Real-Time Grasp Detection Using Convolutional Neural Networks

Presenter: 杨雪 洪雨盈 魏毓瞳 李蕴哲

20240312

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

- General scene understanding requires complex visual tasks such as segmenting a scene into component parts, recognizing what those parts are, and disambiguating between visually similar objects. Due to these complexities, visual perception is a large bottleneck in real robotic systems.

- Robot interaction with the physical world needs: general purpose robots need to be able to interact with and manipulate objects in the physical world. Robotic grasp detection lags far behind human performance.

the role of the AI and machine learning

- Data Processing: Complex features can be automatically learned from large amounts of training data using deep learning. This is critical for understanding the shape, size, and possible grip points of an object.
- Accuracy Improvement: The accuracy of grasp detection can be significantly improved through deep learning model training and optimization. This is critical for improving the success rate of robots operating in complex environments.
- Real-Time Processing: Methods utilizing deep learning can significantly increase processing speeds, enabling robots to make grasping decisions at real-time or near real-time speeds.

AncoraSIR.com

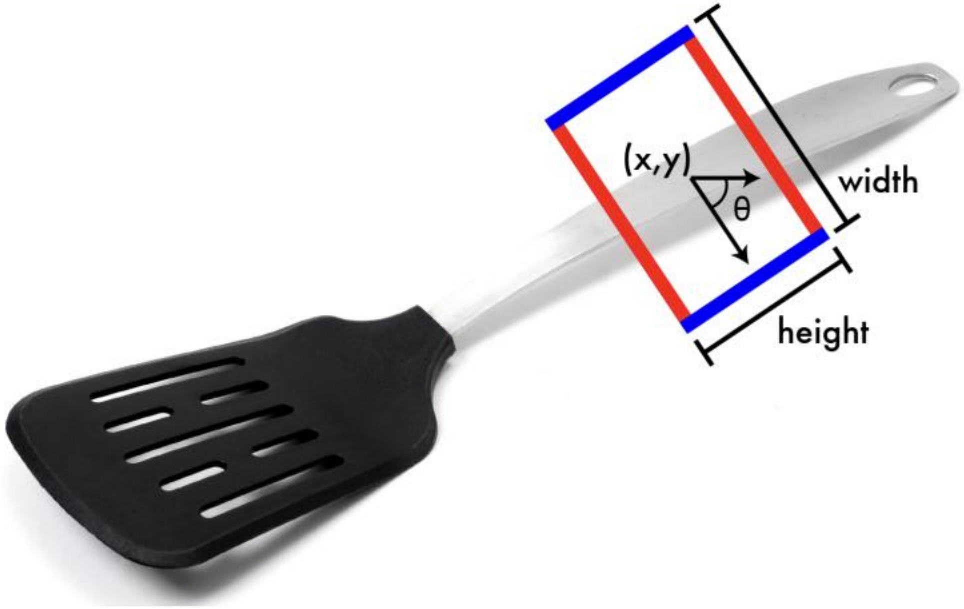
Problem Setting

The **core problem** is to determine a method for safely picking up and holding an object, given its image. So the robot can grasp the object without causing damage or dropping it.

Grasp Representation:

A five-dimensional grasp representation, with terms for location, size, and orientation. The blue lines demark the size and orientation of the gripper plates. The red lines show the approximate distance between the plates before the grasp is executed.

Problem Setting



g={x,y,θ,h,w}

(x,y): The center of the rectangle (grasp center).

θ: The orientation of the rectangle relative to the horizontal axis.

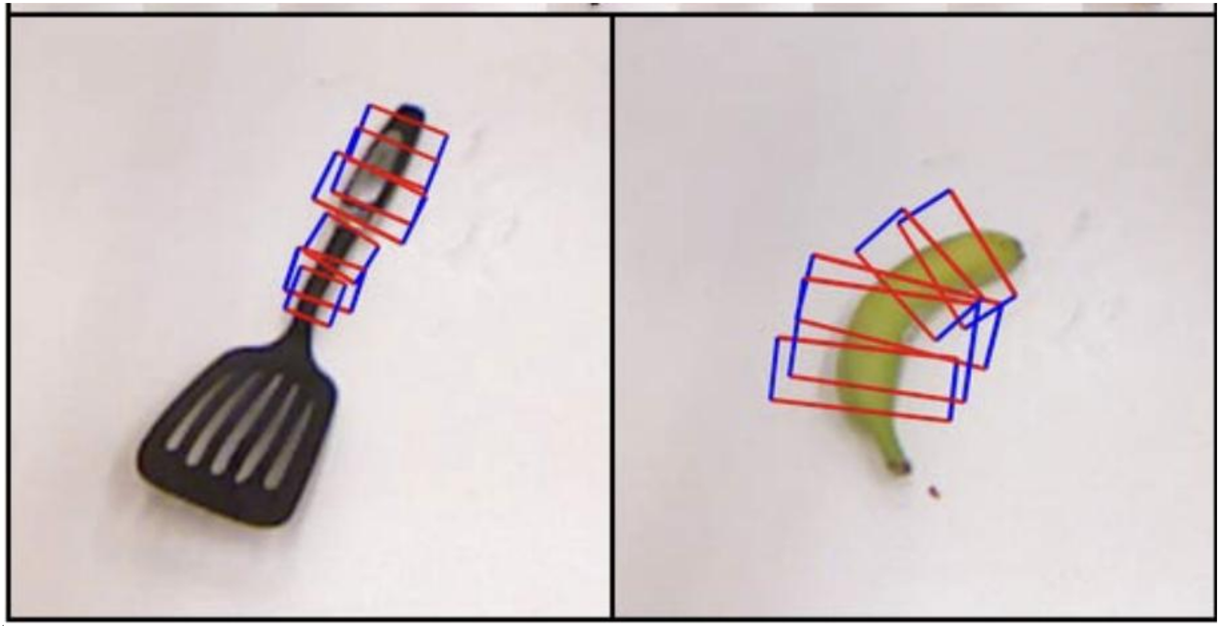
h: The height of the rectangle (gripper opening).

w: The width of the rectangle (gripper plate size).

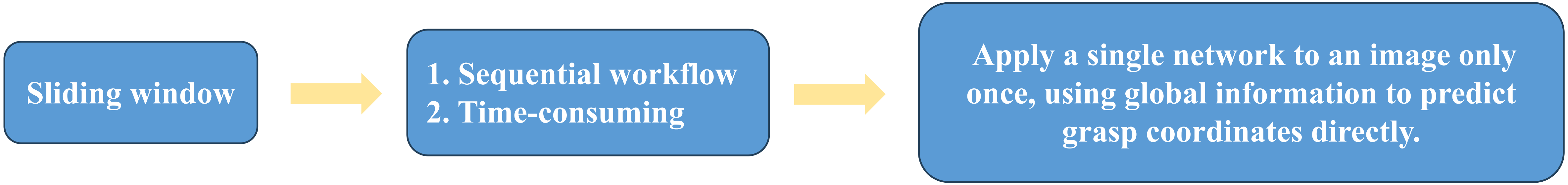
Limitations of Prior Work

Based on evaluation of the Cornell grasp detection dataset performed on the The latest findings on this dataset run at 13.5 seconds per frame with 75% accuracy.

