Per-Pixel Classification is Not All You Need for Semantic Segmentation

Bowen Cheng, Alexander G. Schwing, Alexander Kirillov ---- NIPS 2021

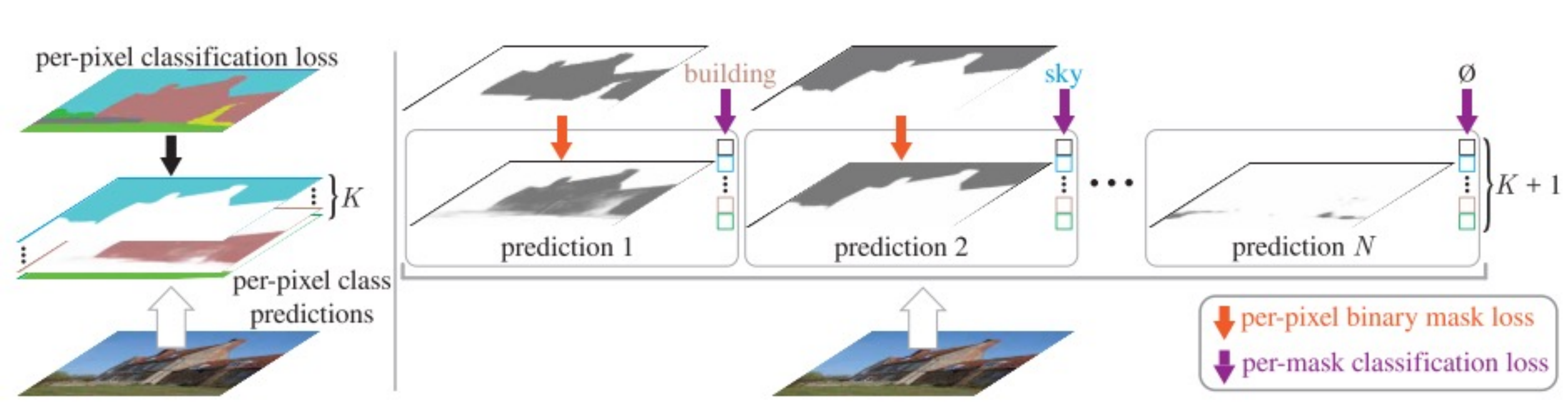
Fang Zhiyuan

2021.12/05

Prior knowledge

Per-pixel classification vs. mask classification

- **Per-pixel classification** applies the same classification loss to each location
  - Often used in Semantic Segmentation
- **Mask classification** predicts a set of binary masks and assigns a single class to each mask
  - Often used in Instance Segmentation



MaskFormer

- Semantic/Panoptic Segmentation algorithm based on Mask Classification
- Converts any existing per-pixel classification model into a mask classification
- Solves both semantic- and instance-level segmentation tasks in a unified manner
  - Do not change the model, losses, and training procedure

