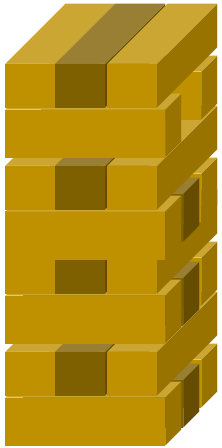
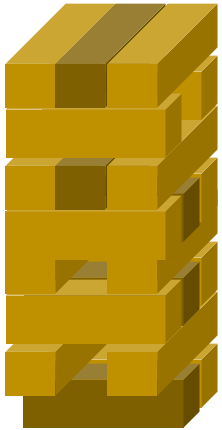
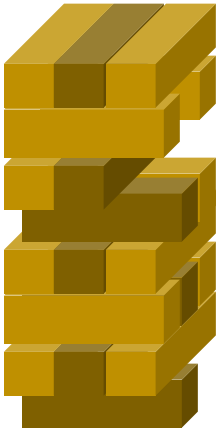
Visual Learning for Jenga Tower Stability Prediction

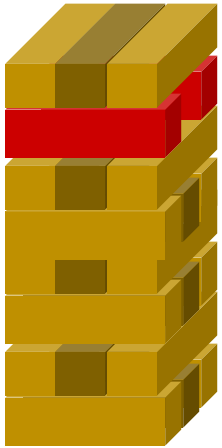
Aditya Agarwal (adityaa2)
Tanmay Agarwal (tanmaya)
Sarthak Ahuja (sarthaka)

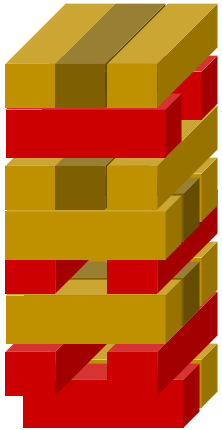


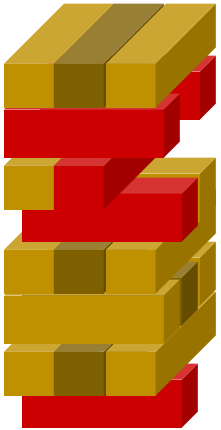




While playing a game of Jenga, which block will you choose to remove in each of above tower configurations?

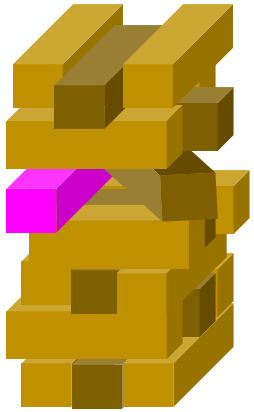




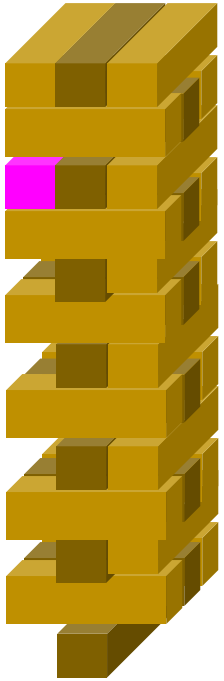


As humans, we have an intuitive understanding of the rules that dictate our physical world (physics)

Blocks under consideration



(a)



(b)

Though things are not so obvious or deterministic when

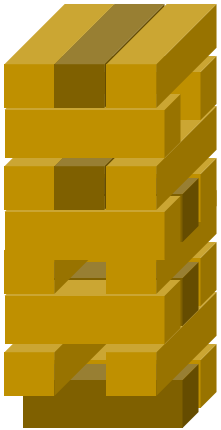
(a) there is noise in the structure, (b) physics becomes imperceivable to humans

