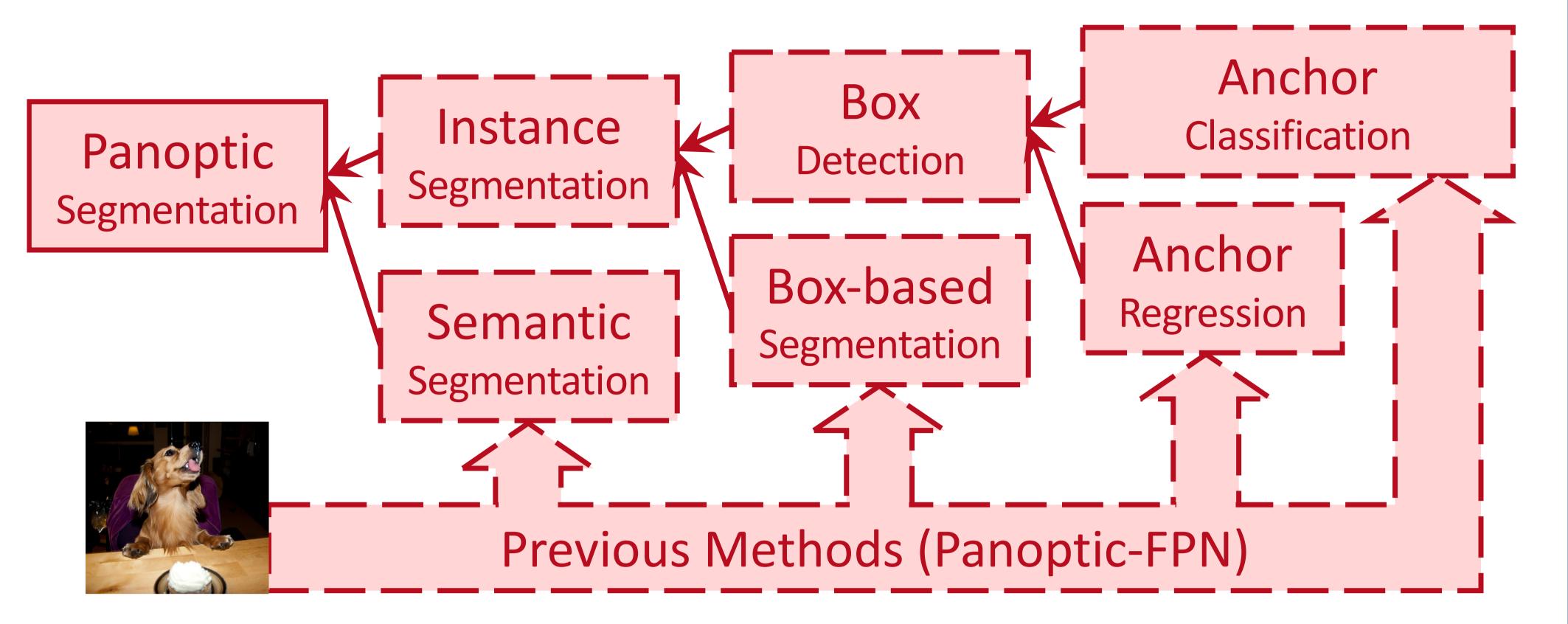
MAX-DEEPLAB: END-TO-END PANOPTIC SEGMENTATION WITH MASK TRANSFORMERS



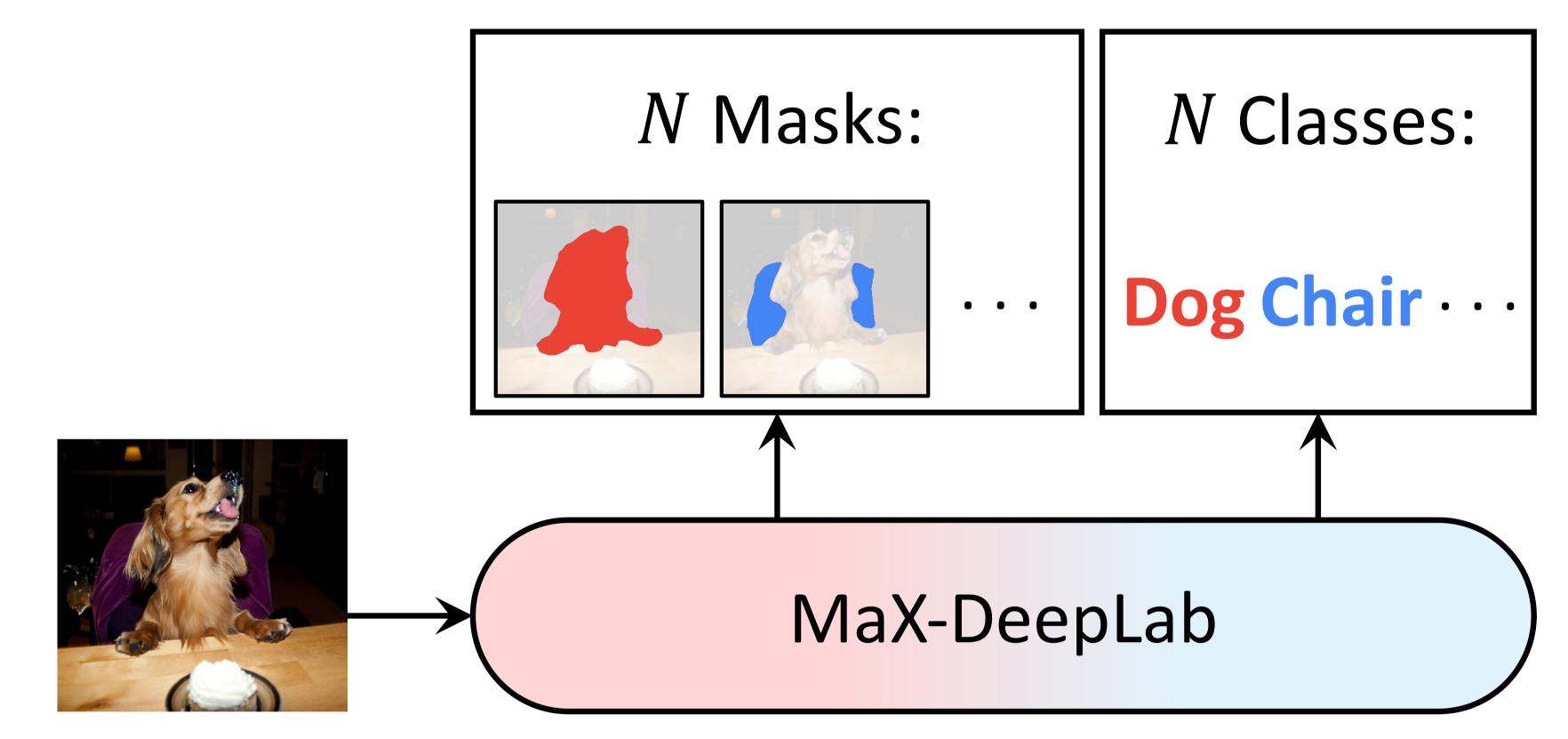
JOHNS HOPKINS UNIVERSITY

MOTIVATION

• Previous methods rely on a tree of surrogate sub-tasks. Although these sub-tasks are tackled by area experts, they fail to comprehensively solve the target task.

