

# Panoptic SegFormer: Delving Deeper into Panoptic Segmentation with Transformers

Zhiqi Li<sup>1</sup>, Wenhui Wang<sup>2</sup>, Enze Xie<sup>3</sup>, Zhiding Yu<sup>4</sup>,  
Anima Anandkumar<sup>4,5</sup>, Jose M. Alvarez<sup>4</sup>, Ping Luo<sup>3</sup>, Tong Lu<sup>1</sup>

<sup>1</sup>Nanjing University <sup>2</sup>Shanghai AI Laboratory <sup>3</sup>The University of Hong Kong <sup>4</sup>NVIDIA <sup>5</sup>Caltech

lzq@smail.nju.edu.cn wangwenhai@pjlab.org.cn xieenze@hku.hk zhidingy@nvidia.com  
aanandkumar@nvidia.com josea@nvidia.com pluo@cs.hku.hk lutong@nju.edu.cn

## Abstract

Panoptic segmentation involves a combination of joint semantic segmentation and instance segmentation, where image contents are divided into two types: things and stuff. We present Panoptic SegFormer, a general framework for panoptic segmentation with transformers. It contains three innovative components: an efficient deeply-supervised mask decoder, a query decoupling strategy, and an improved post-processing method. We also use Deformable DETR to efficiently process multi-scale features, which is a fast and efficient version of DETR. Specifically, we supervise the attention modules in the mask decoder in a layer-wise manner. This deep supervision strategy lets the attention modules quickly focus on meaningful semantic regions. It improves performance and reduces the number of required training epochs by half compared to Deformable DETR. Our query decoupling strategy decouples the responsibilities of the query set and avoids mutual interference between things and stuff. In addition, our post-processing strategy improves performance without additional costs by jointly considering classification and segmentation qualities to resolve conflicting mask overlaps. Our approach increases the accuracy 6.2% PQ over the baseline DETR model. Panoptic SegFormer achieves state-of-the-art results on COCO test-dev with 56.2% PQ. It also shows stronger zero-shot robustness over existing methods. The code is released at <https://github.com/zhiqi-li/Panoptic-SegFormer>.

## 1. Introduction

Semantic segmentation and instance segmentation are two important and related vision tasks. Their underlying connections recently motivated panoptic segmentation as a unification of both the tasks [6]. In panoptic segmentation, image contents are divided into two types: things and stuff.

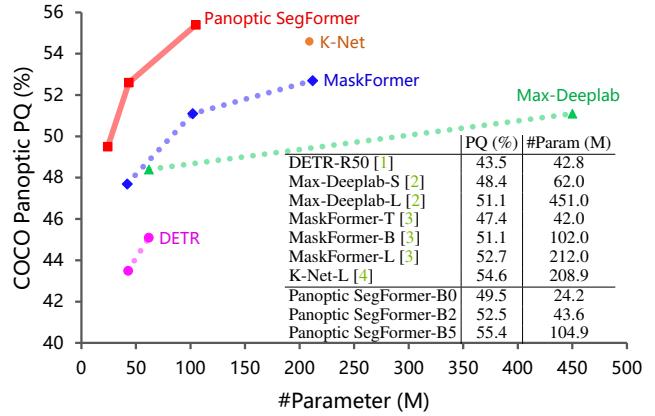
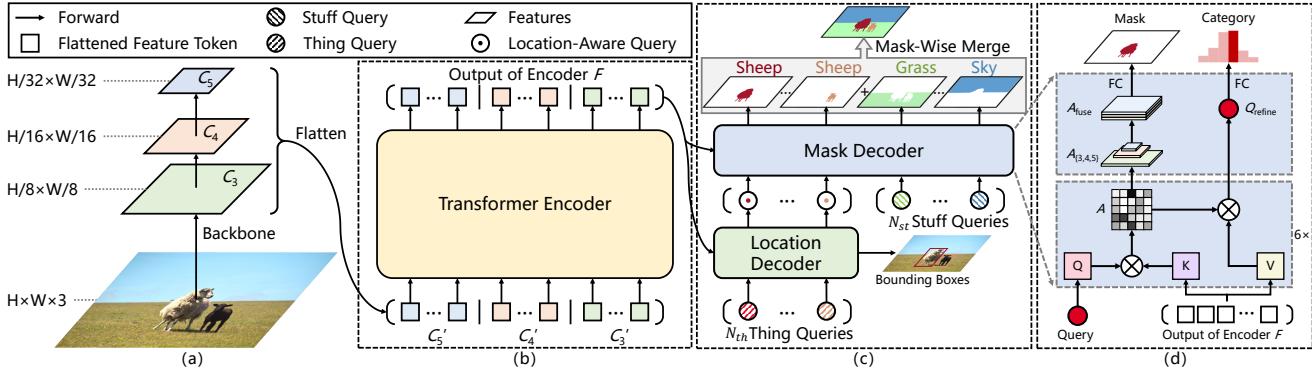


Figure 1. Comparison to the prior arts in panoptic segmentation methods on the COCO val2017 split. Panoptic SegFormer models outperform the other counterparts among different models. Panoptic SegFormer (PVTv2-B5 [5]) achieves 55.4% PQ, surpassing previous methods with significantly fewer parameters.

Things refer to countable instances (e.g., person, car) and each instance has a unique id to distinguish it from the other instances. Stuff refers to the amorphous and uncountable regions (e.g., sky, grassland) and has no instance id [6].

Recent works [1-3] attempt to employ transformers to handle both things and stuff through a query set. For example, DETR [1] simplifies the workflow of panoptic segmentation by adding a panoptic head on top of an end-to-end object detector. Unlike previous methods [6,7], DETR does not require additional handcrafted pipelines [8, 9]. While being simple, DETR also causes some issues: (1) It requires a lengthy training process to converge; (2) Because the computational complexity of self-attention is squared with the length of the input sequence, the feature resolution of DETR is limited. So that it uses an FPN-style [1, 10] panoptic head to generate masks, which always suffer low-fidelity boundaries; (3) It handles things and stuff equally, yet representing them with bounding boxes, which may be suboptimal for stuff [2, 3]. Although DETR achieves excellent perfor-



**Figure 2. Overview of Panoptic SegFormer.** Panoptic SegFormer is composed of backbone, encoder, and decoder. The backbone and the encoder output and refine multi-scale features. Inputs of the location decoder are  $N_{th}$  thing queries and the multi-scale features. We feed  $N_{th}$  thing queries from the location decoder and  $N_{st}$  stuff queries to the mask decoder. The location decoder aims to learn reference points of queries, and the mask decoder predicts the final category and mask. Details of the decoder will be introduced below. We use a mask-wise merging method instead of the commonly used pixel-wise argmax method to perform inference.

mance on the object detection task, its superiority on panoptic segmentation has not been well demonstrated. In order to overcome the defects of DETR on panoptic segmentation, we propose a series of novel and effective strategies that improve the performance of transformer-based panoptic segmentation models by a large margin.

**Our approach.** In this work, we propose Panoptic SegFormer, a concise and effective framework for panoptic segmentation with transformers. Our framework design is motivated by the following observations: 1) Deep supervision matters in learning high-quality discriminative attention representations in the mask decoder. 2) Treating things and stuff with the same recipe [1] is suboptimal due to the different properties between things and stuff [6]. 3) Commonly used post-processing such as pixel-wise argmax [1–3] tends to generate false-positive results due to extreme anomalies. We overcome these challenges in Panoptic SegFormer framework as follows:

- We propose a mask decoder that utilizes multi-scale attention maps to generate high-fidelity masks. The mask decoder is deeply-supervised, promoting discriminative attention representations in the intermediate layers with better mask qualities and faster convergence.
- We propose a query decoupling strategy that decomposes the query set into a thing query set to match things via bipartite matching and another stuff query set to process stuff with class-fixed assign. This strategy avoids mutual interference between things and stuff within each query and significantly improves the qualities of stuff segmentation. Kindly refer to Sec. 3.3.1 and Fig. 3 for more details.
- We propose an improved post-processing method to generate results in panoptic format. Besides being more efficient than the widely used pixel-wise argmax method, our method contains a mask-wise merging strategy that considers both classification probability and predicted mask

qualities. Our post-processing method alone renders a 1.3% PQ improvement to DETR [1].

We conduct extensive experiments on COCO [11] dataset. As shown in Fig. 1, Panoptic SegFormer significantly surpasses prior arts such as MaskFormer [3] and K-Net [4] with much fewer parameters. With deformable attention [12] and our deeply-supervised mask decoder, our method requires much fewer training epochs than previous transformer-based methods (24 vs. 300+) [1, 3]. In addition, our approach also achieves competitive performance with current methods [13, 14] on the instance segmentation task.

## 2. Related Work

**Panoptic Segmentation.** Panoptic segmentation becomes a popular task for holistic scene understanding [6, 15–17]. The panoptic segmentation literature mainly treats this problem as a joint task of instance segmentation and semantic segmentation where things and stuff are handled separately [18, 19]. Kirillov *et al.* [6] proposed the concept of and benchmark of panoptic segmentation together with a baseline that directly combines the outputs of individual instance segmentation and semantic segmentation models. Since then, models such as Panoptic FPN [7], UPSNet [9] and AUNet [20] have improved the accuracy and reduced the computational overhead by combining instance segmentation and semantic segmentation into a single model. However, these methods approximate the target task by solving the surrogate sub-tasks, therefore introducing undesired model complexities and suboptimal performance.

Recently, efforts have been made to unify the framework of panoptic segmentation. Li *et al.* [21] proposed Panoptic FCN where the panoptic segmentation pipeline is simplified with a “top-down meets bottom-up” two-branch design similar to CondInst [22]. In their work, things and stuff are jointly modeled by an object/region-level kernel

branch and an image-level feature branch. Several recent works represent things and stuff as queries and perform end-to-end panoptic segmentation via transformers. DETR [1] predicts the bounding boxes of things and stuff and combines the attention maps of the transformer decoder and the feature maps of ResNet [23] to perform panoptic segmentation. Max-Deeplab [2] directly predicts object categories and masks through a dual-path transformer regardless of the category being things or stuff. On top of DETR, MaskFormer [3] used an additional pixel decoder to refine high spatial resolution features and generated the masks by multiplying queries and features from the pixel decoder. Due to the computational complexity of self attention [24], both DETR and MaskFormer use feature maps with limited spatial resolutions for panoptic segmentation, which hurts the performance and requires combining additional high-resolution feature maps in final mask prediction. Unlike the methods mentioned above, our query decoupling strategy deals with things and stuff with separate query sets. Although thing and stuff queries are designed for different targets, they are processed by the mask decoder with the same workflow. Prediction results of these queries are in the same format so that we can process them in an equal manner during the post-processing procedure. One concurrent work [4] employs a similar line of thinking to use dynamic kernels to perform instance and semantic segmentation, and it aims to utilize unified kernels to handle various segmentation tasks. In contrast to it, we aim to delve deeper into the transformer-based panoptic segmentation. Due to the different nature of various tasks, whether a unified pipeline is suitable for these tasks is still an open problem. In this work, we utilize an additional location decoder to assist things to learn location clues and get better results.

