CNN BASED REGION PROPOSALS FOR EFFICIENT OBJECT DETECTION

Jawadul H. Bappy and Amit K. Roy-Chowdhury

Department of Electrical and Computer Engineering, University of California, Riverside, CA 92521

ABSTRACT

In computer vision, object detection is addressed as one of the most challenging problems as it is prone to localization and classification error. The current best-performing detectors are based on the technique of finding region proposals in order to localize objects. Despite having very good performance, these techniques are computationally expensive due to having large number of proposed regions. In this paper, we develop a high-confidence region-based object detection framework that boosts up the classification performance with less computational burden. In order to formulate our framework, we consider a deep network that activates the semantically meaningful regions in order to localize objects. These activated regions are used as input to a convolutional neural network (CNN) to extract deep features. With these features, we train a set of class-specific binary classifiers to predict the object labels. Our new region-based detection technique significantly reduces the computational complexity and improves the performance in object detection. We perform rigorous experiments on PASCAL, SUN, MIT-67 Indoor and MSRC datasets to demonstrate that our proposed framework outperforms other state-of-the-art methods in recognizing objects.

Index Terms— Region Proposal, Object Recognition, CNN, Object Localization.

1. INTRODUCTION

Over the years, the problem of recognizing objects in images has been studied extensively but still remains a challenging task. Most of the feature based object detection algorithms perform poorly in the face of variability of illumination, deformation, background clutter and occlusion. In recent years, the study of deep learning has been a growing interest due to its superior performance in several recognition tasks, for instance, activity recognition [1], object detection [2], and scene classification [3]. In this paper, we consider how CNNs can be used to *jointly propose regions and recognize object categories from these regions*.

In computer vision, the recent state-of-the-art methods present some promising approaches (e.g. R-CNN [2]) that demonstrate outstanding performance in terms of detection accuracy. However, this approach has large computational complexity in order to classify a large number proposed regions. One common technique used in existing detectors is selective search [4] to generate object proposals. However, this method performs exhaustive search and proposes large number of regions from an image. It is computationally expensive especially when deep features are extracted from a large number of proposed regions. In [5], it is observed that semantically meaningful regions can be spotted at deeper layers. Through the activation of receptive fields of the final convolutional layer, semantic regions are proposed to localize objects which can be used in object recognition. These semantic regions are very useful as they are related to objects in an image.

Towards this goal, we formulate our object detection framework using two CNN architectures. We call these two deep networks - *Region-Net* and *Object-CNN* (O-CNN) as shown in Fig. 1. Region-Net is used to generate the semantic regions for object proposals that are taken as input to the O-CNN in order to extract deep features. The information flow from region-Net to O-CNN helps us to build better object detector. More importantly, the whole framework performs very efficiently in terms of *computational time* when compared to the recent state-of-the-art methods. Fig. 1 illustrates our whole framework (details will be found on later sections). Our experimental results presented in Sec. 4 will attest the efficacy of our proposed approach.

Framework Overview and Main Contributions. In this paper, our goal is to design a *region-based object detection* framework for recognizing objects. Our proposed framework is formulated in two stages. In first stage, we use a CNN architecture, Region-Net with five convolutional layers to generate semantic regions for detection. We exploit the features from the final convolutional layer of the CNN to activate the receptive fields (RFs) to obtain the regions where an object might appear. Then, these proposed regions are used as input to O-CNN architecture (with five convolutional layers and two fully connected layers) to extract the deep features processed from the last fully connected (fc) layer (please see Sec. 3 for details). It is to be noted that same or different model can be used to learn the parameters for the two CNN architectures mentioned above. Finally, the deep features are used to model the class specific binary classifiers in order to predict labels. All steps are shown in Fig. 1. At the very end, bounding box regression method [6] is used to reduce the localization error.

