Multi-view self-supervised deep learning framework for solving 6D pose estimation problem

Presenter:Ren Shize

Li Zhidong

Jiang Meng

Zeng Yuqi

Zhao Xuda

2023.3.2

仿生设计与学习实验室

Bionic Design & Learning Lab

main DL

In Collaboration with MIT

AncoraSIR.com

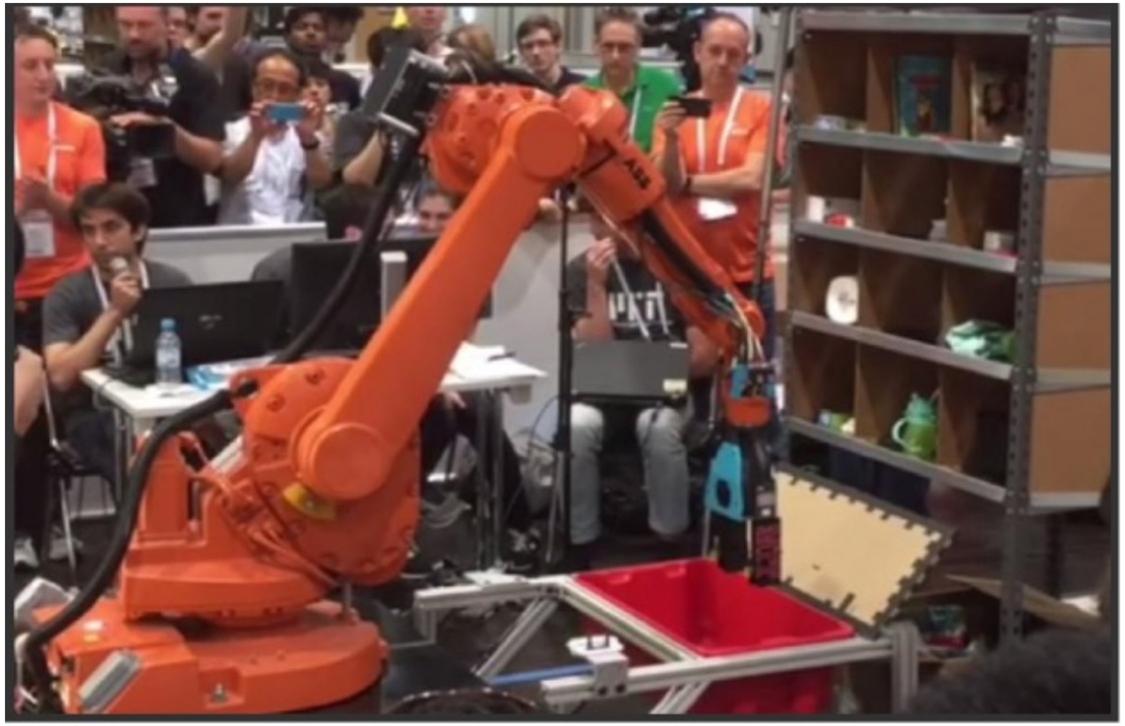
SUSTech

Southern University of Science and Technology

Motivation and Main Problem

The description of problem being solved

Robot warehouse automation has attracted significant interest in recent years, perhaps most visibly in the Amazon Picking Challenge (APC）.Amazon, in collaboration with the academic community, has led a recent effort to define two such tasks: 1) picking an instance of a given a product ID out of a populated shelf and place it into a tote; and 2) stowing a tote full of products into a populated shelf.



AncoraSIR.com

Motivation and Main Problem

Why is the problem important?

- A fully autonomous warehouse pick-and-place system requires robust vision that reliably recognizes and locates objects amid cluttered environments, self-occlusions, sensor noise, and a large variety of objects. That's a huge improvement to general-purpose robot autonomy
- Warehouse automation technologies, satisfying the growing demand of e-commerce and providing faster, cheaper delivery.

Motivation and Main Problem

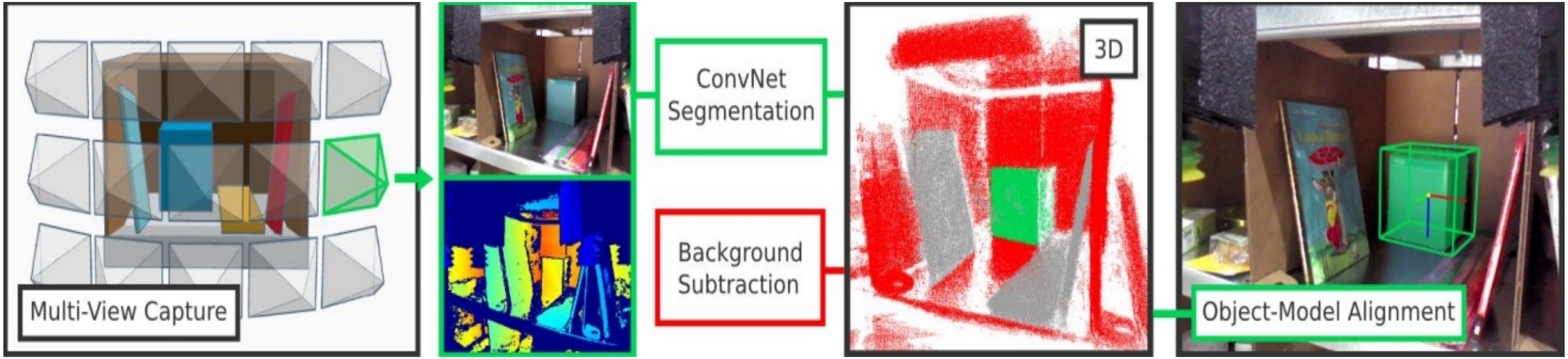
Technical challenges arising from the problem

- Cluttered environments: shelves and totes may have multiple objects and could be arranged as to deceive vision algorithms (e.g., objects on top of one another).
- Self-occlusion: due to limited camera positions, the system only sees a partial view of an object.
- Missing data: commercial depth sensors are unreliable at capturing reflective, transparent, or meshed surfaces, all common in product packaging.
- Small or deformable objects: small objects provide fewer data points, while deformable objects are difficult to align to prior models.
- Speed: the total time dedicated to capturing and processing visual information is under 20 seconds

Motivation and Main Problem

The role of the AI and machine learning in tackling this problem

In the proposed approach, we segment and label multiple views of a scene with a fully convolutional neural network, and then fit pre-scanned 3D object models to the resulting segmentation to get the 6D object pose.



AncoraSIR.com

Motivation and Main Problem

High-level idea of why prior approaches didn't already solve

- While the 2015 APC winning team used a histogram backprojection method with manually defined features, recent work in computer vision has shown that deep learning considerably improves object segmentation
- In pose estimation,existing frameworks such as LINEMOD or MOPED work well under certain assumptions such as objects sitting on a table top with good illumination, but underperform when confronted with the limited visibility, shadows, and clutter imposed by the APC scenario.