**End-to-end Object Detection.** The recent popular end-to-end object detection frameworks have inspired many other related works [13, 25]. DETR [1] is arguably the most representative end-to-end object detector among these methods. DETR models the object detection task as a dictionary lookup problem with learnable queries and employs an encoder-decoder transformer to predict bounding boxes without extra post-processing. DETR greatly simplifies the conventional detection framework and removes many hand-crafted components such as Non-Maximum Suppression (NMS) [26, 27] and anchors [27]. Zhu *et al.* [12] proposed Deformable DETR, which further reduces the memory and computational cost through deformable attention layers. In this work, we adopt deformable attention [12] for the improved efficiency and convergence over DETR [1].

### 3. Methods

#### 3.1. Overall Architecture

As illustrated in Fig. 2, Panoptic SegFormer consists of three key modules: transformer encoder, location decoder,

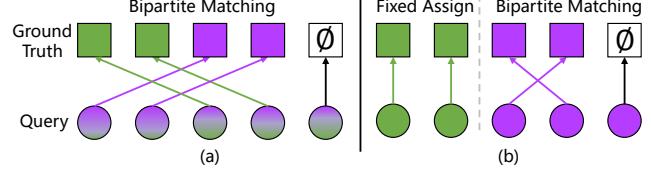


Figure 3. (a) Methods [1–3] adopt one query set to match things (purple squares) and stuff (green squares) jointly. (b) We use one thing query set (purple circles) to target things through bipartite matching and one stuff query set (green circles) to predict stuff by a class-fixed assign strategy.  $\emptyset$  is assigned to not-matched queries.

and mask decoder, where (1) the transformer encoder is applied to refine the multi-scale feature maps given by the backbone, (2) the location decoder is designed to capturing location clues of things, and (3) the mask decoder is for final classification and segmentation.

Our architecture feeds an input image  $X \in \mathbb{R}^{H \times W \times 3}$  to the backbone network, and obtains the feature maps  $C_3$ ,  $C_4$ , and  $C_5$  from the last three stages, of which the resolutions are  $1/8$ ,  $1/16$  and  $1/32$  compared to the input image, respectively. We project the three feature maps to the ones with 256 channels by a fully-connected (FC) layer, and flatten them into feature tokens  $C'_3$ ,  $C'_4$ , and  $C'_5$ . Here, we define  $L_i$  as  $\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}}$ , and the shapes of  $C'_3$ ,  $C'_4$ , and  $C'_5$  are  $L_1 \times 256$ ,  $L_2 \times 256$ , and  $L_3 \times 256$ , respectively. Next, using the concatenated feature tokens as input, the transformer encoder outputs the refined features of size  $(L_1 + L_2 + L_3) \times 256$ . After that, we use  $N_{\text{th}}$  and  $N_{\text{st}}$  randomly initialized things and stuff queries to describe things and stuff separately. Location decoder refines  $N_{\text{th}}$  thing queries by detecting the bounding boxes of things to capture location information. The mask decoder then takes both things and stuff queries as input and predicts mask and category at each layer.

During inference, we adopt a mask-wise merging strategy to convert the predicted masks from final mask decoder layer into the panoptic segmentation results, which will be introduced in detail in Sec. 3.5.

#### 3.2. Transformer Encoder

High-resolution and the multi-scale features maps are important for the segmentation tasks [7, 21, 28]. Since the high computational cost of self-attention layer, previous transformer-based methods [1, 3] can only process low-resolution feature maps (*e.g.*, ResNet  $C_5$ ) in their encoders, which limits the segmentation performance. Different from these methods, we employ the deformable attention [12] to implement our transformer encoder. Due to the low computational complexity of the deformable attention, our encoder can refine and involve positional encoding [24] to high-resolution and multi-scale feature maps  $F$ .

#### 3.3. Decoder

In this section, we introduce our query decoupling strategy firstly, and then we will explain the details of our loca-

tion decoder and mask decoder.

### 3.3.1 Query Decoupling Strategy

We argue that using one query set to handle both things and stuff equally is suboptimal. Since there are many different properties between them, things and stuff is likely to interfere with each other and hurt the model performance, especially for PQ<sup>st</sup>. To prevent things and stuff from interfering with each other, we apply a query decoupling strategy in Panoptic SegFormer, as shown in Fig. 3. Specifically,  $N_{\text{th}}$  thing queries are used to predict things results, and  $N_{\text{st}}$  stuff queries target stuff only. Using this form, things and stuff queries can share the same pipeline since they are in the same format. We can also customize private workflow for things or stuff according to the characteristics of different tasks. In this work, we use an additional location decoder to detect individual instances with thing queries, and this will assist in distinguishing between different instances [6]. Mask decoder accepts both thing queries and stuff queries and generates the final masks and categories. Note that, for thing queries, ground truths are assigned by bipartite matching strategy. For stuff, We use a class-fixed assign strategy, and each stuff query corresponds to one stuff category.

Thing and stuff queries will output results in the same format, and we handle these results with a uniform post-processing method.

### 3.3.2 Location Decoder

Location information plays an important role in distinguishing things with different instance ids in the panoptic segmentation task [22, 28, 29]. Inspired by this, we employ a location decoder to introduce the location information of things into the learnable queries. Specifically, given  $N_{\text{th}}$  randomly initialized thing queries and the refined feature tokens generated by transformer encoder, the decoder will output  $N_{\text{th}}$  location-aware queries.

In the training phase, we apply an auxiliary MLP head on top of location-aware queries to predict the bounding boxes and categories of the target object. We supervise the prediction results with a detection loss  $\mathcal{L}_{\text{det}}$ . The MLP head is an auxiliary branch, which can be discarded during the inference phase. The location decoder follows Deformable DETR [12]. Notably, the location decoder can learn location information by predicting the mass centers of masks instead of bounding boxes. This box-free model can still achieve comparable results to our box-based model.

### 3.3.3 Mask Decoder

As shown in Fig. 2 (d), the mask decoder is proposed to predict the categories and masks according to the given queries. The queries  $Q$  of the mask decoder are the location-aware thing queries from the location decoder or the class-fixed stuff queries. The keys  $K$  and values  $V$  of the mask decoder are projected from the refined feature tokens  $F$

from the transformer encoder. We first pass thing queries through the mask decoder, and then fetch the attention map  $A \in \mathbb{R}^{N \times h \times (L_1+L_2+L_3)}$  and the refined query  $Q_{\text{refine}} \in \mathbb{R}^{N \times 256}$  from each decoder layer, where  $N = N_{\text{th}} + N_{\text{st}}$  is the whole query number,  $h$  is the number of attention heads, and  $L_1+L_2+L_3$  is the length of feature tokens  $F$ .

Similar to methods [1, 2], we directly perform classification through a FC layer on top of the refined query  $Q_{\text{refine}}$  from each decoder layer. Each thing query needs to predict probabilities over all thing categories. Stuff query only predicts the probability of its corresponding stuff category.

At the same time, to predict the masks, we first split and reshape the attention maps  $A$  into attention maps  $A_3$ ,  $A_4$ , and  $A_5$ , which have the same spatial resolution as  $C_3$ ,  $C_4$ , and  $C_5$ . This process can be formulated as:

$$(A_3, A_4, A_5) = \text{Split}(A), \quad A_i \in \mathbb{R}^{\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}} \times h}, \quad (1)$$

where  $\text{Split}(\cdot)$  denotes the split and reshaping operation. After that, as illustrated in Eq. (2), we upsample these attention maps to the resolution of  $H/8 \times W/8$  and concatenate them along the channel dimension,

$$A_{\text{fused}} = \text{Concat}(A_1, \text{Up}_{\times 2}(A_2), \text{Up}_{\times 4}(A_3)), \quad (2)$$

where  $\text{Up}_{\times 2}(\cdot)$  and  $\text{Up}_{\times 4}(\cdot)$  mean the 2 times and 4 times bilinear interpolation operations, respectively.  $\text{Concat}(\cdot)$  is the concatenation operation. Finally, based on the fused attention maps  $A_{\text{fused}}$ , we predict the binary mask through a  $1 \times 1$  convolution.

Previous literature [12] argues that the reason for slow convergence of DETR is that attention modules equally pay attention to all the pixels in the feature maps, and learning to focus on sparse meaningful locations requires plenty of effort. We use two key designs to solve this problem in our mask decoder: (1) Using an ultra-light FC head to generate masks from the attention maps, ensuring attention modules can be guided by ground truth mask to learn where to focus on. This FC head only contains 200 parameters, which ensures the semantic information of attention maps is highly related to the mask. Intuitively, the ground truth mask is exactly the meaningful region on which we expect the attention module to focus. (2) We employ deep supervision in the mask decoder. Attention maps of each layer will be supervised by the mask, the attention module can capture meaningful information in the earlier stage. This can highly accelerate the learning process of attention modules.

### 3.4. Loss Function

During training, our overall loss function of Panoptic SegFormer can be written as:

$$\mathcal{L} = \lambda_{\text{things}} \mathcal{L}_{\text{things}} + \lambda_{\text{stuff}} \mathcal{L}_{\text{stuff}}, \quad (3)$$

where  $\mathcal{L}_{\text{things}}$  and  $\mathcal{L}_{\text{stuff}}$  are loss for things and stuff, separately.  $\lambda_{\text{things}}$  and  $\lambda_{\text{stuff}}$  are hyperparameters.

**Things Loss.** Following common practices [1, 30], we search the best bipartite matching between the prediction set and the ground truth set. Specifically, we utilize Hungarian algorithm [31] to search for the permutation with the minimum matching cost, which is the sum of the classification loss  $\mathcal{L}_{\text{cls}}$ , detection loss  $\mathcal{L}_{\text{det}}$  and the segmentation loss  $\mathcal{L}_{\text{seg}}$ . The overall loss function for the thing categories is accordingly defined as follows:

$$\mathcal{L}_{\text{things}} = \lambda_{\text{det}} \mathcal{L}_{\text{det}} + \sum_i^{D_m} (\lambda_{\text{cls}} \mathcal{L}_{\text{cls}}^i + \lambda_{\text{seg}} \mathcal{L}_{\text{seg}}^i), \quad (4)$$

where  $\lambda_{\text{cls}}$ ,  $\lambda_{\text{seg}}$ , and  $\lambda_{\text{loc}}$  are the weights to balance three losses.  $D_m$  is the number of layers in the mask decoder.  $\mathcal{L}_{\text{cls}}^i$  is the classification loss that is implemented by Focal loss [27], and  $\mathcal{L}_{\text{seg}}^i$  is the segmentation loss implemented by Dice loss [32].  $\mathcal{L}_{\text{det}}$  is the loss of Deformable DETR that used to perform detection.