The Problem

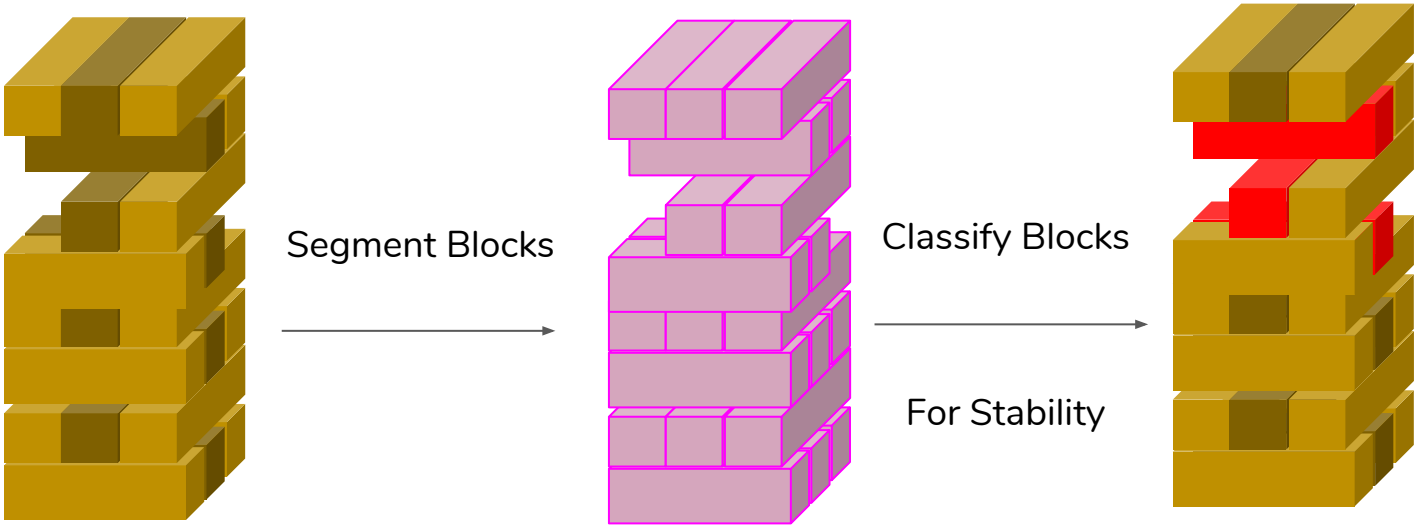
In this project we explore if we can enable a robot to learn this **physical intuition** needed to play a game of Jenga using **visual perception**

This super-power may lead to a variety of skills our robots are currently lacking -

1. rapid assessment of unfamiliar situations
2. dexterous manipulation of objects
3. creative use of tools