This means that there will be a delay of 13.5 seconds between when the robot views the scene and when it determines where its gripper is moving.



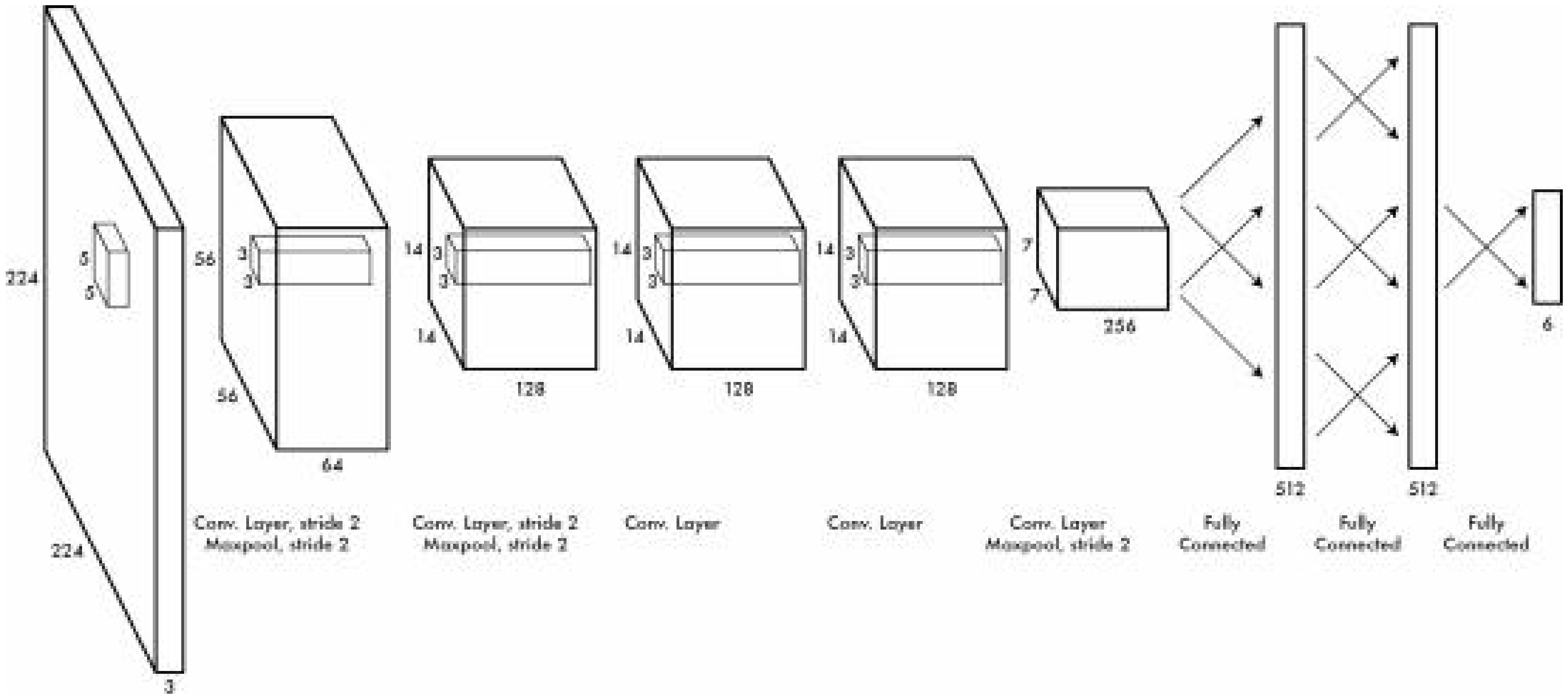
Proposed Approach / Algorithm / Method



Sliding window

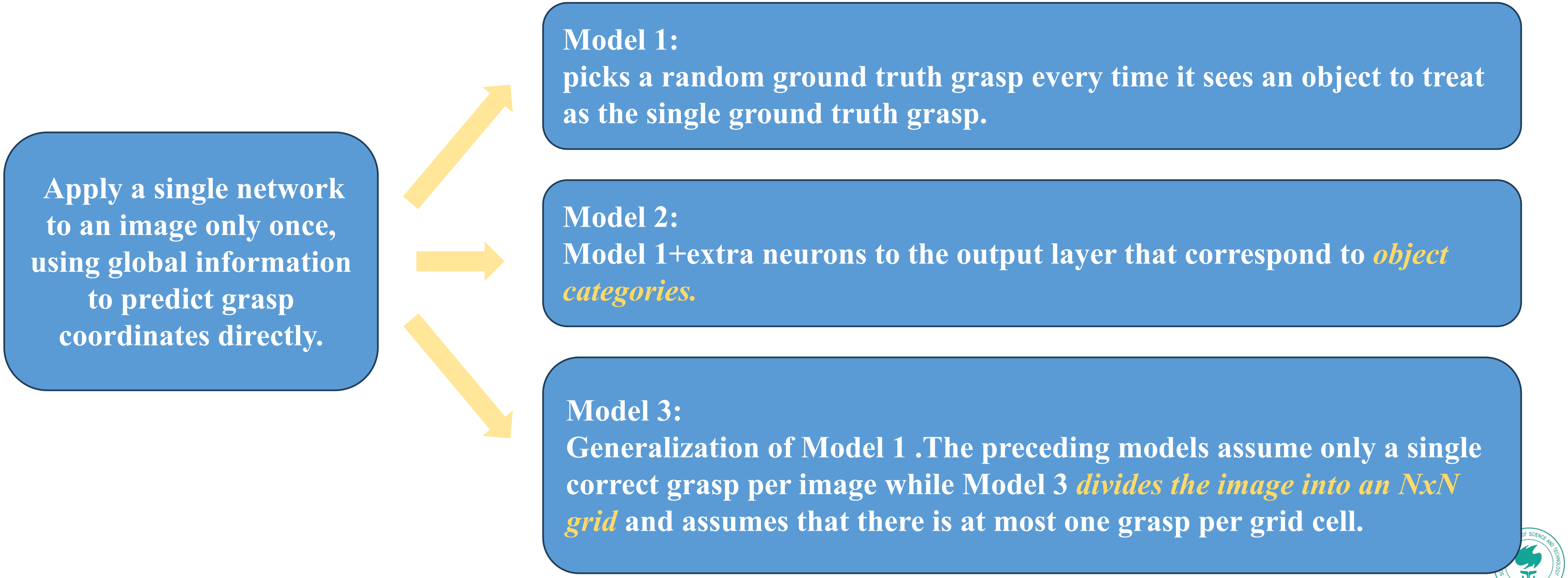
1. Sequential workflow
2. Time-consuming

Apply a single network to an image only once, using global information to predict grasp coordinates directly.