Context / Related Work / Limitations of Prior Work

Other papers and key limitations

Zhang, H., Long, P., Zhou, D., Qian, Z., Wang, Z., Wan, W., . . . Pan, J. (2016). *DoraPicker: An autonomous picking system for general objects*. Ithaca: Cornell University Library, arXiv.org. Retrieved from https://www.proquest.com/working-papers/dorapicker-autonomous-picking-system-general/docview/2077047016/se-2

- The perception method proposed in this paper currently has a low precision for transparent or reflective objects as they are invisible to our RGBD camera.
- If the target object is very close to other objects, the robot may simultaneously pick up the target object with a non-target item. This is because our current pipeline does not have a scheme to separate them. In addition, while picking up a very large object from the bin as shown in Figure 10(b), the robot may get stuck inside the bin.
- Again, for our current method, we require the target object to be not occluded by other objects. If most part of the target object is not observed by our RGBD sensor, the perception component will fail to detect and locate the object.

Problem Setting

Key definitions

1.Multi-view Self-supervised Deep Learning:

A type of machine learning technique that uses multiple views of the same object to train a deep learning model in a self-supervised manner, without the need for explicit annotation or labeling of the data.

2.6D Pose Estimation:

6D Pose Estimation refers to the process of estimating the 3D position and orientation of an object in a 3D space relative to a camera. The 6Dpose estimation problem is challenging because it requires precise and accurate localization of the object in a 3D space.

3.RGB-D data:

RGB-D data refers to a type of data that contains both color (RGB) and depth (D) information of a scene or object. RGB-D data is typically captured using specialized sensors such as Microsoft&apos;s Kinect or Intel&apos;s RealSense cameras, which use structured light or time-of-flight technology to capture depth information.

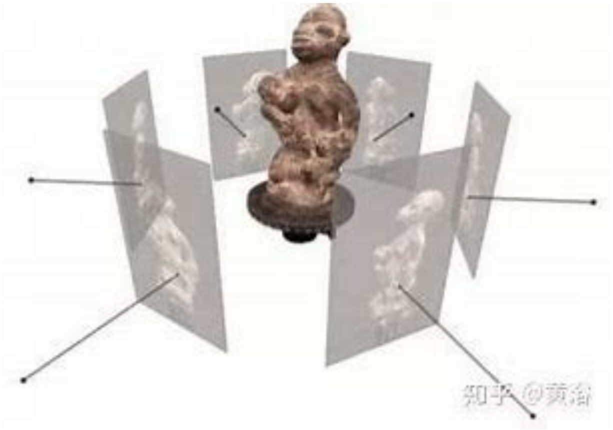
4.Fully Convolutional Networks:

Fully Convolutional Networks (FCNs) are a type of deep neural network architecture that is commonly used for image segmentation tasks.

Proposed Approach / Algorithm / Method

Multi-view self-supervised deep learning framework for solving 6D pose estimation problem

- Multi-view image generation
- self-supervised learning
- pose estimation



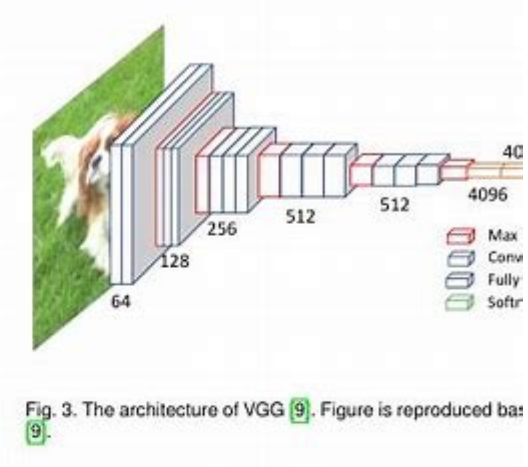
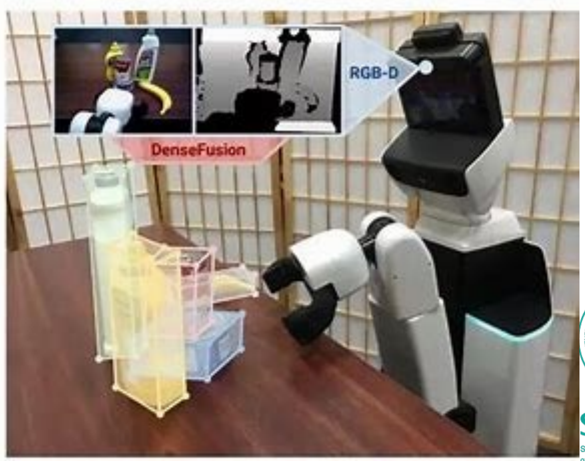


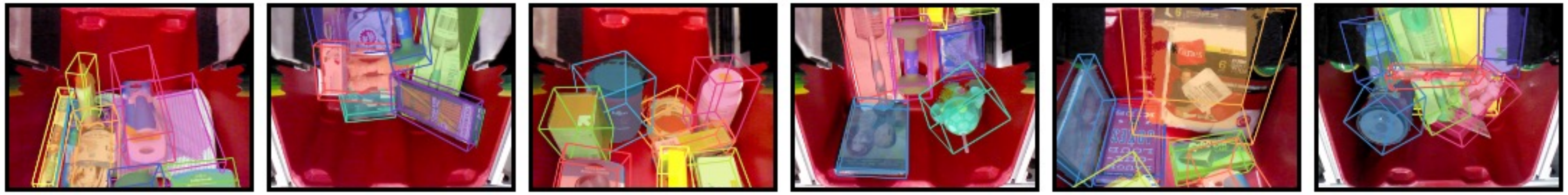
Fig. 3. The architecture of VGG [9]. Figure is reproduced bas
[9].



Proposed Approach / Algorithm / Method

Multi-view image generation

- Capture a set of multi-view images of objects from multiple perspectives.
- Then, the images were preprocessed, including background removal, alignment, and cropping.
- Next, they used this preprocessed set of images to perform self-supervised learning. Specifically, they used images from different views to construct a self-supervised task for each object.



