Neurocomputing 506 (2022) 290–299

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Real-time panoptic segmentation with relationship between adjacent pixels and boundary prediction

Xiaoliang Zhang, Hongliang Li *, Lanxiao Wang, Haoyang Cheng, Heqian Qiu, Wenzhe Hu, Fanman Meng, Qingbo Wu *

School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

ARTICLE INFO

Article history:
Received 16 January 2022
Revised 19 July 2022
Accepted 24 July 2022
Available online 26 July 2022

Keywords:
Panoptic segmentation
Graph convolution
Fully convolution
Relationship between adjacent pixels
Boundary prediction

ABSTRACT

Panoptic segmentation has recently received increasing attention since it generates coherent scene segmentation by unifying semantic and instance segmentation. The most popular methods for panoptic segmentation are currently based on an instance segmentation framework with a semantic segmentation branch in parallel. However, these methods are too bloated for real-world applications. In this paper, we propose a simple yet effective fully convolutional network for fast panoptic segmentation. Instead of directly generating the mask for each instance, we leverage a simple graph convolutional layer to construct a pixel relationship head to predict the relationship between two adjacent pixels and determine whether they belong to the same instance. Besides, we leverage boundary information to enhance supervision information and help our method distinguish adjacent objects. Combining predicted category labels for each pixel from the semantic segmentation branch, we can generate a unified panoptic segmentation mask in a parameter-free step. We demonstrate our method's effectiveness on MS COCO dataset and Cityscapes dataset, which obtain competitive results.

1. Introduction

Unlike semantic segmentation [1–9] classifies each pixel into a fixed set of categories without differentiating between different instances, or instance segmentation [10–18] only segments each foreground instance with a semantic label, panoptic segmentation [19] is considered an emerging and challenging task, which needs to generate a coherent scene segmentation for both foreground "things" and background "stuff", which attracts increasing attention and has become one of the most crucial computer vision tasks. This task needs to not only segment each foreground instance ("things") precisely, but also classify all background pixels ("stuff") correctly. Therefore, panoptic segmentation can be considered a task combining semantic and instance segmentation, and can be widely adopted in real-world applications, such as robots, video surveillance, autonomous driving, *etc.* This task requires speed as well as performance. Panoptic segmentation is first proposed by *Kirillov et al.* [19]. A representative approach to completing the task is Panoptic FPN [20]. Adapting the framework of Mask R-CNN [11], one of the most dominant instance segmentation methods, Panoptic FPN [20] adds a fully convolutional network as a branch for "stuff" semantic segmentation and then generates panoptic segmentation results in a parameter-free post-processing method. Inspired by this, most subsequent methods regard panoptic segmentation as a combination of instance and semantic segmentation.

However, this structure is too bloated to be applied in the real world. Firstly, based on Mask R-CNN [11], operations like RPN [21] and RoIAlign [11] are time-consuming. Secondly, predicting the target category is unnecessary, which can be entirely replaced by semantic segmentation. Thirdly, due to the overlap of predicted bounding boxes, massive areas are repeatedly segmented, whereas panoptic segmentation only needs to output a unified segmentation mask. Lastly, some areas are segmented in the instance and semantic segmentation branches redundantly. These problems lead to a bloated network and huge computational costs, making panoptic segmentation far from real-world applications.

In this paper, we propose a fully convolutional neural network for panoptic segmentation based on the relationship between adjacent pixels. Specifically, if a pixel belongs to an instance, we determine whether its adjacent pixels belong to the same instance. Thus, we model it as a binary classification problem and utilize a

* Corresponding authors.
E-mail addresses: xlzhang@std.uestc.edu.cn (X. Zhang), hlli@uestc.edu.cn (H. Li), lanxiao.wang@std.uestc.edu.cn (L. Wang), chenghaoyang@std.uestc.edu.cn (H. Cheng), hqqiu@std.uestc.edu.cn (H. Qiu), wenzhe-hu@std.uestc.edu.cn (W. Hu), fmmeng@uestc.edu.cn (F. Meng), qbwu@uestc.edu.cn (Q. Wu).

https://doi.org/10.1016/j.neucom.2022.07.078
0925-2312/

graph convolutional layer to predict the relationship. According to the above process, we can generate class-agnostic masks. In addition, we leverage the boundary information to enhance the supervision and help our method distinguish adjacent instances. In the inference stage, the semantic segmentation branch predicts class labels for pixels. We combine them with the fore-mentioned class-agnostic masks and then generate a unified panoptic segmentation mask in a parameter-free module. Due to the simplicity of a fully convolutional network structure and omitting instance segmentation as a pre-task, our method can save massive time and decrease computational cost, achieving fast panoptic segmentation while ensuring good performance.

To sum up, the main contributions in this paper are highlighted as follow:

- We propose a simple but effective method for panoptic segmentation. Our method is a fully convolutional neural network and can achieve fast panoptic segmentation while ensuring good performance.
- To predict the relationship between two adjacent pixels, we propose a pixels relationship head to model these relationships by leveraging a graph convolutional layer by judging whether a pixel's adjacent pixel belongs to the same instance as this pixel.
- To enhance supervision information and help our method distinguish adjacent instances, we propose a boundary prediction head to suppress the relationship value on the boundary and improve panoptic segmentation accuracy.
- A parameter-free fusion module is proposed to generate a unified panoptic segmentation mask for each input image. Without training, this step combines the predicted class-agnostic instance mask with the semantic segmentation mask to develop a unified panoptic segmentation mask.
- Experiments on MS COCO dataset [22] and Cityscapes dataset [23] prove that our method achieves fast panoptic segmentation while still guaranteeing that our performance is competitive.

2. Related work

2.1. Semantic segmentation

Semantic segmentation is a fundamental computer vision task, which needs to predict a class label for each pixel in an image. Many works [1–5,9,6–8] have achieved outstanding performance in this field. FCNs [3] performs up-sampling through deconvolution and implements end-to-end convolutional neural network training for the first time in semantic segmentation task. SegNet [4] employs the encoder-decoder structure and gradually recovers resolution to capture more detailed information. PSPNet [5] exploits spatial pyramid pooling to cascade feature maps of different scales to achieve the fusion of multi-layer semantic features. DeepLab [6–8] uses atrous spatial pyramid pooling to combine multi-scale information and dramatically improves semantic segmentation performance. DANet [9] adaptively integrates local features in spatial and channel dimensions, respectively.

