Image Segmentation Using FCN and DeepLab

Bo Peng[1], **Yiwu Zhong**[1]
{bpeng28, yzhong52}@wisc.edu
CS 760 Machine Learning Course Project

Abstract

Semantic Segmentation is an important task in the field of computer vision, whose goal is to label each pixel according to which object or class it belongs to. In this project, we studied whether sophisticated learning methods provide better predictive accuracy than simple ones and how predictive performance varies when changing the training set size. Experiment results showed that better predictive accuracy could be acquired as we increase the size of training dataset. Also a more sophisticated model outperforms the simple one.

Introduction

Deep Convolutional neural networks (DCNNs) (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014) have achieved significant breakthrough on image classification. They are powerful visual models to yield hierarchies of features which will be used to implement classification tasks.

Usually, DCNNs have four main operations: Convolution, activation function, pooling, and fully-connected layers. Passing an image through a series of these operations and outputs a feature vector containing the probabilities for each class label. Note that in this setup, DCNNs categorize an image as a whole; that is, assign a single label to an entire image. Regarding image semantic segmentation tasks, the model needs to determine the class of a single pixel, which means understanding images at a pixel level, also known as dense prediction.

However, regular DCNNs such as the AlexNet (Krizhevsky, Sutskever, and Hinton 2012) and VGG (Simonyan and Zisserman 2014) are not suitable for dense prediction tasks. There are two inherent drawbacks of DCNNs. First, these models reduce feature resolution, caused by consecutive pooling operation and convolution striding, which allows DCNNs to learn increasingly abstract feature representations. As the consequence, these layers end up producing highly decimated feature vectors that lack rich and sharp spatial details. Second, fully-connected layers of DCNNs require resizing input images into fixed sizes and will change the spatial information, which is not suitable for dense prediction.

In order to implement dense prediction tasks and reduce the computing load as low as possible, researcher introduced segmentation network (Ronneberger, Fischer, and Brox 2015; Badrinarayanan, Kendall, and Cipolla 2015)with convolution, down-sampling and up-sampling layers, which is also called encoder and decoder networks.

Fully convolutional networks (FCN) (Long, Shelhamer, and Darrell 2014) has a great achievement in semantic segmentation, which takes input of arbitrary size and produces output image with same size as input, after the up-sampling processing and concatenating processing skip architecture.

However, consecutive pooling layers and convolution striding, also known as down-sampling operators, still exist in FCN. These down-sampling operators reduce spatial information and thus impede dense prediction tasks, where detailed spatial information is desired. To overcome this problem, DeepLab methods (Chen et al. 2014; 2016; 2017) remove the down-sampling operator from the last few max pooling layers of DCNNs, and instead advocate the use of atrous convolution. Atrous convolution is able to control the resolution at which feature responses are computed within DCNNS without requiring learning extra parameters. Also, Atrous Spatial Pyramid Pooling (ASPP) method was introduced in DeepLab, which corresponds to another fact of the existence of objects at multiple scales. And Conditional Random Field (CRF) has been broadly used in semantic segmentation to improve the localization of object boundaries as post-processing after neural networks.

In this project, we aimed at studying the underlying explanation for different predictive ability of FCN and DeepLab. More specifically, we will talk about how the predictive accuracy is influenced by the training-set size. Also we are trying to figure out whether a sophisticated image segmentation model (e.g., DeepLab) outperforms a relatively simple one (e.g., FCN).

The remaining part of this project report will at first go through the theory of FCN and DeepLab in the next section. Then, we will explain the experiments conducted in this project, followed by the section discussing the results. Finally, we made our conclusion regarding the predictive ability of FCN and DeepLab.

Methods

FCN

The key insight of Fully convolutional networks (FCN) (Long, Shelhamer, and Darrell 2014) is to replace fully connected networks with fully convolutional networks that take input of arbitrary size. Fully convolutional networks are followed by up-sampling layers which use deconvolution and bilinear interpolation to produce correspondingly-sized output.

But the semantic information from a single up-sampling layers is too coarse to predict pixels correctly. In order to improve efficient inference and learning, FCN introduced a skip architecture, as shown in Figure 1, which combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer. Fully convolutional networks, up-sampling layers and the skip architecture guarantee FCN to be applied to spatially dense prediction tasks and generate relatively fine prediction with more details instead of coarse prediction.

