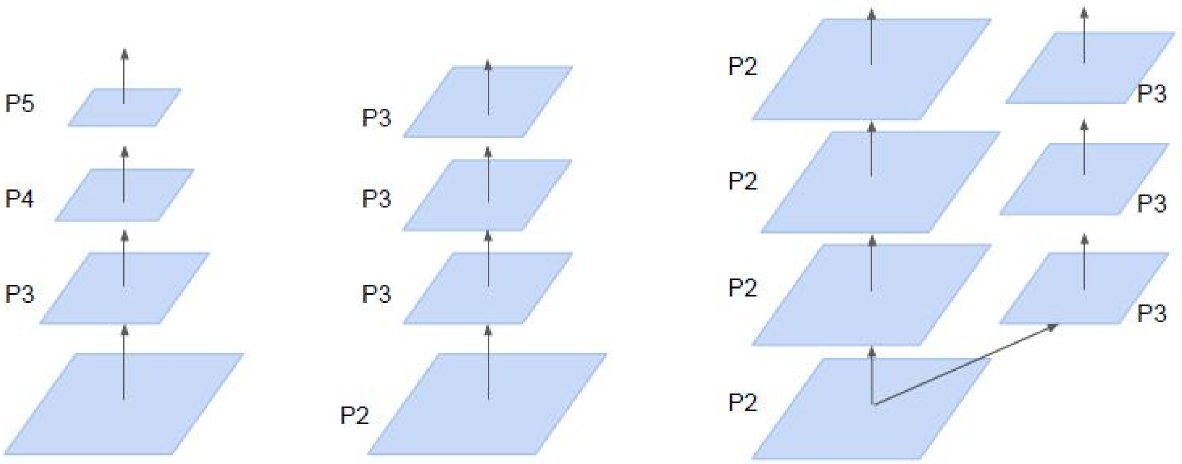
Revisiting Multi-Scale Feature Fusion for Semantic Segmentation

Tianjian Meng [1] Golnaz Ghiasi [1] Reza Mahjorian [2] Quoc V. Le [1] Mingxing Tan [1]
[1] Google Research [2] Waymo Inc.
{mengtianjian, tanmingxing}@google.com

Background Knowledge

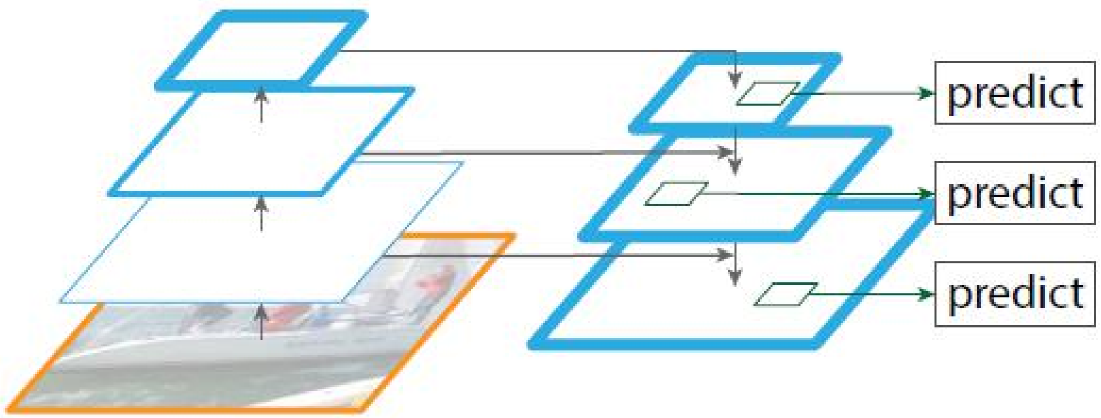
- Atrous Conv.
  - Problem
    - Hardware unfriendly
    - Long-ranged information could be irrelevant
  - Commonly Used Structure



