The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

Improving Panoptic Narrative Grounding by Harnessing Semantic Relationships and Visual Confirmation

Tianyu Guo[*], **Haowei Wang**[*], **Yiwei Ma**, **Jiayi Ji**[†], **Xiaoshuai Sun**

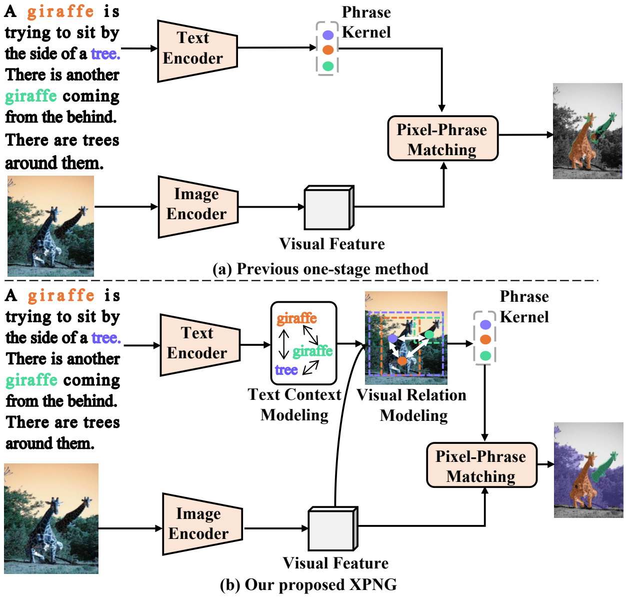
Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China,
Xiamen University, 361005, P.R. China
{guotianyu, wanghaowei, yiweima}@stu.xmu.edu.cn, jjyxmu@gmail.com, xssun@xmu.edu.cn

Abstract

Recent advancements in single-stage Panoptic Narrative Grounding (PNG) have demonstrated significant potential. These methods predict pixel-level masks by directly matching pixels and phrases. However, they often neglect the modeling of semantic and visual relationships between phrase-level instances, limiting their ability for complex multi-modal reasoning in PNG. To tackle this issue, we propose XPNG, a "differentiation-refinement-localization" reasoning paradigm for accurately locating instances or regions. In XPNG, we introduce a Semantic Context Convolution (SCC) module to leverage semantic priors for generating distinctive features. This well-crafted module employs a combination of dynamic channel-wise convolution and pixel-wise convolution to embed semantic information and establish inter-object relationships guided by semantics. Subsequently, we propose a Visual Context Verification (VCV) module to provide visual cues, eliminating potential space biases introduced by semantics and further refining the visual features generated by the previous module. Extensive experiments on PNG benchmark datasets reveal that our approach achieves state-of-the-art performance, significantly outperforming existing methods by a considerable margin and yielding a 3.9-point improvement in overall metrics. Our codes and results are available at our project webpage: https://github.com/TianyuGoGO/XPNG.

1 Introduction

Recently, the growing interest in multimodal research (Fei 2022; Li et al. 2022a; Chen et al. 2022; Jing et al. 2020; Ma et al. 2022, 2023; Ji et al. 2022; Huang et al. 2023; Zhao et al. 2023; Wu et al. 2023) at the intersection of computer vision and natural language processing has driven the development of systems that can understand and describe the world as humans do. Panoptic Narrative Grounding (PNG) (González et al. 2021) is an emerging visually-grounded language understanding task that aims to locate and segment all instances of objects and regions in an image, corresponding to a given text description using binary pixel masks. This task goes beyond conventional grounding tasks, such as Referring Expression Segmentation (RES) (Cheng et al. 2021; Li, Bu, and Cai 2021; Liao et al. 2022; Liu et al. 2021; Luo et al. 2020a), by involving the joint understanding of multi-modal information and necessitating many-to-many language-vision alignment, which adds complexity to the task.



(a) Previous one-stage method

(b) Our proposed XPNG

Figure 1: A comparison between the previously mentioned one-stage method, PPMN (Ding et al. 2022), and our proposed XPNG framework. PPMN overlooks modeling relationship, resulting in an inability to distinguish between two "giraffes". In contrast, our proposed XPNG achieves accurate segmentation. Notably, it has only a slightly higher parameter count compared to PPMN, as shown in Tab. 1.

Previous PNG research primarily involves a two-stage paradigm. These models first use pre-trained panoptic segmentation models (Kirillov et al. 2019) to generate a set of candidate masks for a given image and then transform these candidates into features, which are ranked via cross-modal matching with language features. However, two-stage models (González et al. 2021) have limitations: they are less efficient due to the separated segmentation and matching processes, and they interact with language in a matching way,

[*] These authors contributed equally.

[†] The corresponding author.

which cannot rectify inaccurate segmentation. In contrast, single-stage models (Wang et al. 2023b; Ding et al. 2022) address these limitations by incorporating text as a condition and fusing visual features with textual information to directly predict pixel-level masks, improving overall accuracy and efficiency.

Despite the success of single-stage models, challenges remain in the context of Panoptic Narrative Grounding (PNG). Existing single-stage models primarily focus on cross-modal relationships between phrases and pixels, calculating similarity based on phrase representations and pixel values as binary mask results. However, they often overlook interactions among instances referred to by phrases and their corresponding regions. This limitation restricts the models' reasoning capabilities, leading to segmentation errors. As shown in Fig. 1, for example, when an image contains two distinct "giraffes" and is accompanied by a long text description, it becomes difficult to accurately identify the correspondence between phrases and image regions solely based on isolated phrase features. Similar issues can arise in complex textual or visual scenes. The key to solving this problem is to consider relationships among different instances or regions.

Modeling these relationships is not a simple task, as it involves a three-step reasoning process. First, the textual semantic context and visual features must be fused and interacted with to fully understand the textual meaning and generate a distinctive representation, such as differentiating between "a giraffe" and "another giraffe" in Fig. 1. Next, instances should be used to further explore visual cues, refine the instance representation and obtain a semantic representation kernel with sufficient discrimination to accurately locate the target. Finally, the semantic representation kernel is utilized to complete target localization.

