Masked-attention Mask Transformer for Universal Image Segmentation (Mask2Former)

Presented by Louie Lu, Chongyi Zheng, Mingcheng Hu

Summary of the Arguments

1. One Model Rule Them All
2. Improve Training Efficiency
3. Better Results

Argument #1 One Model Rule Them All

- Less training time
- Reuse one model one multiple image segmentation tasks

Argument #2 Improve Training Efficiency

Memory requirement reduced by 3x

Balance computation with model performance, especially for small objects

Argument #3 Better Results

The Mask2Former outperforms SegFormer on both Cityscapes and ADE20K datasets. Particularly, it obtains a 6% mIOU gain on the latter, setting a new-state-of-the-art by 2021 without being computational burdensome.

Cityscapes Dataset	Mask2Former	SegFormer
Crop Size	512*1024	1024*1024
Training Iterations	90K	160K
Do Infernce on	1024*2048 (whole image)	1024*1024
Batch Size	16	8
mIOU	84.5(Swin-B)	84
Params	107M (Swin-B)	84.7M

Semantic Segmentation on ADE20K val

15	**Mask2Former** (Swin-L-FaPN, multiscale)	57.7	×	Masked-attention Mask Transformer for Universal Image Segmentation	2021
22	**Mask2Former** (Swin-L-FaPN)	56.4	×	Masked-attention Mask Transformer for Universal Image Segmentation	2021
39	**SegFormer-B5** (MS, 87M #Params, ImageNet-1K pretrain)	51.8	×	SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers	2021

		panoptic model				instance model		semantic model	
method	backbone	PQ (s.s.)	PQ (m.s.)	AP^{Th}_{pan}	$mIoU_{pan}$	AP	AP50	mIoU (s.s.)	mIoU (m.s.)
Panoptic-DeepLab [11]	R50	60.3	-	32.1	78.7	-	-	-	-
	X71 [15]	63.0	64.1	35.3	80.5	-	-	-	-
	SWideRNet [9]	66.4	67.5	40.1	82.2	-	-	-	-
Panoptic FCN [31]	Swin-L†	65.9	-	-	-	-	-	-	-
Segmenter [45]	ViT-L†	-	-	-	-	-	-	-	81.3
SETR [64]	ViT-L†	-	-	-	-	-	-	-	82.2
SegFormer [59]	MiT-B5	-	-	-	-	-	-	-	84.0
Mask2Former (ours)	R50	62.1	-	37.3	77.5	37.4	61.9	79.4	82.2
	R101	62.4	-	37.7	78.6	38.5	63.9	80.1	81.9
	Swin-T	63.9	-	39.1	80.5	39.7	66.9	82.1	83.0
	Swin-S	64.8	-	40.7	81.8	41.8	70.4	82.6	83.6
	Swin-B†	66.1	-	42.8	82.7	42.0	68.8	**83.3**	**84.5**
	Swin-L†	**66.6**	-	**43.6**	**82.9**	**43.7**	**71.4**	**83.3**	84.3

ADE20K Dataset	Mask2Former	SegFormer
Crop Size	640*640	512*512
Training Iterations	90K	160K
Batch Size	16	16
mIOU	57.3 (Swin-L)	51.8
Params	215M	84.7M
Flops	403G	183.3G

Summary of the Arguments

1. One Model Rule Them All
2. Improve Training Efficiency
3. Better Results

SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers

Authors: Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo

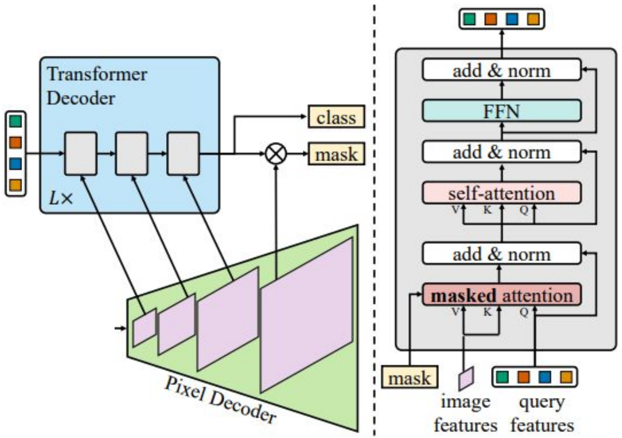
Submitted May 31, 2021; Last Revised October 28, 2021