AncoraSIR.com

Proposed Approach / Algorithm / Method



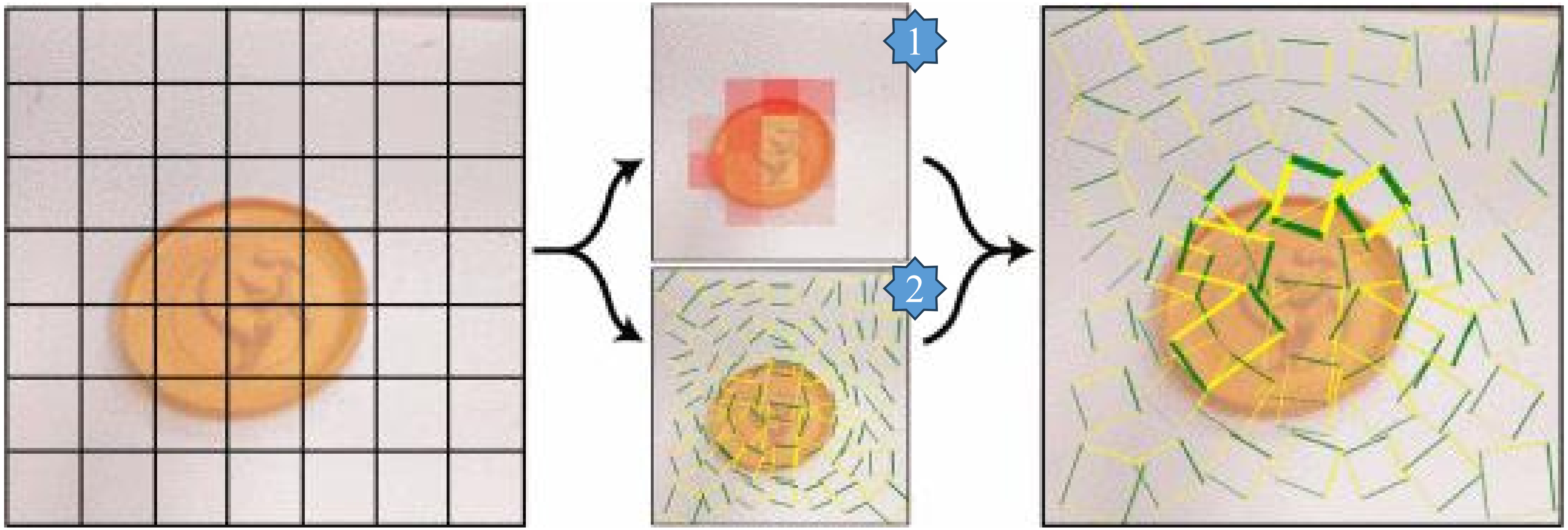
Apply a single network to an image only once, using global information to predict grasp coordinates directly.

Model 1:
picks a random ground truth grasp every time it sees an object to treat as the single ground truth grasp.

Model 2:
Model 1+extra neurons to the output layer that correspond to *object categories.*

Model 3:
Generalization of Model 1 .The preceding models assume only a single correct grasp per image while Model 3 *divides the image into an NxN grid* and assumes that there is at most one grasp per grid cell.

Proposed Approach / Algorithm / Method



Model 3 visualization:

For each cell in the grid, the model predicts

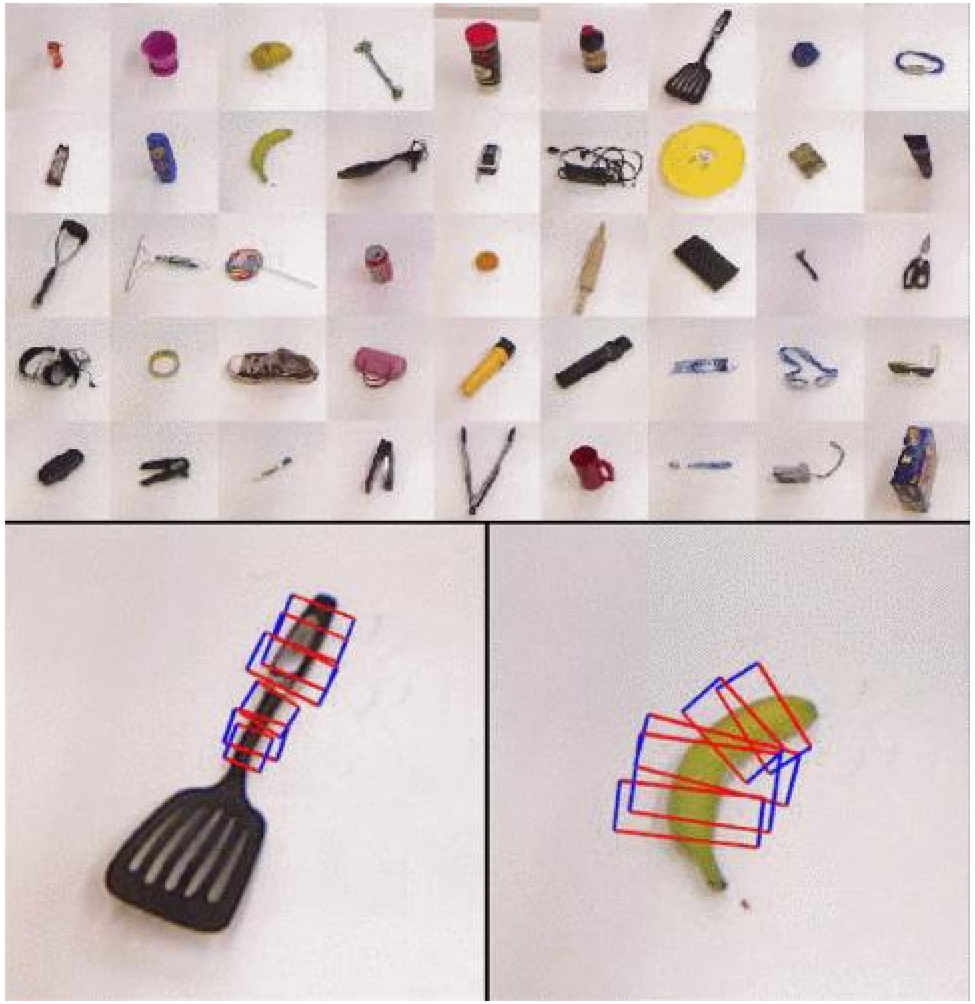
① a bounding box centered at that cell

②a probability that this grasp is a true grasp for the object in the image.

