PEDESTRIAN DETECTION USING MASK-RCNN

Ganta Supriya
Department of Computer Scienceand Engineering
Amrita Vishwa Vidyapeetham
Coimbatore, India
cb.en.u4cse19617@cb.students.amrita.edu

Abstract

This study offers a generic, adaptable, and conceptually straightforward framework for object instance segmentation. Fundamental computer vision problems, such as recognizing pedestrians, have a wide range of applications in the industries of security, surveillance, autonomous vehicles, and robotics. Our method effectively locates things in a picture while also producing a top-notch segmentation mask for each object. By adding a branch for predicting an object mask in tandem with the existing branch for bounding box recognition, the technique known as Mask R-CNN expands Faster R-CNN. Faster R-CNN is run at 5 frames per second while Mask R-CNN adds only a little overhead. Occlusions still operate as a substantial obstacle even if persons detection is currently regarded to be a widely used technique. Here, we are striving to improve the detection performance in every way in order to more effectively manage the occlusion problem.

1. Introduction

Over a brief period, advancements in computer vision have improved both object detection and semantic segmentation results significantly. These improvements were primarily due to strong baseline systems; most notably, the Fully Convolutional Network and Fast/Faster R-CNN frameworks for respective semantic segmentation and object detection needs. These techniques provide adaptability, resilience, quick training timeframes for the ease of learning regardless of technical expertise along with efficient performance during inference whilst maintaining conceptual simplicity throughout development phases.

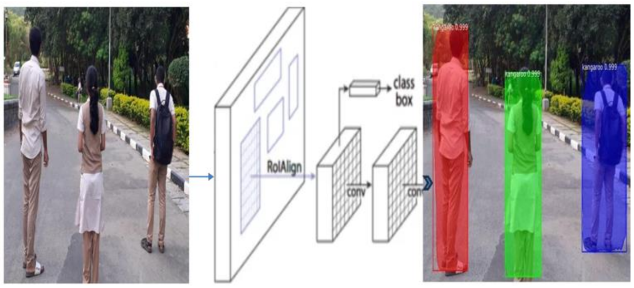


Figure 1. The **Mask R-CNN** framework for instance segmentation.

The main objective is to design a supportive framework like that found in segmentation while looking into how accurately detecting all objects within an image can simultaneously lead towards better instance segmentation results given their interconnected importance. Combining fundamental principles from both traditional computer vision tasks such as object detection- focused on categorizing individual objects using bounding boxes while localizing each accurately with those of semantic segmentation- aimed at classifying pixels under fixed categories while disregarding differentiating factors among object occurrences can improve overall performance.

Our method, A variant of Faster R-CNN called ***Mask R-CNN*** may forecast object masks in addition to the type and location of the objects. This can help with more accurate object tracking across numerous cameras. Mask R-CNN may be utilized to recognize individuals in each camera view and follow them across by connecting their masks between various viewpoints across many cameras. Mask R-CNN may also be used to segment objects in photos and movies at the pixel level. It can precisely determine an object's boundaries and separate them from the background.

