Nguyen Thanh, Kien, Fookes, Clinton, & Sridharan, Sridha
(2016)
Deeper and wider fully convolutional network coupled with conditional random fields for scene labeling.
In Sharma, G & Pereira, F (Eds.) *Proceedings of the 23rd IEEE International Conference on Image Processing (ICIP).*
Institute of Electrical and Electronics Engineers Inc., United States of America, pp. 1344-1348.

https://doi.org/10.1109/ICIP.2016.7532577

DEEPER AND WIDER FULLY CONVOLUTIONAL NETWORK COUPLED WITH CONDITIONAL RANDOM FIELDS FOR SCENE LABELING

Kien Nguyen, Clinton Fookes, Sridha Sridharan

Image and Video lab, SAIVT Research Program
Queensland University of Technology, Brisbane, Australia

ABSTRACT

Deep convolutional neural networks (DCNNs) have been employed in many computer vision tasks with great success due to their robustness in feature learning. One of the advantages of DCNNs is their representation robustness to object locations, which is useful for object recognition tasks. However, this also discards spatial information, which is useful when dealing with topological information of the image (e.g. scene labeling, face recognition). In this paper, we propose a deeper and wider network architecture to tackle the scene labeling task. The depth is achieved by incorporating predictions from multiple early layers of the DCNN. The width is achieved by combining multiple outputs of the network. We then further refine the parsing task by adopting graphical models (GMs) as a post-processing step to incorporate spatial and contextual information into the network. The new strategy for a deeper, wider convolutional network coupled with graphical models has shown promising results on the PASCAL-Context dataset.

***Index Terms*—** Scene labeling, Scene parsing, Deep learning, Conditional Random Fields

1. INTRODUCTION

Scene understanding, as a core and ultimate problem of high level computer vision, is a task to understand the semantic contents in the observed scene [21, 32, 29]. Among the subtasks of scene understanding, scene labeling has attracted noticeable attention recently. Also known as scene parsing, scene labeling is a task to label every pixel in the image with the corresponding object class which it belongs to. After a perfect scene labeling, every region and every object is delineated and tagged [7]. Scene labeling is a challenging problem as it combines three traditional problems: object detection, segmentation and multi-labels recognition in a single process. In several scenarios, scene labeling is also named semantic segmentation because it faces both segmenting multi-class objects and identifying the category labels of the segmented objects.

From a feature representation point of view, Farabet *et al.* raised two fundamental questions in the context of efficient scene labeling [7]:

- Feature representation: How to produce good internal representations of the visual information (local features)?
- Contextual representation: How to employ contextual information to ensure the self-consistency of the interpretation (global features/relationship)?

Finding a good feature representation is critical to the segmentation task. Most traditional approaches rely on handcrafted features, e.g., color histogram, SIFT, HOG [32, 27]. Recently, deep learning has gained great popularity in learning to represent features for computer vision tasks. Since layer-wise learning algorithms were revised in 2006 [13], Deep Learning in general and large-scale Deep Convolutional Neural Networks (DCNNs) in particular have significantly advanced the performance of computer vision systems, including object detection, object segmentation, object recognition and natural language processing systems [19]. With their built-in hierarchical representation learnt directly from the data rather than human assumption, which is robust to translation, rotation, scale and deformation variation, DCNNs provide an effective response to the first question raised by Farabet *et al.*. In recent years, we have witnessed great success of DCNNs in scene labeling, outperforming all other traditional approaches (i.e. all top 10 approaches on the segmentation challenge in the PASCAL VOC 2012 dataset are CNN-based [1]).

In this paper, we propose a new approach to integrate spatial/contextual information into the network in a deeper and wider manner to improve the labeling accuracy. Our contribution is twofold:

- Firstly, we identify two trends (as detailed in Section 2) in the incorporation of spatial/contextual information into the labeling task and categorise the state-of-the-art approaches in the light of these trends.
- Secondly, we propose a deeper and wider convolutional neural network coupled with graphical modeling as a post-processing step for the labeling task. The proposed approach will been shown to effectively incorporate spatial/contextual information, which in turn increases the accuracy of the segmentation/labeling task.

The remainder of this paper is organized as follows: Section 2 presents our analysis of the state-of-the-art trends in integrating spatial information into DCNNs; Section 3 and 4 describe the background on Fully Convolutional Neural Networks and Fully Connected Conditional Random Fields respectively; Section 5 explains our proposed approach; Section 6 discusses the experimental results; and the paper is concluded in Section 7.

