Towards Real-Time Panoptic Narrative Grounding by an End-to-End Grounding Network

Haowei Wang[1*], Jiayi Ji[1*], Yiyi Zhou[1, 2], Yongjian Wu[4], Xiaoshuai Sun[1, 2, 3†]

[1]Media Analytics and Computing Lab, Department of Artificial Intelligence,
School of Informatics, Xiamen University, 361005, China.
[2]Institute of Artificial Intelligence, Xiamen University, China.
[3]Fujian Engineering Research Center of Trusted Artificial Intelligence Analysis and Application, Xiamen University, China.
[4]Tencent Youtu Lab, Shanghai, China
wanghaowei@stu.xmu.edu.cn, jjyxmu@gmail.com, zhouyiyi@xmu.edu.cn, littlekenwu@tencent.com, xssun@xmu.edu.cn

Abstract

Panoptic Narrative Grounding (PNG) is an emerging cross-modal grounding task, which locates the target regions of an image corresponding to the text description. Existing approaches for PNG are mainly based on a two-stage paradigm, which is computationally expensive. In this paper, we propose a one-stage network for real-time PNG, termed End-to-End Panoptic Narrative Grounding network (EPNG), which directly generates masks for referents. Specifically, we propose two innovative designs, i.e., Locality-Perceptive Attention (LPA) and a bidirectional Semantic Alignment Loss (SAL), to properly handle the many-to-many relationship between textual expressions and visual objects. LPA embeds the local spatial priors into attention modeling, i.e., a pixel may belong to multiple masks at different scales, thereby improving segmentation. To help understand the complex semantic relationships, SAL proposes a bidirectional contrastive objective to regularize the semantic consistency inter modalities. Extensive experiments on the PNG benchmark dataset demonstrate the effectiveness and efficiency of our method. Compared to the single-stage baseline, our method achieves a significant improvement of up to 9.4% accuracy. More importantly, our EPNG is 10 times faster than the two-stage model. Meanwhile, the generalization ability of EPNG is also validated by zero-shot experiments on other grounding tasks. The source codes and trained models for all our experiments are publicly available at https://github.com/Mr-Neko/EPNG.git.

Introduction

Panoptic Narrative Grounding (González et al. 2021) is a new challenging task that locates the target instances of an image corresponding to the text description via binary pixel masks. Its main challenges not only lie in the joint understanding of multi-modal information but also in many-to-many language-vision alignment, i.e., grounding all related instances or amorphous regions mentioned in the text description. This property also makes it different from a similar grounding task called Referring Expression Segmentation (RES) (Hu, Rohrbach, and Darrell 2016; Yu et al. 2018;
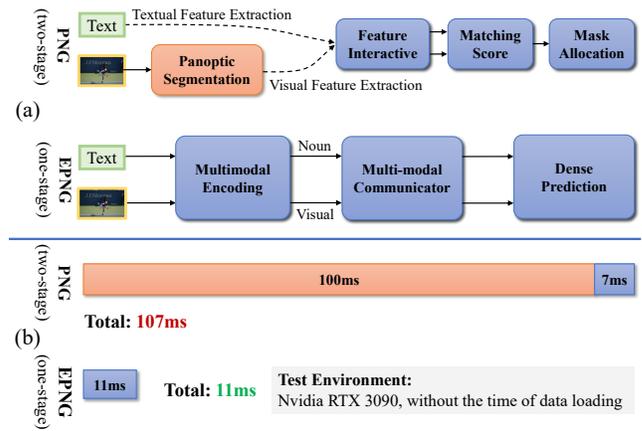
Figure 1: Comparison of pipeline and inference speed between the proposed EPNG and two-stage PNG. (a) EPNG jointly processes visual and text information to generate referred masks in a one-stage fashion, while PNG relies on mask proposals. (b) Our single-stage EPNG is 10x faster than the two-stage approach, enabling real-time deployment.

Ye et al. 2019; Shi et al. 2018), which segments only one instance per expression.

Gonzalez *et al.* (González et al. 2021) first explore this task and propose a preliminary solution in a two-stage fashion, as illustrated in Fig. 1 (a). First, the pre-trained panoptic segmentation models like PFPN (Kirillov et al. 2019a) are used to provide a set of candidate masks of the given image. Secondly, these masks are further transformed into convolution features and then ranked by cross-modal matching. Overall, with the help of panoramic segmentation models, this two-stage solution defines PNG as a mask-text matching problem, greatly reducing the difficulty of prediction.

However, this solution still suffers from two limitations. On the one hand, such a two-stage approach requires offline feature extraction, storage, and alignment, which is inevitably time-consuming. This limitation poses a huge obstacle to real-time applications, *e.g.*, text-to-image retrieval, and video matting. On the other hand, the pre-trained panoramic segmentation model requires massive mask an-

notations, which place a greater burden on the already expensive expenditure of PNG. More importantly, the performance of these panoptic segmentation models also limits the upper bound of PNG models.

To solve the above problems, a natural way is to design an efficient single-stage network for end-to-end training from scratch. However, this solution also encounters two challenges that are critical for PNG. First, in PNG, each pixel can be subordinated to different masks, which is greatly different from panoptic segmentation (Kirillov et al. 2019a). This property makes the model need to capture visual semantics from macro- to micro-views. However, existing methods only focus on global modeling and overlook local information, resulting in limited performance. Second, PNG involves more complicated relationships than other grounding or segmentation tasks (Liu, Wang, and Yang 2017; Luo and Shakhnarovich 2017; Yu et al. 2017). In each example, multiple nouns of an expression may correspond to the same mask, or one noun may refer to multiple masks. This case further increases the difficulty of vision-language alignment.