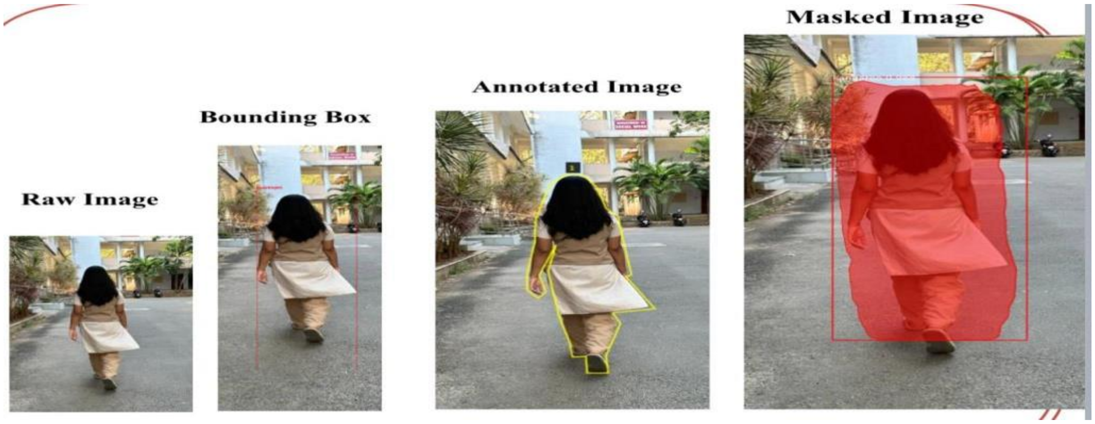


Figure 2. The flow of **OBJECT DETECTION** using **Mask R-CNN.** These are results of own data set. These results are based on ResNet-101, achieving a *mask* AP of 35.7 and running at 5 fps. Masks are shown in color, and bounding box, category, and confidences are also shown.

Related Work

R-CNN:

In order to successfully recognize bounding-box objects, the Region-based CNN (R-CNN) method focuses on a manageable number of potential object regions and evaluates convolutional networks separately for each RoI. Roi Pool was added to R-CNN to enable attention to Roi's on feature maps, resulting in increased speed and accuracy. By using a Region Proposal Network (RPN) to understand the attention mechanism, faster R-CNN progressed this stream. Faster R-CNN is the current top framework in many benchmarks since it is adaptable and robust to numerous subsequent enhancements.

Annotation:

Image annotating is a common aspect of the training process for computer vision models that handle image data for object identification, classification, segmentation, and other applications. For example, a collection of photos that have been labelled and annotated to identify and classify specific objects is required to train an object identification model. Auto annotation technologies often use pre-trained algorithms that can accurately annotate photos. For difficult annotation jobs like constructing segment masks, which take a lot of time, their annotations are crucial. In these situations, auto-annotate tools support manual annotation by offering a foundation from which additional annotation can be carried out. Any errors in the labels are duplicated as well because image annotation establishes the criteria that the model strives to follow. As a result, accurate image annotation is one of the most crucial computer vision jobs since it provides the framework for training neural networks.

Instance Segmentation:

Many techniques to instance segmentation are based on segment suggestions due to the success of R-CNN. Prior techniques were re-ordered into bottom-up phases. Following works, including Deep Mask, choose out segment candidates for classification by Fast R-CNN. These techniques start with segmentation, which is sluggish and less precise, before moving on to recognition. Like this, Dai et al. suggested a multi-stage complex cascade that predicts segment proposals from bounding-box suggestions, then follows up with classification. Our approach, which is simpler and more adaptable, is based on the parallel prediction of masks and class labels.

The success of semantic segmentation motivates the development of a new family of instance segmentation systems. These techniques aim to divide pixels belonging to the same category into distinct instances starting with per-pixel classification results (for example, FCN outputs). Mask R-CNN is built on an instance-first strategy as opposed to these methods' segmentation-first approaches. We anticipate that both methodologies will eventually be investigated in greater depth.

Mask R-CNN:

The idea of Mask R-CNN is straightforward: Faster R-CNN outputs a class label and a bounding-box offset for each candidate item; we then add a third branch that outputs the object mask. Thus, the concept of Mask R-CNN is simple and intuitive. The additional mask output, however, differs from the class and box outputs and necessitates the extraction of an object's considerably more precise spatial arrangement. The major component of Fast/Faster R-CNN, pixel-to-pixel alignment, is then introduced as one of the essential elements of Mask R-CNN.

Faster R-CNN:

The quicker R-CNN has two stages. Candidate object bounding boxes are suggested in the first stage, known as a Region Proposal Network (RPN). The second stage, which is essentially Fast R-CNN, executes classification and bounding-box regression after extracting features using Roi Pool from each candidate box. For quicker inference, the features used by both steps can be combined. For the most recent, thorough comparisons of Faster R-CNN and other frameworks.

Mask R-CNN: The same two-stage technique, with the same first stage (RPN), is used by Mask R-CNN. In the second stage, Mask R-CNN additionally produces a binary mask for each RoI in addition to class and box offset predictions. In contrast, most modern systems depend on mask predictions for classification. Our method is in accordance with Fast R-CNN, which simultaneously performs bounding-box classification and regression (which ended up greatly simplifying the multi-stage pipeline of original R-CNN).