2. RELATED WORK

Incorporating spatial/contextual information into the parsing task not only provides self-consistency of the interpretation, but also improves the meaningful layout of the scene. To address this problem and deal with the second question mentioned above, the research community has been focusing on two main trends.

Trend 1: modifying the DCNNs themselves to incorporate larger relationship with the surrounding pixels. Farabet *et al.* [7] represented the raw input image in a 3-scale Laplacian pyramid before feeding them to three 3-stage convolutional networks. The outputs of three convolutional networks are concatenated, enabling large context to be integrated into local decisions, still remaining manageable in terms of parameters/dimensionality. Long *et al.* [25] casted the network into fully convolutional by replacing the last fully connected neural layers, which have fixed dimensions and throw away spatial coordinates, by 1×1 convolutional layers. Hariharan *et al.* [11] also collect features from intermediate layers in a similar approach to form so-called hypercolumn feature vectors for each pixel. Differently, the authors in [12, 8, 9] extracted two types of Region-CNN features: region features extracted from proposal bounding boxes and segment features extracted from the raw image content masked by the segments. However, Dai *et al.* [4] showed that using the masks in the image content as in [12, 8, 9] may lead to artificial boundaries. Liu *et al.* added global context to the DCCN by using the average feature for a layer to augment the features at each location [24].

Trend 2: integrating graphical models into the DCNNs to further refine the segmentation. A number of approaches apply DCNNs and CRFs as two separate approaches, then combining outputs. Farabet *et al.* employs a 2-layer neural network to combine deep learning features extracted by DCNNs and graphical model features extracted by CRFs [6, 7]. In the same vein, Kekec *et al.* use one DCNN to learn the CRF-type contextual information and another DCNN to learn visual features, then combine them for scene labeling [15]. Other researchers have applied DCNNs and CRFs in cascading order, which means using the output of the other as input. Liu *et al.* [23] oversegmented the input image into superpixels, calculated deep convolutional features, before feeding them into a CRF. A Structured SVM is employed to learn the parameters of the CRF model. Recently, the label assignment probability computed by the DCNN at each pixel location is employed to model the unary potential of the corresponding pixel. Both [7] and [2] simultaneously classified each pixel of the image densely by a multi-scale DCNN and oversegmented the image with superpixels. The predictors from the multi-scale DCNNs are then aggregated in each superpixel by computing the average class distribution within the superpixel. The aggregated probabilities of assigning classes to each pixel are then fed into a pairwise CRF model as unary potentials. Similarly, Chen et al. also employed the dense segmentation map computed by a DCNN to model the unary potential, but using a fully connected pairwise CRF model [3]. Guosheng *et al.* learn both unary and edge potentials of CRF models [22].

In this paper, we propose a new approach to take advantage of both trends discussed above. We propose a new architecture of the FCN to integrate deeper and wider spatial and contextual information in the deep network. The deeper and wider approach is further refined by a fully connected Conditional Random Fields (CRFs) post-processing step to improve the overall performance. The approach most closely related to ours is by Chen *et al.* [3] in which they employed a VGG-16 pre-trained model to extract deep features coupled with a fully connected pairwise CRF model for the labeling task. Our work is significantly different since we propose a deeper and wider FCN for extracting deep features. We will show in the experiments that our approach outperforms their work.

3. FULLY CONVOLUTIONAL NEURAL NETWORK (FCN) FOR SEMANTIC SEGMENTATION

Convolutional neural networks are powerful at visual modeling hierarchies of features. Networks such as LeNet [20], AlexNet [17] and its deeper successors [31, 30] have shown great success in various computer vision tasks [19]. These networks consist of multiple convolutional layers (convolution + nonlinear activation + pooling) followed by multiple fully connected layers [20, 17, 31, 30]. These fully connected layers have fixed dimensions and discard spatial coordinates. This, on the one hand, is useful for high level tasks such as recognition due to the robustness to locations of the objects, but on the other hand, is detrimental to lower level tasks such as segmentation. Observing that these fully connected layers can be viewed as convolutions with kernels that cover their entire input regions, Long *et al.* proposed to replace those fully connected layers with equivalent convolutions to cast the network into a fully convolutional network for the semantic segmentation task [25].