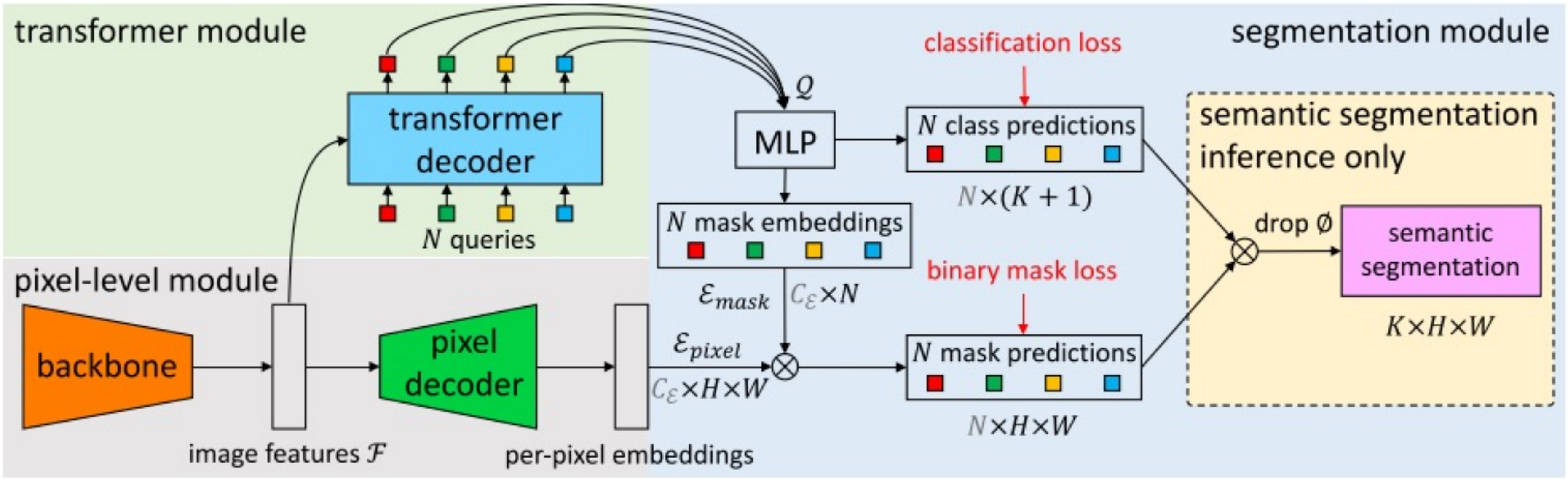
MaskFormer

Mask classification formulation

- Partitioning/grouping the image into ***N*** regions
- Output $z = \{(p_i, m_i)\}_{i=1}^{N}$
  - $p_i \in \Delta^{K+1}$ is the probability distribution, K categories + Ø (no object)
  - $m_i \in [0,1]^{H\times W}$
- Ground truth segment $z^{gt} = \left\{\left(c_i^{gt}, m_i^{gt}\right) \middle| c_i^{gt} \in \{1, \cdots, K\}, m_i^{gt} \in \{0,1\}^{H\times W}\right\}_{i=1}^{N_{gt}}$
- Associating each region as a whole with some distribution with matching σ
- Loss: $\mathcal{L}_{mask-cls}(z, z^{gt}) = \sum_{j=1}^{N}\left[-\log p_{\sigma(j)}(c_j^{gt}) + 1_{c_j^{gt}\neq\emptyset}\mathcal{L}_{mask}(m_{\sigma(j)}, m_j^{gt})\right]$

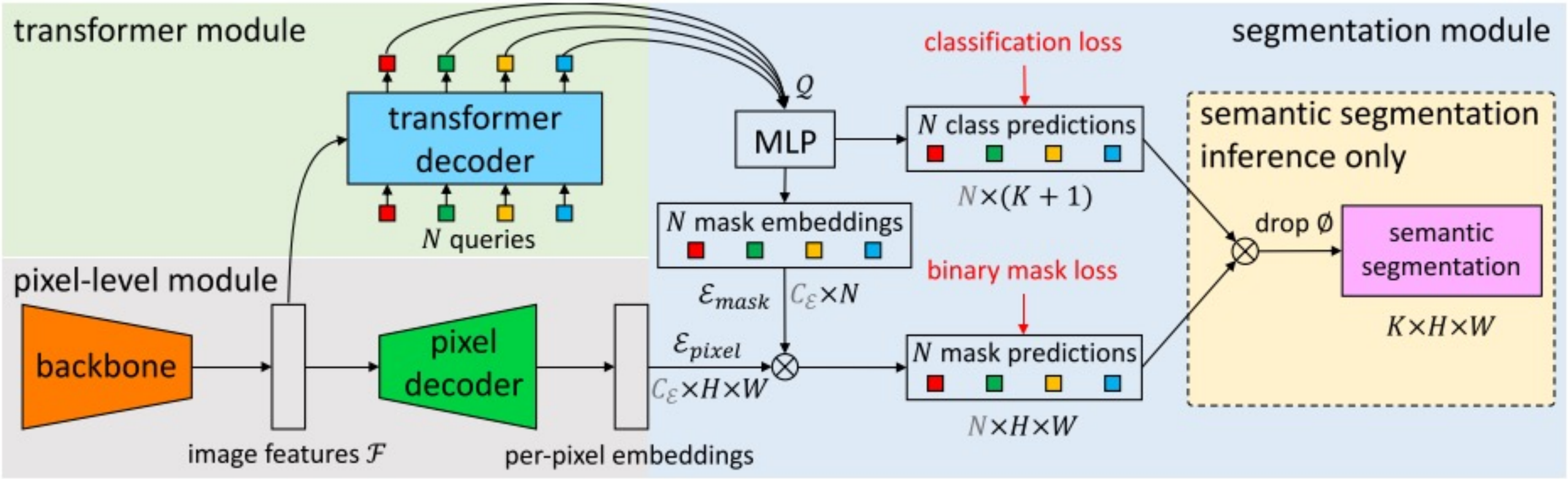
MaskFormer

- Transformer module
- Pixel-level module
- Segmentation module (training)
  - Linear classify -> class probability predictions $\{p_i \in \Delta^{K+1}\}_{i=1}^{N}$
  - MLP -> mask embeddings $\varepsilon_{mask} \in C_{\varepsilon} \times N$
  - $\mathcal{L}_{cls}$: Cross-entropy loss, $\mathcal{L}_{mask} = \lambda_{focal} \cdot l_{focal} + \lambda_{dice} \cdot l_{dice}$