In this paper, we propose a novel End-to-End Panoptic Narrative Grounding network (EPNG) for real-time panoptic narrative grounding, as shown in Fig. 1. Specifically, EPNG adopts a visual encoder to extract the features of the given image, based on which a decoder is deployed to predict masks for different noun phrases.

To enhance local semantic modeling, we introduce Locality-Perceptive Attention (LPA) to enhance grid features via neighborhood interactions based on their spatial priors. In LPA, different attention heads are allowed to perceive visual information in different receptive fields, thus achieving multi-scale modeling. To ensure the semantic consistency of many-to-many relationships in PNG, we design a new bidirectional Semantic Alignment Loss (SAL), which uses one modality as an anchor to eliminate the deviation of similar semantic tokens of the other modality. With these innovative designs, EPNG is superior in cross-modal reasoning while keeping real-time inference.

Conclusively, the contributions of our work are as below:

- We propose a real-time End-to-End Panoptic Narrative Grounding network (EPNG), which greatly reduces computation overhead via unifying cross-modal alignment and mask prediction in one forward structure.

- We propose two novel designs, namely Locality-Perceptive Attention (LPA) and bidirectional Semantic Alignment Loss (SAL). LPA enhances visual features at different scales to understand complex cross-modal relationships. SAL regularizes the semantic consistency problem by performing contrastive learning between pixels and noun phrases.

- On the benchmark dataset, EPNG is on par with or even better than existing two-stage methods, while its inference is 10 times faster. In addition, it requires no additional mask annotations for pre-training.

Related Work

Panoptic Segmentation

Panoptic segmentation aims to entirely understand scenes containing things and stuff. Following the benchmark proposed by (Kirillov et al. 2019b), the earlier methods treated it as the combination of things masks and stuff masks (de Geus, Meletis, and Dubbelman 2018), *e.g.*, PFPN (Kirillov et al. 2019a), Panoptic-DeepLab (Cheng et al. 2020a), and UPSNet (Xiong et al. 2019). Recently, things and stuff are expected to be treated uniformly (Carion et al. 2020; Wang et al. 2021; Cheng, Schwing, and Kirillov 2021). To eliminate the difference between things and stuff, part of those like PFCN (Li et al. 2021), K-Net (Zhang et al. 2021), and Panoptic SegFormer (Li et al. 2022) try to use the kernel to represent things and stuff uniformly and generate masks by the convolution on feature maps, which obtain significant performance. Benefiting from those methods, our model utilizes word features as reliable kernels to get corresponding masks through the convolution on multi-modal features.

Referring Expression Segmentation

Recently, multi-modal applications have received a lot of attention and made significant progress (Ji et al. 2022b, 2021; Ma et al. 2022b,a; Zhou et al. 2019). Among them, as a prevalent task in multi-modal communities, Referring Expression Segmentation (RES) (Hu, Rohrbach, and Darrell 2016; Yu et al. 2018; Ye et al. 2019; Shi et al. 2018) is to segment a referent based on the understanding of a related short phrase. In sequential order, previous models (Li et al. 2018; Liu et al. 2017; Margffoy-Tuay et al. 2018) obtain a set of proposals by a general method of segmentation and pick up a better one that is described by the given short phrase. With the strength of leveraging visual information, however, the upper bound of those methods is seriously restricted by the performance of the segmentation models. After that, a batch of methods is developed for refining segmentation masks by a single-stage network (Ye et al. 2019; Liu et al. 2019; Zhou et al. 2021), which brings higher rates of false positive segmentation. In summary, RES is an incomplete task with the neglect of stuff and many-to-many relationships between natural language and images. Additionally, whether things and stuff or the many-to-many relationships should be considered in Panoptic Narrative Grounding.

Panoptic Narrative Grounding

The existing method (González et al. 2021) handles it with a two-stage paradigm, which first obtains a lot of candidate panoptic masks by a pre-trained panoptic segmentation model. With those candidates, a scoring module is used to assign plural masks to referred phrases. This paradigm achieves impressive performance, nevertheless, the expensive computation cost and space cost on the stage of segment becomes the barrier to real-time. Because of the reasons above, we propose an End-to-End Panoptic Narrative Grounding network (EPNG) to generate the corresponding mask directly from the noun phrases.
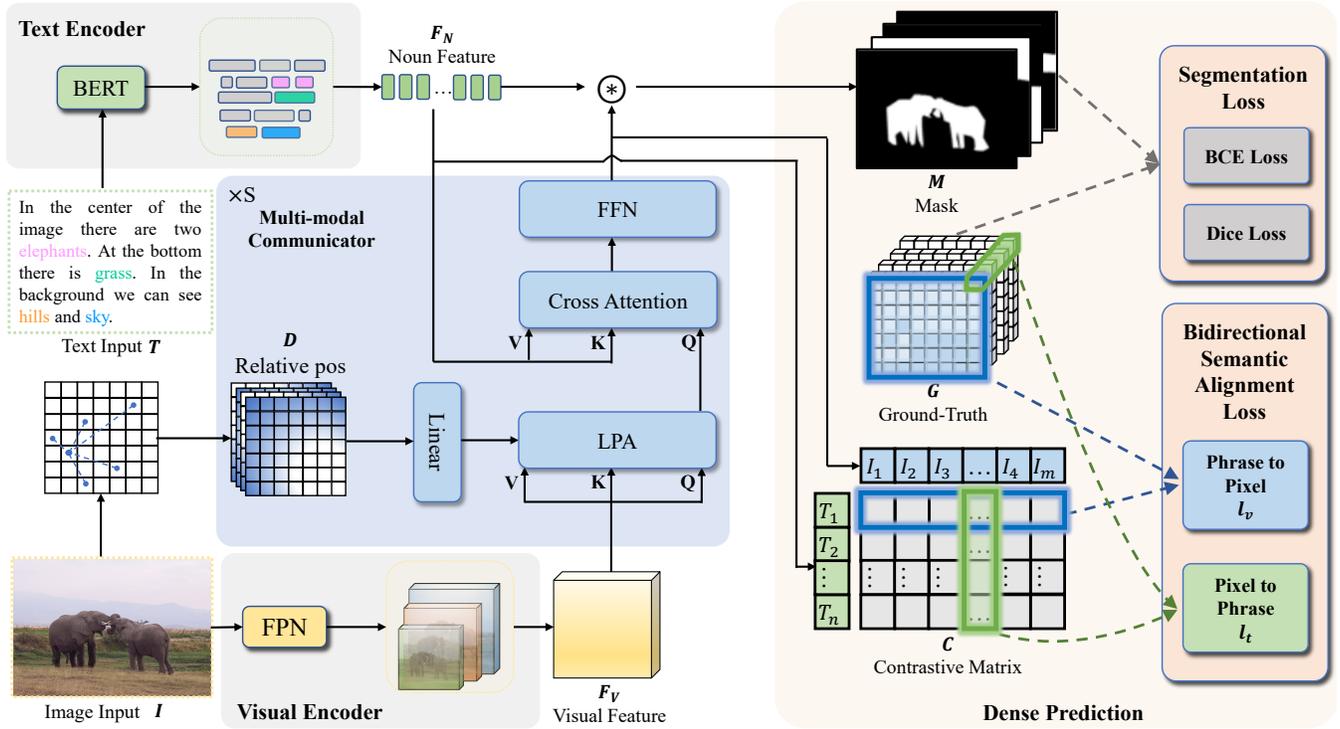
Figure 2: The framework of the proposed EPNG. The solid lines denote the pipeline of EPNG, while the dotted lines represent the loss computation during training. During the pipeline, a Multi-modal Encoding module is used to extract the features. Then a Multi-modal Communicator fuses multi-modal features with a Cross Attention module and the proposed LPA. Finally, traditional segmentation loss and the proposed SAL are set to improve the quality of segmentation and align the multi-modal information.