2.2. Instance segmentation

Compared with object detection, instance segmentation requires not only to locate all instances in an image correctly but also to segment each instance precisely. On one hand, the dominant instance segmentation methods [10–15,24,25] get bounding boxes or region proposals first, then segment each instance within the bounding box. Mask R-CNN [11] is the most representative method, which extends on two-stage object detection method Faster R-CNN [21] by adding a branch for predicting instance masks. Following Mask R-CNN [11], Mask Scoring R-CNN [12], PANet [13] and HTC [14] exhibit competitive performance. Based on one-stage object detection methods, FCOS [26] and RetinaNet [27], CenterMask [15], and YOLACT [16] achieves fast instance segmentation and succeeds in performance.

On the other hand, some instance segmentation methods do not need object detection as a pre-task. SpatialEmbeddings [17] clusters the spatial embedding of pixels that belong to the same instance. Jointly learning affinity pyramid and semantic class labeling, SSAP [18] hierarchically generates instance segmentation masks. Instead of predicting instance masks, PolarMask [28] and PolyTransform [29] compute a contour for each instance. Conditioned on instances, CondInst [30] employs dynamic instance-aware networks. According to the instance's location and size, SOLO [31,32] assigns an instance category to each pixel.

2.3. Panoptic segmentation

Kirillov et al. [19] first proposed Panoptic segmentation, which can be considered a semantic and instance segmentation combination. Panoptic FPN [20] extends Mask R-CNN [11] by adding an independent branch for semantic segmentation and generates panoptic results by a heuristic post-processing step. Later studies [33–43] follow this paradigm, which is called the top-down panoptic segmentation method. UPSNet [34] adopts a panoptic segmentation head to solve the panoptic segmentation via pixel-wise classification. OANet [35] utilizes a spatial ranking module to deal with the occlusion problem between the predicted instances. AUNet [36] exploits the underlying relationship between foreground objects and background contents. BANet [38] models the intrinsic interaction between instance segmentation and semantic segmentation and handles occlusion for panoptic segmentation. BGRNet [39] mines the inter-modular and intramodular relationships between foreground and background classes by incorporating graph structure into the conventional panoptic segmentation network. SOGNet [41] models overlap relationships among instances and resolve them for panoptic segmentation.

On the contrary, other panoptic segmentation methods [44–46] do not need an extra object detector for foreground instance segmentation but a group operation after semantic segmentation to generate instance masks. Based on DeepLabV3+ [8], Panoptic-DeepLab [45] adopts dual-decoder and dual-ASPP structures specific to semantic segmentation and instance segmentation. Axial-DeepLab [46] proposes a position-sensitive axial-attention model to reduce computation complexity and perform attention within a larger or even global region. Pixels votes for the likely regions that contain instance centroids, PCV [44] generates panoptic segmentation results according to FCN-style semantic segmentation and the consensus among pixel-wise votes. These methods are called the bottom-up panoptic segmentation methods.

3. Proposed method

In this paper, we intend to refer to the relationship between two adjacent pixels to achieve fast panoptic segmentation. If a pixel belongs to an instance, we need to judge whether its adjacent pixels belong to the same instance. Specifically, we discuss this relationship in a 3×3 region as shown in Fig. 1, and it can be defined as follow:

$$r_{i,j} = \begin{cases} 1 & \mathbf{i}, \mathbf{j} \in \mathbf{A} \\ 0 & \text{else} \end{cases} \tag{1}$$

where $\mathbf{A}$ represents an instance. If pixel $\mathbf{i}$ and $\mathbf{j}$ belong to the same instance $\mathbf{A}$, the value of the relationship between these two pixels $r_{i,j}$ equals 1, otherwise 0.

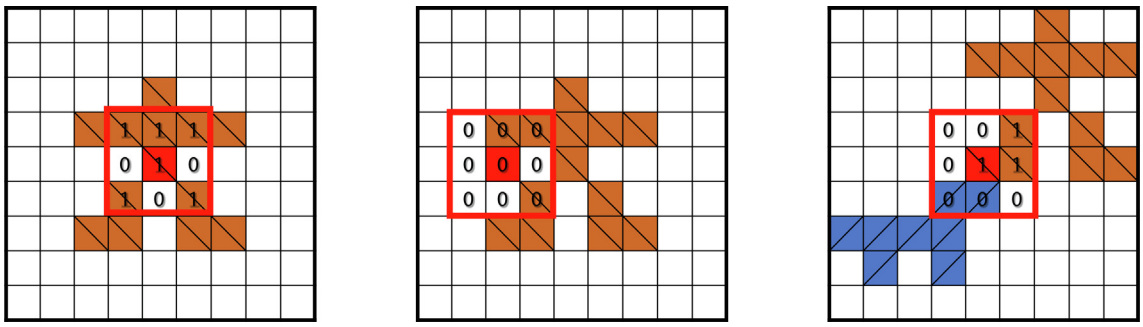

Fig. 1. Examples of the relationship between adjacent pixels. This relationship indicates whether two adjacent pixels belong to the same instance. As we can see from the left example on the left, a red pixel belongs to an instance **A**. Within a 3 × 3 region around this red pixel, some pixels also belong to instance **A** as the red pixel. These values of relationships between these pixels with the red pixel are set to 1 while others are set to 0 because those pixels belong to the background. The middle example shows that a red pixel does not belong to any instance. Therefore, although some pixels within the 3 × 3 region belong to an instance, all the relationship values are set to 0. In the right example, a red pixel belongs to an instance **A**. However, there are some pixels within the 3 × 3 region belonging to another instance **B**. These values of relationships between these pixels with the red pixel are also set to 0.

3.1. Network architecture

Panoptic-DeepLab [45] is a simple, strong, and fast baseline for panoptic segmentation. This paper follows Panoptic-DeepLab [45] and adopts the same shared encoder, decoupled ASPP modules, decoupled decoder modules specific to each task, and task-specific prediction heads. We replace the instance centre prediction head and instance centre regression head with our designed adjacent-pixels relationship head by constructing a graph convolutional layer. In addition, we add a boundary prediction head to predict the boundary of instances. Our objective is to achieve a faster panoptic segmentation while ensuring performance. The overall architecture is shown in Fig. 2.

