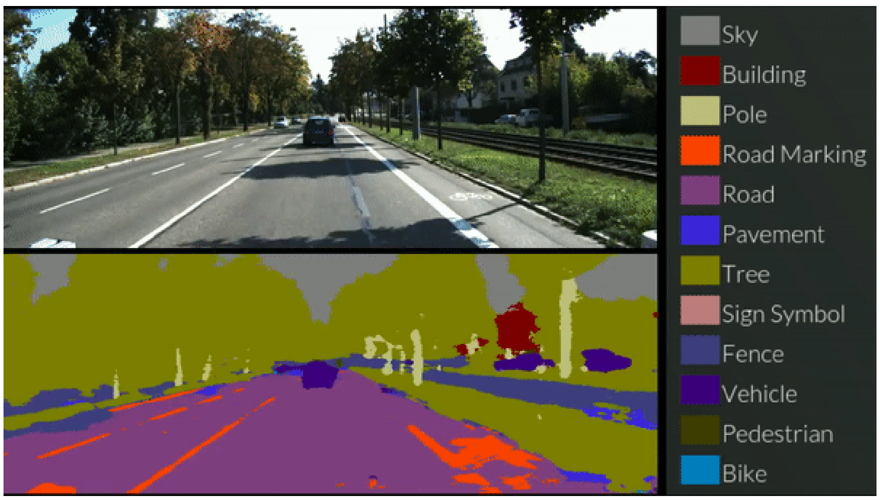
Review: SegNet (Semantic Segmentation)

Encoder Decoder Architecture, Using Max Pooling Indices to Upsample, Outperforms FCN, DeepLabv1, DeconvNet

Sik-Ho Tsang Follow
Feb 10 · 4 min read



SegNet by Authors (https://www.youtube.com/watch?v=CxanE_W46ts)

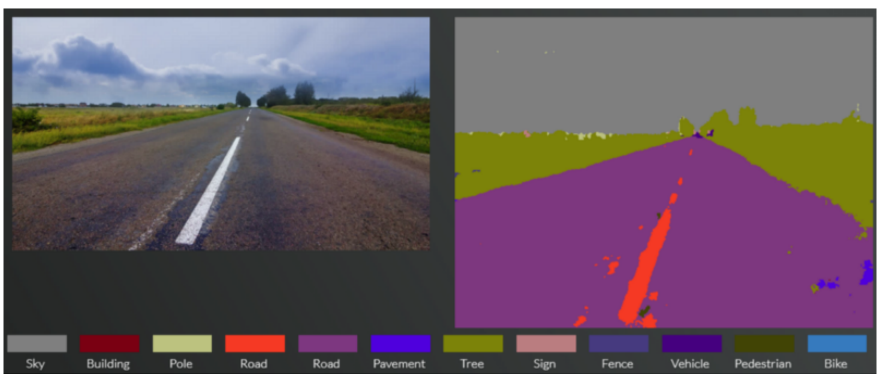
In this story, **SegNet**, by **University of Cambridge**, is briefly reviewed. Originally, it was submitted to 2015 CVPR, but at last it is not being published in CVPR (But it's **2015 arXiv** tech report version and still got over **100 citations**). Instead, it is published in **2017 TPAMI** with more than **1800 citations**. And right now the first author has become the Director of Deep Learning and AI in Magic Leap Inc. (Sik-Ho Tsang @ Medium)

Below is the demo from authors:

SegNet by Authors (https://www.youtube.com/watch?v=CxanE_W46ts)

There is also an interesting demo that we can choose a random image or even upload our own image to try the SegNet. I have tried as below:

