A Deep Framework for Automatic Annotation with Application to Retail Warehouses

Kanika Mahajan, Anima Majumder*, Harika Nanduri, Swagat Kumar
Tata Consultancy Services
Bengaluru, India
(mahajan.kanika,anima.majumder,harika.nanduri,swagat.kumar)@tcs.com

Abstract

The paper presents a novel deep learning framework for automatic annotation and segmentation of densely cluttered objects in a warehouse application use-case as specified by the Amazon Robotics Challenge (ARC) 2017. This framework addresses two challenges of the competition: (1) reducing the amount of manual labour involved in generating a large number of annotated data that could be used for training a deep network and, (2) achieving good segmentation accuracy in a very limited amount of training time ($\leq$ 30 minutes). These two problems are solved by proposing a deep architecture comprising of Residual Network and Feature Pyramidal based convolutional neural network that helps to retain primitive features along with higher level features obtained from each successive layer. In addition, a framework is proposed using this network to automatically generate a large annotated dataset having different degrees of clutters to carry out multi-class semantic segmentation after training with this machine generated dataset. The proposed framework is shown to provide better segmentation accuracy with lesser training time as compared to the existing state-of-the-art architectures such as PSPNet and Mask R-CNN. The overall working of the proposed architecture is explained by creating a new dataset from the objects specified by the ARC competition. An extensive experiment is also performed using the MIT-Princeton database [22]. Our *TCS-ARC-Dataset* [13]is made available online for the convenience of readers.

1 Introduction

The E-commerce companies like Amazon deploy thousands of wheeled mobile robots [19] to move goods within their warehouses. However, it still requires several hundred people in each warehouse to do things like pulling items from shelves and placing them into packaging boxes to be shipped to the user. Robots that can automatically pick and place items would boost the efficiency of operation by reducing the reliance on human workers which is very expensive in highly competitive e-commerce market with very narrow profit margins. The development of such robots require expertise in multiple domains. In this paper, we are focusing on the perception part of the system responsible for identifying and segmenting objects in a dense clutter. The problem is challenging due to uncertainties arising from varying illumination, partial occlusion, changing shape and sizes as well as varying view point.

Amazon has been organizing Amazon Picking / Robotics Challenge (APC/ARC) [18] [6] [15] consecutively over last three years in order to spur and encourage research and de-

velopment in this direction. While most of the participating teams in 2015 APC competition relied on traditional feature-based image processing methods for object recognition [1] [7], the APC 2016 event witnessed the dominance of deep learning algorithms for object recognition [6] [8]. The challenge in ARC 2017 was made more difficult by providing new set of objects (around 15 of them) only 45 minutes prior to the actual demonstration. So, the participating team were required to generate sufficient templates, annotate them and train a deep network to achieve good accuracy within this duration. It became apparent that there was need for automating the data annotation process and at the same time, reduce the training time for achieving a specified level of accuracy. While one could always use more number of GPUs to reduce training time, it is still necessary to automate the data generation process which takes considerable amount of time and effort to produce a decent size of dataset necessary for training.

This paper primarily targets the two aforesaid points and proposes an automatic object annotation and a multi-class object segmentation approach based on the concept of Feature Pyramid Network (FPN) [10] integrated in a deep framework. The proposed FPN based deep network generates a comprehensive feature vector, selectively collected from successive convolutional layers that preserve features from primitive shape information, such as lines, corners to much higher and better represented features. Residual Network (ResNet) [9] has been used as a base network in our deep framework as it has been experimentally proven to be outperforming networks, like VGG-16, both in-terms of computational complexity as well as recognition accuracy. The proficiency of the proposed multi-class semantic segmentation framework has also been compared with the current state of the art techniques, such as Pyramid Scene Parsing Network (PSPNet) [23] and Mask R-CNN [5]. Unlike, the proposed approach, PSPNet uses Spatial Pyramid Pooling (SPP) [3] module to learn surrounding contextual information for resolving ambiguities in class labels. For instance, while it learns to segment a 'river' from a 'house', it also learns that a house surrounded by a river could be a 'boat house'. However, such contextual correlations may not be useful in our application where the objects belong to 'household retail objects' and are to be segmented based on their intra-class variability. In-fact this dependency on contextual information may often mislead the classifier in warehouse scenarios. Moreover, incorporation of the SPP module makes the PSPNet computationally more intensive as compared to our approach. Mask R-CNN, which is considered as the current state of the art technique, also has certain important limitations. It uses regional proposal to estimate anchors (computationally very expensive) and detects the bounding box using a regression loss that always take 5-10 epochs to learn the network. Whereas, the proposed approach does semantic segmentation using the softmax loss, which is much faster.