End-to-End Panoptic Narrative Grounding Network

In this section, we give a detailed description of our EPNG, of which the framework is illustrated in Fig. 2. The input images and descriptions are first processed by the visual and text encoders, respectively. A multi-modal fusion module is further deployed for image-text interaction, based on which a dense prediction head is used to predict masks.

Problem Definition

Unlike the existing two-stage PNG, the proposed one-stage PNG is free of mask proposals, which generates the mask directly based on the expressions and images. We formulate it as a cross-modal dense prediction task.

Specifically, given an image $\mathbf{I}$ and the corresponding text $\mathbf{T}$, the goal of PNG is to find the nouns $\mathbf{N} = \{n_\ell\}_{\ell=0}^L$ that each pixel i belongs to, where n_ℓ is the ℓ-th noun and L denotes the number of the noun phrases. Then the probability of the obtained mask $\mathbf{M} \in \{0,1\}$ is formulated as:

$$p(\mathbf{M}) = \prod_{i \in \mathbf{I}} \prod_{\ell=0}^{L} p(i|\mathbf{I}, \mathbf{T}, n_\ell). \quad (1)$$

Multi-modal Encoding

Visual Encoder Given an image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we first adopt a visual backbone (Lin et al. 2017) to extract the multi-scale visual features, $e.g.$, $\mathbf{F}_{v1} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_1}$, $\mathbf{F}_{v2} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_2}$, and $\mathbf{F}_{v3} \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_3}$. Then we obtain the final visual feature $\mathbf{F_v} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$ by:

$$\mathbf{F_v} = \text{concat}\left[\text{Down}(\mathbf{F}_{v1}); \mathbf{F}_{v2}; \text{Up}(\mathbf{F}_{v3})\right]. \quad (2)$$

where $\text{Up}(\cdot)$ denotes $2\times$ *upsampling*, $\text{Down}(\cdot)$ denotes $2\times$ *downsampling* and concat $[\cdot]$ denotes *feature concatenation*.

Text Encoder Given a sentence $\mathbf{T}$, we follow (González et al. 2021) to adopt a pre-trained BERT (Kenton and Toutanova 2019) to extract the word embeddings $\mathbf{F_T} = \{v_t\}_{t=0}^{|T|}$, where v_t denotes the embedding of t-th word. After that, we filter out the noun phrases according to the annotations given by (González et al. 2021) and then obtain the phrase features by average-pooling the word embeddings in each phrase. These features are then projected by a linear layer, making their feature dimension consistent with the visual features. As a result, the phrase embedding is denoted as $\mathbf{F_N} = \{f_{n_\ell}\}_{n_\ell=0}^{L} \in \mathbb{R}^{L \times C}$, where n_ℓ represents the ℓ-th noun phrases, and L is the number of phrases.

Multi-modal Communicator

Based on the visual feature $\mathbf{F_v}$ and the textual feature $\mathbf{F_N}$, Multi-Modal Communicator is designed for cross-modal interaction and fusion. It consists of S serial identical layer, and each layer is composed of two modules called Locality-Perceptive Attention (LPA) and Cross Attention (CA).

Locality-Perceptive Attention Similar to self-attention (Vaswani et al. 2017), LPA aims to improve the input features via modeling their inter-relationships. As argued in (Cheng et al. 2020b), local information is important for the visual segmentation tasks. Then EPNG, going a step further, presents multi-scale local modeling. Each pixel in an image may belong to different masks at the same time. For example, a pixel in cloth may also belong to a person. However, the standard self-attention treats all tokens in the feature map equally. To this end, we reinforce the role of neighborhood information of each pixel when in attention modeling, following (Wu, Wu, and Huang 2021; Ji et al. 2022a).

