D$, $z_v \in \mathbb{R}^{N \times D}$, $N = \frac{H}{4} \times \frac{W}{4}$, Up denotes $4\times$ upsampling, W_v and W_t are two learnable matrices to transform F_c and F_s into the same feature dimension D, b_v and b_t are two learnable biases.

Given a transformed textual feature z_t and a set of transformed pixel-level features z_v, a contrastive loss is adopted to optimize the relationship between two modalities, where z_t is encouraged to be similar with its corresponding z_v and dissimilar with other irrelevant z_v. With the similarity measured by dot product, the text-to-pixel contrastive loss can be formulated as:

$$
L_{con}^i(z_t, z_v^i) = \begin{cases} -\log \sigma(z_t \cdot z_v^i), & i \in \mathcal{P}, \\ -\log(1 - \sigma(z_t \cdot z_v^i)), & i \in \mathcal{N}, \end{cases}
\tag{9}
$$

$$
L_{con}(z_t, z_v) = \frac{1}{|\mathcal{P} \cup \mathcal{N}|} \sum_{i \in \mathcal{P} \cup \mathcal{N}} L_{con}^i(z_t, z_v^i),
\tag{10}
$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the class of "1" and "0" in the ground truth, $|\mathcal{P} \cup \mathcal{N}|$ is the cardinality, σ is the sigmoid function. Finally, to obtain the final segmentation results, we reshape $\sigma(z_t \cdot z_v)$ into $\frac{H}{4} \times \frac{W}{4}$ and upsample it back to the original image size.

4. Experimental results

Our proposed framework is built on different image encoders (*e.g.*, ResNet-50, ResNet-101 [13]) and compared with a series of state-of-the-art methods. To evaluate the effectiveness of each component in our method, we conduct extensive experiments on three benchmarks, including RefCOCO [19], RefCOCO+ [19], and G-Ref [32].

4.1. Datasets

RefCOCO [19] is one of the largest and most commonly used datasets for referring image segmentation. It contains 19,994 images with 142,210 referring expressions for 50,000 objects, which are collected from the MSCOCO [24] via a two-player game [19]. The dataset is split into 120,624 train,

10,834 validation, 5,657 test A, and 5,095 test B samples, respectively. According to statistics, each image contains two or more objects and each expression has an average length of 3.6 words.

RefCOCO+ [19] dataset contains 141,564 language expressions with 49,856 objects in 19,992 images. The dataset is split into train, validation, test A, and test B with 120,624, 10,758, 5,726, and 4,889 samples, respectively. Compared with RefCOCO dataset, some kinds of absolute-location words are excluded from the RefCOCO+ dataset, which could be more challenging than the RefCOCO dataset.

G-Ref [36] includes 104,560 referring expressions for 54,822 objects in 26,711 images. Unlike the above two datasets, natural expressions in G-Ref are collected from Amazon Mechanical Turk instead of a two-player game. The average length of sentences is 8.4 words, which have more words about locations and appearances. It is worth mentioning that we adopt UNC partition in this paper.

4.2. Implementation Details

Experimental Settings. We initiate the text and image encoder with CLIP [39], and adopt ResNet-50 [13] as the image encoder for all ablation studies. Input images are resized to 416×416. Due to the extra `[SOS]` and `[EOS]` tokens, and the input sentences are set with a maximum sentence length of 17 for RefCOCO and RefCOCO+, and 22 for G-Ref. Each Transformer Decoder layer has 8 heads, and the feed-forward hidden dimension is set to 2048. We train the network for 50 epochs using the Adam optimizer with the learning rate $\lambda = 0.0001$. The learning rate is decreased by a factor of 0.1 at the 35th epoch. We train the model with a batch size of 64 on 8 Tesla V100 with 16 GPU VRAM.

During inference, we upsample the predicted results back to the original image size and binarize them at a threshold of 0.35 as the final result. No other post-processing operations are needed.

Metrics. Following previous works [5, 22, 25, 33], we adopt two metrics to verify the effectiveness: IoU and Precision@X. The IoU calculates intersection regions over union regions of the predicted segmentation mask and the ground truth. The Precision@X measures the percentage of test images with an IoU score higher than the threshold $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, which focuses on the location ability of the method.