**Stuff Loss.** We use a fixed matching strategy for stuff. Thus there is a one-to-one mapping between stuff queries and stuff categories. The loss for the stuff categories is similarly defined as:

$$\mathcal{L}_{\text{stuff}} = \sum_i^{D_m} (\lambda_{\text{cls}} \mathcal{L}_{\text{cls}}^i + \lambda_{\text{seg}} \mathcal{L}_{\text{seg}}^i), \quad (5)$$

where  $\mathcal{L}_{\text{cls}}^i$  and  $\mathcal{L}_{\text{seg}}^i$  are the same as those in Eq. (4).

### 3.5. Mask-Wise Merging Inference

Panoptic Segmentation requires each pixel to be assigned a category label (or void) and instance id (ignored for stuff) [6]. One challenge of panoptic segmentation is that it requires generating non-overlap results. Recent methods [1–3] directly use pixel-wise argmax to determine the attribution of each pixel, and this can solve the overlap problem naturally. Although pixel-wise argmax strategy is simple and effective, we observe that it consistently produces false-positive results due to the abnormal pixel values.

Unlike pixel-wise argmax resolves conflicts on each pixel, we propose the mask-wise merging strategy by resolving the conflicts among predicted masks. Specifically, we use the confidence scores of masks to determine the attribution of the overlap region. Inspired by previous NMS methods [28], the confidence scores take into both classification probability and predicted mask qualities. The confidence score of the  $i$ -th result can be formulated as:

$$s_i = p_i^\alpha \times \text{average}(\mathbb{1}_{\{m_i[h,w]>0.5\}} m_i[h,w])^\beta, \quad (6)$$

where  $p_i$  is the most likely class probability of  $i$ -th result.  $m_i[h,w]$  is the mask logit at pixel  $[h,w]$ ,  $\alpha, \beta$  are used to balance the weight of classification probability and segmentation qualities.

As illustrated in Algorithm 1, mask-wise merging strategy takes  $c$ ,  $s$ , and  $m$  as input, denoting the predicted categories, confidence scores, and segmentation masks, respectively. It outputs a semantic mask  $\text{SemMsk}$  and an instance

---

#### Algorithm 1: Mask-Wise Merging

---

```

def MaskWiseMergeing(c, s, m):
    # category c ∈ ℝN
    # confidence score s ∈ ℝN
    # mask m ∈ ℝN×H×W
    SemMsk = np.zeros(H, W)
    IdMsk = np.zeros(H, W)
    order = np.argsort(-s)
    id = 0
    for i in order:
        m_i = m[i]>0.5 & (SemMsk==0)
        if s[i]< tcnf or  $\frac{m_i}{m[i]>0.5} < t_{keep}$ :
            continue
        SemMsk[m_i] = c[i]
        IdMsk[m_i] = id
        id += 1
    return SemMsk, IdMsk

```

---

id mask  $\text{IdMsk}$ , to assign a category label and an instance id to each pixel. Specifically,  $\text{SemMsk}$  and  $\text{IdMsk}$  are first initialized by zeros. Then, we sort prediction results in descending order of confidence score and fill the sorted predicted masks into  $\text{SemMsk}$  and  $\text{IdMsk}$  in order. Then we discard the results with confidence scores below  $t_{\text{cls}}$  and remove the overlaps with lower confidence scores. Only remained non-overlap part with a sufficient fraction  $t_{\text{keep}}$  to origin mask will be kept. Finally, the category label and unique id of each mask are added to generate non-overlap panoptic format results.

## 4. Experiments

We evaluate Panoptic SegFormer on COCO [11] and ADE20K dataset [33], comparing it with several state-of-the-art methods. We provide the main results of panoptic segmentation and instance segmentation. We also conduct detailed ablation studies to verify the effects of each module. Please refer to Appendix for implementation details.

### 4.1. Dataset

We perform experiments on COCO 2017 datasets [11] without external data. The COCO dataset contains 118K training images and 5k validation images, and it contains 80 things and 53 stuff. We further demonstrate the generality of our model on the ADE20K dataset [33], which contains 100 things and 50 stuff.

### 4.2. Main Results

**Panoptic segmentation.** We conduct experiments on COCO val set and test-dev set. In Tab. 1 and Tab. 2, we report our main results, comparing with other state-of-the-art methods. Panoptic SegFormer attains 49.6% PQ on COCO val with ResNet-50 as the backbone and single-scale input, and it surpasses previous methods K-Net [4]

Method	Backbone	Epochs	PQ	PQ <sup>th</sup>	PQ <sup>st</sup>	#P	#F
Panoptic FPN [7]	R50	36	41.5	48.5	31.1	-	-
SOLov2 [28]	R50	36	42.1	49.6	30.7	-	-
DETR [1]	R50	325	43.4	48.2	36.3	42.9	248
Panoptic FCN [21]	R50	36	43.6	49.3	35.0	37.0	244
K-Net [4]	R50	36	47.1	51.7	40.3	-	-
MaskFormer [3]	R50	300	46.5	51.0	39.8	45.0	181
Panoptic SegFormer	R50	12	48.0	52.3	41.5	51.0	214
Panoptic SegFormer	R50	24	<b>49.6</b>	<b>54.4</b>	<b>42.4</b>	51.0	214
DETR [1]	R101	325	45.1	50.5	37.0	61.8	306
Max-Deeplab-S [2]	Max-S [2]	54	48.4	53.0	41.5	61.9	162
MaskFormer [3]	R101	300	47.6	52.5	40.3	64.0	248
Panoptic SegFormer	R101	24	<b>50.6</b>	<b>55.5</b>	<b>43.2</b>	69.9	286
Max-Deeplab-L [2]	Max-L [2]	54	51.1	57.0	42.2	451.0	1846
Panoptic FCN [36]	Swin-L <sup>†</sup>	36	51.8	58.6	41.6	-	-
MaskFormer [3]	Swin-L <sup>†</sup>	300	52.7	58.5	44.0	212.0	792
K-Net [4]	Swin-L <sup>†</sup>	36	54.6	60.2	46.0	208.9	-
Panoptic SegFormer	Swin-L <sup>†</sup>	24	<b>55.8</b>	<b>61.7</b>	<b>46.9</b>	221.4	816
Panoptic SegFormer	PVTv2-B5 <sup>†</sup>	24	55.4	61.2	46.6	104.9	349

Table 1. Experiments on COCO val set. #P and #F indicate number of parameters (M) and number of FLOPs (G). Panoptic SegFormer (R50) achieves 49.6% PQ on COCO val, surpassing previous methods such as DETR [1] and MaskFormer [3] over 6.2% PQ and 3.1% PQ respectively. <sup>†</sup> notes that backbones are pre-trained on ImageNet-22K.

Method	Backbone	PQ	PQ <sup>th</sup>	PQ <sup>st</sup>	SQ	RQ
Max-Deeplab-L [2]	Max-L [2]	51.3	57.2	42.4	82.5	61.3
Innovation [35]	ensemble	53.5	61.8	41.1	<b>83.4</b>	63.3
MaskFormer [3]	Swin-L <sup>†</sup>	53.3	59.1	44.5	82.0	64.1
K-Net [4]	Swin-L <sup>†</sup>	55.2	61.2	46.2	82.4	66.1
Panoptic SegFormer	R50	50.2	55.3	42.4	81.9	60.4
Panoptic SegFormer	R101	50.9	56.2	43.0	82.0	61.2
Panoptic SegFormer	Swin-L <sup>†</sup>	<b>56.2</b>	<b>62.3</b>	<b>47.0</b>	82.8	<b>67.1</b>
Panoptic SegFormer	PVTv2-B5 <sup>†</sup>	55.8	61.9	46.5	83.0	66.5

Table 2. Experiments on COCO test-dev set. <sup>†</sup> notes that backbones are pre-trained on ImageNet-22K.

Method	Backbone	PQ	PQ <sup>th</sup>	PQ <sup>st</sup>	SQ	RQ
BGRNet [37]	R50	31.8	-	-	-	-
Auto-Panoptic [38]	ShuffleNetV2 [39]	32.4	-	-	-	-
MaskFormer [3]	R50	34.7	32.2	<b>39.7</b>	76.7	42.8
MaskFormer [3]	R101	35.7	34.5	38.0	77.4	43.8
Panoptic SegFormer	R50	<b>36.4</b>	<b>35.3</b>	38.6	<b>78.0</b>	<b>44.9</b>

Table 3. Panoptic segmentation results on ADE20K val set.

and DETR [1] over 2.5% PQ and 6.2% PQ, respectively. Except for the remarkable performance, the training of Panoptic SegFormer is efficient. Under 1× training strategy (12 epochs), Panoptic SegFormer (R50) achieves 48.0% PQ that outperforms MaskFormer [3] that training 300 epochs by 1.5% PQ. Enhanced by vision transformer backbone Swin-L [34], Panoptic SegFormer attains a new record of 56.2% PQ on COCO test-dev without bells and whistles, surpassing MaskFormer [3] over 2.9% PQ. Our method even surpasses the previous competition-level method Innovation [35] over 2.7 % PQ. We also obtain comparable performance by employing PVTv2-B5 [5], while the model parameters and FLOPs are reduced significantly compared to Swin-L. Panoptic SegFormer also outperforms MaskFormer by 1.7% PQ on ADE20K dataset [33], see Tab. 3.

Method	Backbone	AP <sup>seg</sup>	AP <sup>seg</sup> <sub>S</sub>	AP <sup>seg</sup> <sub>M</sub>	AP <sup>seg</sup> <sub>L</sub>
Mask R-CNN [40]	R50	37.5	21.1	39.6	48.3
SOLov2 [28]	R50	38.8	16.5	41.7	56.2
K-Net [4]	R50	38.6	19.1	42.0	<b>57.7</b>
SOLQ [25]	R50	39.7	21.5	42.5	53.1
HTC [14]	R50	39.7	22.6	42.2	50.6
QueryInst [13]	R50	40.6	<b>23.4</b>	42.5	52.8
Ours (w/o crop)	R50	40.4	21.1	43.8	54.7
Ours (w/ crop)	R50	<b>41.7</b>	21.9	<b>45.3</b>	56.3

Table 4. Instance segmentation on COCO test-dev set.

