THE APPLICATION OF CONVOLUTIONAL NEURAL NETWORKS TO SEMANTIC SEGMENTATION

Arda Akça Büyük* **ardaakcabuyuk@gmail.com**
Computer Science/Faculty of Engineering

Ceyda Ömür** **ceydaomur@sabanciuniv.edu**
Computer Science/Faculty of Engineering and Natural Sciences

Gökberk Yar** **gokberkyar@sabanciuniv.edu**
Computer Science/Faculty of Engineering and Natural Sciences

Hasan Ocak** **ocakhasan@sabanciuniv.edu**
Computer Science/Faculty of Engineering and Natural Sciences

Işıl Dereli** **isildereli@sabanciuniv.edu**
Computer Science/Faculty of Engineering and Natural Sciences

Oğuz Çelik** **oguzcelik@sabanciuniv.edu**
Computer Science/Faculty of Engineering and Natural Sciences

Sena Korkut* **sena123korkut123@gmail.com**
Computer Science/Faculty of Engineering

Advisor
Assoc. Prof. Dr. Mehmet Keskinöz **keskinoz@sabanciuniv.edu**
Electronics Engineering

** Bilkent University*
*** Sabancı University*

Abstract

An image is a set of different pixels and each pixel has many different characteristics such as color, intensity and texture. Image segmentation is a process of partitioning a digital image into multiple segments that share similar attributes. It is typically used to locate objects and boundaries in images. Pixels that are nearby to each other and share the same color or pattern or gentle gradient of brightness are grouped into a single object. In that way we create a pixel-wise mask for each object in the image to identify the shape and boundary of each object. In our project, the aim is to perceive the impact of training datasets in human segmentation and compare the accuracy of existing models trained with appropriate datasets.

Keywords: Benchmark**,** Segmentation, CNN, Image Data, PSPNet, FCN

1. Introduction

Image segmentation is a process of taking an image as input and outputting multiple regions which have similar pixel-wise patterns. An over-simplified example could be labeling human like pixels as black and rest as white. Of course, today's standards are beyond this binary segmentation. To be exact 59 classes exists in State-of-Art Fully Connected Convolutional Network (FCN) [1] architecture. Application of image segmentation varies from self-driving cars to robots to medical imaging since it provides some degree of human level of visual awareness to the machines. As a benchmark project, different datasets (VOC12, Cityscapes, ADE20K) [2][3][4] were used in order to compare how well same Convolutional Neural Networks (CNN) architecture does when dataset is changed. In addition to the dataset change, different measures of success (pixel-wise accuracy, mean accuracy, mean IoU score and weighted IoU) are used to determine the model's success. Different CNN architectures of FCN and Pyramid Scene Parsing Network (PSPNET) [5] are taken under consideration and being tested under varies dataset and success measurements. For a curious reader, U-Net [6] is not examined in this benchmark because it focuses on medical images and the datasets relevant to study do not contain medical images. This benchmark's focus is to address suitable CNN architecture or architectures due to similarity of the input to the datasets that are tested in this benchmark. While focusing on this goal, benchmark also tries to address suitable CNN depths for tested architectures and accuracy trade-off due to the less depth. Different architectures beyond CNN or different dataset performances are not examined in this study so pre-deep learning techniques, feature extraction or selection machine learning techniques are not part of the study. In the study, due to the lack of computational power, official publicly available pre-trained weights are used for each architecture which is pre-trained for that specific dataset, ie FCN50 on VOC12.

2. Benchmark

2.1 Datasets

2.1.1 VOC2012

VOC2012 is the 8th version of the dataset for visual object recognition and segmentation challenge. It has been created by Mark Everingham and John Winn in 2005 and updated annually to 2012. There are 20 different classes (person, car, boat…) in the dataset. Half of the VOC2012 dataset consists train and validation set, and the other half consists test set. The images in training data have an annotation file with a bounding box and a class label [7]. Also, most of the images are annotated with pixel-wise segmentation of each object for the segmentation task [7].

There are 2913 images and 6929 objects in train and validation sets. As it can be seen on Table 1, there are 887 human images and 1733 objects. Each image in the segmentation subset has class segmentation and object segmentation for accurate segmentation. In class segmentation, each pixel in an image is labeled as a class or background to mask different classes from each other. In object segmentation part, pixels are labeled with an object number such as first, second, third object etc. which can be used to obtain a class or background [2]. Object segmentation is used to separate instances of the same class. Ground truth segmentations are very accurate but there might be some wrongly labeled pixels. There are bordering regions with a width of five pixels which can have a background or object. These regions are marked with a 'void' label which shows that corresponding pixels might be any class.