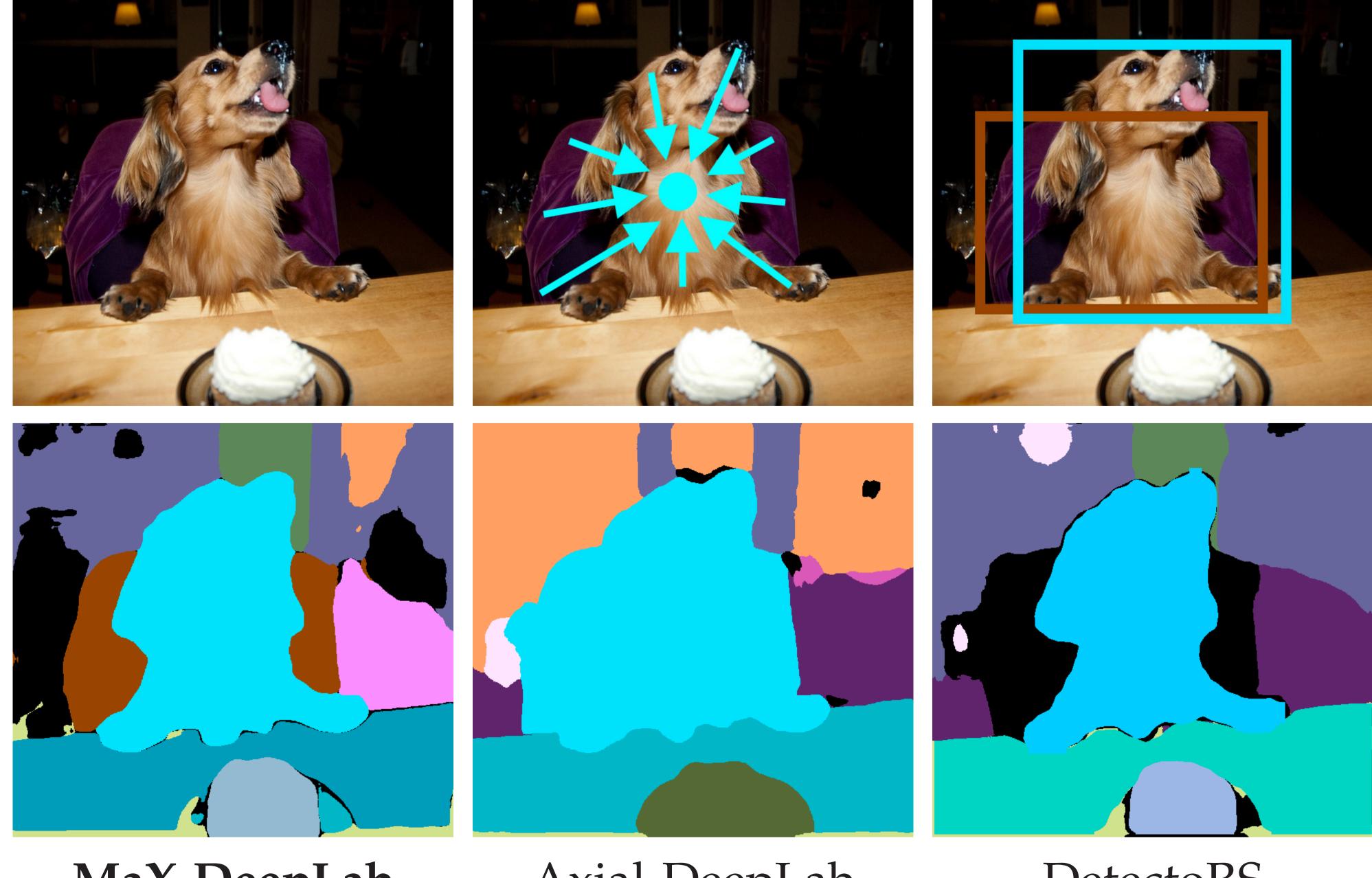


 Our mask transformer enables end-to-end panoptic segmentation for the first time by directly predicting a set of object masks and their semantic classes.



 MaX-DeepLab correctly segments a dog sitting on a chair, while other state-of-the-art methods fail because

(a) the *centers* of the dog and the chair are close to each other, (b) the *boxes* of the dog and the chair overlap a lot.



MaX-DeepLab **End-to-End**

Axial-DeepLab Center-Based

DetectoRS Box-Based

Huiyu Wang¹ Yukun Zhu² Hartwig Adam² Alan Yuille¹ Liang-Chieh Chen² ¹Johns Hopkins University ²Google Research

Code: https://github.com/google-research/deeplab2

FORMULATION

• Ground Truth as a set of class-labeled masks:

$$\{y_i\}_{i=1}^K = \{(m_i, c_i)\}_{i=1}^K,$$
(1)

K: #GT, m_i : masks, c_i : classes (thing & stuff).

• **Prediction** in the exact same form:

$$\{\hat{y}_i\}_{i=1}^N = \{(\hat{m}_i, \hat{p}_i(c))\}_{i=1}^N, \qquad (2)$$

N: a constant size of predictions (e.g. 128 for COCO), \hat{m}_i : predicted masks after pixel-wise softmax, $\hat{p}_i(c)$: class probabilities (thing, stuff, no object \emptyset).

SIMPLE INFERENCE

• Predict a class-ID for each mask:

$$\hat{c}_i = \arg\max\hat{p}_i(c) \,. \tag{3}$$

• Predict a mask-ID for each pixel:

$$\hat{z}_{h,w} = \arg\max_{i} \hat{m}_{i,h,w} , \qquad (4)$$

 $\forall h \in \{1, 2, \dots, H\}, \quad \forall w \in \{1, 2, \dots, W\}.$

TRAINING WITH PQ-STYLE LOSS

• Inspired by Panoptic Quality (PQ) decomposition:

 $PQ = RQ \times SQ$,

(5)

RQ: recognition quality, SQ: segmentation quality. • A PQ-style similarity metric is defined between a ground

$$\underbrace{\operatorname{sim}(y_i, \hat{y}_j)}_{\approx \operatorname{PQ}} = \underbrace{\hat{p}_j(c_i)}_{\approx \operatorname{RQ}} \times \underbrace{\operatorname{Dice}(m_i, \hat{m}_j)}_{\approx \operatorname{SQ}} . \tag{6}$$

• Match predictions to GTs with the metric:

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_N}{\arg \max} \sum_{i=1}^{K} \sin(y_i, \hat{y}_{\sigma(i)}), \qquad (7)$$

 $\sigma \in \mathfrak{S}_N$: a permutation of N elements,

 $\hat{\sigma}$: the optimal permutation with maximum similarity.

