MaX-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers

Huiyu Wang[1] Yukun Zhu[2] Hartwig Adam[2] Alan Yuille[1] Liang-Chieh Chen[2]

[1]Johns Hopkins University [2]Google Research

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

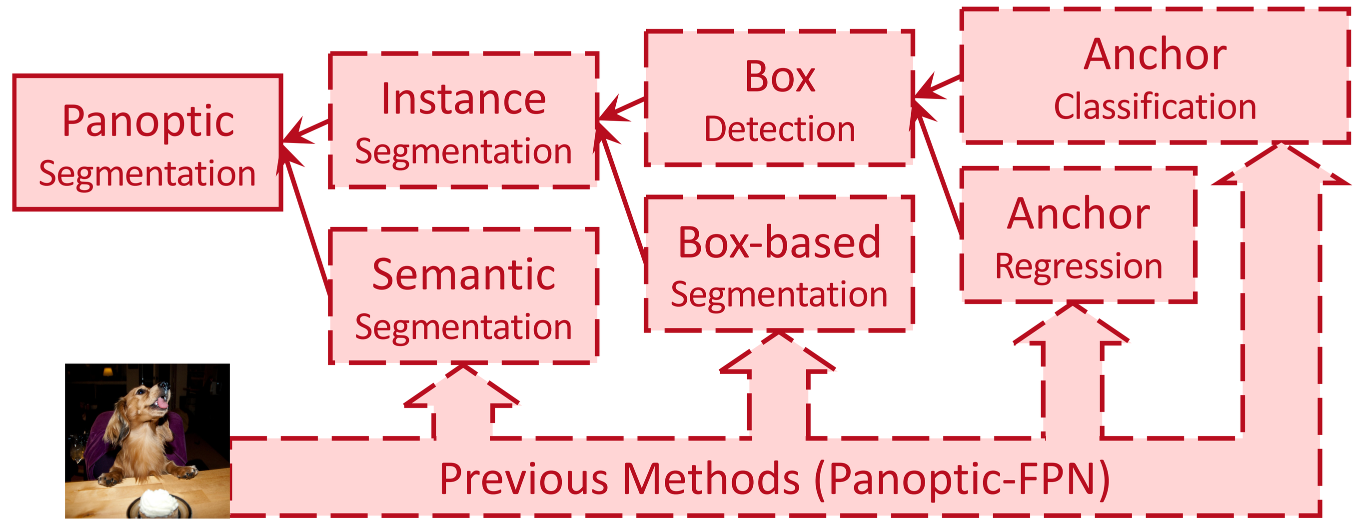
JOHNS HOPKINS UNIVERSITY

Google Research

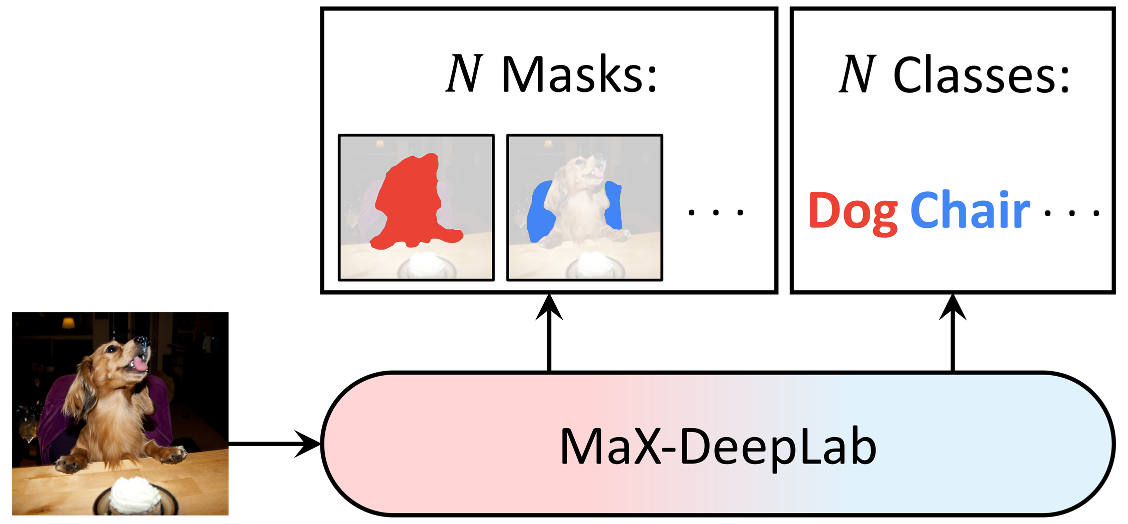
CVPR VIRTUAL JUNE 19-25

Motivation

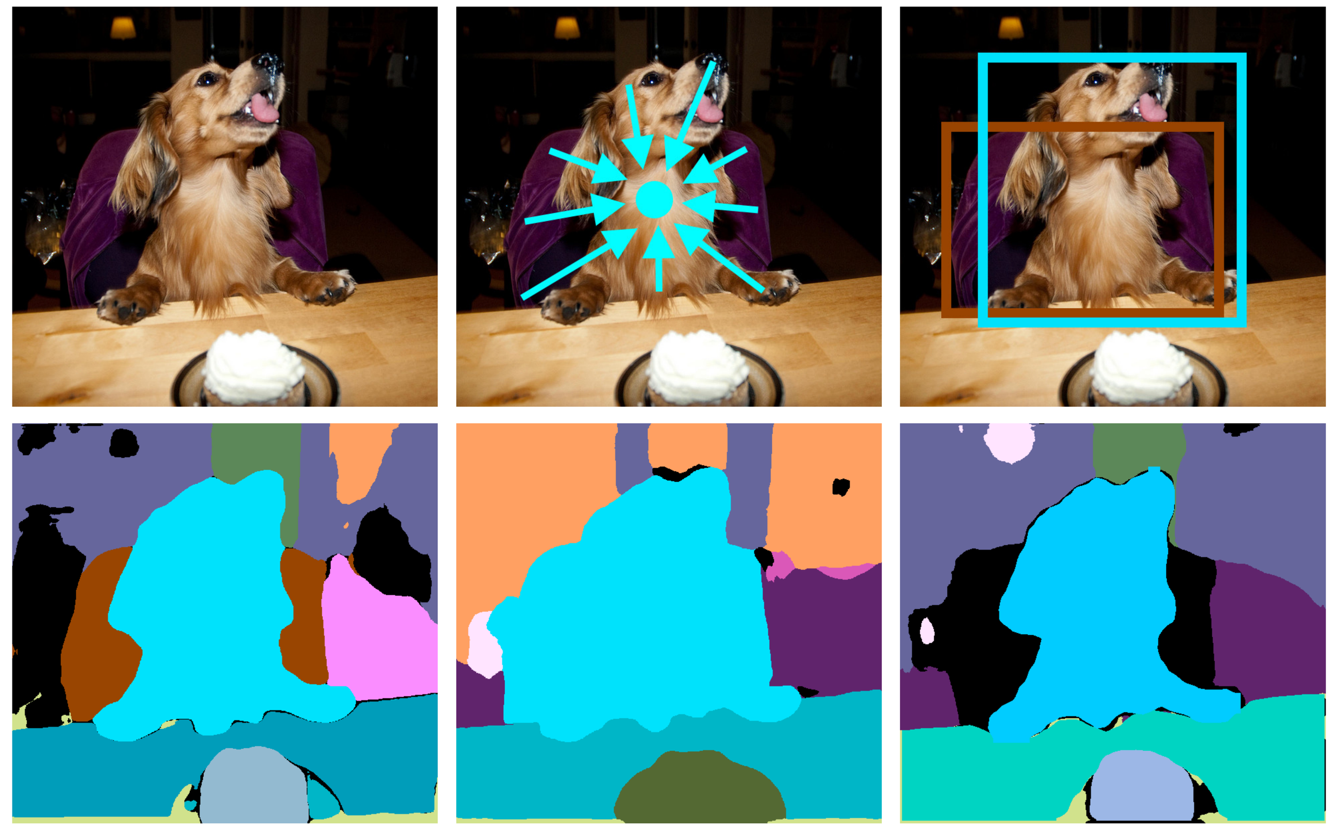
- **Previous methods rely on a tree of surrogate sub-tasks.** Although these sub-tasks are tackled by area experts, they fail to comprehensively solve the target task.



- **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

Formulation

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

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

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

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

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

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 $\varnothing$).