Specifically, in the features $\mathbf{F}^i$, the 2D spatial coordinates of the m-th and n-th vectors are denoted as (x_m, y_m) and (x_n, y_n), where the superscript i indicates that the feature map is the output of the layer i. Then we calculate the Euclidean Distance between these two coordinates:

$$\mathbf{D}_{m,n} = \sqrt{(x_m - x_n)^2 + (y_m - y_n)^2}, \qquad (3)$$

where $\mathbf{D} \in \mathbb{R}^{(H \times W) \times (H \times W)}$, and we truncate the values in $\mathbf{D}$ with an upper bound 2 to explicitly inject the local receptive information. Afterward, for an attention head j in LPA, we transform distance matrix into a coefficient matrix $\mathbf{R}^j \in \mathbb{R}^{(H \times W) \times (H \times W)}$, obtained by:

$$\mathbf{R}^j = \mathbf{W}^j \mathbf{D}. \qquad (4)$$

The obtained matrix $\mathbf{R}^j$ is used to re-weight the attention, which is given by:

$$\mathbf{A}^j = \text{Softmax}\left(\frac{(\mathbf{F}^i \mathbf{W}_Q^j)(\mathbf{F}^i \mathbf{W}_K^j)^T}{\sqrt{d_k}} \otimes \mathbf{R}^j\right), \qquad (5)$$

where the projections $\mathbf{W}_Q^j \in \mathbb{R}^{d \times \frac{d}{h}}$ and $\mathbf{W}_K^j \in \mathbb{R}^{d \times \frac{d}{h}}$ are weight matrices, and d_k is a scaling factor. The subscript j represents the j-th head, and the number of heads h is set to 8. $\otimes$ represents an element-wise product. In this way, we naturally embed local information into attention modeling.

Next, we sum the features using the attention weights to obtain the results for head j, and aggregate all the results:

$$\text{Head}^j = \mathbf{A}^j(\mathbf{F}^i \mathbf{W}_V^j), \qquad (6)$$

$$\text{LPA}(\mathbf{F}^i, \mathbf{F}^i, \mathbf{F}^i) = \text{concat}(\text{Head}^1, \cdots, \text{Head}^h)\mathbf{W}_O, \qquad (7)$$

where $\mathbf{W}_V^j \in \mathbb{R}^{d \times \frac{d}{h}}$, and $\mathbf{W}_O \in \mathbb{R}^{d \times d}$ are projection matrices.

$$\mathbf{F}^{i'} = \text{LN}\left(\text{LPA}\left(\mathbf{F}^i, \mathbf{F}^i, \mathbf{F}^i\right) + \mathbf{F}^i\right), \qquad (8)$$

where LN $(\cdot)$ means the *layer normalization* (Ba, Kiros, and Hinton 2016), and shortcut connection (He et al. 2016) is applied after the LPA module.

The difference between LPA and Multi-Head Attention (MHA). As shown in Eq. 5, the definition of LPA is based on Multi-head Attention (Vaswani et al. 2017), but it still has an obvious difference in principle. MHA treats all tokens in the feature map equally and excels at capturing long-range dependencies. However, local information is inevitably ignored in this process. We introduce the local prior of each grid, which is obtained from the distance matrix to attention modeling. Note that, these local priors can be dynamically adjusted by the coefficient matrix $\mathbf{R}^j$ in Eq. 4.

Cross-modal Attention Following (Yu et al. 2019), we use Cross-modal Attention for modality interactions:

$$\mathbf{F}^{i+1} = \text{FFN}\left(\text{LN}\left(\text{MHA}\left(\mathbf{F}^{i'}, \mathbf{F}_\mathbf{N}, \mathbf{F}_\mathbf{N}\right) + \mathbf{F}^{i'}\right)\right), \quad (9)$$

where FFN $(\cdot)$ denotes feed-forward network, and $\mathbf{F}_\mathbf{N}$ is the noun features.

Dense Prediction

Given the fused feature after multi-modal communicator, $\mathbf{F} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$, we upsample it to a tensor shape of $\frac{H}{4} \times \frac{W}{4} \times C$. Afterward, we apply each noun phrase feature as a kernel to convolve $\mathbf{F}$. The final masks $\mathbf{M}$ are obtained by:

$$\mathbf{M} = \text{Up}\left(\text{Sigmoid}\left(\mathbf{F}_\mathbf{N} * \mathbf{F}\right)\right), \qquad (10)$$

where $*$ represents convolution operation, and *Sigmoid* transforms the results to $(0, 1)$. After upsampling, we set a threshold to force $\mathbf{M} \in \{0, 1\}$.

Training loss Since panoptic narrative grounding is a segmentation task, we try different *seg.* losses in existing tasks, including BCE loss and Dice loss (Milletari, Navab, and Ahmadi 2016). For BCE loss, the loss function is formulated as

$$L_{BCE} = \sum_{\hat{y_i} \in \mathbb{M}} -(y_i \cdot \log(\hat{y_i}) + (1 - y_i) \cdot \log(1 - \hat{y_i})),$$

$$(11)$$

where $\hat{y_i}$ is the prediction of the i-th pixel and y_i is the ground-truth. Dice loss is defined by

$$L_{Dice} = 1 - \frac{2|M \bigcap G|}{|M| + |G|}, \qquad (12)$$

where M is the generated mask and G is the ground truth, the value of which all belongs to $\{0, 1\}$. Considering these loss functions are designed for single-modal tasks (He et al. 2017), we propose a new loss called bidirectional Semantic Alignment Loss (SAL) for regularizing the semantic consistency between modalities.