4.3. Ablation Study

The proposed CRIS framework consists of two main parts, *i.e.*, text-to-pixel contrastive learning and vision-language decoder. To investigate each component in our method, we conduct extensive experiments on the validation set of three widely used datasets.

Effectiveness of Contrastive Learning & Vision-Language Decoder. Firstly, we remove the parts of the

Table 1. **Ablation studies on validation set of three benchmarks.** $Con.$ denotes the proposed text-to-pixel contrastive learning. $Dec.$ denotes the proposed vision-language decoder. n denotes the number of layers in the vision-language decoder. We set $Num = 3$ as the default. "Params" and "FPS" denote the parameter complexity (M) and inference speed, respectively. Given an image $I \in \mathbb{R}^{416 \times 416 \times 3}$, they are calculated on a Tesla V100 GPU. Gray lines denote the baseline network.

Dataset	$Con.$	$Dec.$	n	IoU	Pr@50	Pr@60	Pr@70	Pr@80	Pr@90	Params	FPS
RefCOCO	-	-	-	62.66	72.55	67.29	59.53	43.52	12.72	131.86	27.30
	✓	-	-	64.64	74.89	69.58	61.70	45.50	13.31	134.22	25.79
	-	✓	1	66.31	77.66	72.99	65.67	48.43	14.81	136.07	23.02
	✓	✓	1	68.66	80.16	75.72	68.82	51.98	15.94	138.43	22.64
	✓	✓	2	69.13	80.96	76.60	69.67	52.23	16.09	142.64	20.68
	✓	✓	3	**69.52**	**81.35**	**77.54**	**70.79**	**52.65**	16.21	146.85	19.22
	✓	✓	4	69.18	80.99	76.74	69.32	52.57	**16.37**	151.06	18.26
RefCOCO+	-	-	-	50.17	54.55	47.69	40.19	28.75	8.21	131.86	27.30
	✓	-	-	53.15	58.28	53.74	46.67	34.01	9.30	134.22	25.79
	-	✓	1	54.73	63.31	58.89	52.46	38.53	11.70	136.07	23.02
	✓	✓	1	59.97	69.19	64.85	58.17	43.47	13.39	138.43	22.64
	✓	✓	2	60.75	70.69	66.83	60.74	45.69	13.42	142.64	20.68
	✓	✓	3	**61.39**	**71.46**	**67.82**	**61.80**	**47.00**	**15.02**	146.85	19.22
	✓	✓	4	61.15	71.05	66.94	61.25	46.98	14.97	151.06	18.26
G-Ref	-	-	-	49.24	53.33	45.49	36.58	23.90	6.92	131.86	25.72
	✓	-	-	52.67	59.27	52.45	44.12	29.53	8.80	134.22	25.33
	-	✓	1	51.46	58.68	53.33	45.61	31.78	10.23	136.07	22.57
	✓	✓	1	57.82	66.28	60.99	53.21	38.58	13.38	138.43	22.34
	✓	✓	2	58.40	67.30	61.72	54.70	39.67	13.40	142.64	20.61
	✓	✓	3	**59.35**	**68.93**	**63.66**	**55.45**	**40.67**	**14.40**	146.85	19.14
	✓	✓	4	58.79	67.91	63.11	55.43	39.81	13.48	151.06	17.84

text-to-pixel contrastive learning and vision-language decoder from the framework to build our baseline, which is same as the naive setting in Figure 2.(c). As illustrated in Table 1, we introduce the contrastive learning scheme, which significantly increases the IoU accuracy of 1.98%, 2.98%, and 3.43% than the baseline network on three datasets, respectively. This superior performance gain proves that the contrastive loss can encourage the model to explicitly pull closer linguistic and relevant pixel-level visual representations and push away other irrelevances for learning fine-structured multi-modal corresponding information.

Besides, we evaluate the performance of the proposed vision-language decoder. Compared with the baseline network, we use one layer in the decoder, bringing 3.65%, 4.56%, and 2.22% IoU improvements on RefCOCO, RefCOCO+, and G-Ref, respectively. In particular, the self-attention operation can help the model sufficiently capture long-range dependencies across each pixel, which is helpful for understanding complex scenarios. Furthermore, each word encoded by the text encoder is used in the cross-attention operation, which can propagate fine-grained semantic information from textual features to pixel-level features to generate more discriminative visual representations and

obtain more accurate segmentation masks.