In this paper, we propose a novel approach called XPNG, which constructs a "differentiation-refinement-localization" reasoning paradigm to accurately locate instances or regions. First, we introduce a Semantic Context Convolution (SCC) module to create distinctive representation kernels. Relying solely on the phrase's semantic features are insufficient, so we need to leverage visual information to further enhance this capability. To achieve this, following a dynamic channel-wise convolution to embed visual information into the text, a pixel-wise convolution models relationships between instances or regions, resulting in distinctive representation kernels. Subsequently, we design a Visual Context Verification (VCV) module to further refine the feature kernel with visual clues, ultimately obtaining a comprehensive semantic kernel. This modeling step creates a customized representation, perfectly integrating an instance's semantic and appearance information. Finally, the obtained semantic kernel can be used to match the target.

In summary, our contributions are three-fold as follows:

- We propose a three-step reasoning paradigm, XPNG, which constructs a "differentiation-refinement-localization" reasoning paradigm to enhance the model's cross-modal reasoning capabilities.
- We introduce the Semantic Context Convolution (SCC) to leverage prior semantic information for improving feature discriminability, and the Visual Context Verification (VCV) to incorporate geometric information, eliminating biases and further refining features.
- Our experimental results demonstrate that XPNG achieves a new state-of-the-art segmentation performance with a score of 63.3%, surpassing the current state-of-the-art method PPMN by 3.9%.

2 Related Work

2.1 Panoptic Segmentation

The field of panoptic segmentation (Kirillov et al. 2019; Hu et al. 2021, 2023) has recently gained significant attention due to its ability to assign a semantic label and instance ID to each pixel. More recently, with the emergence of Transformer (Vaswani et al. 2017), some methods (Zhang et al. 2021; Cheng, Schwing, and Kirillov 2021; Li et al. 2022b; Wang et al. 2021) have adopted an end-to-end set prediction objective, using attention blocks to generate panoptic masks. The proposal of these methods enables panoptic segmentation to be used for various application tasks, including autonomous navigation (Kiran et al. 2021; Moosavi et al. 2021), augmented reality (Alhaija et al. 2017), and virtual reality (Giannitrapani, Trucco, and Murino 1999). In contrast to these methods, PNG (González et al. 2021) aims to generate panoptic segmentation for an image using dense narrative captions.

2.2 Referring Expression Segmentation

Referring Expression Segmentation (RES) involves predicting foreground pixels for the object described by a given referring expression. One-stage frameworks (Suo et al. 2021; Li and Sigal 2021; Hu, Rohrbach, and Darrell 2016) have been proposed. To model semantic relationships between vision and language, recent methods (Ding et al. 2021; Feng et al. 2021; Jiao et al. 2021; Li and Sigal 2021; Yang et al. 2022; Luo et al. 2020b)incorporate complex cross-attention mechanisms inspired by the powerful abilities of Transformers (Vaswani et al. 2017) for capturing long-range dependencies.

2.3 Panoptic Narrative Grounding

PNG (González et al. 2021) proposed a two-stage paradigm for handling the task at hand. In the first stage, a pre-trained panoptic segmentation model (Kirillov et al. 2019; González et al. 2023) is used to generate a large number of candidate panoptic masks. Subsequently, a scoring module is employed to assign plural masks to referred phrases. Although this approach achieves impressive performance, the computation and space costs incurred during the segmentation stage pose a barrier to real-time implementation. The one-stage models (Ding et al. 2022; Yang et al. 2023; Wang et al. 2023a,b; Hui et al. 2023; Lin et al. 2023) provide an improved approach by overcoming the limitations of the traditional methods. They achieve this by incorporating the textual information as a condition and fusing it with visual features to predict pixel-level masks directly. This integration leads to a boost in accuracy and efficiency, thereby enhancing the performance of the overall system.

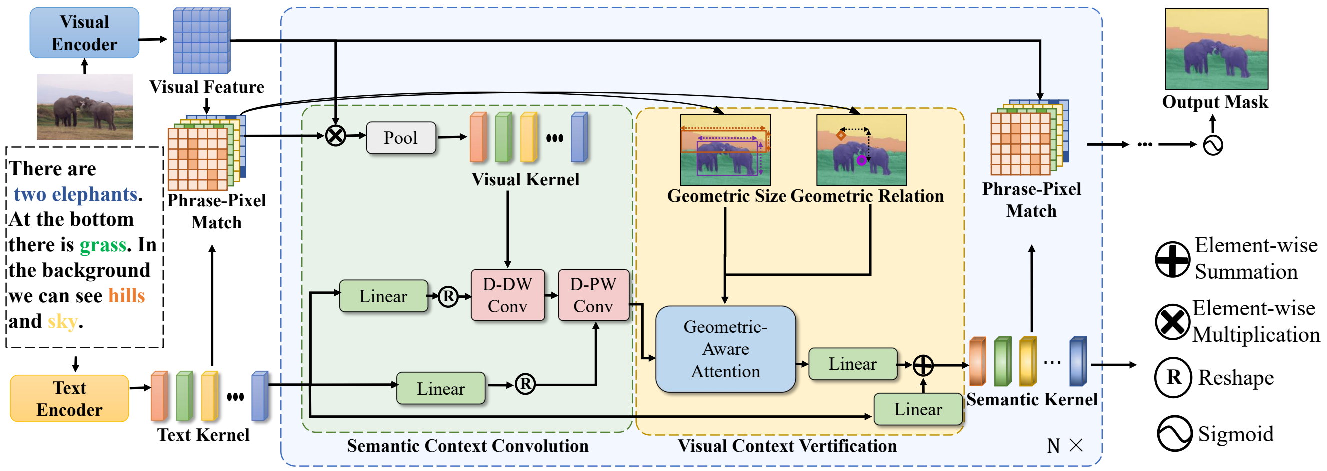


Figure 2: Overview of our proposed XPNG. We employ a visual encoder to extract visual features map F_v. For the linguistic modality, we use a text encoder to extract noun phrase features F_n. Our model consists of multiple iterative stages. First, we utilize the Semantic Context Convolution (SCC) module to generate a discriminative kernel K_v. Next, we employ the Visual Context Verification (VCV) module, which leverages the geometric information of masks, to eliminate semantic ambiguity arising from context guidance.