Following the approach in [25], we adapted the pre-trained models available including AlexNet [17], GoogLeNet [31] and VGG 16-layer net [30] to get a fully convolutional network. The network consists of five convolutional layers with each followed by a pooling layer. The outputs are achieved by upsampling and combining predictions at different levels. Three outputs are employed here: FCN-32s by 32× upsampling the prediction from the pool 5 layer, FCN-16s by 16× upsampling the summation of 2× upsampled prediction from the pool 5 layer and prediction from the pool 4 layer, and FCN-8s by 8× upsampling the summation of the summation in FCN-16s with prediction from the pool 3 layer.

4. FULLY CONNECTED CRFS FOR SEMANTIC SEGMENTATION

Fully connected CRFs are a type of discriminative undirected probabilistic graphical model which models the relationships of every pixel pair in the image. These models are effective in modeling the spatial information of the image, which is useful for the semantic segmentation task [16, 3]. Assume that we have a set of input images, $I = \{I_1, ..., I_N\}$, and its set of corresponding pixel-level image labelings, $X = \{X_1, ..., X_N\}$. Both sets I and X are random fields. There are k label classes for labeling each pixel, $L = \{l_1, ..., l_k\}$. By the fundamental theorem of random fields [10], a conditional random field (I, X) is characterised by a Gibbs distribution,

$$P(X|I) = \frac{1}{Z(I)} exp(-\Sigma_{c \in C_G} \Phi_c(X_c|I)), \tag{1}$$

where $G = (V, E)$ is a graph on X and each clique, c, in a set of cliques, C_G, in G induces a potential, Φ_c [18]. The Gibbs energy of a labeling $x \in L^N$ is $E(x|I) = \Sigma_{c \in C_G} \Phi_c(X_c|I)$. In a fully connected pairwise CRF model, G is the complete graph on X and C_G is the set of all unary and pairwise cliques. The corresponding Gibbs energy is,

$$E(x) = \Sigma_i \psi_u(x_i) + \Sigma_{i<j} \psi_p(x_i, x_j), \tag{2}$$

where i and j ranges from 1 to N. There are two factors affecting the energy of a labeling: unary potential, $\psi_u(x_i)$, and pairwise potential, $\psi_p(x_i, x_j)$. While the unary potential represents how likely a node takes on a label, the edge pairwise potential represents how likely the labels of two pixels agree. The unary potential normally incorporates shape, texture, location and color descriptor [16] or hand-crafted features such as SIFT [26] and HOG [5]. The pairwise edge potentials can be modeled as linear combinations of Gaussians,

$$\Psi_p(x_i, x_j) = \mu(x_i, x_j) \Sigma_{m=1}^{K} w^{(m)} k^{(m)}(f_i, f_j), \tag{3}$$

where each $k^{(m)}$ is a Gaussian kernel. For scene labeling, a contrast-sensitive two-kernel potential, combining an appearance kernel and a smoothness kernel has been proposed [16].

While performing inference, the best labeling is found by a Maximum A Posterior (MAP) approach,

$$\hat{x} = argmax_x P(x|I). \tag{4}$$

Since a fully connected CRF is capable of modeling the relationships between all pairs of pixels in the image, it better represents long-range, or even scene-level spatial relationships between pixels in the image. However, the cost of computation of modeling the relationship between all pairs of pixels in the image, makes traditional inference impractical even on a low resolution image. To deal with computational expense, Krahenbuhl and Koltun [16] employed a mean field approximation to the CRF distribution. This approximation reduces the complexity of message passing from quadratic to linear, resulting in an approximate inference algorithm for fully connected CRFs that is linear in the number of variables N and sublinear in the number of edges in the model.