Loss Function for Masking

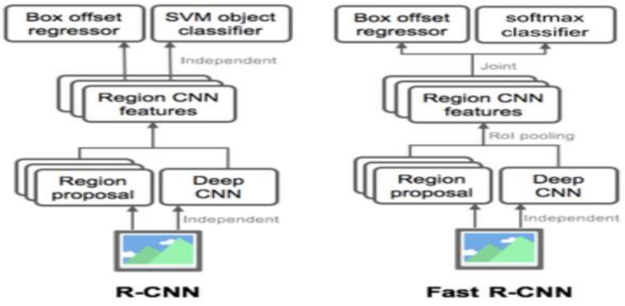
The loss function is a multi-task loss:

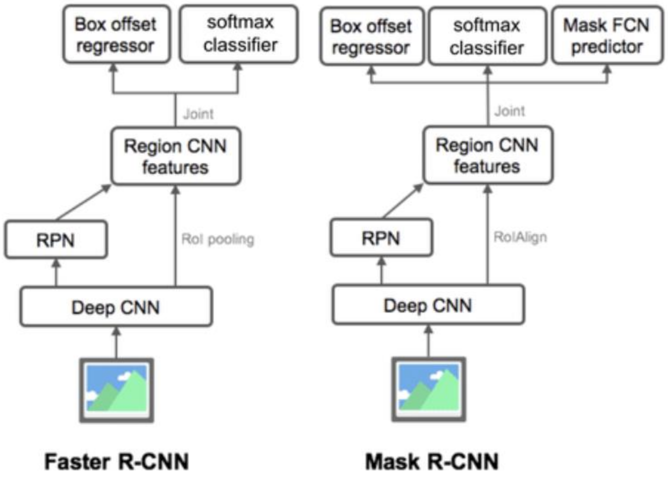
$$L = L_{cls} + L_{box} + L_{mask}$$

Like Faster R-CNN Lbox, Lcls is the classification loss. Like Faster R-CNN Lmask, the bounding box loss loss of the binary mask. For each RoI, this mask branch outputs Km2, which are K binary masks of mm resolution, each of which represents K classes.

Mask Representation: The spatial layout of an input object is encoded bmask. In contrast to class labels or box offsets, which fully connected (fc) layers must eventually collapse into brief output vectors, retrieving the spatial structure of masks can thus be addressed easily by the pixel-to-pixel correlation offered by convolutions.

We specifically use an FCN to predict a m m mask from each RoI. This prevents the specific spatial layout of each layer in the mask branch from being collapsed into a vector representation that lacks spatial dimensions. Our fully convolutional representation requires fewer parameters and is more accurate than earlier approaches that re-sort to fc layers for mask prediction.





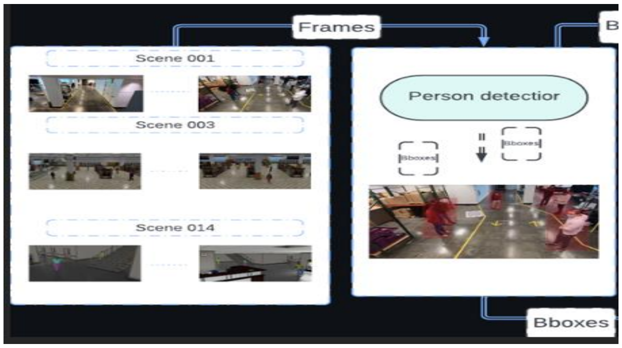


Figure 3. The architecture of **OBJECT DETECTION** using **Mask R-CNN.** These are results of AI City Challenge data set and frames have been extracted from data.

Implementation Details

Data Set: -

we have collected our own dataset from our college for our reference with a size 300 images and also, we have collected the dataset from AICITY challenge 2023 with a size of 17 GB (10GB training +7GB testing)

Dataset Link :

http://www.aicitychallenge.org/2023-track1-download/

Video frame extraction:

With a frame rate of 60 frames per second, video frames from seven cameras are extracted. However, as the provided ground-truth frame annotations are extracted at a frame rate of 10 frames per second, we should compare the relevant frame pictures between extracted frames and frame annotations when assessing the performance.