The proposed framework works in three steps. First, the deep network is trained on a small set of manually annotated dataset to act as a binary classifier, that can segment foreground objects from its background. This binary classifier is then used to automatically generate labeled template for any object placed on a trained background. Second step involves generating synthetic clutters by superimposing individual templates. Third step involves training the proposed network using this machine generated cluttered object dataset for multi-class segmentation. It has to be noted that, the multi-class segmentation module is not making use of any trained models from the binary classifier, except the automatically generated dataset. In short, the main contributions made in this paper are as follows: (1) a deep network architecture is proposed for automatic generation of annotated dataset. (2) A framework using the above deep network is proposed for recognizing and segmenting retail household objects in a dense clutter. This network architecture provides better accuracy

compared to the PSPNet and the Mask R-CNN with lesser training time. (3) A new dataset for objects specified by ARC 2017 competition is created using this approach and will be made available online for community use. (4) A rotating platform with multiple cameras is designed to automatically collect a large amount of data in a very short time. The whole process of data collection and network training is performed within the time constraint of 45 minutes in ARC 2017 event.

The rest of this paper is organized as follows. A brief overview of related work is provided in the next section. The proposed deep network architecture and the method for generating automatic annotations is explained in Section 3. The details of experiment and analysis of results is presented in Section 4 followed by conclusion and direction for future work in Section 5.

2 Related Work

Numerous research work have been done in the field of robotic perception. However, covering the entire literature in this domain is beyond the scope of this paper. Thus, we have concentrated only on those research works which are directly related to the object segmentation and recognition techniques needed for warehouse automation. The area of automatic annotation came into limelight in the recent past, only after continuous progress of deep learning based object recognition approaches. Few research works in this direction include [7] [21] [14] and [9]. Grossmann et al. [7] first presented a deep neural network based foreground and background segmentation method using Pascal VOC database. In a recent paper Milan et al. [14] used RefineNet [9], a deep convolutional neural network specifically designed for autonomous picking which can be fine-tuned with a limited training image set. However, the segmentation performance of that approach was not very precise and also required human intervention for correction of wrongly segmented data. In another work Zeng et al. [21] estimated object pose using RGBD data. They have segmented and labeled multiple views of objects with a fully convolutional neural network. 6D object pose was obtained by fitting pre-scanned 3D object models to the resulting segmentation. One important limitation of this approach was that, it must have pre-scanned 3D model for each of the objects. The approach is hence not suitable to use for new set of objects where shape may be entirely different. There are few other conventional methods, such as FCN [12] which can be directly used for semantic segmentation. DeepLab [16] uses a combination of the few strongly labeled and many weakly labeled images to semantically segment objects. They have also used dilated convolutions to prevent excessive downscaling of the input images.

3 Proposed Method

We propose a new framework that enables us to generate a semantically stronger comprehensive feature vector by selectively collecting features from different convolutional layers while preserving primitive information, such as lines and corners. FPN as shown in Figure.1 is generated by consolidating features from multiple CNN layers which undergoes interpolation in order to generate equivalent dimensional feature vectors followed by concatenation at three different points. The architecture of the proposed deep network is explained in Section 3.1 and detailed description of the end-to-end multi-class segmentation approach is given in Section 3.2.