Proposed Approach / Algorithm / Method

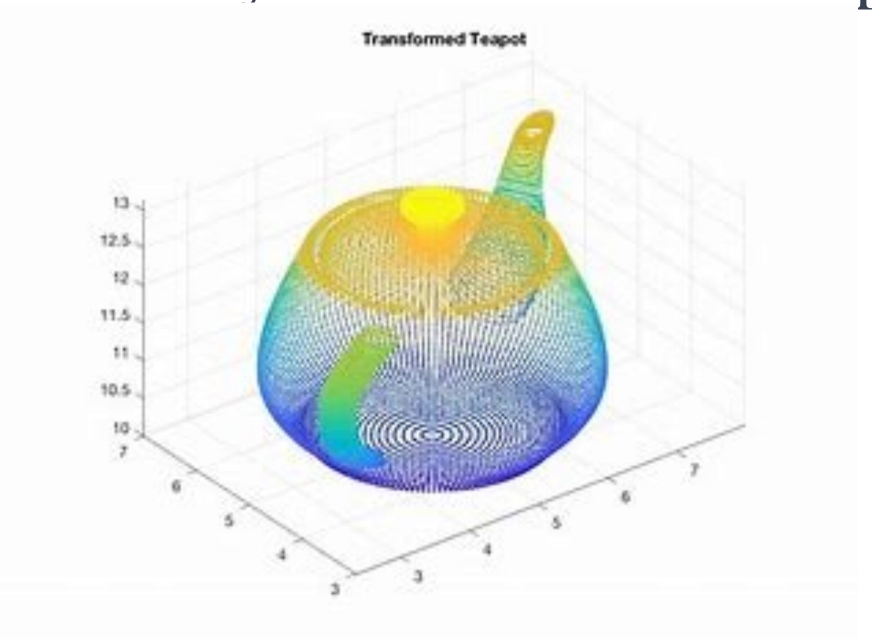
Self-supervised learning

- In a pair of images from different views, one image is used as the "query image" and the other as the "positive sample". Then, by rendering the query image using a 3D model, a "synthetic image" is obtained as the "negative sample".
- Next, the three images are input into the deep neural network for training, with the goal of minimizing the distance between the positive and negative samples.
- During the training process, they used some techniques to improve the training effect, including image enhancement, loss function design, and weight sharing.

Proposed Approach / Algorithm / Method

pose estimation

- During the training process, they employed several techniques to enhance the training effectiveness, including image augmentation, loss function design, and weight sharing. They also utilized Iterative Closest Point (ICP) algorithm to optimize the pose estimation results, which eventually led to the final 6D pose estimation outcome.



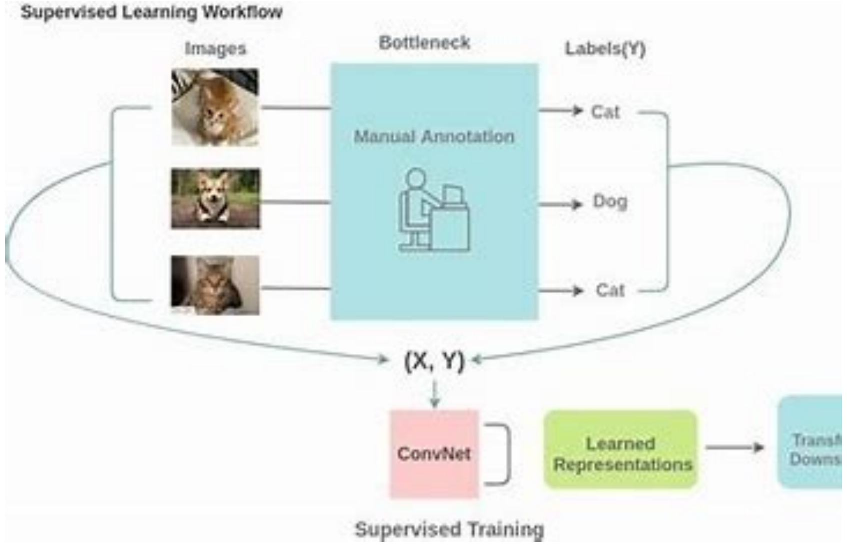
Theory

Self-supervised learning

Self-supervised learning is a machine learning technique that learns representations from unlabeled data without requiring manual annotation. The theory of self-supervised models involves designing appropriate tasks or objective functions for the model to learn useful feature representations.

Common self-supervised learning include predicting missing parts, rotation prediction, colorizing images, etc.

In these tasks, the model generates additional information or labels from the input data and uses it to train the model.



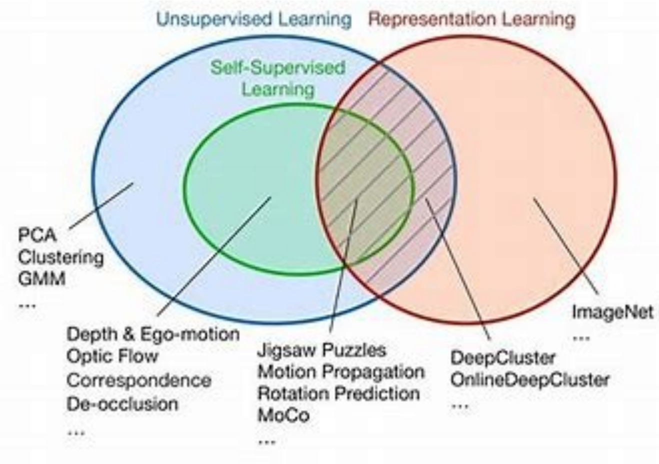
The advantage of self-supervised models is that they can use unlabeled data for training, reducing the need for labeled data. Additionally, self-supervised models can learn from a wider range of data sources, improving the model's generalization performance.

AncoraSIR.com

Theory

Self-supervised learning

However, the training process of self-supervised models also faces some challenges. For example, how to design effective self-supervised tasks, how to balance the model's training across different tasks, and how to use the results of self-supervised training for downstream tasks. These issues require in-depth research and practice.