Table 2: Statistics of the segmentation image sets.

	train		val		trainval		test	
	img	**obj**	**img**	**obj**	**img**	**obj**	**img**	**obj**
Aeroplane	88	108	90	110	178	218	–	–
Bicycle	65	94	79	103	144	197	–	–
Bird	105	137	103	140	208	277	–	–
Boat	78	124	72	108	150	232	–	–
Bottle	87	195	96	162	183	357	–	–
Bus	78	121	74	116	152	237	–	–
Car	128	209	127	249	255	458	–	–
Cat	131	154	119	132	250	286	–	–
Chair	148	303	123	245	271	548	–	–
Cow	64	152	71	132	135	284	–	–
Diningtable	82	86	75	82	157	168	–	–
Dog	121	149	128	150	249	299	–	–
Horse	68	100	79	104	147	204	–	–
Motorbike	81	101	76	103	157	204	–	–
Person	442	868	445	865	887	1733	–	–
Pottedplant	82	151	85	171	167	322	–	–
Sheep	63	155	57	153	120	308	–	–
Sofa	93	103	90	106	183	209	–	–
Train	83	96	84	93	167	189	–	–
Tvmonitor	84	101	74	98	158	199	–	–
Total	1464	3507	1449	3422	2913	6929	–	–

Table 1. Statistics of the segmentation image sets [2].

In Figure 1**, a** is the original training image, in **b** image is segmented with many different class labels such as human, horse, background. 'Void' (cream colored) label is also used to show border regions and label difficult objects. In the image **c**, each object instance is separately segmented [2].

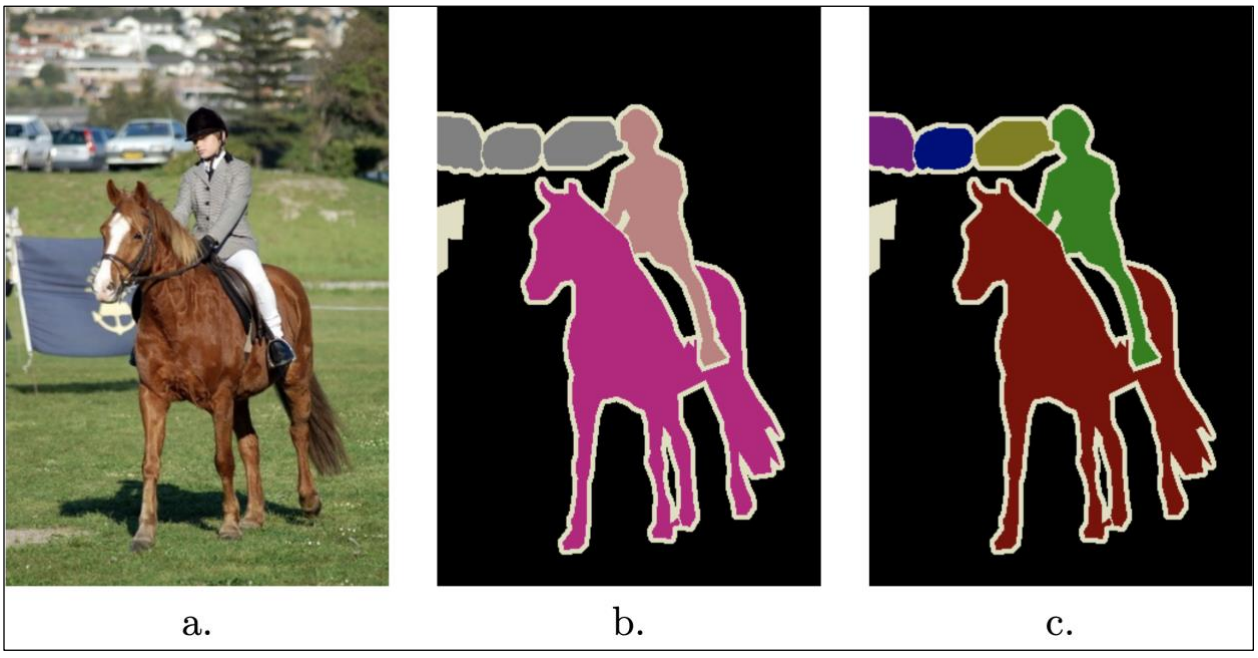

Figure 1. Example of segmentation ground truth [2].