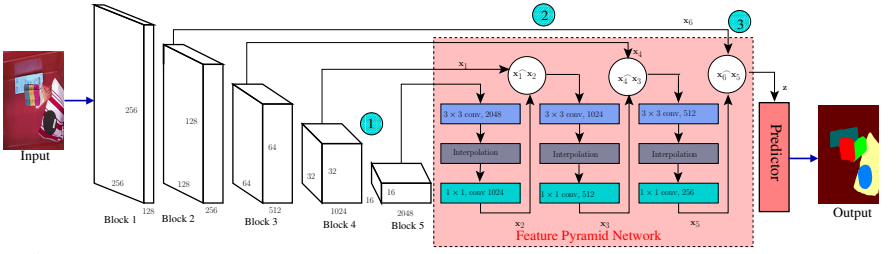


Figure 1: Overall architectural diagram of the proposed network. Features from different layers are convolved followed by interpolation to generate equivalent dimensional feature vector and then concatenated in three different steps. The final concatenated feature vector $\mathbf{z}$ is then used for FCN based semantic segmentation. ResNet is used as a base network. In contrast to PSPNet, (1) No dilation is used in the convolutional layers of the last two blocks; (2) FPN is used to exploit multi-level features for high-resolution prediction (3) Each feature space is smoothened before concatenation.

3.1 Network Architecture

The presented pyramidal feature generation network is motivated by the concept of FPN proposed in [10]. In our proposed object annotation framework, ResNet consisting of 50 and 101 convolutional layers [7] are used as building blocks. As, it has been shown in the Figure 1 that, a set of residual blocks (block 2, block 3, block 4, and block 5) except the first one is used to make the final prediction. Outcome of the last layer at each of these blocks has been considered as the reference set of feature maps, since the deepest layer of each block contains the strongest features [10]. This approach thus ensures the inclusion of different levels of features, starting from primitive features to much stronger and better representational features. First block has been omitted from the pyramid just to avoid an increased memory requirement. Moreover, it is expected that, the output features at first block are not significant enough to represent meaningful patterns. The FPN is shown as component 2 in the Figure 1. The chosen set of blocks have strides of (4, 8, 16, 32) in case of ResNet-50 and for ResNet-101 it is (4, 8, 8, 8) pixels with respect to the input image. At the region 3 as shown in the Figure 1, the derived feature vectors gets concatenated to retain the aforesaid features. Just before concatenation, we have used an interpolation layer to ensure that same dimensional feature vector is coming from each block.

3.2 End-to-End approach for Semantic Segmentation

3.2.1 Image Acquisition

To facilitate an automatic and speedy image acquisition technique, and also to avoid any possibility of redundant data in our dataset, we have designed and developed an automatic image acquisition rig as shown in the Figure 2. It is an in-house production and has 5 Foscams of 519 series mounted to it. It can capture 400 images of an object (placed on a rotating platform), at various scale and orientation in a single rotation of the rotating platform. For this presented work, we have obtained 300 images per object with varying orientations, scales and positions. The resolution of the images captured are 1920×1080. Later, these images are cropped to the size of 512×512 before passing to the annotation network for training. All these images were captured in a uniform (red) back-

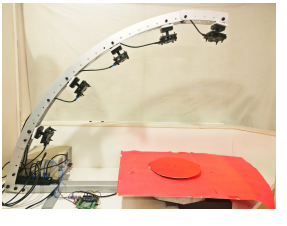


Figure 2: In-house developed rig to capture multiple images at a time with varying orientations and scale. It helps in speedy image acquisition and ensures non-redundancy in generated data.

ground so as to promote binary classification to generate ground truth data for these captured images. Manual intervention is limited to changing the pose of object 2-3 times for capturing this dataset. Once the images are captured, 200 random single-object images are annotated manually and passed to the proposed annotation network for training our pixel-wise binary classifier.

3.2.2 Ground Truth Generation

As noted previously, in this work we aim to develop an automatic object annotation module that needs to be highly accurate in order to be able to accurately classify and segment any object in a cluttered environment. Starting with only few manually labeled training images, we train the proposed deep network as a binary classifier to generate ground truth masks of any new object placed on a familiar background (red in this case). The generated ground truth masks are further refined by integrating the proposed annotation framework with Single Shot Multibox Detection (SSD) [11] network, to makes sure that, no pixel outside the detected object boundary gets classified as an object. This module thus act as a refinement network. SSD is chosen over other bounding box detectors, such as Faster RCNN as it is computationally much faster and yet achieves similar detection accuracy due to elimination of both bounding box proposals and the subsequent pixel or feature re-sampling stage [11]. This trained binary segmentation model can then be used to automatically generate ground truth masks for any new object on the given background. These generated masks are later used as ground truth for multi-class segmentation.