AncoraSIR.com

domains and baseline

vision system

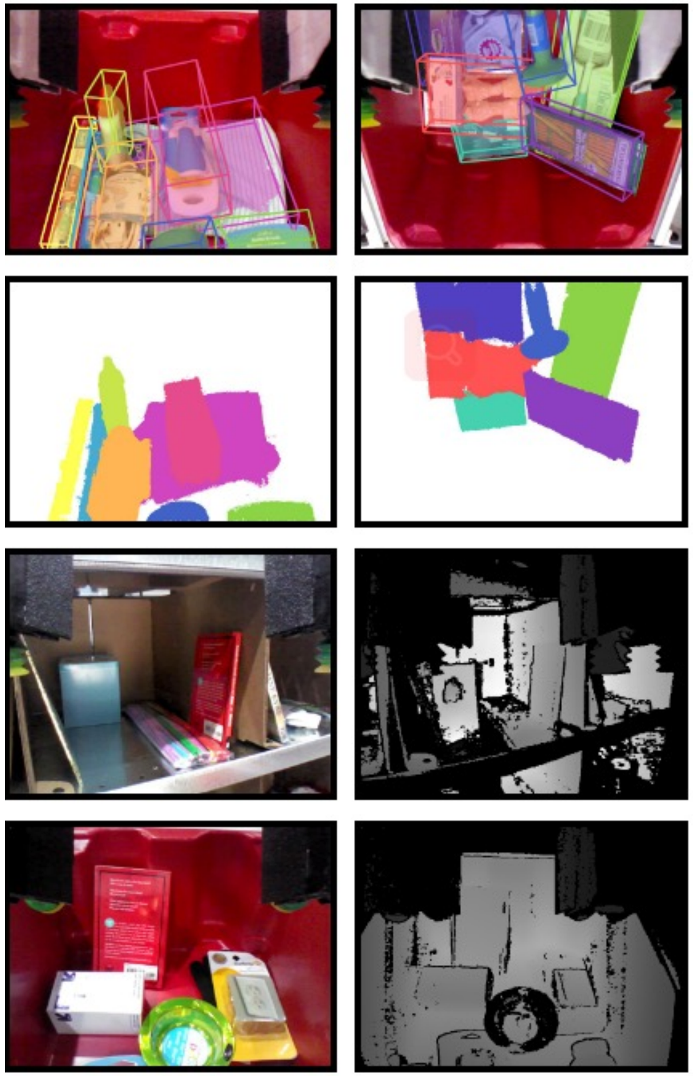
- A robust multi-view vision system to estimate the 6D pose of objects;

self-supervised method

- A self-supervised method that trains deep networks by automatically labeling training data;

benchmark dataset

- A benchmark dataset for estimating object poses.

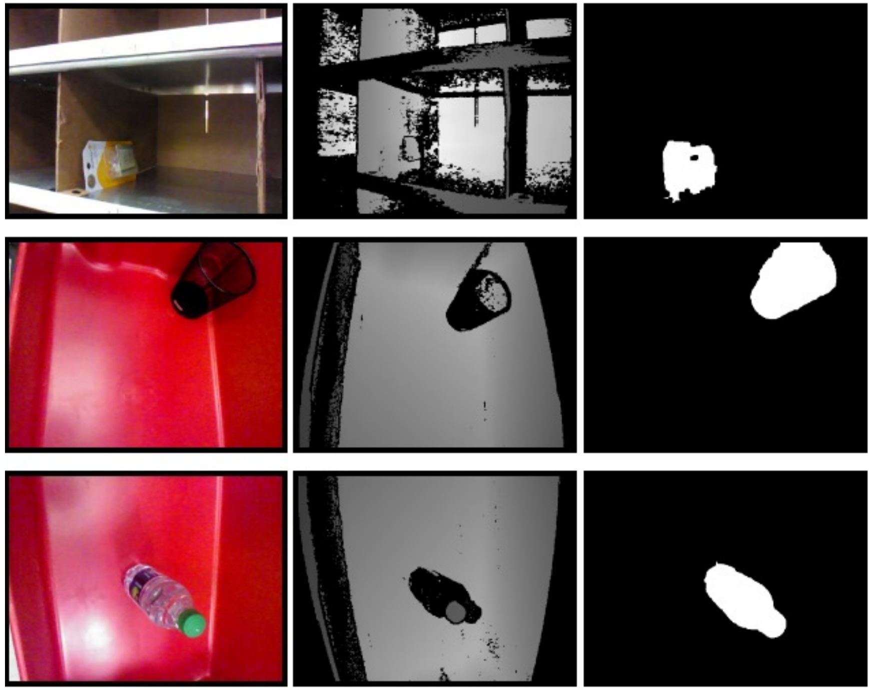


domains and baseline

- Object segmentation:extend the state-of-the art deep learning architecture used for image segmentation.

- Pose estimation: The first aligns 3D CAD models to 3D point clouds with algorithms,The second uses more elaborated local descriptors.

- Benchmark for 6D pose estimation: produce a large benchmark dataset with scenarios from APC 2016.



AncoraSIR.com

evaluate their approach

- They evaluate variants of their method in different scenarios in the benchmark dataset to understand :
- (1) how segmentation performs under different input modalities and training dataset sizes
- (2) how the full vision system performs.

the quantitative and qualitative results

Among them,this figure clearly shows the power of this visual ability. In such a cluttered environment, the new visual algorithm can accurately and efficiently identify different objects. This vision algorithm estimates the 6D poses of objects robustly under challenging scenarios.

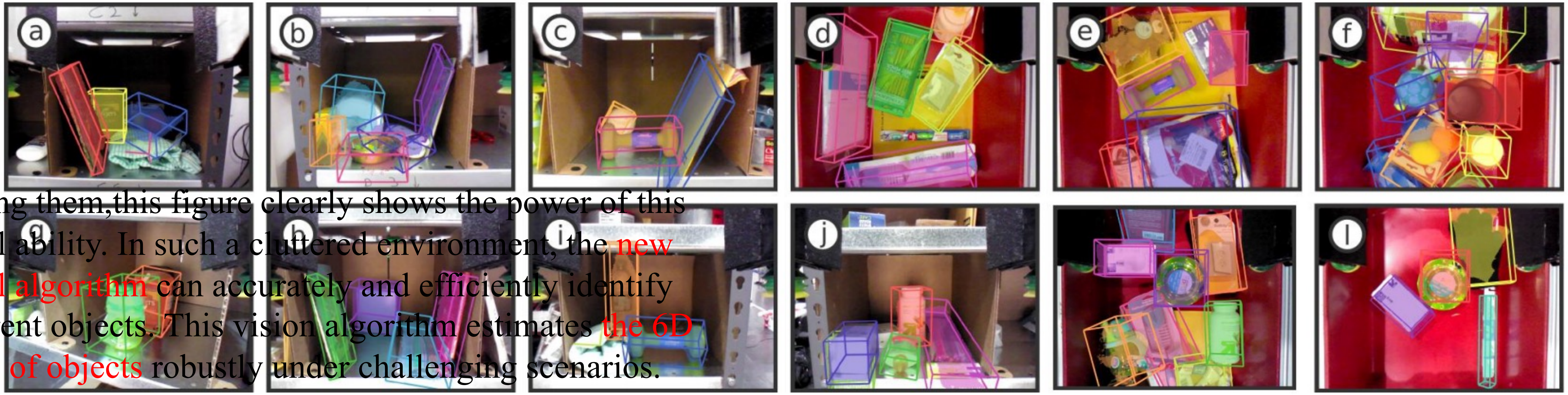


Fig. 8. Example results from our vision system. 6D pose predictions are highlighted with a 3D bounding box. For deformable objects (cloth in a,c,i) we output the center of mass. We additionally illustrate successful pose predictions for objects with missing depth (water bottle, black bin, green sippy cup, green bowl)

