Uncertainty-aware Instance Segmentation using Dropout Sampling

Doug Morrison*
ACRV, QUT
Brisbane, Australia
doug.morrison@roboticvision.org

Anton Milan, Epameinondas Antonakos
Amazon
Berlin, Germany
{antmila, antonak}@amazon.de

Abstract

Vision is an integral part of many robotic systems, and especially so when a robot must interact with its environment. In such cases, decisions made based on erroneous visual detections can have disastrous consequences. Hence, being able to accurately measure the uncertainty associated with visual information is highly important for making informed decisions. However, this uncertainty is often not captured by classic computer vision systems or metrics. In this paper we address the task of instance segmentation in a robotics context, where we are concerned with uncertainty associated with not only the class of an object (semantic uncertainty) but also its location (spatial uncertainty). We apply dropout sampling to the state-of-the-art instance segmentation network Mask-RCNN to provide estimates of both semantic uncertainty and spatial uncertainty. We show that a metric which combines both uncertainty measures provides an estimate of uncertainty which improves over either one individually. Additionally, we apply our technique to the ACRV Probabilistic Object Detection dataset where it achieves a score of 14.65.

1. Introduction

When interacting with objects in their environments, robots tend to rely heavily on vision systems. While the specific format may be task dependent – e.g. semantic segmentation for robotic bin picking [15, 19], object detection for self-driving cars [10] or affordance prediction for grasping [18, 20] – in all cases acting on incorrect visual information will inevitably result in a failure. Hence, in robotic applications, providing a well calibrated measure of uncertainty along with visual information is critical.

In many cases, robotic application make use of off-the-shelf computer vision algorithms, such as RefineNet for instance segmentation [11, 15, 19], SSD [13, 16] or YOLO [9, 21] for object detection or Mask-RCNN for instance segmentation [8]. However, these deep neural networks do not give well calibrated estimates of uncertainty in their class predictions and do not estimate spatial uncertainty at all [6, 7].

Furthermore, while large, public computer vision datasets and their associated evaluation metrics [1, 12, 22] have been crucial in driving computer vision research, existing performance metrics such as mean average precision (mAP) don't account for uncertainty in the predictions, relying on hard thresholds instead. An exception to this is the newly proposed probability-based detection quality (PDQ) measure and associated ACRV Probabilistic Object Detection dataset [7], which considers both semantic and spatial uncertainty as part of the calculated score.

In this paper we propose a method for probabilistic inference for instance segmentation. To achieve this, we apply dropout sampling [3, 16] to the state-of-the-art instance segmentation network Mask-RCNN [8] to provide estimates of both semantic and spatial uncertainty. While both measures can be used independently as measures of uncertainty, we propose a hybrid uncertainty metric which combines both into a further-improved estimate of uncertainty on a segment-wise basis. To demonstrate the effectiveness of our approach, we apply our method to the ACRV Probabilistic Object Detection dataset where it achieves a PDQ score of 14.65.

2. Related Work

2.1. Uncertainty Estimation using Dropout Sampling

Classically, many machine learning models lend themselves to fairly simple methods for uncertainty estimation [23]. However, this is not necessarily true for deep CNNs, due to their lack of ability to accurately predict their uncertainty [2, 3, 6, 25]. In particular, deep neural networks which tend to be overconfident in their predictions, meaning that the class score may not be a good proxy for model uncertainty. This phenomenon is also demonstrated in the top row of Fig. 1. For individual detections, the network is always