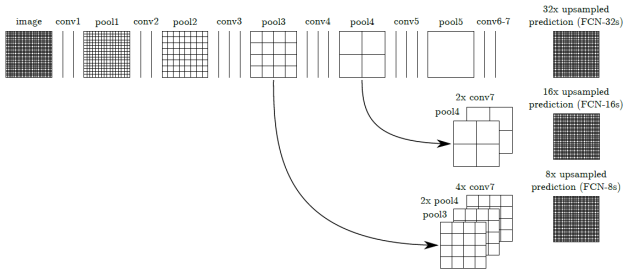


Figure 1: Skip Architecture of FCN

DeepLab

DeepLab methods advocate the use of atrous convolution to prevent reduced spatial resolution on feature maps. And Atrous Spatial Pyramid Pooling (ASPP) method was introduced to corresponds another fact of the existence of objects at multiple scales. Also, after the dense prediction operated by networks, Conditional Random Field (CRF) was used as post-processing to improve the localization of object boundaries.

Atrous Convolution Atrous convolution (Yu and Koltun 2015; Chen et al. 2016) proved to be a powerful tool in dense prediction tasks. It allows us to explicitly control the resolution at which feature responses are computed within Deep Convolutional Neural Networks. It also allows us to effectively enlarge the field of view of filters by enlarging the rate of inserting holes between filter weights, to incorporate larger context without increasing the number of parameters or the amount of computation. Figure 2 shows the atrous convolution process.

ASPP Atrous spatial pyramid pooling (ASPP) (Chen et al. 2017) robustly segments objects at multiple scales. ASPP uses multiple parallel atrous convolutional layers with different sampling rates to probe an incoming convolutional feature layer, thus capturing objects as well as image context at multiple scales

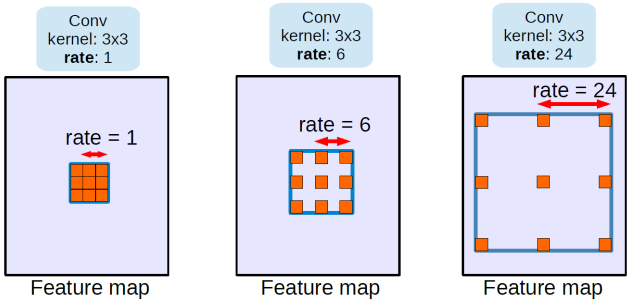


Figure 2: Atrous Convolution

CRF Deeplab improves the localization of object boundaries by combining methods from DCNNs and probabilistic graphical models. The commonly deployed combination of max-pooling and down-sampling in DCNNs achieves invariance but has a toll on localization accuracy. Deeplab overcomes this by combining the responses at the final DCNN layer with a fully connected CRF (Chen et al. 2014), which is shown both qualitatively and quantitatively to improve localization performance.

Experiments

In this project, we conducted image segmentation by training and testing the two state-of-art convolutional neural networks (i.e., FCN and DeepLab). All the experiments were conducted with tensorflow on a single NVIDIA Tesls K20m GPU.

Dataset

We used PASCAL VOC 2011 (Everingham et al. a) for training and validation, the size of training set and validation set are 1112 and 1111, respectively. To test the performance of both FCN and DeepLab, part of PASCAL VOC 2012 (Everingham et al. b) dataset are used for testing, which contain a total of 690 images that are not included in PASCAL VOC 2011.

Model Training and Testing

For comparative analysis, we took a pre-trained FCN model (Zou 2017) as the baseline of this project. This pre-trained model has been trained on the training set described in last section with 8 pixel stride nets (i.e., FCN-8s).

First of all, we directly outputted the prediction results of the testing set based on FCN-8s, and evaluate the predictive accuracy as a baseline. Secondly, we randomly select part of the training-set samples to build new training sets with various sizes, like 20% and 60%. Together with the original entire training set (i.e., size = 100%), we trained 3 DeepLab models (Silva 2018) on these 3 training sets, of which the size varies. It should be noted that the validation set remains the same for all models during the training process.