the quantitative and qualitative results

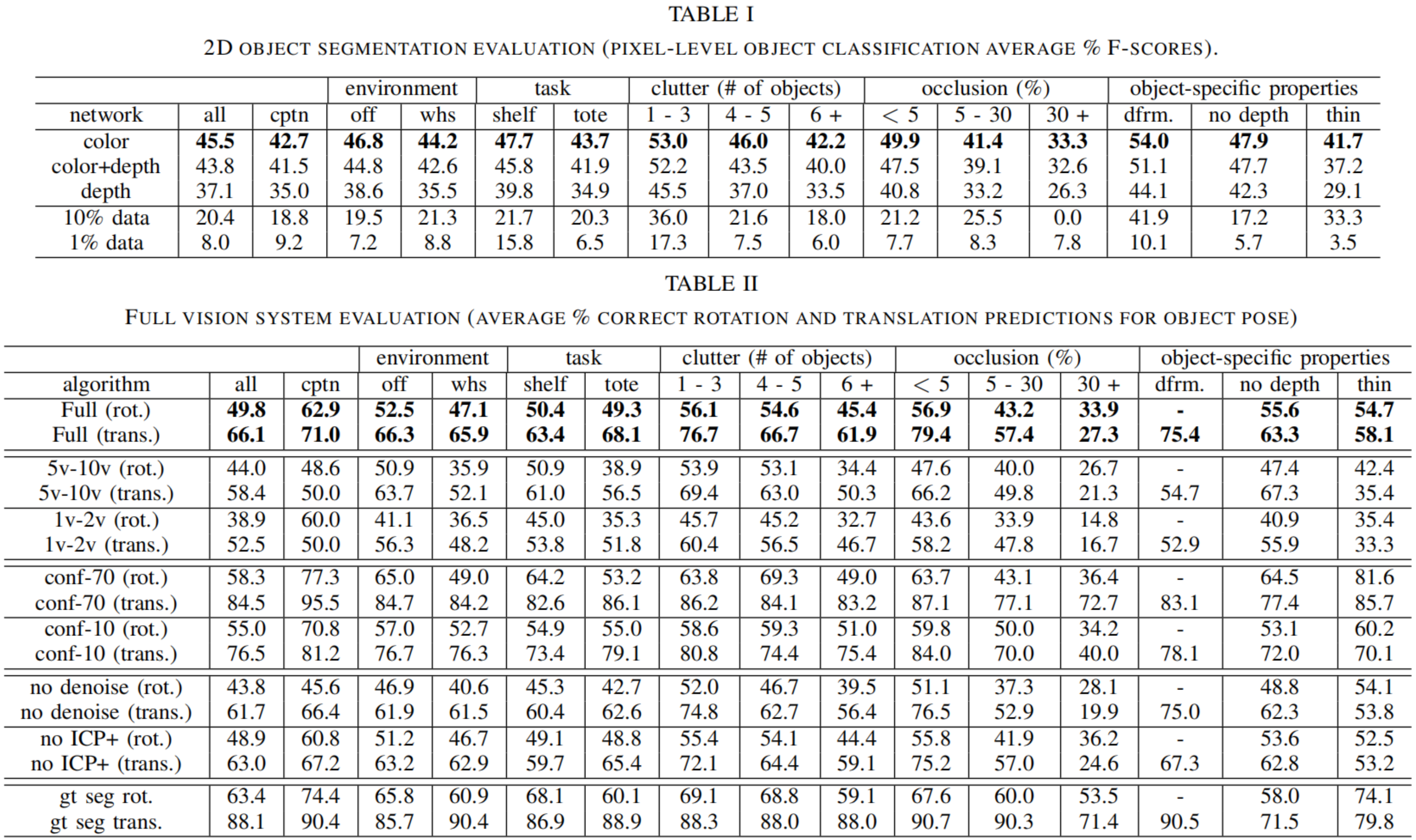


TABLE I

2D OBJECT SEGMENTATION EVALUATION (PIXEL-LEVEL OBJECT CLASSIFICATION AVERAGE % F-SCORES).

			environment		task		clutter (# of objects)			occlusion (%)			object-specific properties		
network	all	cptn	off	whs	shelf	tote	1 - 3	4 - 5	6 +	< 5	5 - 30	30 +	dfrm.	no depth	thin
color	**45.5**	**42.7**	**46.8**	**44.2**	**47.7**	**43.7**	**53.0**	**46.0**	**42.2**	**49.9**	**41.4**	**33.3**	**54.0**	**47.9**	**41.7**
color+depth	43.8	41.5	44.8	42.6	45.8	41.9	52.2	43.5	40.0	47.5	39.1	32.6	51.1	47.7	37.2
depth	37.1	35.0	38.6	35.5	39.8	34.9	45.5	37.0	33.5	40.8	33.2	26.3	44.1	42.3	29.1
10% data	20.4	18.8	19.5	21.3	21.7	20.3	36.0	21.6	18.0	21.2	25.5	0.0	41.9	17.2	33.3
1% data	8.0	9.2	7.2	8.8	15.8	6.5	17.3	7.5	6.0	7.7	8.3	7.8	10.1	5.7	3.5

TABLE II

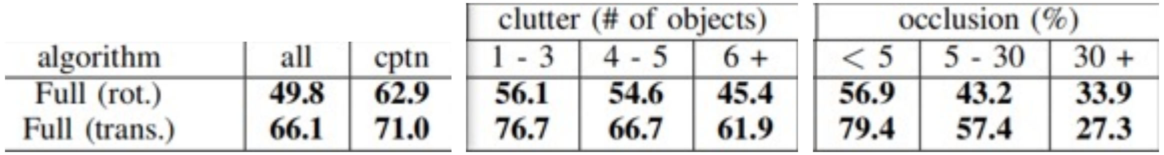
FULL VISION SYSTEM EVALUATION (AVERAGE % CORRECT ROTATION AND TRANSLATION PREDICTIONS FOR OBJECT POSE)