Bidirectional Semantic Alignment Loss As mentioned above, PNG has complex many-to-many relationships, *i.e.*, a mask may belong to several noun phrases or *vice versa*. However, the above segmentation loss only considers the one-to-one interaction between the phrases and mask, while ignoring the semantic connection between them. Inspired by (Kamath et al. 2021), we design a SAL to guarantee semantic consistency, which encourages the multi-modal features with the same semantics to be similar.

Specifically, we first adopt noun phrases as anchors to improve the semantic consistency within visual features. For i-th noun phrase $\mathbf{F}_N^i \in \mathbb{R}^C$ and the ground-truth $G^i \in \{0, 1\}^{H \times W}$, the collection of pixels with the class of "1" is considered as the positive set, while the one of "0" is gathered as the negative set. By increasing the similarity within the positive set, multi-modal information is forced to be aligned. Considering all nouns together, the loss is introduced as follows:

$$l_v = \frac{1}{L}\sum_{i=0}^{L}\frac{1}{|G^+|}\sum_{j \in G^+} -\log\left(\frac{\exp\left(\mathbf{F}_n^i \cdot \mathbf{F}^j/\tau\right)}{\sum_{k \in G}\exp\left(\mathbf{F}_n^i \cdot \mathbf{F}^k/\tau\right)}\right), \quad (13)$$

Method	Average Recall					Inference Time	Params			Training Data
	All	Thing	Stuff	Single	Plural	All	Stage-1	Stage-2	All	All
PNG	55.4	56.2	54.3	56.2	48.8	107ms	21.0M	240.3M	261.3M	2.1M
Baseline (ours)	40.3	34.5	50.5	42.3	31.4	**9.5ms**	-	-	**76.5M**	**0.8M**
EPNG (ours)	49.7	45.6	55.5	50.2	45.1	11ms	-	-	**76.5M**	**0.8M**
EPNG* (ours)	**58.0**	**54.8**	**62.4**	**58.6**	**52.1**	11ms	-	-	**76.5M**	2.1M

Table 1: Comparison of the EPNG and the previous two-stage method PNG (González et al. 2021). Baseline means the same design as EPNG except for LPA and SAL. EPNG* is trained with the same data as PNG.

Communicator	Average Recall				
	All	Thing	Stuff	Single	Plural
w/o PE	42.8	36.5	52.1	44.2	32.5
PE	46.7	43.0	52.9	47.6	42.4
SPE	46.1	41.8	52.2	46.6	42.2
DPE	45.9	41.5	52.0	46.5	40.3
RPE	45.3	42.1	52.2	46.9	41.5
LPA	**49.7**	**45.6**	**55.5**	**50.2**	**45.1**

Table 2: Ablation study of the LPA module, where "w/o PE" denotes no position embedding, and "PE" (Vaswani et al. 2017), "SPE" (Liu et al. 2021), "DPE" (Zhu et al. 2020), and "RPE" (Dosovitskiy et al. 2020) mean different relative position embedding from other methods.

Loss	Average Recall				
	All	Thing	Stuff	Single	Plural
BCE + Dice	43.7	38.5	49.6	43.7	37.6
BCE + Dice + SAL	**49.7**	**45.6**	**55.5**	**50.2**	**45.1**

Table 3: Ablation study of the SAL.

Dataset		Type	mIoU	p@0.3	p@0.5
RefCOCO	testA	random	9.2	2.5	0.1
		zero-shot	28.0	42.5	12.4
	testB	random	9.8	3.3	0.3
		zero-shot	17.7	22.4	7.8
RefCOCO+	testA	random	9.3	2.6	0.2
		zero-shot	27.7	41.1	11.9
	testB	random	10.5	3.8	0.3
		zero-shot	20.6	27.4	9.4
RefCOCOg	test	random	9.9	4.0	0.6
		zero-shot	27.4	40.1	16.3

Table 4: Zero-shot results of EPNG on RES. EPNG is not trained with RES data.

where G^+ is the positive set, denotes the class of "1" in the ground truth, and G is the ground-truth. τ is a temperature coefficient.

Next, We adopt pixel features as anchors to improve the semantic consistency within noun features. Similarly, the loss is introduced as follows:

$$l_t = \frac{1}{|G|}\sum_{i=0}^{|G|}\frac{1}{|T^+|}\sum_{j\in T^+} -\log\left(\frac{\exp\left(\mathbf{F}^i \cdot \mathbf{F}_n^j/\tau\right)}{\sum_{k\in T}\exp\left(\mathbf{F}^i \cdot \mathbf{F}_n^k/\tau\right)}\right), \quad (14)$$

where T^+ is the positive gather of the noun set, and T is the whole set, where "1" denotes the pixel belonging to this noun phrase. We combine Eq. 13 and Eq. 14 as the SAL loss function.

During the training, we use the summation of Dice loss, BCE loss, and SAL:

$$L = \lambda_1 L_{BCE} + \lambda_2 L_{Dice} + \lambda_3 L_{SAL}, \quad (15)$$

where λ_1, λ_2 and λ_3 are the hyper-parameters.

Experiment

Datasets

We train and compare our model with the existing method on the Panoptic Narrative Grounding dataset (González et al.

2021). It is consist of images and the corresponding text. Unlike the brief phrase in other datasets such as RefCOCO (Yu et al. 2016), the texts of PNG are long and are a narrative of all items in the complete image and their relationships. It often has hundreds of words and more complex semantic information. The noun-level segmentations are provided for each text. It encompasses both the thing and the stuff, similar to panoptic segmentation. The difference is that the thing will include both the singular and the plural according to the semantics, which also brings more difficulty for vision-text alignment. The dataset includes a total of 133,103 training images and 8,380 test images with 875,073 and 56,531 segmentation annotations, respectively.