Finally, combining the proposed contrastive loss and vision-language decoder, the IoU and Precision are significantly better than the baseline solely with the contrastive loss or decoder module, which further achieves large margins at about 4% - 8% on three datasets. The reason of this obvious complementary phenomenon is that the contrastive loss can guide the decoder to find the more informative emphasis and transfer this knowledge to more accurate pixel-level visual representations, which boosts the ability of cross-modal matching and generates precise segmentation masks.

Numbers of Layers in Decoder. In Table 1, the results illustrate the effect of utilizing different number of layers in the vision-language decoder. When the visual representations are sequentially processed by more layers, our model can consistently get better IoU of 69.52%, 61.39%, and 59.35% on three benchmarks, respectively. The setting of $n = 1$ may not taking full advantage of the multi-modal corresponding information from both vision and language. Meanwhile, the setting of $n = 4$ introduces more parameters, which could increase the risk of over-fitting. Considering the performance and efficiency, we set $n = 3$ as the default in our framework.

Table 2. **Comparisons with the state-of-the-art approaches on three benchmarks.** We report the results of our method with various visual backbones. "⋆" denotes the post-processing of DenseCRF [20]. "-" represents that the result is not provided. IoU is utilized as the metric.

Method	Backbone	RefCOCO			RefCOCO+			G-Ref	
		val	test A	test B	val	test A	test B	val	test
RMI⋆ [25]	ResNet-101	45.18	45.69	45.57	29.86	30.48	29.50	-	-
DMN [33]	ResNet-101	49.78	54.83	45.13	38.88	44.22	32.29	-	-
RRN⋆ [22]	ResNet-101	55.33	57.26	53.95	39.75	42.15	36.11	-	-
MAttNet [50]	ResNet-101	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
NMTree [26]	ResNet-101	56.59	63.02	52.06	47.40	53.01	41.56	46.59	47.88
CMSA⋆ [49]	ResNet-101	58.32	60.61	55.09	43.76	47.60	37.89	-	-
Lang2Seg [4]	ResNet-101	58.90	61.77	53.81	-	-	-	46.37	46.95
BCAN⋆ [15]	ResNet-101	61.35	63.37	59.57	48.57	52.87	42.13	-	-
CMPC⋆ [16]	ResNet-101	61.36	64.53	59.64	49.56	53.44	43.23	-	-
LSCM⋆ [17]	ResNet-101	61.47	64.99	59.55	49.34	53.12	43.50	-	-
MCN [30]	DarkNet-53	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
CGAN [29]	DarkNet-53	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69
EFNet [7]	ResNet-101	62.76	65.69	59.67	51.50	55.24	43.01	-	-
LTS [18]	DarkNet-53	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
VLT [5]	DarkNet-53	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
CRIS (Ours)	ResNet-50	69.52	72.72	64.70	61.39	67.10	52.48	59.35	59.39
CRIS (Ours)	ResNet-101	**70.47**	**73.18**	**66.10**	**62.27**	**68.08**	**53.68**	**59.87**	**60.36**

4.4. Main Results

We compare our proposed approach, CLIP-Driven Referring Image Segmentation, with a series of state-of-the-art methods on three commonly used datasets. As illustrated in Table 2, our method surpasses other methods on each split of all datasets even though we utilize a shallow ResNet-50 [13].

On the RefCOCO dataset, our model significantly outperforms the state-of-the-art Vision Language Transformer [5] by 4.82%, 4.89% and 3.37% on three splits, respectively, which indicates that our model effectively transfer the knowledge of the CLIP model from image-level to pixel-level, enhancing the ability of cross-modal matching.

Besides, in Table 2, our method achieves remarkable performance gains of about 4~8% than a series of state-of-the-art works on the more challenging RefCOCO+ dataset. These obvious improvements over them suggest that our method can adequately leverage the powerful knowledge of the CLIP to accurately focus the region referred by the given language expression.

Furthermore, on another more complex G-Ref dataset where the average length of referring expressions is complicated, our proposed method consistently achieve notable improvement of around 5% IoU than the state-of-the-art Locate then Segmentation [18]. As shown in Table 2, the results demonstrate that our proposed approach manages to understand long and complex sentences that contain more information and more emphases, and simultaneously perceive the corresponding object. Apart from that, longer referring

expressions could contain complex scenarios, which need a strong ability to model the global contextual information. Our proposed vision-language decoder is suitable to enhance the holistic understanding of vision and language features.