- http://mi.eng.cam.ac.uk/projects/segnet/demo.php



The segmentation result for a road scene image that I found from internet

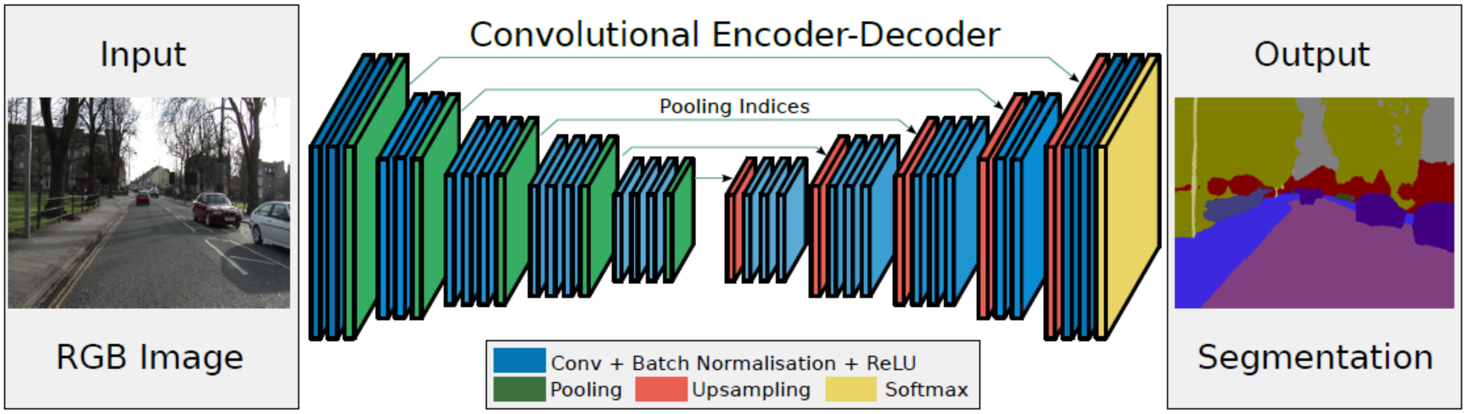
· · ·

Outline

1. **Encoder Decoder Architecture**
2. **Differences from DeconvNet and U-Net**
3. **Results**

· · ·

1. Encoder Decoder Architecture



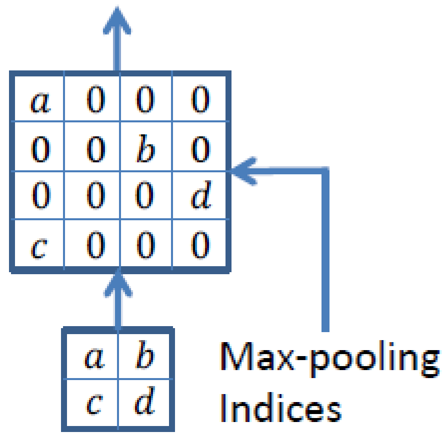
SegNet: Encoder Decoder Architecture

- SegNet has an **encoder** network and a corresponding **decoder** network, followed by a final pixelwise classification layer.

1.1. Encoder

- At the encoder, convolutions and max pooling are performed.
- There are 13 convolutional layers from VGG-16. (The original fully connected layers are discarded.)
- While doing 2×2 max pooling, the corresponding max pooling indices (locations) are stored.

1.2. Decoder



Upsampling Using Max-Pooling Indices

- At the decoder, upsampling and convolutions are performed. At the end, there is softmax classifier for each pixel.
- During upsampling, the max pooling indices at the corresponding encoder layer are recalled to upsample as shown above.
- Finally, a K-class softmax classifier is used to predict the class for each pixel.

. . .

2. Differences from DeconvNet and U-Net

DeconvNet and U-Net have similar structures as SegNet.

2.1. Differences from DeconvNet

- Similar upsampling approach called unpooling is used.
- However, there are fully-connected layers which make the model larger.

2.2. Differences from U-Net

- It is used for biomedical image segmentation.
- Instead of using pooling indices, the entire feature maps are transfer from encoder to decoder, then with concatenation to perform convolution.
- This makes the model larger and need more memory.

. . .

3. Results

- Two datasets are tried. One is CamVid dataset for Road Scene Segmentation. One is SUN RGB-D dataset for Indoor Scene Segmentation.

