Paper List:

- Sematic Segmentation
  1. Semantic Segmentation by Early Region Proxy
  2. Deep Hierarchical Semantic Segmentation
  3. High Quality Segmentation for Ultra High-resolution Images
  4. Masked-Attention Mask Transformer for Universal Image Segmentation
  5. Rethinking Semantic Segmentation: A Prototype View

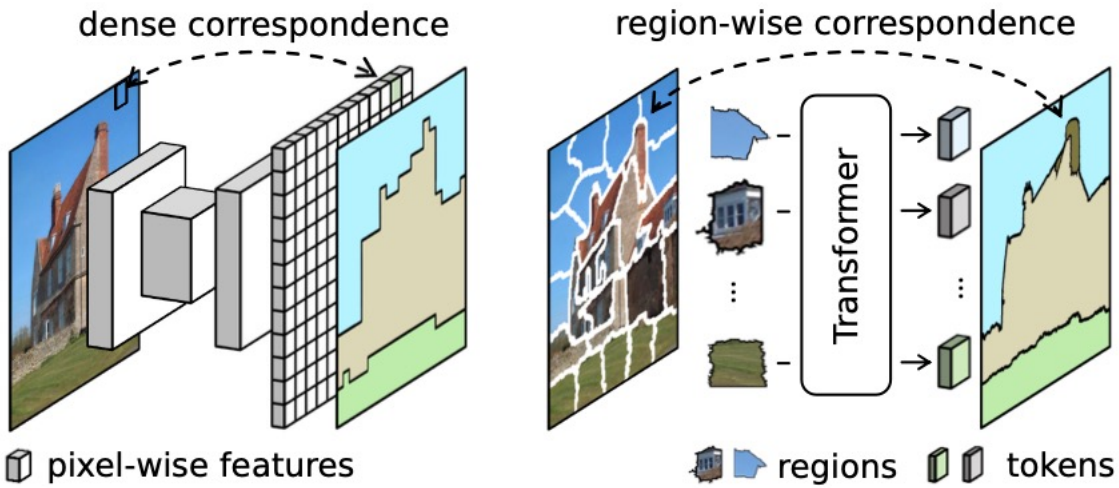
- Panoptic Segmentation
  1. Panoptic, Instance and Semantic Relations: A Relational Context Encoder to Enhance Panoptic Segmentation
  2. CMT-DeepLab: Clustering Mask Transformers for Panoptic Segmentation
  3. Panoptic SegFormer: Delving Deeper Into Panoptic Segmentation With Transformers

Semantic Segmentation

Semantic Segmentation by Early Region Proxy

Yifan Zhang Bo Pang Cewu Lu*
Shanghai Jiao Tong University
{zhangyf_sjtu, pangbo, lucewu}@sjtu.edu.cn

Semantic Segmentation



Pilot experiment (superpixels)

backbone	method	#params.	FLOPs	ADE20K	Cityscapes
ViT-Ti/16	Baseline	5.7M	3.8G	39.0 / 37.8	72.3 / 68.1
	Ours	-	-	**40.9** (+1.9)	**74.1** (+1.8)
ViT-S/16	Baseline	22.0M	14.9G	45.4 / 44.2	**76.1** / 71.8
	Ours	-	-	**46.0** (+0.6)	75.9 (−0.2)
ViT-B/16	Baseline	86.6M	58.8G	47.1 / 45.6	**78.5** / 75.1
	Ours	-	-	**47.3** (+0.2)	77.3 (−1.2)

Classify regions

$$\hat{\mathbf{y}}_i[c] = \frac{|\{\mathbf{p} \in \mathbf{s}_i \mid \hat{y}(\mathbf{p}) = c\}|}{|\mathbf{s}_i|}, c \in \{0, 1, \ldots, K-1\} \quad (1)$$

Motivation:

- limited context
- **coarse prediction**
  1. induced by the inflexible *regular* layout of network features

Semantic Segmentation

Pipeline

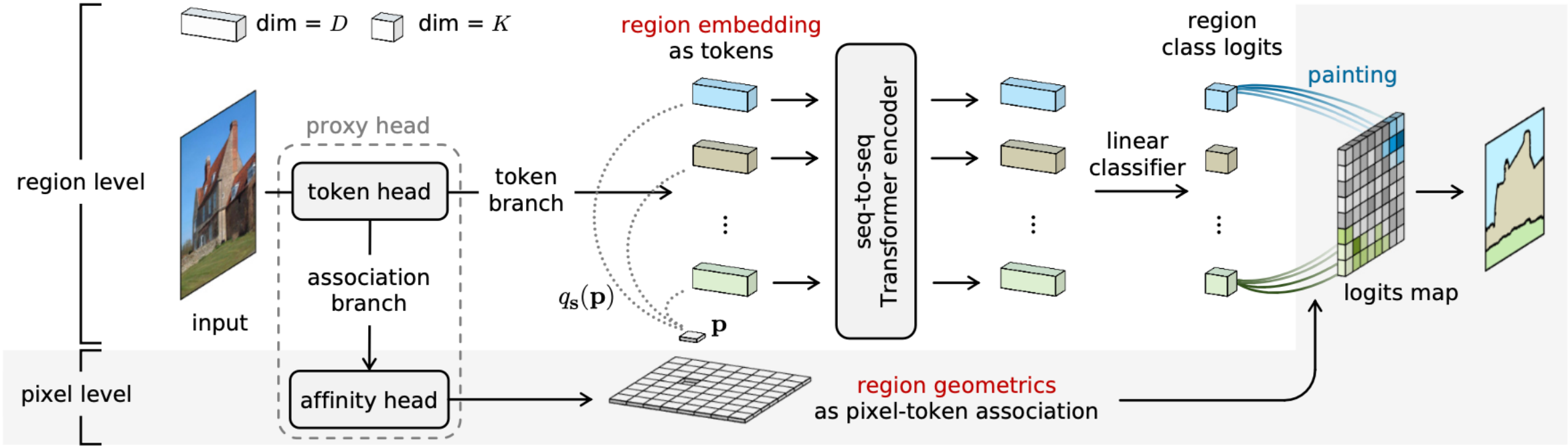
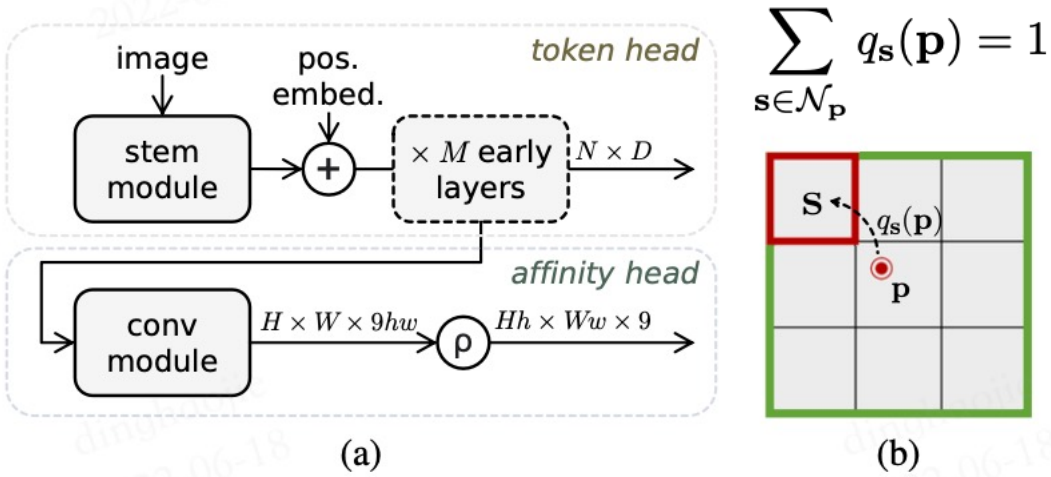


Figure 3. **Overview of our RegProxy approach.** The sequence-to-sequence Transformer encoder computes on region embeddings in the form of tokens, which serve as *proxies* of specific regions whose geometrics are described by the class-agnostic pixel-token association. Notably, we model global context completely at region-level without any typical "feature map". The region embedding and its geometrical description are jointly learned using the *proxy head*. A single linear classifier is adopted for *per-region* prediction. The region class logits are simply "painted" to the output plane according the corresponding region geometrics to yield final segmentation result.