The predictions are weighted by this probability predicting multiple grasps for an object.

AncoraSIR.com

Theory



Compared to the sliding window method, this method scaled the network to generate a global understanding for the whole image.

①For items with multiple graspable points, grasping strategies can be generated by applying this network only once.

②Global information of the image provides good object classification.

AncoraSIR.com

Experimental Setup

Pretrain：

Datasets： ImageNet

Tasks： Classification

Train：

Datasets： Cornell grasping dataset

Tasks： Grasping strategy prediction

Hardware setups：

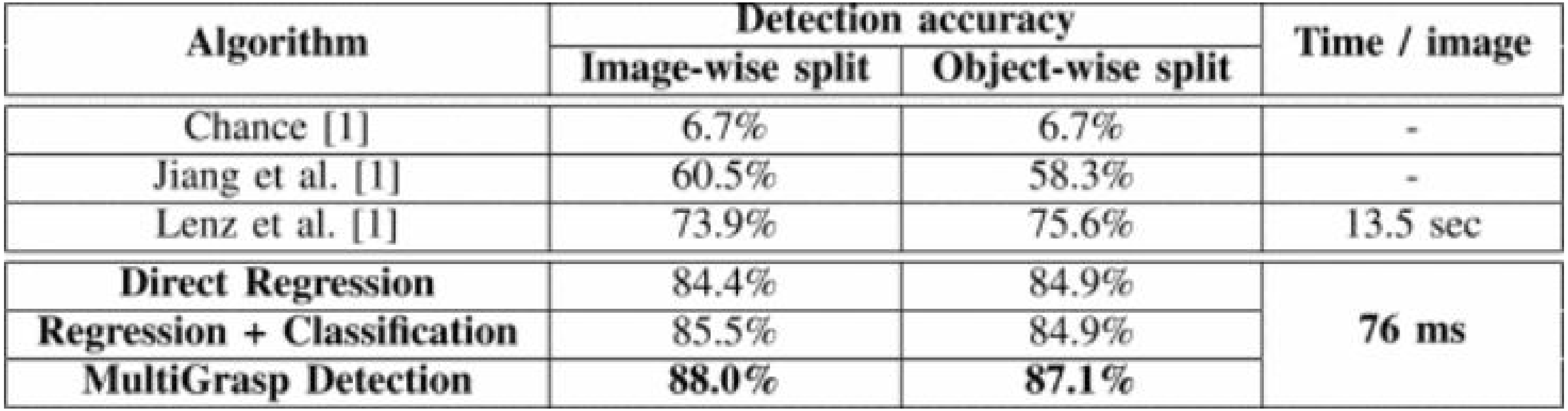
Nvidia Tesla K20 GPU

As no actual grasping work contained, there is no actual robotic platform illustrated in this article.

Experimental Setup

Baseline：

Proposed：



Algorithm	Detection accuracy		Time / image
	Image-wise split	Object-wise split	
Chance [1]	6.7%	6.7%	-
Jiang et al. [1]	60.5%	58.3%	-
Lenz et al. [1]	73.9%	75.6%	13.5 sec
Direct Regression	84.4%	84.9%	**76 ms**
Regression + Classification	85.5%	84.9%	
MultiGrasp Detection	**88.0%**	**87.1%**	

Experimental Results

In Table 1 we compare our results to previous work using their self-reported scores for the rectangle metric accuracy.

Across the board our models outperform the current state-of-the-art both in terms of accuracy and speed.

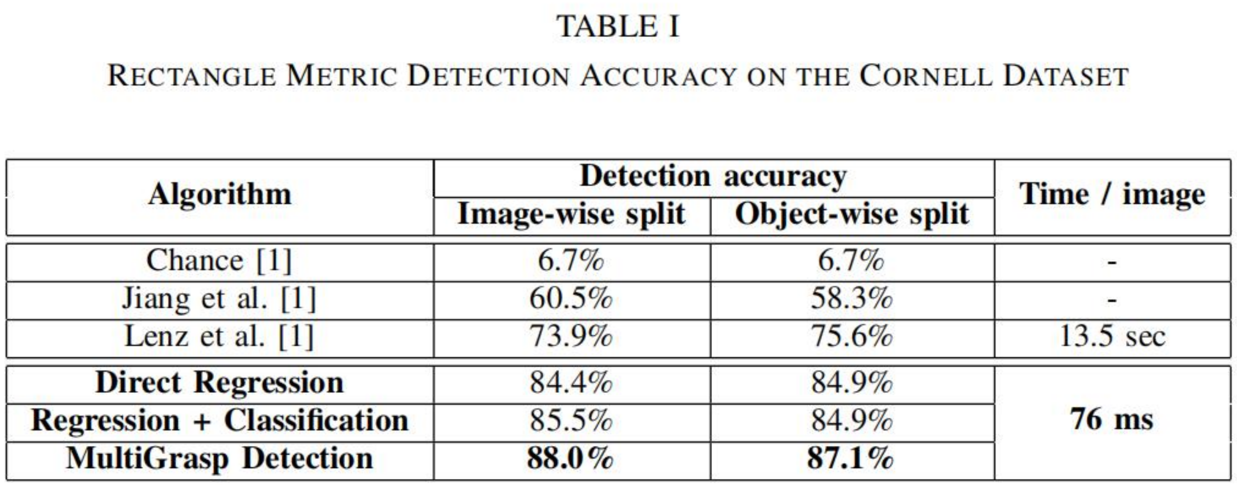


TABLE I

RECTANGLE METRIC DETECTION ACCURACY ON THE CORNELL DATASET

Algorithm	**Detection accuracy**		**Time / image**
	Image-wise split	**Object-wise split**	
Chance [1]	6.7%	6.7%	-
Jiang et al. [1]	60.5%	58.3%	-
Lenz et al. [1]	73.9%	75.6%	13.5 sec
Direct Regression	84.4%	84.9%	**76 ms**
Regression + Classification	85.5%	84.9%	
MultiGrasp Detection	**88.0%**	**87.1%**	

Experimental Results

Direct Regression

The direct regression model sets a new baseline for performance in grasp detection. It achieves around 85 percent accuracy in both image-wise and object-wise splits, ten percentage points higher than the previous best.