			environment		task		clutter (# of objects)			occlusion (%)			object-specific properties		
algorithm	all	cptn	off	whs	shelf	tote	1 - 3	4 - 5	6 +	< 5	5 - 30	30 +	dfrm.	no depth	thin
Full (rot.)	**49.8**	**62.9**	**52.5**	**47.1**	**50.4**	**49.3**	**56.1**	**54.6**	**45.4**	**56.9**	**43.2**	**33.9**	**-**	**55.6**	**54.7**
Full (trans.)	**66.1**	**71.0**	**66.3**	**65.9**	**63.4**	**68.1**	**76.7**	**66.7**	**61.9**	**79.4**	**57.4**	**27.3**	**75.4**	**63.3**	**58.1**
5v-10v (rot.)	44.0	48.6	50.9	35.9	50.9	38.9	53.9	53.1	34.4	47.6	40.0	26.7	-	47.4	42.4
5v-10v (trans.)	58.4	50.0	63.7	52.1	61.0	56.5	69.4	63.0	50.3	66.2	49.8	21.3	54.7	67.3	35.4
1v-2v (rot.)	38.9	60.0	41.1	36.5	45.0	35.3	45.7	45.2	32.7	43.6	33.9	14.8	-	40.9	35.4
1v-2v (trans.)	52.5	50.0	56.3	48.2	53.8	51.8	60.4	56.5	46.7	58.2	47.8	16.7	52.9	55.9	33.3
conf-70 (rot.)	58.3	77.3	65.0	49.0	64.2	53.2	63.8	69.3	49.0	63.7	43.1	36.4	-	64.5	81.6
conf-70 (trans.)	84.5	95.5	84.7	84.2	82.6	86.1	86.2	84.1	83.2	87.1	77.1	72.7	83.1	77.4	85.7
conf-10 (rot.)	55.0	70.8	57.0	52.7	54.9	55.0	58.6	59.3	51.0	59.8	50.0	34.2	-	53.1	60.2
conf-10 (trans.)	76.5	81.2	76.7	76.3	73.4	79.1	80.8	74.4	75.4	84.0	70.0	40.0	78.1	72.0	70.1
no denoise (rot.)	43.8	45.6	46.9	40.6	45.3	42.7	52.0	46.7	39.5	51.1	37.3	28.1	-	48.8	54.1
no denoise (trans.)	61.7	66.4	61.9	61.5	60.4	62.6	74.8	62.7	56.4	76.5	52.9	19.9	75.0	62.3	53.8
no ICP+ (rot.)	48.9	60.8	51.2	46.7	49.1	48.8	55.4	54.1	44.4	55.8	41.9	36.2	-	53.6	52.5
no ICP+ (trans.)	63.0	67.2	63.2	62.9	59.7	65.4	72.1	64.4	59.1	75.2	57.0	24.6	67.3	62.8	53.2
gt seg rot.	63.4	74.4	65.8	60.9	68.1	60.1	69.1	68.8	59.1	67.6	60.0	53.5	-	58.0	74.1
gt seg trans.	88.1	90.4	85.7	90.4	86.9	88.9	88.3	88.0	88.0	90.7	90.3	71.4	90.5	71.5	79.8

Discussion of Results

Conclusions and Insights

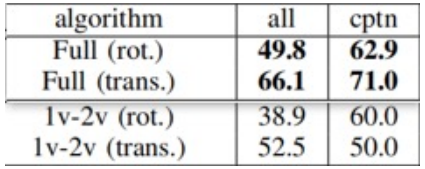
The conclusions drawn by the author are:



			clutter (# of objects)			occlusion (%)		
algorithm	all	cptn	1 - 3	4 - 5	6 +	< 5	5 - 30	30 +
Full (rot.)	**49.8**	**62.9**	**56.1**	**54.6**	**45.4**	**56.9**	**43.2**	**33.9**
Full (trans.)	**66.1**	**71.0**	**76.7**	**66.7**	**61.9**	**79.4**	**57.4**	**27.3**

1. Multiple views robustly address occlusion and heavy clutter in the warehouse setting.

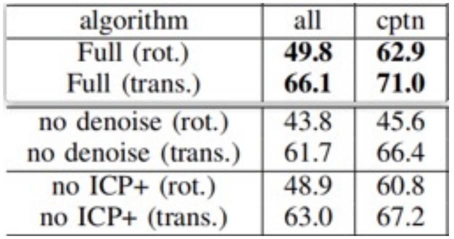
2. Contrast between the performance of the algorithm using a single view of the scene, versus multiple views of the scene.



algorithm	all	cptn
Full (rot.)	**49.8**	**62.9**
Full (trans.)	**66.1**	**71.0**
1v-2v (rot.)	38.9	60.0
1v-2v (trans.)	52.5	50.0

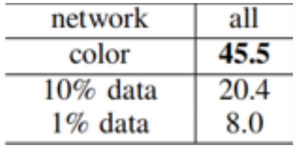
3. The denoising step proves important for achieving good results.

4. Without the pre-processing steps to ICP, we observe a drop in prediction



algorithm	all	cptn
Full (rot.)	**49.8**	**62.9**
Full (trans.)	**66.1**	**71.0**
no denoise (rot.)	43.8	45.6
no denoise (trans.)	61.7	66.4
no ICP+ (rot.)	48.9	60.8
no ICP+ (trans.)	63.0	67.2

5. The volume of training data is strongly correlated with performance.



network	all
color	**45.5**
10% data	20.4
1% data	8.0

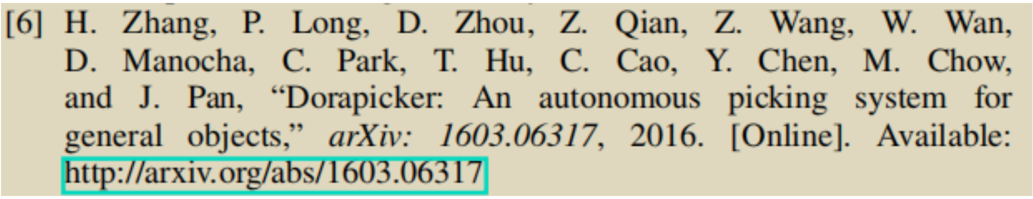
Discussion of Results

Supports

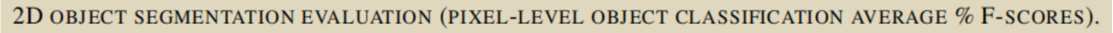
The results and references support the conclusion well.

The results support the conclusion by the experimental statistics. The previous page showed that the conclusions and insights are drawn from the tables, which include detailed data to fully support the conclusion.

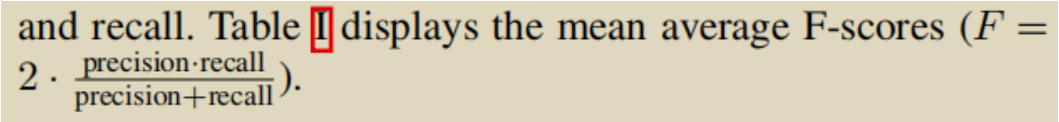
The reference we chose did not show statistics, but there are some statements that support the conclusion. “Again, for our current method, we require the target object to be not occluded by other objects. If most part of the target object is not observed by our RGBD sensor, the perception component will fail to detect and locate the object.” This means the application of the multiple views that increased the accuracy of clutter and occlusion is critical.