3.2.3 Artificial Clutter Generation

It is expected in the warehouses that the objects are to be picked from a cluttered tote/box. We have thus synthetically generated different degree of clutters and their corresponding masks by using automatically generated masks of single object images. Few such resultant clutters are shown in the Figure 3. These artificial clutters are obtained by applying data augmentation techniques, such as rotation, translation and scaling. Multiple degrees of clutter are generated by using different permutation and combinations of items along with varying pose and orientation.

This technique results in generating the final dataset with 18000 cluttered images. Among these 12000 images are used to fine-tune the proposed multi-class segmentation network and the remaining 6000 are used for validating the performance of the proposed approach.

Figure 3: Few examples of synthetically generated cluttered environment is shown here.

3.2.4 Multi-class Segmentation

A flow diagram of the proposed approach is shown in the Figure 4. In this flow diagram, we have shown that, once we synthetically generate sufficiently large amount of artificial clutters, we train the proposed network to obtain a multi-class object spatial classification model. Unlike, the state-of-the-art technique PSPNet [23], this approach ensures accurate segmentation of the objects.The promising aspect of PSPNet was the use of SPP module, whereas, we opted to use FPN. This has two fold advantage over SPP: firstly, the multi-scale feature map prediction is almost twice as fast as

SPP and secondly, SPP module [3] of PSPNet increases accuracy by bringing in contextual information to differentiate between confusing categories. Since the objects in the bin or the stack are highly uncorrelated, surrounding context does not add much information. In fact, inclusion of SPP may sometimes mislead the classifier. To further increase the speed of our network, unlike PSPNet, dilation [20] (which is computationally very intense) is not used in any of the convolutional layers. Dilation simply enables a network to have higher receptive field and it is not a requirement in the current scenario, as the distance from object to camera location is not varying. The proposed framework, specifically designed for objects in warehouses is hence an improvement over the state of the art technique, PSPNet, both in terms of faster computation and segmentation accuracy. Even when comparing with the current state of the art technique: Mask R-CNN, the proposed annotation module is computationally more efficient and performs better in the given warehouse scenario. Experimental results and the performance comparisons with PSPNet and Mask R-CNN have been presented in Section 4.

4 Experiment

The proficiency of the proposed annotation framework and the multi-class object segmentation module is validated by performing different experiments on both synthetically generated cluttered objects as well as on real-world clutters. We have trained multi-class object segmentation model using the automatically generated annotations and cluttered images. Also, we have statistically compared our network with two state-of-the art techniques, PSPNet and Mask R-CNN, through rigorous experimentations. In this section we have demonstrated some of those experimental results.

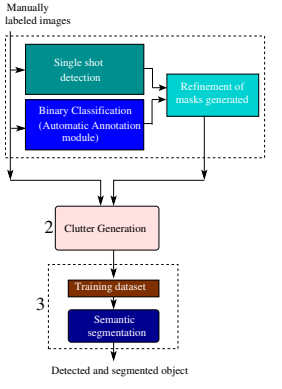


Figure 4: Flow diagram of the approach. (1) Semi-supervised technique to automatically annotate training data set with SSD running in parallel to refine the generated masks; (2) Automatic data generation of cluttered environment; (3)Multi-class image segmentation trained on data generated in (1) and (2).

4.1 Loss Function

The proposed architecture uses softmax loss at its main branch. Apart from this loss, an auxiliary loss is also applied at the 4^{th} block of the architecture to train the final binary classifier. Like PSPNet [23], we let the auxiliary loss to back-propagate all previous layers. This deeply supervised learning strategy for ResNet-based network optimizes the learning process. The weightage for auxiliary loss is set to 0.4. However, the auxiliary branch is not used in the testing phase.