Annotation: -

Annotation is important at the time of detection. very good annotation leads to greater accuracy for the Mask R-CNN. We manually annotated 7000 frames from AI City Challenge (10 GB training dataset) using LabelImg Annotation tool and also we annotated 300 images from our own dataset. We also have used annotation tools like LabelME Annotation tool , VGG Annotator but results using those tools were not accurate

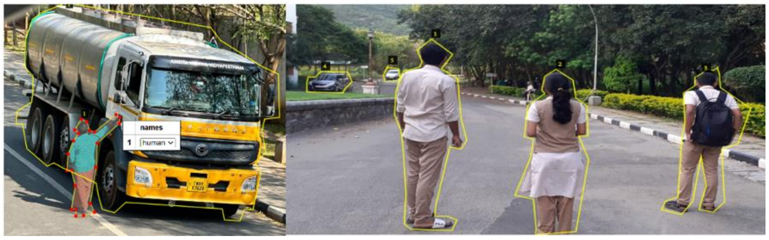


Figure 3. these are annotations for our own dataset using VGG Annotator

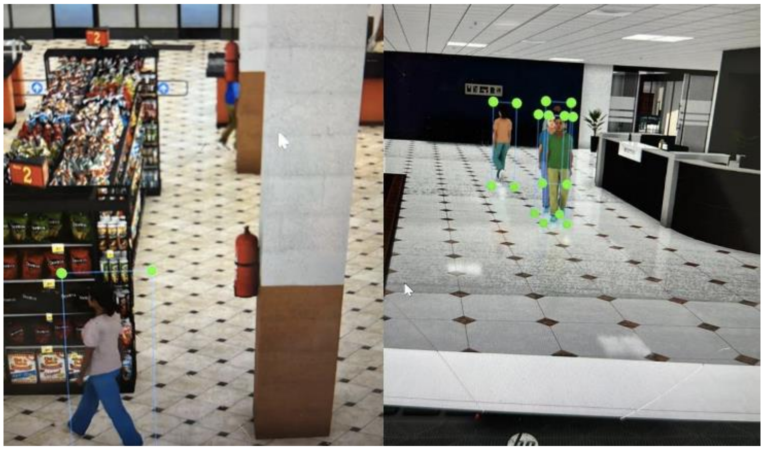


Figure 3. These are annotations for Ai City challenge dataset using VGG Annotator

Training: -

After extracting the frames from the video (10 frames per second), we manually annotated the frames in the xml format using the Labeling annotation tool. These xml files were then trained by us using the tensor flow background and the Quda 10.2 and Quda 10.4 environments to produce our Mask RCNN trained model with annotated frames of 7000. As a result, the results are more accurate, and the masking of the human was obvious with the test data set.

The precision of the Mask RCNN will increase if we train with more annotations about 10000 frames.

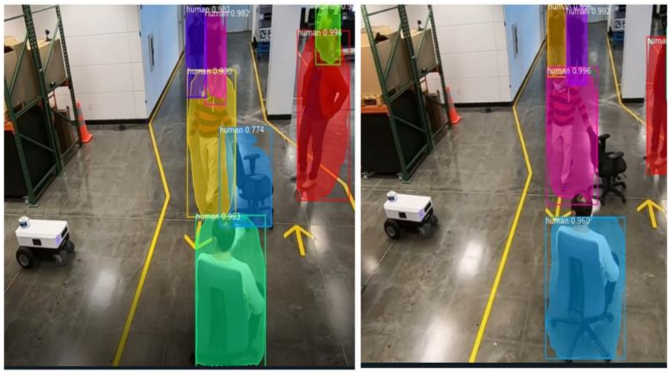


Fig4.1 Fig4.2

Figure 4. These are results of Mask R-CNN for AI City challenge testing dataset and also the comparison between fig4.1&fig4.2 the accuracy increased in fig4.2

Experiments: -

We thoroughly evaluate Mask R-CNN against the state of the art and carry out rigorous ablations on the AI City dataset. The typical COCO metrics, such as AP (averaged over IoU thresholds), AP50, AP75, and APS, APM, and APL (AP at various scales), are reported. AP is evaluating employing mask IoU, it should be stated. We train utilizing the union of 7000 train photos, much like in earlier work.