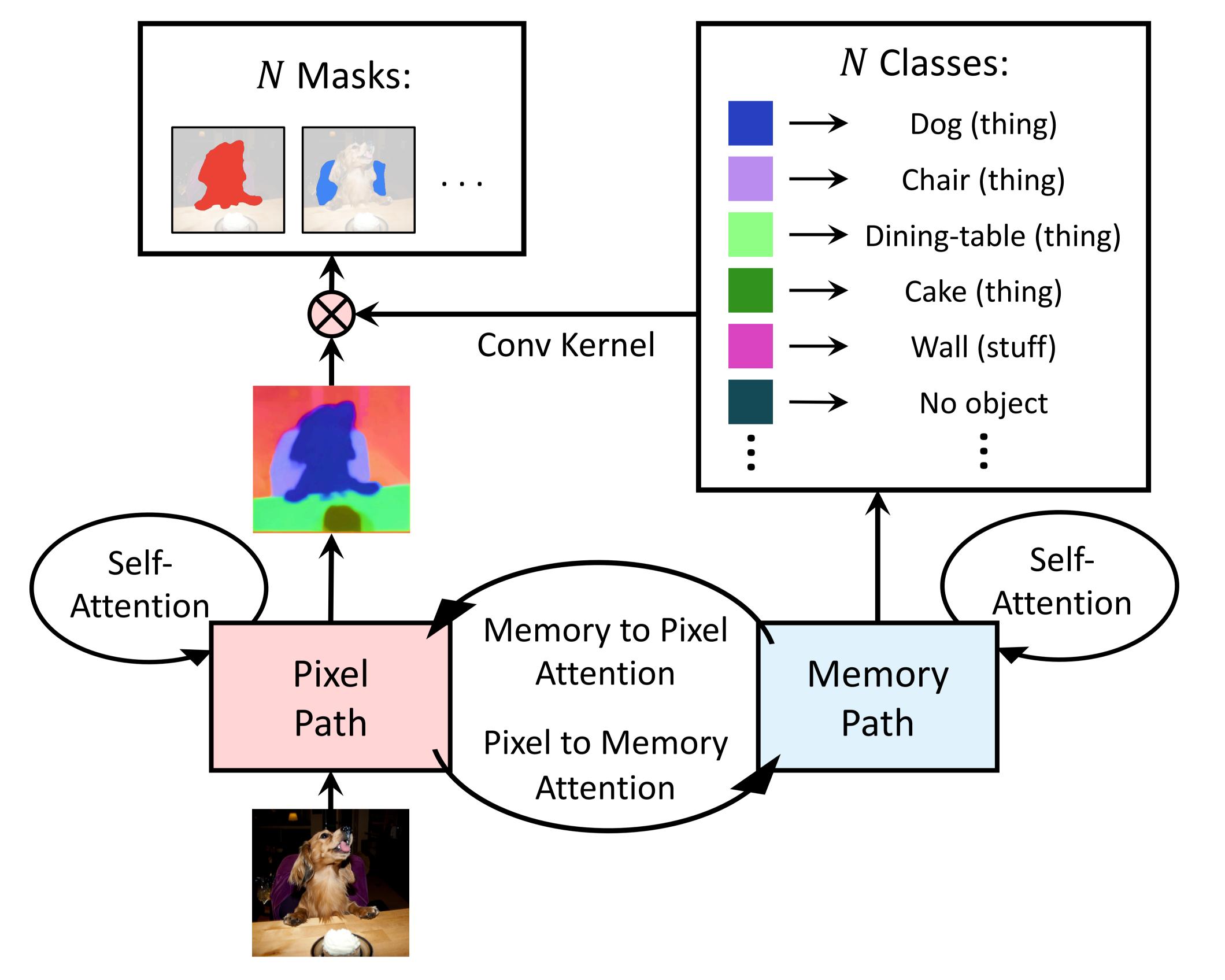
• Optimize the model with the same metric:

$$\max_{\theta} \sum_{i=1}^{K} \underbrace{\operatorname{sim}(y_i, \hat{y}_{\hat{\sigma}(i)})}_{\approx \operatorname{PQ}} = \max_{\theta} \sum_{i=1}^{K} \underbrace{\hat{p}_{\hat{\sigma}(i)}(c_i)}_{\approx \operatorname{RQ}} \times \underbrace{\operatorname{Dice}(m_i, \hat{m}_{\hat{\sigma}(i)})}_{\approx \operatorname{SQ}}.$$
 (8)

The mask and the class should be correct at the same time. Please read the paper for auxiliary loss terms.