MaskFormer

- Segmentation module (inference)
- **General**
  - For each output $\{(p_i, m_i)\}_{i=1}^{N}$,
    - $\underset{i:c_i \neq \emptyset}{\arg\max}\, p_i(c_i) \cdot m_i[h,w]$, $c_i = \underset{c \in \{1,\ldots,K,\emptyset\}}{\arg\max}\ p_i(c)$
  - Post process for panoptic segmentation: filter out low-confidence, NMS
- **Semantic**
  - Drop $\emptyset$ and simple matrix multiplication



Experiments

Dataset

- Semantic Segmentation
  - ADE20K, COCO-Stuff-10K, Cityscapes, Mapillary Vistas
  - 8 V100 GPUs
- Panoramic segmentation
  - COCO (64 V100 GPUs)
  - ADE20K-Panoptic (8 V100 GPUs)

Experiments

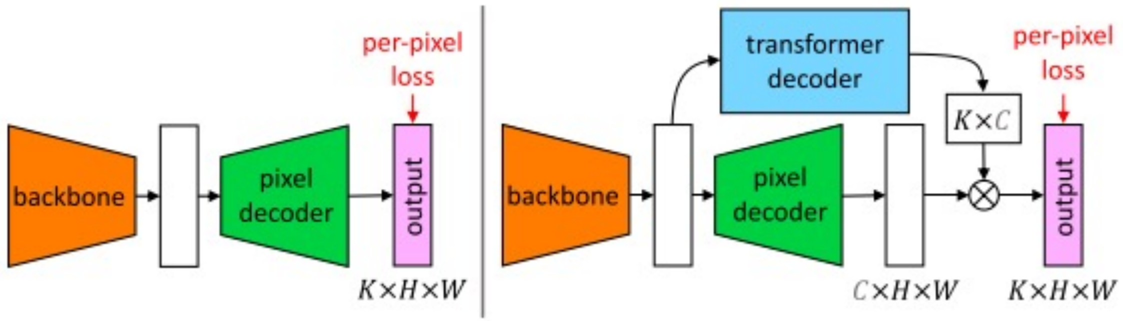
Semantic Segmentation on ADE20K

	method	backbone	crop size	mIoU (s.s.)	mIoU (m.s.)	#params.	FLOPs	fps
CNN backbones	OCRNet [50]	R101c	520 × 520	-	45.3	-	-	-
	DeepLabV3+ [9]	R50c	512 × 512	44.0	44.9	44M	177G	21.0
		R101c	512 × 512	45.5	46.4	63M	255G	14.2
	MaskFormer (ours)	R50	512 × 512	44.5 ±0.5	46.7 ±0.6	41M	53G	24.5
		R101	512 × 512	45.5 ±0.5	47.2 ±0.2	60M	73G	19.5
		R101c	512 × 512	**46.0** ±0.1	**48.1** ±0.2	60M	80G	19.0
Transformer backbones	SETR [53]	ViT-L†	512 × 512	-	50.3	308M	-	-
	Swin-UperNet [29, 49]	Swin-T	512 × 512	-	46.1	60M	236G	18.5
		Swin-S	512 × 512	-	49.3	81M	259G	15.2
		Swin-B†	640 × 640	-	51.6	121M	471G	8.7
		Swin-L†	640 × 640	-	53.5	234M	647G	6.2
	MaskFormer (ours)	Swin-T	512 × 512	46.7 ±0.7	48.8 ±0.6	42M	55G	22.1
		Swin-S	512 × 512	49.8 ±0.4	51.0 ±0.4	63M	79G	19.6
		Swin-B	640 × 640	51.1 ±0.2	52.3 ±0.4	102M	195G	12.6
		Swin-B†	640 × 640	52.7 ±0.4	53.9 ±0.2	102M	195G	12.6
		Swin-L†	640 × 640	**54.1** ±0.2	**55.6** ±0.1	212M	375G	7.9