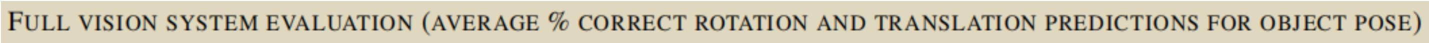
[6] H. Zhang, P. Long, D. Zhou, Z. Qian, Z. Wang, W. Wan, D. Manocha, C. Park, T. Hu, C. Cao, Y. Chen, M. Chow, and J. Pan, “Dorapicker: An autonomous picking system for general objects,” *arXiv: 1603.06317*, 2016. [Online]. Available: http://arxiv.org/abs/1603.06317



2D OBJECT SEGMENTATION EVALUATION (PIXEL-LEVEL OBJECT CLASSIFICATION AVERAGE % F-SCORES).



and recall. Table I displays the mean average F-scores ($F = 2 \cdot \frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}$).



FULL VISION SYSTEM EVALUATION (AVERAGE % CORRECT ROTATION AND TRANSLATION PREDICTIONS FOR OBJECT POSE)

Discussion of Results

Critique and Limitations

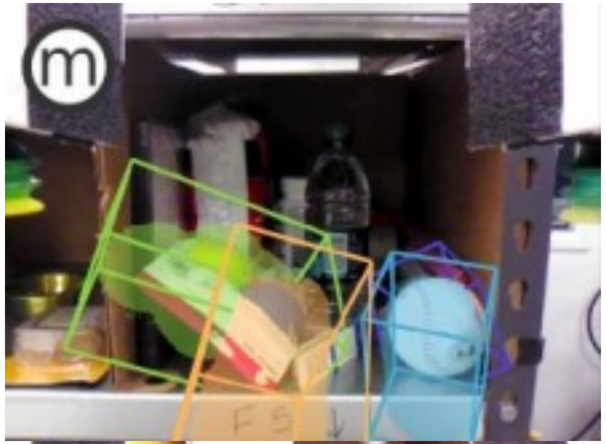
The usage scenarios of the paper is in the Amazon Picking Challenge, so the limitation of the technology is transparent: it should not cost much. However, it is a huge project to set quantities of cameras is the warehouse. Also, it requires a large amount of training data to train the deep neural network.

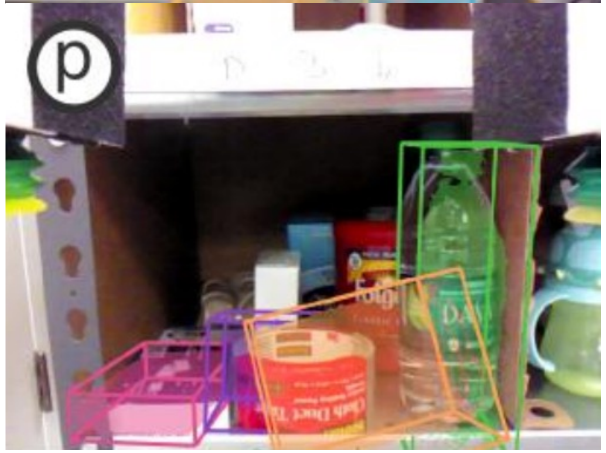
Apart from this, the amount of computation should be brought into account. In this paper, they just take the shelf and the tote. However, in real case, there must be countless objects should be recognized and predicted. Therefore, the amount of computation would be astonishing.

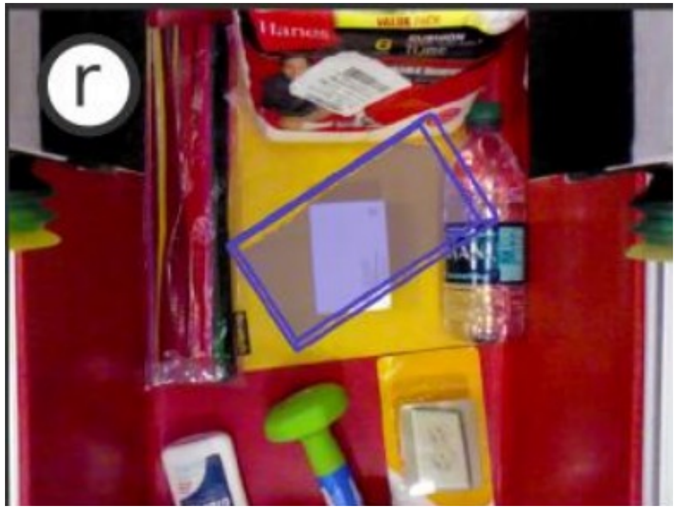
Future Work for Paper / Reading

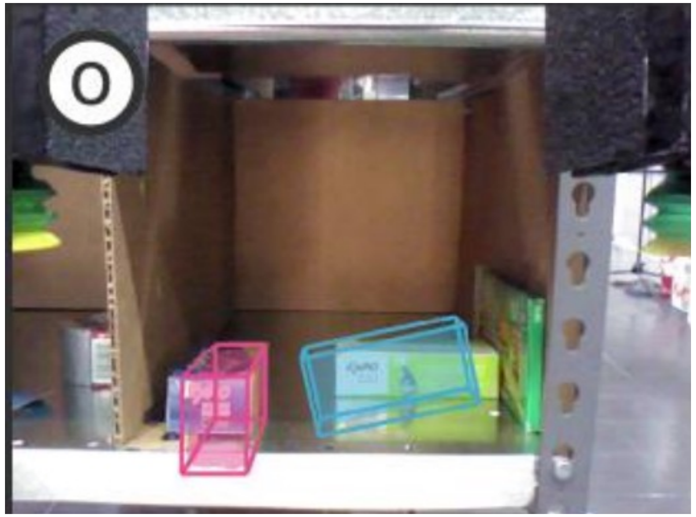
Interesting Questions:

- FCN(Fully Convolutional Network) may be incomplete: under **heavy occlusion** or **clutter** or **poor illumination*(m, p)***
- Objects color textures are confused with each other. ***(r)***









- Model fitting for cuboid objects often confuses corner alignments. ***(o)***

AncoraSIR.com

Future Work for Paper / Reading

How to Improve:

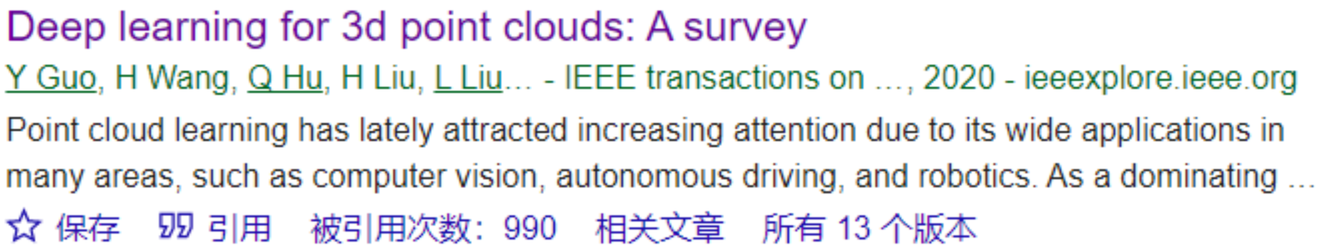
- Make the most out of every constraint:
- Using bin and shelf mode
- The volume of training data is strongly correlated with performance

- Designing robotic and vision systems hand-in-hand:
- Robotic arms allow us to precisely fuse multiple views and improve performance in cluttered environments
- Configure different sensing for different effectors.