4.2 Automatic Annotation and Multi-Class Object Segmentation Results

Figure 6 shows some test results of the automatic annotation model. It can be observed from the illustrated examples, that even transparent objects (like plastic bottle) and objects with colors close to the background color (like barbie book) get segmented very precisely. To further validate the accuracy of the proposed automatic annotation model, we train our multi-class object segmentation framework using only the automatically generated cluttered images along-with the corresponding automatically generated annotations. Training image

Figure 5: Multi-class semantic segmentation results of few images with different degree of clutters are shown here. The clutters are generated synthetically by using automatically generated mask of individual objects. Comparison have been made with the state of the art techniques; Mask R-CNN and PSPNet.

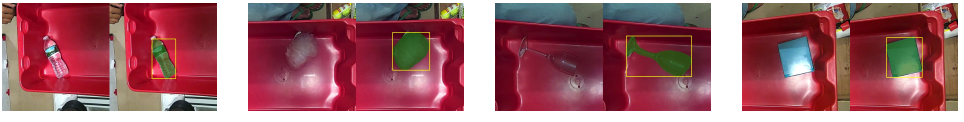


Figure 6: Figure illustrates few examples of binary mask generated using the proposed network. It can be observed that even for transparent objects the network can generate precise segmented region.

set containing 12000 cluttered images are used for this purpose. A set of 40 different objects given by ARC-2017 has been used through out these experiments. The hardware configurations and various other experimental details are jotted down in the Table 1. Semantic segmentation results of few synthetically generated cluttered images with increasing degree of clutters are shown in the Figure 5. The resultant segmented images are obtained for three different networks: Mask R-CNN, PSPNet and the Proposed net. To verify the effectiveness of the proposed approach we have also tested it with real cluttered object images. Some of such cluttered images and their segmentation results using all the three model are shown in the Figure 9. It can be observed from both the above two Figures that, even objects with approximately 50% occlusion also gets segmented accurately using the proposed model. Statistical analysis in terms of precision, recall and f-measure, obtained for all the 40 different classes of objects when tested using all the three trained models (Mask R-CNN, PSPNet and the Proposed Net) are shown in the Figure 7. Table[3] compares the performance of our architecture with PSPNet and Mask R-CNN in terms of overall recognition accuracy, time

Table 1: Training and testing setup. Training is done using GPU-machine: NVIDIA Quadro P6000 with two 24GB GPUs. The parameters setting and the training-testing details are given here.

Method	Training Setup	Training time (s)	# Training images	Batch Size	# Test images	Image size
Proposed	Quadro	500		10		512×512
Mask R-CNN	P-6000	1800	12×10^3	8	6×10^3	600×600
PSPNet		1400		5		473×473

Table 2: Performance comparison of the proposed network with the state of the art techniques: Mask R-CNN and PSPNet. The measurement criterion are Mean- Intersection Over Union (mIOU), average pixel accuracy, Precision, Recall and F-measure for 40 different classes of objects.

Base Network	Method	mIOU %	Pixel Accuracy %	Precision %	Recall %	F1 Score %
ResNet 50	Mask R-CNN	69.95	90.34	66.87	56.45	59.94
	PSPNet	81.65	95.64	91.06	92.46	91.26
	Proposed Net	87.25	96.76	93.08	93.08	93.08
ResNet 101	Mask R-CNN	43.53	70.23	55.38	48.54	49.70
	PSPNet	80.71	95.08	89.74	88.21	88.87
	Proposed Net	86.58	96.54	92.05	90.27	91.05

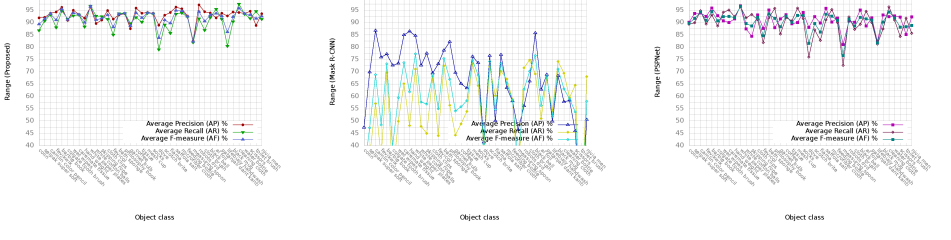
Table 3: Performance comparison of the proposed network with the state of the art techniques: Mask R-CNN and PSPNet when trained using RseNet-50 as base network. The comparison is made in terms of Average pixel accuracy, training time of each image during forward pass, backward pass and combining forward as well as backward pass.