With 200 epochs training for DeepLabs, we obtained 3 DeepLab models with different predictive ability. To study

how the predictive accuracy vary with the training-set size, we tested the 3 models on the same testing set and evaluated the predictive accuracy respectively. Figure 3, 4, and 5 show the training process in terms of training loss versus the global training steps, where cross entropy is used to indicate the training loss. As shown by the training process, all models converges after a number of iterations.

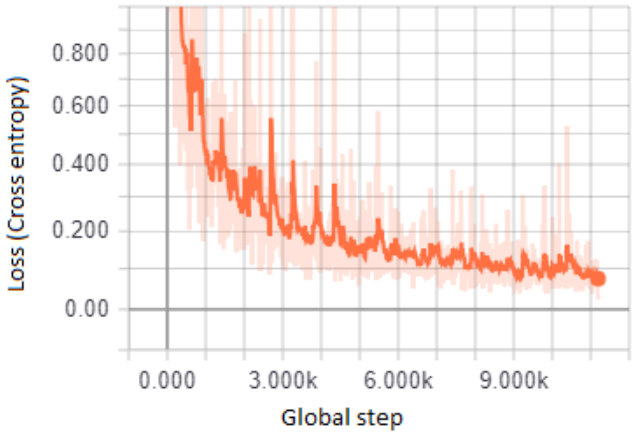


Figure 3: Training loss with training-set size = 20%

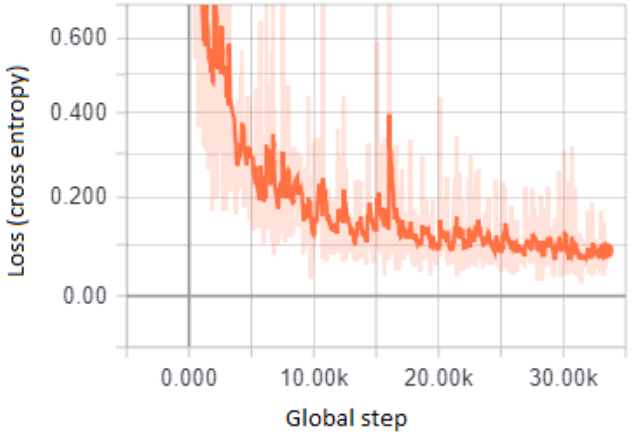


Figure 4: Training loss with training-set size = 60%

It is worth noting that the total global steps varies for 3 models even though we trained all models with the same 200 epochs. This is because each model was trained on training sets with different sizes.

Results and Discussion

Testing Performance

Image Segmentation Figure 6 presents the semantic segmentation results of one sample image from the testing set that was predicted by a pre-trained FCN-8s on the 100% training set, and DeepLab trained on 20%, 60%, and 100%, respectively. Through visual inspection, the predictive ability of DeepLab is going up as we increase the training-set

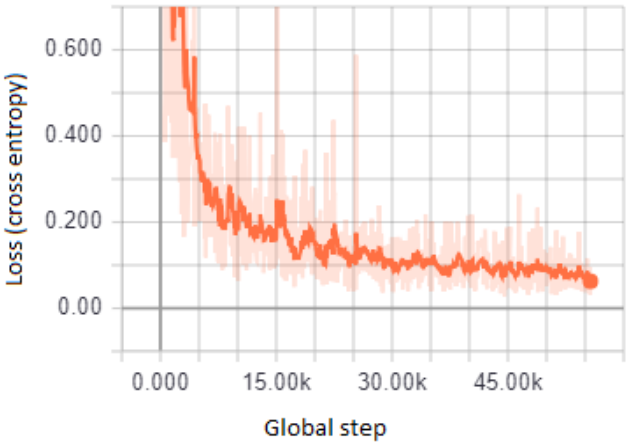
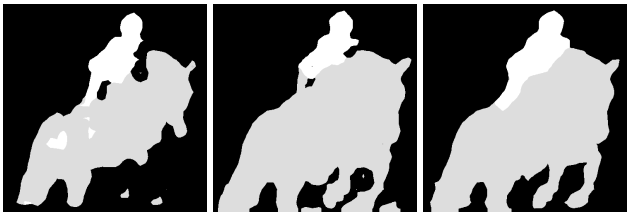


Figure 5: Training loss with training-set size = 100%

(a) Raw image (b) Ground truth (c) FCN-8s (100%)



(d) DeepLab (20%) (e) DeepLab (60%) (f) DeepLab (100%)

Figure 6: Image segmentation predictions

size. Additionally, all models failed to differentiate the reins from the person and the horse. In the predicted image produced by FCN-8s model, both the person and the horse pixels were not well determined since some part of both objects were mixed. Compared with FCN-8s, DeepLab trained on 100% training set is able to tell the person apart from the horse better than FCN-8s does. However, DeepLab trained on a small training set did not perform well compared with DeepLab with 100% training set, as part of the person and the horse pixels were determined as background.