Main Contribution: In our object detection framework, we show how a CNN can be used to propose regions by activating the receptive fields. This technique generates regions that are semantically meaningful and related to objects. More importantly, the number of proposed regions is significantly reduced when compared to current state-of-the-art region pro-

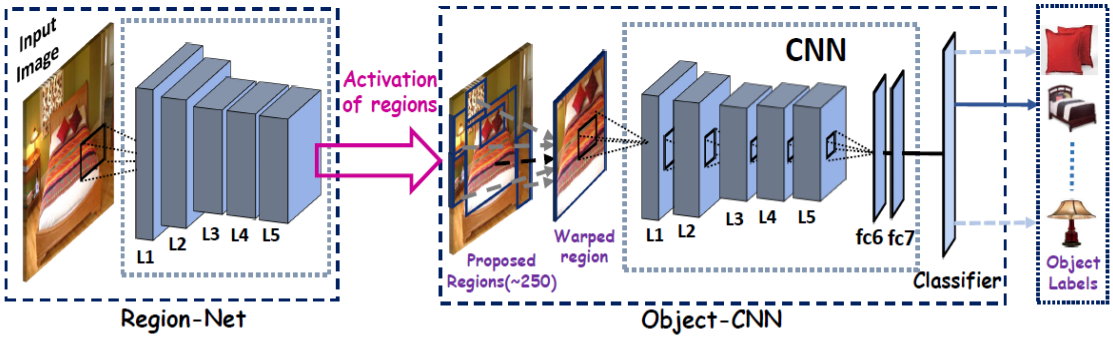


Fig. 1. Overall framework of our object detection model where two deep networks are used. We call these two CNN architectures 'Region-Net' and 'Object-CNN' respectively. Using the convolutional features of final layer of Region-Net, semantic regions are activated to localize the objects. Then, the features from activated regions are extracted to predict the labels.

posal approaches. This reduction leads to significantly less *computational cost* and our approach outperforms the existing state-of-the-art object detectors.

2. PRIOR WORK

Object detection is one of the fundamental problems in computer vision and has been studied for years to make it efficient and faster. Most of the object recognition methods involve edge (or contour) [7, 8] and patch [9, 10] based feature extraction. In object detection, some of the efficient techniques exploit sliding window [11] and boosting [12]. In [6], objects are represented using mixtures of deformable part models (DPM). In [13], authors show that DPM can be formulated as convolutional neural networks. Recently, convolutional neural network (CNN) has become dominating model due to its outstanding performance in object detection [2, 14, 15]. In [14, 15], deep neural network based regression approaches on bounding boxes are used to localize the objects.

Regions with CNN (R-CNN) presented in [2] shows promising results in detecting objects. However, finding regions by selective search [4] and then classifying approximately 2000 warped regions are indeed computationally expensive. In [16], authors demonstrate object detection using an additional spatial pyramid pooling layer. In [17], Region Proposal Network (RPN) is introduced to propose object proposals for better detection. It is shown in [5] that CNN supports different levels of representations (edges, texture, objects and scenes) of an image at different layers. The estimation of receptive fields (RFs) and their activation patterns can be used to visualize objects in different layers of the CNN model. In this paper. we propose a new approach to select regions from the activation of RFs at different units of the final convolutional layer of a CNN. It is computationally faster since it only gives fewer number of regions ($\sim$ 250) with compared to selective search ($\sim$ 2400) presented in [2]. We then demonstrate how the chosen regions can lead to higher object detection accuracy.

3. OBJECT RECOGNITION FRAMEWORK

Our framework follows three steps. Firstly, we generate region proposals to localize the objects. Secondly, we extract features

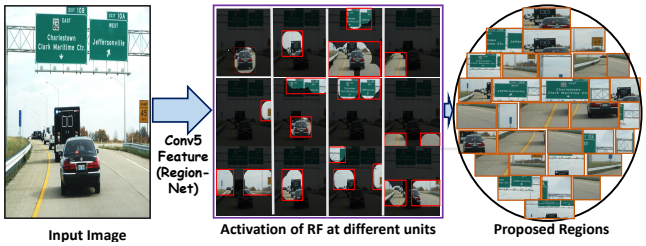


Fig. 2. Representation of region proposal approach for localizing objects. Given an image, receptive fields at 5^{th} convolutional layer produce the activation regions. Bounding boxes have been placed around the segmentation using contour approximation method to obtain regions. Best viewable in color.

of the proposed regions using deep convolutional neural network. Finally, we train a set of linear binary SVM classifier for an object class. The overall framework is shown in Fig. 1.