2.1.2 Cityscapes

Cityscapes dataset is a benchmark suite and evaluation server to train pixel-level and instance-level semantic labeling. It has been constructed by M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele in 2016. It contains semantic images of urban streets and 30 classes such as person, road, car, truck, etc. It includes images of various cities, different weather conditions, several months, different backgrounds and frames. The dataset has 3 different types of annotations which are semantic, instance-wise and dense pixel annotations. This dataset has "5000 annotated images with fine annotations" and "20000 annotated images with coarse annotations" [3]. Fine annotations are high quality dense pixel-wise annotations and on the other hand, coarse annotations are coarser polygonal annotations.

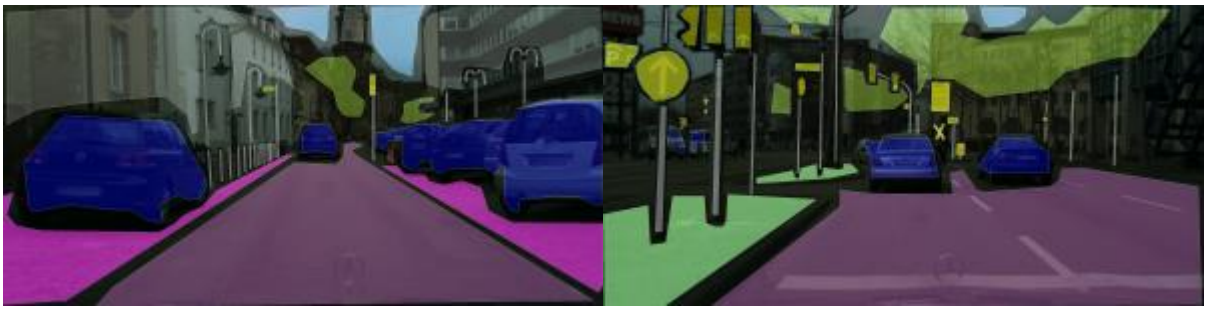

Figure 2 & 3. Example of coarse annotations which overlayed colors encode semantic classes [3].

Figure 4 & 5. Example of fine annotations which overlayed colors encode semantic classes [3].

Cityscapes dataset's size, scene variability, complexity and annotation richness make it a source beyond previous works for semantic segmentation [3].

	#humans [10^3]	#vehicles [10^3]	#h/image	#v/image
Ours (fine)	24.4	**41.0**	**7.0**	**11.8**
KITTI	6.1	30.3	0.8	4.1
Caltech	**192**[1]	-	1.5	-

Table 2. Average and absolute number of examples for Cityscapes, KITTI, and Caltech [3].

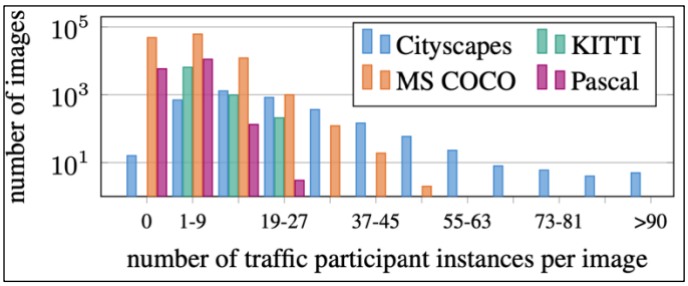

Table 3. Scene complexity statistics [3].

2.1.3 ADE20K

Fully Annotated Image Dataset (ADE20K) is a dataset which is constructed by Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso and Antonio Torralba in a scene parsing article. While constructing ADE20K dataset, their aim was to have a large set with different scene categories with intense annotations for all the visual concepts. ADE20K contains 20.210 in its training set, 2.000 images in its validation test and lastly 3.000 images in the test set. In the dataset, there are 3.169 class labels annotated and 2.693 of them are object classes while 476 of them are the classes of parts of the objects. Images are annotated in a detailed manner with objects and many of the objects are also annotated with their own components. Although many objects are annotated with their parts particularly in its validation set, images in the training set are not annotated that exhaustively [4].

Figure 6. Images in the dataset annotated in detail [4].