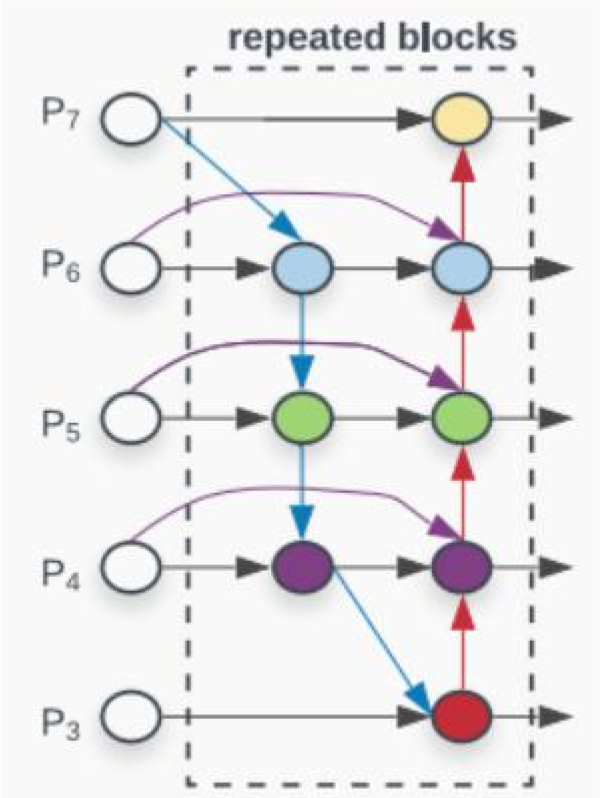
(a) Conventional P3-P5 (b) DeepLab with atrous convs (c) HRNet