3 Method

In Sec. 3.1, we first present a comprehensive overview of the feature extraction process for both visual and linguistic patterns. Following this, Sec. 3.2 introduces the Semantic Context Convolution (SCC) module, which effectively constructs segmentation kernels using image features. Subsequently, in Sec. 3.3, we describe the Visual Context Verification (VCV) module, which refines the kernels by incorporating the geometric information between objects and facilitating interactions between the visual and language modalities through a control network. The entire pipeline of our proposed model is illustrated in Fig. 2.

3.1 Features Extraction

Visual Encoder As illustrated in Fig. 2, given an image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ with original dimensions H and W, we first employ a Feature Pyramid Network (FPN) (Lin et al. 2017) with a ResNet-101 backbone (He et al. 2016) to extract multi-scale visual features. These features are represented as $\mathbf{F}_{v1} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$, $\mathbf{F}_{v2} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$, $\mathbf{F}_{v3} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$, and $\mathbf{F}_{v4} \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C}$. Given the significance of position information, we incorporate pixel position encoding (Ding et al. 2022) into the visual features. Following this step, we employ an FPN neck (Kirillov et al. 2019) to aggregate features from different layers, resulting in F_v, which contains both multi-scale information and position information.

Text Encoder Given a sentence $\mathbf{T}$, we adopt the approach from (González et al. 2021) and use a pre-trained BERT model (Devlin et al. 2018) to extract token embeddings $\mathbf{F_t} = \{f_i\}_{i=0}^{|T|}$, where f_i denotes the embedding of the i-th token. We then filter out noun phrases based on the annotations provided by (González et al. 2021; Pont-Tuset et al. 2020) and generate phrase features by average-pooling the token embeddings within each phrase. We employ a linear layer to transform these phrase features, ensuring they have the same dimensions as the visual features. The noun features are represented as $\mathbf{F_n} = \{f_\ell\}_{\ell=0}^{L} \in \mathbb{R}^{L \times C}$, where f_ℓ corresponds to the ℓ-th noun phrase, and L indicates the total number of phrases.

3.2 Semantic Context Convolution

Previous methods (González et al. 2021; Wang et al. 2023b) incorporate text as a condition and fuse visual features with textual information to directly predict pixel-level masks, thereby enhancing overall accuracy and efficiency. However, these noun phrases and their corresponding representations are treated independently throughout the process, without any interaction. Such interactions are crucial for understanding semantics and relationships within a scene, as discussed in Sec. 1. To model these relationships, we introduce a module, called Semantic Context Convolution (SCC), which leverages the prior information from the text modality to construct internal relationships between phrases, thus facilitating the establishment of semantic relationships between related regions. Inspired by (Zhang et al. 2021), the Semantic Context Convolution (SCC) module in Sec. 3.2 and the Visual Context Verification (VCV) module in Sec. 3.3 both involve multiple stages.

Generation of Visual Kernels The core idea of our approach is to model the relationships between objects (regions). First, we need to identify the visual representations corresponding to each noun, referred to as visual kernels. Directly using text (semantic) kernels is not suitable, as they lack visual characteristics such as appearance and spatial location, leading to information loss. For the stage s, we use the segment kernels K^{s-1} from the stage $(s-1)$ and visual features F_v to obtain the visual kernels K_v^s.

$$M^{s-1} = \text{sigmoid}\left(K^{s-1} * F_v\right), \tag{1}$$

$$M^{s-1} = \begin{cases} 0, & M^{s-1} \leq 0.5 \\ 1, & M^{s-1} > 0.5 \end{cases}, \tag{2}$$

where $*$ denotes the convolution operation. After predicting the results for each mask M^{s-1} in the stage $s-1$, the threshold should be filtered to obtain the final mask graph M^{s-1}, where $M^{s-1} \in \mathbb{R}^{L \times H \times W}$. H and W are the resolutions of masks. L is the number of masks. Specifically, the segment kernels K^s of the 0-th stage are derived from F_n in Sec. 3.1, *e.g.*, $K^0 = F_n$.

The subsequent step consists of performing element-wise multiplication and summation between the masked result M^{s-1} and the visual features F_v, followed by averaging the outcome to obtain the visual convolution kernels K_v^s, $K_v^s \in \mathbb{R}^{L \times C}$, where L is the number of nouns. Each kernel corresponds to an object.

$$K_v^s = \text{Pool}\left(M^{s-1} \otimes F_v\right). \tag{3}$$

Leveraging Semantic Priors for Visual Feature Interaction Our approach begins by using semantic priors to guide the interaction of visual features in images. This helps the model to understand the potential relationships between objects in the image, such as a "person" and a "bicycle" is likely to have a relationship. To achieve this, we enhance the visual kernels K_v^s by incorporating language modality across channels and pixels. Following (Hu et al. 2023), the visual kernels are then passed through a channel-wise convolution block and a pixel-wise convolution block, with parameters derived from K^{s-1}. This can be expressed as:

$$\begin{cases} W_{Channel}^s = \text{reshape }\left(K^{s-1} W_c^s\right), \\ W_{Point}^s = \text{reshape }\left(K^{s-1} W_p^s\right), \end{cases} \tag{4}$$

where $W_c^s \in \mathbb{R}^{C \times K}$, $W_p^s \in \mathbb{R}^{C \times L}$, $W_{Channel}^s \in \mathbb{R}^{L \times 1 \times K}$, $W_{Point}^s \in \mathbb{R}^{L \times L \times 1}$. Here, L represents the number of objects, K is the size of W_{Point}^s, and C is the feature dimension of segment kernels K^{s-1}.