Simple Inference

- **Predict a class-ID for each mask:**

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

- **Predict a mask-ID for each pixel:**

$$\hat{z}_{h,w} = \arg\max_i \hat{m}_{i,h,w}\,, \quad \forall h \in \{1,2,\ldots,H\}, \quad \forall w \in \{1,2,\ldots,W\}\,. \tag{4}$$

Training with PQ-Style Loss

- **Inspired by Panoptic Quality (PQ) decomposition:**

$$\mathrm{PQ} = \mathrm{RQ} \times \mathrm{SQ}\,, \tag{5}$$

RQ: recognition quality, SQ: segmentation quality.

- **A PQ-style similarity metric** is defined between a ground truth class-labeled mask and a predicted mask:

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

- **Match predictions to GTs with the metric:**

$$\hat{\sigma} = \arg\max_{\sigma \in \mathfrak{S}_N} \sum_{i=1}^K \mathrm{sim}(y_i, \hat{y}_{\sigma(i)})\,, \tag{7}$$

$\sigma \in \mathfrak{S}_N$: a permutation of N elements,
$\hat{\sigma}$: the optimal permutation with maximum similarity.

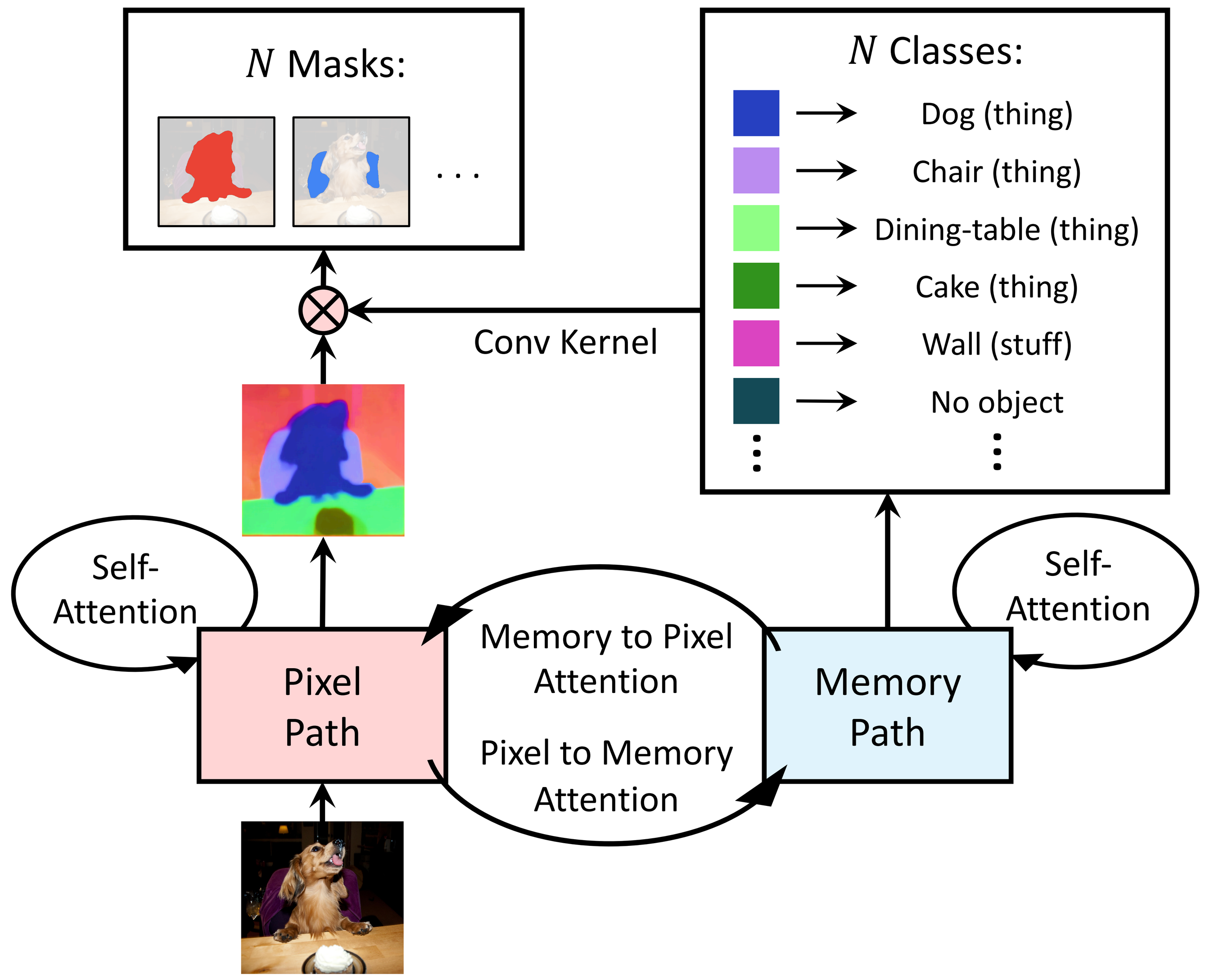
- **Optimize the model with the same metric:**