3.1. CamVid dataset for Road Scene Segmentation

Method	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Side-walk	Bicyclist	Class avg.	Global avg.	mIoU	BF
SfM+Appearance [28]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	53.0	69.1	n/a*	
Boosting [29]	61.9	67.3	91.1	71.1	58.5	92.9	49.5	37.6	25.8	77.8	24.7	59.8	76.4	n/a*	
Dense Depth Maps [32]	85.3	57.3	95.4	69.2	46.5	**98.5**	23.8	44.3	22.0	38.1	28.7	55.4	82.1	n/a*	
Structured Random Forests [31]	n/a											51.4	72.5	n/a*	
Neural Decision Forests [64]	n/a											56.1	82.1	n/a*	
Local Label Descriptors [65]	80.7	61.5	88.8	16.4	n/a	98.0	1.09	0.05	4.13	12.4	0.07	36.3	73.6	n/a*	
Super Parsing [33]	87.0	67.1	96.9	62.7	30.1	95.9	14.7	17.9	1.7	70.0	19.4	51.2	83.3	n/a*	
SegNet (3.5K dataset training - 140K)	**89.6**	**83.4**	96.1	**87.7**	52.7	96.4	**62.2**	**53.45**	**32.1**	**93.3**	**36.5**	**71.20**	**90.40**	60.10	46.84
CRF based approaches															
Boosting + pairwise CRF [29]	70.7	70.8	94.7	74.4	55.9	94.1	45.7	37.2	13.0	79.3	23.1	59.9	79.8	n/a*	
Boosting+Higher order [29]	84.5	72.6	**97.5**	72.7	34.1	95.3	34.2	45.7	8.1	77.6	28.5	59.2	83.8	n/a*	
Boosting+Detectors+CRF [30]	81.5	76.6	96.2	78.7	40.2	93.9	43.0	47.6	14.3	81.5	33.9	62.5	83.8	n/a*	

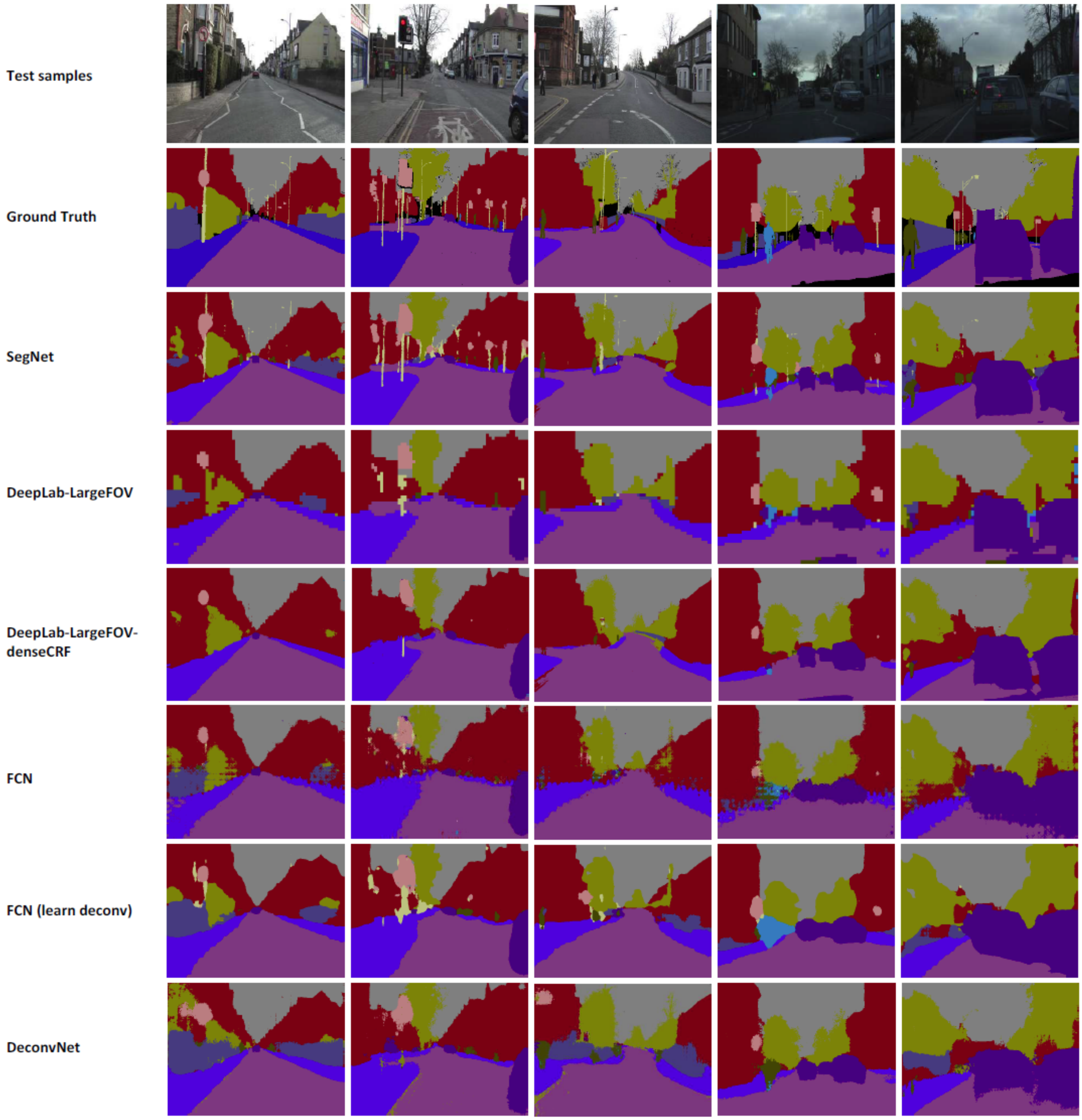
Compared With Conventional Approaches on CamVid dataset for Road Scene Segmentation