3.1.1. Encoder-decoder and semantic segmentation head

We utilize the same structure as the shared encoder, decoupled ASPP modules, decoupled decoder modules specific to each task, and semantic segmentation head in Panoptic-DeepLab [45] because we find this design can ensure both good performance and high speed.

3.1.2. Adjacent-pixels relationship head

Graph Convolutional Network [47] is proposed to model relationships in images. Therefore, we adopt a graph convolutional layer to model adjacent-pixels relationships as shown in Fig. 2. Given an adjacency graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ with edges $\mathcal{E}$ among nodes $\mathcal{V}$, where nodes $\mathcal{V}$ represents 9 pixels within a 3 × 3 region and edges $\mathcal{E}$ represents relationships between the centre pixel and other pixels within the 3 × 3 region. Specifically, the calculation process is as follows:

$$r_{i,j'} = \sigma(A_{i,j}\beta(z_i)) \tag{2}$$

where $z_i \in \mathbf{Z} = \{z_1, z_2, \cdots, z_N\}$ represents features of each pixels, which is also the node $\mathcal{V}$ in graph $\mathcal{G}$. And N is the number of pixels, which is also the number of the node $\mathcal{V}$. $\beta(\cdot)$ represents the value convolution as shown in Fig. 2. σ denotes the sigmoid activation function, which limits $r_{i,j'} \in (0,1)$. $A \in \mathbb{R}^{3\times3}$ is the adjacency matrix for defining similarity between adjacent pixels. Therefore, $A_{i,j}$ represents similarity between pixel **i** and pixel **j**, where pixel **i** is the center pixel within the 3 × 3 region and pixel **j** is the adjacent pixel of

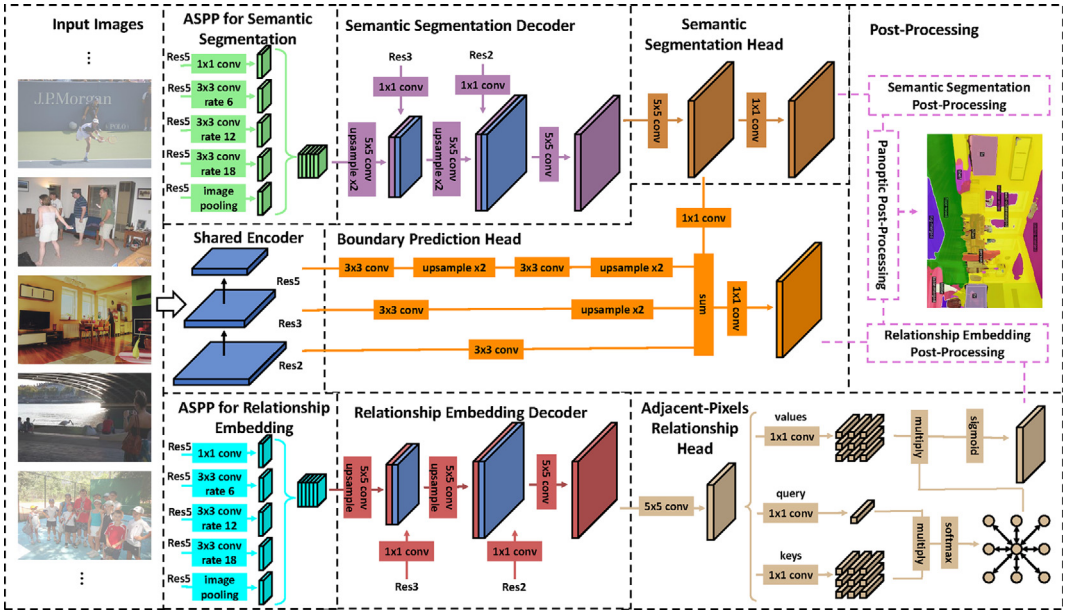

Fig. 2. The overall architecture of our method. Given an input image, the shared encoder is applied to provide feature maps at multi-scales. The independent Atrous Spatial Pyramid Pooling (ASPP) and light-weight decoder are employed for semantic segmentation and relationship embedding. The semantic segmentation head is aimed to predict category labels for both "things" and "stuff" pixels. The adjacent-pixels relationship head outputs relationships between every two adjacent pixels. The boundary prediction is used to output boundary prediction. A parameter-free module is used to generate panoptic segmentation results during inference. The solid arrow indicates that the feature is transferred during training and inference, while the dashed arrow indicates that the feature is only transferred during inference.

pixel **i**. To construct the adjacency matrix A, we define the similarity between two adjacent nodes x_i and x_j by:

$$A_{i,j} = softmax\Big(\theta(z_i)^T \phi(z_j)\Big) \tag{3}$$

where $\theta(\cdot)$ and $\phi(\cdot)$ respectively represent the key convolution and the query convolution as shown in Fig. 2.

In addition, only considering relationships within a 3×3 region cannot model long-range relationships, which will make it difficult for our method to distinguish between two adjacent instances. However, enlarging the region, such as 9×9, will bring larger time and computing resources consumption, resulting in the loss of speed. Thus, we follow the idea of Dilated Convolutions [48] to implement a graph convolutional layer with a dilation rate, which dilation rate is set to 2 in this paper. In this way, we can model a relationship between two non-adjacent pixels while saving time and computing resources. A detailed discussion about the dilation rate is reported in Section 4.4.3.

3.1.3. Boundary prediction head

As shown in Fig. 2, we fuse features from the shared encoder at different scales by element-wise addition after repeated 3×3 convolutions and upsampling operations. Besides, we integrate semantic features into the boundary prediction head by a simple 1×1 convolution because rich high-level semantic information is beneficial to boundary prediction. The last 1×1 convolution is used to predict a predicted boundary map. This tiny module does not bring much computational cost while helping our method enhance supervision information and distinguish adjacent objects.