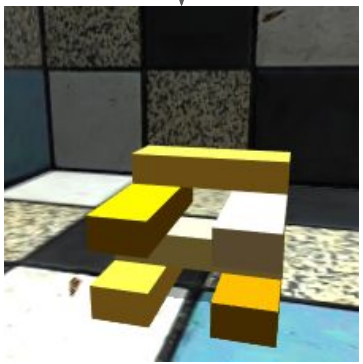
Approach



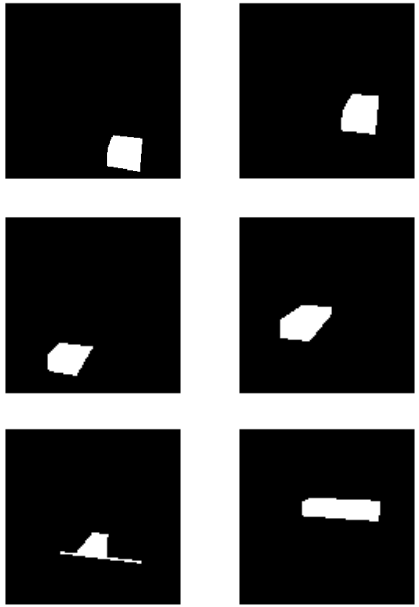
Data Generation

SEGMENTATION NETWORK

RANDOMLY GENERATE JENGA STABLE CONFIGURATION



4 <= HEIGHT <= 8
16 CAMERA ANGLES



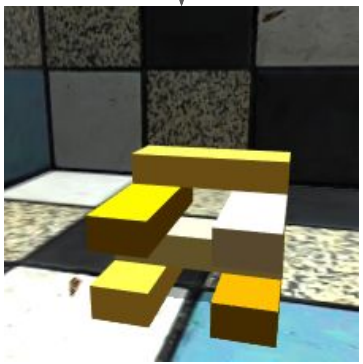
Segmentation Masks

Data Generation

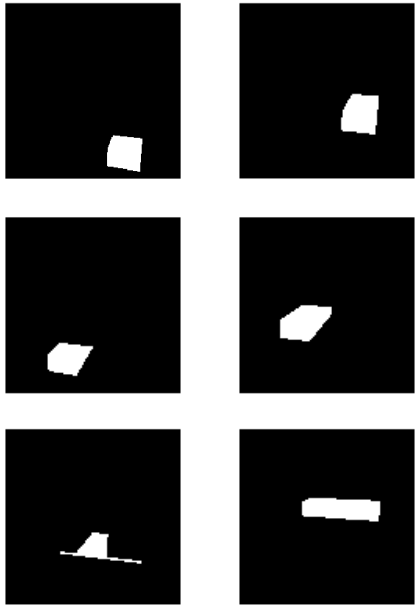
SEGMENTATION NETWORK

STABILITY NETWORK

RANDOMLY GENERATE JENGA STABLE CONFIGURATION



4 <= HEIGHT <= 8
16 CAMERA ANGLES

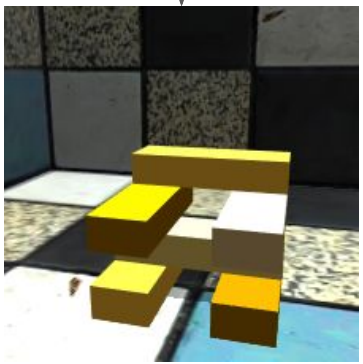


Segmentation Masks

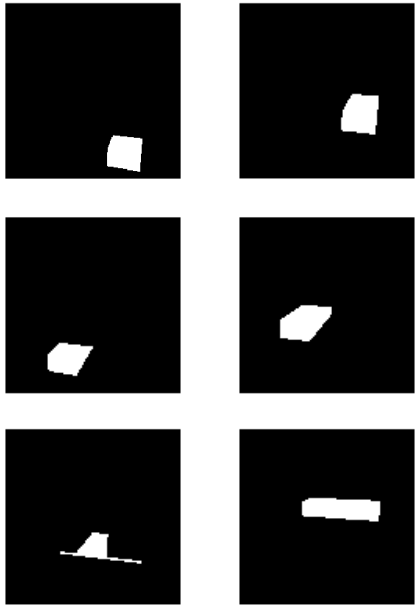
RANDOMLY SELECT BLOCK TO BE REMOVED

Data Generation

RANDOMLY GENERATE JENGA STABLE CONFIGURATION



4 <= HEIGHT <= 8
16 CAMERA ANGLES



Segmentation Masks

RANDOMLY SELECT BLOCK TO BE REMOVED

GENERATE TRUE LABEL AFTER REMOVAL BY RUNNING AN ACTUAL SIMULATION

Data Generation

	Positive Scenarios	Negative Scenarios
Training Data	**3426**	**3426**
Validation Data	**856**	**856**
Testing Data	**1066**	**1434**

4 <= HEIGHT <= 8
16 CAMERA ANGLES

NETWORK ARCHITECTURE

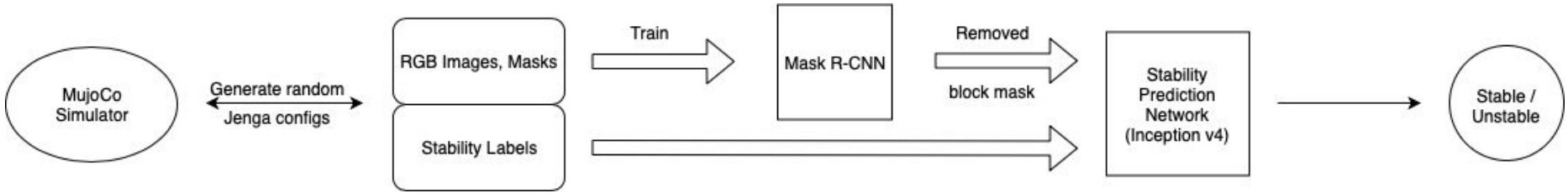
$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}}$$

$$\mathcal{L}(\{p_i\}, \{t_i\}) = \frac{1}{N_{\text{cls}}} \sum_i \mathcal{L}_{\text{cls}}(p_i, p_i^*) + \frac{\lambda}{N_{\text{box}}} \sum_i p_i^* \cdot L_1^{\text{smooth}}(t_i - t_i^*)$$