- As shown above, SegNet obtains very good results for many classes. It also got the highest class average and global average.

Network/Iterations	40K				80K				>80K				Max iter
	G	C	mIoU	BF	G	C	mIoU	BF	G	C	mIoU	BF	
SegNet	88.81	59.93	50.02	35.78	89.68	69.82	57.18	42.08	90.40	71.20	60.10	46.84	140K
DeepLab-LargeFOV [3]	85.95	60.41	50.18	26.25	87.76	62.57	53.34	32.04	88.20	62.53	53.88	32.77	140K
DeepLab-LargeFOV-denseCRF [3]	not computed								89.71	60.67	54.74	40.79	140K
FCN	81.97	54.38	46.59	22.86	82.71	56.22	47.95	24.76	83.27	59.56	49.83	27.99	200K
FCN (learnt deconv) [2]	83.21	56.05	48.68	27.40	83.71	59.64	50.80	31.01	83.14	64.21	51.96	33.18	160K
DeconvNet [4]	85.26	46.40	39.69	27.36	85.19	54.08	43.74	29.33	89.58	70.24	59.77	52.23	260K

Compared With Deep Learning Approaches on CamVid dataset for Road Scene Segmentation

- SegNet obtains highest global average accuracy (G), class average accuracy (C), mIOU and Boundary F1-measure (BF). It outperforms FCN, DeepLabv1 and DeconvNet.



Qualitative Results

3.2. SUN RGB-D Dataset for Indoor Scene Segmentation

- Only RGB is used, depth (D) information are not used.