Main Results: -

Mask R-CNN outputs are visualized in Figures 4and 5.Mask R-CNN achieves good results even under challenging conditions. In Figure 4, 5we compare our Mask R-CNN baseline and FCIS+++ [26]. FCIS+++ exhibits systematic rtifacts on overlapping instances, suggesting that it is challenged by the fundamental difficulty of instance segmentation. And, in Figure 4.1 and 4.2

In the first image of the 4th figure shows that a chair is detected as a human with an accuracy of **77 (**4000 images are annotated) and in the second image the accuracy is improved, and it didn’t detect the chair as human.

also, in Figure 5.1 and 5.2

In the first image of the 5th figure shows bikes are also detected as humans along with girl, and in the second image the accuracy is improved, and it didn’t detect the bikes as human.

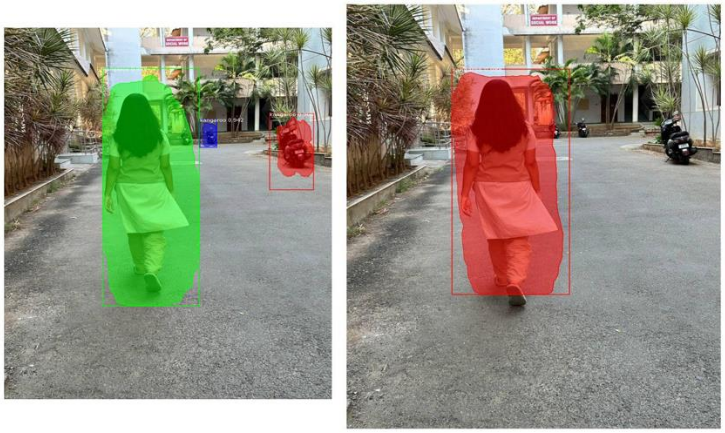


Fig5.1 Fig5.2

Figure 5. These are results of Mask R-CNN for our own dataset the comparison between fig5.1&fig5.2 the accuracy increased in fig5.2

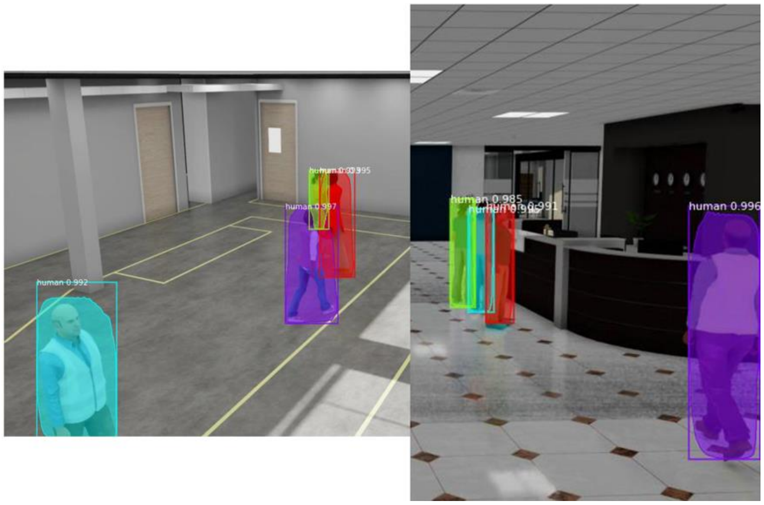


Figure 6. these are results from AI city Challenge training dataset we can see that all people are overlapping. Even though the accuracy is high.

Occlusion problem: -

The occlusion issue in object detection can be solved with the help of the well-liked and efficient architecture known as Mask R-CNN (Region-based Convolutional Neural Networks with Masking). Incorporating instance segmentation, Mask R-CNN is a development of the Faster R-CNN framework that not only detects objects but also offers pixel-level segmentation masks for each identified object.

In figure 6 we can see that all people are overlapping. Even though the accuracy is high. In contrast to conventional bounding box-based object detection techniques, Mask R-CNN can manage occlusion more successfully by creating pixel-level masks. Even when objects are completely or partially obscured, instance segmentation masks can nevertheless clearly define their borders.

By using the instance segmentation masks offered by Mask R-CNN, objects can be segmented down to the pixel level, including occluded areas. As a result, the model can distinguish between various objects in a scene and handle occlusion better.