Semantic Segmentation

Region Proxy

- H x W = N → number of Tokens
- (h, w) → Patch Size
- p → per pixel
- X M early layers → from ViT



$$\sum_{\mathbf{s}\in\mathcal{N}_\mathbf{p}} q_\mathbf{s}(\mathbf{p}) = 1$$

Figure 4. **More details of the RegProxy model.** (a) Illustration of the proxy head, where + stands for element-wise addition and ρ stands for reshape and rearrange of dimensions. (b) Describing region geometrics by local pixel-region association.

Prediction

$$\mathbf{Y} = \left[\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{N-1}\right]^\top \in \mathbb{R}^{N\times K},$$

$$\mathbf{Y}'[u,v] = \sum_{\mathbf{s}\in\mathcal{N}_\mathbf{p}} \mathbf{y}(\mathbf{s}) \cdot q_\mathbf{s}(\mathbf{p}),$$

Loss

- Cross Entropy

Semantic Segmentation

method	FLOPs	#params.	ADE20K (SS/MS)		Cityscapes (SS/MS)	
ViT-Ti/16	3.8G	5.7M	39.0	39.8	72.3	74.1
+Mask.T	+1.0G	+1.0M	38.1 (−0.9)	38.8 (−1.0)	-	-
+Ours	+0.1G	+0.1M	**42.1** (+3.1)	**43.1** (+3.3)	76.5 (+4.2)	77.7 (+3.6)
ViT-S/16	14.9G	22.0M	45.4	45.9	76.1	78.0
+Mask.T	+4.2G	+4.1M	45.3 (−0.1)	46.9 (+1.0)	-	-
+Ours	+0.2G	+0.2M	**47.6** (+2.2)	**48.4** (+2.5)	79.8 (+3.7)	81.5 (+3.5)
ViT-B/16	58.8G	86.6M	47.1	48.1	78.5	80.5
+UperNet	+336G	+57.6M	47.9 (+0.8)	49.5 (+1.4)	79.6 (+1.1)	80.9 (+0.4)
+Mask.T	+17.1G	+16.0M	48.7 (+1.6)	50.1 (+2.0)	-	80.6 (+0.1)
+Ours	+0.7G	+0.7M	**49.8** (+2.7)	**50.5** (+2.4)	**80.9** (+2.4)	**82.2** (+1.7)
ViT-L/16	325.0G	304.3M	50.7	51.8	78.4	80.7
+Mask.T	+44.5G	+28.5M	51.8 (+1.1)	**53.6** (+1.8)	-	81.3 (+0.6)
+Ours	+0.9G	+1.8M	**52.9** (+2.2)	53.4 (+1.6)	**81.4** (+3.0)	**82.7** (+2.0)
SETR [53]	325.1G	305.6M	48.1	48.8	77.9	-
+MLA	+8.7G	+4.0M	48.6 (+0.5)	50.3 (+1.5)	77.2 (−0.7)	-
+PUP	+97.5G	+11.7M	48.6 (+0.5)	50.1 (+1.3)	79.3 (+1.4)	-

* In green are the gaps of at least **+2.0** mIoU.

Table 3. **Compare different integrations of vision Transformer.** We report the results of the baseline, the state-of-the-art Mask Transformer from Segmenter [38], UperNet [47] and our RegProxy. We also report numbers from SETR [53] for reference.

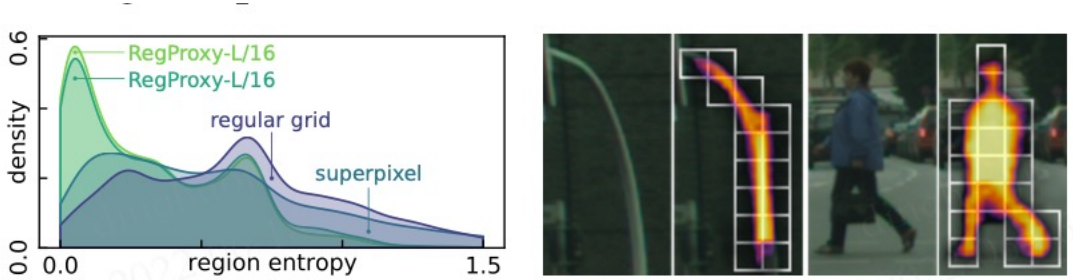


Figure 6. **(Left) Distribution of region entropies** estimated by kernel density estimation (KDE). Regions with entropy of 0 are ignored. **(Right) Geometrics of the *class-agnostic* regions** and its corresponding tokens (marked using white cell).

depth M	0	3	4	5	6	9	12	baseline
ADE20K	46.3	47.1	47.0	**47.2**	46.8	46.6	45.7	45.0
Cityscapes	76.8	**79.0**	78.8	78.2	77.7	76.1	75.2	75.4

Table 7. **Ablation on depth of token features used for region learning**. We report single scale mIoU results of RegProxy-S/16 on ADE20K and Cityscapes using half of the training schedule.

Semantic Segmentation

method	backbone	FLOPs	#params.	mIoU	FPS
FCN [27]	MobileNetV2	39.6G	9.8M	19.7 / -	64.4
D.LabV3+ [6]	MobileNetV2	69.4G	15.4M	34.0 / -	43.1
SegFormer [48]	MiT-B0	8.4G	3.8M	37.4 / 38.0	50.5
Segmenter [38]	ViT-Ti/16	4.9G	6.7M	38.1 / 38.8	-
RegProxy	ViT-Ti/16	3.9G	5.8M	**42.1** / **43.1**	38.9
OCRNet [50]	HRNet-W18	55G	12M	39.3 / 40.8	18.9
SegFormer [48]	MiT-B1	16G	14M	42.2 / 43.1	47.7
Segmenter [38]	ViT-S/16	19G	26M	45.3 / 46.9	29.8
MaskFormer [7]	Swin-T	55G	42M	46.7 / 48.8	22.1
RegProxy	ViT-S/16	15G	22M	**47.6** / 48.5	32.1
RegProxy	R26+ViT-S/32	16G	36M	**47.8** / **49.1**	28.5
OCRNet [50]	HRNet-W48	165G	71M	43.2 / 44.9	17.0
D.LabV3+ [6]	ResNet-101	255G	63M	45.5 / 46.4	14.1
D.LabV3+ [6]	ResNeSt-200	345G	88M	- / 48.4	-
Segmenter [38]	ViT-B/16	76G	103M	48.7 / 50.1	14.6
RegProxy	ViT-B/16	59G	87M	**49.8** / **50.5**	20.1
DPT [32]	DPT-Hybrid	-	123M	- 49.0	-
SETR [53]	ViT-L/16	422G	318M	48.6 / 50.1	4.5
SegFormer [48]	MiT-B5	183G	85M	51.0 / 51.8	9.8
Segmenter [38]	ViT-L/16	370G	333M	51.8 / **53.6**	-
RegProxy	R50+ViT-L/32	82G	323M	51.0 / 51.7	12.7
RegProxy	ViT-L/16	326G	306M	**52.9** / 53.4	6.6

* All models in the last group except DPT use a larger 640 × 640 crop.

method	backbone	crop	FLOPs	#params.	mIoU
D.LabV3+ [6]	ResNet-18	769^2	992G	12M	76.3 / 77.9
OCRNet [50]	HRNet-W18	full	424G	12M	78.6 / 80.5
SegFormer [48]	MiT-B0	768^2	52G	4M	75.3 / -
SegFormer [48]	MiT-B1	1024^2	244G	14M	78.5 / 80.0
RegProxy	ViT-Ti/16	768^2	69G	6M	76.5 / 77.7
RegProxy	ViT-S/16	768^2	270G	23M	**79.8** / **81.5**
OCRNet [50]	HRNet-W48	full	1297G	70M	80.7 / 81.9
Auto-D.Lab [25]	NAS-F48	769^2	-	44M	- / 80.4
Axial-D.Lab [42]	Axial-D.Lab-XL	-	2447G	173M	- / 81.1
D.LabV3+ [6]	ResNeSt-200	full	-	88M	- / **82.7**
SETR [53]	ViT-B/16	768^2	-	98M	79.5 / -
SETR [53]	ViT-L/16	768^2	-	318M	79.3 / 82.2
Segmenter [38]	ViT-B/16	768^2	1344G	103M	- / 80.6
Segmenter [38]	ViT-L/16	768^2	-	337M	79.1 / 81.3
RegProxy	ViT-B/16	768^2	1064G	88M	81.0 / 82.2
RegProxy	ViT-L/16	768^2	-	307M	**81.4** / **82.7**

Table 5. **Comparison to state-of-the-art methods on Cityscapes `val` split.** The "full" crop indicates the whole image inference, while others indicate the *sliding window* protocol.