Network/Iterations	80K				140K				>140K				Max iter
	G	C	mIoU	BF	G	C	mIoU	BF	G	C	mIoU	BF	
SegNet	70.73	30.82	22.52	9.16	71.66	37.60	27.46	11.33	72.63	44.76	31.84	12.66	240K
DeepLab-LargeFOV [3]	70.70	41.75	30.67	7.28	71.16	42.71	31.29	7.57	71.90	42.21	32.08	8.26	240K
DeepLab-LargeFOV-denseCRF [3]	not computed								66.96	33.06	24.13	9.41	240K
FCN (learnt deconv) [2]	67.31	34.32	24.05	7.88	68.04	37.2	26.33	9.0	68.18	38.41	27.39	9.68	200K
DeconvNet [4]	59.62	12.93	8.35	6.50	63.28	22.53	15.14	7.86	66.13	32.28	22.57	10.47	380K

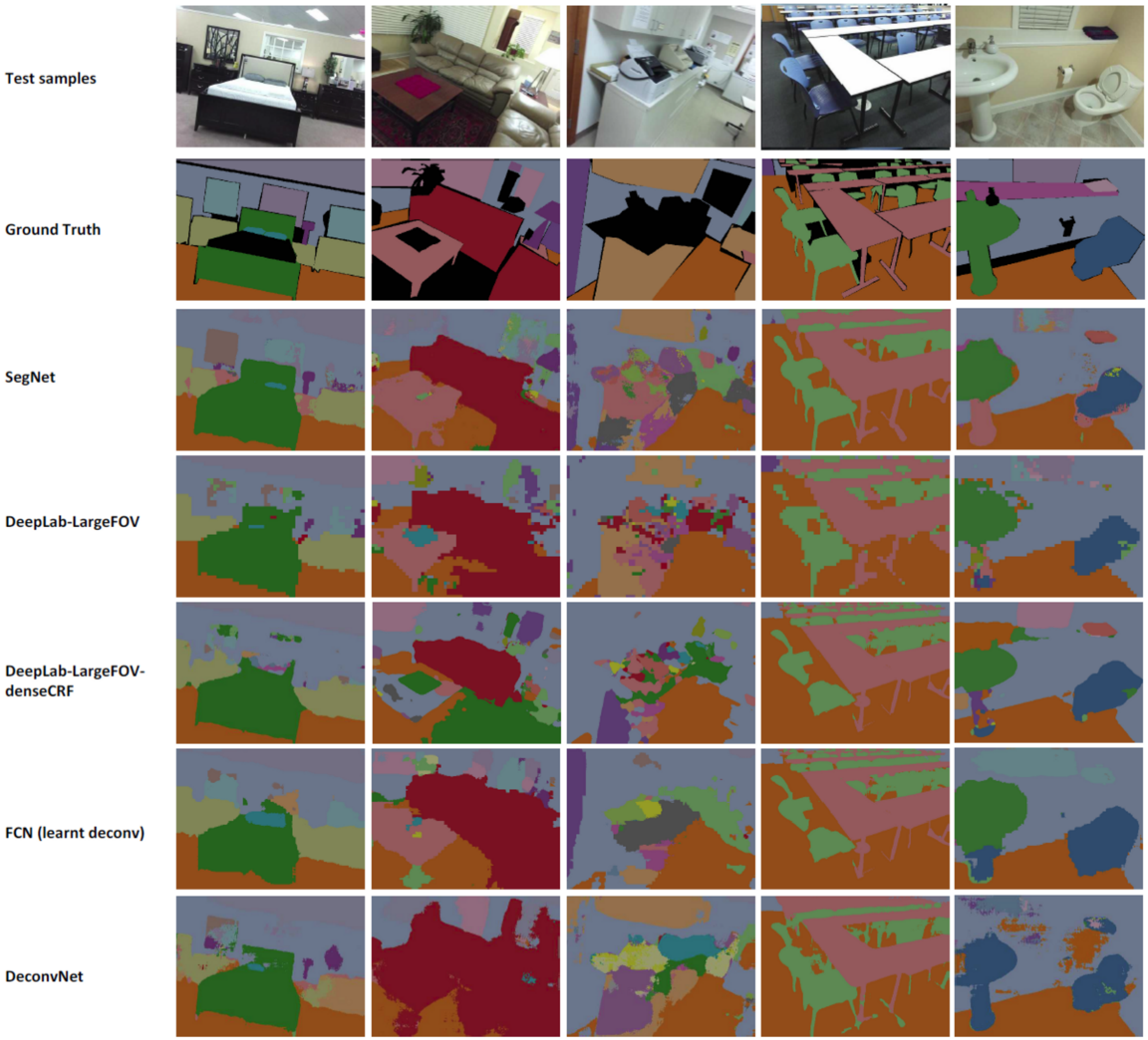
Compared With Deep Learning Approaches on SUN RGB-D Dataset for Indoor Scene Segmentation

- Again, SegNet outperforms FCN, DeconvNet, and DeepLabv1.
- SegNet only got a bit inferior to DeepLabv1 for mIOU.