Presented By: Alex Georgiev, David Zhang, Pan Lu

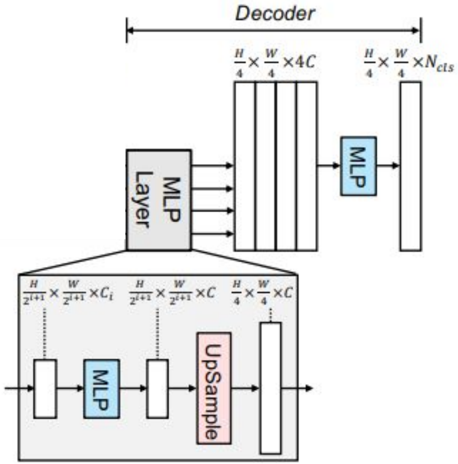
Arguments

Simple and lightweight design

- No widely-used tricks, such as auxiliary losses
- No positional encoding, so no interpolation when dealing with higher resolution images
- Lightweight decoder only has at most 3.3M parameters whereas theirs has ~20M
  - Our decoder only consist of MLP layers while theirs uses a transformer



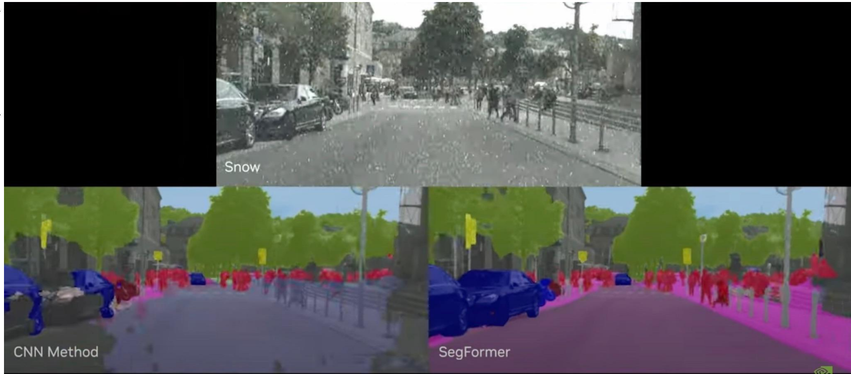
vs



Can be used for latency-critical real-time applications

- Our B0 model achieves a high mIoU and high FPS with a much lower number of FLOPS and only 3.8M parameters
- Robust to common corruptions such as weather conditions

	Method	Encoder	Params ↓	ADE20K			Cityscapes		
				Flops ↓	FPS ↑	mIoU ↑	Flops ↓	FPS ↑	mIoU ↑
Real-Time	FCN [1]	MobileNetV2	9.8	39.6	64.4	19.7	317.1	14.2	61.5
	ICNet [11]	-	-	-	-	-	-	30.3	67.7
	PSPNet [17]	MobileNetV2	13.7	52.9	57.7	29.6	423.4	11.2	70.2
	DeepLabV3+ [20]	MobileNetV2	15.4	69.4	43.1	34.0	555.4	8.4	75.2
	SegFormer (Ours)	MiT-B0	**3.8**	**8.4**	**50.5**	**37.4**	125.5	15.2	**76.2**
				-	-	-	51.7	26.3	75.3
				-	-	-	31.5	37.1	73.7
				-	-	-	**17.7**	**47.6**	71.9



Comparable performance despite earlier publication

Table 1: Ablation studies related to model size, encoder and decoder design.

(a) Accuracy, parameters and flops as a function of the model size on the three datasets. "SS" and "MS" means single/multi-scale test.

Encoder	Params		ADE20K		Cityscapes		COCO-Stuff	
Model Size	Encoder	Decoder	Flops ↓	mIoU(SS/MS) ↑	Flops ↓	mIoU(SS/MS) ↑	Flops ↓	mIoU(SS) ↑
MiT-B0	3.4	0.4	8.4	37.4 / 38.0	125.5	76.2 / 78.1	8.4	35.6
MiT-B1	13.1	0.6	15.9	42.2 / 43.1	243.7	78.5 / 80.0	15.9	40.2
MiT-B2	24.2	3.3	62.4	46.5 / 47.5	717.1	81.0 / 82.2	62.4	44.6
MiT-B3	44.0	3.3	79.0	49.4 / 50.0	962.9	81.7 / 83.3	79.0	45.5
MiT-B4	60.8	3.3	95.7	50.3 / 51.1	1240.6	82.3 / 83.9	95.7	46.5
MiT-B5	81.4	3.3	183.3	51.0 / 51.8	1460.4	82.4 / 84.0	111.6	46.7