Semantic Segmentation

Deep Hierarchical Semantic Segmentation

Liulei Li[1,5*], Tianfei Zhou[2], Wenguan Wang[3†], Jianwu Li[1], Yi Yang[4]

[1] Beijing Institute of Technology [2] ETH Zurich [3] ReLER, AAII, University of Technology Sydney [4] CCAI, Zhejiang University [5] Baidu Research

https://github.com/0liliulei/HieraSeg

Semantic Segmentation

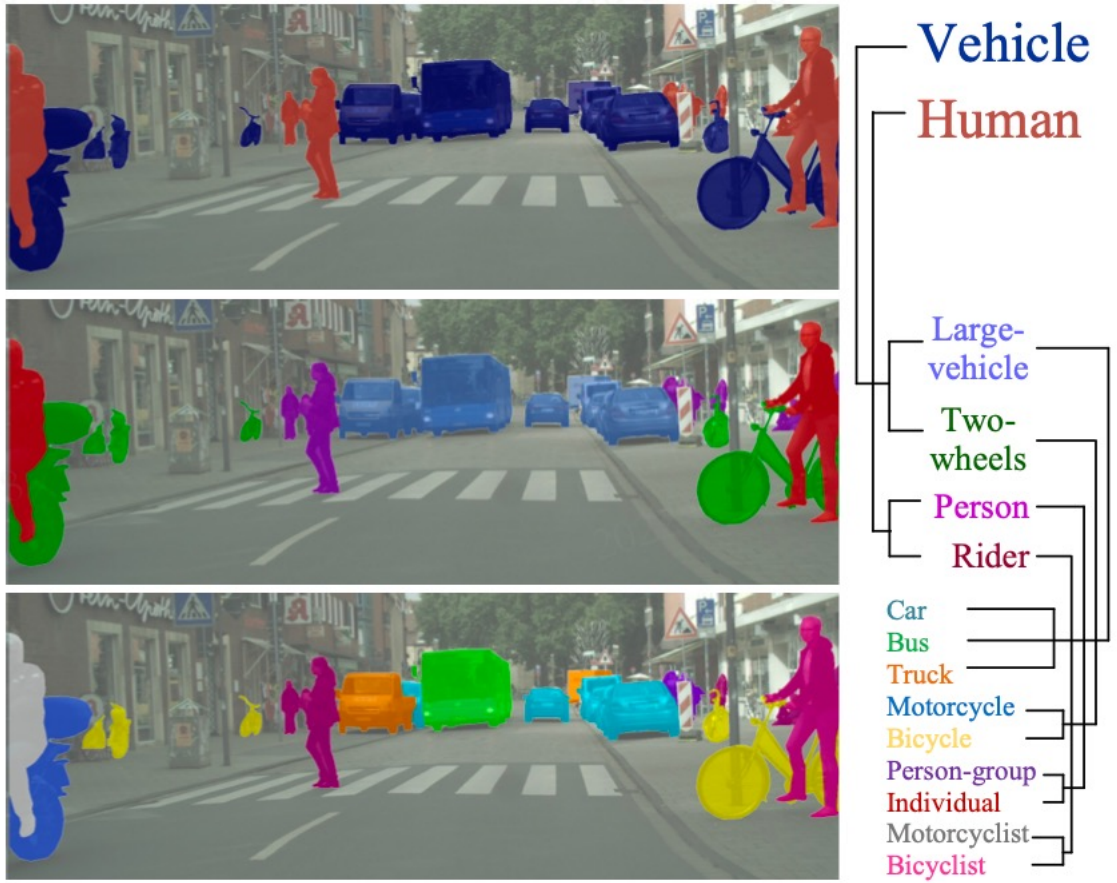


Figure 1. **Hierarchical semantic segmentation** explains visual scenes with multi-level abstraction (*left*), by considering structured class relations (*right*). The class taxonomy is borrowed from [58].

Definition 3.2.1 (Positive $\mathcal{T}$-Property). *For each pixel, if a class is labeled positive, all its ancestor nodes* (i.e., *superclasses) in* $\mathcal{T}$ *should be labeled positive.*

Definition 3.2.2 (Negative $\mathcal{T}$-Property). *For each pixel, if a class is labeled negative, all its child nodes* (i.e., *subclasses) in* $\mathcal{T}$ *should be labeled negative.*

Definition 3.2.3 (Positive $\mathcal{T}$-Constraint). *For each pixel, if* v *class is labeled positive, and* u *is an ancestor node* (i.e., *superclass) of* v*, it should hold that* $s_v \leq s_u$.

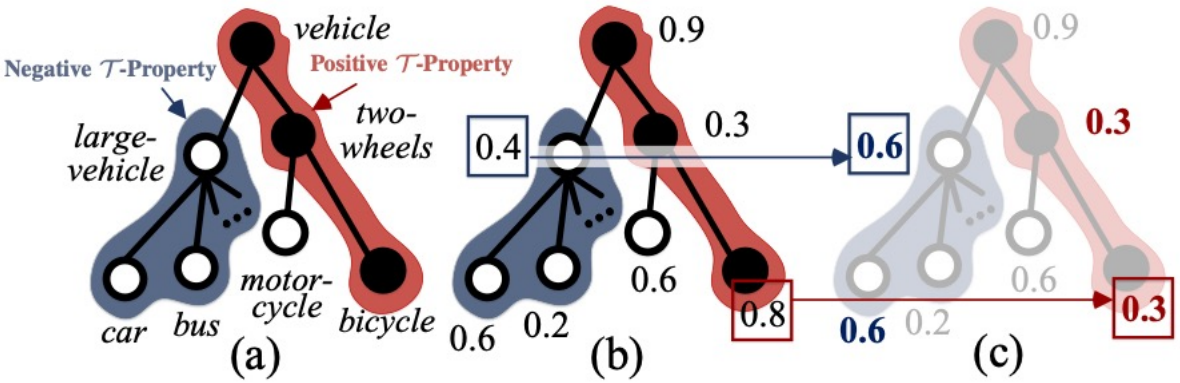
Definition 3.2.4 (Negative $\mathcal{T}$-Constraint). *For each pixel, if* v *class is labeled negative, and* u *is a child node* (i.e., *subclass) of* v*, it should hold that* $1 - s_v \leq 1 - s_u$.

$$\mathcal{L}^{\text{BCE}}(\boldsymbol{s}) = \sum_{v \in \mathcal{V}} -\hat{l}_v \log(s_v) - (1-\hat{l}_v)\log(1-s_v).$$

Semantic Segmentation

Definition 3.2.3 (Positive $\mathcal{T}$-Constraint). *For each pixel, if v class is labeled positive, and u is an ancestor node (*i.e., *superclass) of v, it should hold that $s_v \leq s_u$.*

Definition 3.2.4 (Negative $\mathcal{T}$-Constraint). *For each pixel, if v class is labeled negative, and u is a child node (*i.e., *subclass) of v, it should hold that $1 - s_v \leq 1 - s_u$.*



$$\begin{cases} p_v = \min\limits_{u \in \mathcal{A}_v} (s_u) & \text{if } \hat{l}_v = 1, \\ 1 - p_v = \min\limits_{u \in \mathcal{C}_v} (1 - s_u) = 1 - \max\limits_{u \in \mathcal{C}_v} (s_u) & \text{if } \hat{l}_v = 0, \end{cases} \tag{4}$$

$$\mathcal{L}^{\mathrm{TM}}(\boldsymbol{p}) = \sum_{v \in \mathcal{V}} -\hat{l}_v \log(p_v) - (1 - \hat{l}_v) \log(1 - p_v),$$

$$\mathcal{L}^{\mathrm{FTM}}(\boldsymbol{p}) = \sum_{v \in \mathcal{V}} -\hat{l}_v (1 - p_v)^{\gamma} \log(p_v) - (1 - \hat{l}_v)(p_v)^{\gamma} \log(1 - p_v),$$