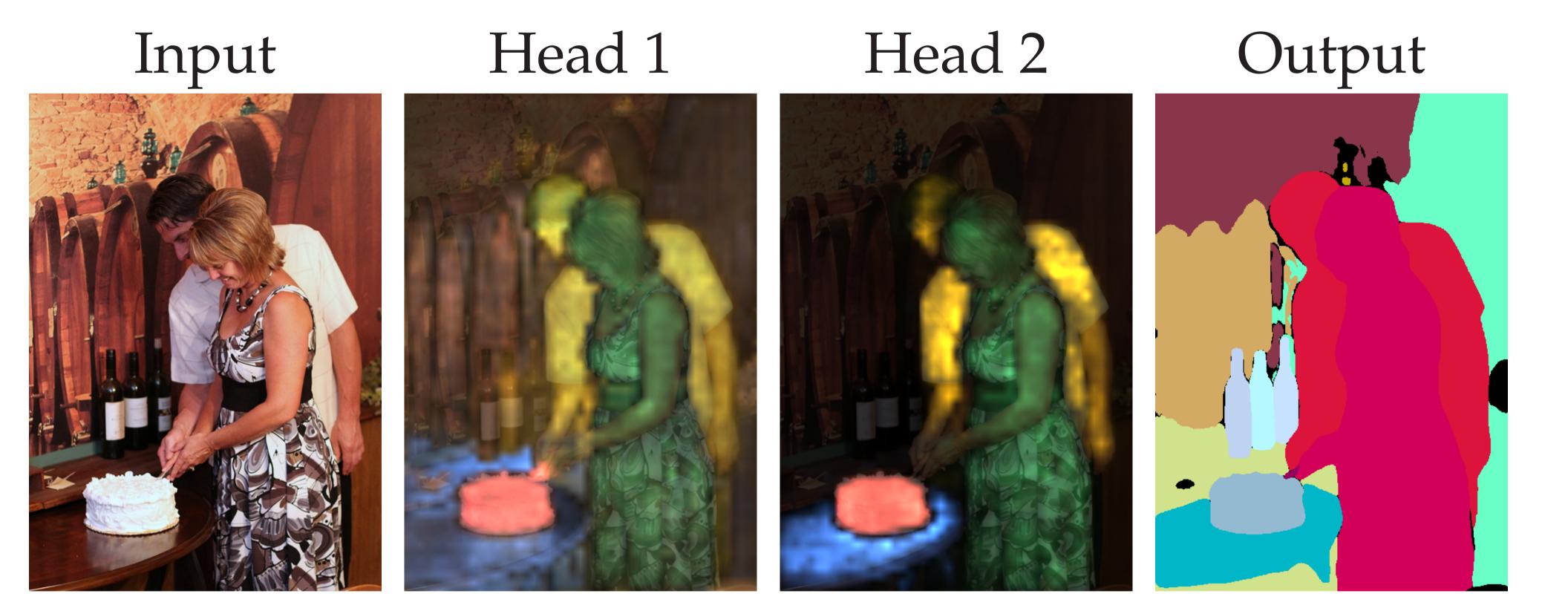
ARCHITECTURE

• An overview of the mask transformer architecture.



We augment the pixel-path CNN with a global memory path, enabling any CNN layer to read and write the memory. We adopt all four types of attention between the two paths.

ATTENTION MAPS



Two people (woman, man) cutting a cake on a table.