Experiments

Compare with baseline

	Cityscapes (19 classes)		ADE20K (150 classes)		COCO-Stuff (171 classes)		ADE20K-Full (847 classes)	
	mIoU	PQ^{St}	mIoU	PQ^{St}	mIoU	PQ^{St}	mIoU	PQ^{St}
PerPixelBaseline	77.4	58.9	39.2	21.6	32.4	15.5	12.4	5.8
PerPixelBaseline+	**78.5**	60.2	41.9	28.3	34.2	24.6	13.9	9.0
MaskFormer (ours)	**78.5** (+0.0)	**63.1** (+2.9)	**44.5** (+2.6)	**33.4** (+5.1)	**37.1** (+2.9)	**28.9** (+4.3)	**17.4** (+3.5)	**11.9** (+2.9)



(a) PerPixelBaseline (b) PerPixelBaseline+

Experiments

Panoptic Segmentation on COCO

	method	backbone	PQ	PQ^{Th}	PQ^{St}	SQ	RQ	#params.	FLOPs	fps
CNN backbones	DETR [4]	R50 + 6 Enc	43.4	48.2	36.3	79.3	53.8	-	-	-
	MaskFormer (DETR)	R50 + 6 Enc	45.6	50.0 (+1.8)	39.0 (+2.7)	80.2	55.8	-	-	-
	MaskFormer (ours)	R50 + 6 Enc	**46.5**	**51.0** (+2.8)	**39.8** (+3.5)	**80.4**	**56.8**	45M	181G	17.6
	DETR [4]	R101 + 6 Enc	45.1	50.5	37.0	79.9	55.5	-	-	-
	MaskFormer (ours)	R101 + 6 Enc	**47.6**	**52.5** (+2.0)	**40.3** (+3.3)	**80.7**	**58.0**	64M	248G	14.0
Transformer backbones	Max-DeepLab [42]	Max-S	48.4	53.0	41.5	-	-	62M	324G	7.6
		Max-L	51.1	57.0	42.2	-	-	451M	3692G	-
	MaskFormer (ours)	Swin-T	47.7	51.7	41.7	80.4	58.3	42M	179G	17.0
		Swin-S	49.7	54.4	42.6	80.9	60.4	63M	259G	12.4
		Swin-B	51.1	56.3	43.2	81.4	61.8	102M	411G	8.4
		Swin-B†	51.8	56.9	**44.1**	81.4	62.6	102M	411G	8.4
		Swin-L†	**52.7**	**58.5**	**44.0**	**81.8**	**63.5**	212M	792G	5.2

Experiments

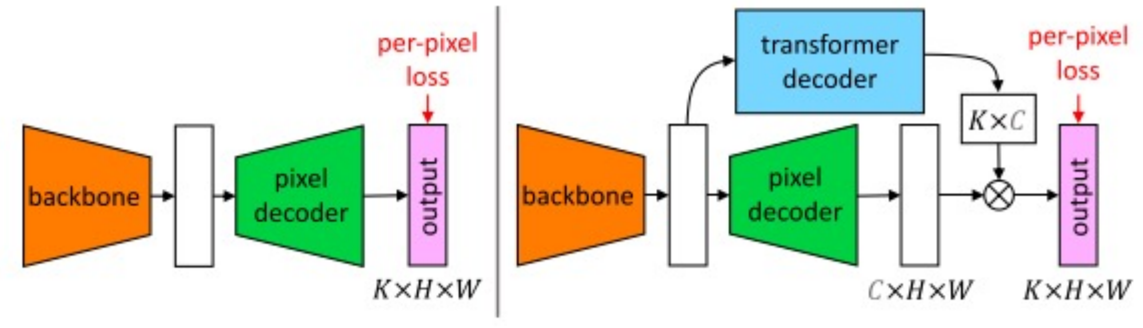
Ablation Study

(a) Per-pixel *vs*. mask classification.

	mIoU	PQ^{St}
PerPixelBaseline+	41.9	28.3
MaskFormer-fixed	**43.7** (+1.8)	**30.3** (+2.0)

(b) Fixed *vs*. bipartite matching assignment.

	mIoU	PQ^{St}
MaskFormer-fixed	43.7	30.3
MaskFormer-bipartite (ours)	**44.2** (+0.5)	**33.4** (+3.1)



(a) PerPixelBaseline (b) PerPixelBaseline+