*Work performed while interning at Amazon Berlin

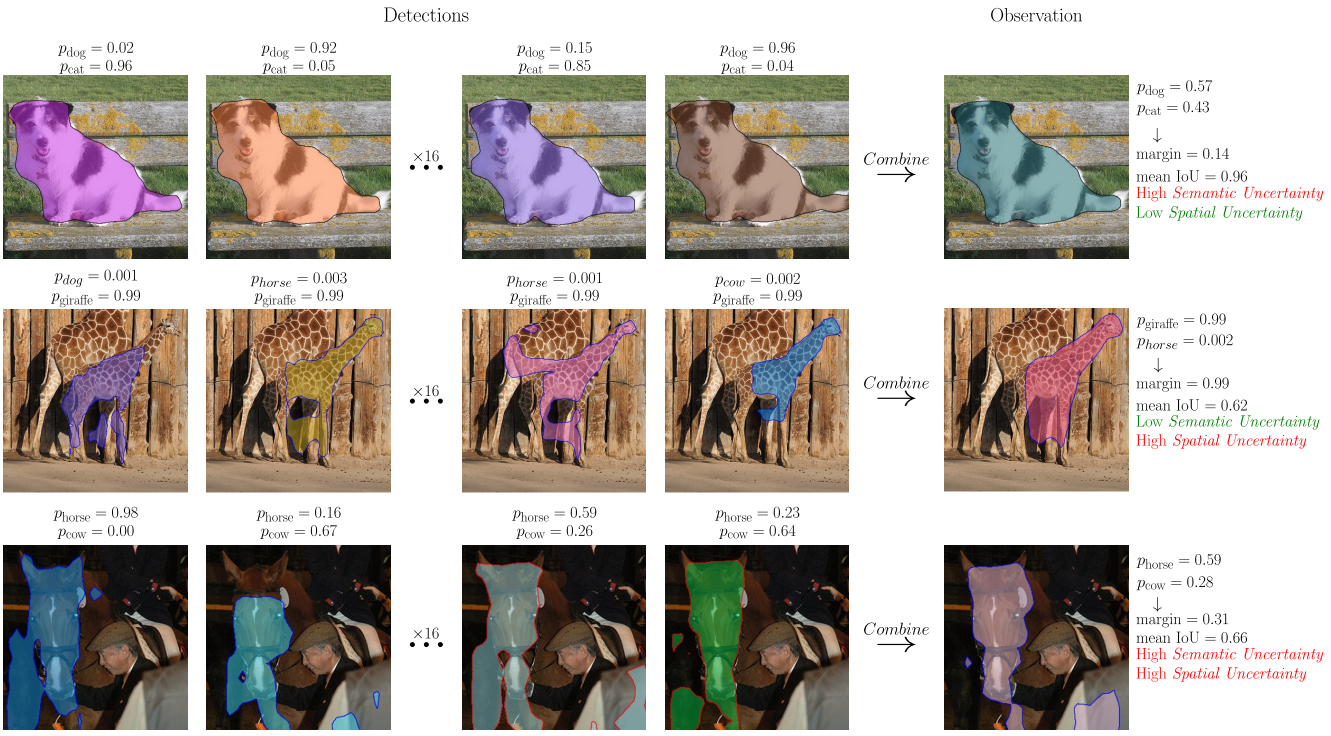

Figure 1. Detections from multiple forward passes of Mask-RCNN are combined based on their spatial affinity (IoU), resulting in a single observation which is the mean of all detections. **Top**: The segment is inconsistently classified as *dog* or *cat*, resulting in high *semantic uncertainty*, but the segment mask is consistent between detections resulting in low *spatial uncertainty*. **Middle**: The segment is always correctly classified correctly with high confidence, resulting in low *semantic uncertainty*, however, the inconsistent segmentation mask results in high *spatial uncertainty*. Such a case would not be captured by an uncertainty metric that relies on only *semantic uncertainty*. **Bottom**: The segment has both high *semantic uncertainty* and high *spatial uncertainty*.

certain that the class is either one of *dog* or *cat*, however the uncertainty in the detection is only captured after averaging the scores from multiple forward passes.

Bayesian Neural Networks [14] provide one method to predict output uncertainty, however come with prohibitively high computational overhead [3, 16]. Gal and Ghahramani [3] propose to use dropout in the final layers of their deep learning model at test time over multiple forward passes to approximate Bayesian inference over the parameters of the neural network, showing that this more effectively captures uncertainty about a given input within the model. Using this *Dropout Sampling* technique, Gal et al. [4] were able to perform Active Learning for image classification using deep convolution neural networks (CNNs). This approach has also been applied to the tasks of melanoma detection [5] and object detection using LiDAR data [2].