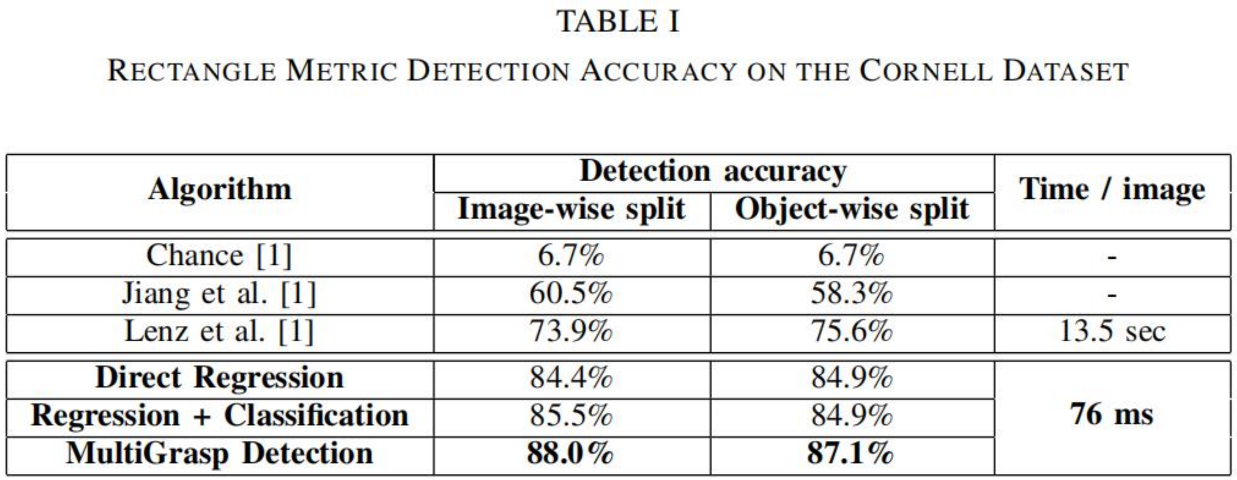


TABLE I

RECTANGLE METRIC DETECTION ACCURACY ON THE CORNELL DATASET

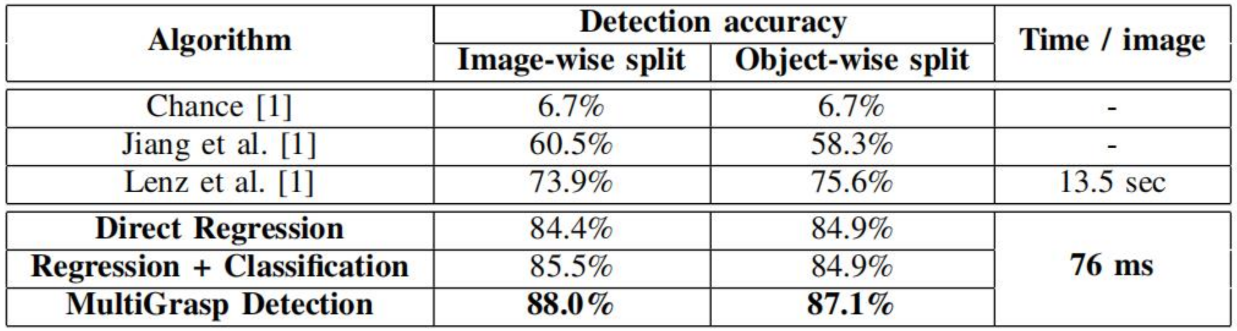
Algorithm	Detection accuracy		Time / image
	Image-wise split	Object-wise split	
Chance [1]	6.7%	6.7%	-
Jiang et al. [1]	60.5%	58.3%	-
Lenz et al. [1]	73.9%	75.6%	13.5 sec
Direct Regression	84.4%	84.9%	**76 ms**
Regression + Classification	85.5%	84.9%	
MultiGrasp Detection	**88.0%**	**87.1%**	

Experimental Results

At test time the direct regression model runs in 76 milliseconds per batch, with a batch size of 128 images. While this amounts to processing more than 1,600 images per second, latency matters more than throughput in grasp detection so we report the per batch number as 13 fps..

TABLE I

Rectangle Metric Detection Accuracy on the Cornell Dataset

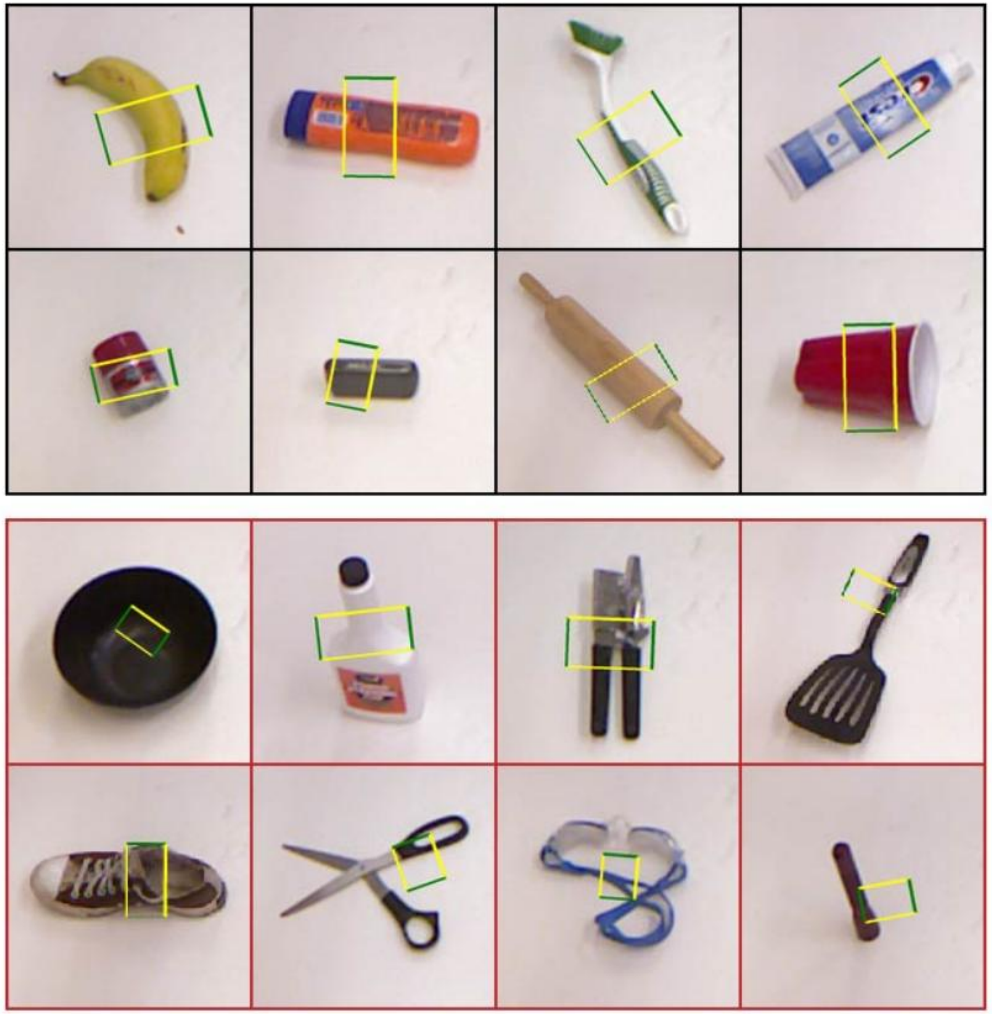


Algorithm	Detection accuracy		Time / image
	Image-wise split	Object-wise split	
Chance [1]	6.7%	6.7%	-
Jiang et al. [1]	60.5%	58.3%	-
Lenz et al. [1]	73.9%	75.6%	13.5 sec
Direct Regression	84.4%	84.9%	**76 ms**
Regression + Classification	85.5%	84.9%	
MultiGrasp Detection	**88.0%**	**87.1%**	