As demonstrated in Figure 6, the first row shows the original image while the second and third row show the segmented versions and the annotated parts of the segmented objects respectively. Additionally, if the color difference for diverse objects are considered, there is a large color gap between different categories of objects whereas same or close object categories have a small color gap between them.

Figure 7. The annotation and labeled objects done by the expert [4].

Images of ADE20K dataset are collected from LabelMe, Places and SUN datasets so that 900 different categories defined in SUN database can be included. The annotation was made by just one expert with LabelMe interface (see Figure 7).

2.2 CNN Architectures

2.2.1 Pyramid Scene Parsing Network (PSPNet)

Pyramid scene parsing network is a model, which has been developed by Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia in 2017 and aim of the model to assign each pixel in the image a category label and understand the scene completely with the help of scene parsing [5]. State-of-the-art scene parsing methods can predict the label, shape, location of each object according to the context of the scene. For instance, even though a house and a boathouse are shapely similar, if there is a river in the scene, the model would predict the object as a boathouse. The difference between PSPNet and other scene-of-the-art methods is that creator of PSPNet extended the pixel-level feature to the "specially designed global pyramid pooling one" and offered an optimization method with deeply supervised loss.

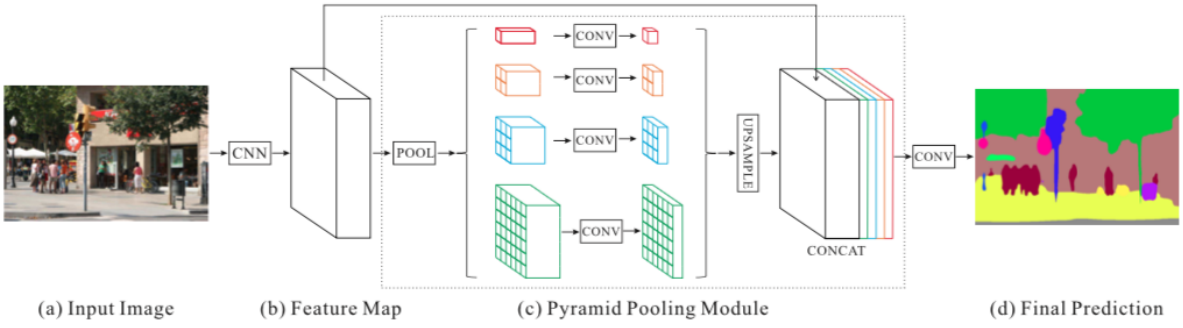

Figure 8. Overview of PSPNet [5].

How PSPNet works is as follows: given an input image (a), CNN is used to get the feature map of the last convolutional layer (b). After that, Pyramid Pooling Module is applied to get different sub-region representations, followed by upsampling and concatenation layers to form the final feature representation, which carries both local and global context information in (c). At the end, the representation is fed into a convolution layer to get the final per-pixel prediction (d) [5].

Zhao et al. developed deeper neural networks, which are beneficial to large scale data classification, to further analyze PSPNet. There are four depths of 50, 101, 152 and 269. In our research we choose PSPNet 101 and PSPNet 50 to analyze the accuracy rate.

Method	Mean IoU(%)	Pixel Acc.(%)
PSPNet(50)	41.68	80.04
PSPNet(101)	41.96	80.64
PSPNet(152)	42.62	80.80
PSPNet(269)	**43.81**	**80.88**

Table 4. Deeper pre-trained model gets higher accuracy rate (pixel-wise accuracy (Pixel Acc.) and mean of class-wise intersection over union are used (Mean IoU)) [5].

PSPNet came in first place in ImageNet Scene Parsing Challenge 2016 (see Table 2) and the ADE20K dataset is used in the challenge. For evaluation pixel-wise accuracy (Pixel Acc.) and mean of class-wise intersection over union are used (Mean IoU).

Rank	Team Name	Final Score (%)
1	**Ours**	**57.21**
2	Adelaide	56.74
3	360+MCG-ICT-CAS_SP	55.56
-	(our single model)	(55.38)
4	SegModel	54.65
5	CASIA_IVA	54.33
-	DilatedNet [40]	45.67
-	FCN [26]	44.80
-	SegNet [2]	40.79

Table 5. Results of ImageNet scene parsing challenge 2016. The label "Ours" is PSPNet 50. The final score is the mean of Mean IoU and Pixel Acc. [5].