3.2. Training and inference

3.2.1. Joint training

We use the relationship between adjacent pixels instead of being supervised by instance-wise annotations during training. Binary cross-entropy loss is used for the relationship embedding head. The potential problem is that, for most images, there are more background pixels than foreground pixels, which will lead to negative samples dominating the training procedure. Therefore, we apply online hard example mining (OHEM) to balance negative and positive samples in the relationship embedding head. We use mean-squared and dice loss [49] for the boundary prediction head to supervise the boundary learning because dice loss [49] can alleviate the class-imbalance problem. For semantic segmentation head, we use traditional cross-entropy loss. The multi-task loss function $\mathscr{L}$ is defined as follow:

$$\begin{aligned}
\mathscr{L} &= \mathscr{L}_{Semantic} + \mathscr{L}_{Relationship} + \mathscr{L}_{Boundary} \\
\mathscr{L}_{Semantic} &= \mathscr{L}_{CE}(\mathscr{F}_S(\mathbf{X}), \mathscr{GT}_S) \\
\mathscr{L}_{Relationship} &= \mathscr{L}_{BCE}(\mathscr{F}_R(\mathbf{X}), \mathscr{GT}_R) \\
\mathscr{L}_{Boundary} &= \mathscr{L}_{MSE}(\mathscr{F}_B(\mathbf{X}), \mathscr{GT}_B) + \mathscr{L}_{DICE}(\mathscr{F}_B(\mathbf{X}), \mathscr{GT}_B)
\end{aligned} \tag{4}$$

where $\mathscr{L}_{Semantic}$ represents semantic segmentation loss, $\mathscr{L}_{Boundary}$ is for boundary prediction head, and $\mathscr{L}_{Relationship}$ is for adjacent-pixels relationship head. $\mathscr{L}_{CE}$ denotes the cross-entropy loss, $\mathscr{L}_{BCE}$ denotes the binary cross-entropy loss, $\mathscr{L}_{MSE}$ denotes the mean-squared loss, while $\mathscr{L}_{DICE}$ denotes the dice loss. $\mathscr{GT}_S, \mathscr{GT}_R$ and $\mathscr{GT}_B$ are ground truths for semantic segmentation, adjacent pixels relationships prediction and boundary prediction. $\mathscr{F}_S, \mathscr{F}_R$ and $\mathscr{F}_B$ are semantic segmentation part (including ASPP for semantic segmentation, semantic decoder and semantic segmentation head), Relationship embedding part (including ASPP for relationship embedding, relationship embedding decoder and relationship embedding head) and boundary prediction head respectively. $\mathbf{X}$ is the input feature.

3.2.2. Panoptic inference

Panoptic segmentation aims to generate a unified segmentation mask for foreground "things" and background "stuff", requiring the model to assign instance ids and category labels to all pixels. If a pixel belongs to the background, its instance id should be ignored. To generate qualified panoptic segmentation masks, we design a parameter-free fusion step. Firstly, for every two adjacent pixels, we determine whether they belong to the same instance, and predict a corresponding relationship value. Because the predicted relationship value between two adjacent pixels **i** and **j** $r_{i,j'}$ is in [0,1], the threshold is set to 0.5. If $r_{i,j'} \geqslant 0.5$, pixel **i** and **j** are considered to belong to the same instance, otherwise not. Next, we generate a class-agnostic instance mask using these relationships. However, this procedure is commonly modeled as a time-consuming graph partition problem. In order to save time, we replace it with a simpler and faster method. For pixel **i**, we assume that if there are more than n pixels (we set n to 5 in this paper) within its corresponding 3×3 region belonging to the same instance, we consider pixel **i** as a foreground pixel. So we can obtain a binary foreground mask, and this segmentation problem is transformed into a problem of finding the connected domain of a binary image. We find that this operation saves a lot of calculation time while ensuring performance. Related ablation study will be discussed in Section 4.4.2. Then, we find that if multiple objects are together, our method cannot distinguish each object well. So we introduce boundary information to distinguish adjacent objects. Specifically, we use the boundary information to suppress the value of relationship $r'_{i,j}$ on the boundary as follow:

$$r''_{i,j} = r'_{i,j}(1 - \sigma(b_i)) \tag{5}$$

where b_i represents the output of boundary prediction head on pixel **i** and σ denotes the sigmoid activation function. Finally, we assign a corresponding category label for each class-agnostic instance. If an instance corresponds to multiple category labels, we divide this instance into multiple instances to ensure an instance corresponds to a category label. Of course, any region under an area threshold will be removed whether foreground or background. Different area threshold settings will be discussed in Section 4.4.4. Going through the above steps, we can obtain a coherent scene segmentation.

4. Experiments

4.1. Datasets and metrics

4.1.1. MS COCO

The MS COCO dataset [22] is one of the most important computer vision datasets widely used in object detection, instance segmentation, keypoint detection, and panoptic segmentation. It consists of 80 categories with pixel-wise instance mask annotation and 53 categories with pixel-wise semantic mask annotation. Moreover, there are 118 K images for `train` subset, 5 K images for `val` subset, 20 K images for `test-dev` subset, and 20 K images for `test-challenge` subset.

4.1.2. Metrics

We use Panoptic Quality (PQ) [19] as the metric for panoptic segmentation. It is defined by recognition quality (RQ) and segmentation quality (SQ):

$$\begin{aligned}
\mathrm{PQ} = &\underbrace{\frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP|}}_{\text{segmentationquality(SQ)}} \\
&\times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognitionquality(RQ)}}
\end{aligned} \tag{6}$$

where IoU(p, g) represents the intersection-over-union between a predicted panoptic segmentation mask p and the ground truth g, TP are matched pairs of segments, FP is the unmatched predictions, and FN is the unmatched ground truth segments. Additionally, PQ^{th} (average over "things" categories), and PQ^{st} (average over "stuff" categories) are also reported.

4.2. Implementation details

Our models are trained using PyTorch on 8 GPUs (NVIDIA GeForce RTX 3090). ResNet [50] with Deformable Convolution [51] is used as our backbone network. In particular, we optimize our model with ADAM [52] without weight decay. For MS COCO dataset [22], we resize the input images to 960 pixels at the longest side and train our models with crop size 640 × 640 with the batch size is set to 64. The initial learning rate is set to 0.0005 while training iterations is 200 K. All our experiments are conducted on Detectron2 [53].