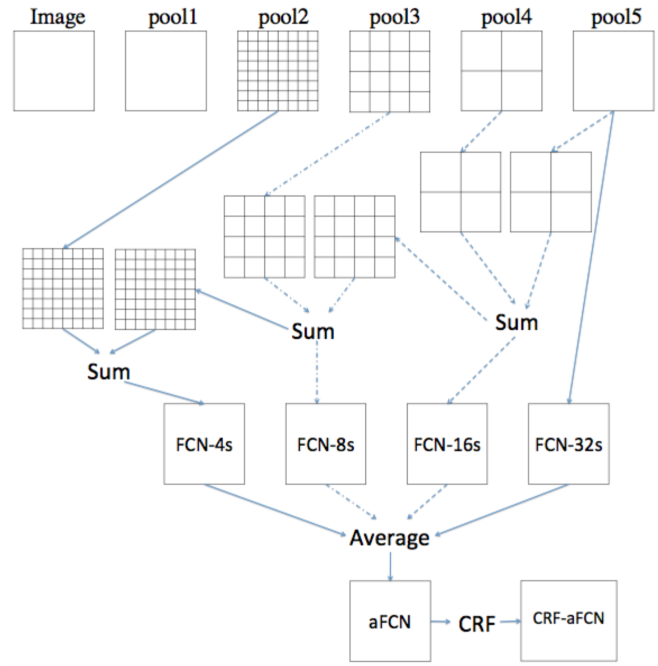

Fig. 1. Our proposed deep and wide fully convolutional neural network coupled with graphical modeling (CRF) for the scene labeling task.

5. THE PROPOSED DEEP AND WIDE APPROACH

In this paper, we propose a new architecture of the fully convolutional neural network to incorporate spatial and contextual information in a deeper and wider manner. This deep and wide fully convolutional neural network is then coupled with a fully connected Conditional Random Field to further refine the labeling task. The proposed approach is illustrated in Figure 1. Three highlights of the proposed approach are explained as follows.

Deeper: In a FCN, while later layers with large receptive fields are robust to object locations (which is useful for object recognition tasks), they discard spatial information, which is useful when dealing with topological information of the image (e.g. labeling). The earlier layers are finer and retain more details than the later and coarser layers. Incorporating early layers into the prediction has been shown to improve the overall segmentation accuracy [25]. In this paper, we propose to go deeper than [25] by incorporating predictions from the last four pooling layers into the final segmentation prediction step as shown in Figure 1. Following the same naming convention in [25], we call this FCN-4s. This output is established by summarising: (1) $4\times$ upsampled prediction from the pool2 layer, and (2) $2\times$ upsampled of the sum in the FCN-8s prediction.

Wider: While four prediction outputs FCN-32s, FCN-16s, FCN-8s and FCN-4s have managed to model spatial and contextual information at different levels, combining them for the final prediction will incorporate more clues for the labeling task. Hence we propose to extend the width of the prediction by going wider by averaging four prediction outputs to produce an average (aFCN). This averaging combines clues from multiple hierarchical levels to effectively identify the class of each pixel in the image.

Graphical modeling: To further model the spatial and contextual information in the network, we append a CRF as a post-processing step. A fully connected CRF approach, as discussed in Section 4, is employed to widely model the relationship between any pair of pixels in the image. The unary potentials are modeled as the label assignment probabilities at each pixel as computed by aFCN. The edge potentials are calculated based on appearance and smoothness kernel as discussed in Section 4.

Combining the above three innovations leads to a deep and wide fully convolutional neural network coupled with graphical modeling (CRFaFCN) for the scene labeling task.

6. EXPERIMENTAL RESULTS

PASCAL-Context dataset augments PASCAL VOC 2010 dataset with annotations for 500+ additional categories, allowing diverse tasks towards comprehensively parsing the images [28]. The dataset contains pixel-wise annotations for 10,103 images in the Training and Validation subsets of PASCAL VOC 2010 dataset. Due to the imbalance of the classes appearing in the dataset, similar to [28], we choose 59 most frequent classes for our experiments. The remaining classes are classified as background in this research. The scene labeling task has been performed on training subsets (4,998 images) and testing subsets (5,105 images) of the PASCAL-context dataset.

In this research, we employed a public framework called CAFFE [14]. Caffe is a clean and modifiable C++ framework with state-of-the-art deep learning algorithms for training and deploying general-purpose convolutional neural networks and other deep models efficiently on commodity architectures.

We first implemented the baseline as a fully convolutional neural network with three different outputs: FCN-32s, FCN-16s and FCN-8s as described in Section 3. The segmentation accuracy metric used here is the average pixel accuracy of the images. The segmentation accuracies achieved for three baseline outputs FCN-32s, FCN-16s and FCN-8s are 66.7%, 66.9% and 67.2% respectively.

Table 1. Performance of different approaches and the proposal for the semantic segmentation task on the PASCAL-context dataset.