There are more analyzes with other datasets such as VOC2012 and Cityscapes. PSPNet 101, which has been trained on Cityscapes, achieved 80.2% accuracy and PSPNet 101, which has been trained on VOC2012 achieved 82.6% accuracy. The visual differences between models trained on different datasets can be seen in Figure 9, Figure 10 and Figure 11.

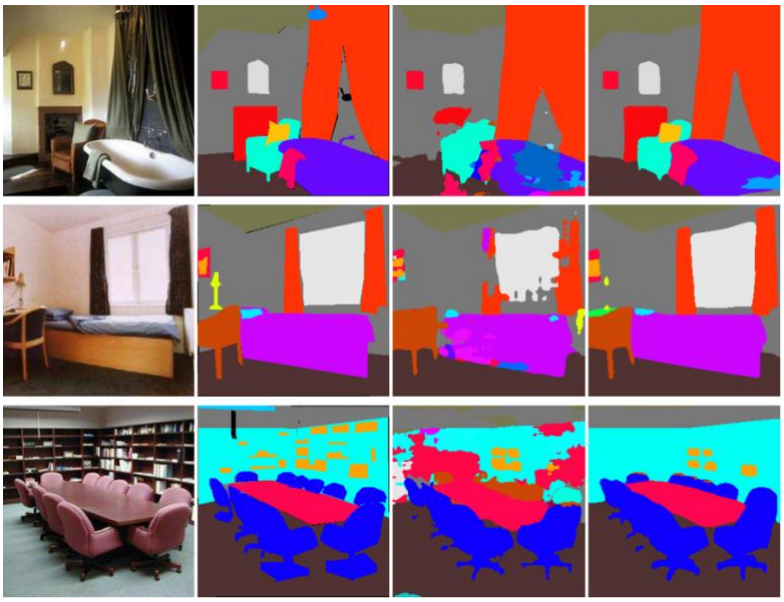

Figure 9. PSPNet 50, trained on ADE20K [5].

Figure 10. PSPNe101t, trained on VOC2012 [5].

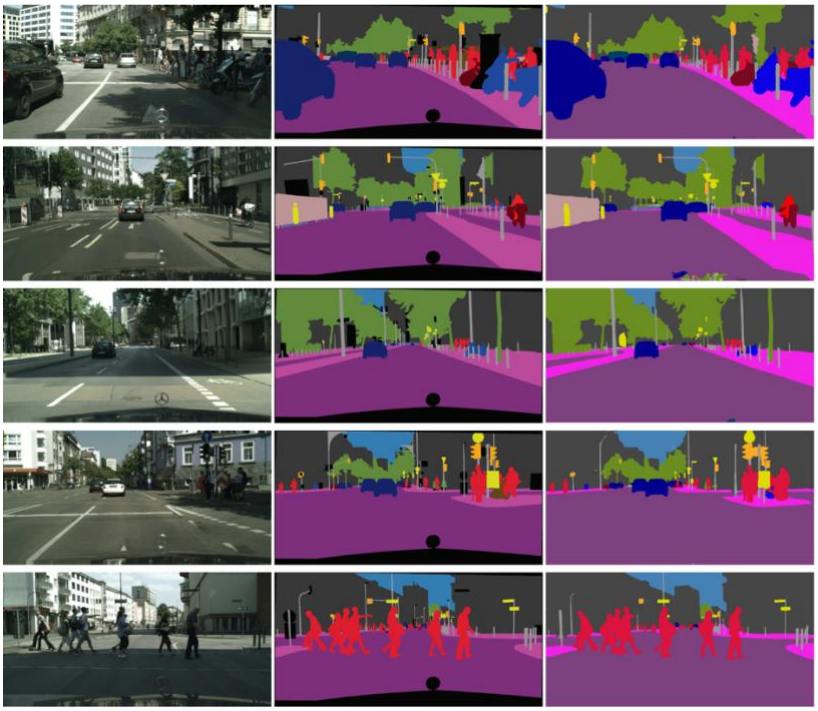

Figure 11. PSPNet 101, trained on Cityscapes [5].

2.2.2 Fully Convolutional Networks for Semantic Segmentation (FCN)

Convolutional networks in image classification, output one label for the image. FCN is a network which uses the same network for making prediction at every pixel.