4.3. Quantitative results

As we can see from Table 1, we compared our method with previous state-of-the-art panoptic segmentation methods on COCO2017 `test-dev` subset. With ResNet-50 and ResNet-101, our method can respectively achieve 34.6 and 35.7 on *PQ*. Obviously, our method is not the best in performance, but we are more concerned about the trade-off between inference time, model size, computational complexity and performance. Table 2 demonstrates that our method achieves the fastest inference with competitive performance compared with other panoptic segmentation methods. "*" represents this result is reproduced in our environment settings. Speed means the time that inference an image, Param means the total number of model parameters, and GFLOPs means floating-point operands, which is used to measure the complexity of model. Especially, compared with Panoptic-DeepLab [45], our method has fewer calculations when the parameters are close.

In Fig. 3, we compare our visual results with Panoptic FPN [20] and Panoptic-DeepLab [45] on COCO2017 `val` subset. The first column is for input images, the second is ground truth, and the last three columns are the visual results of Panoptic FPN [20], Panoptic-DeepLab [45], and our method, respectively.

In addition, we analyzed the time consumed by each part of our method, and the result is shown in Table 3. We use ResNet-50 as the backbone for a fair comparison. Compared with Panoptic-DeepLab [45], our post-processing spends less time. We attribute our fast panoptic segmentation to our fully convolutional network structure and simple panoptic segmentation inference process.

4.4. Ablation study

In this paper, we perform ablation study on COCO2017 `val` subset with ResNet-50 as our backbone network.

4.4.1. Component analysis

In this ablation study, we discuss the design of network architecture. As shown in Table 4, we first analyze the choice between traditional convolution graph convolution, and the result shows that the graph convolution is a better choice. Besides, we only use adjacency matrix to determine whether two adjacent pixels belong to the same instance, and the result also shows our graph convolution have a better performance. Next, the result proves that adding a boundary prediction head can help our method to obtain a 0.5 gain in *PQ*. Then, we compared whether to join dice loss [49] in the boundary prediction head. The experiment results show that this operation is effective because dice loss [49] is insensitive to the number of positive/negative pixels while boundary pixels are much smaller than non-boundary loss. Finally, to balance positive samples and negative samples, we adopt OHEM, and the experiment result shows this is necessary.

4.4.2. Inference operation

In this ablation study, we discuss the inference operation. We assume that if there are more than n pixels within a 3 × 3 region recognized to belong the same instance, we consider the centre pixel as a foreground pixel. As shown in Table 5, the results show that if we set n = 5, the performance achieves the best.

4.4.3. Dilation rate

This ablation study compared different dilation rates in graph convolutional layer. As shown in Table 6, dilation rate is 1 means a normal graph convolutional layer. The result shows when dilation rate is set to 2, our method can achieve a balance between performance and speed.

4.4.4. Area threshold settings

In this ablation study, we compared the effects of different area threshold settings on the experimental results. As we can see in Table 7, PQ^{th} achieves the best when foreground area threshold is set to 512. Similarly, according to Table 8, PQ^{st} is the best when background area threshold is set to 5,120.

4.5. Experiments on Cityscapes Dataset

The Cityscapes Dataset [23] is designed for understanding urban street scenes and is widely applied in tasks such as semantic

Table 1
Comparison with state-of-the-art methods on COCO2017 `test-dev` subset.

Method	Backbone	*PQ*	*SQ*	*RQ*	PQ^{th}	SQ^{th}	RQ^{th}	PQ^{st}	SQ^{st}	RQ^{st}
top-down panoptic segmentation methods										
JSIS [54]	ResNet-50	27.2	71.9	35.9	29.6	71.6	39.4	23.4	72.3	30.6
FPSNet [55]	ResNet-50-FPN	29.8	74.3	37.9	33.1	76.1	41.6	24.9	71.5	32.4
SSPS [56]	ResNet-50-FPN	32.6	74.3	42.0	35.0	74.8	44.8	29.0	73.6	30.6
TASCNet [57]	ResNet-50-FPN	40.7	78.5	50.1	47.0	80.6	57.1	31.0	75.3	39.6
Panoptic FPN [20]	ResNet-101-FPN	**40.9**	78.5	**50.1**	**48.3**	**81.7**	**58.3**	29.7	73.7	37.3
bottom-up panoptic segmentation methods										
DeeperLab [58]	Xception-71	34.3	77.1	43.1	37.5	77.5	46.8	29.6	76.4	37.4
Panoptic-DeepLab [45]	ResNet-50	35.2	-	-	-	-	-	-	-	-
Panoptic-DeepLab [45]	Xception-71	38.9	-	-	-	-	-	-	-	-
SSAP [18]	ResNet-101	36.9	**81.1**	46.0	40.1	81.6	48.5	33.2	**79.7**	40.8
PCV [44]	ResNet-50	37.7	77.8	47.3	40.7	78.7	50.7	33.1	76.3	42.0
ours	ResNet-50	34.6	79.0	42.6	35.0	80.0	42.8	33.9	77.6	42.4
ours	ResNet-101	35.7	79.7	43.8	36.0	80.6	43.9	**35.1**	78.3	**43.5**

Table 2
Comparison with state-of-the-art methods on COCO2017 `val` subset. "*" represents this result is reproduced in our environment settings.

Method	Backbone	Input Size	*PQ*	Speed(*ms*)	Params(M)	GFLOPs
SSPS [56]	ResNet-50-FPN	576 × 864	32.4	42.6	-	-
DeeperLab [58]	Xception-71	640 × 640	33.8	119.0	-	-
PCV [44]	ResNet-50	1333 × 800	37.5	176.5	-	-
RealTimePan [59]	ResNet-50-FPN	1333 × 800	37.1	63.0	-	-
UPSNet [34]	ResNet-50-FPN	1333 × 800	**42.5**	171.0	-	-
Panoptic-DeepLab [45]*	ResNet-50	640 × 640	35.1	52.3	**30.3**	82.5
Panoptic FPN [20]*	ResNet-50-FPN	1333 × 800	39.4	62.9	48.5	231.5
Panoptic–CenterMask [15]*	VoVNet-39-FPN	1333 × 800	39.7	69.8	99.2	515.8
Panoptic-FCN [60]*	ResNet-50-FPN	1333 × 800	41.1	80.6	37.0	251.0
ours	ResNet-50	640 × 640	34.0	**37.7**	31.5	**52.8**

Fig. 3. Visual examples of panoptic segmentation with previous state-of-the-art methods on COCO2017 `val` subset.