Semantic Segmentation

tree-triplet loss

$$\mathcal{L}^{\mathrm{TT}}(\boldsymbol{i}, \boldsymbol{i}^{+}, \boldsymbol{i}^{-}) = \max\{\langle \boldsymbol{i}, \boldsymbol{i}^{+} \rangle - \langle \boldsymbol{i}, \boldsymbol{i}^{-} \rangle + m, 0\},$$

$$m = m_{\varepsilon} + 0.5 m_{\tau}$$
$$m_{\tau} = (\psi(\hat{v}_{\chi}, \hat{v}_{\chi}^{-}) - \psi(\hat{v}_{\chi}, \hat{v}_{\chi}^{+}))/2D,$$

Semantic Segmentation

Method		Backbone	mIoU^2↑	mIoU^1↑
DeepLabV2 [10] [CVPR17]		ResNet-101	-	70.22
PSPNet [105] [CVPR17]		ResNet-101	-	80.91
PSANet [106] [ECCV18]		ResNet-101	-	80.96
PAN [40] [ArXiv18]		ResNet-101	-	81.12
DeepLabV3+ [13] [ECCV18]		ResNet-101	92.16	82.08
DANet [25] [CVPR19]		ResNet-101	-	81.52
Acfnet [100] [ICCV19]		ResNet-101	-	81.60
CCNet [35] [ICCV19]		ResNet-101	-	81.08
HANet [17] [CVPR20]		ResNet-101	-	81.82
HRNet [79] [TPAMI20]		HRNet-W48	92.12	81.96
OCRNet [98] [ECCV20]		HRNet-W48	92.57	82.33
MaskFormer [16] [NeurIPS21]		Swin-Small	92.96	82.57
HSSN	DeepLabV3+	ResNet-101	93.31	83.02
	OCRNet	HRNet-W48	93.92	83.37
	MaskFormer	Swin-Small	**94.39**	**83.74**

Table 2. **Hierarchical semantic segmentation results** (§4.2) the `val` set of Cityscapes [18].

$\mathcal{L}^{\text{FTM}}$ Eq. 6	$\mathcal{L}^{\text{TT}}$ Eq. 7	Mapillary Vistas 2.0 mIoU^3↑	mIoU^2↑	mIoU^1↑	Pascal-Person-Part mIoU^3↑	mIoU^2↑	mIoU^1↑
		81.86	68.17	37.43	93.58	83.04	67.84
✓		84.17	69.62	39.17	96.33	86.72	72.89
	✓	83.06	68.61	38.29	95.92	86.03	72.27
✓	✓	**85.27**	**71.40**	**40.16**	**97.69**	**88.20**	**75.44**

Table 5. **Analysis of essential components** on Mapillary Vistas 2.0 [58] `val` and PASCAL-Person-Part [87] `test` (§4.4).

Loss	Mapillary Vistas 2.0 mIoU^3↑	mIoU^2↑	mIoU^1↑	Pascal-Person-Part mIoU^3↑	mIoU^2↑	mIoU^1↑
CCE	81.86	68.17	37.43	93.58	83.04	67.84
BCE	81.56	67.61	37.26	93.12	82.55	67.38
Focal	82.63	68.48	38.09	94.07	83.66	68.42
TM	83.48	69.13	38.69	95.32	85.99	72.17
FTM	84.17	69.62	39.17	96.33	86.72	72.89
Full	**85.27**	**71.40**	**40.16**	**97.69**	**88.20**	**75.44**

Table 6. **Analysis of focal tree-min loss $\mathcal{L}^{\text{FTM}}$** on Mapillary Vistas 2.0 [58] `val` and PASCAL-Person-Part [87] `test` (§4.4).

Semantic Segmentation

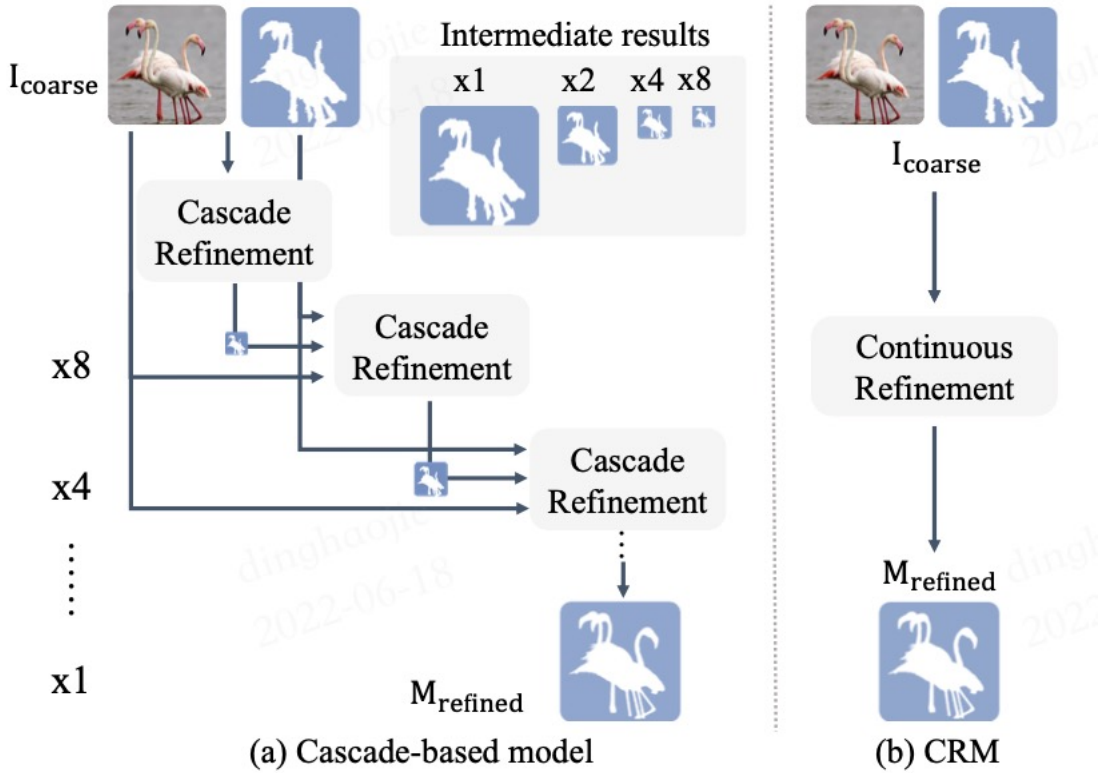
High Quality Segmentation for Ultra High-resolution Images

Tiancheng Shen[1] Yuechen Zhang[1] Lu Qi[1] Jason Kuen[2]
Xingyu Xie[3] Jianlong Wu[4] Zhe Lin[2] Jiaya Jia[1,5]
[1]The Chinese University of Hong Kong [2]Adobe Research [3]Peking University
[4]Shandong University [5]SmartMore