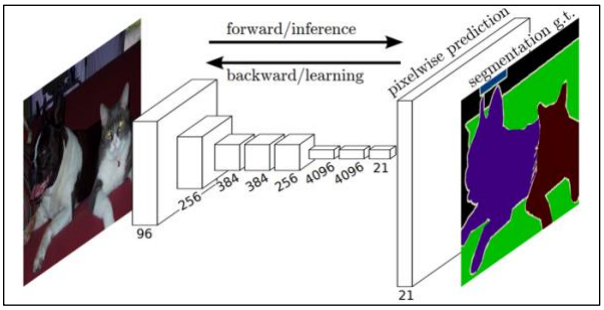

Figure 12. Fully Convolutional Networks can be used efficiently for pixel-size image segmentation [1].

In this network, unlike the classification task, 1*1 convolutions are used instead of fully connected layers to get an image whose size is smaller than the input size and which will be used at the upsample part. This process is shown in Figure 13.

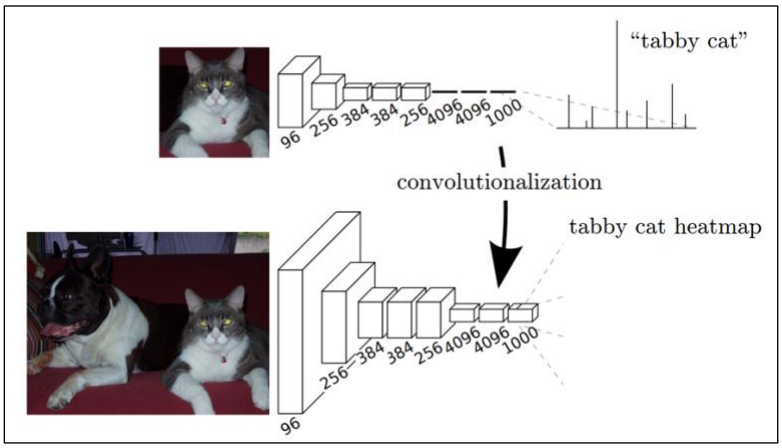

Figure 13. Classification to pixel-size prediction [1].

In the upsampling part of the network, the pooled layer of the image is upsampled to get the same size image as input. After the 5th pooling layer in the convolution part, the upsampled prediction is called FCN-32s (see Figure 14). But as the image going smaller, the location information is also lost. To avoid that, the output from some certain pooling layers (pool3, pool4 and pool5) are used and combine with each other. FCN16-s consists of fusing of 2*2 upsampling of pool5 and pool4. FCN-8s consists of fusing of 2*2 upsampling of pool4 and pool3. In this way, the loss of information is becoming smaller.

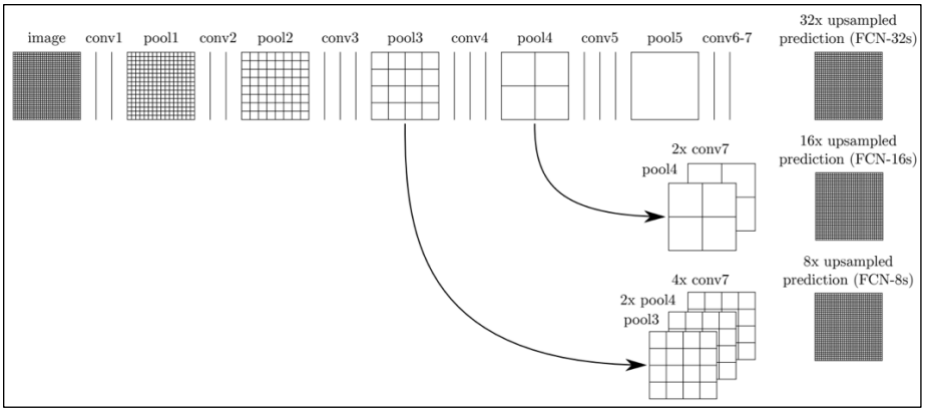

Figure 14. Fusing the layers with each other to get more accurate segmentation [1].

FCN-8s results better in the same image among the others because of less location information. The results are shown in Figure 15.

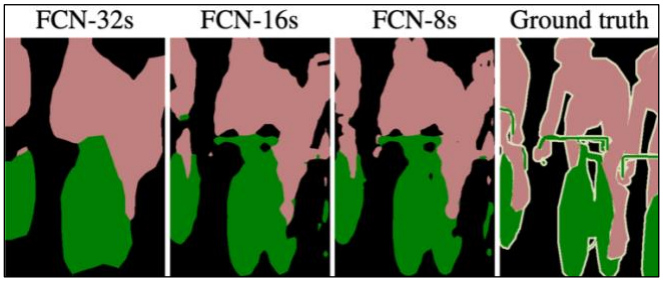

Figure 15. FCN-8s has the best result [1].