Table 3
Analysis about the inference time consumed by each part.

	Time(*ms*)					
Panoptic-DeepLab [45]	Backbone	Sem Part		Inst Part	Post Processing	All
	26.5	1.8		3.6	20.9	52.8
ours	Backbone	Sem Part	Rela Part	Boun Part	Post Processing	All
	30.6	2.2	2.8	1.2	0.9	37.7

Table 4
Component analysis of our method on COCO2017 `val`.

Convolution	Adjacency Matrix	Graph Convolution	Boun Pred Head	Dice Loss	OHEM	*PQ*	*SQ*	*RQ*
√						32.6	77.9	40.4
	√					31.8	78.1	39.5
		√				33.1	**78.6**	40.9
		√	√			33.6	78.5	41.4
		√	√	√		33.8	78.3	41.8
		√	√	√	√	**34.0**	78.4	**41.9**

Table 5
Ablation study about inference operation on COCO2017 `val`.

Number of Pixels	*PQ*	*SQ*	*RQ*
1	33.5	78.2	41.3
3	33.6	78.3	41.5
5	**34.0**	**78.4**	**41.9**
7	33.3	77.5	41.4
9	32.4	76.5	40.8

Table 6
Ablation study about dilation rate on COCO2017 `val` subset.

Dilation Rate	*PQ*	Speed(*ms*)
1	33.7	**35.1**
2	34.0	37.7
3	34.0	42.2
4	**34.1**	46.3

Table 7
Ablation study about foreground area threshold on COCO2017 `val` subset.

Threshold	PQ^{th}	SQ^{th}	RQ^{th}
128	33.8	78.7	41.5
256	**34.1**	78.9	**41.8**
512	34.0	79.2	41.5
768	33.5	**79.6**	40.7

Table 8
Ablation study about background area threshold on COCO2017 `val` subset.

Threshold	PQ^{st}	SQ^{st}	RQ^{st}
1,280	32.3	77.1	40.5
2,560	33.4	77.3	41.8
5,120	**33.8**	**77.6**	**42.1**
7,680	33.7	78.1	41.7

segmentation, instance segmentation and panoptic segmentation. It mainly includes 8 foreground categories and 11 background categories. There are 5,000 images with high-quality annotations, where 2,975 images are for `train` subset, 1,525 images are for `test` subset, and 500 images are for `val` subset.

Experiments settings are similar to experiments on MS COCO [22] dataset, and ResNet-50 is used as our backbone network. The initial learning rate is set to 0.001, while training iterations is 90 K. During training, we first resize the image to 4096 on the longer side, and the shorter side is randomly sampled from [512,2048]. Then, we crop the image to 1024 $\times$ 2048. The batch size is set to 32. At the inference time, the input size is set to 512 $\times$ 1024.

The experiment results are shown in Table 9. "*" represents this result is reproduced in our environment settings. Our method can achieve 55.1 in Cityscapes `val` subset in *PQ* with ResNet-50, while the inference time is 156.8 ms. Considering both performance and inference time, our method shows competitiveness with other methods. Visualization results are shown in Fig. 4.

5. Conclusion

In this paper, we propose a simple yet effective method for panoptic segmentation. During training, instead of directly supervising the instance mask for each instance, we leverage the relationships between each adjacent pixel. During inference, we predict the relationship between two adjacent pixels to generate class-agnostic instance masks. According to predicted category labels from the semantic segmentation head, we can obtain unified panoptic segmentation masks. Nevertheless, our method cannot distinguish each object precisely if multiple objects are piling with the same semantic category, which we need to solve in future research. We hope our method can provide some help for future research.

Table 9
Results on Cityscapes `val` subset. "*" represents this result is reproduced in our environment settings.

Method	Backbone	*PQ*	Speed(*ms*)	Device
DIN [61]	ResNet-50	53.8	-	-
DeeperLab [58]	Xception-71	56.5	308.6	Tesla V100
PCV [44]	ResNet-50	54.2	182.8	GeForce RTX 1080 Ti
Panoptic FPN [20]*	ResNet-50-FPN	57.0	461.3	GeForce RTX 3090
Panoptic-DeepLab [45]*	ResNet-50	**58.6**	261.3	GeForce RTX 3090
ours	ResNet-50	55.1	**156.8**	GeForce RTX 3090

Fig. 4. Visual examples of panoptic segmentation on Cityscapes `val` subset.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was support in part by National Key R&D Program of China (2021ZD0112001) and Nation Natural Science Foundation of China (No. 61831005, 61871087, 61971095).