Semantic Segmentation

Compare with Cascade Model

- Reduce computation cost
- Reconstruct more details



Semantic Segmentation

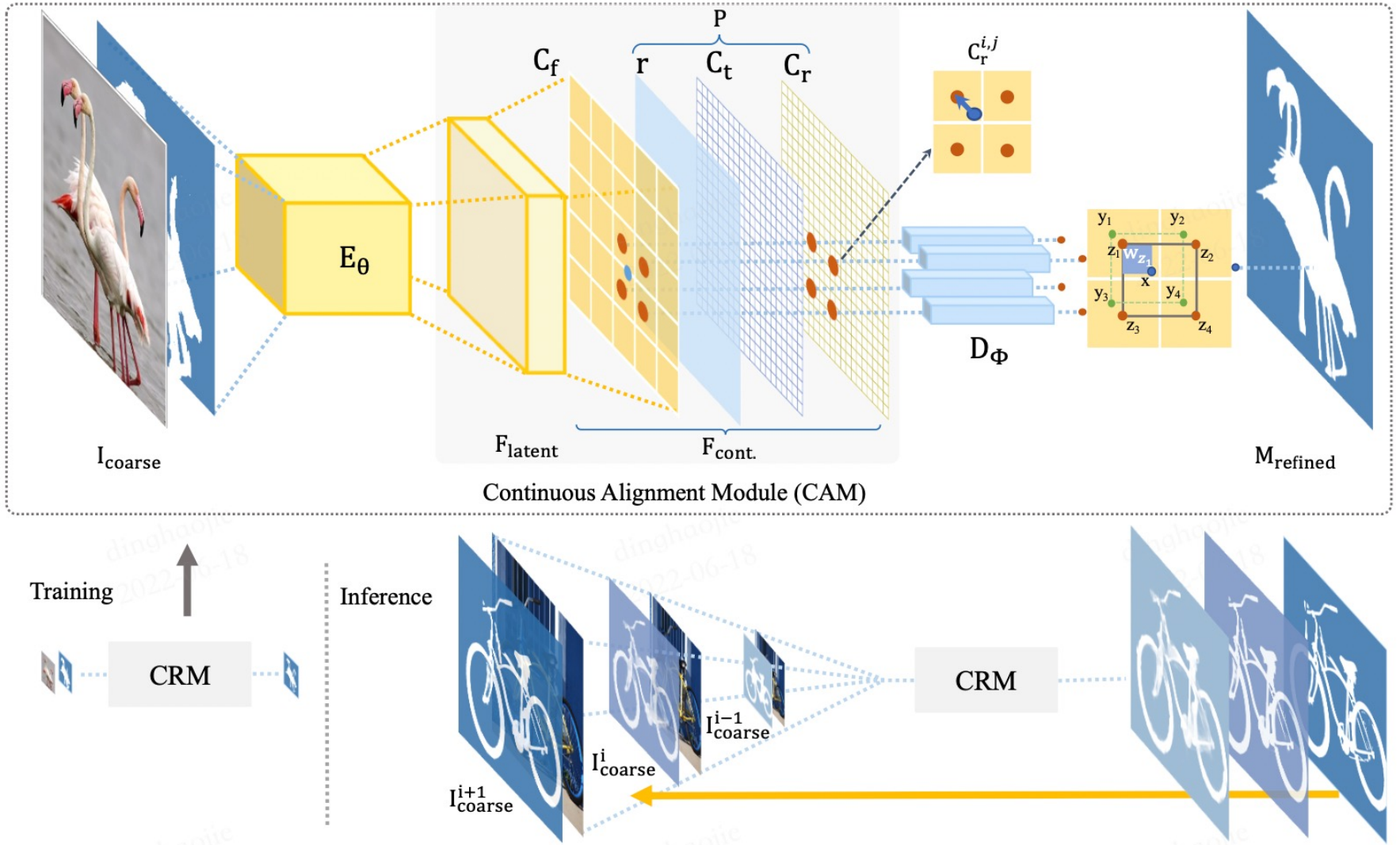


Figure 3. The general framework of CRM. The upper part is the structure of the model. The lower part is the training and testing process of CRM. From the lower part, we can also see the resolution gap between low-resolution training and high-resolution testing.

Semantic Segmentation

IoU/mBA	Coarse Mask	SegFix [53]	MGMatting [51]	CascadePSP [9]	CRM(Ours)
FCN-8s [30]	72.39/53.63	72.69/55.21	72.31/57.32	77.87/67.04	**79.62/69.47**
DeepLabV3+ [5]	89.42/60.25	89.95/64.34	90.49/67.48	**92.23**/74.59	91.84/**74.96**
RefineNet [27]	90.20/62.03	90.73/65.95	90.98/68.40	92.79/74.77	**92.89/75.50**
PSPNet [55]	90.49/59.63	91.01/63.25	91.62/66.73	93.93/75.32	**94.18/76.09**
Average Improve.	0.00/0.00	0.47/3.30	0.73/6.10	3.58/14.05	**4.01/15.12**

Table 1. IoU and mBA results on the BIG dataset comparing with other mask refinement methods. Coarse mask is from FCN, DeepLabV3+, RefineNet and PSPNet. Best results are noted with **bold**. Average Improve. represents average improvement based on coarse mask.

Method (IoU/mBA)	Time(s)	FLOPs(G)	Params(M)
CasPSP (93.9/75.3) [9]	620	26518	67.62
CRM (94.2/76.1)	425	2536	9.27
CRM* (93.9/76.3)	259	1331	9.27

Semantic Segmentation

IoU/mBA	w/o CAM&Impl.	w CAM&Impl.
0.125	92.68/63.70	93.07/65.61
0.25	93.49/69.23	93.88/71.41
0.5	93.85/73.43	94.15/74.95
1.0	93.94/75.42	94.18/76.09

Table 5. The effect of CRM and inference resolutions with PSP-Net [55]'s output as coarse mask. Impl. denotes implicit function.

CAM	Impl.	IoU	mBA
×	×	93.94	75.42
√	×	93.99	75.93
×	√	93.96	75.55
√	√	94.18	76.09

Table 6. The ablation study about CAM and implicit function with PSPNet [55]'s output as coarse mask.

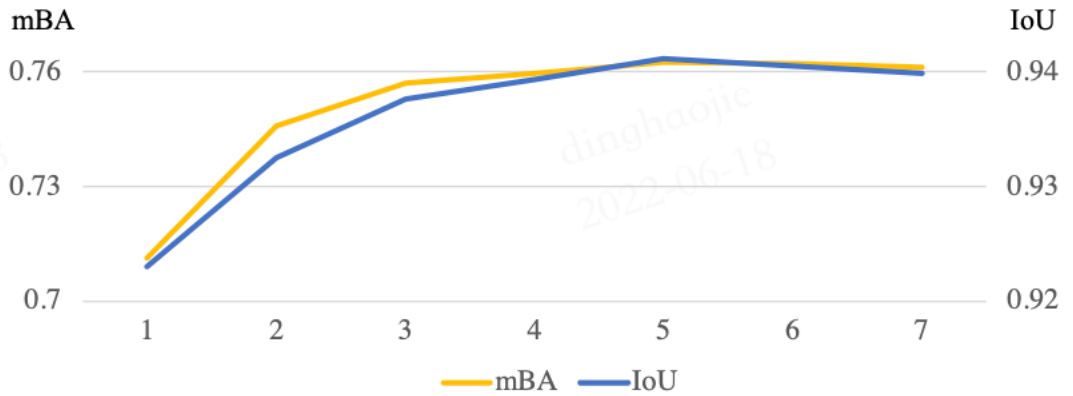


Figure 7. The effect of inference's continuity. The horizontal axis represents the number of uniformly sampled points between 0 and 1. The sampled points are rescale ratios of input.

Panoptic Segmentation

Panoptic, Instance and Semantic Relations: A Relational Context Encoder to Enhance Panoptic Segmentation

Shubhankar Borse * Hyojin Park * Hong Cai Debasmit Das Risheek Garrepalli
Fatih Porikli
Qualcomm AI Research †
{sborse, hyojinp, hongcai, debadas, rgarrepa, fporikli}@qti.qualcomm.com

Panoptic Segmentation

Contributions

- capture the relations among semantic classes and instances
- automatically focus on more important instances while generating relational features
- a universal module