Experimental Results

Examples of correct (top) and incorrect (bottom) grasps from the direct regression model. Some incorrect grasps (e.g. the can opener) may actually be viable while others (e.g. the bowl) are clearly not.

Predicting average grasps works well with certain types of objects, such as long, thin objects like markers or rolling pins. This model fails mainly in cases where average grasps do not translate to viable grasps on the object, for instance with circular objects like flying discs.



AncoraSIR.com

Experimental Results

Regression + Classification

We can extend our base detection model to simultaneously perform classification without sacrificing detection accuracy. Our model can correctly predict the category of an object it has previously seen 9 out of 10 times. When shown novel objects our model predicts the correct category more than 60 percent of the time. By comparison, predicting the most common class would give an accuracy of 17.7 percent.

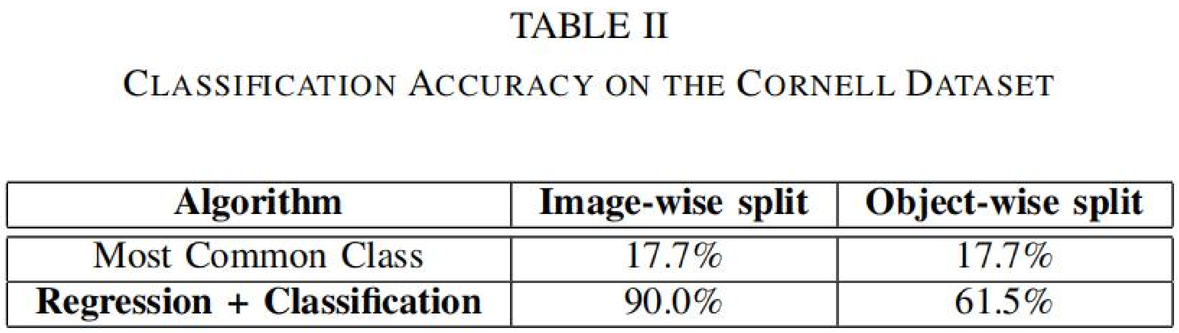


TABLE II

CLASSIFICATION ACCURACY ON THE CORNELL DATASET

Algorithm	**Image-wise split**	**Object-wise split**
Most Common Class	17.7%	17.7%
Regression + Classification	90.0%	61.5%

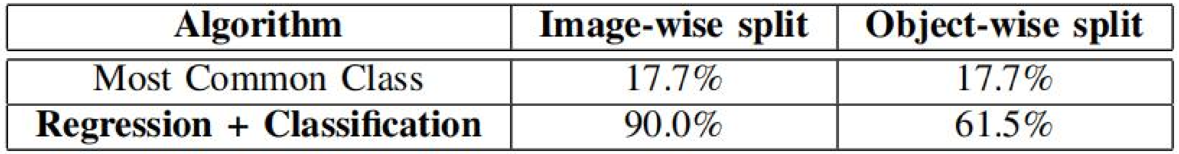
Experimental Results

Even with the added classification task the combined model maintains high detection accuracy. It has identical performance on the object-wise split and actually performs slightly better on the image-wise split.

This model establishes a strong baseline for combined grasp detection and object classification on the Cornell dataset.

TABLE II

CLASSIFICATION ACCURACY ON THE CORNELL DATASET



Algorithm	**Image-wise split**	**Object-wise split**
Most Common Class	17.7%	17.7%
Regression + Classification	90.0%	61.5%

Experimental Results

MultiGrasp

The MultiGrasp model outperforms our baseline direct regression model by a significant margin. For most objects. MultiGrasp gives very similar results to the direct regression model.

MultiGrasp has a very similar architecture to the direct regression model and operates at the same real-time speeds. With a grasp detection accuracy of 88 percent and a processing rate of 13 frames per second.

Experimental Results

MultiGrasp

The MultiGrasp model outperforms our baseline direct regression model by a significant margin. For most objects. MultiGrasp gives very similar results to the direct regression model.