Background Knowledge

- FPN vs BiFPN



FPN



BiFPN

*EfficientDet:Scalable and Efficient Object Detection

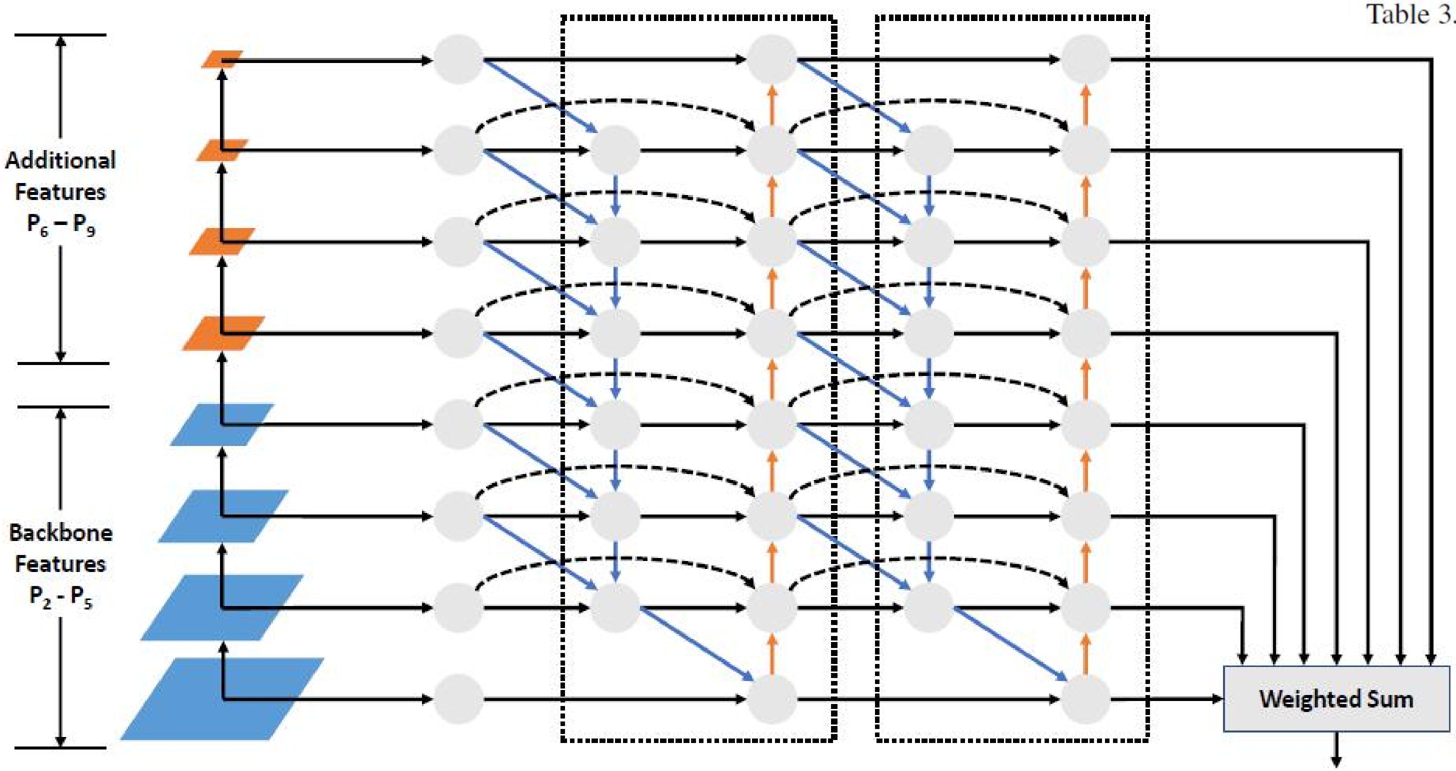
Introduction

- Motivation
  - High resolution features require expensive compution and memory
  - Regular conv.s are difficult to obtain large receptive fields
  - Semantics of each pixel depend on both nearby and far-away context
- Contributions
  - Deeper feature extractor
  - Richer feature fusion

Method

Model	Encoder		Decoder	
	Width	Depth	# channels	# repeats
ESeg-Lite-S	0.4	0.6	64	1
ESeg-Lite-M	0.6	1.0	80	2
ESeg-Lite-L	1.0	1.0	96	3
ESeg-S	1.0	1.1	96	4
ESeg-M	1.4	1.8	192	5
ESeg-L	2.0	3.1	288	6

Table 3. **Network size scaling configurations**.



$$O = \sum_i \frac{e^{w_i}}{\sum_j e^{w_j}} \cdot \mathrm{Upsample}(P_i)$$

Figure 3. **ESeg network architecture**. The backbone [36] extracts $\{P_2 - P_5\}$ feature maps from the raw input images; Four additional feature maps $\{P_6 - P_9\}$ are added on top of these backbone features with simple average pooling. The decoder perform bidirectional multi-scale feature fusion [38] to strength the internal representations for each feature map. All feature maps are upsampled and combined with weighted sum to generate the final per-pixel prediction.

Experiments

- Setting
  - 8 TPU 16 BS
  - Cosine LR decay
  - OHEM Strategy
  - Metrics: mIoU and pixACC (pixel accuracy)
  - Dataset: Cityscapes, ADE20K

Experiments

- Compare with SOTA

Rank	Model	Mean IoU (class)	Category mIoU	Time (ms)	Extra Training Data	Paper	Code	Result	Year	Tags
1	HRNetV2 + OCR +	84.5%			✓	Segmentation Transformer: Object-Contextual Representations for Semantic Segmentation			2019	hrnet
2	Lawin+	84.4%			×	Lawin Transformer: Improving Semantic Segmentation Transformer with Multi-Scale Representations via Large Window Attention			2022	Transformer
3	EfficientPS	84.21%			✓	EfficientPS: Efficient Panoptic Segmentation			2020	
4	Panoptic-DeepLab	84.2%			✓	Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation			2019	

Model	val mIoU w/o extra data	val mIoU w/ extra data	Params	Ratio	FLOPs	Ratio
ESeg-S	**80.1**	**81.7**	**6.9M**	**1x**	**34.5B**	**1x**
Auto-DeepLab-S [24]	79.7	-	10.2M	1.5x	333B	9.7x
PSPNet (ResNet-101) [50]	79.7	-	65.9M	9.6x	2018B	59x
OCR (ResNet-101) [45]	79.6	-	-	-	-	-
DeepLabV3+ (Xception-71) [8]	79.6	-	43.5M	6.3x	1445B	42x
DeepLabV3+ (ResNeXt-50) [53]	79.5	81.4	-	-	-	-
DeepLabV3 (ResNet-101) [6]	78.5	-	58.0M	8.4x	1779B	52x
ESeg-M	**81.6**	**83.7**	**20.0M**	**1x**	**112B**	**1x**
HRNetV2-W48 [35]	81.1	-	65.9M	3.3x	747B	6.7x
OCR (HRNet-W48) [45]	81.1	-	-	-	-	-
ACNet (ResNet-101) [14]	80.9	-	-	-	-	-
Naive-Student [3]	80.7	83.4	147.3M	7.3x	3246B	29x
Panoptic-DeepLab (X-71) [10]	80.5	82.5	46.7M	2.3x	548B	4.9x
DeepLabV3 (ResNeSt-101) [48]	80.4†	-	-	-	-	-
Auto-DeepLab-L [24]	80.3	-	44.4M	2.2x	695B	6.2x
HRNetV2-W40 [35]	80.2	-	45.2M	2.3x	493B	4.1x
Auto-DeepLab-M [24]	80.0	-	21.6M	1.1x	461B	4.1x
DeepLabV3 (ResNeSt-50) [48]	79.9†	-	-	-	-	-
OCNet (ResNet-101) [46]	79.6	-	-	-	-	-
ESeg-L	**82.6**	**84.8**	**70.5M**	**1x**	**343B**	**1x**
SegFormer-B5 [40]	82.4	-	84.7M	1.2x	1460B	4.3x