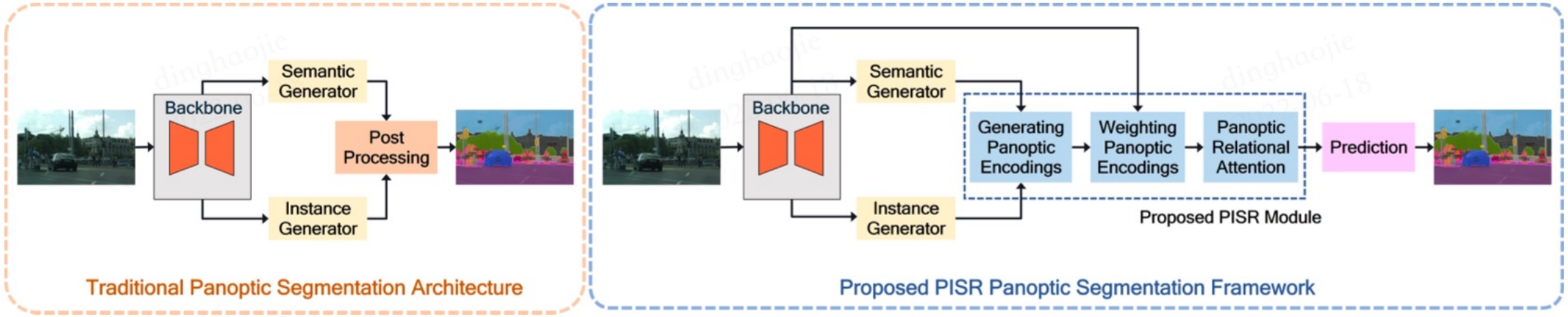
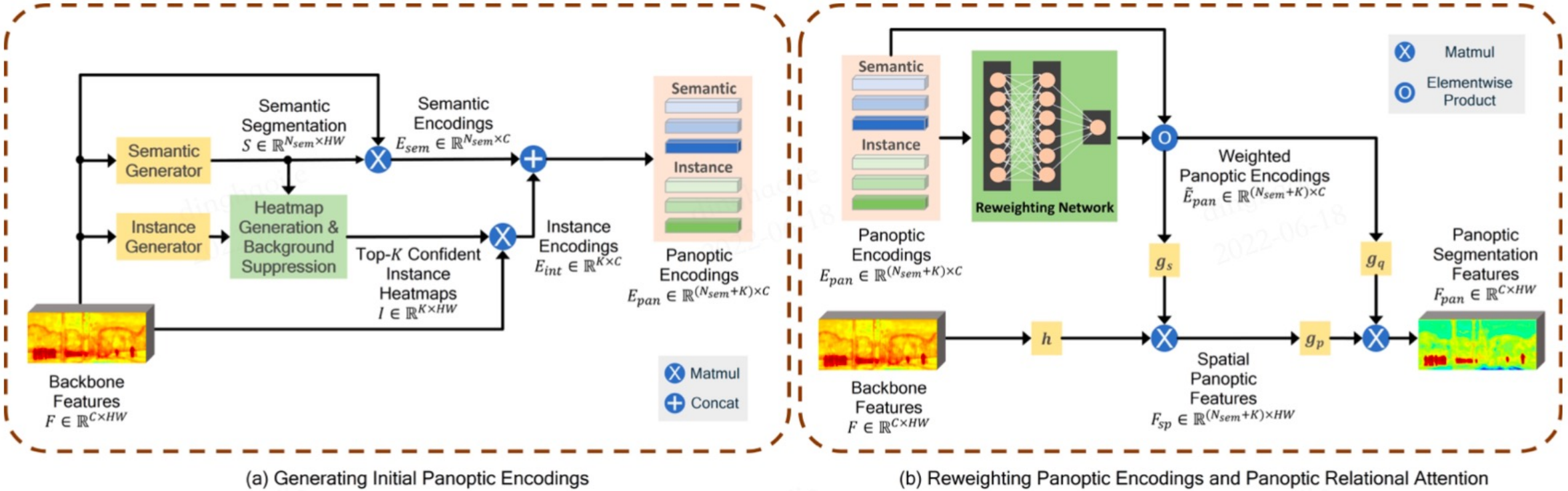


Figure 2. Left: Conventional panoptic segmentation architecture. Right: Our proposed Panoptic, Instance, and Semantic Relations (PISR) framework that can work with any base panoptic segmentation model.

Panoptic Segmentation



(a) Generating Initial Panoptic Encodings

(b) Reweighting Panoptic Encodings and Panoptic Relational Attention

Figure 3. (a) Details on how the initial panoptic encodings are generated. (b) Details on how PISR reweights the initial encodings and subsequently applies two-stage attention by employing PRA to generate the final panoptic segmentation features.

Panoptic Segmentation

Split	Method	Backbone	PQ	PQ^{th}	PQ^{st}
Val	UPSNet [48]	RN50-FPN	42.5	48.5	33.4
	AUNet [26]	RN50-FPN	39.6	**49.1**	25.2
	CIAE (640) [15]	RN50-FPN	39.5	44.4	33.1
	COPS [1]	RN50	38.4	40.5	35.2
	OANet [30]	RN50-FPN	39.0	48.3	24.9
	AdaptIS [39]	RN50	35.9	40.3	29.3
	SSAP [14]	RN50	36.5	40.1	32.0
	LPSNet [19]	RN50	39.1	43.9	30.1
	Panoptic-FPN [21]	RN50-FPN	39.2	46.6	27.9
	Panoptic-FPN + PISR	RN50-FPN	**42.7**	**48.7**	**33.6**
	Panoptic-DL [11]	RN50	35.5	37.8	32.0
	Panoptic-DL [11] + OCR	RN50	37.2	38.9	35.7
	Panoptic-DL [11] + PISR	RN50	**38.8**	**40.6**	**36.2**
	Panoptic-DL [11]	HR48	37.8	-	-
	Panoptic-DL [11] + PISR	HR48	**40.7**	**42.6**	**37.7**
	Panoptic-FCN [27]	Swin-L	52.1	58.5	42.3
	Pan-SegFormer [28]	PvTv2-B5	**54.1**	**60.4**	**44.6**
	Pan-SegFormer [28]	PvTv2-B2	52.6	58.2	43.3
	Max-DeepLab [41]	Max-L	51.1	57.0	42.2
	MaskFormer [12]	Swin-L	52.7	58.5	44.0
	UPerNet [47]	Swin-L	50.3	55.7	42.1
	UPerNet [47] + PISR	Swin-L	**52.9**	**58.9**	**43.8**
Test	Panoptic-FCN [27]	Swin-L	52.7	59.4	42.5
	Refine [38]	RNX101-FPN	51.5	59.6	39.2
	Max-DeepLab [41]	Max-L	51.3	57.2	43.4
	UPerNet [47]	Swin-L	50.9	56.7	42.3
	UPerNet [47] + PISR	Swin-L	**53.2**	**59.2**	**44.2**

Table 2. Quantitative evaluation on the COCO validation and test sets, in terms of PQ, PQ^{th}, and PQ^{st}. RN and HR48 indicate ResNet and HRNet-w48, respectively. Gray rows are our models introduced in this paper. Best numbers are highlighted in bold.

Method	Backbone	PQ	PQ^{th}	PQ^{st}	SQ	RQ
Panoptic-FCN [27]	RN50	30.1	34.1	27.3	-	-
BGRNet [45]	RN50	31.8	34.1	27.3	-	-
Auto-pan. [46]	SV2	32.4	33.5	30.2	-	-
MaskFormer [12]	RN50	34.7	32.5	38.0	76.3	41.7
MaskFormer + PISR	RN50	36.1	34.7	39.0	78.3	44.3
MaskFormer	RN101	35.7	34.5	38.0	77.4	43.8
MaskFormer + PISR	RN101	**37.0**	**35.6**	**39.7**	**79.9**	**45.2**

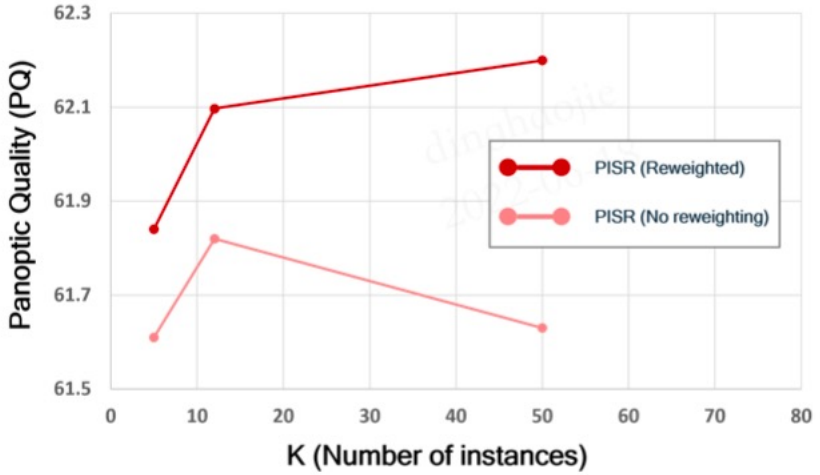
Table 3. Quantitative evaluation on the ADE20K validation set. RN and SV2 indicates ResNet and ShuffleNetV2, respectively. Gray rows are new models (ours) introduced in this paper. Best numbers are highlighted in bold.

Panoptic Segmentation

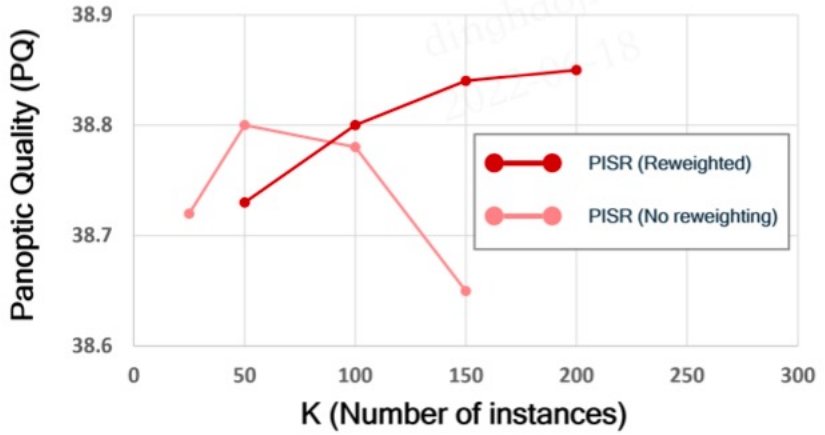
Model	PQ	mIoU	iIoU	Δ
Panoptic-DL [11]	59.9	78.5	62.5	0
Panoptic-DL + Concatenation	60.6	79.1	62.6	+ 1.4
Panoptic-DL + Elementwise Product	60.4	78.2	63.0	+ 0.7
Panoptic-DL + OCR	60.7	79.5	62.4	+ 1.7
Panoptic-DL + PISR (w/o Reweighting)	61.8	79.6	64.0	+ 4.5
Panoptic-DL + PISR	**62.2**	**80.2**	**64.4**	**+ 5.9**

Table 5. Comparing alternative ways to process semantic and instance features. Δ is the sum of all gains in PQ, mIoU, and iIoU.