References

[1] J. Liu, X. Xu, Y. Shi, C. Deng, M. Shi, RELAXNet: Residual Efficient Learning and Attention Expected Fusion Network for Real-Time Semantic Segmentation, Neurocomputing 474 (14) (2021) 115–127.
[2] S. Yi, J. Li, X. Liu, X. Yuan, CCAFFMNet: Dual-Spectral Semantic Segmentation Network with Channel-Coordinate Attention Feature Fusion Module, Neurocomputing.
[3] E. Shelhamer, J. Long, T. Darrell, Fully Convolutional Networks for Semantic Segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 640–651.
[4] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[5] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid Scene Parsing Network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 2881–2890.
[6] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 834–848.
[7] L.C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking Atrous Convolution for Semantic Image Segmentation (2017). arXiv:1706.05587.
[8] L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, in: Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018, pp. 801–818.
[9] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual Attention Network for Scene Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 3146–3154.
[10] C. Shang, H. Li, F. Meng, H. Qiu, Q. Wu, L. Xu, K.N. Ngan, Instance-Level Context Attention Network for Instance Segmentation, Neurocomputing 472 (1) (2022) 124–137.
[11] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2020) 386–397.
[12] Z. Huang, L. Huang, Y. Gong, C. Huang, X. Wang, Mask Scoring R-CNN, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 6409–6418.
[13] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path Aggregation Network for Instance Segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 8759–8768.
[14] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C.C. Loy, D. Lin, Hybrid Task Cascade for Instance Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 4974–4983.
[15] Y. Lee, J. Park, CenterMask: Real-Time Anchor-Free Instance Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2020, pp. 13906–13915.
[16] D. Bolya, C. Zhou, F. Xiao, Y.J. Lee, YOLACT: Real-Time Instance Segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 2019, pp. 9157–9166.
[17] D. Neven, B.D. Brabandere, M. Proesmans, L.V. Gool, Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Bandwidth, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 8837–8845.
[18] N. Gao, Y. Shan, Y. Wang, X. Zhao, Y. Yu, M. Yang, K. Huang, SSAP: Single-Shot Instance Segmentation with Affinity Pyramid, IEEE Trans. Circuits Syst. Video Technol. 31 (2) (2020) 661–673.
[19] A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollár, Panoptic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 9404–9413.
[20] A. Kirillov, R. Girshick, K. He, P. Dollár, Panoptic Feature Pyramid Networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 6399–6408.
[21] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149.
[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common Objects in Context, in: Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 2014, pp. 740–755.
[23] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The Cityscapes Dataset for Semantic Urban Scene Understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 3213–3223.
[24] Z. Cai, N. Vasconcelos, Cascade R-CNN: High Quality Object Detection and Instance Segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 43 (5) (2021) 1483–1498.
[25] X. Zhang, H. Li, F. Meng, Z. Song, L. Xu, Segmenting Beyond the Bounding Box for Instance Segmentation, IEEE Trans. Circuits Syst. Video Technol.
[26] Z. Tian, C. Shen, H. Chen, T. He, FCOS: Fully Convolutional One-Stage Object Detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 2019, pp. 9627–9636.
[27] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2020) 318–327.
[28] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, P. Luo, PolarMask: Single Shot Instance Segmentation with Polar Representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2020, pp. 12193–12202.
[29] J. Liang, N. Homayounfar, W.C. Ma, Y. Xiong, R. Hu, R. Urtasun, PolyTransform: Deep Polygon Transformer for Instance Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2020, pp. 9131–9140.
[30] Z. Tian, C. Shen, H. Chen, Conditional Convolutions for Instance Segmentation, in: Proceedings of the European Conference on Computer Vision, Glasgow, UK, 2020, pp. 282–298.
[31] X. Wang, T. Kong, C. Shen, Y. Jiang, L. Li, SOLO: Segmenting Objects by Locations, in: Proceedings of the European Conference on Computer Vision, Glasgow, UK, 2020, pp. 649–665.
[32] X. Wang, R. Zhang, T. Kong, L. Li, C. Shen, SOLOv2: Dynamic and Fast Instance Segmentation, in: Advances in Neural Information Processing Systems, Virtual, 2020, pp. 17721–17732.
[33] F. Jie, Q. Nie, M. Li, M. Yin, T. Jin, Atrous Spatial pyramid Convolution for Object Detection with Encoder-Decoder, Neurocomputing 464 (13) (2021) 107–118.
[34] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, R. Urtasun, UPSNet: A Unified Panoptic Segmentation Network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2019, pp. 8818–8826.
[35] H. Liu, C. Peng, C. Yu, J. Wang, X. Liu, G. Yu, W. Jiang, An End-to-End Network for Panoptic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2019, pp. 6172–6181.
[36] Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, X. Wang, Attention-Guided Unified Network for Panoptic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2019, pp. 7026–7035.
[37] Q. Li, X. Qi, P.H. Torr, Unifying Training and Inference for Panoptic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2020, pp. 13320–13328.
[38] Y. Chen, G. Lin, S. Li, O. Bourahla, Y. Wu, F. Wang, J. Feng, M. Xu, X. Li, BANet: Bidirectional Aggregation Network with Occlusion Handling for Panoptic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2020, pp. 3793–3802.
[39] Y. Wu, G. Zhang, Y. Gao, X. Deng, K. Gong, X. Liang, L. Lin, Bidirectional Graph Reasoning Network for Panoptic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2020, pp. 9080–9089.
[40] J. Lazarow, K. Lee, K. Shi, Z. Tu, Learning Instance Occlusion for Panoptic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2020, pp. 10720–10729.
[41] Y. Yang, H. Li, X. Li, Q. Zhao, J. Wu, Z. Lin, SOGNet: Scene Overlap Graph Network for Panoptic Segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 2020, pp. 12637–12644.
[42] R. Mohan, A. Valada, EfficientPS: Efficient Panoptic Segmentation, Int. J. Comput. Vision 129 (5) (2021) 1551–1579.
[43] L. Porzi, S.R. Bulò, A. Colovic, P. Kontschieder, Seamless Scene Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2019, pp. 8277–8286.
[44] H. Wang, R. Luo, M. Maire, G. Shakhnarovich, Pixel Consensus Voting for Panoptic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2020, pp. 9464–9473.
[45] B. Cheng, M.D. Collins, Y. Zhu, T. Liu, T.S. Huang, H. Adam, L.C. Chen, Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2020, pp. 12475–12485.
[46] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, L.C. Chen, Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation, in: Proceedings of the European Conference on Computer Vision, Glasgow, UK, 2020, pp. 108–126.
[47] T.N. Kipf, M. Welling, Semi-Supervised Classification with Graph Convolutional Networks, in: Proceedings the International Conference on Learning Representations, Toulon, France, 2017, pp. 1–14.