Table 4. **Performance comparison on CityScapes.** † denotes results using multi-scale evaluation protocol. All our models are evaluated in single-scale evaluation protocol.

Model	mIoU	PixAcc
ESeg-M	**46.0**	**81.3**
OCR (ResNet-101) [45]	44.3/45.3†	-
HRNetV2-W48 [35]	43.1/44.2†	-
Auto-DeepLab-M [24]	42.2†	81.1†
PSPNet (ResNet-101) [50]	42.0†	80.6†
Auto-DeepLab-S [24]	40.7†	80.6†
ESeg-L	**48.2**	81.8
DeepLabV3 (ResNeSt-101) [48]	46.9†	82.1†
ACNet (ResNet-101) [14]	45.9†	82.0†
OCR (HRNet-W48) [45]	44.5/45.5†	-
OCNet (ResNet-101) [46]	45.5†	-
DeepLabV3 (ResNeSt-50) [48]	45.1†	81.2†
Auto-DeepLab-L [24]	44.0†	81.7†
SETR [51]	46.3	-
Swin-S [26]	49.3†	-
SegFormer-B4 [40]	50.3	-

Table 5. **Performance comparison on ADE20K.** † denotes results using multi-scale evaluation protocol. All our models are evaluated in single-scale evaluation protocol. Recent Transformer-based models are marked in gray.