- *Concatenation :* simply concat S, I, and F
- *Elementwise Product :* S and I are first concatenated and match the dimensions of F. Finally product with F
- *OCR :* apply the OCR module to the semantic features



(a) Top-K analysis on Cityscapes val



(b) Top-K analysis on COCO val

Panoptic Segmentation

CMT-DeepLab: Clustering Mask Transformers for Panoptic Segmentation

Qihang Yu[1*] Huiyu Wang[1] Dahun Kim[2] Siyuan Qiao[3] Maxwell Collins[3] Yukun Zhu[3]
Hartwig Adam[3] Alan Yuille[1] Liang-Chieh Chen[3]
[1]Johns Hopkins University [2]KAIST [3]Google Research

Panoptic Segmentation

Motivation:

- Object queries tends to focus on only a few locations
- The pixels only have *one* chance to communicate with the object queries in the final output

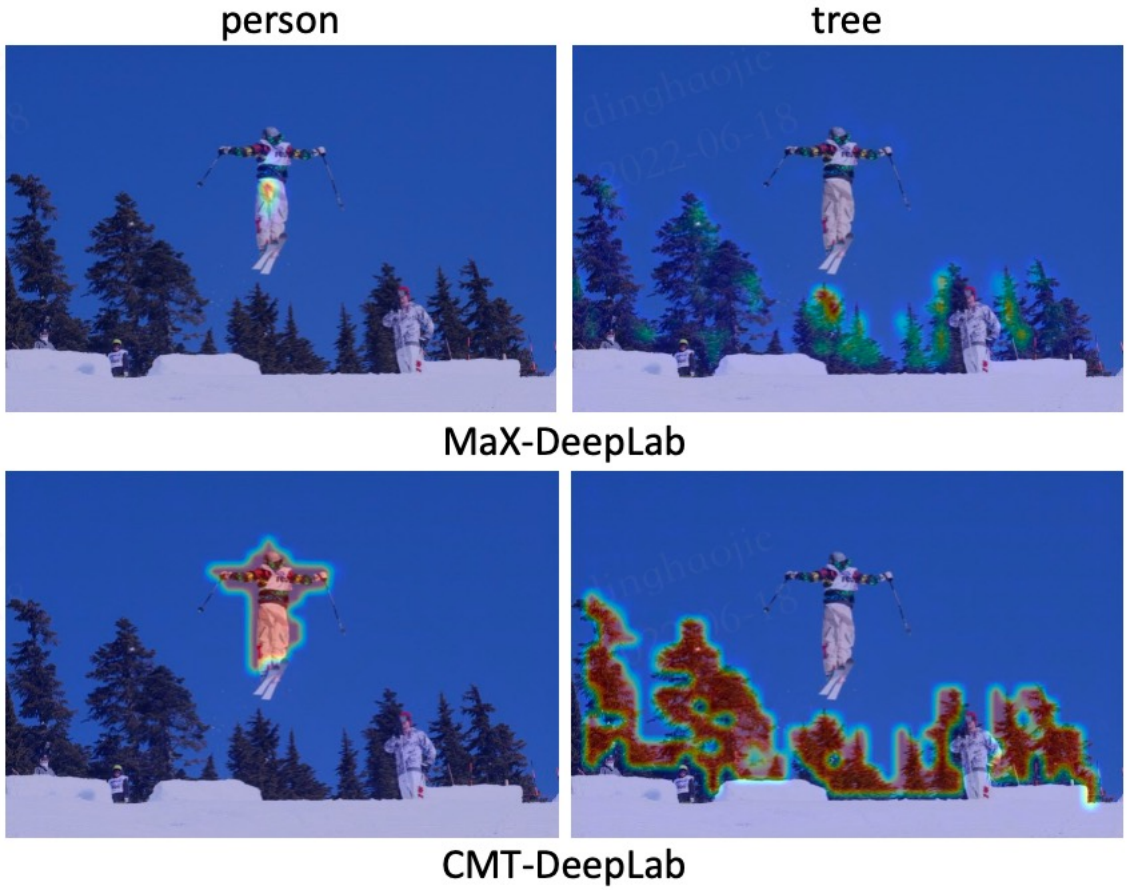
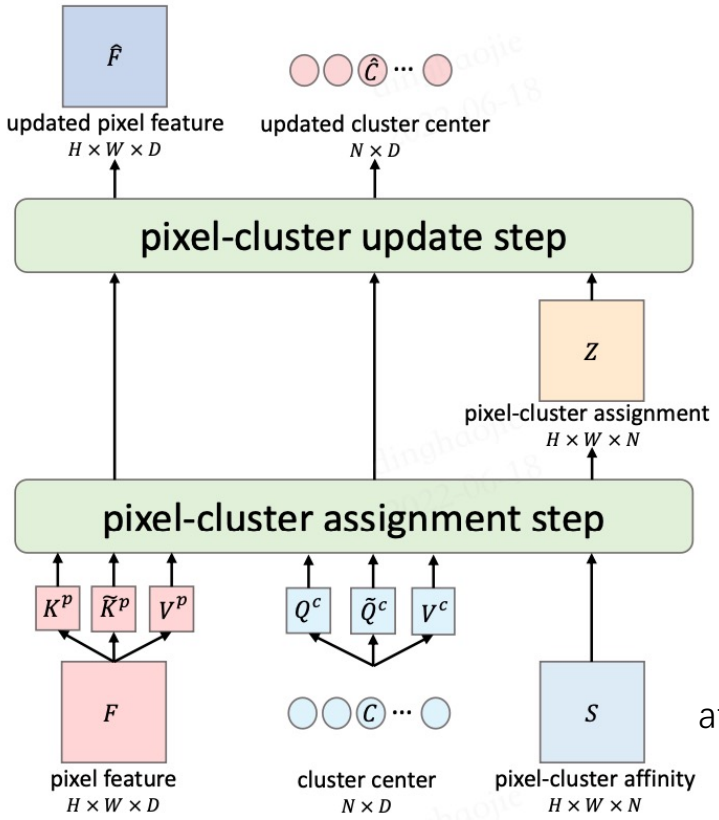


Figure 1. Our CMT-DeepLab generates denser cross-attention maps than MaX-DeepLab [82]. The visualization is based on the last transformer layer with averaged multi-head attentions.

Panoptic Segmentation



affinity logits

Figure 3. A visual illustration of Clustering Mask Transformer layer, where three variables are updated in a dynamic manner based on the clustering results: pixel features, cluster centers, and pixel-cluster affinity. Details of assignment and update steps are illustrated in Fig. 4.