Metrics For quantitative analysis of the predictive accuracy, we used pixel accuracy and the region intersection over union (IoU) to evaluate the performance of FCN and Deeplab on the testing set. Specifically, Let n_{ij} be the number of pixels of class i predicted as class j, where there are n_{cl} different classes, and let $t_i = \sum_j n_{ij}$ be the total number of pixels of class i. As a result,

- pixel accuracy $= \sum_i n_{ii} / \sum_i t_i$
- mean IoU $= (1/n_{cl}) \sum_i n_{ii} / (t_i + \sum_j n_{ji} - n_{ii})$

Predictive Accuracy Table 1 gives the performance of FCN and Deeplab in terms of mean IoU and pixel accuracy on testing set from VOC2012. As highlighted in the table, the DeepLab model trained on the entire training set (i.e., size = 100%) achieved the best performance on both mean IoU and pixel accuracy, compared with DeepLab model trained on partial training set (i.e., 20% and 60%) and FCN model on the entire training set. In addition, it should be noted that, even for DeepLab models trained on training-set of size 20% or 60%, DeepLab outperforms FCN in terms of the mean pixel accuracy and IoU.

Table 1: Predictive Accuracy on Testing Set

	Mean Pixel Accuracy	Mean IoU
FCN-8s (100%)	0.8115	0.5625
DeepLab (20%)	0.8542	0.5935
DeepLab (60%)	0.8808	0.6423
DeepLab (100%)	**0.8861**	**0.6579**

Comparative Analysis

Predictive Accuracy versus Training-set Size In this project, we studied how the predictive accuracy vary as a function of training-set size. Figure 7 shows the learning curve of DeepLab with x-axis denoting the training-set size and y-axis denoting the mean IoU and pixel accuracy. As shown by the learning curve, we found that as training-set size increases, the predictive accuracy on testing data of the DeepLab model also grows correspondingly.

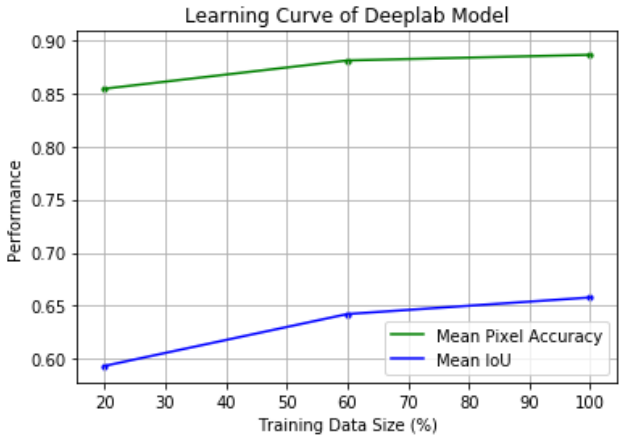


Figure 7: Learning Curve of DeepLab

Simple Model versus Sophisticated Model Figure 8 shows the mean pixel accuracy and IoU produced by FCN and DeepLab, both of which were trained on 100% training set. As demonstrated by Figure 8, sophisticated models have better predictive accuracy than simple ones, as the result of more reasonable network architectures. Although FCN has a great achievement in semantic segmentation, it has 2 main inherent drawbacks in its architecture. First, inevitably, down-sampling operations make the network lose a great amount of spatial information when feeding forward. The feature maps in last few layers correspond to a large region of original image, which lose too many details, like the boundary of objects. And it is too hard to recover the lost information. Second, FCN does not incorporate prior probability and ignores the constrains for boundary spatial information, which should be used to enrich the details of boundary between different objects.

