Semantic Image Segmentation with PSPNet and Dense CRF

Ge Shi
University of Massachusetts
geshi@umass.edu

Shuying Guan
University of Massachusetts
sguan@umass.edu

Xian Yang
University of Massachusetts
xianyang@umass.edu

Abstract

Semantic image segmentation has always been an popular topic in computer vision, where we assign every pixel its unique category label. However, classic methods only predict locally rather than integrate global information which comes from previous predicted labels. In order to gain a robust model to generate a better result, we will basically follow the pyramid scene parsing network to dig into its benefits and optimize its result through dense conditional random field. Our main contribution are threefold. We study and simulate the experiment in [10] with Tensorflow; We integrate the PSPNet[10] with CRF[5] as post-processing; We compared the used method with others and discussed its limits and possible improvement method.

1. Introduction

Deep Convolutional Neural Networks had long been a popular frame to deal with image classification task such as the pixel-level semantic image segmentation in the field of computer vision. However, basic models always omit some helpful information which could been obtained based on global scene category. This limitation of local prediction is always triggered by the similar appearance of objects even they actually belong to different categories without a good understanding of the given scene. So as to build a descriptor for a variety of scenes and integrate useful information of global featues, *Zhao et al.*[10] proposes the pyramid scene parsing network to make more reliable predictions.

Nevertheless, with a decent improvement achieved by PSPNet, it is still a trade off between classification accuracy and localization precision of objects. In some specific situations, such as autonomous driving, we might urge precise localization with less loss of spatial details rather than merely accurate classification. So we will basically utilize the method proposed by textitKrahenbuhl et al. 2011[4] as a post-processing metric to optimize the result from PSPNet. In conclude, this project mainly focus on two major part:

- Understand and implement the pyramid scene pooling net on the Cityscapes dataset, assess the result based on the pixel-level accuracy and eventually discuss the advantages and disadvantages of PSPNet on the chosen dataset
- Post-process the output from PSPNet with Dense CRF with a Mean Field Approximation and evaluate the potential improvement through this approach

2. Related Work

Krahenbuhl et al. 2011[4] propose an efficient conditional random field metric to exclude the limitation of traditional adjacency CRF, which implements a fully connected CRF to establish the pairwise energy potential based on all possible pairs of pixels in images. In his algorithm, CRF distribution will be approximated by a mean field through a Gaussian filter in feature space. The result demonstrates that long dependency CRF could significantly improve the accuracy of pixel-level classifiers.

Chen et al. 2016[1] combine the post processing CRF metric proposed by *Krahenbuhl et al.*[4] and deep convolutional neural networks to dig into the image segmentation problem on the PASCAL VOC-2012 dataset. Also, 'atrous' algorithm has been utilized to accelerate the computation of deep convolutional neural networks in a much more simpler scheme.

Noh et al. 2015[9] presents several drawbacks of fully connected networks. Since the receptive field is always fixed and predefined for a specific FCN architecture, the objects which have relatively smaller size could be omitted in the training process. Also, the objects which have relatively larger size could be labeled in different patches, where fragmented labels could be assigned in one single object. In order to conquer these shortcomings, they train a deep deconvolution network consists of deconvolution, unpooling and ReLU layers which will also handle different objects of different scales.

Liu et al. 2015[7] propose a new architecture called ParseNet to avoid locally ambiguous prediction. To achieve this, a joint prediction of all pixels in a image, which is also based on global context, will be made to replace the classic prediction method built on receptive regions or objects. In

order to obtain the global context vector during the training process, this approach will pool the feature map for a layer over the entire image. This global average pooling with FCN has been proved to be simple and robust on the PASCAL VOC2012 dataset.

Zhao et al.[10] first propose the pyramid scene parsing network to overcome the limitation caused by the local prediction of traditional convolutional neural network. To achieve a better quality result based on global prior contextual information, a hierachical model which contains information built on different scales of receptive fields and different sub-regions. They also apply this model on PASCAL VOC and Cityscapes to explore the superiority over a sequence of old networks such as FCN and dilated network.