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}}$$

$$\mathcal{L}_{\text{cls}}(p_i, p_i^*) = -p_i^* \log p_i - (1 - p_i^*) \log(1 - p_i)$$

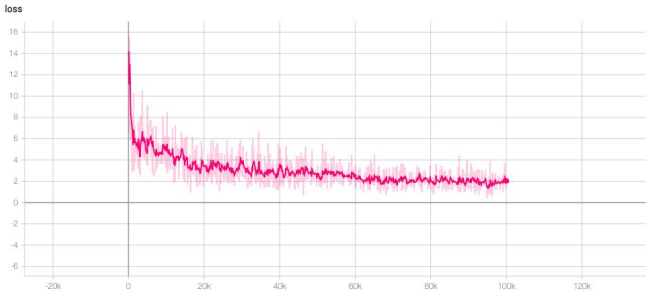
$$\mathcal{L}_{\text{mask}} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} \left[y_{ij} \log y_{ij}^k + (1 - y_{ij}) \log(1 - y_{ij}^k) \right]$$



$$H_p(q) = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

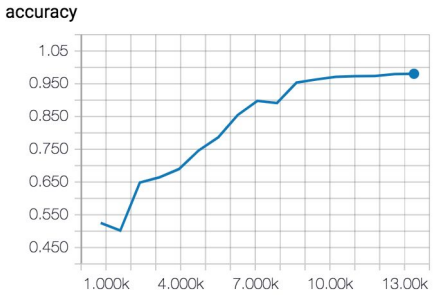
Quantitative Analysis: Test Set

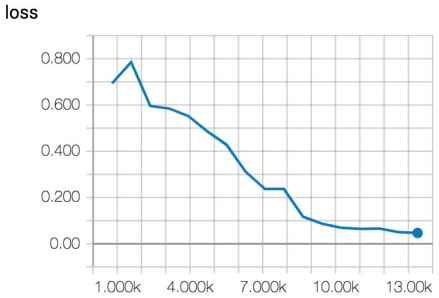
Mask-RCNN



	Test mAP
BBOX	0.4388
Segmentation	0.4254

Stability Network

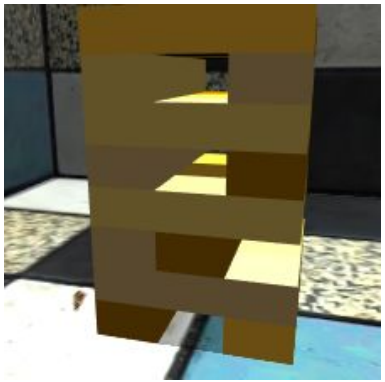




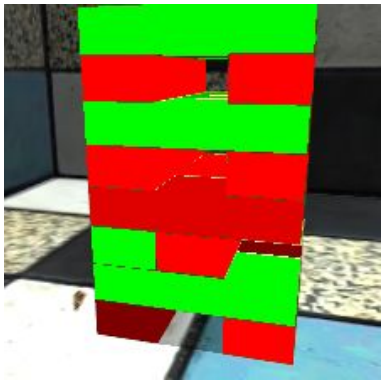
	Accuracy
Validation Data	0.98
Testing Data	0.97

Qualitative Analysis : Test Set

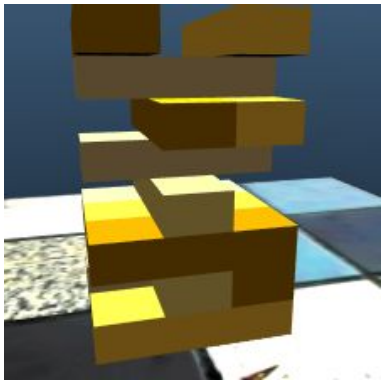
Input Image



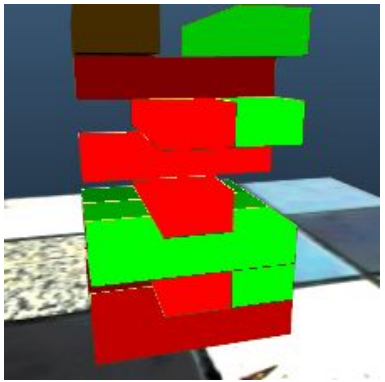
Stability prediction



Input Image



Stability prediction



- Notice that the blocks at all levels are classified accurately for stability across multiple Jenga configurations