Occlusion Handling: Mask R-CNN can accurately forecast masks for occluded objects by learning to correlate object components that are partially visible owing to occlusion. Mask R-CNN can infer the existence and placement of obscured objects by considering the contextual information offered by the mask predictions.

End-to-End Training: In order to jointly optimize for object detection and instance segmentation, Mask R-CNN is trained from beginning to end. Due to the model's ability to learn from and adapt to occlusion scenarios during training, it is better able to handle occluded objects during inference.

In Figure 7 the Mask RCNN model detects the humans who are not completely visible the camera u can see the results in which at the right and left corner of the image have detected the legs of the human

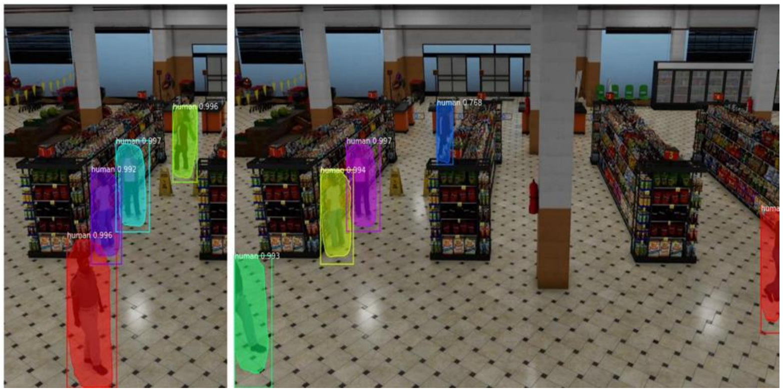
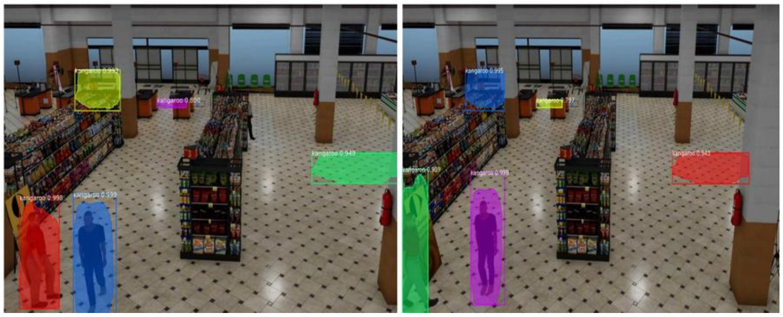


Figure 7. here some of the results of Mask RCNN with excellent accuracy where it detects the parts of the human.

Exceptional Case Results :-



Conclusion: -

The real-world surveillance video dataset and method for instantaneous anomaly identification in video surveillance systems were both discussed in this study. Future study on multiple conventional anomaly detection will be influenced by the datasets used in the experimental evaluation. Major implementations of the suggested approach indicate that the sub-action description offers thorough details on human actions. It lessens misclassifications brought on by a bigger number of activities made up of varying levels of distinct sub-actions. Our proposed method also localises and accurately detects the actions of a large number of people at a minimal computational cost. The suggested Ex-Mask R-CNN performs better than other Mask R-CNNs at face mask recognition. We can later modify our suggested strategy for large data platforms to work with different datasets.

References: -

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.

[3] A. Arnab and P. H. Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *CVPR*, 2017.

[4] M. Bai and R. Urtasun. Deep watershed transform for instance segmentation. In *CVPR*, 2017.

[5] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, 2016.

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[7] 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 *CVPR*, 2016.

[8] J. Dai, K. He, Y. Li, S. Ren, and J. Sun. Instance-sensitive fully convolutional networks. In *ECCV*, 2016.

[9] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*, 2015.

[10] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016.

[11] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In *NIPS*, 2016.

[12] R. Girshick. Fast R-CNN. In *ICCV*, 2015.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[14] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. In *CVPR*, 2015.

[15] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*. 2014.

[16] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015.

[17] Z. Hayder, X. He, and M. Salzmann. Shape-aware instance segmentation. In *CVPR*, 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*. 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[20] J. Hosang, R. Benenson, P. Dollár, and B. Schiele. What makes for effective detection proposals? *PAMI*, 2015.