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

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

Architecture

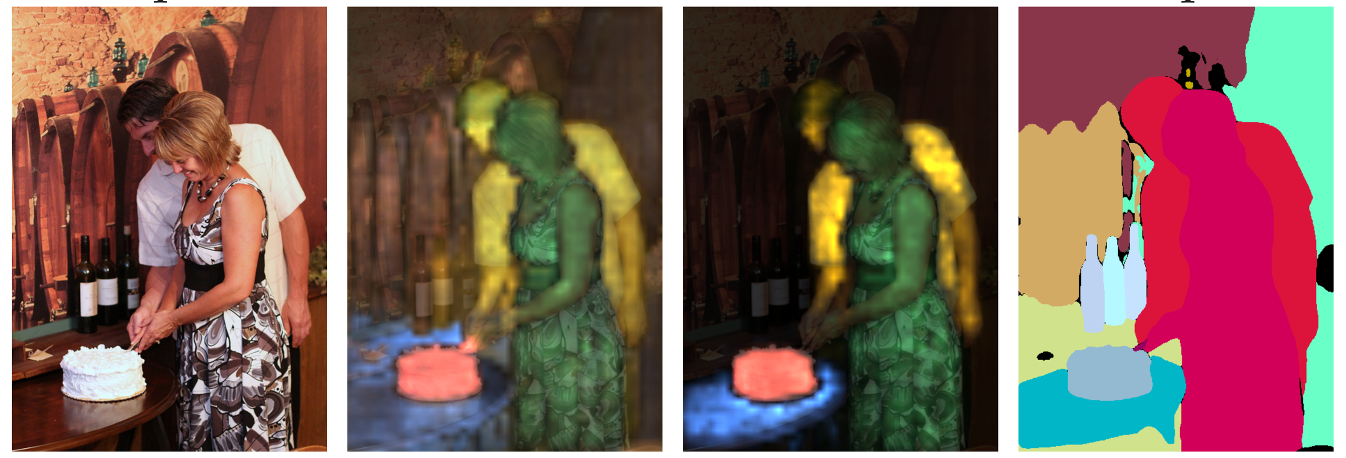
- **An overview of the mask transformer architecture.**



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

Input | Head 1 | Head 2 | Output



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

Results

- **COCO test-dev set**, TTA: Test-time augmentation.

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					
Panoptic-DeepLab	X-71	✓	41.4	45.1	35.9
Axial-DeepLab-L	AX-L		43.6	48.9	35.6
Axial-DeepLab-L	AX-L	✓	44.2	49.2	36.8
MaX-DeepLab-S	MaX-S		49.0	54.0	41.6
MaX-DeepLab-L	MaX-L		**51.3**	57.2	42.4

- **Architecture Ablations.**

M2P Attn	M2M Attn	P2M Attn	P2P Axial Attention	Large Input	PQ	PQ[Th]	PQ[St]
✓					44.7	48.5	39.0
✓	✓				45.0	48.9	39.2
✓	✓	✓			45.7	49.8	39.4
✓	✓	✓	✓		47.8	51.9	41.5
✓	✓	✓	✓	✓	**49.4**	**54.5**	**41.8**

- **Runtime (ms)**, on a V100 GPU.

Method	Backbone	Input Size	Runtime	PQ
Panoptic-DeepLab	X-71	641^2	74	38.8
MaX-DeepLab-S	MaX-S	641^2	**67**	**46.7**
DETR	RN-101	1333×800	128	46.0
Panoptic-DeepLab	X-71	1025^2	132	39.6
MaX-DeepLab-S	MaX-S	1025^2	**131**	**49.0**

Visualizations

Input | Output | Input | Output | Input | Output