Implementation Details

Experimental Settings In this paper, we follow (González et al. 2021) to use the ResNet-101 as our visual backbone, which is pre-trained on the ImageNet (Krizhevsky, Sutskever, and Hinton 2012). BERT is used as the text backbone. During the training process, all the backbones are frozen except for the last two layers of ResNet-101. For parallel training, we increase the input image resolution to 640×640, so the shapes of the last three layers are $20\times20\times256$, $40\times40\times256$, and $80\times80\times256$, respectively. Moreover, the dimension of text features is 768. The number of attention heads is 8 and the hidden dimension is 2048. Besides, the number of Layers S is 3. In terms of hyper-parameters, we use $\lambda_1 = 2$, $\lambda_2 = 2$ and $\lambda_3 = 1$ to balance the final loss. We set the initial learning rate $\eta = 1e^{-5}$ which is half decayed by every 5 epochs, and fix $\eta = 5e^{-7}$ after 10 epochs. The batch size is 32. We train it on 4 RTX3090 GPUs, which cost 20 hours in total. The optimizer is Adam.
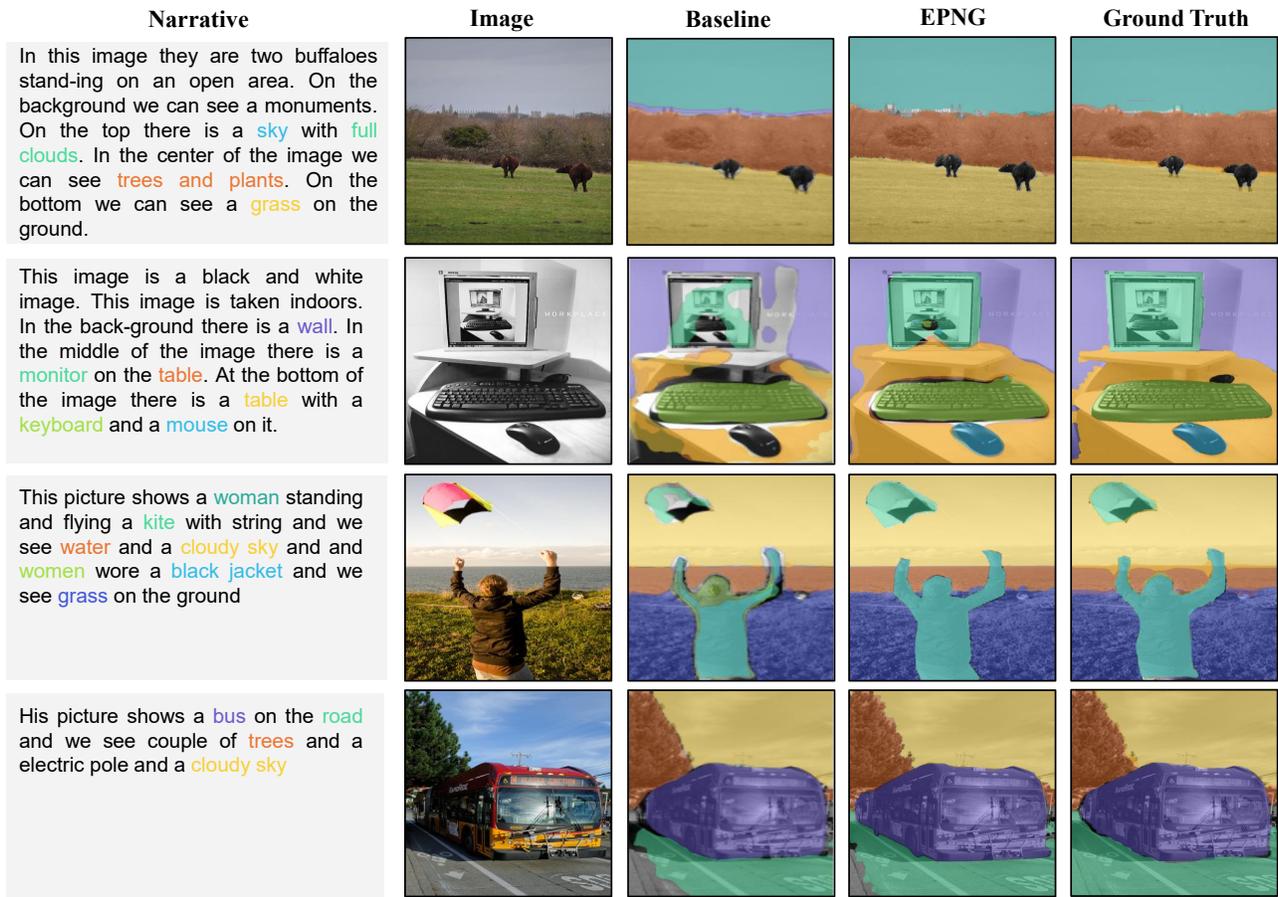
Figure 3: Visualization of EPNG. We mark the same color between the nouns in the narrative and the referred pixels.
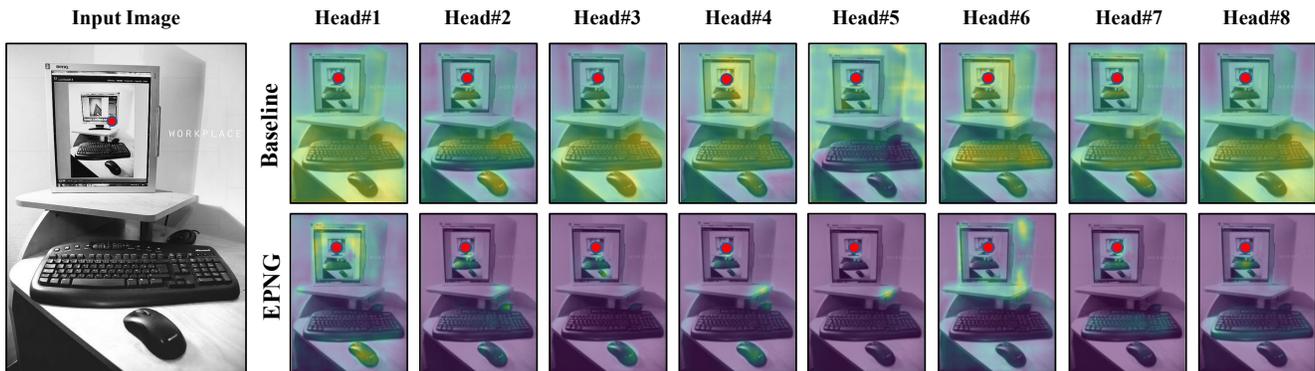
Figure 4: Comparison of the attention of the sampled point for the LPA and MHA from the top layer. The red point represents the sampled point.