Experiments

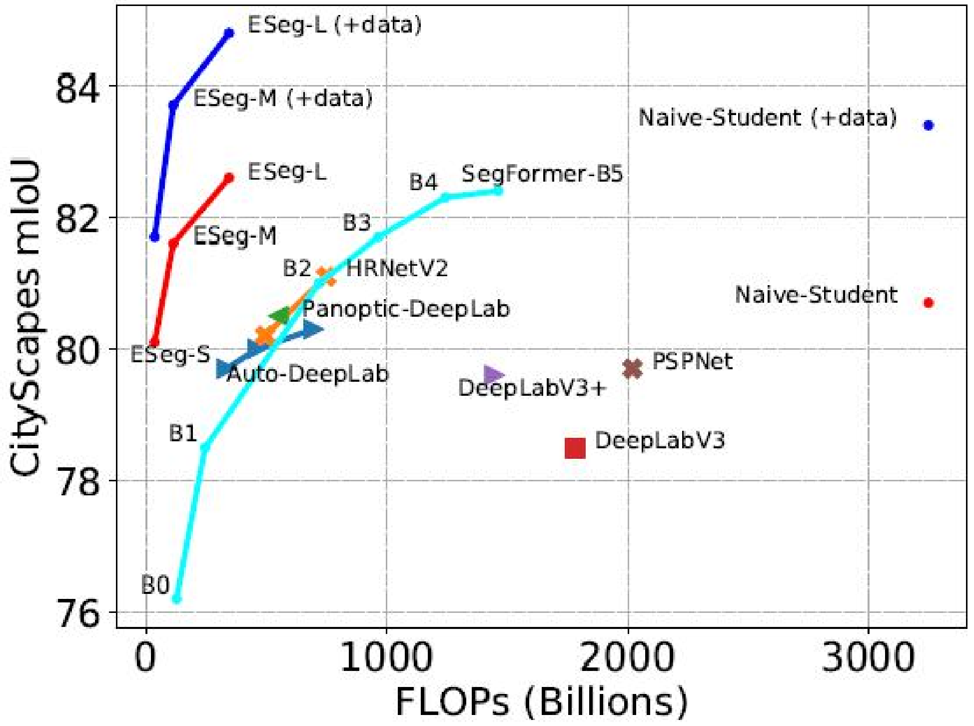


Figure 1. **Model Sizes vs. CityScapes validation mIoU**. All models in the figure are using single-scale evaluation protocol. `+data` denotes using extra data for pretraining and self-training. The FLOPs are calculated at 1024×2048 input resolution. Our proposed ESeg models are much simpler, yet still outperform previous models by better quality and less computation cost.

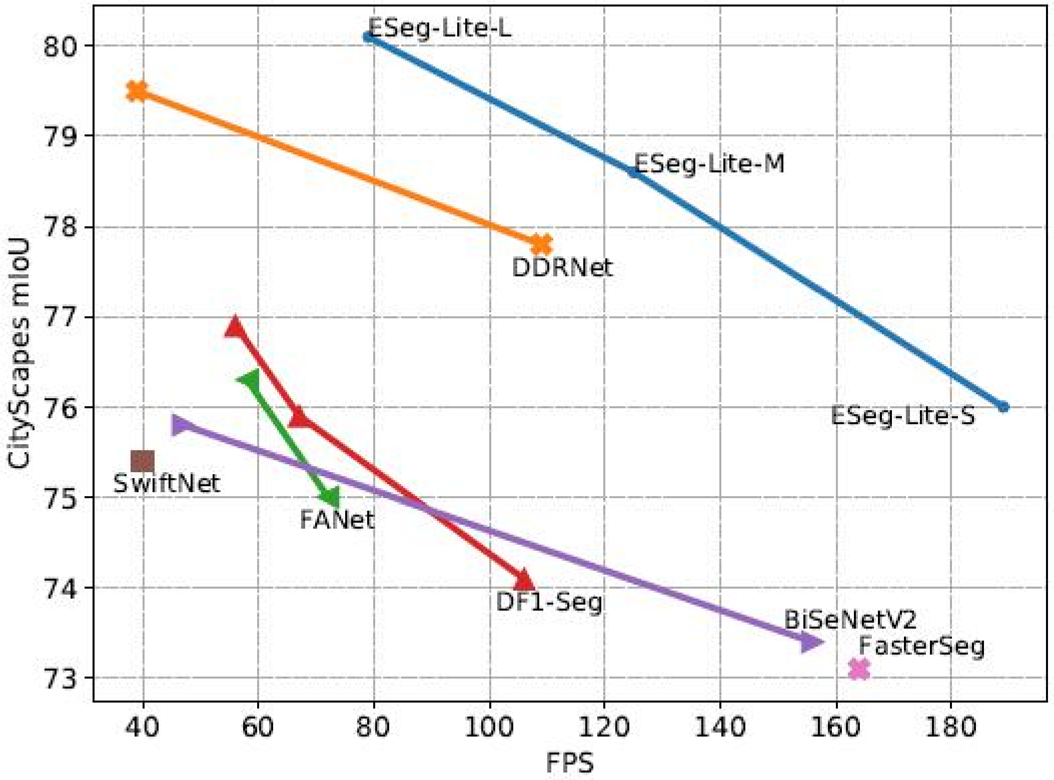


Figure 4. **Inference speed vs. CityScapes validation mIoU**. Real-time ESeg family of models outperform previous models by a large margin with much faster speed.

Experiments

- Ablation Study

Encoder	Decoder	mIoU	FLOPs
EfficientNet-B1	BiFPN (w/o atrous)	**80.1**	**34.5B**
	DeepLabV3+ (w/ atrous)	79.4	91.8B
	DeepLabV3+ (w/o atrous)	78.8	49.9B
ResNet-50	BiFPN (w/o atrous)	78.9	188.0B
	DeepLabV3+ (w/ atrous)	77.8	324.3B
	DeepLabV3+ (w/o atrous)	77.4	230.3B

Table 8. **Encoder and decoder choices.** All models are trained with exactly the same training settings. BiFPN outperforms DeepLabV3+ [8] regardless whether atrous convolutions are used.