	Epochs	PQ	#Params	FLOPs	FPS
baseline (DETR [1])	325	43.4	42.9M	247.5G	4.9
+ mask-wise merging	325	44.7	42.9M	247.5G	6.1
++ ms deformable attention [12]	50	47.3	44.9M	618.7G	2.7
+++ mask decoder	24	48.5	51.0M	214.8G	7.8
++++ query decoupling	24	49.6	51.0M	214.2G	7.8

Table 5. We increase the panoptic segmentation performance of DETR [1] (R50 [23]) from 43.4% PQ to 49.6% PQ with fewer training epochs, less computation cost, and faster inference speed.

**Instance segmentation.** Panoptic SegFormer can be converted to an instance segmentation model by just discarding stuff queries. In Tab. 4, we report our instance segmentation results on COCO test-dev set. We achieve results comparable to the current state-of-the-art methods such as QueryInst [13] and HTC [14], and 1.8 AP higher than K-Net [4]. Using random crops during training boosts the AP by 1.3 percentage points.

### 4.3. Ablation Studies

First, we show the effect of each module in Tab. 5. Compared to baseline DETR, our model achieves better performance, faster inference speed and significantly reduces the training epochs. We use Panoptic SegFormer (R50) to perform ablation experiments by default.

#Layer	PQ	PQ <sup>th</sup>	PQ <sup>st</sup>
0	47.0	50.0	42.5
1	47.7	51.1	42.5
2	48.1	51.8	42.5
6*	49.2	53.5	42.6
6	49.6	54.4	42.4

Table 6. Ablate location decoder.

**Effect of Location Decoder.** Location decoder assists queries to capture the location information of things. Tab. 6 shows the results with varying the number of layers in the location decoder. With fewer location decoder layers, our model performs worse on things, and it demonstrates that learning location clues through the location decoder is beneficial to the model to handle things better. \* notes we predict mass centers rather than bounding boxes in our location decoder, and this box-free model achieves comparable results (49.2% PQ vs. 49.6% PQ).

**Mask-wise Merging.** As shows in Tab. 7, we compare our mask-wise merging strategy against pixel-wise argmax strategy on various models. We use both Mask PQ and Boundary PQ [41] to make our conclusions more credible. Models with mask-wise merging strategy always performs better. DETR with mask-wise merging outperforms origin DETR by 1.3% PQ [1]. In addition, our mask-wise merging is 20% less time-consuming than DETR’s pixel-wise

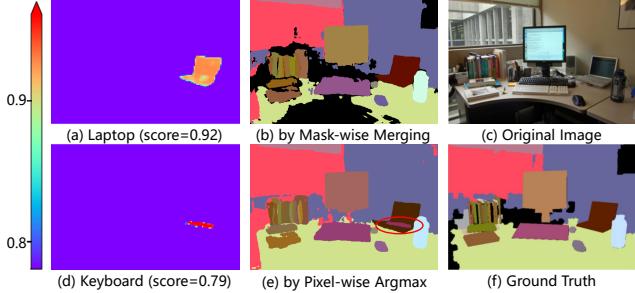


Figure 4. While using pixel-wise argmax, the keyboard is covered on the laptop (noted by the red circle in (e)). However, the laptop has a higher classification probability than the keyboard. The pixel-wise argmax strategy fails to use this important clue. Masks logits were generated through DETR-R50 [1].

Method	Mask PQ			Boundary PQ [41]		
	PQ	SQ	RQ	PQ	SQ	RQ
DETR (p)	43.4	79.3	53.8	32.8	71.0	45.2
DETR (m)	44.7	80.2	54.7	33.7	71.1	46.5
D-DETR-MS (p)	46.3	80.0	56.5	37.1	72.1	50.2
D-DETR-MS (m)	47.3	81.1	56.8	38.0	72.3	51.0
MaskFormer (p)	45.6	80.2	55.8	-	-	-
MaskFormer (p*)	46.5	80.4	56.8	36.8	72.5	49.8
MaskFormer (m)	46.8	80.4	57.2	37.6	72.6	51.1
Panoptic SegFormer (p)	48.4	80.7	58.9	39.3	73.0	52.9
Panoptic SegFormer (m)	49.6	81.6	59.9	40.4	73.4	54.2

Table 7. Effect of mask-wise merging strategy. The table shows the results of models with different post-processing methods, and the backbone is ResNet-50. “(p)” refers to using pixel-wise argmax as the post-processing method. “(p\*)” considers both class probability and mask prediction probability in its pixel-wise argmax strategy [3]. Models with “(m)” that employ mask-wise merging always perform better in both Mask PQ and Boundary PQ [41] than pixel-wise argmax method.

Method	PQ	PQ <sup>th</sup>	PQ <sup>st</sup>	AP <sup>box</sup>	AP <sup>seg</sup>
DETR [1]	43.4	48.2	36.3	38.8	31.1
D-DETR-MS [12]	47.3	52.6	39.0	45.3	37.6
Panoptic FCN [21]	43.6	49.3	35.0	36.6	34.5
Ours (Joint Matching)	48.5	<b>54.5</b>	39.5	44.1	37.7
Ours (Query Decoupling)	<b>49.6</b>	54.4	<b>42.4</b>	<b>45.6</b>	<b>39.5</b>

Table 8. Effect of query decoupling strategy. PQ and AP scores of various panoptic segmentation models on COCO val2017.

argmax since DETR uses more tricks in its code, such as merging stuff with the same category and iteratively removing masks with small areas. Fig. 4 shows one typical fail case of using pixel-wise argmax.

**Mask Decoder.** Our proposed mask decoder converges faster since the ground truth masks guide the attention module to focus on meaningful regions. Fig. 5 shows the convergence curves of several models. We only supervise the last layer of the mask decoder while not employing deep supervision. We can observe that our method achieves 49.6% PQ with training for 24 epochs, and longer training has little effect. However, D-DETR-MS needs at least 50 epochs to achieve better performance. Deep supervision is vital for

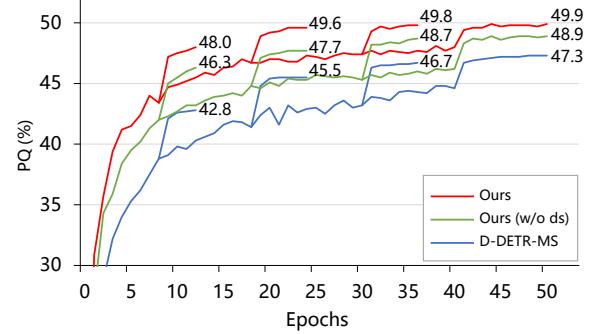


Figure 5. Convergence curves of Panoptic SegFormer and D-DETR-MS. We train models with different training schedules. “w/o ds” refers that we do not employ deep supervision in the mask decoder. The learning rate is reduced where the curves leap.

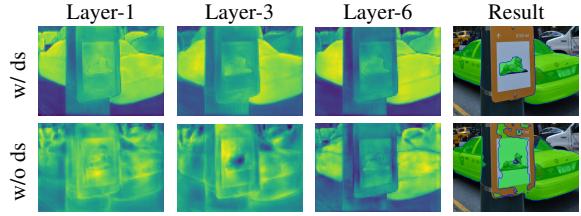


Figure 6. Attention maps of different layers in the mask decoder. “ds” refers to deep supervision.

our mask decoder to perform better and converge faster. Fig. 6 shows the attention maps of different layers in the mask decoder, and the attention module focuses on the target car in the previous layer when using deep supervision. The attention maps are very similar to the final predicted masks, since masks are generated by attention maps with a lightweight FC head.

Since our mask decoder can generate masks from each layer, we evaluate the performance of each layer in the mask decoder, see Tab. 10. During inference, using the first two layers of mask decoder will be on par with the whole mask decoder. It also infers faster because the computational cost decreases. PQ<sup>th</sup> is hardly affected by the number of layers, PQ<sup>st</sup> performs a little poorly in the first layer. The reason is that the location decoder has made additional refinements to the thing queries.

**Effect of Query Decoupling Strategy.** We compare our proposed query decoupling strategy with previous DETR’s matching method (described here as “joint matching”) [1], as shown in Tab. 8. Following DETR, joint matching uses a set of queries to target both things and stuff and feeds all queries to both location decoder and mask decoder. For our proposed query decoupling strategy, we use thing queries to detect things through bipartite matching and use location decoder to refine them. Stuff queries are assigned through class-fixed assign strategy. For a fair

Layer	PQ	PQ <sup>th</sup>	PQ <sup>st</sup>	Fps
1st	48.8	54.3	40.5	10.6
2nd	49.5	54.5	42.0	9.8
3rd	49.6	54.5	42.3	9.3
Last	49.6	54.4	42.4	7.8

Table 10. Results of each layer in the mask decoder.

Method	Clean Mean		Blur				Noise				Digital				Weather			
	Motion	Defoc	Glass	Gauss	Gauss	Impul	Shot	Speck	Bright	Contr	Satur	JPEG	Snow	Spatt	Fog	Frost		
Panoptic FCN (R50)	43.8	26.8	22.5	23.7	14.1	25.0	28.2	20.0	28.3	31.9	39.4	24.3	38.0	22.9	20.0	29.6	35.3	25.3
MaskFormer (R50)	47.0	29.5	24.9	28.1	16.4	29.5	31.2	24.7	30.9	34.8	42.5	27.5	41.2	22.0	20.4	31.0	38.5	27.7
D-DETR (R50)	47.6	30.3	25.6	28.7	16.8	29.7	32.5	24.9	31.4	35.9	43.1	28.6	41.3	24.5	21.7	31.7	39.7	28.7
Ours (R50)	50.0	32.9	26.9	30.2	17.5	31.6	35.5	27.9	35.4	38.6	45.7	31.3	43.9	29.0	24.3	35.0	41.9	32.3
MaskFormer (Swin-L)	52.9	41.7	37.3	38.0	30.4	39.3	42.3	42.5	42.8	45.3	49.7	43.9	49.4	39.7	35.2	45.2	48.8	37.9
Ours (Swin-L)	<b>55.8</b>	<b>47.2</b>	41.3	41.5	34.3	42.7	48.6	49.5	48.8	50.3	53.8	50.1	53.5	46.9	44.8	51.5	53.3	44.3
Ours (PVTv2-B5)	55.6	47.0	41.5	41.1	36.1	42.5	48.4	49.6	48.4	50.4	53.5	50.8	53.0	46.2	42.4	50.3	52.9	44.3

Table 11. Panoptic segmentation results on COCO-C. To ease the workload of the experiment, we use a subset of 2000 images from the COCO val2017. The third column is the average results on 16 types of corruption data.

comparison, both the joint matching strategy and our query decoupling strategy employ 353 queries. We can observe that our proposed strategy highly boost PQ<sup>st</sup>. In addition, panoptic segmentation model can perform instance segmentation by utilizing its thing results only. However, previous panoptic segmentation methods always perform poorly on instance segmentation task even though the two tasks are closely related. Tab. 8 shows both panoptic segmentation and instance segmentation performance of various methods. Our query decoupling strategy can achieve sota performance on panoptic segmentation task while obtaining a competitive instance segmentation performance. In short, query decoupling strategy achieves higher PQ<sup>st</sup> and AP<sup>seg</sup> compared to joint matching. We analyze the experimental results of joint matching and find that if one query prefers things more, the precision of stuff results detected by it will be lower, see Fig. 7. Each point represents the Thing-Preference and Stuff-Precision corresponding to each query, and the specific definitions are in Appendix. The red line is the linear regression of these points. When using one query set to detect things and stuff together, it will cause interference within each query. Our query decoupling strategy prevents things and stuff from interfering within the same query.