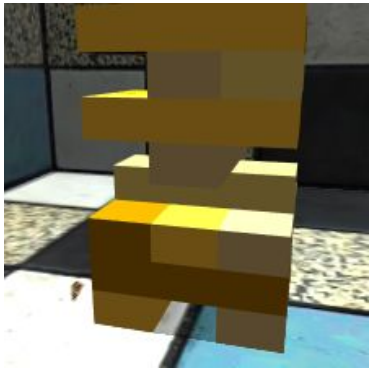
GOOD, to remove

BAD, to remove

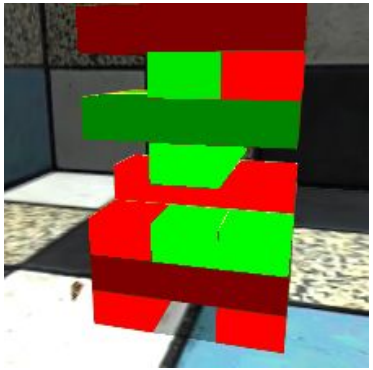
Qualitative Analysis : Test Set (continued)

Camera Angle 1

Input Image

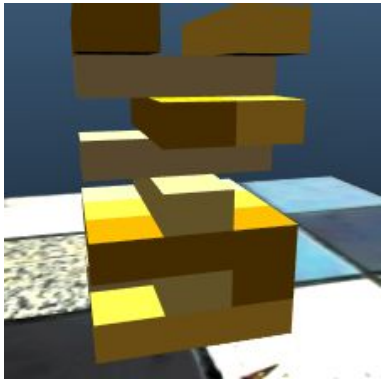


Stability prediction

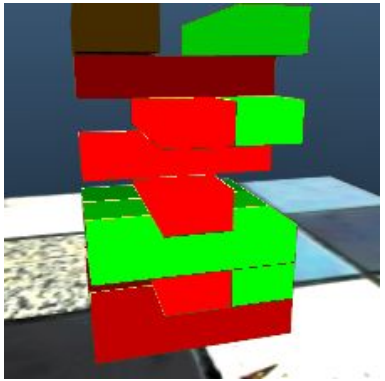


Camera Angle 2

Input Image



Stability prediction

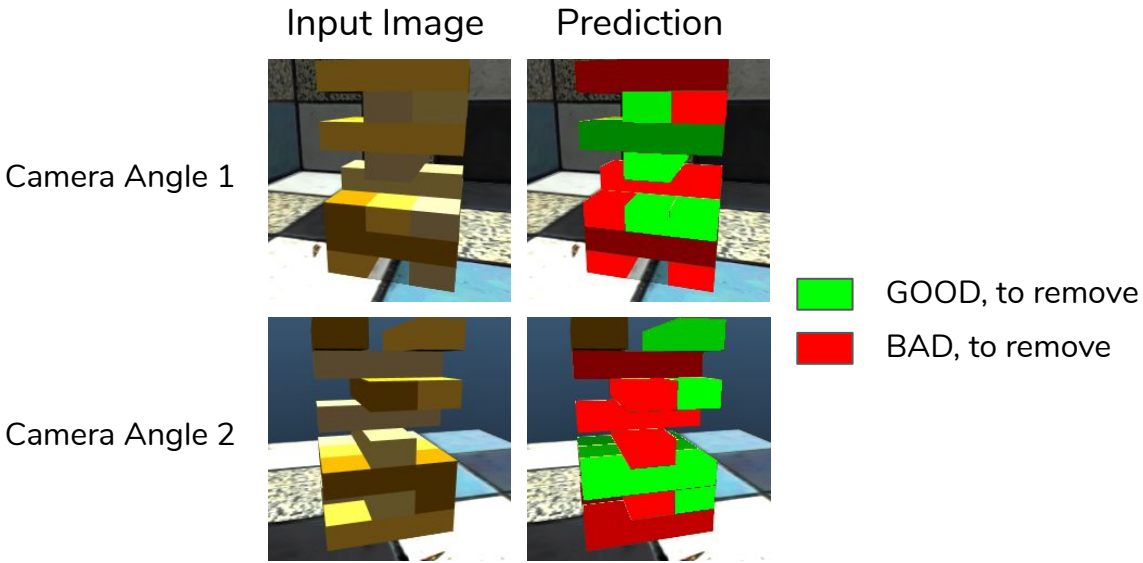


- For same configuration and same blocks, predictions can differ across camera angles
- <u>Enhancement :</u> Combine predictions from multiple angles