Region Proposal. Given an image, our region proposal approach provides some regions to localize the objects. It is shown in [5] that using the actual or empirical size of the receptive field (RF), semantically meaningful regions can be spotted at deeper layers. We use 5^{th} convolutional layer features of a CNN to activate the possible regions for object localization using empirical receptive field for each unit. We obtain approximately 400 regions on average from 256 units of Region-Net. Contour approximation method has been used to put a bounding box around the activated region. Fig. 2 demonstrates the visualization of the activated regions of an image using receptive field. We observe from Fig. 2 that all the objects of an image have been almost covered by the activation of RF of all the units from 5^{th} convolutional layer of the CNN. We approximately keep 250 regions from around 400 regions obtained from all the units of 5^{th} layer of the Region-Net. In order to do that, we first calculate Intersection over Union (IoU) among the bounding boxes of region proposals. If $IoU > \lambda$, we keep one from them and adjust the bounding boxes by multiplying a factor β.

Feature Extraction. For each region, we warp the image to fixed pixel size (227×227) that is required to make it compatible with the CNN. With each warped region, we extract features from a deep network with five convolutional and two

Methods	VOC2010 [21] dataset		SUN [18] dataset		MIT-67 [19] dataset		MSRC [20] dataset	
	accuracy	N_R	accuracy	N_R	accuracy	N_R	accuracy	N_R
DPM [6]	24.78%	-	18.79%	-	19.61%	-	48.20%	-
R-CNN [2]	50.46%	~ 2400	36.28%	~ 2400	32.07%	~ 2400	76.22%	~ 2400
R'-CNN	50.80%	20	37.63%	$240-260$	32.88%	$240-260$	76.79%	10
R'-CNN-FT	53.11%	20	$\mathbf{38.72}\%$	$240-260$	$\mathbf{32.95}\%$	$240-260$	-	-

Table 1. Mean average precision (mAP) of state-of-the-art methods and our method on VOC2010 [21], SUN [18], MIT-67 Indoor [19] and MSRC [20] datasets. Here, N_R denotes number of region proposals. With number of regions being 1-3 orders less, our proposed method R'-CNN is significantly faster than R-CNN while achieving similar or better performance.

fully connected layers. We consider the feature from last fully connected layer (fc7). For each region, we have the feature vector with 4096-dimension. These vectors are then fed into class-specific binary classifier.

Training Classifier. In order to detect objects, we train class specific binary classifier. For example, consider an object 'person', so the classifier will only give us the probability of a person given an input image. Let $f \in \Re^{4096\times 1}$ be the feature vector from the last fully connected layer (fc7) of the CNN. So, each binary classifier will compute the probability, $P(O_i = 1|f)$ where i denotes the object class. Now, it is very important to choose the training samples very carefully in order to train the classifiers. For each binary classifier, we split the training set into positive and negative examples. With all regions proposals, we have the bounding boxes with different IoU overlap with ground-truth bounding box. We select the $IoU \leq 0.3$ as negative examples. We only consider the ground-truth bounding boxes for each class as positive examples. With training features and labels, we get one linear SVM model per class. To fit the large training data we implement standard hard negative mining method [6, 22].

Bounding-Box Regression. To reduce the object localization error, we adopt the regression method as presented in [6]. We first train and then predict a new detection bounding box with pool5 features of O-CNN given a proposed region using linear regression model as in [2].

4. EXPERIMENT

In this section, we evaluate our proposed framework on four challenging datasets- VOC2010 [21], SUN [18], MIT-67 Indoor [19] and MSRC [20] datasets. These datasets are appropriate for our experiment since each image contains multiple objects with different categories.