3. Pyramid Scene Parsing Network

3.1. Problem Statement

The major drawback of current FCN methods is they are not able to effectively utilize the global context information as prior [10]. This may lead to conflict classification, for example, a boat on the river is mistaken as a car. In order to exploit global context information, we adopt pyramid scene parsing network to aggregate different-region-based context.

The motivation of our approach comes from the observation and analysis of the performance FCN methods and notice the following three challenging tasks.

- **Relationship Mismatch** There exists visual context that universally co-occur which is important for complex scene parsing. For example, an airplane tends to fly in the sky rather than on a road.
- **Confusion Categories** There are many confusing class label pairs such as building and house with similar appearance. FCN may predict two parts of an unique object as different class labels.
- **Inconspicuous Classes** There are several small size things in complex scene like streetlight and signboard which are easy to be ignored. Contrarily, there also exists big size object that cannot be included in one receptive field of view which causes discontinuous prediction.

To exclude these common mistakes, we find mining the contextual relationship and global information from different receptive fields can be effective to improve the performance of semantic segmentation.

3.2. Pyramid Pooling Module

In a deep neural network, how exhaustively the context information are used can be roughly indecated by the size of receptive field. To propose an effective global prior representation, PSPNet fused information from different subregions with various size of receptive fields. the pyramid pooling module is meant to set up a hierarchical global prior and information with different scales among different subregions are included in it.

The pyramid pooling module takes the final-layer-feature-map of deep neural network as input and fuses features under four different pyramid scales. The first level uses global pooling to generate the coarsest single bin output. The following pyramid levels do pooling on different sub-regions of the feature map and forms pooled representation with varied size from different locations. Then we use 1×1 convolution layer after each pyramid level to reduce the dimension of context to maintain the weight of global feature. We directly upsample the low-dimension featuer maps to get the same size features as the original feature map through bilinear interpolation. A the end, different levels of features with the original feature map are concatenated to form the final pyramid pooling global feature.

3.3. Network Architecture

With the pyramid pooling module, the proposed pyramid scene parsing network (PSPNet) as illustrated in 1. Given an input image in (1a), we use a pretrained ResNet [3] model with the dilated network strategy to extract the feature map. The final feature map size is 1/8 of the input image, as shown in Fig. (1b). On top of the map, we use the pyramid pooling module shown in (c) to gather context information. Using our 4-level pyramid, the pooling kernels cover the whole, half of, and small portions of the image. They are fused as the global prior. Then we concatenate the priors with the original feature map in the final part of (1c). It is followed by a convolution layer to generate the final prediction map in (1d). As stated in [7], PSPNet can be used in end-to-end learning and the pyramid pooling module and the local FCN feature can be optimized simultaneously without increasing its computational cost.

3.4. Evaluation Protocol

In order to evaluate the performance of Pyramid Scene Parsing Network, we will implement the standard Jaccard Index (known as intersection-over-union metric), which is commonly applied to assess the pixel-level semantic labeling accuracy. It is defined as:

$$IoU = \frac{TP}{TP + FP + FN}$$

where TP is the number of **Truth Positive** pixels, FP is the number of **False Positive** pixels and FN is the number of **False Negative** pixels.