Performance measure	Mask R-CNN	PSPNet	Proposed net
Average pixel accuracy	90.34%	95.64%	96.76%
Forward pass	115.20 ms	108.066 ms	63.622 ms
Backward pass	192.8 ms	172.048 ms	82.650 ms
Forward-Backward pass	308 ms	280.460 ms	146.598 ms

taken during forward pass and time for backward pass. Forward-backward pass of the proposed network is comparatively much faster (approximately 1.91 times faster than PSPNet and approximately 2.1 times faster than Mask R-CNN) when trained using ResNet-50 as base network.

The error convergence plots for all the three approaches, when trained for multi-class object segmentation with base networks ResNet-50 and ResNet-101, are shown in Figure 8. It is clearly observed that, the learning of the proposed model is quite better than other two networks in terms of faster learning rate as well as lower error. Even the overall recognition accuracy of the proposed net is better in-comparison to other two approaches. An overall recognition accuracy of 96.76% is achieved using proposed model which is marginally higher than PSPNet with 95.64% and significantly much higher than Mask R-CNN that has an overall accuracy of 90.34% A comparative analysis in terms of mean precision, recall and f-measure obtained for 40 class spatial classification is shown in the Table 2. The analyses are done for both ResNet-50 as well as ResNet-101. In both the cases, the proposed approach significantly outperforms other two approaches. We have further performed extensive experiments on a MIT-Princeton Database [22] (created using same set of 40 objects given in ARC

Figure 7: Statistical comparison of the proposed multi-class segmentation model with PSPNet and Mask R-CNN. Plots for precision, recall and f-measure obtained for 40 different class of objects are demonstrated here. ResNet-50 is used as a base network. Same set of test and training data have been used for all the three networks. The observations clearly demonstrates the proficiency of the proposed multi-class object segmentation approach.

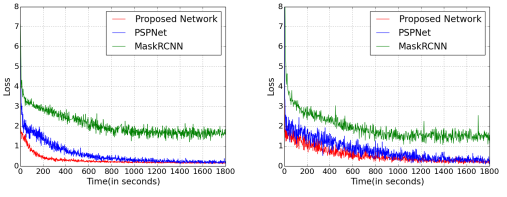


Cluttered image | Mask R-CNN | PSPNet | Proposed

Figure 9: Segmentation comparison for real cluttered images using Mask R-CNN, PSPNet and the Proposed architecture.

2017). MIT-Princeton database has few images per class (around 16-24 images for an object). However, In case of our database, we have used 12000 cluttered images for training the deep networks remaining 6000 images are used for testing. In order to maintain a consistency and to have a better training, we have applied augmentation and generated same number of cluttered images using the given images in MIT-Princeton database. Distribution of training and testing data are also done accordingly. The mean average precision, recall and f-measure obtained for the proposed approach when trained and tested for 40 different object classes are 88.84%, 88.53% and 88.54% respectively. Whereas, the corresponding results for the Mask R-CNN and the PSPNet are 78.88%, 69.17%, 73.01% and 86.52%, 84.83%, 85.58%.

Figure 8: Plots show loss functions for the proposed multi-class object segmentation model, PSPNet and Mask R-CNN when trained using ResNet-50 and ResNet-101.



5 Conclusion

We have presented an end to end semi-supervised methodology for automatic annotation and multi-class segmentation of object images in an warehouse scenario. Our proposed method has three fold utility. Firstly it automates the process of ground truth generation, giving near-perfect binary masks. Secondly, human efforts are reduced not only in the task of ground truth generation but also by eliminating the need of data collection in a cluttered environment, ensuring non-redundancy in generated data. The results show that the network trained on artificially created clutters works more efficiently (both in terms of segmentation accuracy as well as computation speed) as compared to the state of the art techniques. The test datasets contain both synthetically generated clutters as well as real clutters. We have also performed the experiments on MIT-Princeton database and compared the results with state of the art techniques to validate the proficiency of the proposed approach. As a future work, the proposed work can be transformed into a completely unsupervised technique by using weakly labeled dataset with cascaded classifiers at the initial stage instead of using the manually annotated dataset for binary classification.