PASCAL VOC 2010 Dataset. We evaluate our results on VOC2010 [21] dataset that contains 20 object categories with 11321 annotated images.

SUN Dataset. In SUN [18] dataset, each image contains an average of 7 different object categories. For our experiment we choose 14200 annotated images that have 120 object categories to evaluate our object detection performance.

MIT-67 Indoor Dataset. Even though MIT-67 indoor [19] scene dataset is largely used for scene classification but it also provides large varieties of object categories. To evaluate object detection performance, we select all the 2743 annotated images to recognize 70 object categories.

MSRC Dataset. We perform our experiment on MSRC [20] dataset with the ground truth provided in [23].

Experimental Setup. In this work, we use pre-trained model *ILSVRC2012* [24] to extract the features from CNN. We use caffe [25] to implement the CNN architecture. For the convenience of notation, we call our whole framework *(Region-Net + Object-CNN)* as R'-CNN in the rest of the paper.

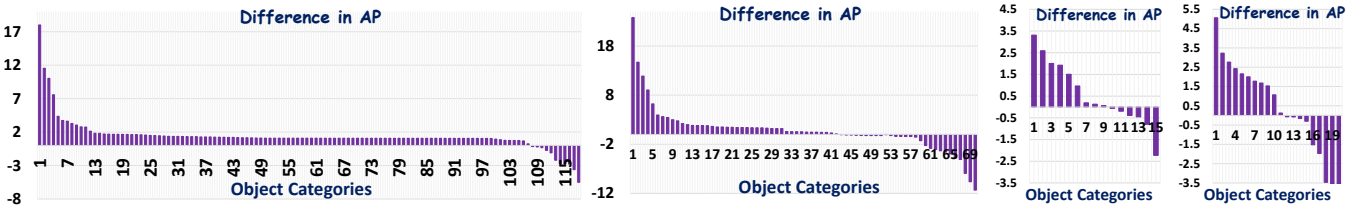
As discussed in Sec. 3, we choose approximately 250 regions from more than 400 regions using the parameters λ and β. We empirically choose the value of λ and β as 0.9 and 1.1. In MSRC dataset, 1-3 objects are present in an image and most of the objects occupy more than 50% area of the whole image. So, we sort out the top 10 activated regions with larger area from finally selected 250 regions for MSRC dataset. We also choose 20 regions on VOC2010 in similar way.

Training Data. To train the binary classifiers, we consider the groundtruth bounding boxes as positive samples. Bounding box of any region with $IoU \leq 0.3$ (compared to the ground-truth box) is considered as negative samples.

In this work, we also split the dataset to form training and validation set in order to fine-tune (FT) the parameters of O-CNN. Fine-tuning is required to adjust the CNN parameters to new datasets. We fine-tune our O-CNN with respect to VOC2010 [21], SUN [18] and MIT-67 datasets [19]. To fine-tune the network, we consider $IoU \geq 0.5$ between the bounding boxes of a proposed region and ground-truth as positive and rest of the regions as negatives. With $50k$ iterations we fine-tune the CNN parameters using stochastic gradient descent solver [25]. We do not fine-tune the parameters on MSRC since this dataset does not contain enough samples.

Evaluation Criteria. We calculate the average precision (AP) of each category comparing with the ground truth. Precision depends on both correct labeling and localization (overlap between object detection box and ground truth box). Let the computed bounding box of an object be O_b and the ground truth box be G_b, then the overlap ratio, $OR = \frac{O_b \cap G_b}{O_b \cup G_b}$. $OR \geq 0.5$ is considered as correct localization of an object if the label of the object is also correct. To calculate the AP of each object category, we perform thresholding on the detection confidence score. From the average precision (AP) of each category, we calculate the mean AP over all the categories.