We employ the text-conditioned parameters of convolution $W_{Channel}^s$ and W_{Point}^s for different purposes. W_{Point}^s is used to exchange information between different visual kernels, while $W_{Channel}^s$ is used for self-interaction of each visual kernel. This operation embeds semantic features into the visual kernels, enhancing the discriminability of each feature. Through cross-object interaction with W_{Point}^s, the visual kernels K_v^s engage in semantic-level interaction, reinforcing themselves through related objects and further clarifying instance relationships. Finally, a residual module (He et al. 2016) is used to preserve sufficient visual semantics:

$$K_v^s = W_{Point}^s * \left(W_{Channel}^s * K_v^s\right) + K_v^s. \tag{5}$$

3.3 Visual Context Vertification

In the previous section, we established preliminary relationships between objects using semantic priors and strengthened the visual features accordingly. However, relying solely on such semantic priors can easily introduce biases, a common phenomenon in deep learning. For instance, when presented with an image of "a person standing next to a horse", merely using text-based priors to link "person" and "horse" may lead the model to mistakenly assume that "the person is riding the horse", even though they do not have a strong relationship. To overcome such issues, we need to search for visual cues within the image to correct relationships acquired solely through semantics. This necessitates the integration of visual information to guide instance relationship modeling. Furthermore, incorporating visual information provides strong signals, such as attributes and spatial relationships, which facilitate the model's understanding of the scene. To achieve this, we propose the Visual Context Verification (VCV) module, which includes new geometric attributes for each object corresponding to the mask. In addition to the attribute features inherent in the visual features, we need to incorporate geometric information, such as size and location, into relationship modeling, as it has been proven to be crucial for relationship modeling (Herdade et al. 2019).

Obtaining Geometric Information The geometric information of each visual kernel corresponds to the geometric information of the instance mask it represents. First, we calculate the scale of the mask. The conventional approach would be to find the maximum differences in horizontal and vertical coordinates as the width and height, respectively. However, we found that the presence of outliers within the mask is particularly severe, causing the calculated width and height to significantly exceed expectations. To avoid this phenomenon, we innovatively adopt the following method to calculate the scale of the k-th mask:

$$H_n^s = \sum_{i=1}^{h} \text{Max}\left(M_{i1}^s, M_{i2}^s, \ldots M_{iw}^s\right), \tag{6}$$

$$W_n^s = \sum_{i=1}^{w} \text{Max}\left(M_{1i}^s, M_{2i}^s, \ldots M_{hi}^s\right), \tag{7}$$

where M^s is the mask of the n-th object, and h and w are the dimensions of the mask. H_n^s and W_n^s represent the height and width of the n-th instance. When outliers are present, the distances between outliers and instances are disregarded, allowing the calculation of the effective scale of the mask.

Various approaches can be employed to determine the location of the mask. In this paper, we use the centroid as its location. The centroid (X_n^s, Y_n^s) of the n-th object is defined as follows:

$$X_n^s = \frac{\iint_{M^\ell} x^n M^s\left(x^n, y^n\right) dx dy}{\sum_{x^n, y^n}^{w,h} M^s\left(x^n, y^n\right)}, \tag{8}$$

$$Y_n^s = \frac{\iint_{M^s} y^n M^s\left(x^n, y^n\right) dx dy}{\sum_{x^n, y^n}^{w,h} M^s\left(x^n, y^n\right)}, \tag{9}$$

where (x^n, y^n) is the coordinates of each point on the n -th mask. $M^s\left(x^n, y^n\right)$ is the mask value corresponding to point (x^n, y^n), h and w are the scale of M^s.

Geometric Relationship Reasoning Inspired by (Herdade et al. 2019), we transform the aforementioned geometric information into geometric relationships between objects and use them as prior knowledge to guide the inter-object relationships. A common approach to model visual object relationships is self-attention (Vaswani et al. 2017). The definition of an attention weight matrix based on appearance is as follows:

$$\omega^s = \frac{(K_v^s W_Q^s) \cdot (K_v^s W_K^s)^T}{\sqrt{d_k}}, \tag{10}$$

where W_Q and W_K are learned projection matrices, d_k is the features dimension of K^s, and ω is the attention weights for the appearance features.

In this paper, we need to incorporate size and position information as prior knowledge into attention modeling. For a given instance m and instance n, we calculate the displacement vector $\lambda(m, n)$, which represents their geometric relationship. We calculate $\lambda(m, n)$ based on m-th object's geometric features $(X_m^s, Y_m^s, W_m^s, H_m^s)$ and n-th object's geometric features $(X_n^s, Y_n^s, W_n^s, H_n^s)$ as:

$$\lambda(m,n) = \log\left(\frac{|X_m^s - X_n^s|}{W_m^s}, \frac{|Y_m^s - Y_n^s|}{H_m^s}, \frac{W_n^s}{W_m^s}, \frac{H_n^s}{H_m^s}\right), \tag{11}$$

where X_m^s, Y_m^s, W_m^s, and H_m^s correspond respectively to the centroid coordinates, width, and height of the m-th object. The geometric attention weights are then calculated as:

$$\omega_g^s = \mathrm{ReLU}\left(\mathrm{Emb}(\lambda) W_g^s\right), \tag{12}$$

where Emb$(\cdot)$ is a sinusoid function to calculate a high-dimensional embedding for the scalar λ. The combining geometry and appearance attention weights of the m-th and n-th objects $\omega_{ga}^s(m, n)$ are normalized as:

$$\omega_{ga}^s(m,n) = \frac{\omega_g^s(m,n)\exp\left(\omega^s(m,n)\right)}{\sum_{l=1}^{L} \omega_g^s(m,l)\exp\left(\omega^s(m,l)\right)}, \tag{13}$$

where ω^s are the appearance-based attention weights from Eq. 10, $\omega^s(m, n)$ represents the appearance-based attention weights between the n-th and m-th objects. ω_g^s are the new combined attention weights. $\omega_{ga}^s(m, n)$ represents the combining geometry and appearance attention weights between the n-th and m-th objectives.