References

[1] Clemens Eppner, Sebastian Höfer, Rico Jonschkowski, Roberto Martin Martin, Arne Sieverling, Vincent Wall, and Oliver Brock. Lessons from the amazon picking challenge: Four aspects of building robotic systems. In *Robotics: Science and Systems*, 2016.

[2] Etienne Grossmann, Amit Kale, and Christopher Jaynes. Towards interactive generation of" ground-truth" in background subtraction from partially labeled examples. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005.*, pages 325–332. IEEE, 2005.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, pages 346–361. Springer, 2014.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *arXiv preprint arXiv:1703.06870*, 2017.

[6] Carlos Hernandez, Mukunda Bharatheesha, Wilson Ko, Hans Gaiser, Jethro Tan, Kanter van Deurzen, Maarten de Vries, Bas Van Mil, Jeff van Egmond, Ruben Burger, et al. Team delft's robot winner of the amazon picking challenge 2016. In *Robot World Cup*, pages 613–624. Springer, 2016.

[7] Rico Jonschkowski, Clemens Eppner, Sebastian Höfer, Roberto Martín-Martín, and Oliver Brock. Probabilistic multi-class segmentation for the amazon picking challenge. In *International Conference on Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ*, pages 1–7. IEEE, 2016.

[8] Swagat Kumar, Anima Majumder, Samrat Dutta, Rekha Raja, Sharath Jotawar, Ashish Kumar, Manish Soni, Venkat Raju, Olyvia Kundu, Ehtesham Hassan Laxmidhar Behera, et al. Design and development of an automated robotic pick & stow system for an e-commerce warehouse. *arXiv preprint arXiv:1703.02340*, 2017.

[9] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. *arXiv preprint arXiv:1612.03144*, 2016.

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[12] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[13] Kanika Mahajan, Anima Majumder, Harika Nanduri, and Swagat Kumar. Tcs arc 2017 warehouse object dataset. https://doi.org/10.6084/m9.figshare.6848738.v1, July 2018.

[14] A Milan, T Pham, K Vijay, D Morrison, AW Tow, L Liu, J Erskine, R Grinover, A Gurman, T Hunn, et al. Semantic segmentation from limited training data. *arXiv preprint arXiv:1709.07665*, 2017.

[15] D Morrison, AW Tow, M McTaggart, R Smith, N Kelly-Boxall, S Wade-McCue, J Erskine, R Grinover, A Gurman, T Hunn, et al. Cartman: The low-cost cartesian manipulator that won the amazon robotics challenge. *arXiv preprint arXiv:1709.06283*, 2017.

[16] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015.

[17] Songtao Wu, Shenghua Zhong, and Yan Liu. Deep residual learning for image steganalysis. *Multimedia Tools and Applications*, pages 1–17, 2017.

[18] Peter R Wurman and Joseph M Romano. The amazon picking challenge 2015. *IEEE Robotics and Automation Magazine*, 22(3):10–12, 2015.

[19] Peter R Wurman, Raffaello D'Andrea, and Mick Mountz. Coordinating hundreds of cooperative, autonomous vehicles in warehouses. *AI magazine*, 29(1):9, 2008.

[20] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. *arXiv preprint arXiv:1705.09914*, 2017.

[21] Andy Zeng, Kuan-Ting Yu, Shuran Song, Daniel Suo, Ed Walker, Alberto Rodriguez, and Jianxiong Xiao. Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge. In *IEEE International Conference on Robotics and Automation (ICRA), 2017*, pages 1386–1383. IEEE, 2017.

[22] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois Robert Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, Nima Fazeli, Ferran Alet, Nikhil Chavan Dafle, Rachel Holladay, Isabella Morona, Prem Qu Nair, Druck Green, Ian Taylor, Weber Liu, Thomas Funkhouser, and Alberto Rodriguez. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018.

[23] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *arXiv preprint arXiv:1612.01105*, 2016.