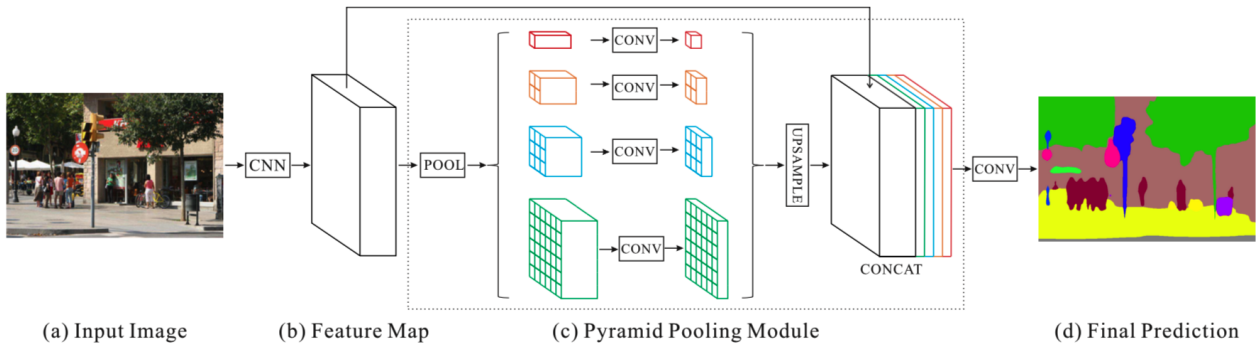


Figure 1: Overview of PSPNet Architecture

4. Dense Conditional Random Field

4.1. Fully Connected Random Field

Based on the efforts made by *Krahenbuhl et al. 2011*[4],we know that even though deep convolutional neural networks have been long proved to finish image classification task with decent accuracy, it is still a concern when the pooling layers in deep neural network introduce increased invariance and reduce the locality of predictions. That is, reliable label predictions of objects could have always been obtained in DCNN, however, the prediction near edges is sometimes obscure, which makes it hard to locate the exact location of objects. To accentuate the localization accuracy of prediction, we will implement a fully connected conditional random field metric (also known as Dense CRF) which as well eliminate the limitation of traditional adjancency CRF. In order to recover some local structures that are omiited in the DCNN, this long range CRF which considers all possible pair of pixels in a Gaussian kernel will eventually achieve our goal.

4.2. Energy Function

The energy function of the dense CRF is defined as:

$$E(x) = \sum_i \phi_i(x_i) + \sum_{i,j} \phi_{i,j}(x_i, x_j)$$

In our implementation, the unary potentials is well defined by the output of the PSPNet as $-logP(x_i)$, where $P(x_i)$ denote the probability map for each label at pixel i. And the color similarity among every pair of pixels (i and j) will be quantified as the pairwise potentials. Since every possible pair of pixels has been considered no matter the distance, this graph built based on this model is indeed fully connected. Also, to define the pairwise potential that is dependent on distance and color intensities of a pair of pixel, $\phi_{i,j}(x_i, x_j)$ is defined as:

$$\omega_1 exp(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2}) + \omega_2 exp(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2})$$

4.3. Mean Field Approximate Inference

As to gain a efficient approximation to the maximum posterior efficiency, it is suggested by *Krahenbuhl et al. 2011*[4] to use a mean field approximation. The algorithm is shown as following:

Algorithm 1 Mean Field Approximation

1: **Initialization** Q; where $Q_i(x_i) = \frac{1}{Z_i} exp(-\phi_u(x_i))$
2: **while** not converge **do**
3: $\tilde{Q}_i^{(m)}(l) \leftarrow \sum_{j \neq i} k^{(n)}(f_i, f_j) Q_j(l)$ for all m
4: $\hat{Q}_i(x_i) \leftarrow \sum_{l \in L} \mu^{(m)}(x_i, l) \sum_m \omega^{(m)} \tilde{Q}_i^{(m)}(l)$
5: $Q_i^{(m)}(l) \leftarrow exp(\phi_\mu(x_i) - \hat{Q_i}(x_i))$
6: Normalize $Q_i(x_i)$
7: **end while**

5. Experiments & Result

Our method takes pretrained model as input and implements PSPNet and CRF based on tensorflow for prediction. We evaluate our proposed method on urban scene understanding dataset Cityscapes [2]. In the following sections, we'll introduce the dataset and analyse the results using PSPNet and CRF with different value of parameters.

5.1. Cityscapes

Cityscapes is a dataset focuses on urban street scenes semantic understanding. It contains 5000 annotated images with fine high quality pixel-level annotations collected from 50 different cities of different time. It has 19 classes of objects with varying scene layout and background. As for the annotation policies, some background visible 'through'

some foreground object is considered to be part of the foreground like tree leaves in front of house or sky (everything tree), transparent car windows (everything car). The images are divided into sets with numbers 2,975, 500, and 1,525 for training, validation and testing. We took experiment on fine annotated images with our PSPNet and CRF combined mothods and take FCN as baseline comparison algorithm.