Wall	Floor	Cabinet	Bed	Chair	Sofa	Table	Door	Window	Bookshelf	Picture	Counter	Blinds
83.42	93.43	63.37	73.18	75.92	59.57	64.18	52.50	57.51	42.05	56.17	37.66	40.29
Desk	Shelves	Curtain	Dresser	Pillow	Mirror	Floor mat	Clothes	Ceiling	Books	Fridge	TV	Paper
11.92	11.45	66.56	52.73	43.80	26.30	0.00	34.31	74.11	53.77	29.85	33.76	22.73
Towel	Shower curtain	Box	Whiteboard	Person	Night stand	Toilet	Sink	Lamp	Bathtub	Bag		
19.83	0.03	23.14	60.25	27.27	29.88	76.00	58.10	35.27	48.86	16.76		

Class Average Accuracy for Different Classes

- Higher accuracy for large-size classes.
- Lower accuracy for small-size classes.



Qualitative Results

3.3. Memory and Inference Time

Network	Forward pass(ms)	Backward pass(ms)	GPU training memory (MB)	GPU inference memory (MB)	Model size (MB)
SegNet	422.50	488.71	6803	**1052**	117
DeepLab-LargeFOV [3]	**110.06**	**160.73**	**5618**	1993	83
FCN (learnt deconv) [2]	317.09	484.11	9735	1806	539
DeconvNet [4]	474.65	602.15	9731	1872	877

Memory and Inference Time

- SegNet is slower than FCN and DeepLabv1 because SegNet contains the decoder architecture. And it is faster than DeconvNet because it does not have fully connected layers.
- And SegNet has low memory requirement during both training and testing. And the model size is much smaller than FCN and DeconvNet.

. . .

References

[2015 arXiv] [SegNet]
SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling

[2017 TPAMI] [SegNet]
SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

My Previous Reviews

Image Classification
[LeNet] [AlexNet] [ZFNet] [VGGNet] [SPPNet] [PReLU-Net] [STN] [DeepImage] [GoogLeNet / Inception-v1] [BN-Inception / Inception-v2] [Inception-v3] [Inception-v4] [Xception] [MobileNetV1] [ResNet] [Pre-Activation ResNet] [RiR] [RoR] [Stochastic Depth] [WRN] [FractalNet] [Trimps-Soushen] [PolyNet] [ResNeXt] [DenseNet] [PyramidNet]

Object Detection
[OverFeat] [R-CNN] [Fast R-CNN] [Faster R-CNN] [DeepID-Net] [R-FCN] [ION] [MultiPathNet] [NoC] [G-RMI] [TDM] [SSD] [DSSD] [YOLOv1] [YOLOv2 / YOLO9000] [YOLOv3] [FPN] [RetinaNet] [DCN]

Semantic Segmentation
[FCN] [DeconvNet] [DeepLabv1 & DeepLabv2] [ParseNet] [DilatedNet] [PSPNet] [DeepLabv3]

Biomedical Image Segmentation
[CUMedVision1] [CUMedVision2 / DCAN] [U-Net] [CFS-FCN] [U-Net+ResNet]

Instance Segmentation
[DeepMask] [SharpMask] [MultiPathNet] [MNC] [InstanceFCN] [FCIS]

Super Resolution
[SRCNN] [FSRCNN] [VDSR] [ESPCN] [RED-Net] [DRCN] [DRRN] [LapSRN & MS-LapSRN]