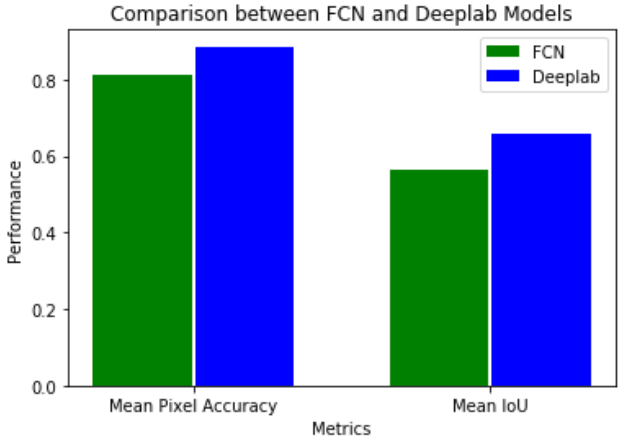


Figure 8: Comparison between FCN and Deeplab

To overcome these 2 drawbacks, DeepLab methods remove the down-sampling operator from the last few max pooling layers, and instead advocate the use of atrous convolution which is able to enlarge receptive field area without requiring extra computation. The benefit is that atrous convolution reserves rich spatial information and thus can generate dense prediction. Also, probabilistic graphical models are introduced into DeepLab as post-processing after neural networks. Fully connected CRFs focus on the probability that single pixel is classified as a certain class, and the differences between pixels or super-pixels. With the prior knowledges obtained from existing data, DeepLab can use it to help classify pixels. And it is more likely that two pixels belong to different classes, if the distance between two pixels or super-pixels is larger, which is determined by spatial distance in image and color values.

Confusion Matrix Analysis After training and testing our 4 models: FCN, DeepLab with 20%, 60%, 100% training data, we count prediction and ground truth label for each pixel in the testing set and output the confusion matrix which has dimension 21 by 21. And then we normalized each column as shown in Figure 9, Figure 10, Figure 11, and Figure 12.

The confusion matrix contains 21 classes: class 0 is "background", class 1 is "aeroplane", class 2 is "bicycle" class 15 is "person", etc. Each column in confusion matrix represents the ground truth label of certain class. And each row in confusion matrix represents the prediction for certain class. The entries with deeper color means the value is closer to the probability of 1.

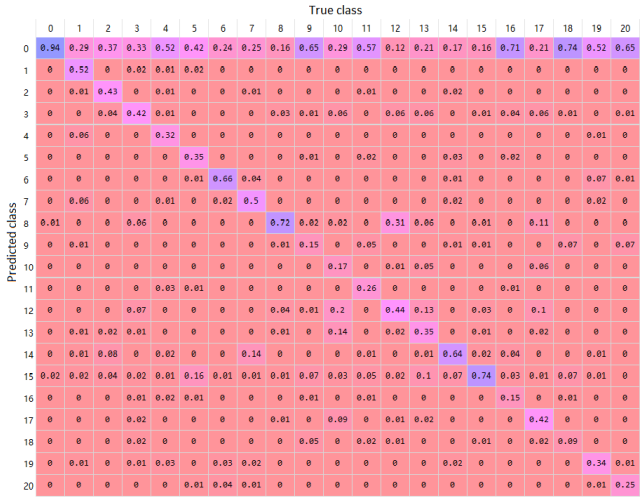


Figure 9: Confusion Matrix of FCN

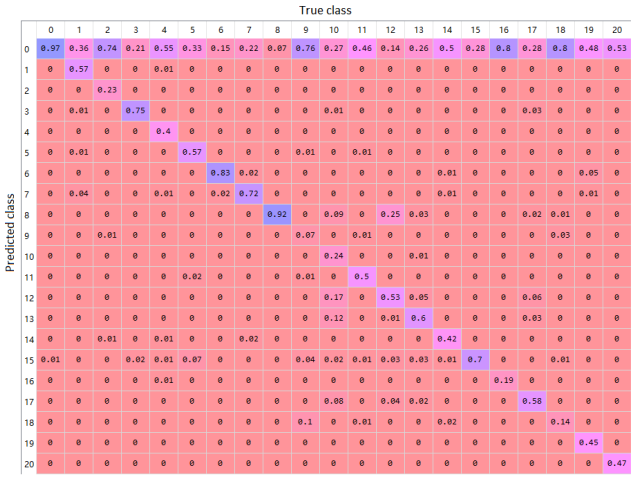


Figure 10: Confusion Matrix of Deeplab (20%)