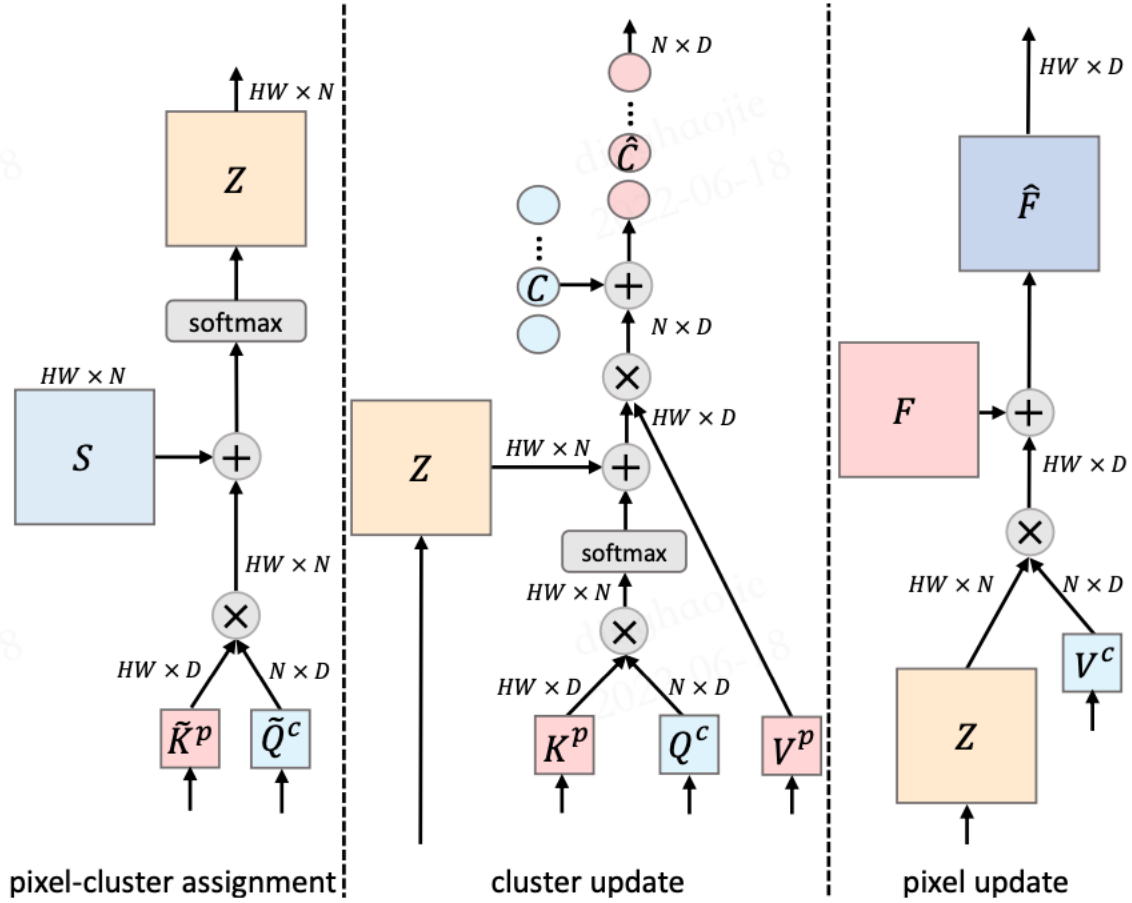


Figure 4. Detailed visual illustration of pixel-cluster assignment (left), cluster centers update (middle), and pixel features update (right). The tensor shapes are specified for illustration.

Panoptic Segmentation

Location-Sensitive Clustering

$$\hat{\mathbf{C}} = \mathrm{Conv}(\mathrm{Concat}(\mathbf{C}, r^c)),$$
$$\hat{\mathbf{F}} = \mathrm{Conv}(\mathrm{Concat}(\mathbf{F}, r^p)),$$
$$r^c = [r^{c,h}, r^{c,w}] \in \mathbb{R}^{N \times 2M}$$

$$\begin{aligned}
\mathcal{L}_{\mathrm{ext}} &= \frac{1}{4K}\sum_{i=1}^{K}(|\min(r_i^{c,h}) - \min(y_i^h)| + |\max(r_i^{c,h}) - \max(y_i^h)| \\
&\quad + |\min(r_i^{c,w}) - \min(y_i^w)| + |\max(r_i^{c,w}) - \max(y_i^w)|), \\
\mathcal{L}_{\mathrm{cen}} &= \frac{1}{2K}\sum_{i=1}^{K}(|\mathrm{avg}(r_i^{c,h}) - \mathrm{avg}(y_i^h)| + |\mathrm{avg}(r_i^{c,w}) - \mathrm{avg}(y_i^w)|), \\
\mathcal{L}_{\mathrm{loc}} &= \mathcal{L}_{\mathrm{ext}} + \mathcal{L}_{\mathrm{cen}}, \qquad (13)
\end{aligned}$$

Contrastive loss

$$\mathcal{L}^{insdis} = \sum_{a \in A} \frac{-1}{|P(a)|} \sum_{p \in P(a)} \log \frac{\exp(f_a \cdot f_p/\tau)}{\sum_{b \in A} \exp(f_a \cdot f_b/\tau)},$$

Panoptic Segmentation

	PQ	PQ^{Th}	PQ^{St}
baseline	46.2	50.0	40.5
+ clustering transformer	47.1	51.0	41.1
+ pixel-wise contrastive loss	**47.5**	**51.1**	**42.1**

(a) **CMT-DeepLab: clustering update**.

	PQ	PQ^{Th}	PQ^{St}
baseline	46.2	50.0	40.5
+ ref. mask pred.	46.6	50.3	40.9
+ coord-conv	**46.9**	**50.6**	**41.3**

(b) **CMT-DeepLab: location-senseitive clustering**.

clustering update	location	decoder	params	PQ	PQ^{Th}	PQ^{St}
			61.9M	46.2	50.0	40.5
✓			61.9M	47.5	51.1	42.1
	✓		65.5M	46.9	50.6	41.3
		✓	91.0M	47.1	51.3	40.9
✓		✓	91.0M	48.1	51.9	42.2
✓	✓	✓	94.9M	**48.4**	**52.1**	**42.8**

(c) **CMT-DeepLab: architecture**.

ImageNet-22K	RFN	mask-wise merge	PQ	PQ^{Th}	PQ^{St}
			48.4	52.1	42.8
✓			49.3	53.3	43.4
✓	✓		50.1	**54.8**	43.0
✓	✓	✓	**50.6**	**54.8**	**44.3**

(d) **CMT-DeepLab: pretraining, post-processing, scaling**.

Table 2. CMT-DeepLab ablation experiments. Baseline is labeled with grey color. Results are reported in accumulative manner.

Panoptic Segmentation

method	backbone	TTA	params	val-set PQ	val-set PQ^{Th}	val-set PQ^{St}	test-dev PQ	test-dev PQ^{Th}	test-dev PQ^{St}
box-based panoptic segmentation methods									
Panoptic-FPN [47]	R101			40.3	47.5	29.5	-	-	-
UPSNet [89]	R50			42.5	48.5	33.4	-	-	-
UPSNet [89]	R50	✓		43.2	49.1	34.1	-	-	-
UPSNet [89]	DCN-101 [24]	✓		-	-	-	46.6	53.2	36.7
DETR [10]	R101		61.8M	45.1	50.5	37.0	46.0	-	-
DetectoRS [71]	RX-101 [88]	✓		-	-	-	49.6	57.8	37.1
center-based panoptic segmentation methods									
Panoptic-DeepLab [19]	X-71 [23]		46.7M	39.7	43.9	33.2	-	-	-
Panoptic-DeepLab [19]	X-71 [23]	✓	46.7M	41.2	44.9	35.7	41.4	45.1	35.9
Axial-DeepLab-L [83]	AX-L [83]		44.9M	43.4	48.5	35.6	43.6	48.9	35.6
Axial-DeepLab-L [83]	AX-L [83]	✓	44.9M	43.9	48.6	36.8	44.2	49.2	36.8
end-to-end panoptic segmentation methods									
MaX-DeepLab-S [82]	MaX-S [82]		61.9M	48.4	53.0	41.5	49.0	54.0	41.6
MaX-DeepLab-L [82]	MaX-L [82]		451M	51.1	57.0	42.2	51.3	57.2	42.4
MaskFormer [20]	Swin-B‡ [62]		102M	51.8	56.9	44.1	-	-	-
MaskFormer [20]	Swin-L‡ [62]		212M	52.7	58.5	44.0	53.3	59.1	44.5
CMT-DeepLab	Axial-R50‡ [83]		94.9M	53.0	57.7	45.9	53.4	58.3	46.0
CMT-DeepLab	Axial-R104‡		135.2M	54.1	58.8	**47.1**	54.5	59.6	46.9
CMT-DeepLab	Axial-R104‡-RFN		270.3M	55.1	60.6	46.8	55.4	61.0	**47.0**
CMT-DeepLab (iter 200k)	Axial-R104‡-RFN		270.3M	**55.3**	**61.0**	46.6	**55.7**	**61.6**	46.8

Table 1. Results comparison on COCO val and test-dev set. **TTA:** Test-time augmentation. ‡: ImageNet-22K pretraining. We provide more comparisons with concurrent works in the supplementary materials.