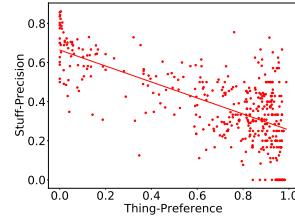


Figure 7. Things-Preference vs. Stuff-Precision.

#### 4.4. Robustness to Natural Corruptions

Panoptic segmentation has promising applications in many fields, such as autonomous driving. Model robustness is one of the top concerns of autonomous driving. In this experiment, we evaluate the robustness of our model to disturbed data. We follow [42] and generate COCO-C, which extends the COCO validation set to include disturbed data generated by 16 algorithms from blur, noise, digital and weather categories. We compare our model to Panoptic FCN [21], D-DETR-MS and MaskFormer [3]. The results are shown in Tab. 11. We calculated the mean results of disturbed data on COCO-C. Using the same backbone, our model always performs better than others. Previ-

ous literature [43–45] found that transformer-based model has stronger robustness on image classification and semantic segmentation tasks. Our experimental results also show that the transformer-based backbone (Swin-L and PVTv2-B5) can bring better robustness to the model. However, for tasks requiring a more complex pipeline, such as panoptic segmentation, we argue that the design of the task head also plays an important role for the robustness of the model. For example, Panoptic SegFormer (Swin-L) has an average result of 47.2% PQ on COCO-C, outperforming MaskFormer (Swin-L) by 5.5% PQ, higher than their gap (2.9% PQ) on clean data. We posit it is due to our transformer-based mask decoder having stronger robustness than the convolution-based pixel decoder of MaskFormer.

## 5. Conclusion

**Limitation.** This work relies on deformable attention to process multi-scale features, and the speed is a little slow. Our model is still hard to handle features with a larger spatial shape and does not perform well for small targets.

**Discussion.** Recently, the segmentation field attempted to use a uniform pipeline to process various tasks, including semantic segmentation, instance segmentation, and panoptic segmentation. However, we think that complete unification is conceptually exciting but not necessarily a suitable strategy. Given the similarities and differences among the various segmentation tasks, “seek common ground while reserving differences” is a more reasonable guiding ideology. With query decoupling strategy, we can handle things and stuff in the same paradigm since they are represented as queries. In addition, we can also design customized pipelines for things or stuff. Such a flexible strategy is more suitable for various segmentation tasks. At present, task-specific designs still bring better performance. We encourage the community to further explore the unified segmentation frameworks and expect that Panoptic SegFormer can inspire future works.

## 6. Acknowledge

This work is supported by the Natural Science Foundation of China under Grant 61672273 and Grant 61832008. Ping Luo is supported by the General Research Fund of HK

No.27208720 and 17212120. Wenhui Wang and Tong Lu are corresponding authors.

## Appendix

### A. Implementation Details

#### A.1. Panoptic SegFormer.

Our settings mainly follow DETR [1] and Deformable DETR [12] for simplicity. The hyper-parameters in deformable attention are the same as Deformable DETR [12]. We use Channel Mapper [12,46] to map dimensions of the backbone’s outputs to 256. The location decoder contains 6 deformable attention layers, and the mask decoder contains 6 vanilla cross-attention layers [24]. The spatial positional encoding is the commonly used fixed absolute encoding that is the same as DETR. The window size of Swin-L [34] we used is 7. Since we equally treat each query.  $\lambda_{\text{things}}$  and  $\lambda_{\text{stuff}}$  are dynamically adjusted according to the relative proportion of things and stuff in each image, and their sum is 1.  $\lambda_{\text{cls}}$ ,  $\lambda_{\text{seg}}$ , and  $\lambda_{\text{det}}$  in Eq. (4) are set to 2, 1, 1, respectively.

During the training phase, the predicted masks that be assigned  $\emptyset$  will have a weight of zero in computing  $\mathcal{L}_{\text{seg}}$ . While using the mass center of instance to replace the bounding box, we only use L1 loss to supervise the mass center of predicted mask and mass center of ground truth. We employ a threshold 0.5 to obtain binary masks from soft masks. Threshold  $t_{\text{cnf}}$  and  $t_{\text{keep}}$  are 0.25 (0.3) and 0.6, respectively.  $\alpha$  and  $\beta$  in Eq. (6) are 1 and 2, respectively. All experiments are trained on one NVIDIA DGX node with 8 Tesla V100 GPUs.

By default, for COCO dataset [11], We train our models with 24 epochs, a batch size of 1 per GPU, a learning rate of  $1.4 \times 10^{-4}$  (decayed at the 18th epoch by a factor of 0.1, learning rate multiplier of the backbone is 0.1). We use a multi-scale training strategy with the maximum image-side not exceeding 1333 and the minimum image size varying from 480 to 800, and random crop augmentations is applied during training. The number of thing queries  $N_{\text{th}}$  is set to 300. Stuff queries have the equal number of stuff classes, and it is 53 in COCO.

For the ADE20K dataset [33], we train our model with 100 epochs (decayed at 80th epoch), image size varying from 512 to 2048. Since ADE20K contains 50 stuff, we use 50 stuff queries. Other settings are the same to COCO dataset.

**FPS and FLOPs.** FPS in Tab.5 is measured on a V100 GPU with a batch size of 1. “DETR” and “DETR+mask wise merging” are from Detectron2 [47] and DETR’s implementation. Others are from Mmdet [46] and our own implementation. Our framework is slightly more efficient than DETR. FLOPs of DETR are measured from Detectron2 on an average of 100 images.

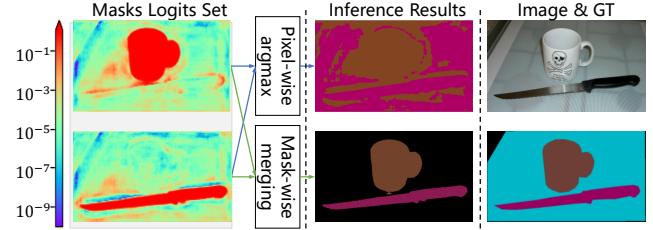


Figure B.1. **Pixel-wise Argmax vs. Mask-wise Merging.** We use DETR-R50 to compare the results generated through pixel-wise argmax and mask-wise merging. Firstly, DETR-R50 detects a cup and a knife from the image. When using pixel-wise argmax, other pixels (dining table) are incorrectly filled with “cup” or “knife”. It mistakenly believes that the largest mask logit is the correct result, regardless of its value. However, our mask-wise merging strategy generates the correct results since we binarize each mask.

### A.2. Deformable DETR for Panoptic Segmentation

Following DETR for panoptic segmentation, we transplanted the panoptic head of DETR to Deformable DETR. To ensure consistency, we only generate the attention maps with the spatial shape of 32s. When using single scale deformable DETR, the process of generating attention maps is the same as DETR. When using multi-scale deformable DETR, we only multiply queries and the features (from C5) to generate attention maps. Other settings of deformable DETR for object detection are kept unchanged. We apply iterative bounding box refinement as the default setting for Deformable DETR. We use 300 queries and this brings huge computation costs, although this model achieves pretty good performance.

## B. Discussion

We will deliver more ablation studies, more detailed analysis in this section.

### Effect of Deformation

**Attention.** To ablate the effect of deformable attention, we extend Deformable DETR on panoptic segmentation. Table B.1. “D-”, “SS” and “MS” refers to “Deformation with the panoptic head formable”, single-scale and multi-scale. For more implementation details, please refer to the Appendix A. As shown in Tab. B.1, multi-scale deformable attention improves 2.9% PQ compared to DETR. Multi-scale attention outperforms single-scale attention by 5.7% PQ, highlighting the important role of multi-scale features for segmentation task.

### B.1. Post-processing Method

**Defects of Pixel-wise Argmax.** Pixel-wise argmax only considers the mask logits of each pixel. It has multiple issues that may lead to incorrect results. First of all, the

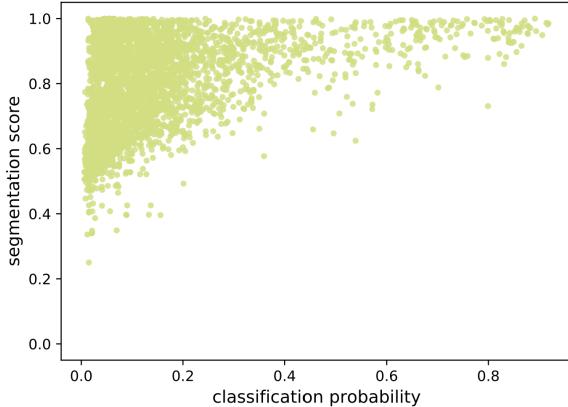


Figure B.2. The joint distribution for classification probability and segmentation score. We can observe that segmentation scores can be high while the masks have low classification probability.

Post-Processing Method	PQ	PQ <sup>th</sup>	PQ <sup>st</sup>
Pixel-wise Argmax	48.4	53.2	41.3
Heuristic Procedure [6]	48.4	54.3	39.4
Mask-wise Mering	49.6	54.4	42.4

Table B.2. The results of Panoptic SegFormer (R50) with different post-processing methods. Because the heuristic procedure always prefers things, it has the lowest PQ<sup>st</sup>.