Note that the diagonal entries of confusion matrix represent true positive; that is to say, those pixels are classified correctly. So it is not difficult to find that the most deep color entries lie on the diagonal of each matrix, which means the most pixels are classified correctly. However, it cannot be denied that there are still a great amount of pixels classified as class 0, the "background". That is because the boundary of objects are difficult to predict and they become blurred to some extent, especially some small objects or slender body, like the wheels of bicycle.

The interesting thing is that from FCN to DeepLab 20%, 60% and 100%, less and less deep color entries scatter outside the diagonal. This fact demonstrates that, from the detailed perspective of each class prediction, the overall predictive accuracy is improved by more sophisticated structure and more training data.

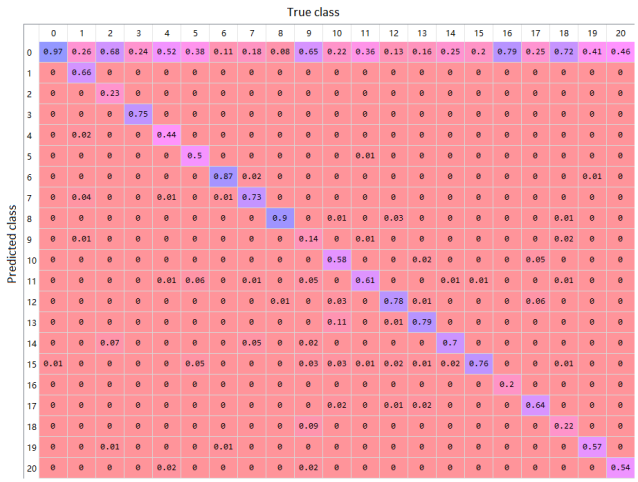


Figure 11: Confusion Matrix of Deeplab (60%)

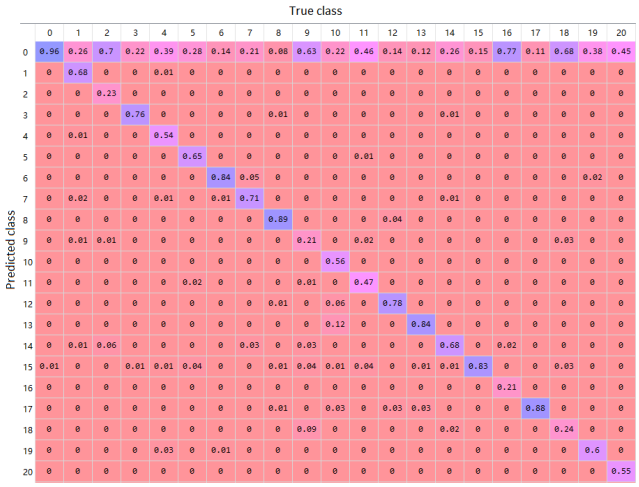


Figure 12: Confusion Matrix of Deeplab (100%)

Conclusion

Our project tests and compares the predictive accuracy between different methods and different training set sizes. We choose pixel accuracy, mean IoU and confusion matrix as our metrics. The final results turn out that the more sophisticated methods with reasonable structures and more training data we use, the more accurate the prediction will be.

Details about this project, including the code, model, and dataset, can be found at `https://github.com/YiwuZhong/Image_Segmentation_Deeplab_FCN`.

References

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2015. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR* abs/1511.00561.

Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2014. Semantic image segmentation with

deep convolutional nets and fully connected crfs. *CoRR* abs/1412.7062.

Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR* abs/1606.00915.

Chen, L.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *CoRR* abs/1706.05587.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. 1097–1105.

Long, J.; Shelhamer, E.; and Darrell, T. 2014. Fully convolutional networks for semantic segmentation. *CoRR* abs/1411.4038.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. *CoRR* abs/1505.04597.

Silva, T. 2018. *DeepLab_V3 Image Semantic Segmentation Network*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Yu, F., and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *CoRR* abs/1511.07122.

Zou, Y. 2017. *A TensorFlow Implementation of FCN*.