2.2. Dropout Sampling for Object Detection

While the above method is only applicable to classification tasks using deep neural networks, Miller et al. [16] extend the idea of Dropout Sampling to the SSD object detection network [13] – a more challenging task since each forward pass results in multiple object detections which must then be matched and combined. By clustering the object detections from multiple forward passes based on spatial and semantic similarity, they showed that object detection performance could be improved in both closed- and open-set conditions by rejecting detections based on both spatial and semantic uncertainty between sets of grouped detections. Using the same techniques, Miller et al. [17] evaluate different strategies for merging detections when using Dropout Sampling in an object detection scenario. In this work we adapt the method of [16] for an instance segmentation task, where objects are detected using a pixelwise mask rather than just a bounding box.

3. Method

In order to perform probabilistic instance segmentation, we adapt the state-of-the-art instance segmentation network Mask-RCNN [8] to use Dropout Sampling at inference time, similar to [3, 16]. Note, however, that our methodology is general and can easily be applied to any instance segmentation network. By combining segments from multiple forward passes, similar to [16, 17], we are able to provide uncertainty estimates that incorporate both semantic and spatial uncertainty.

3.1. Dropout Sampling for Instance Segmentation

To capture both semantic and spatial uncertainty using Mask-RCNN, we adapt prior work on sampling-based uncertainty techniques for object detection [16] to the task of instance segmentation using Mask-RCNN. To achieve this, we apply dropout to the fully-connected layers of Mask-RCNN, which are responsible for providing class scores and bounding box locations for each detection in the image. The locations of masks are then dependent on the set of highest-scoring bounding boxes, so are also similarly effected by this procedure.

Following the notation of [16, 17], each forward pass of Mask-RCNN on an image provides a set of instance detections $S = \{D_1, ..., D_k\}$. In our case, each detection $D_i = \{\mathbf{s}, \mathbf{b}, \mathbf{m}\}$, comprises a distribution of softmax scores for each class $\mathbf{s} = (s_1, ..., s_m)$, a bounding box $\mathbf{b} = (x_1, y_1, x_2, y_2)$ and a pixel-wise mask $\mathbf{m}$. By performing N forward passes of the network ($N = 16$ in our experiments), we obtain a set of samples $\mathbb{S} = \{S_1, ..., S_n\}$, each of which contains a set of detections as described above. This process is illustrated in Fig. 1.

The detections from all forward passes are grouped into a set of individual observations $\mathbb{O} = \{O_1, ..., O_j\}$ based on their spatial affinity. Ideally, each observation should represent a single object within the image. To achieve this, we use the Basic Sequential Algorithmic Scheme (BSAS) [17, 24], whereby detections are clustered together into observations in a sequential fashion if their mask intersection-over-union (IoU) is greater than some threshold θ. Specifically, if the IoU between a detection and every detection in an existing observation is greater than θ, the detection is added to that observation. If the detection matches no existing observations, a new observation is created. An individual observation $O = (\bar{\mathbf{s}}, \bar{\mathbf{b}}, \bar{\mathbf{m}})$ is parameterised by the mean softmax scores, bounding box and mask across all detections within the observation.

A distinct advantage of using pixelwise segments rather than bounding boxes as in [16] to combine detections is that two significantly overlapping segments are much less likely to represent the same object than two overlapping bounding boxes, especially in the case of many tightly grouped or irregularly shaped objects. As such, we empirically find that a lower IoU threshold of, e.g. $\theta = 0.5$ (compared to $\theta = 0.95$ [16]), can be used without incorrectly combining objects or incorrectly rejecting matching segments.