FCN networks are also tested in PASCAL VOC 2011 and 2012, ADE20K (MIT SceneParse) and the Cityscapes dataset. The results are in Table 1.

	mean IU VOC2011 test	mean IU VOC2012 test	inference time
R-CNN [12]	47.9	-	-
SDS [17]	52.6	51.6	~ 50 s
FCN-8s	**62.7**	**62.2**	~ **175 ms**

Table 6. Scores of FCN-8s in Pascal VOC [1].

	train	val	coarse	sub	Classes IoU	Classes iIoU	Categories IoU	Categories iIoU
FCN-32s	✓	✓			61.3	38.2	82.2	65.4
FCN-16s	✓	✓			64.3	41.1	84.5	69.2
FCN-8s	✓	✓			65.3	41.7	85.7	70.1
FCN-8s	✓	✓		2	61.9	33.6	81.6	60.9
FCN-8s		✓			58.3	37.4	83.4	67.2
FCN-8s			✓		58.0	31.8	78.2	58.4

Table 7. Score of FCNs in Cityscapes dataset [3].

Networks	Pixel Acc.	Mean Acc.	Mean IoU	Weighted IoU
FCN-8s	71.32%	40.32%	0.2939	0.5733
SegNet	71.00%	31.14%	0.2164	0.5384
DilatedVGG	73.55%	44.59%	0.3231	0.6014
DilatedResNet-34	76.47%	45.84%	0.3277	0.6068
DilatedResNet-50	76.40%	45.93%	0.3385	0.6100
Cascade-SegNet	71.83%	37.90%	0.2751	0.5805
Cascade-DilatedVGG	74.52%	45.38%	0.3490	0.6108

Table 8. Score of ADE20K dataset on various models [5].

In conclusion, FCN is a useful network for the image segmentation task because it makes the loss smaller with the fusing part.

3. Research Methodology

Our three objectives were to compare the successes of

- the same models trained with different datasets
- the same architecture with different depths
- different models trained on the same dataset

on particularly human segmentation. In order to achieve these objectives, we downloaded models that are implemented with Keras. Since training these models is a tremendously costly operation on both time and computational power, we used pretrained models to test them with our human picture data. The models that we tested with human pictures were

- PSPNet50 (Pyramid Scene Parsing Network), trained with ADE20k dataset
- PSPNet101, trained with VOC2012 dataset
- PSPNet101, trained with Cityscapes dataset
- FCN (Fully Convolutional Network), trained with VOC2012 dataset

First, we coded a little Python script to make our models runnable, therefore testable (see Figure 16).

```python
import keras_segmentation

# load the pretrained model trained on ADE20k dataset
# model = keras_segmentation.pretrained.pspnet_50_ADE_20K()

# load the pretrained model trained on Cityscapes dataset
# model = keras_segmentation.pretrained.pspnet_101_cityscapes()

# load the pretrained model trained on Pascal VOC 2012 dataset
model = keras_segmentation.pretrained.pspnet_101_voc12()

out = model.predict_segmentation(
    inp="roger.jpg",
    out_fname="roger-PSPNet101-VOC2012-out.png"
)
```

Figure 16. Python script that segments an image with the selected model

Then, we prepared a little test data of 4 human images by downloading them online. And we tried to achieve our objectives in 3 steps.

1. After inputting our test data into two PSPNet101s (one trained with VOC2012 dataset, the other trained with Cityscapes dataset), we ended up with two segmented images that make us able to simply evaluate and compare the results for our first objective.

Figure 17. Outputs of PSPNet101s trained with Cityscapes and VOC2012 datasets.

2. For our second objective, we inputted our test data into PSPNet50 and PSPNet101. Since these models are structurally the same but their numbers of layers differ, the difference between their outputs gave us intuitive deductions about how the number of layers influence the performance of the model.