Metrics Following (González et al. 2021), we use Average Recall (González et al. 2021) as our metric. Specifically, we calculate Intersection over Union (IoU) between masks and the ground truth. We use the integral of the IoU curve as the final metric. Additionally, we simultaneously analyze this measure for the thing, stuff, single, and plural.

Quantitative Analysis

Comparison with the state-of-the-arts. We first evaluate the overall performance of the model using the Average Re-

Referring	Image	Prediction	Ground Truth
the man playing tennis			
man			
front left chicken			
a lady wearing jeans and a pink and gray north face jacket			
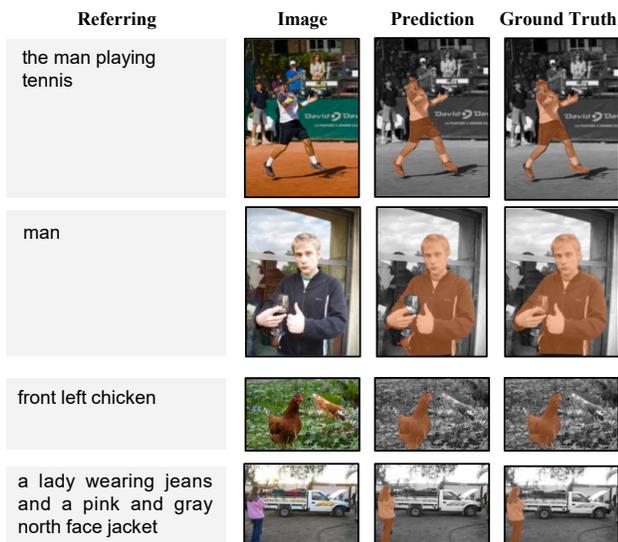
Figure 5: The visualization of zero-shot setting for RES.

call metric, as shown in Tab. 1. In Tab. 1, we also introduce a baseline model for comparison, which is the same as EPNG except for LPA and SAL. The performance of the one-stage baseline is much inferior to the two-stage PNG. It is because single-stage models are trained end-to-end from scratch, suffering from greater training difficulties. After adding LPA and SAL, the proposed EPNG can bring up to 23.3% gain on all the masks, 32.2% on the thing, and 9.9% on the stuff, respectively. This fully demonstrates the effectiveness of the proposed method. In terms of the inference speed, we set the batch size to 12, and calculate the average inference time on each image. Our EPNG has a significant benefit, being 10× faster than the two-stage model and using only 38% of its parameters, allowing for model deployment on edge devices.

To make a fair comparison with PNG, we adopt FPN with a ResNet-101 backbone pre-trained with Panoptic Feature Pyramid Network (Kirillov et al. 2019a) on MS COCO. Our EPNG achieves better performance (*i.e.*, 2.6 points of performance gain) at 10× inference speed, which better demonstrates the contribution of our EPNG.

Ablation Study   To verify the contribution of our proposed LPA and SAL, we conduct ablation experiments on the two modules, respectively. In Tab. 2, we compare LPA with other position embeddings. As can be seen that the method that does not use any location and distance information performs the worst, demonstrating the importance of location information for this task. Meanwhile, our LPA achieves the best performance than other relative position embeddings, which are widely used to capture location information.

To verify the efficiency of SAL, we perform two experiments with and without it, of which results are given in Tab. 3. It can be seen that SAL brings a significant improvement of 6.0 points to EPNG, which fully validates our motivation that the improvement of semantic consistency can well improve the segmentation accuracy in PNG.

Zero-Shot Study for RES   Meanwhile, we validate our model's generalization by conducting zero-shot experiments on the datasets of the RES task, *e.g.,* RefCOCO (Yu et al. 2016), RefCOCO+ (Yu et al. 2016), and RefCOCOg (Mao et al. 2016; Nagaraja, Morariu, and Davis 2016). These datasets are built based on MS COCO (Lin et al. 2014). We use the feature of the whole phrase as the text feature. The results are shown in Tab. 4. By comparing with the fully randomized model, we can find that the zero-shot performance of EPNG is significantly improved. Even on the more complex dataset *i.e.,* RefCOCO+, our performance is close to some early supervised REC models, like (Liu et al. 2017), mIoU of which is 30.48 and 29.5 on the testA and testB of RefCOCO+, respectively.

Qualitative Analysis

Visualization   As shown in Fig. 3, we present some typical grounding results from EPNG compared to the ground truth. Compared to the baseline, our method can generate more accurate masks, especially for the edge parts. This further proves the effectiveness of our method.

Attention Visualization   To figure out the role of LPA in our proposed method, we visualize its attention weights during inference in Fig. 4. Compared with the baseline, LPA presents diverse local attention patterns with miscellaneous scopes, which brings a powerful capability to model local semantic relationships. This also validates the motivation of LPA, and our proposed method does allow the model to focus locally and improve the accuracy of attention.