4.5. Qualitative Analysis

Visualization. As illustrated in Figure 4, we present some visualization results with different setting, which demonstrates the benefits of each component in our proposed method. Firstly, compared with our full model, the baseline network without the contrastive learning and vision-language decoder generates worse segmentation masks, because the baseline network fails to interwind referring expressions with the corresponding regions. Secondly, the setting of (d) and (e) can obtain more accurate segmentation results than the baseline network, but the model is still confused in some hard regions. Finally, our proposed full model can generate high-quality segmentation masks, which demonstrates the effectiveness of our proposed method, *i.e.*, CRIS.

Failure Cases. We visualize some insightful failed cases in Figure 5. One type of failure is caused by the ambiguity of the input expression. For the top left example in Figure 5, the expression of "yellow" is not enough to describe the region of the man in the yellow snowsuit. Besides, for the top right example, some failures are also caused by the wrong label. It is obvious that the top region is unrelated to "fingers". As shown in the bottom left example, the boundaries of the referent cannot be accurately segmented, but this issue can be alleviated by introducing other technologies, such as the
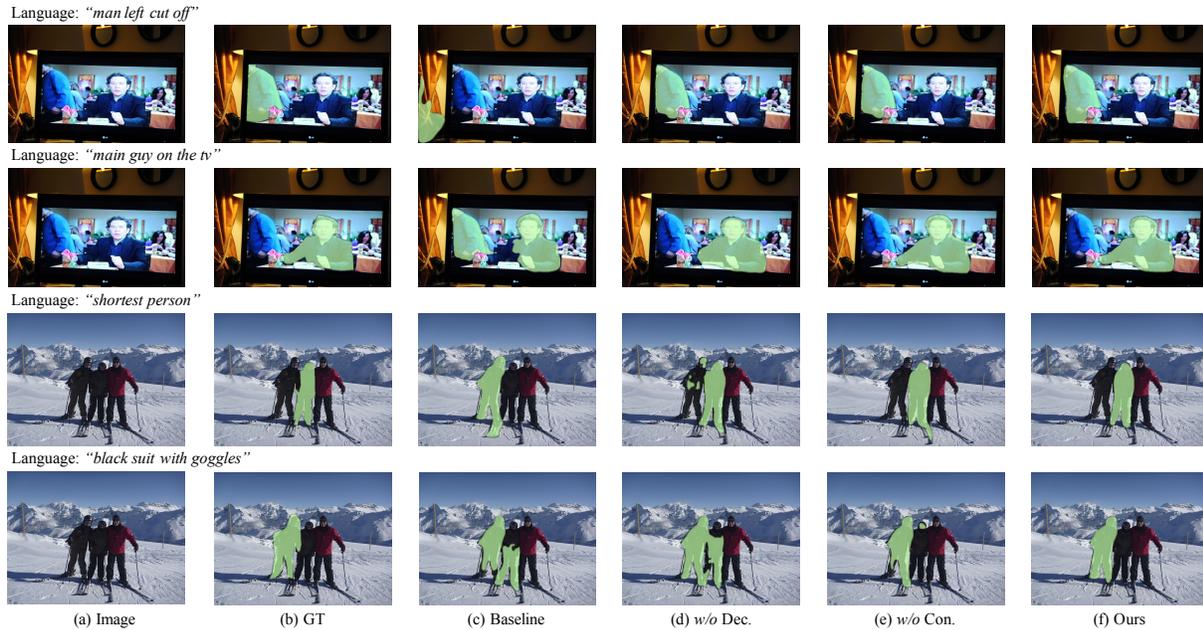
Figure 4. **Qualitative examples with different settings.** (a) the input image. (b) the ground turth. (c) the baseline network. (d) CRIS without Vision-Language Decoder. (e) CRIS without Contrastive Learning. (f) our proposed CRIS. *Best viewed in color.*
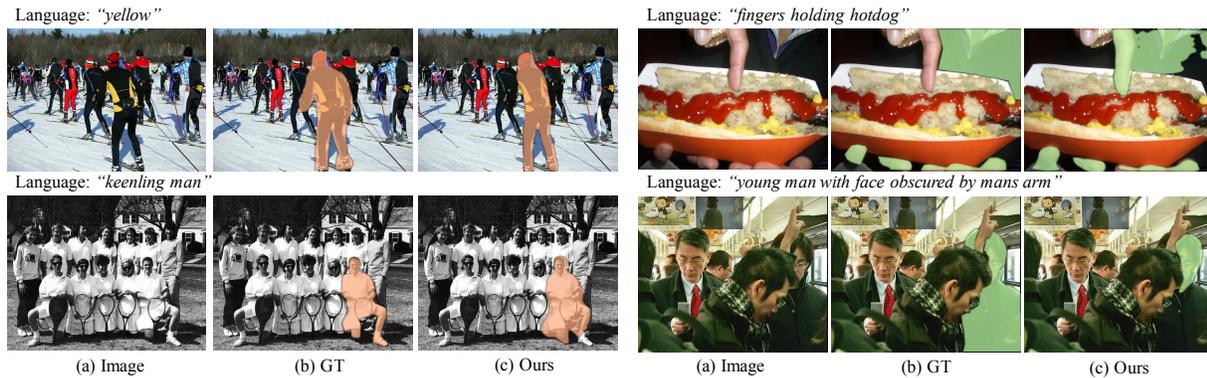