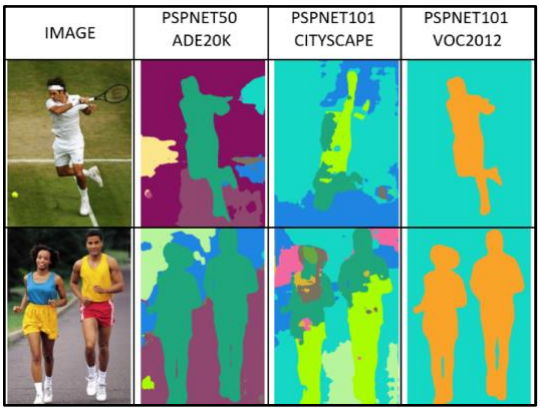

Figure 18. Outputs of PSPNet50 and PSPNet101s

3. Lastly, to be able to compare FCN and PSPNet101 trained on the same dataset, we took a glance at FCN paper and realized that there are samples which are the outputs of the model trained with VOC2012. So, we inputted the same image from the paper into our PSPNet101 model in order to infer how different models succeed regardless of the dataset that they are trained with.

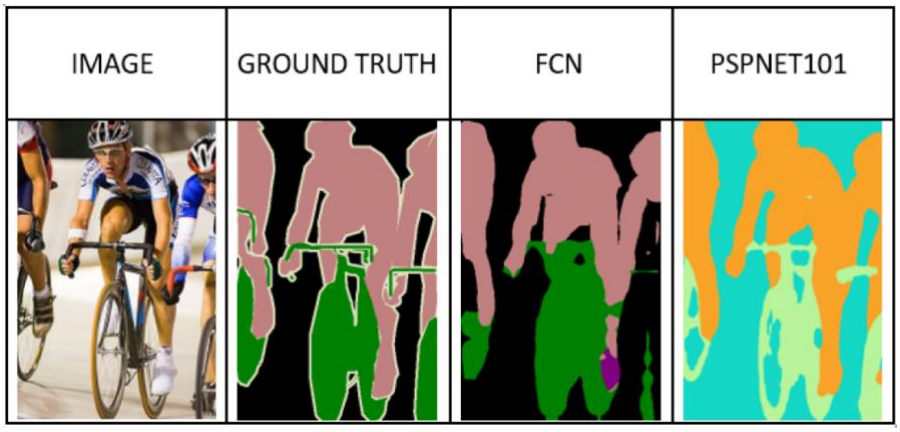

Figure 19. Outputs of FCN and PSPNet trained on both VOC2012 dataset [1].

All of our coding work is done on PyCharm CE. The models are downloaded from GitHub [8]. Our image data is in the format JPEG, outputs are in the format PNG.

4. Discussion and Conclusion

- PSPNet 101 has a better performance than FCN when they are both trained with the same dataset (VOC2012). PSPNet 101 resulted as a more accurate segmentation to the ground truth.

- PSPNet 101 has a higher accuracy when it is trained with VOC2012 dataset instead of Cityscape dataset.

- PSPNet 101 model which is trained by VOC2012 dataset has a higher accuracy than PSPNet 50 model trained with ADE20k dataset.

- PSPNet 50 model trained with ADE20k dataset has a higher accuracy than PSPNet 101 trained with Cityscape dataset.

- PSPNet 101 has a better accuracy than PSPNet 50 because deeper pre-trained model gets a higher performance [2].

References

1. J. Long, E. Shealhamer and T. Darrell "Fully Convolutional Networks for Semantic Segmentation, Mar 2015. Accessed on July. 3. 2019. [Online] Available: https://arxiv.org/pdf/1411.4038.pdf

2. M. Everingham and J. Winn, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Development Kit," *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Development Kit*, 18-May-2012. [Online]. Available: https://pjreddie.com/media/files/VOC2012_doc.pdf.

3. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

4. B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic Understanding of Scenes Through the ADE20K Dataset," International Journal of Computer Vision, vol. 127, no. 3, pp. 302–321, 2018.

5. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

6. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, 2015.

7. M. Everingham, L. Van Gool, Chris Williams, John Winn, and Andrew Zisserman, "'The {PASCAL} {V}isual {O}bject {C}lasses {C}hallenge 2012 {(VOC2012)} {R}esults,'" *Visual Object Classes Challenge 2012 (VOC2012)*, 18-May-2012. [Online]. Available: http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html.

8. Divamgupta, "divamgupta/image-segmentation-keras," *GitHub*, 10-Jun-2019. [Online]. Available: https://github.com/divamgupta/image-segmentation-keras. [Accessed: 05-Aug-2019].