[48] F. Yu, V. Koltun, Multi-Scale Context Aggregation by Dilated Cnvolutions, in: Proceedings the International Conference on Learning Representations, San Juan, Puerto Rico, 2016, pp. 1–13.
[49] F. Milletari, N. Navab, S.-A. Ahmadi, V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, in: Proceedings the International Conference on 3D Vision, Stanford, CA, USA, 2016, pp. 565–571.
[50] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2016, pp. 770–778.
[51] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable Convolutional Networks, in: Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017, pp. 764–773.
[52] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: Proceedings the International Conference on Learning Representations, San Diego, CA, USA, 2015, pp. 1–15.
[53] Y. Wu, A. Kirillov, F. Massa, W.Y. Lo, R. Girshick, Detectron2, URL: https://github.com/facebookresearch/detectron2 (2019).
[54] D. De Geus, P. Meletis, G. Dubbelman, Panoptic Segmentation with a Joint Semantic and Instance Segmentation Network (2018). arXiv:1809.02110.
[55] D. de Geus, P. Meletis, G. Dubbelman, Fast Panoptic Segmentation Network, IEEE Robot. Autom. Lett. 5 (2) (2020) 1742–1749.
[56] M. Weber, J. Luiten, B. Leibe, Single-Shot Panoptic Segmentation, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, NV, USA, 2020, pp. 8476–8483.
[57] J. Li, A. Raventos, A. Bhargava, T. Tagawa, A. Gaidon, Learning to Fuse Things and Stuff (2018). arXiv:1812.01192.
[58] T.J. Yang, M.D. Collins, Y. Zhu, J.J. Hwang, T. Liu, X. Zhang, V. Sze, G. Papandreou, L.C. Chen, DeeperLab: Single-Shot Image Parser (2019). arXiv:1902.05093.
[59] R. Hou, J. Li, A. Bhargava, A. Raventos, V. Guizilini, C. Fang, J. Lynch, A. Gaidon, Real-time panoptic segmentation from dense detections, IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual (2020) 8523–8532.
[60] Y. Li, H. Zhao, X. Qi, L. Wang, Z. Li, J. Sun, J. Jia, Fully Convolutional Networks for Panoptic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2021, pp. 214–223.
[61] A. Arnab, P.H.S. Torr, Pixelwise Instance Segmentation with a Dynamically Instantiated Network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 441–450.

Xiaoliang Zhang received his B.E. degree in Electronic Science and Technology at the University of Electronic Science and Technology of China (UESTC) in 2015, and the M.D. degree in Electronics and Communications Engineering in Southwest Jiaotong University (SWJTU) in 2018. Now he is working for his Ph.D. degree in Signal and Information Processing at UESTC. His main research interests include computer vision and machine learning, especially the application of deep learning on instance segmentation.

Hongliang Li (Senior Member, IEEE) received his Ph.D. degree in Electronics and Information Engineering from Xi'an Jiaotong University, China, in 2005. From 2005 to 2006, he joined the visual signal processing and communication laboratory (VSPC) of the Chinese University of Hong Kong (CUHK) as a Research Associate. From 2006 to 2008, he was a Postdoctoral Fellow at the same laboratory in CUHK. He is currently a Professor in the School of Electronic Engineering, University of Electronic Science and Technology of China. His research interests include image segmentation, object detection, image and video coding, visual attention, and multimedia processing.
Dr. Li has authored or co-authored numerous technical articles in well-known international journals and conferences. He is a co-editor of a Springer book titled "Video segmentation and its applications". Dr. Li is involved in many professional activities. He received the 2019 and 2020 Best Associate Editor Awards for IEEE Transactions on Circuits and Systems for Video Technology (TCSVT). He served as a Technical Program Chair for VCIP2016 and PCM2017, General Chairs for ISPACS 2017 and ISPACS2010, a Publicity Chair for IEEE VCIP 2013, a Local Chair for the IEEE ICME 2014, and a TPC Member for a number of international conferences, such as, ICME 2013, ICME 2012, ISCAS 2013, PCM 2007, PCM 2009, and VCIP 2010. He was an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology. He serves as an Associate Editor of Journal on Visual Communication and Image Representation, and an Area Editor of Signal Processing: Image Communication (Elsevier Science).

Lanxiao Wang received her B.E. degree in Electronics Information Engineering at the University of Electronic Science and Technology of China (UESTC) in 2019. Now she is working for her Ph.D. degree under the supervision of Prof. Li. in Information and Communication Engineering at UESTC, Chengdu, China.
Her main research interests include computer vision and machine learning, especially the application of deep learning on scene analysis and multimodal representation learning.

Haoyang Cheng received the B.E. degree in electronic and information engineering from the University of Electronic Science and Technology of China (UESTC) in 2019, where he is currently pursuing the Ph.D. degree in information and communication engineering. His research interests include self-supervised learning, contrastive learning and continual learning.

Heqian Qiu received the B.Sc. degree in electronic information science and technology from Shanxi Datong University, Datong, China, in 2015. She is currently pursuing the Ph.D. degree under the supervision of Prof. H. Li., University of Electronic Science and Technology of China, Chengdu, China.
Her research interests include object detection, multimodal representative learning, computer vision, and machine learning.

Wenzhe Hu received his B.E. degree in University of Electronic Science and Technology of China (UESTC) in 2020. He is pursuing his Master degree in information and communication engineering in USETC. He is currently working on multimodal interpretation and representation in computer vision.

Fanman Meng (Member, IEEE) received the Ph.D. degree in signal and information processing from the University of Electronic Science and Technology of China, Chengdu, China, in 2014. From 2013 to 2014, he was a Research Assistant with the Division of Visual and Interactive Computing, Nanyang Technological University, Singapore. He is currently Professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China. He has authored or co-authored numerous technical articles in well-known international journals and conferences. His current research interests include image segmentation and object detection.
Dr. Meng is a member of the IEEE Circuits and Systems Society. He was a recipient of the Best Student Paper Honorable Mention Award at the 12th Asian Conference on Computer Vision, Singapore, in 2014, and the Top 10% Paper Award at the IEEE International Conference on Image Processing, Paris, France, in 2014.

Qingbo Wu (Member, IEEE) received the Ph.D. degree in signal and information processing from the University of Electronic Science and Technology of China in 2015. From February 2014 to May 2014, he was a Research Assistant with the Image and Video Processing (IVP) Laboratory, Chinese University of Hong Kong. From October 2014 to October 2015, he served as a Visiting Scholar with the Image and Vision Computing (IVC) Laboratory, University of Waterloo. He is currently an Associate Professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China. His research interests include image/video coding, quality evaluation, perceptual modeling and processing.