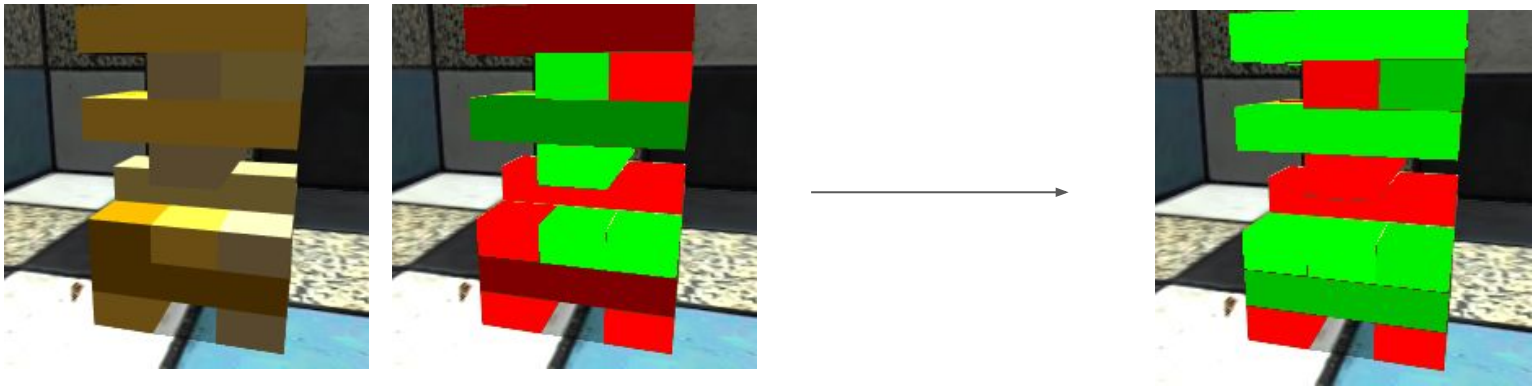
GOOD, to remove

BAD, to remove

Qualitative Analysis : Test Set (continued)



Enhancement : Combining Camera Inputs



- Perform Majority Voting Across Camera Angles

GOOD, to remove

BAD, to remove

Qualitative Analysis : Complete Pipeline

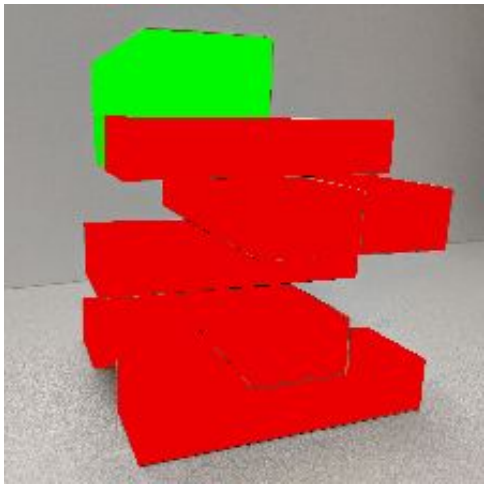
- Block Masks detected with high confidence
- Some blocks missed from detection
- Stability classification done accurately for detected blocks
- Enhancement : Train Mask-RCNN further with higher learning rate *

GOOD, to remove

BAD, to remove

Fazeli, N., Oller, M., Wu, J., Wu, Z., Tenenbaum, J. B., & Rodriguez, A. (2019). *See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion*. Retrieved from http://robotics.sciencemag.org

Qualitative Analysis : Real World Images



GOOD, to remove

BAD, to remove

- Top most blocks accurately classified for stability
- Accuracy decreases for blocks in between
- <u>Enhancement :</u> Improve quality of synthetic data for better transference to real world *

GOOD, to remove

BAD, to remove

Fazeli, N., Oller, M., Wu, J., Wu, Z., Tenenbaum, J. B., & Rodriguez, A. (2019). *See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion.* Retrieved from http://robotics.sciencemag.org