$\alpha$	$\beta$	PQ	PQ <sup>th</sup>	PQ <sup>st</sup>
1	0	48.7	53.5	41.3
0	1	44.4	52.1	32.7
1	1	49.3	54.1	42.1
1	2	49.6	54.4	42.4
1	3	49.7	54.5	42.2

Table B.3. The weights of the classification score and segmentation score determine the priority of masks. We can observe that employing both of them will perform better. According to the results on multiple models, we choose  $\alpha=1, \beta=2$  as our default setting.

pixel value generated from argmax may be extremely small, as shown in Fig. B.1, which will generate plenty of false-positive results. The second issue is that the pixel with max mask logit may be the suboptimal result, as shown in Fig. 4 of the paper. This kind of error frequently appears in the segmentation maps generated by pixel-wise argmax. MaskFormer [3] alleviates this problem by multiplying the classification probability by the masks logits. But this kind of error will still exist.

**Heuristic Procedure.** The heuristic procedure [6] was the first proposed post-processing method of panoptic segmentation. It uses different strategies to handle things and stuff separately. Pixel-wise argmax was still used in its stuff workflow. One apparent defect of this method is that it solves the overlap problem of stuff and things by always

$t_{cnf}$ \ $t_{keep}$	0.9	0.8	0.7	0.6	0.5
0.20	48.9	<b>49.5</b>	<b>49.6</b>	<b>49.6</b>	49.4
0.25	48.9	<b>49.6</b>	<b>49.7</b>	<b>49.7</b>	<b>49.5</b>
0.30	48.8	<b>49.5</b>	<b>49.6</b>	<b>49.6</b>	<b>49.5</b>
0.35	48.3	49.1	49.2	49.2	49.1
0.40	47.4	48.1	48.2	48.2	48.1

Table B.4. We use two thresholds  $t_{cnf}$  and  $t_{keep}$  in our mask-wise merging. We evaluate the results by combining different thresholds with Panoptic SegFormer (R50) to verify whether our algorithm is sensitive to these thresholds. Results higher than 49.5 are displayed in bold.

#Head	PQ	PQ <sup>th</sup>	PQ <sup>st</sup>
1	49.2	54.0	42.0
8	49.6	54.4	42.4

Table B.5. We varied the number of heads in our mask decoder. More heads can bring slight performance improvements.

preferring things. This is an unfair way of dealing with stuff. Tab. B.2 shows that PQ<sup>st</sup> of using heuristic procedure is lower than other methods because all stuff are treated unfairly.

**Masks-wise Merging.** Post-processing of panoptic segmentation aims to solve the overlap problem between masks. Although pixel-wise argmax uses an intuitive method to solve the overlap problem, it has defects mentioned above. We solve the overlap problem by giving different masks different priorities. Mask-wise merging guarantees that high-quality masks have higher priority by sorting the masks with confidence scores. This strategy ensures that low-quality instances will not cover high-quality instances. In order to be able to effectively distinguish the quality of the masks, we consider both classification probability and segmentation score as the confidence score of each mask. The segmentation score average( $\mathbb{1}_{\{m_i[h,w]>0.5\}} m_i[h,w]$ ) represents the confidence of the overall segmentation quality of the mask. Tab. B.3 shows the results of varying  $\alpha$  and  $\beta$  in Eq.6. Applying both classification probability and segmentation scores always have better performance. Fig. B.2 shows the relationship of classification probability and segmentation score. While one mask has a low classification probability ([0, 0.4]), it may have a large segmentation score. Large segmentation score means it has many pixels with high logits and this may generate false-positive results through pixel-wise argmax since its classification probability is pretty low.

Our mask-wise merging needs two thresholds to filter out undesirable results. Tab. B.4 shows that our algorithm is not very dependent on the choice of threshold  $t_{cnf}$  and  $t_{keep}$ . Tab. B.4

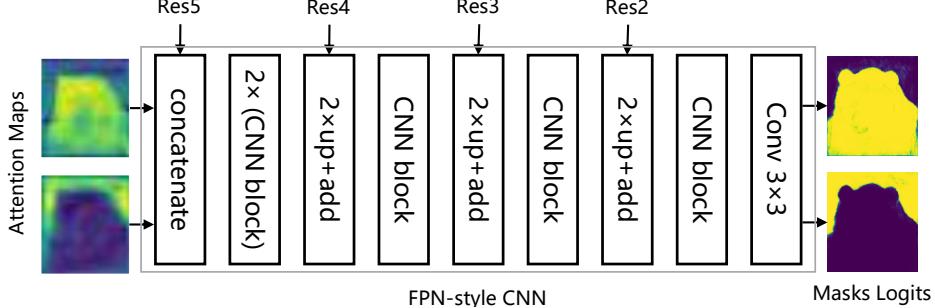


Figure B.3. Illustration of DETR’s panoptic head. CNN block consists of  $3 \times 3$  convolution, GN, and ReLU.

$P_t$	#Query	Stuff			Things		
		TP	TP+FP	Precision	TP	TP+FP	Precision
[0.0, 0.1)	44	7318	10060	0.73	136	222	0.61
[0.1, 0.2)	17	839	1308	0.64	140	198	0.71
[0.2, 0.3)	4	121	212	0.57	53	69	0.77
[0.3, 0.4)	11	339	646	0.52	252	368	0.68
[0.4, 0.5)	10	211	446	0.47	212	365	0.58
[0.5, 0.6)	15	327	684	0.48	477	903	0.53
[0.6, 0.7)	24	339	810	0.42	1001	1465	0.68
[0.7, 0.8)	40	400	1094	0.37	2019	3255	0.62
[0.8, 0.9)	83	539	1586	0.34	6325	9687	0.65
[0.9, 1.0]	105	309	1029	0.30	11252	16724	0.67
Total	353	10742	17875	0.60	21867	33256	0.66

Table B.6. We divide 353 queries into ten groups according to their  $P_t$ . For each group, we calculate their precision on stuff and things. Queries with higher  $P_t$  have very low precision when they predict stuff. This demonstrates that things may interfere with the prediction of stuff and using queries to target both things and stuff is suboptimal.

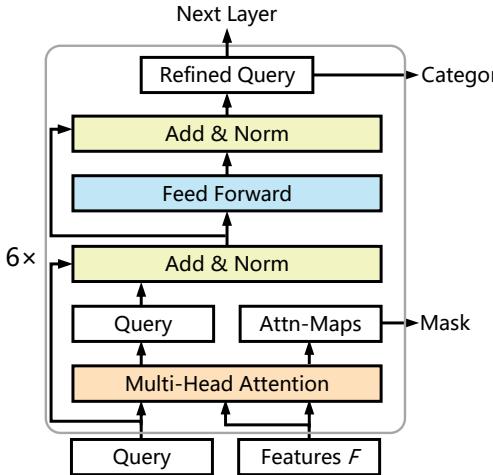


Figure B.4. **Architecture of mask decoder.** Attn-Maps notes attention maps.

Although our proposed mask-wise merging strategy has achieved better results than other post-processing methods, it also has several shortcomings. First of all, we binarize the mask through a fixed threshold. This may cause one pixel to be easily assigned a void label because the values of

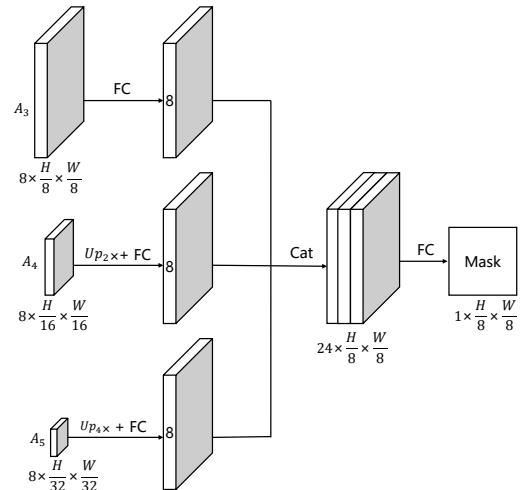
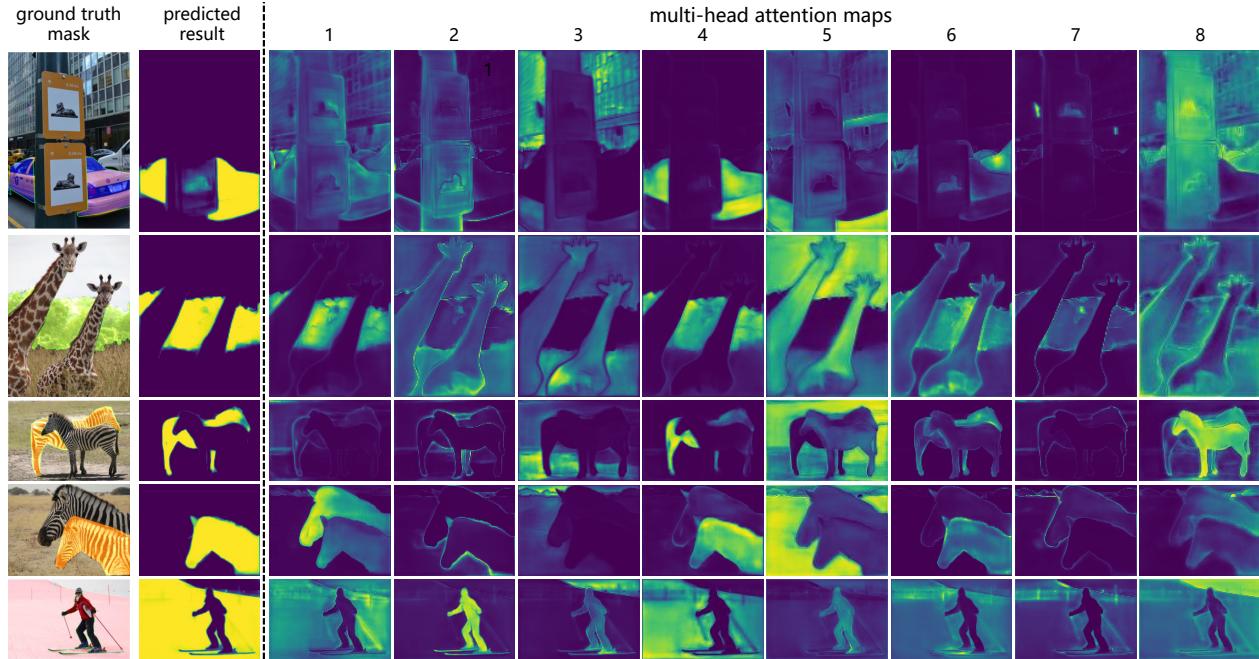
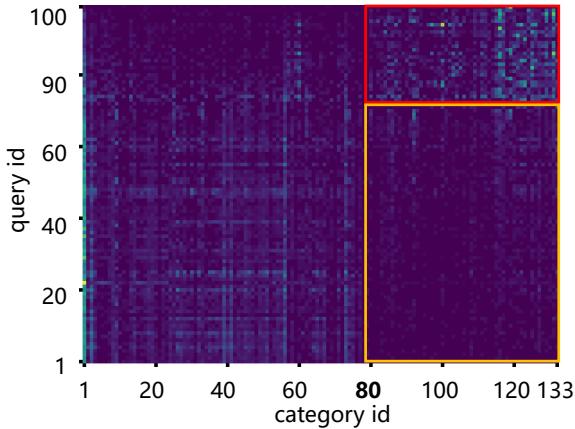


Figure B.5. Illustration of the module that generates mask from multi-scale multi-head attention maps.  $Up_{2x}$  means upsampling by two times. FC notes fully connected layer. Cat notes concatenate. While using 8 heads in the attention module, this module only contains 200+ parameters.

all candidate instances at this pixel are below the threshold. Secondly, our strategy highly depends on the accuracy of



**Figure B.6. Visualization of multi-head attention maps and corresponding outputs from mask decoder.** Different heads have different preferences. Head 4 and Head 1 pay attention to foreground regions, and Head 8 prefers regions that occlude foreground. Head 5 always pays attention to the background that is around the foreground. Through the collaboration of these heads, Panoptic SegFormer can predict accurate masks. The 3rd row shows an impressive result of a horse that is highly obscured by the other horse.