AncoraSIR.com

Extended Readings

Papers that use this paper as a reference:



Deep learning for 3d point clouds: A survey

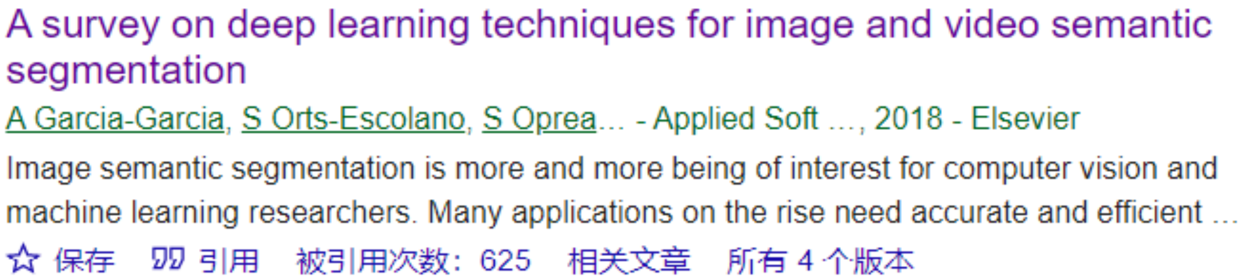
Y Guo, H Wang, Q Hu, H Liu, L Liu... - IEEE transactions on ..., 2020 - ieeexplore.ieee.org

Point cloud learning has lately attracted increasing attention due to its wide applications in many areas, such as computer vision, autonomous driving, and robotics. As a dominating ...

☆ 保存 ⁹⁹ 引用 被引用次数：990 相关文章 所有 13 个版本

(1)3D Point Clouds:

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., & Bennamoun, M. (2020). Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *43*(12), 4338-4364.



A survey on deep learning techniques for image and video semantic segmentation

A Garcia-Garcia, S Orts-Escolano, S Oprea... - Applied Soft ..., 2018 - Elsevier

Image semantic segmentation is more and more being of interest for computer vision and machine learning researchers. Many applications on the rise need accurate and efficient ...

☆ 保存 ⁹⁹ 引用 被引用次数：625 相关文章 所有 4 个版本

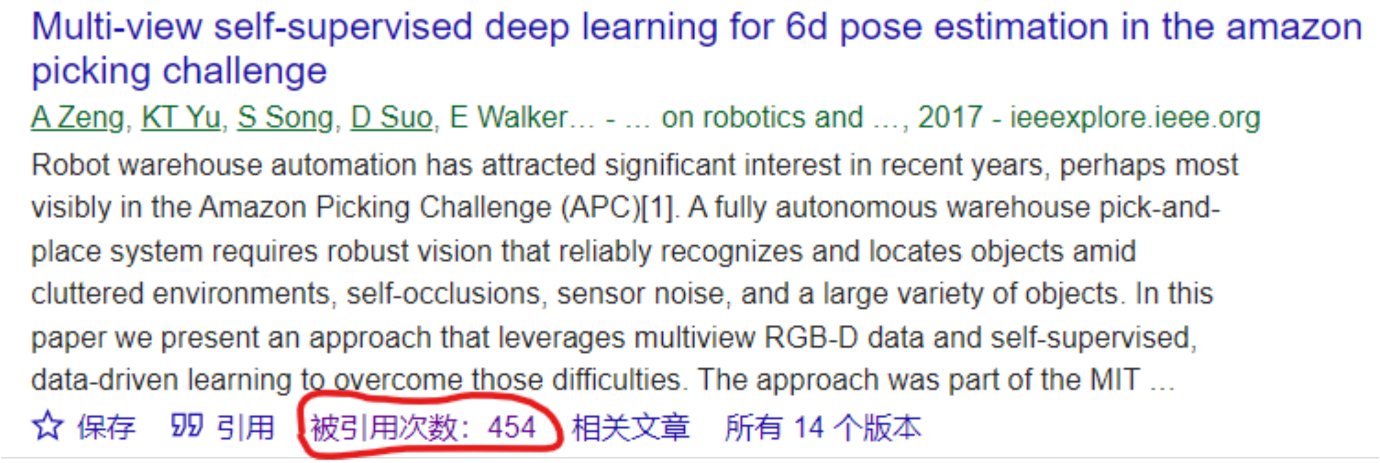
(2) Segmentation & Deep Learning for Image and Video:

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., & Garcia-Rodriguez, J. (2018). A survey on deep learning techniques for image and video semantic segmentation. Applied Soft Computing, 70, 41-65.

Extended Readings

How to find:

- Web of Science(IF 影响因子): *https://www.webofscience.com/wos/woscc/basic-search*
- Dimensions: *https://app.dimensions.ai/discover/publication*
- PubMed: *https://pubmed.ncbi.nlm.nih.gov/*
- Google Scholar: *https://scholar.google.com/*



Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge

A Zeng, KT Yu, S Song, D Suo, E Walker… - … on robotics and …, 2017 - ieeexplore.ieee.org

Robot warehouse automation has attracted significant interest in recent years, perhaps most visibly in the Amazon Picking Challenge (APC)[1]. A fully autonomous warehouse pick-and-place system requires robust vision that reliably recognizes and locates objects amid cluttered environments, self-occlusions, sensor noise, and a large variety of objects. In this paper we present an approach that leverages multiview RGB-D data and self-supervised, data-driven learning to overcome those difficulties. The approach was part of the MIT …

Summary

- **Mainly discuss**:

multi-view self-supervised deep learning approach for 6D pose estimation

- **Why important and hard**:

Fully autonomous warehouse pick-and-place system require robust vision, where the environment is complex

- **Key limitation of prior work**:

constrain of data set, limitation of supervised learning method, deficiency of pose optimization algorithm

- **Key insights of the proposed work:**

make the most out of constraint, integrated design of visions and robots

- **Demonstrate by this insight:**

increased the accuracy of clutter and occlusion is critical

AncoraSIR.com

Thank you for listening!

Reporters:

Ren Shize

Li Zhidong

Jiang Meng

Zeng Yuqi

Zhao Xuda

Q & A

AncoraSIR.com