We use ω_{ga}^s instead of the appearance attention weight matrix ω^s for self-attention operations and name it Geometric Aware Attention. We use Geometric Aware Attention to enhance visual kernels. The final segment kernels K^s are obtained by fusing the segment kernels K^{s-1} and visual kernels K_v^s using a residual structure:

$$K^s = \omega_{ga}^s \cdot (K_v^s W_K^s) W_2^s + K^{s-1} W_1^s. \tag{14}$$

Assuming a total of s iterations. After the last iteration, we use the final segmentation kernels K^s and visual features F_v to obtain masks M^s similar to Eq. 1 and Eq. 2.

3.4 Training loss

The employed loss function is a composite of Dice loss (Milletari, Navab, and Ahmadi 2016) and BCE loss (Milletari, Navab, and Ahmadi 2016). Specifically, the two types of losses are defined as follows:

$$\overline{\mathcal{L}}_{\text{bce}} = -\frac{1}{NHW} \sum_{n=1}^{N} \sum_{i=1}^{H\times W} \mathcal{L}_{\text{bce}}\left(M^{n,i}, Y^{n,i}\right), \tag{15}$$

where M is the generated masks and G is the ground truth, N is the number of masks, $H \times W$ is the number of points.

$$L_{Dice} = 1 - \frac{2|M \bigcap G|}{|M| + |G|}, \tag{16}$$

where M is the generated masks and G is the ground truth, the value of which all belongs to $\{0, 1\}$.

During the training, we use the summation of Dice loss (Milletari, Navab, and Ahmadi 2016) and BCE loss (Milletari, Navab, and Ahmadi 2016).

$$L = \lambda_1 \overline{\mathcal{L}}_{bce} + \lambda_2 L_{Dice}, \tag{17}$$

where λ_1 and λ_2 are the hyper-parameters. $\lambda_1 = 1$, $\lambda_2 = 1$.

4 Experiment

4.1 Datasets

We trained and evaluated our model on the Panoptic Narrative Grounding (PNG) dataset (González et al. 2021), which contains images and their corresponding narratives with pixel-level segmentation annotations for related phrases. Unlike datasets that have only one target corresponding to each short phrase, such as RefCOCO (Milletari, Navab, and Ahmadi 2016), the narratives in PNG often contain hundreds of words and more complex semantics. Each description consists of an average of 5.1 objects. In total, the PNG dataset comprises 133,103 training images and 8,380 test images, accompanied by 875,073 and 56,531 segmentation annotations, respectively.

4.2 Implementation Details

In our experiment, we employ the FPN (Lin et al. 2017) with ResNet101 (He et al. 2016) as the backbone, pre-trained on the Panoptic segmentation task using the MS COCO (Lin et al. 2014) dataset. We utilize the official implementation to ensure consistency with previous works. During training, the FPN parameters are frozen. Images are resized so that the short side is 800 pixels while maintaining the aspect ratio, and the long side is 1333 pixels. For language input, we use the BERT (Devlin et al. 2018) model to convert descriptive captions into tokens with 768-dimensional vectors. The maximum token length is set to 230. We employ the Adam optimizer with an initial learning rate of $1e - 4$, which is halved every two epochs after the tenth epoch. The learning rate for BERT is set to $1e-5$. The number of iteration update stages is set to 3. During inference, we average the masks of all tokens in each noun phrase to obtain the final results. All experiments are conducted on an A100 GPU with a batch size of 11.

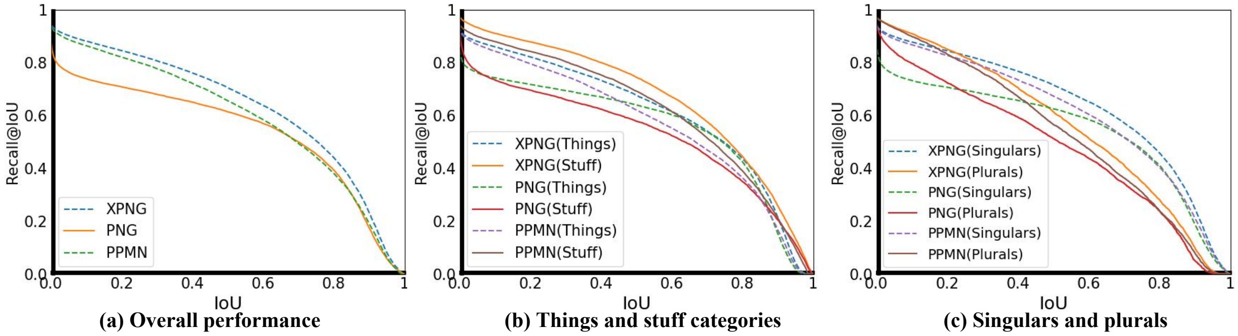


Figure 3: Average Recall Curve for our XPNG method performance (a) compared to the state-of-the-art methods, and (b) things and stuff categories, as well as (c) singular and plural noun phrases.