VISUALIZATIONS



I







Google Research



RESULTS

Method Backbone TTA PQ PQ^{Th} PQ^{St} Box-based panoptic segmentation methods Panoptic-FPN RN-101 40.9 48.3 29.7 DETR RN-101 46.0 - - - UPSNet DCN-101 ✓ 46.6 53.2 36.7 DetectoRS RX-101 ✓ 49.6 57.8 37.1 Box-free panoptic segmentation methods - 43.6 48.9 35.6 Axial-DeepLab-L AX-L ✓ 44.2 49.2 36.8 MaX-DeepLab-L AX-L ✓ 44.2 49.2 36.8 MaX-DeepLab-L MaX-S 49.0 54.0 41.6 MaX-DeepLab-L	• COCO test-dev set, TTA: Test-time augmentation.							
Panoptic-FPN RN-101 40.9 48.3 29.7 DETR RN-101 46.0 - - UPSNet DCN-101 ✓ 46.6 53.2 36.7 DetectoRS RX-101 ✓ 49.6 57.8 37.1 Box-free panoptic segmentation methods Panoptic-DeepLab X-71 ✓ 41.4 45.1 35.9 Axial-DeepLab-L AX-L ✓ 44.2 49.2 36.8 Max-DeepLab-L AX-L ✓ 44.2 49.2 36.8 Max-DeepLab-L MaX-S 49.0 54.0 41.6 Max-DeepLab-L MaX-L ✓ 44.7 48.5 39.0 Max-DeepLab-L MaX-L ✓ 44.7 48.5 39.0 ✓ ✓ ✓ ✓ 45.7 49.8 39.2 ✓ ✓ ✓ ✓ 45.7 49.8 39.4 ✓ ✓ ✓ ✓ 47.8 51.9 41.5 ✓ ✓ ✓ ✓ ✓ 49.4 54.5 </td <td>Method</td> <td>Backbone</td> <td>TTA</td> <td>PÇ</td> <td>)]</td> <td>COTh</td> <td>PQSt</td>	Method	Backbone	TTA	PÇ)]	COTh	PQ St	
$\begin{array}{c c c c c c c c c c c c c c c c c c c $	Box-based panoptic segmentation methods							
UPSNet DetectoRS DCN-101 RX-101 \checkmark 46.6 53.2 36.7 37.1 Box-free panoptic segmentation methods Panoptic-DeepLab X-71 \checkmark 41.4 45.1 35.9 43.6 Axial-DeepLab-L AX-L 43.6 48.9 35.6 43.6 Axial-DeepLab-L AX-L \checkmark 44.2 49.2 36.8 Max-DeepLab-L AX-L \checkmark 44.2 49.2 36.8 Max-DeepLab-L MaX-S 49.0 54.0 41.6 Max-DeepLab-L MaX-L \checkmark 44.2 49.2 36.8 Max-DeepLab-L MaX-S 49.0 54.0 41.6 Max-DeepLab-L MaX-S 49.0 54.0 41.6 Max-DeepLab-L MaX-S 49.0 54.0 41.6 Max-DeepLab-L MaX-Y 92 42.4 42.4 • Architecture Ablations. Input PQ PQ Th PQ St \checkmark \checkmark \checkmark 44.7 48.5 39.0 \checkmark \checkmark \checkmark 45.7 49.4 54.5 <td< td=""><td>Panoptic-FPN</td><td>RN-101</td><td></td><td>40.</td><td>9</td><td>48.3</td><td>29.7</td></td<>	Panoptic-FPN	RN-101		40.	9	48.3	29.7	
DetectoRS RX-101 \checkmark 49.6 57.8 37.1 Box-free panoptic segmentation methods Panoptic-DeepLab X-71 \checkmark 41.4 45.1 35.9 Axial-DeepLab-L AX-L 43.6 48.9 35.6 Axial-DeepLab-L AX-L \checkmark 44.2 49.2 36.8 MaX-DeepLab-L MaX-S 49.0 54.0 41.6 MaX-DeepLab-L MaX-L \checkmark 44.2 49.2 36.8 MaX-DeepLab-L MaX-S 49.0 54.0 41.6 MaX-DeepLab-L MaX-L \checkmark 44.2 49.2 36.8 MaX-DeepLab-L MaX-S 49.0 54.0 41.6 MaX-DeepLab-L MaX-Y \checkmark 42.4 42.4 • Architecture Ablations. \blacksquare \blacksquare \blacksquare \blacksquare \blacksquare \blacksquare \checkmark \checkmark \checkmark 44.7 48.5 39.0 \checkmark \checkmark \checkmark \checkmark 44.7 48.5 39.0 \checkmark \checkmark \checkmark \checkmark 45.7	DETR	RN-101		46.	0			