Figure 5. **Qualitative examples of failure cases.** *Best viewed in color.*

refine module. Finally, occlusion could cause failure cases, which is a challenging problem in many vision tasks.

5. Conclusion

In this paper, we have investigated to leverage the power of Contrastive Language-Image Pretraining (CLIP) models to achieve text-to-pixel alignment for referring image segmentation. And, we have proposed an end-to-end CLIP-Driven Referring Image Segmentation (CRIS) framework to well transfer the knowledge of the CLIP model. Compared with the direct fine-tuning, our proposed framework not only inherit the strong cross-modal matching ability of the CLIP, but also learn ample fine-structured visual representa-tions. The designed vision-language decoder can adaptively propagate sufficient semantic information of the language expression into pixel-level visual features, promoting consis-tency between two modalities. Furthermore, the introduced text-to-pixel contrastive learning can explicitly interwind the text representation and relevant pixel-level visual features, learning fine-grained multi-modal corresponding informa-tion. Extensive ablation studies on three commonly used datasets have verified the effectiveness of each proposed component, and our approach significantly outperforms pre-vious state-of-the-art methods without any post-processing.

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. 4

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[4] Yi-Wen Chen, Yi-Hsuan Tsai, Tiantian Wang, Yen-Yu Lin, and Ming-Hsuan Yang. Referring expression object segmentation with caption-aware consistency. *arXiv preprint arXiv:1910.04748*, 2019. 2, 7

[5] 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. 3, 5, 7

[6] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2

[7] 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. 3, 7

[8] 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. 1

[9] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 3

[10] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7519–7528, 2019. 1

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[12] 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. 1, 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 5, 7

[14] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016. 1, 3

[15] 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. 3, 7

[16] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10488–10497, 2020. 2, 7

[17] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *European Conference on Computer Vision*, pages 59–75. Springer, 2020. 2, 7

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9858–9867, 2021. 3, 7

[19] 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. 5

[20] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011. 7

[21] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv e-prints*, pages arXiv–1607, 2016. 4

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018. 1, 3, 5, 7

[23] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. *arXiv preprint arXiv:2203.04181*, 2022. 3

[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 *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[25] 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, 3, 5, 7

[26] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4673–4682, 2019. 7

[27] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31, 2018. 4

[28] 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. 3

[29] 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. 7

[30] 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. 3, 7

[31] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 2

[32] 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. 5

[33] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–645, 2018. 1, 3, 5, 7

[34] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 2

[35] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 2

[36] 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. 5

[37] Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems*, 33:4489–4500, 2020. 3

[38] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 5

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[41] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 3

[42] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–54, 2018. 2, 3

[43] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4858–4862, 2021. 2

[44] 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. 3, 4

[45] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 3

[46] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[47] Tianyi Wu, Yu Lu, Yu Zhu, Chuang Zhang, Ming Wu, Zhanyu Ma, and Guodong Guo. Ginet: Graph interaction network for scene parsing. In *European Conference on Computer Vision*, pages 34–51. Springer, 2020. 1

[48] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 3

[49] 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, 7

[50] 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, 3, 7