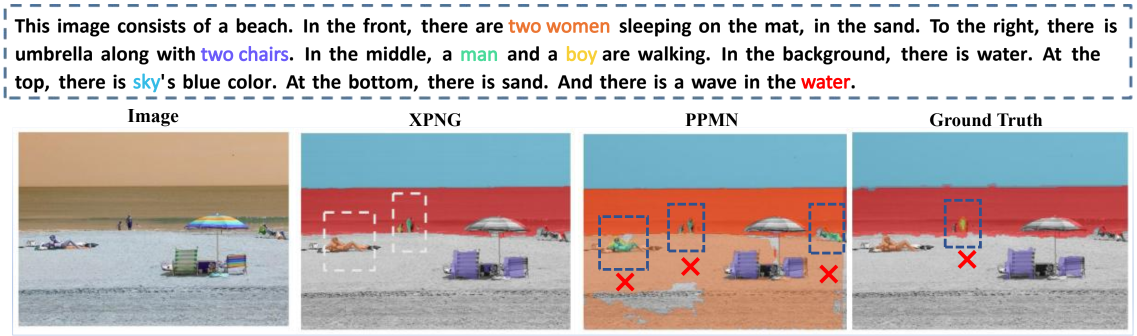

This image consists of a beach. In the front, there are two women sleeping on the mat, in the sand. To the right, there is umbrella along with two chairs. In the middle, a man and a boy are walking. In the background, there is water. At the top, there is sky's blue color. At the bottom, there is sand. And there is a wave in the water.

Figure 4: Visualizations of XPNG's predictions. We use the same color to mark the masks with their corresponding phrases. In particular, we use white dashed boxes to highlight the areas where XPNG performs well, and use blue dashed boxes to highlight the areas where Ground Truth and PPMN exhibit poor performance.

4.3 Comparison with State-of-the-Art Methods

To evaluate the performance of our XPNG, we conduct experiments on the PNG benchmark (González et al. 2021) as shown in Tab. 1. XPNG achieved a 3.9% improvement, which sets a new SOTA on the benchmark. The results of our study demonstrate the effectiveness of our one-stage approach XPNG. Specifically, when compared to the one-stage SOTA method, *i.e.*, PPMN (Ding et al. 2022), XPNG achieved significant improvements in average recall evaluation metrics for overall, things, stuff, singular, and plural segmentation by 3.9%, 3.9%, 3.7%, 4.0%, 2.4%, Fig. 3 illustrates their recall performance in detail. The proposed XPNG with SCC and VCV shows a more powerful ability in the grounding of multi phrases. At the same time, it can be observed that EPNG (Wang et al. 2023b) utilizes a lightweight encoder, which results in fewer parameters compared to PPMN and XPNG. However, this comes at the cost of a significant performance gap.

4.4 Ablation

In order to validate the potential benefits of our proposed SCC and VCV, we performed ablation studies on the PNG benchmark with various designs.

SCC kernels vs. other kernels To investigate the effect of the SCC, we conduct experiments to examine the perfor-

Method	Segmentation Average Recall					Params
	All	Thing	Stuff	Single	Plural	
MCN	54.2	48.6	61.4	56.6	38.8	**40.19M**
PNG	55.4	56.2	54.3	56.2	48.8	261.3M
EPNG	49.7	45.6	55.5	50.2	45.1	76.5M
PPMN	59.4	57.2	62.5	60.0	54.0	94.4M
XPNG[†]	62.3	59.6	66.0	63.0	**56.7**	95.57M
XPNG	**63.3**	**61.1**	**66.2**	**64.0**	56.4	95.57M

Table 1: The comparison of XPNG with the state-of-the-art methods. XPNG[†] results from frozen training of bert network parameters and FPN network parameters. XPNG results from FPN network parameter freezing and bert encoder participating in training.

Method	Segmentation Average Recall				
	Overall	Thing	Stuff	Single	Plural
Text Kernel	59.4	57.2	62.5	60.0	54.0
Visual Kernel	60.1	57.3	64.1	60.6	55.7
SCC Kernel	**62.3**	**59.6**	**66.0**	**63.0**	**56.7**

Table 2: The ablation study of the influence of semantic context on the performance.

mance of SCC kernels, visual kernels, and text kernels. SCC kernels represent kernels output by the SCC, visual kernels represent kernels without Pixel-wise and Channel-wise convolution, and text kernels correspond to kernels processed directly from the BERT output. In Tab. 2, we assess the impact of different types of segmentation kernels on the model. SCC yields a 2.9% and 2.2% improvement compared to text-only kernels and visual-only kernels, respectively. This can be attributed to the inherent differences between modalities and the lack of interaction between visual targets. The independent information of a single modality is insufficient for addressing the challenging PNG task. In contrast, SCC fuses cross-modal information at the object level, connecting the two modalities within related instances.

VCV	SA	Overall	Thing	Stuff	Single	Plural
✗	✗	58.2	55.1	65.2	58.8	52.5
✗	✓	61.3	58.5	65.3	62.0	55.5
✓	✗	**62.3**	**59.6**	**66.0**	**63.0**	**56.7**

Table 3: The ablation study of the influence of geometry guided relations on the performance.

Stages	Segmentation Average Recall				
	Overall	Thing	Stuff	Single	Plural
1	60.1	57.2	64.1	60.6	55.3
2	62.0	59.3	65.7	62.6	56.5
3	**62.3**	**59.6**	**66.0**	**63.0**	**56.7**
4	62.2	59.4	**66.0**	62.8	56.4

Table 4: The ablation the number of stages.

Dataset		Type	p@0.3	p@0.4	p@0.5
RefCOCO	testA	PPMN	25.7	19.1	13.3
		XPNG	**30.6**	**27.0**	**23.0**
	testB	PPMN	22.7	16.3	10.7
		XPNG	**33.9**	**27.3**	**20.7**
RefCOCO+	testA	PPMN	25.9	19.4	13.4
		XPNG	**27.1**	**23.3**	**19.9**
	testB	PPMN	24.6	18.1	12.2
		XPNG	**28.7**	**23.0**	**18.0**
RefCOCOg	test	PPMN	19.2	14.7	10.8
		XPNG	**25.0**	**21.0**	**17.3**

Table 5: Zero-shot results of XPNG on RES. XPNG is not trained with RES data. We average the IoU of every case as the mIoU.