Zero-Shot for RES   Additionally, Fig. 5 shows some qualitative results of our zero-shot study on RES. Our method can achieve very accurate segmentation, which fully demonstrates its transferability. For example, in the second case, our model identifies the segmentation of "man", which is wrong even in the ground truth. Due to the ability of the finest-grained and complex semantic understanding, EPNG can handle more general scenarios with greater potential.

Conclusion

In this paper, we proposed an End-to-End Panoptic Narrative Grounding network (EPNG) for real-time inference. To better handle the many-to-many relationships between pixels and phrases, two innovative designs are proposed, namely Local-Sensitive Attention (LPA) and bidirectional Semantic Alignment Loss (SAL), respectively. Extensive experiments show that our proposed EPNG achieves significant performance gains compared to the baseline. More importantly, compared to the two-stage model, EPNG achieves competitive performance with faster inference (10x) and fewer parameters. Furthermore, we conducted zero-shot experiments on RES and achieved surprising performance. These results demonstrate the excellent generalization of our model and also provide a reference for a subsequent unified vision-language segmentation framework.

Acknowledgments

References

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer Normalization. arXiv:1607.06450.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Cheng, B.; Collins, M. D.; Zhu, Y.; Liu, T.; Huang, T. S.; Adam, H.; and Chen, L.-C. 2020a. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12475–12485.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34: 17864–17875.

Cheng, H. K.; Chung, J.; Tai, Y.-W.; and Tang, C.-K. 2020b. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8890–8899.

de Geus, D.; Meletis, P.; and Dubbelman, G. 2018. Panoptic Segmentation with a Joint Semantic and Instance Segmentation Network. arXiv:1809.02110.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

González, C.; Ayobi, N.; Hernández, I.; Hernández, J.; Pont-Tuset, J.; and Arbeláez, P. 2021. Panoptic Narrative Grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1364–1373.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, R.; Rohrbach, M.; and Darrell, T. 2016. Segmentation from natural language expressions. In *European Conference on Computer Vision*, 108–124. Springer.

Ji, J.; Huang, X.; Sun, X.; Zhou, Y.; Luo, G.; Cao, L.; Liu, J.; Shao, L.; and Ji, R. 2022a. Multi-Branch Distance-Sensitive Self-Attention Network for Image Captioning. *IEEE Transactions on Multimedia*.

Ji, J.; Luo, Y.; Sun, X.; Chen, F.; Luo, G.; Wu, Y.; Gao, Y.; and Ji, R. 2021. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1655–1663.

Ji, J.; Ma, Y.; Sun, X.; Zhou, Y.; Wu, Y.; and Ji, R. 2022b. Knowing what to learn: a metric-oriented focal mechanism for image captioning. *IEEE Transactions on Image Processing*, 31: 4321–4335.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. MDETR-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1780–1790.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Kirillov, A.; Girshick, R.; He, K.; and Dollár, P. 2019a. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6399–6408.

Kirillov, A.; He, K.; Girshick, R.; Rother, C.; and Dollár, P. 2019b. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9404–9413.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Li, R.; Li, K.; Kuo, Y.-C.; Shu, M.; Qi, X.; Shen, X.; and Jia, J. 2018. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5745–5753.

Li, Y.; Zhao, H.; Qi, X.; Wang, L.; Li, Z.; Sun, J.; and Jia, J. 2021. Fully convolutional networks for panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 214–223.

Li, Z.; Wang, W.; Xie, E.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; Luo, P.; and Lu, T. 2022. Panoptic SegFormer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1280–1289.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, C.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; and Yuille, A. 2017. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 1271–1280.

Liu, D.; Zhang, H.; Wu, F.; and Zha, Z.-J. 2019. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4673–4682.

Liu, J.; Wang, L.; and Yang, M.-H. 2017. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, 4856–4864.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Luo, R.; and Shakhnarovich, G. 2017. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7102–7111.

Ma, Y.; Ji, J.; Sun, X.; Zhou, Y.; Wu, Y.; Huang, F.; and Ji, R. 2022a. Knowing what it is: Semantic-enhanced Dual Attention Transformer. *IEEE Transactions on Multimedia*.

Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; and Ji, R. 2022b. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 638–647.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.

Margffoy-Tuay, E.; Pérez, J. C.; Botero, E.; and Arbeláez, P. 2018. Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 630–645.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. IEEE.

Nagaraja, V. K.; Morariu, V. I.; and Davis, L. S. 2016. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, 792–807. Springer.

Shi, H.; Li, H.; Meng, F.; and Wu, Q. 2018. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 38–54.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H.; Zhu, Y.; Adam, H.; Yuille, A.; and Chen, L.-C. 2021. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5463–5474.

Wu, C.; Wu, F.; and Huang, Y. 2021. DA-Transformer: Distance-aware Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2059–2068.

Xiong, Y.; Liao, R.; Zhao, H.; Hu, R.; Bai, M.; Yumer, E.; and Urtasun, R. 2019. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8818–8826.

Ye, L.; Rochan, M.; Liu, Z.; and Wang, Y. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10502–10511.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1307–1315.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, 69–85. Springer.

Yu, L.; Tan, H.; Bansal, M.; and Berg, T. L. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7282–7290.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6281–6290.

Zhang, W.; Pang, J.; Chen, K.; and Loy, C. C. 2021. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34: 10326–10338.

Zhou, Y.; Ji, R.; Luo, G.; Sun, X.; Su, J.; Ding, X.; Lin, C.-W.; and Tian, Q. 2021. A real-time global inference network for one-stage referring expression comprehension. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhou, Y.; Ji, R.; Sun, X.; Su, J.; Meng, D.; Gao, Y.; and Shen, C. 2019. Plenty is plague: Fine-grained learning for visual question answering. *IEEE transactions on pattern analysis and machine intelligence*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.