4. Results

4.1. Uncertainty Estimation

Similar to [4, 16], we can easily compute a semantic certainty from the average softmax scores $\bar{\mathbf{s}}$ of an observation O, i.e.:

$$u_{sem}(O) = \max(\bar{\mathbf{s}}) \tag{1}$$

This corresponds to examples where the class label is uncertain, such as shown in the top row of Fig. 1.

However, this metric does not capture spatial uncertainty of an observation, for example as in the middle row of Fig. 1, where an instance may be classified with high confidence but the precise location of the segment is uncertain. To overcome this, we introduce a spatial uncertainty measure for each observation by calculating the mean IoU between the observation mask $\bar{\mathbf{m}}$ and the mask $\mathbf{m}$ of every one of n detections within the observation O:

$$u_{spl}(O) = \frac{1}{|O|} \sum_{i=1}^{|O|} \text{IoU}(\mathbf{m}_i, \bar{\mathbf{m}}) \tag{2}$$

Additionally, a third measure of uncertainty that we can consider is the fraction of the forward passes in which an Observation appears, i.e. (where N is the total number of forward passes):

$$u_n(O) = \frac{|O|}{N} \tag{3}$$

In the top row of Fig. 2 compare each of these three metrics against their ability to predict a successful instance segmentation result (i.e. a true positive detection) against the ACRV Probabilistic Detection validation set. What we observe is that no one metric captures uncertainty in a well calibrated way on this dataset as they consider only one metric individually. To effectively capture both *semantic uncertainty* and *spatial uncertainty* about an observation, we propose a hybrid uncertainty metric which combines uncertainty metrics.

In the bottom row of Fig. 2 we plot the calibration for three hybrid metrics which combine those above. Firstly, we weight the semantic uncertainty by the fraction of forward passes in which the observation appears:

$$u_{sem\_w}(O) = u_{sem}(O) \cdot u_n(O), \tag{4}$$

secondly we compute a weighted mean IoU in the same way:

$$u_{spl\_w}(O) = u_{spl}(O) \cdot u_n(O), \tag{5}$$

and finally a hybrid metric which combines all three:

$$u_h(O) = u_{sem}(O) \cdot u_{spl}(O) \cdot u_n(O) \tag{6}$$

We observe a much improved calibration of predictive uncertainty with the combined metric u_h which incorporates both semantic and spatial uncertainty. Potentially, the calibration could be improved further with more forward passes of the network, however this comes with a large computational overhead.