With vs. without geometric guidance In Tab. 3, we conduct an evaluation of the influence of geometry-guided relations on the performance of the PNG task. The interaction is necessary, even if the traditional self-attention brings 3.1% improvement. Meanwhile, with more space information, VCV achieves better performance, which outperforms SA 1.0% and baseline 4.1% on "Overall" sets. From this, it can be concluded that the information exchange of the visual kernels is of great significance for understanding the objects' relationship and adding geometric relationships can better guide the establishment of objects' relationships.

The number of stages We employ cascaded iteration stages for XPNG. Hence, we investigate the impact of varying the number of stages on XPNG's performance in Tab. 4. It can be observed that as the number of stages increases, the performance of XPNG also improves. XPNG achieves its best performance with an average recall of 62.3% on "Overall" when the number of stages is set to 3. However, as the number of stages further increases to 4, more interactions result in overfitting, leading to a decline in performance. Consequently, we ultimately adopt a 3-stage configuration for the final model.

4.5 Zero-Shot Study for RES

Simultaneously, we evaluate our model's generalization capability by conducting zero-shot experiments on the datasets of the RES task, such as RefCOCO (Yu et al. 2016), RefCOCO+ (Yu et al. 2016), and RefCOCOg (Mao et al. 2016; Nagaraja, Morariu, and Davis 2016). We utilize the feature of the entire phrase as the text feature. The results are presented in Tab. 5. By comparing with the SOTA, PPMN, we observe that the zero-shot performance of XPNG is significantly enhanced.

4.6 Qualitative Analysis

Visualization In Fig. 4, We present the qualitative results of our proposed XPNG. We generated corresponding masks for each phrase marked in each text of each image and visualized them on a single image. Surprisingly, our proposed XPNG outperforms ground truth in terms of producing more accurate segmentation results. We demonstrate the visualization results of XPNG and PPMN in complex contexts and visual environments.

In Fig. 4, there are complex representations of multiple similar objects in the scene. The white dashed box on the right indicates the targets of "boy" and "man". XPNG infers that there might be someone near the seaside based on the information of "water" and "two chairs", and uses the geometric position relationship between "boy" and "man" to perform a clearer boundary segmentation for them. However, PPMN performs poorly in scenarios involving complex references due to the lack of interaction between objects.

5 Conclusion

In this paper, we propose a novel one-stage framework, XPNG, which effectively models object relationships through contextual semantic information and object geometry properties. Our approach includes a Semantic Context Convolution (SCC) module that models the relationships between nouns and aggregates visual features. This module provides rich semantic prior information, which enhances the model's ability to comprehend the various elements of a scene. Moreover, our Visual Context Verification (VCV) module combines the geometric information of target objects or regions to minimize bias and precisely focus on the correct targets. Empirical results demonstrate that XPNG surpasses current state-of-the-art methods by 3.9% in terms of performance.

Acknowledgments

This work was supported by National Key R&D Program of China (No.2022ZD0118201), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62072389), the National Natural Science Fund for Young Scholars of China (No. 62302411), China Postdoctoral Science Foundation (No. 2023M732948), the National Natural Science Foundation of China (No. 62072386, No. 62176222, No. 62176223, No. 62176226, No. 62072387, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001).

References

Alhaija, H. A.; Mustikovela, S. K.; Mescheder, L.; Geiger, A.; and Rother, C. 2017. Augmented reality meets deep learning for car instance segmentation in urban scenes. In *BMVC*, volume 1, 2.

Chen, W.; Hong, D.; Qi, Y.; Han, Z.; Wang, S.; Qing, L.; Huang, Q.; and Li, G. 2022. Multi-Attention Network for Compressed Video Referring Object Segmentation. In *ACM MM*, 4416–4425.

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

Cheng, Y.; Wang, R.; Yu, J.; Zhao, R.-W.; Zhang, Y.; and Feng, R. 2021. Exploring Logical Reasoning for Referring Expression Comprehension. In *ACM MM*, 5047–5055.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2021. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 16321–16330.

Ding, Z.; Ding, Z.-h.; Hui, T.; Huang, J.; Wei, X.; Wei, X.; and Liu, S. 2022. PPMN: Pixel-Phrase Matching Network for One-Stage Panoptic Narrative Grounding. In *ACM MM*, 5537–5546.

Fei, Z. 2022. Efficient Modeling of Future Context for Image Captioning. In *ACM MM*, 5026–5035.

Feng, G.; Hu, Z.; Zhang, L.; and Lu, H. 2021. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, 15506–15515.

Giannitrapani, R.; Trucco, A.; and Murino, V. 1999. Segmentation of underwater 3D acoustical images for augmented and virtual reality applications. In *Oceans' 99*, volume 1, 459–465. IEEE.

González, C.; Ayobi, N.; Hernández, I.; Hernández, J.; Pont-Tuset, J.; and Arbeláez, P. 2021. Panoptic narrative grounding. In *ICCV*, 1364–1373.

González, C.; Ayobi, N.; Hernández, I.; Pont-Tuset, J.; and Arbeláez, P. 2023. PiGLET: Pixel-level Grounding of Language Expressions with Transformers. *TPAMI*, 45: 12206–12221.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Herdade, S.; Kappeler, A.; Boakye, K.; and Soares, J. 2019. Image captioning: Transforming objects into words. *NeurIPS*, 32.

Hu, J.; Cao, L.; Lu, Y.; Zhang, S.; Wang, Y.; Li, K.; Huang, F.; Shao, L.; and Ji, R. 2021. Istr: End-to-end instance segmentation with transformers. *arXiv*.

Hu, J.; Huang, L.; Ren, T.; Zhang, S.; Ji, R.; and Cao, L. 2023. You Only Segment Once: Towards Real-Time Panoptic Segmentation. In *CVPR*, 17819–17829.

Hu, R.; Rohrbach, M.; and Darrell, T. 2016. Segmentation from natural language expressions. In *ECCV*, 108–124. Springer.