Object Recognition Performance. Table 1 provides the mean AP of the state-of-the-art detectors implemented on VOC2010 [21], SUN [18], MIT-67 Indoor [19] and MSRC [20] datasets. We compare our R'-CNN detector with DPM [6] and R-CNN [2]. We implement both DPM and R-CNN



(a) SUN Dataset (b) MIT-67 Indoor Scene Dataset (c) MSRC (d) VOC2010

Fig. 3. Average precision (AP) improvement (percentage) of R'-CNN compared to the R-CNN [2] on (a) SUN [18], (b) MIT-67 [19], (c) MSRC [20], and (d) VOC2010 [21] datasets. The categories are arranged in decreasing order of AP improvement scores.

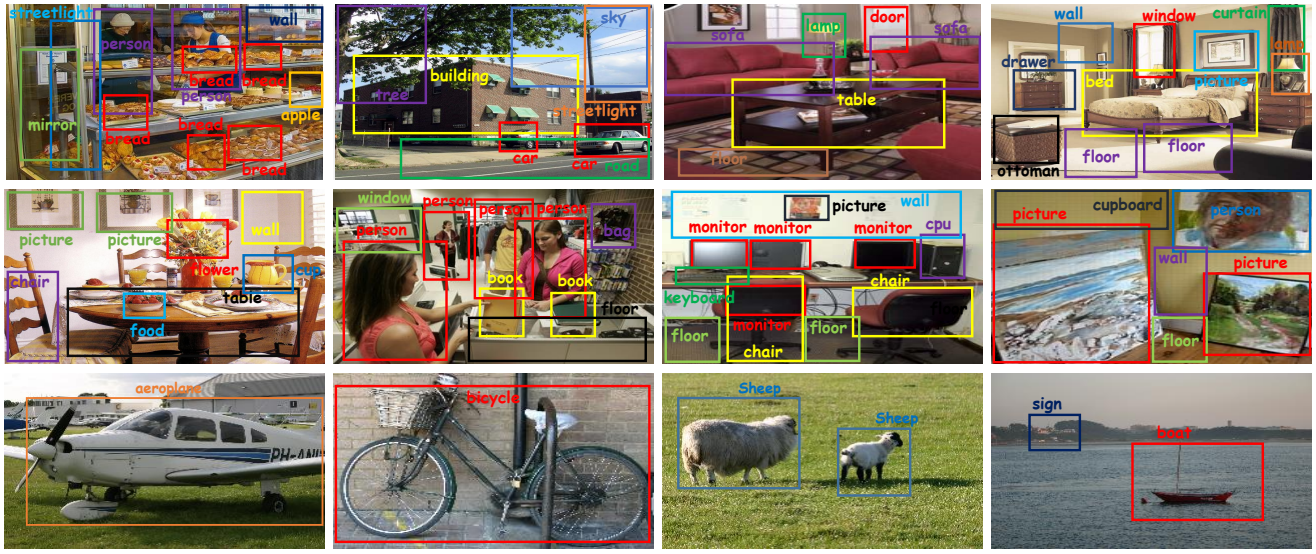


Fig. 4. Some examples showing object detection result. Images in the first, second and third rows are the localizations with label of different object categories on SUN [18], MIT-67 Indoor scene [19] and MSRC [20] datasets respectively. Different colors of box represent different object categories in an image. Sample detections on VOC2010 [21] dataset are shown in supplementary material. Best viewable in color.

on aforementioned datasets. We can see from Table 1 that R'-CNN outperforms both DPM [6] and R-CNN [2]. From Table 1, the best comparable result is found with R-CNN technique [2]. *Our method achieves better or similar performance when compared to R-CNN with significantly less number of region proposals.* With the fine-tuned parameters, our method further improves the accuracy as shown in Table 1. However, no significant change is observed after fine-tuning on MIT-67 Indoor dataset, because most of the classes of the pre-trained model *ILSVRC2012* are aligned with selected object classes.