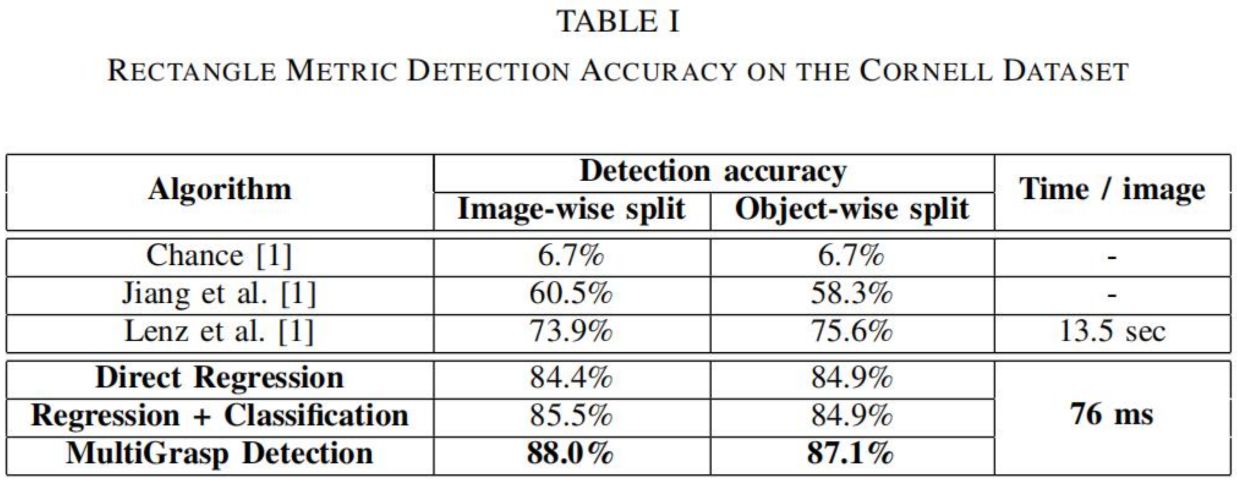


TABLE I

RECTANGLE METRIC DETECTION ACCURACY ON THE CORNELL DATASET

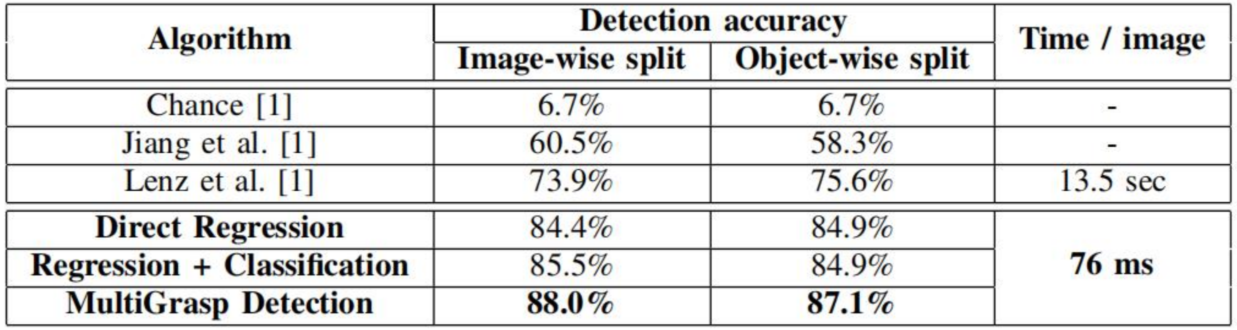
Algorithm	Detection accuracy		Time / image
	Image-wise split	Object-wise split	
Chance [1]	6.7%	6.7%	-
Jiang et al. [1]	60.5%	58.3%	-
Lenz et al. [1]	73.9%	75.6%	13.5 sec
Direct Regression	84.4%	84.9%	**76 ms**
Regression + Classification	85.5%	84.9%	
MultiGrasp Detection	**88.0%**	**87.1%**	

Discussion of Results

The paper present a fast, accurate system for predicting robotic grasps of objects in RGB-D images. Those models improve the state-of-the-art and run more than 150 times faster than previous methods. It show that grasp detection and object classification can be combined without sacrificing accuracy or performance. The MultiGrasp model gets the best known performance on the Cornell Grasping Dataset by combining global information with a local prediction procedure.

TABLE I

RECTANGLE METRIC DETECTION ACCURACY ON THE CORNELL DATASET

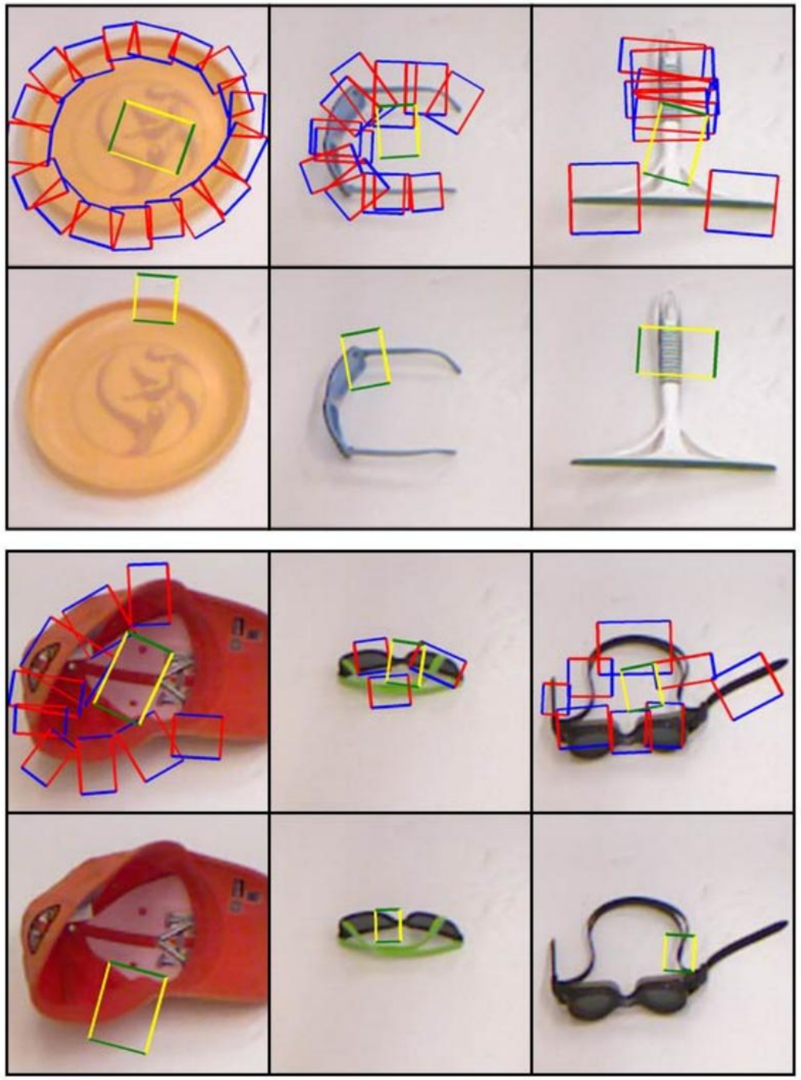


Algorithm	Detection accuracy		Time / image
	Image-wise split	Object-wise split	
Chance [1]	6.7%	6.7%	-
Jiang et al. [1]	60.5%	58.3%	-
Lenz et al. [1]	73.9%	75.6%	13.5 sec
Direct Regression	84.4%	84.9%	**76 ms**
Regression + Classification	85.5%	84.9%	
MultiGrasp Detection	**88.0%**	**87.1%**	

Limitations

The comparative performance of the direct regression model and MultiGrasp.

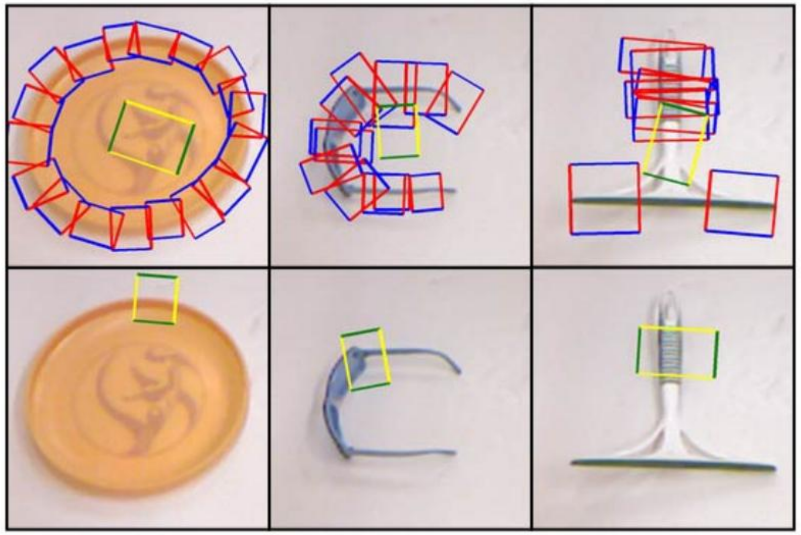
The top two rows show examples where direct regression model fails due to averaging effects but MultiGrasp predicts a viable grasp. The bottom two rows show examples where both models fail to predict good grasps. The ground truth grasps are shown in blue and red on the direct regression model images.