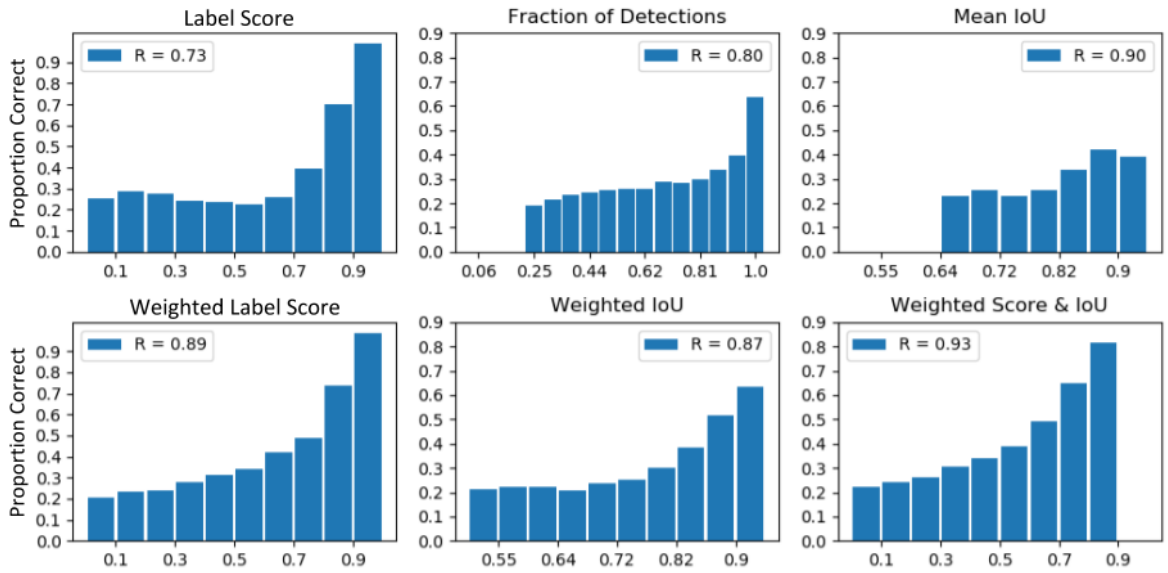

Figure 2. Calibration of various metrics for uncertainty, evaluated against the ACRV Probabilistic Detection validation set. Individual observations are grouped based on their metrics, and compared to the probability of the observation representing a true positive. Top: semantic uncertainty (u_{sem}), length of observation (u_n) and spatial uncertainty (u_{spl}). Bottom: The three hybrid uncertainty metrics u_{spl_w}, u_{sem_w} and u_h. R indicates the Pearson correlation (and hence linearity) of the binned score, indicating the quality of uncertainty calibration.

θ = 0.25	Fraction of Detections 0.25	0.5	0.75	0.9
Score 0.1	7.8	10.6	11.9	13.3
0.3	9.4	11.9	13.6	14
0.5	12.3	14.4	15.1	14.6
0.8	14.4	14.5	14.2	13.9

θ = 0.5	Fraction of Detections 0.25	0.5	0.75	0.9
Score 0.1	7.7	10.3	12.4	13.8
0.3	9.2	12.1	13.8	14.3
0.5	12.2	14.5	15.2	15
0.8	14.3	14.5	14.4	14.1

θ = 0.75	Fraction of Detections 0.25	0.5	0.75	0.9
Score 0.1	7.5	11.3	13.2	13.4
0.3	9.1	12.8	14.0	13.5
0.5	12.0	14.7	14.3	13.3
0.8	13.9	14	13.2	12.1

Figure 3. PDQ score for different thresholds of score (u_{sem}) and fraction of detections (u_n) on the ACRV Probabilistic Object Detection validation set.

4.2. Probabilistic Object Detection

To evaluate our method, we apply it to the ACRV Probabilistic Object Detection dataset, a simulated dataset which utilises a subset of classes from the COCO dataset. This dataset uses the pairwise Probability-based Detection Quality (pPDQ) metric [7] for scoring, which explicitly takes into account spatial and semantic probabilities to compute scores. For our entry we use a Mask-RCNN network which has been trained on COCO data only.

Because the dataset evaluation is based on object detection rather than segmentation, we generate bounding boxes that tightly enclose the masks of each detection. We find that this method provides more accurate localisation than the raw Mask-RCNN bounding box predictions (average *spatial quality* of 0.48 versus 0.37 for the same detections on the ACRV Probabilistic Object Detection dataset). To generate the required probabilistic bounding box representation, we compute the mean bounding box and bounding box covariance for each observation.

Unfortunately, the final score, the PDQ, is still weighted by the number of false positive detections. As a result, including all all uncertain detections, even if well calibrated, will result in a low PDQ score. Hence, we apply two methods to find a threshold for including individual observations that maximises the trade-off between true positives (which generate a pPDQ > 0) and false positives.

First, using the validation dataset, we perform a simple grid search across individual uncertainty metrics, as shown in Fig. 3. The highest performing combination ($\theta = 0.5$, $u_{sem}(O) \geq 0.5$, $u_n(O) \geq 0.75$) achieves a score of 15.2 on the validation set and 13.4 on the full dataset. Below these thresholds the number of false positives increases significantly, and at higher thresholds the number of true positives is significantly decreased, both of which result in lower over-

all scores. Interestingly, any further thresholding based on $u_{sem}(O)$ decreases the overall score.