Approaches	**Accuracy (%)**
FCN-32s [25]	66.7%
FCN-16s [25]	66.9%
FCN-8s [25]	67.2%
FCN-4s	67.7%
aFCN	69.5%
Chen *et al.* [3]	70.8%
CRFaFCN	71.9%

We tested our deeper proposal by incorporating the prediction from the pool2 layer as discussed in Section 5. The FCN-4s achieves a slight improvement in the accuracy to 67.7%. Our wider proposal is tested by averaging four outputs FCN-32s, FCN-16s, FCN-8s and FCN-4s as discussed in Section 5. The aFCN boosts the accuracy to 69.5%.

Lastly, we appended a CRF as a post-processing step to further incorporate spatial and contextual information. The CRFaFCN reaches 71.9%. Topping a CRF further improves the segmentation accuracy, which is in line with previous finding by Chen *et al.* [3]. We also compared the performance with the approach by Chen *et al.* [3] to show that our deeper and wider deep networks yields better results (71.9% vs. 70.8%). The experimental results are presented in Table 1.

The above experiments have shown the effectiveness of going deeper, wider and coupled with graphical-modeling for scene labeling. Our proposed deep and wide fully convolutional neural network coupled with graphical model advances the state-of-the-art baseline.

7. CONCLUSIONS

In this paper, we have analysed the state-of-the-art approaches to incorporate spatial/contextual information into deep convolutional neural networks in the light of two trends: Trend 1 includes approaches modifying the DCNNs themselves to incorporate larger involvement with the surrounding pixels and Trend 2 comprises approaches integrating graphical models into the DCNNs to further refine the segmentation task. Based on this analysis, we combine both trends by proposing a deeper and wider convolutional neural network coupled with graphical modeling as a post-processing step. The proposed approach has been shown to effectively incorporate spatial/contextual information, which in turn increases the accuracy of the labeling task.

8. REFERENCES

[1] Segmentation results: Pascal voc 2012. http://host.robots.ox.ac.uk:8080/leaderboard/. Accessed: 2016-01-20.

[2] J. Alvarez, Y. LeCun, T. Gevers, and A. Lopez. Semantic road segmentation via multi-scale ensembles of learned features. In *Computer Vision ECCV 2012. Workshops and Demonstrations*, volume 7584, pages 586–595. Springer, 2012.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *International Conference on Learning Representations, ICLR 2015*. 2015.

[4] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, June 2005.

[6] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Scene parsing with multiscale feature learning, purity trees, and optimal covers. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, volume 1, pages 575 – 582, 2012.

[7] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions on*, 35(8):1915–1929, Aug 2013.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587, June 2014.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PP(99):1–1, 2015.

[10] J. M. Hammersley and P. E. Clifford. Markov random fields on finite graphs and lattices. Unpublished manuscript, 1971.

[11] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[12] B. Hariharan, P. Arbelez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*, volume 8695, pages 297–312. Springer, 2014.

[13] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, July 2006.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[15] T. Kekec, R. Emonet, E. Fromont, A. Tremeau, and C. Wolf. Contextually constrained deep networks for scene labeling. pages Microsoft; NVIDIA; Qualcomm; Springer –, Nottingham, United kingdom, 2014.

[16] P. Krahenbuhl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems (NIPS) 25th*, Granada, Spain, 2011.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 2, pages 1097 – 1105, Lake Tahoe, NV, United states, 2012.

[18] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 282–289, USA, 2001.

[19] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[20] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[21] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 2036–2043, 2009.

[22] G. Lin, C. Shen, I. D. Reid, and A. van den Hengel. Efficient piecewise training of deep structured models for semantic segmentation. *CoRR*, 2015.

[23] F. Liu, G. Lin, and C. Shen. Crf learning with cnn features for image segmentation. *Pattern Recognition*, 48(10):2983 – 2992, 2015.

[24] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *http://arxiv.org/abs/1506.04579*, 2015.

[25] 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.

[26] D. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157 vol.2, 1999.

[27] A. Lucchi, Y. Li, and P. Fua. Learning for structured prediction using approximate subgradient descent with working sets. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1987–1994, June 2013.

[28] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 891–898, June 2014.

[29] S. F. Shenlong Wang and R. Urtasun. Holistic 3d scene understanding from a single geo-tagged image. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, June 2009.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, May 2015.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[32] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 702–709, 2012.