Huang, L.; Wang, H.; Zeng, J.; Zhang, S.; Cao, L.; Ji, R.; Yan, J.; and Li, H. 2023. Geometric-aware Pretraining for Vision-centric 3D Object Detection. *arXiv*.

Hui, T.; Ding, Z.; Huang, J.; Wei, X.; Wei, X.; Dai, J.; Han, J.; and Liu, S. 2023. Enriching phrases with coupled pixel and object contexts for panoptic narrative grounding. *arXiv*.

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

Jiao, Y.; Jie, Z.; Luo, W.; Chen, J.; Jiang, Y.-G.; Wei, X.; and Ma, L. 2021. Two-stage visual cues enhancement network for referring image segmentation. In *ACM MM*, 1331–1340.

Jing, C.; Wu, Y.; Pei, M.; Hu, Y.; Jia, Y.; and Wu, Q. 2020. Visual-semantic graph matching for visual grounding. In *ACM MM*, 4041–4050.

Kiran, B. R.; Sobh, I.; Talpaert, V.; Mannion, P.; Al Sallab, A. A.; Yogamani, S.; and Pérez, P. 2021. Deep reinforcement learning for autonomous driving: A survey. *TITS*, 23(6): 4909–4926.

Kirillov, A.; He, K.; Girshick, R.; Rother, C.; and Dollár, P. 2019. Panoptic segmentation. In *CVPR*, 9404–9413.

Li, L.; Bu, Y.; and Cai, Y. 2021. Bottom-Up and Bidirectional Alignment for Referring Expression Comprehension. In *ACM MM*, 5167–5175.

Li, M.; and Sigal, L. 2021. Referring transformer: A one-step approach to multi-task visual grounding. *NeurIPS*, 34: 19652–19664.

Li, R.; Xu, C.; Guo, Z.; Fan, B.; Zhang, R.; Liu, W.; Zhao, Y.; Gong, W.; and Wang, E. 2022a. AI-VQA: Visual Question Answering based on Agent Interaction with Interpretability. In *ACM MM*, 5274–5282.

Li, Z.; Wang, W.; Xie, E.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; Luo, P.; and Lu, T. 2022b. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *CVPR*, 1280–1289.

Liao, Y.; Zhang, A.; Chen, Z.; Hui, T.; and Liu, S. 2022. Progressive language-customized visual feature learning for one-stage visual grounding. *TIP*, 31: 4266–4277.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *CVPR*, 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 *ECCV*, 740–755. Springer.

Lin, Y.; Jin, X.-B.; Wang, Q.; and Huang, K. 2023. Context Does Matter: End-to-end Panoptic Narrative Grounding with Deformable Attention Refined Matching Network. *arXiv*.

Liu, S.; Hui, T.; Huang, S.; Wei, Y.; Li, B.; and Li, G. 2021. Cross-modal progressive comprehension for referring segmentation. *TPAMI*, 44(9): 4761–4775.

Luo, G.; Zhou, Y.; Ji, R.; Sun, X.; Su, J.; Lin, C.-W.; and Tian, Q. 2020a. Cascade grouped attention network for referring expression segmentation. In *ACM MM*, 1274–1282.

Luo, G.; Zhou, Y.; Sun, X.; Cao, L.; Wu, C.; Deng, C.; and Ji, R. 2020b. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, 10034–10043.

Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; and Ji, R. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM MM*, 638–647.

Ma, Y.; Zhang, X.; Sun, X.; Ji, J.; Wang, H.; Jiang, G.; Zhuang, W.; and Ji, R. 2023. X-Mesh: Towards Fast and Accurate Text-driven 3D Stylization via Dynamic Textual Guidance. In *ICCV*, 2749–2760.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 11–20.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 565–571. Ieee.

Moosavi, S.; Mahajan, P. D.; Parthasarathy, S.; Saunders-Chukwu, C.; and Ramnath, R. 2021. Driving style representation in convolutional recurrent neural network model of driver identification. *arXiv*.

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

Pont-Tuset, J.; Uijlings, J.; Changpinyo, S.; Soricut, R.; and Ferrari, V. 2020. Connecting vision and language with localized narratives. In *ECCV*, 647–664. Springer.

Suo, W.; Sun, M.; Wang, P.; and Wu, Q. 2021. Proposal-free one-stage referring expression via grid-word cross-attention. *arXiv*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Wang, H.; Ji, J.; Guo, T.; Yang, Y.; Zhou, Y.; Sun, X.; and Ji, R. 2023a. NICE: Improving Panoptic Narrative Detection and Segmentation with Cascading Collaborative Learning. *arXiv*.

Wang, H.; Ji, J.; Zhou, Y.; Wu, Y.; and Sun, X. 2023b. Towards Real-Time Panoptic Narrative Grounding by an End-to-End Grounding Network. *AAAI*, 37(2): 2528–2536.

Wang, H.; Zhu, Y.; Adam, H.; Yuille, A.; and Chen, L.-C. 2021. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, 5463–5474.

Wu, S.; Fei, H.; Ji, W.; and Chua, T.-S. 2023. Cross2StrA: Unpaired Cross-lingual Image Captioning with Cross-lingual Cross-modal Structure-pivoted Alignment. In *ACL*, 2593–2608.

Yang, D.; Ji, J.; Sun, X.; Wang, H.; Li, Y.; Ma, Y.; and Ji, R. 2023. Semi-Supervised Panoptic Narrative Grounding. In *ACM MM*, 7164–7174.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 18155–18165.

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

Zhang, W.; Pang, J.; Chen, K.; and Loy, C. C. 2021. K-net: Towards unified image segmentation. *NeurIPS*, 34: 10326–10338.

Zhao, Y.; Fei, H.; Ji, W.; Wei, J.; Zhang, M.; Zhang, M.; and Chua, T.-S. 2023. Generating Visual Spatial Description via Holistic 3D Scene Understanding. In *ACL*, 7960–7977.