Secondly, we perform a similar threshold search on the hybrid uncertainty metric u_h. We find that a threshold of $u_h(O) \geq 0.4$ gives the best score of 16.2 on the validation set (Fig. 4) and 14.65 on the full dataset. The hybrid metric, which includes both semantic and spatial uncertainty gives us a more accurate way to discern between potential true positives and likely false negatives, resulting in a higher overall PDQ score.

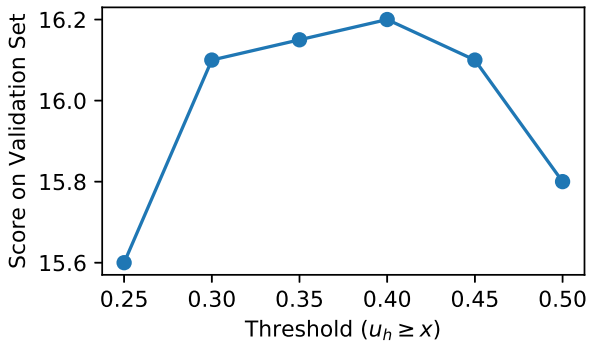

Figure 4. PDQ score for different thresholds on the hybrid detection uncertainty (u_h) on the ACRV Probabilistic Object Detection validation set.

5. Conclusion

We've presented an approach to probabilistic instance segmentation by applying dropout sampling to Mask-RCNN. Our approach is able to give well calibrated hybrid estimate of semantic and spatial certainty, outperforming individual measures based on either one, achieving a PDQ score of 14.65 on the ACRV Probabilistic Object Detection dataset.

References

[1] 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*, 88(2):303–338, 2010.

[2] Di Feng, Xiao Wei, Lars Rosenbaum, Atsuto Maki, and Klaus Dietmayer. Deep Active Learning for Efficient Training of a LiDAR 3D Object Detector. *arXiv preprint arXiv:1901.10609*, 2019.

[3] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[4] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org, 2017.

[5] Marc Górriz, X. Giró-i Nieto, Axel Carlier, and Emmanuel Faure. Cost-effective active learning for melanoma segmentation. In *ML4H: Machine Learning for Health Workshop at NIPS 2017*, 2017.

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330, 2017.

[7] David Hall, Feras Dayoub, John Skinner, Peter Corke, Gustavo Carneiro, and Niko Sünderhauf. Probability-based detection quality (pdq): A probabilistic approach to detection evaluation. *arXiv preprint arXiv:1811.10800*, 2018.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *In proceedings of the IEEE conference on Computer Vision*. IEEE, 2017.

[9] Robert Lee, Serena Mou, Vibhavari Dasagi, Jake Bruce, Jürgen Leitner, and Niko Sünderhauf. Zero-shot sim-to-real transfer with modular priors. *arXiv preprint arXiv:1809.07480*, 2018.

[10] Baojun Li, Shun Liu, Weichao Xu, and Wei Qiu. Real-time object detection and semantic segmentation for autonomous driving. In *MIPPR 2017: Automatic Target Recognition and Navigation*, volume 10608, page 106080P. International Society for Optics and Photonics, 2018.

[11] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[13] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. pages 21–37, 2016.

[14] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[15] Anton Milan, Trung Pham, K Vijay, Douglas Morrison, Adam W Tow, L Liu, J Erskine, R Grinover, A Gurman, T Hunn, et al. Semantic segmentation from limited training data. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1908–1915. IEEE, 2018.

[16] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018.

[17] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[18] Douglas Morrison, Peter Corke, and Jürgen Leitner. Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. In *Robotics: Science and Systems*, 2018.

[19] Douglas Morrison, Adam W 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. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7757–7764. IEEE, 2018.

[20] Douglas Morrison, Peter Corke, and Jürgen Leitner. Multi-view picking: Next-best-view reaching for improved grasping in clutter. In *International Conference on Robotics and Automation (ICRA)*, 2019.

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[23] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[24] Sergios Theodoridis and Konstantinos Koutroumbas. Pattern Recognition 4th Edition. 2008.

[25] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011.