		panoptic model				instance model		semantic model	
method	backbone	PQ (s.s.)	PQ (m.s.)	AP^{Th}_{pan}	$mIoU_{pan}$	AP	AP50	mIoU (s.s.)	mIoU (m.s.)
Panoptic-DeepLab [11]	R50	60.3	-	32.1	78.7	-	-	-	-
	X71 [15]	63.0	64.1	35.3	80.5	-	-	-	-
	SWideRNet [9]	66.4	67.5	40.1	82.2	-	-	-	-
Panoptic FCN [31]	Swin-L†	65.9	-	-	-	-	-	-	-
Segmenter [45]	ViT-L†	-	-	-	-	-	-	-	81.3
SETR [64]	ViT-L†	-	-	-	-	-	-	-	82.2
SegFormer [59]	MiT-B5	-	-	-	-	-	-	-	84.0
Mask2Former (ours)	R50	62.1	-	37.3	77.5	37.4	61.9	79.4	82.2
	R101	62.4	-	37.7	78.6	38.5	63.9	80.1	81.9
	Swin-T	63.9	-	39.1	80.5	39.7	66.9	82.1	83.0
	Swin-S	64.8	-	40.7	81.8	41.8	70.4	82.6	83.6
	Swin-B†	66.1	-	42.8	82.7	42.0	68.8	**83.3**	**84.5**
	Swin-L†	**66.6**	-	**43.6**	**82.9**	**43.7**	**71.4**	**83.3**	84.3

MiT-B5: 84.7M
M2F-Swin-B: 107M

MiT-B4: 64.1M
M2F-Swin-S: 69M

Encoder	Params		ADE20K		Cityscapes		COCO-Stuff	
Model Size	Encoder	Decoder	Flops ↓	mIoU(SS/MS) ↑	Flops ↓	mIoU(SS/MS) ↑	Flops ↓	mIoU(SS) ↑
MiT-B0	3.4	0.4	8.4	37.4 / 38.0	125.5	76.2 / 78.1	8.4	35.6
MiT-B1	13.1	0.6	15.9	42.2 / 43.1	243.7	78.5 / 80.0	15.9	40.2
MiT-B2	24.2	3.3	62.4	46.5 / 47.5	717.1	81.0 / 82.2	62.4	44.6
MiT-B3	44.0	3.3	79.0	49.4 / 50.0	962.9	81.7 / 83.3	79.0	45.5
MiT-B4	60.8	3.3	95.7	50.3 / 51.1	1240.6	82.3 / 83.9	95.7	46.5
MiT-B5	81.4	3.3	183.3	51.0 / 51.8	1460.4	82.4 / 84.0	111.6	46.7

	method	backbone	crop size	mIoU (s.s.)	mIoU (m.s.)	#params.	FLOPs
CNN	MaskFormer [14]	R50	512 × 512	44.5	46.7	41M	53G
		R101	512 × 512	45.5	47.2	60M	73G
	Mask2Former (ours)	R50	512 × 512	47.2	49.2	44M	71G
		R101	512 × 512	**47.8**	**50.1**	63M	90G
Transformer backbones	Swin-UperNet [36, 58]	Swin-L†	640 × 640	-	53.5	234M	647G
	FaPN-MaskFormer [14, 39]	Swin-L†	640 × 640	55.2	56.7	-	-
	BEiT-UperNet [2, 58]	BEiT-L†	640 × 640	-	57.0	502M	-
	MaskFormer [14]	Swin-T	512 × 512	46.7	48.8	42M	55G
		Swin-S	512 × 512	49.8	51.0	63M	79G
		Swin-B	640 × 640	51.1	52.3	102M	195G
		Swin-B†	640 × 640	52.7	53.9	102M	195G
		Swin-L†	640 × 640	54.1	55.6	212M	375G
	Mask2Former (ours)	Swin-T	512 × 512	47.7	49.6	47M	74G
		Swin-S	512 × 512	51.3	52.4	69M	98G
		Swin-B	640 × 640	52.4	53.7	107M	223G
		Swin-B†	640 × 640	53.9	55.1	107M	223G
		Swin-L†	640 × 640	56.1	57.3	215M	403G
		Swin-L-FaPN†	640 × 640	**56.4**	**57.7**	217M	-

MiT-B3 performs better than Mask2-Swin-T with the same number of parameters