The main advantage of using our region proposal technique in object detection is that it is approximately **9** times faster than R-CNN [2] on SUN [18] and MIT-67 [19] datasets, as we only classify around 250 regions instead of 2400 regions presented in [2]. For VOC2010 [21] and MSRC [20] dataset, our method is **120** and **240** times faster than R-CNN respectively. Results are shown in Table 1.

Is the proposed region proposal semantically meaningful? In order to measure the region proposal quality, we also calculate the ratio between the number of false positives (FP) and the number of proposed regions,$FPR = \frac{FP}{N_R}$ (N_R is the number of proposed regions). From our analysis, R-CNN [2] has higher FPR than ours by approximately 0.58%, 1.41%, 0.86% and 0.62% on VOC2010 [21], SUN [18], MIT-67 Indoor scene [19] and MSRC [20] datasets.

Classification Performance on Some Examples. The detection performance of our new region-based object detector is demonstrated in Fig. 3 and Table 1. Fig. 4 shows some of the object detection results using our proposed method. We observe that objects with larger pixel size are more inclined to have correct label and localization.

5. CONCLUSION

In this paper, we propose a novel framework for object localization and recognition where regions are proposed from a CNN. We exploit the interdependence between two deep networks in order to perform better in classifying objects. Our experimental results outperforms other state-of-the-art approaches in object detection. One direction of our future work could be examine how to incorporate context model on top of our proposed model for better detection.

Acknowledgement. This work is partially supported by NSF grants CPS-1544969 and IIS-1316934.

6. REFERENCES

[1] Junik Jang, Youngbin Park, and Il Hong Suh, "Empirical evaluation on deep learning of depth feature for human activity recognition," in *Neural Information Processing*, 2013, pp. 576–583.

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*. IEEE, 2014, pp. 580–587.

[3] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014, pp. 487–495.

[4] Koen EA Van de Sande, Jasper RR Uijlings, Theo Gevers, and Arnold WM Smeulders, "Segmentation as selective search for object recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1879–1886.

[5] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Object detectors emerge in deep scene cnns," *International Conference on Learning Representations*, 2015.

[6] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[7] J. Shotton, A. Blake, and R. Cipolla, "Multiscale categorical object recognition using contour fragments," *PAMI*, vol. 30, pp. 1270–1281, 2008.

[8] K. Schindler and D. Suter, "Object detection by global contour shape," *Pattern Recognition*, vol. 41, pp. 3736–3748, 2008.

[9] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *IJCV*, vol. 77, pp. 259–289, 2008.

[10] B. Ommer and J. Buhmann, "Learning the compositional nature of visual object categories for recognition," *PAMI*, 2010.

[11] Andrea Vedaldi, Varun ulshan, Manik Varma, and Andrew Zisserman, "Multiple kernels for object detection," in *ICCV*, 2009, pp. 606–613.

[12] Zhiquan Qi, Yitian Xu, Laisheng Wang, and Ye Song, "Online multiple instance boosting for object detection," *Neurocomputing*, vol. 74, no. 10, pp. 1769–1775, 2011.

[13] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik, "Deformable part models are convolutional neural networks," in *CVPR*, 2015.

[14] Christian Szegedy, Alexander Toshev, and Dumitru Erhan, "Deep neural networks for object detection," in *NIPS*, 2013, pp. 2553–2561.

[15] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov, "Scalable object detection using deep neural networks," in *CVPR*, 2014, pp. 2155–2162.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *ECCV 2014*, pp. 346–361. 2014.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[18] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," *CVPR*, 2010.

[19] Ariadna Quattoni and Antonio Torralba, "Recognizing indoor scenes," in *CVPR*. IEEE, 2009, pp. 413–420.

[20] Tomasz Malisiewicz and Alexei A Efros, "Improving spatial support for objects via multiple segmentations," *Proc. British Machine Vision Conference*, 2007.

[21] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[22] Xiaofeng Ren and Deva Ramanan, "Histograms of sparse codes for object detection," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3246–3253.

[23] Jian Yao, Sanja Fidler, and Raquel Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," in *CVPR*, 2012, pp. 702–709.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[25] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.