Box-free panoptic segmentation methodsPanoptic-DeepLabX-71✓41.445.135.9Axial-DeepLab-LAX-L✓43.648.935.6Axial-DeepLab-LAX-L✓44.249.236.8MaX-DeepLab-LMaX-S✓49.054.041.6MaX-DeepLab-LMaX-L 51.3 57.242.4• Architecture Ablations.M2PM2MP2MP2P AxialLarge InputPQPQ Th ✓✓44.748.539.0✓✓✓45.048.939.2✓✓✓✓49.454.541.8• Runtime (ms), on a V100 GPU.MethodBackboneInput SizeRuntimePQPanoptic-DeepLabX-71 641^2 7438.8MaX-DeepLab-SMaX-S 641^2 6746.7DETRRN-101 1333×800 12846.0Panoptic-DeepLabX-71 1025^2 13239.6	UPSNet						36.7	
Panoptic-DeepLab X-71 ✓ 41.4 45.1 35.9 Axial-DeepLab-L AX-L AX-L 43.6 48.9 35.6 Axial-DeepLab-L AX-L ✓ 44.2 49.2 36.8 MaX-DeepLab-L AX-L ✓ 44.2 49.2 36.8 MaX-DeepLab-L MaX-S 49.0 54.0 41.6 MaX-DeepLab-L MaX-L 51.3 57.2 42.4 • Architecture Ablations. MaX-L 51.3 57.2 42.4 • Architecture Ablations. Imput PQ PQ Th PQ St Attn Attn Attn Attention Input 44.7 48.5 39.0 ✓ ✓ ✓ ✓ 44.7 48.5 39.0 ✓ ✓ ✓ ✓ 45.0 48.9 39.2 ✓ ✓ ✓ ✓ 45.0 48.9 39.2 ✓ ✓ ✓ ✓ 45.0 48.5 39.0 ✓ ✓ ✓ ✓ ✓ 45.5 41.8	DetectoRS	RX-101		49.	6	57.8	37.1	
Axial-DeepLab-L AX-L 43.6 48.9 35.6 Axial-DeepLab-L AX-L ✓ 44.2 49.2 36.8 MaX-DeepLab-L MaX-S 49.0 54.0 41.6 MaX-DeepLab-L MaX-L 51.3 57.2 42.4 • Architecture Ablations. PQ PQTh PQSt Attn Attn Attn Attention Input PQ PQTh PQSt ✓ ✓ ✓ 44.7 48.5 39.0 45.0 48.9 39.2 ✓ ✓ ✓ ✓ 44.7 48.5 39.0 ✓ ✓ ✓ ✓ 44.7 48.5 39.0 ✓ ✓ ✓ ✓ 44.7 48.5 39.0 ✓ ✓ ✓ ✓ ✓ 45.0 48.9 39.2 ✓ ✓ ✓ ✓ ✓ 45.0 41.5 49.4 54.5 41.8 • Runtime (ms), on a V100 GPU. Mathod Backbone Input Size <	Box-free panoptic segmentation methods							
Axial-DeepLab-LAX-L \checkmark 44.249.236.8MaX-DeepLab-SMaX-S49.054.041.6MaX-DeepLab-LMaX-L51.357.242.4• Architecture Ablations.M2PM2MP2MP2P AxialLarge InputPQ PQ^{Th} PQ^{St} \checkmark \checkmark 44.748.539.0 \checkmark \checkmark \checkmark 44.748.539.0 \checkmark \checkmark \checkmark \checkmark 45.749.839.2 \checkmark \checkmark \checkmark \checkmark \checkmark 45.749.839.4 \checkmark \checkmark \checkmark \checkmark \checkmark 45.749.839.4 \checkmark \checkmark \checkmark \checkmark \checkmark 49.454.541.8• Runtime (ms), on a V100 GPU.Input SizeRuntimePQPQPanoptic-DeepLabX-71 641^2 74 38.8MaX-DeepLab-SMaX-S 641^2 6746.7DETR Panoptic-DeepLabX-71 1333×800 12846.0Panoptic-DeepLabX-71 1025^2 13239.6	Panoptic-DeepLab	X-71		41.	4	45.1	35.9	
MaX-DeepLab-S MaX-S 49.0 54.0 41.6 MaX-DeepLab-L MaX-L 51.3 57.2 42.4 • Architecture Ablations. PQ PQTh PQSt Attn Attn Attn Attention Input PQ PQTh PQSt \checkmark	Axial-DeepLab-L	AX-L		43.	6	48.9	35.6	
MaX-DeepLab-LMaX-L51.357.242.4• Architecture Ablations.M2PM2MP2MP2P AxialLarge InputPQ PQ^{Th} PQ^{St} \checkmark AttnAttnAttnAttentionInputPQ PQ^{Th} PQ^{St} \checkmark \checkmark \checkmark 44.748.539.0 \checkmark \checkmark \checkmark 45.0 48.939.2 \checkmark \checkmark \checkmark \checkmark 45.749.839.4 \checkmark \checkmark \checkmark \checkmark 47.8 51.941.5 \checkmark \checkmark \checkmark \checkmark \checkmark 49.454.541.8• Runtime (ms), on a V100 GPU.Input SizeRuntimePQPanoptic-DeepLabX-71 641^2 7438.8MaX-DeepLab-SMaX-S 641^2 6746.7DETR Panoptic-DeepLabX-71 1333×800 12846.0Panoptic-DeepLabX-71 1025^2 13239.6	Axial-DeepLab-L	AX-L		44.	2	49.2	36.8	