**Figure B.7. The Joint Distribution for Queries and Category in DETR.** We can observe that queries prefer either things or stuff. Although a few queries predict most of the stuff results (within the red box), other queries still generate a considerable proportion of stuff results (within the yellow box). Our experimental results demonstrate that the results in the yellow box are usually of low quality. We sort the query ids for better visualization. Other literature [2] reports similar phenomenon.

confidence scores. If the confidence scores are not accurate, it will produce a low-quality panoptic format mask.

## B.2. Location Decoder

Although we use the location decoder to detect the bounding boxes of things, our workflow is still very different from the previous box-based panoptic segmentation. For example, Panoptic FPN performs instance segmentation with Mask R-CNN style. The two-stage method usually needs to extract regions from the feature based on the bboxes and then use these regions to perform segmentation. The quality of segmentation is heavily dependent on the quality of detection. However, our location decoder is used to assist in learning the location clues of the query and distinguishing different instances. Mask will not have the wrong boundary due to the wrong boundary prediction of the bbox since the bbox does not constrain the mask. We also show that using mass centers of masks to replace bboxes can still learn location clues.

Another valuable function of the location decoder is to help filter out low-quality thing queries during the training and inference phase. This can greatly save memory. Current transformer-based panoptic segmentation methods always consume a lot of GPU memory. For example, MaskFormer takes up more than 20G of GPU memory with a batch size

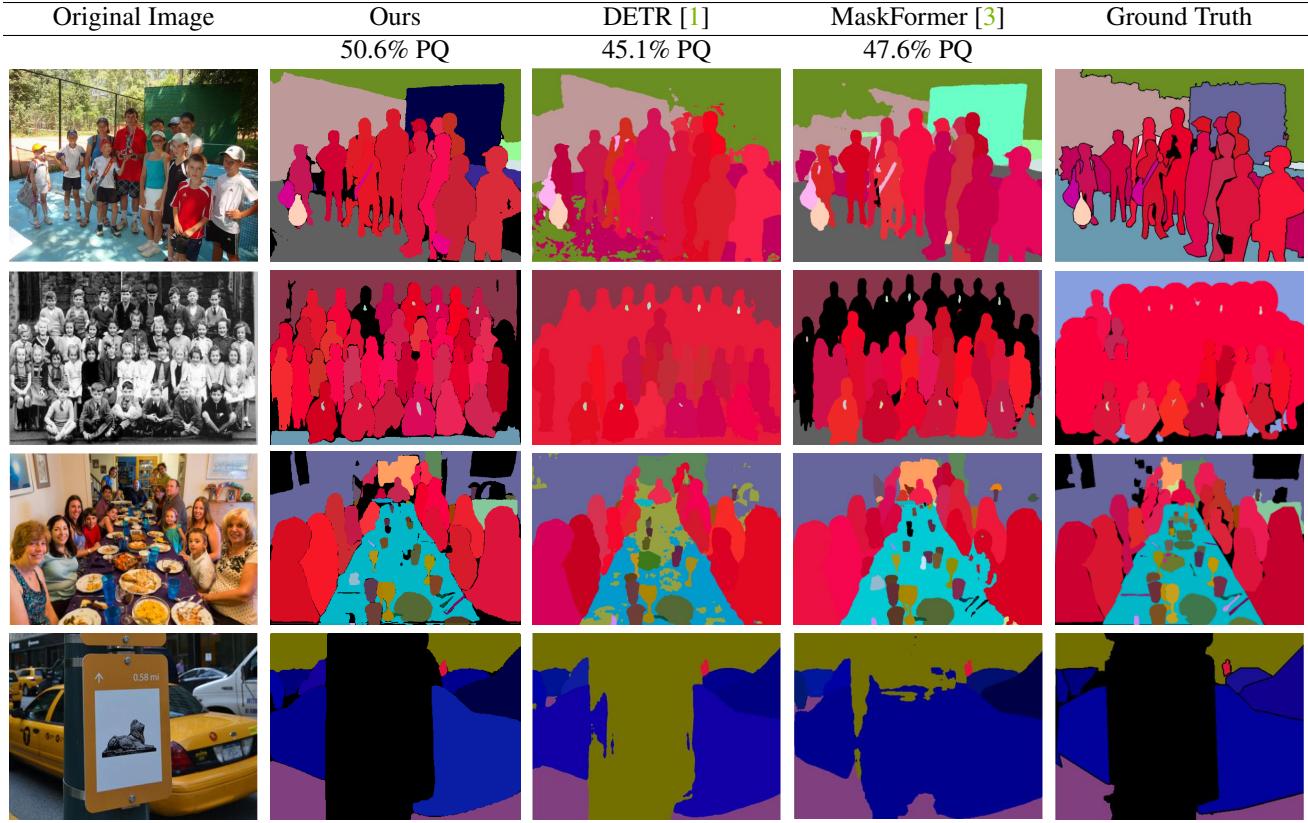


Figure B.8. Comparing visualization results of Panoptic SegFormer with other methods on the COCO `val` set. For a fair comparison, all results are generated with ResNet-101 [23] backbone. The second and fourth row results show that our method still performs well in highly crowded or occluded scenes. Benefits from our mask-wise inference strategy, our results have few artifacts, which often appear in the results of DETR [1] (e.g., dining table of the third row).

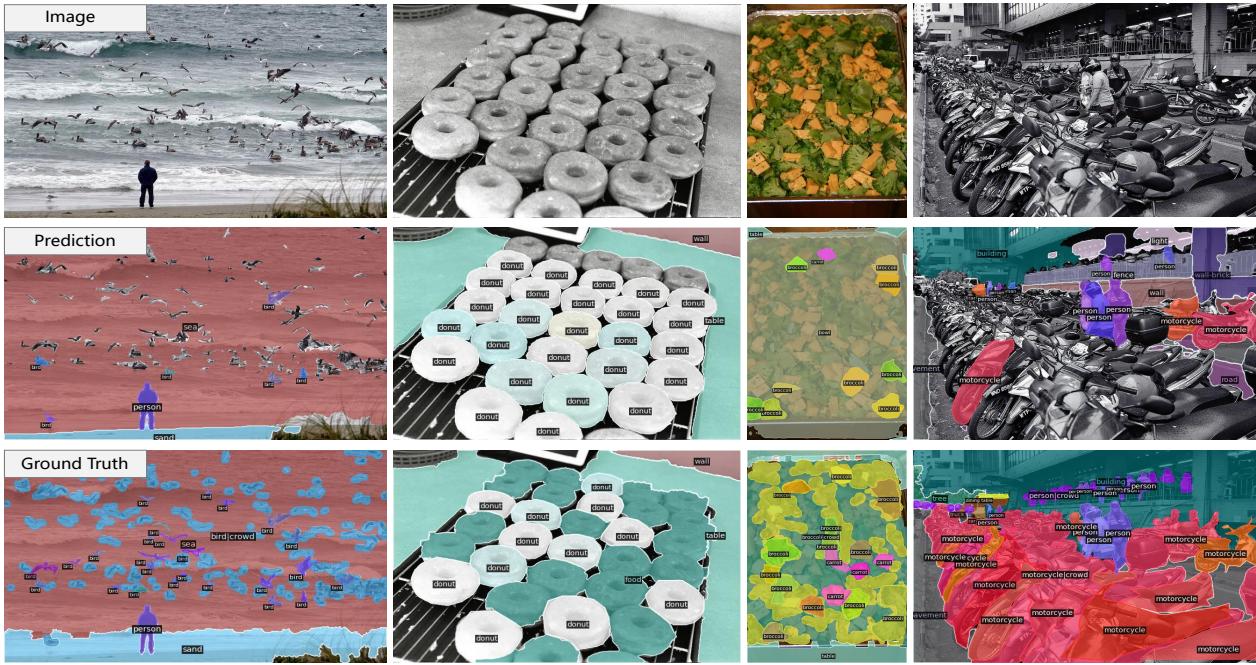


Figure B.9. Failure case of Panoptic SegFormer.



Figure B.10. Visualization results of some complex scenes.

of 1 and R50 backbone. Although these methods have achieved excellent results, they also require high hardware resources. However, our Panoptic SegFormer can be trained with taking up less than 12G memory by using a location decoder to filter out low-quality thing queries. In particular, we use bipartite matching for multiple rounds of matching in the detection phase. The thing query that already be matched will not participate in the next round of matching. After several rounds, we can select partial promising thing queries. Only these promising thing queries will be fed to the mask decoder. with this strategy, the mask decoder usually only needs to handle less than 100 thing and stuff queries.

### B.3. Mask Decoder

Fig. B.3 shows the architecture of DETR’s panoptic head. Although it only contains 1.2M parameters, it has a huge computational burden (about 150G FLOPs). DETR adds ResNet features to each attention map, and this process repeats 100 times since there are 100 attention maps. Fig. B.4 shows the model architecture of our mask decoder. Fig. B.5 shows the process of converting multi-scale multi-head attention maps to mask. We found that discarding the self-attention in the decoder does not affect the effectiveness of the model. The computational cost of our mask decoder is around 30G FLOPs.

**Multi-head attention maps.** Fig. B.6 shows some samples of multi-head attention maps. Through a multi-head attention mechanism, different heads of one query learn

their own attention preference. We observe that some heads pay attention to foreground regions, some prefer boundaries, and others prefer background regions. This shows that each mask is generated by considering various comprehensive information in the image. Tab. B.5 shows that utilizing a multi-head attention mechanism will outperform single-head attention by 0.4% PQ.