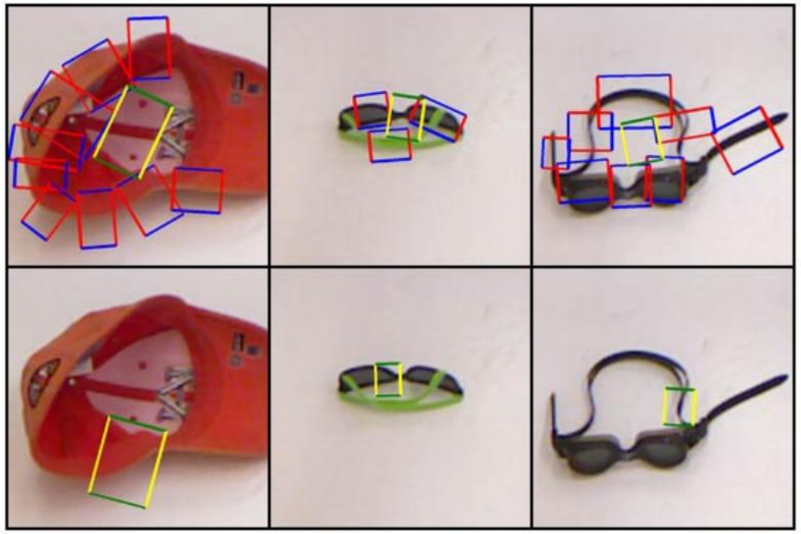
AncoraSIR.com

Limitations

Our direct regression model uses global information about the image to make its prediction, unlike sliding-window approaches. Sliding window classifiers only see small, local patches thus they can not effectively decide between good grasps and are more easily fooled by false positives.



Notably our direct regression model often tries to split the difference between a few good grasps and ends up with a bad grasp. A sliding window approach would never make the mistake of predicting a grasp in the center of a circular object like a flying disc.



Future Work for Paper / Reading

Our Ideas:

- It is claimed that direct regression model often tries to split the difference between a few good grasps and ends up with a bad grasp as a global model. What’s the way to solve this problem？
  - Is it possible to do so by spliting the difference between a few good grasps?
  - How can we figure out the best grasp between the few good grasps?
- What are the actual scenarios that can be applied to these models in real life? How?

Future Work for Paper / Reading

Others' ideas

- This study is based on 2D pictures. However, this means lacking of important shapes or contours of objects in real life. based on this research, we can expand our studies on single-view point cloud.

- using the method in real cases like Human-like Grasp Generation

- When it is in low light condition, the detection performance will be poorer. How to propose a visual enhancement guided grasp detection model to improve detection performance under this situation?

Extended Readings

- Z. Liu, Z. Chen and W. -S. Zheng, "Simulating Complete Points Representations for Single-View 6-DoF Grasp Detection," in IEEE Robotics and Automation Letters, vol. 9, no. 3, pp. 2901-2908, March 2024, doi: 10.1109/LRA.2024.3358757.
    - single-view point cloud grasping
    - https://ieeexplore.ieee.org/document/10414109

- R. Miura, K. Fujita and T. Tasaki, "Increasing the Graspable Objects by Controlling the Errors in the Grasping Points of a Suction Pad Unit and Selecting an Optimal Hand," 2024 IEEE/SICE International Symposium on System Integration (SII), Ha Long, Vietnam, 2024, pp. 196-201, doi: 10.1109/SII58957.2024.10417393.
    - improving the robustness toward errors in the estimation of grasping points
    - https://ieeexplore.ieee.org/document/10417393

- K. Yamamoto, H. Ito, H. Ichiwara, H. Mori and T. Ogata, "Real-Time Motion Generation and Data Augmentation for Grasping Moving Objects with Dynamic Speed and Position Changes," 2024 IEEE/SICE International Symposium on System Integration (SII), Ha Long, Vietnam, 2024, pp. 390-397, doi: 10.1109/SII58957.2024.10417201.
    - Grasping Moving Objects
    - https://ieeexplore.ieee.org/document/10417201

Extended Readings

- E. Balazadeh, M. T. Masouleh and A. Kalhor, "HUGGA: Human-like Grasp Generation with Gripper’s Approach State Using Deep Learning," 2023 11th RSI International Conference on Robotics and Mechatronics (ICRoM), Tehran, Iran, Islamic Republic of, 2023, pp. 854-860, doi: 10.1109/ICRoM60803.2023.10412422.
  - Human-like Grasp Generation based on the method
  - https://ieeexplore.ieee.org/document/10412422

- M. S. Sharif, A. Zorto, V. K. Brown and W. Elmedany, "Scalable Machine Learning Model for Highway CCTV Feed Real-Time Car Accident and Damage Detection," 2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Sakheer, Bahrain, 2023, pp. 341-348, doi: 10.1109/3ICT60104.2023.10391543.
  - employing computer vision algorithms to enhance real-time accident detection and response on highways
  - https://ieeexplore.ieee.org/document/10391543

- M. Niu, Z. Lu, L. Chen, J. Yang and C. Yang, "VERGNet: Visual Enhancement Guided Robotic Grasp Detection Under Low-Light Condition," in IEEE Robotics and Automation Letters, vol. 8, no. 12, pp. 8541-8548, Dec. 2023, doi: 10.1109/LRA.2023.3330664. keywords: {Feature extraction;Grasping;Visualization;Robot
  - proposing a visual enhancement guided grasp detection model to improve detection performance under low-light conditions
  - https://ieeexplore.ieee.org/document/10310093

Summary

Problem:

- The approach to robotic grasp detection so robot can safely pick up and hold an object, given its image.

Why is it important and hard：

- The needs for robots to interaction with physical world
- Scene understanding requires complex visual tasks

key limitation of prior work：

- run for a long time & low accuracy

What is the key insight(s) of the proposed work

- 3 methods based on picking ground truth grasp every time it sees an object to treat as the single ground truth grasp.

Results:

- Running less time with 76 milliseconds per batch, with accuracy up to 88%

AncoraSIR.com

Thanks for listening!

Presenter: 杨雪 洪雨盈 魏毓瞳 李蕴哲

20240312

AncoraSIR.com