5.2. Pyramid Scene Parsing Network

We classify each pixel in the image using Pyramid Scene Parsing Network with pretrained model obtained from [10] and semantic the image according to the classified result. We performed experiment randomly on 21 images in the validation set on Cityscapes dataset. Statistics in Table 1 is cited from [10]. Statistics of other method (FCN[8], Deelplab[1], PSPNet[10])is based on Cityscapes testing set which ours is based on the random 21 images of validation set.

Some of our results are shown in 2. Our PSPNet result is missing some details compared to the ground truth, especially the small objects. They are either downsized or misclassified. The small objects, such as the building and the person between the tree in the second row image, or the slim traffic signs in the third row, is not classified. Other objects, such as vegetation, and fence in the fourth row and the fifth row is dilated. After using CRF as post-processing, the details is missing more due to the the small object would be discarded and reclassified as part of the surrounding object for the search surrounding is missing these objects and conclude the original one is misclassified since CRF is largely based on the surrounding information. However, it is clear that the segmentation is smoother.

Statistics in Table 1 shows that PSPNet[10] is better than FCN[8] and Deelplab[1], especially in class wall, fence, pole, traffic lights and so on. But the class with massive area, such as sky, road and building does not evidently outperform other methods. Our experiment result is sometimes outperformed the one in the paper, mainly because the number of our testing images is much smaller than the one in [10].

5.3. Dense CRF as Post-processing

We successfully implement CRF as post-processing after PSPNet. we read the scores of each pixel on 19 different classes and use softmax function to normalize them to probabilities. The probabilities are transferred to the unary potentials with log space. We designed kernels to (1) penalizes small pieces of segmentation that are spatially isolated – enforces more spatially consistent segmentations (2) create the color-dependent features – because the segmentation that we get from CNN are too coarse and we can use local color features to refine them. This step changes the probabilities of classes for each pixel. Finally we do multiple iterations of refinement using on CRF and find the label with largest probabilities. Figure 3 shows the result on an specific picture.

(3a) is the original picture with abundant information containing foreground and background such as road(dark purple), vegetation(dark green), terrain(light green), person(red), pole(gray), traffic sign(yellow). (3b) shows the result from PSPNet, we can see the poles and persons are obvious with smoothing boundaries. (3c) only utilizes color features and produces more rough boundaries which is meant for a more accurate classification between to objects. And also, the (3d) only penalize spatial distance and it smooth the boundaries and interpolate gaps to form a consistent presentation. (3e, 3f, 3g, 3h) combines the two methods and with iterations from one time to four times. With the number of iterations increase, more details are removed, so we find the CRF methods are not good for tiny or thin objects but lead to more clear boundaries.

6. Discussion

Based on the above analysis, we find there are multiple ways to improve our research.

- **Exhaustive Test and Validation** Because of time limit, we only perform experiments on 21 images in validation dataset. However, to get a more convincing conclusion, we will further do experiments on all the test images.
- **Fine Tuning parameters** We don't do much fine tuning works on the parameters of CRF but only take empirical values of parameters from previous researches.
- **End to End Training** According to [**?**], the PSPNet and deep CNN can be integrated to train end to end model.
- **SSVM Trained CRF** CRF can be trained in the large margin framework using structured support vector machine as stated in [6].

7. Conclusion

As mentioned before, PSPNet[10] is outperformed FCN[8] and DeepLab[1] in almost every class. But it still tend to lose small details, especially those inside other object. CRF doesn't work ideally as expect because it erases some details of tiny objects which is just contrary to the effect of PSPNet. We also discussed our work from different perspectives and proposed several ways for further research.