#### B.4. Advantage of Query Decoupling Strategy

DETR uses the same recipe to predict boxes of things and stuff (To facilitate the distinction between the query decoupling strategy we proposed, we refer to the DETR’s strategy as a joint matching strategy.). However, detecting bboxes for stuff like DETR is suboptimal. We counted the ratio of the area of masks to the area of bboxes on the COCO train2017. The ratios of things and stuff are 52.5% and 9.2%, separately. This shows that bounding boxes can not represent stuff well since the stuff is amorphous and dispersed. We also observe that bbox AP of DETR drops from 42.0 to 38.8 after training on panoptic segmentation. This may be due to the interference of stuff on things, since predicting stuff bboxes needs to re-adapt the model.

Fig. B.7 shows that DETR seems to learn automatic segregation between things and stuff, and each query either prefers things or stuff. However, we argue that this automatically learned segregation is not ideal. If one query prefers things, it will perform poorly when it generates stuff results. This situation is very common, and our following experiments based on Panoptic SegFormer will give detailed data. Following DETR, we use 353 queries to predict things and stuff with the same recipe. Specifically, the input of the location decoder is 353 queries, which will detect both things and stuff. The refined queries are fed to the mask decoder to predict category labels and masks. We define a query’s preference for things as  $P_t$ , which can be calculated by:

$$P_t^i = N_{\text{things}}^i / (N_{\text{things}}^i + N_{\text{stuff}}^i), \quad (7)$$

where  $N_{\text{things}}$  and  $N_{\text{stuff}}$  are the number of things and stuff masks that  $i$ -th query predicted on COCO val set.  $P_t^i > 0.5$  means that  $i$ -th query prefers things more than stuff. The predicted mask is a true positive (TP) if IoU between it and one ground truth mask is larger than 0.5 and the category of them is the same. Then we can calculate the precision of queries’ predicted masks. Tab. B.6 shows relevant statistical results. First, we can observe that the queries that own lower  $P_t$  basically have higher precision. The stuff precision of the queries that have the highest  $P_t$  ([0.9, 1.0]) only is 0.30, which is much lower than the average stuff precision (0.60) on all queries. These erroneous results are mainly due to errors in the predicted category. Queries that have no obvious preference for stuff and things(  $P_t$  in [0.4, 0.6] ) performs poorly both on stuff and things. These results demonstrate that using one query set to predict things and

stuff simultaneously is flawed. This joint matching strategy is suboptimal for stuff.

In order to avoid mutual interference between stuff and things, we propose the query decoupling strategy to handle things and queries with a separate query set. Compared to stuff query, thing query will go through an additional location decoder. However, all queries will produce the outputs in the same format. Things and stuff use the same loss for training, except that things use an additional detection loss. During inference, we can use our mask-wise merging strategy to merge them uniformly. This is different from the previous methods that modeled panoptic segmentation into instance segmentation and semantic segmentation. For example, Panoptic FPN uses one branch to generate things and one branch to generate stuff. The things and stuff generated by Panoptic FPN are in different formats and need different training strategies and post-processing methods. PQ<sup>st</sup> with query decoupling outperforms joint matching strategy by 2.9% PQ and experimental results verify the effectiveness of our method. The stuff precision by using query decoupling is 0.66, better than the joint matching strategy.

### C. Visualization

Fig. B.8 shows our visualization result against DETR and MaskFormer. We use the original codes that they officially implemented. First of all, compared with other methods, we can observe that our results are more consistent with ground truths. Due to the defects of pixel-wise argmax we discussed in Appendix B.1, DETR always generates results with artifacts. MaskFormer performs better because they improved pixel-wise argmax by considering classification probabilities. However, it may still fail in hard cases. For example, it recognizes the billboard as a car in the fourth row. Fig. B.9 shows some failure cases of our model. Firstly, our model may have lower recall when facing crowded scenarios filled with the same things, especially for the small targets. Another typical failure mode is that large stuff with a high confidence score occupies most of the space, causing other things not to be added to the canvas. Fig. B.10 shows the results on some complex scenes.

### D. Various Backbones

We give all the panoptic segmentation results under various backbones, as shown in Tab. D.1. Fig. D.1 shows two training curves with backbone ResNet-101 and Swin-L. With Swin-L, Panoptic SegFormer with training for 24 epochs even performs better than training for 50 epochs.

Backbone	PQ	SQ	RQ	PQ <sup>th</sup>	SQ <sup>th</sup>	RQ <sup>th</sup>	PQ <sup>st</sup>	SQ <sup>st</sup>	RQ <sup>st</sup>
ResNet-50 [23]	49.6	81.6	59.9	54.4	82.7	65.1	42.4	79.9	52.1
ResNet-101 [23]	50.6	81.9	60.9	55.5	83.0	66.3	43.2	80.1	52.9
PVTv2-B0 [5]	49.5	82.4	59.2	55.3	83.3	65.8	40.6	80.9	49.2
PVTv2-B2 [5]	52.5	82.7	62.7	58.5	83.6	69.5	43.4	81.4	52.4
PVTv2-B5 [5]	55.4	82.9	66.1	61.2	84.0	72.4	46.6	81.3	56.5
Swin-L [34]	55.8	82.6	66.8	61.7	83.7	73.3	46.9	80.9	57.0

Table D.1. Panoptic segmentation results on COCO val with various backbones.

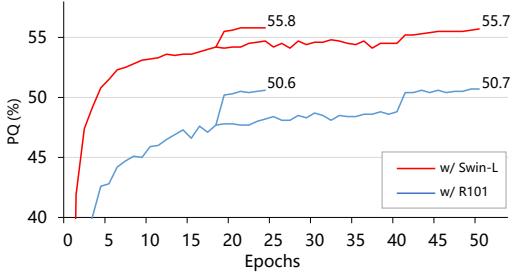


Figure D.1. By using ResNet-101 [23] and Swin-L as the backbone, we train our model for 24 epochs and 50 epochs, separately. We can observe that our model that training for 24 epochs can achieve comparable or even higher results while comparing the models that training for 50 epochs.

## E. Code and Data

We use the official implementations of DETR<sup>1</sup>, MaskFormer<sup>2</sup>, Panoptic FCN<sup>3</sup> to perform additional experiments. The models they provide all can reproduce the same scores they reported in their literature. Deformable DETR is from Mmdet<sup>4</sup>.

## References

- [1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1, 2, 3, 4, 5, 6, 7
- [2] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, 2021. 1, 2, 3, 4, 5, 6
- [3] Bowen Cheng, Alexander G Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 1, 2, 3, 5, 6, 7, 8
- [4] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-Net: Towards unified image segmentation. In *NeurIPS*, 2021. 1, 2, 3, 5, 6
- [5] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Ptv2: Improved baselines with pyramid vision transformer. *arXiv:2106.13797*, 2021. 1, 6, 8
- [6] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019. 1, 2, 4, 5
- [7] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019. 1, 2, 3, 6
- [8] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms—improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017. 1
- [9] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *CVPR*, 2019. 1, 2
- [10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1
- [11] 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 *ECCV*, 2014. 2, 5, 1
- [12] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 2, 3, 4, 6, 7, 1
- [13] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6910–6919, October 2021. 2, 3, 6
- [14] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 2, 6
- [15] Ujwal Bonde, Pablo F Alcantarilla, and Stefan Leutenegger. Towards bounding-box free panoptic segmentation. In *DAGM German Conference on Pattern Recognition*, 2020. 2
- [16] Qizhu Li, Xiaojuan Qi, and Philip HS Torr. Unifying training and inference for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13328, 2020.

<sup>1</sup><https://github.com/facebookresearch/detr>

<sup>2</sup><https://github.com/facebookresearch/MaskFormer>

<sup>3</sup><https://github.com/dvlab-research/PanopticFCN>

<sup>4</sup><https://github.com/open-mmlab/mmdetection>

- [17] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, 2020. 2
- [18] Tien-Ju Yang, Maxwell D Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Pandreou, and Liang-Chieh Chen. Deeperlab: Single-shot image parser. *arXiv:1902.05093*, 2019. 2
- [19] Naiyu Gao, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, and Kaiqi Huang. SSAP: Single-shot instance segmentation with affinity pyramid. In *ICCV*, 2019. 2
- [20] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *CVPR*, 2019. 2
- [21] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *CVPR*, 2021. 2, 3, 6, 7, 8
- [22] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, 2020. 2, 4
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 6, 5, 8
- [24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3, 1
- [25] Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Solq: Segmenting objects by learning queries. *NeurIPS*, 2021. 3, 6
- [26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 3
- [27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 3, 5
- [28] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. SOLOv2: Dynamic and fast instance segmentation. *NeurIPS*, 2020. 3, 4, 5, 6
- [29] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *ECCV*, 2020. 4
- [30] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *CVPR*, 2016. 5
- [31] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5
- [32] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International conference on 3D vision (3DV)*, 2016. 5
- [33] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 5, 6, 1
- [34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*, 2021. 6, 1, 8
- [35] Chongsong Chen, Jiawei Ren, Daisheng Jin, Zhongang Cai, Cunjun Yu, Bairun Wang, Mingyuan Zhang, and Jinyi Wu. Joint coco and mapillary workshop at iccv 2019: Coco panoptic segmentation challenge track technical report: Panoptic htc with class-guided fusion. *SHR*, 56(84.1):67–2, 6
- [36] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Yukang Chen, Lu Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation with point-based supervision. *arXiv preprint arXiv:2108.07682*, 2021. 6
- [37] Yangxin Wu, Gengwei Zhang, Yiming Gao, Xiajun Deng, Ke Gong, Xiaodan Liang, and Liang Lin. Bidirectional graph reasoning network for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9080–9089, 2020. 6
- [38] Yangxin Wu, Gengwei Zhang, Hang Xu, Xiaodan Liang, and Liang Lin. Auto-panoptic: Cooperative multi-component architecture search for panoptic segmentation. *Advances in Neural Information Processing Systems*, 33, 2020. 6
- [39] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 6
- [40] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 6
- [41] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15334–15342, 2021. 6, 7
- [42] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8828–8838, 2020. 8
- [43] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 8
- [44] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. *arXiv preprint arXiv:2103.14586*, 2021. 8
- [45] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan

- Yang. Intriguing properties of vision transformers. *arXiv preprint arXiv:2105.10497*, 2021. 8
- [46] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Zi-wei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDection: Open MMLab detection toolbox and benchmark. *arXiv:1906.07155*, 2019. 1
- [47] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 1