MaX-DeepLab-LMaX-L51.357.242.4• Architecture Ablations.M2PM2MP2MP2P AxialLarge InputPQ PQ^{Th} PQ^{St} \checkmark AttnAttnAttnAttentionInputPQ PQ^{Th} PQ^{St} \checkmark \checkmark \checkmark 44.748.539.0 \checkmark \checkmark \checkmark \checkmark 45.048.939.2 \checkmark \checkmark \checkmark \checkmark 45.7 49.839.4 \checkmark \checkmark \checkmark \checkmark \checkmark 47.851.941.5 \checkmark \checkmark \checkmark \checkmark \checkmark 49.4 54.541.8• Runtime (ms), on a V100 GPU.Max-s64127438.8MaX-DeepLabX-71 641^2 6746.7DETR Panoptic-DeepLabRN-101 1333×800 12846.0Panoptic-DeepLabX-71 1025^2 13239.6	MaX-DeepLab-S	MaX-S					41.6	
M2P AttnM2M AttnP2M AttnP2P Axial AttentionLarge InputPQ PQ PQThPQSt </td <td>MaX-DeepLab-L</td> <td>MaX-L</td> <td></td> <td>51.</td> <td>3</td> <td>57.2</td> <td>42.4</td>	MaX-DeepLab-L	MaX-L		51.	3	57.2	42.4	
Attn Attn Attention Input $1 \cdot Q \cdot 1 \cdot Q \cdot Q$								
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	M2P M2M P2M Attn Attn Attn	P2P Axia Attention	l Larg I Inpu	je it	PQ	PQ Th	PQ St	
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$				4	4.7	48.5	39.0	
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$				4	5.0	48.9	39.2	
\checkmark \checkmark \checkmark 49.4 54.5 41.8 • Runtime (ms), on a V100 GPU.MethodBackboneInput SizeRuntimePQPanoptic-DeepLabX-71 641^2 74 38.8 MaX-DeepLab-SMaX-S 641^2 67 46.7 DETRRN-101 1333×800 128 46.0 Panoptic-DeepLabX-71 1025^2 132 39.6				4	15.7	49.8	39.4	
• Runtime (ms), on a V100 GPU.MethodBackboneInput SizeRuntimePQPanoptic-DeepLabX-71 641^2 7438.8MaX-DeepLab-SMaX-S 641^2 6746.7DETRRN-101 1333×800 12846.0Panoptic-DeepLabX-71 1025^2 13239.6				4	17.8	51.9	41.5	
MethodBackboneInput SizeRuntimePQPanoptic-DeepLabX-71 641^2 7438.8MaX-DeepLab-SMaX-S 641^2 6746.7DETRRN-101 1333×800 12846.0Panoptic-DeepLabX-71 1025^2 13239.6				4	9.4	54.5	41.8	
Panoptic-DeepLabX-71 641^2 7438.8MaX-DeepLab-SMaX-S 641^2 6746.7 DETRRN-101 1333×800 12846.0Panoptic-DeepLabX-71 1025^2 13239.6	• Runtime (ms), on a V100 GPU.							
MaX-DeepLab-SMaX-S 641^2 6746.7DETR Panoptic-DeepLabRN-101 1333×800 12846.0X-71 1025^2 13239.6	Method	Backbone	Input S	bize	Ru	ntime	PQ	
MaX-DeepLab-SMaX-S 641^2 6746.7DETRRN-101 1333×800 12846.0Panoptic-DeepLabX-71 1025^2 13239.6	Panoptic-DeepLab	X-71	-		74			
Panoptic-DeepLabX-71 1025^2 13239.6		MaX-S	641^{2}		67		46.7	
	DETR	RN-101	1333×800			128	46.0	
MaX-DeepLab-S MaX-S 1025 ² 131 49.0	Panoptic-DeepLab	X-71	1025^{2}		132		39.6	
	MaX-DeepLab-S	MaX-S	1025^{2}	2		131	49.0	