Method	road	swalk	build.	wall	fence	pole	tlight	sign	veg.	terrain
FCN[8]	97.4	78.4	89.2	34.9	44.2	47.4	60.1	65.0	91.4	69.3
DeepLab[1]	97.9	81.3	90.3	48.8	47.4	49.6	57.9	67.3	91.9	69.4
PSPNet[10]	98.6	86.2	92.9	50.8	58.8	64.0	75.6	79.0	93.4	72.3
PSPNet(ours)	**97.79**	**84.43**	**92.40**	**64.81**	**57.16**	**61.73**	**73.43**	**87.63**	**92.76**	**79.68**

Method	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
FCN	93.9	77.1	51.4	92.6	35.3	48.6	46.5	51.6	66.8	65.3
DeepLab	94.2	79.8	59.8	93.7	56.5	67.5	57.5	57.7	68.82	70.4
PSPNet(paper)	95.4	86.5	71.3	95.9	68.2	79.5	73.8	69.5	77.2	78.4
PSPNet(ours)	**94.12**	**86.19**	**69.18**	**96.12**	**54.56**	**77.20**	**29.01**	**85.81**	**80.66**	**77.09**

Table 1: Per-class results on Cityscapes with Different Methods

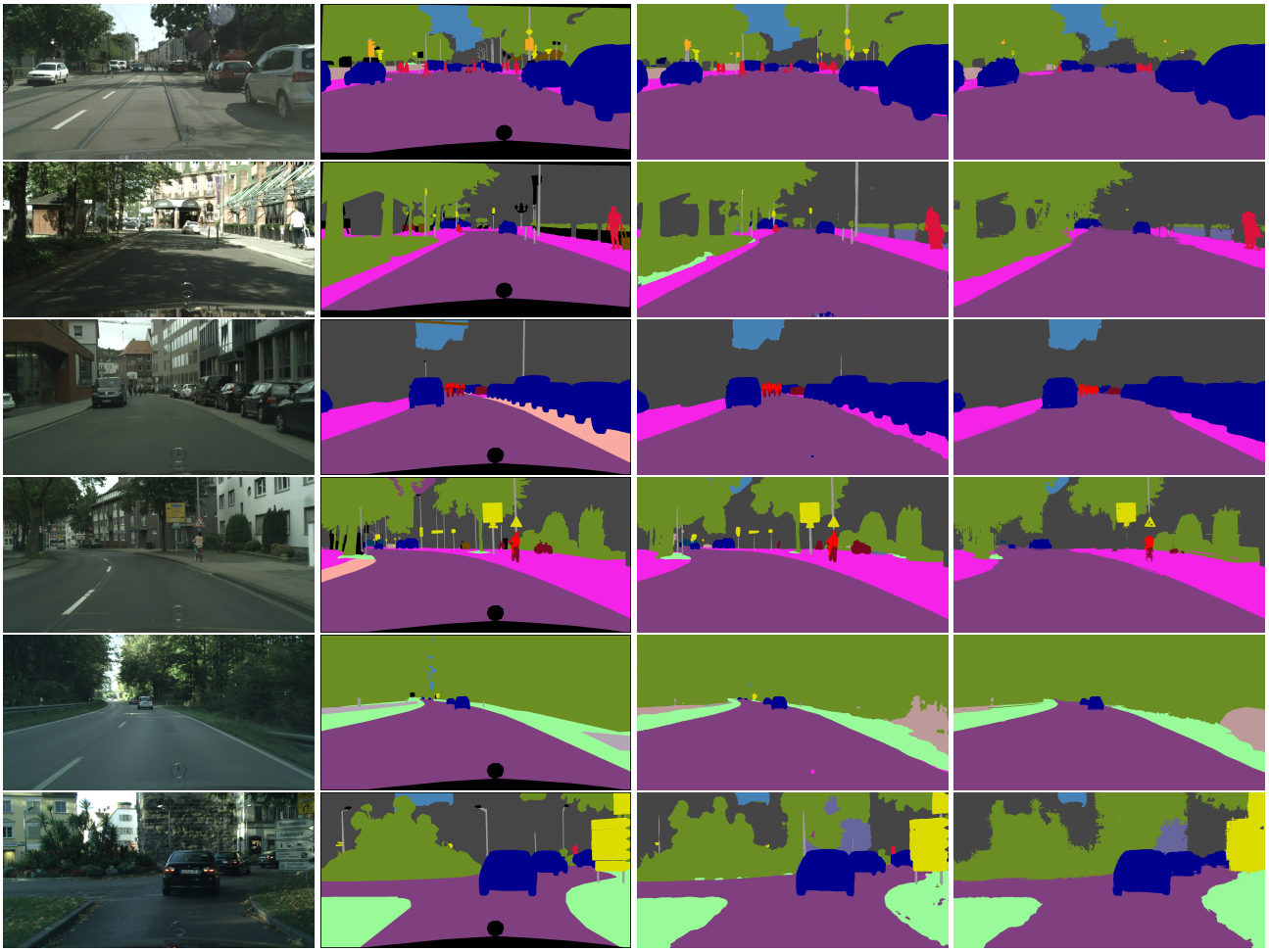
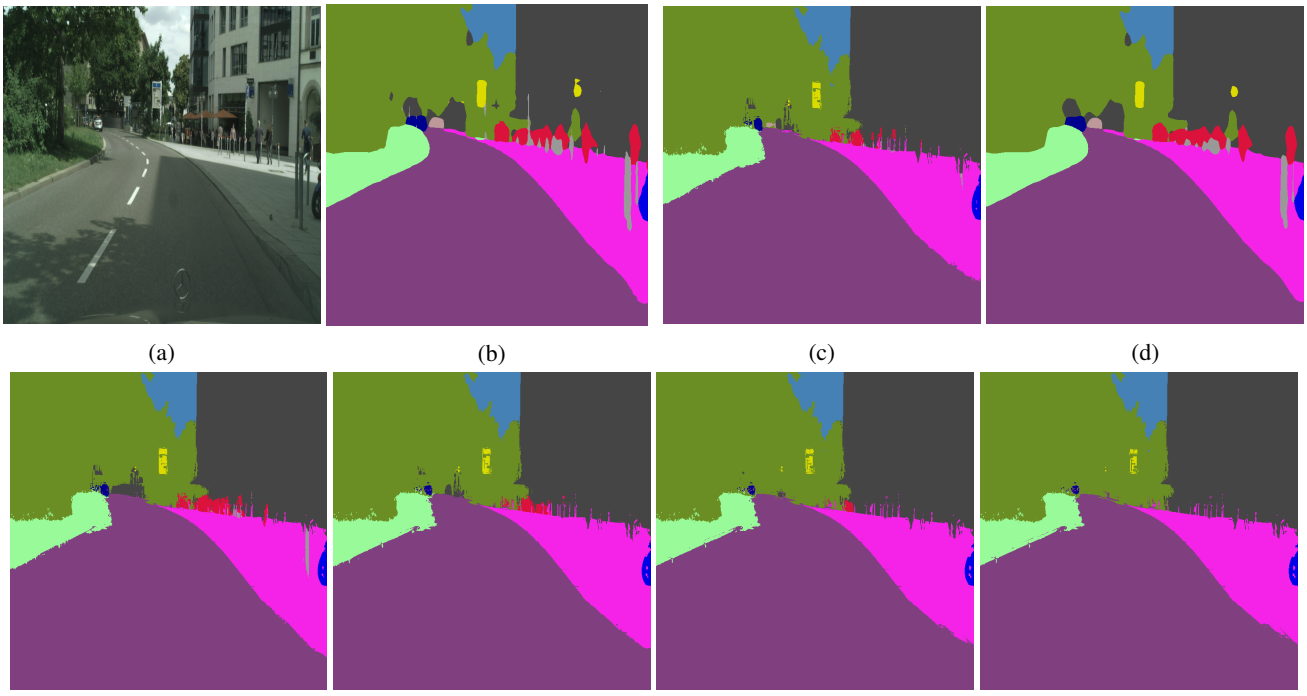


Figure 2: Examples of PSPNet results on Cityscapes dataset.(from left to right: image, ground truth, PSPNet, CRF)

References

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 3: Performance comparison between figures with or without CRF as post-processing and various parameters

[2] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[5] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *CoRR*, abs/1210.5644, 2012.

[6] F. Liu, G. Lin, and C. Shen. Crf learning with cnn features for image segmentation. *Pattern Recognition*, 48(10):2983 – 2992, 2015. Discriminative Feature Learning from Big Data for Visual Recognition.

[7] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[8] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[9] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. *arXiv preprint arXiv:1612.01105*, 2016.