International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Enhanced Robotic Grasp Detection Through Advanced Resnet Variants in Deep Convolutional Neural Networks

Tanya Thukral

Department of Artificial Intelligence and Data Science, Indira Gandhi Delhi Technical University for Women, Delhi

ABSTRACT

The advent of deep learning has propelled significant advancements in the field of robotic manipulation, enabling robots to perform complex grasping tasks with unprecedented accuracy. This study introduces an innovative approach to robotic grasp detection leveraging advanced variants of the Residual Network (ResNet) architecture within Deep Convolutional Neural Networks (DCNNs). Our methodology extends the conventional ResNet framework by integrating novel structural modifications and optimization techniques aimed at enhancing feature representation and learning efficiency. By incorporating dilated convolutions, attention mechanisms, and depth-wise separable convolutions, our proposed ResNet variants significantly improve the grasp detection performance in terms of accuracy, speed, and robustness across diverse object categories and orientations. The effectiveness of our approach is validated through extensive experiments on standard robotic grasp datasets, where our models demonstrate superior performance compared to existing state-of-the-art grasp detection methodologies. Furthermore, our analysis reveals that the advanced ResNet variants are capable of learning more discriminative features from grasp imagery, facilitating more precise localization and orientation prediction for robotic grippers. Additionally, the computational efficiency of our models enables real-time grasp detection, which is critical for practical robotic applications. Advanced techniques in deep learning have significantly propelled the fields of computer vision and natural language processing forward. Despite notable achievements, the integration of deep learning into robotics applications remains relatively limited. This study introduces a novel approach to robotic grasp detection, focusing on predicting the optimal grasping pose for a parallel-plate robotic gripper when encountering unfamiliar objects, utilizing RGB-D imagery of the scene.

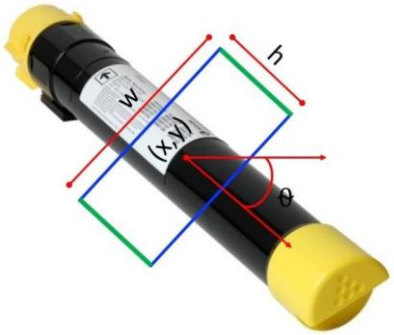
1. INTRODUCTION

Robotic grasping remains a challenging task, significantly trailing human capabilities in this domain. Despite extensive research efforts focused on robotic grasping and manipulation, the quest for real-time grasp detection and planning persists as a formidable challenge. Even with the latest advancements in grasp detection techniques, achieving real-time detection proves elusive. The problem of robotic grasping can be dissected into three sequential phases: grasp detection, trajectory planning, and execution. Grasp detection entails visual recognition, wherein the robot employs sensors, such as 3-D vision systems or RGB-D cameras, to identify graspable objects in its surroundings. The primary objective is to anticipate potential grasps from sensor data and translate pixel values into real-world coordinates. This step is crucial

as subsequent actions heavily rely on the accuracy of these coordinates. The derived
real-world coordinates are then converted into position and orientation for the robot's end-of-arm tooling (EOAT). Subsequently, an optimal trajectory
for the robotic arm is planned to reach the desired grasp position. The planned trajectory is then executed by the robotic arm, utilizing either an open-loop or closed-loop controller. Unlike an open-loop controller, a closed-loop controller continuously receives feedback from the vision system throughout the grasping task. However, incorporating this feedback entails additional computational processing, potentially impacting task speed.

In this study, we address the challenge of identifying optimal grasps from RGB-D imagery. Fig. 1 illustrates a simplified five-dimensional grasp representation for a potentially favorable grasp of a toner cartridge. This representation captures the position and orientation of a parallel plate gripper before executing the grasp on an object. While Jiang et al. [2] introduced a more comprehensive seven-dimensional grasp representation, Lenz et al.
[4] demonstrated that a refined five-dimensional representation can be projected back to a seven-dimensional grasp format suitable for robotic grasp execution . Besides its computational efficiency, this dimensionality reduction enables grasp detection using RGB-D images. Here, we leverage this five-dimensional grasp representation for grasp prediction.

Figure 1.



Our novel approach focuses on detecting effective robotic grasps for parallel plate grippers utilizing the five-dimensional representation. In this paper, we in detail produce the results of a uni-modal grasp where only 3 channels, i.e. RGB or RGD are used to determine a good grasp using RESNET-152 and FC layers. In extension we theoretically discuss
multi-modal grasp where we employ two parallel 50-layer deep convolutional residual neural networks to extract features from RGB-D images.

One network analyzes the RGB component while the other examines the depth channel. The outputs of these networks are combined and fed into another convolutional network to predict grasp configurations. We compare our approach with existing methods in the literature, including a
uni-modal variant of our model that solely utilizes the RGB component. Our experiments are conducted on the established Cornell Grasp Dataset, with sample images displayed in Fig. 2. Results demonstrate that our proposed architecture surpasses current state-of-the-art methods in terms of both accuracy and speed.

Figure 2.

2. BACKGROUND AND RELATED WORKS

Deep learning has made significant strides in advancing computer vision and natural language processing tasks. These achievements have spurred interest among robotics researchers to explore the potential applications of deep learning in addressing complex challenges in robotics. For instance, there has been a shift towards utilizing deep learning features for robot localization instead of hand-engineered features. Deep reinforcement learning is also gaining traction for end-to-end training in robotic arm control tasks. Furthermore, deep learning techniques have led to breakthroughs in multi-view object recognition through camera control, learning dual-arm manipulation tasks, and estimating affordances for autonomous driving.

A notable obstacle in applying deep learning to robotics lies in the requirement for large volumes of labeled training data, which are often lacking in many robotics domains. To mitigate this challenge, transfer learning methods are commonly employed in computer vision. This involves pre-training deep convolutional neural networks on large datasets such as ImageNet, containing 1.2 million images across 1000 categories, before fine-tuning the network on the target dataset. These pre-trained models serve as either initialization or fixed feature extractors for the specific task at hand.

In the realm of robotic grasp prediction, a prevalent approach involves employing a sliding window detection framework. In this framework, a classifier is utilized to assess whether small patches of the input image indicate a viable grasp for an object.

Multiple patches across the image are evaluated, and those with high scores are considered as potential grasps. While effective, this approach suffers from computational inefficiency. For instance, Lenz et al. achieved a 75% accuracy using convolutional neural networks as classifiers, but their method operated at a slow speed of 13.5 seconds per frame, rendering it impractical for

real-time applications. Efforts have been made to accelerate this method, such as passing the entire image through the network at once instead of processing multiple patches.

Considerable research effort has been devoted to leveraging 3-D simulations for identifying effective grasps. While these techniques are potent, many rely on a known 3-D model of the target object to determine an appropriate grasp. However, for versatile robotic applications, the ability to grasp unfamiliar objects without prior knowledge of their 3-D structure is essential. Although Jincheng et al. demonstrated the potential of deep learning for 3-D object recognition and pose estimation, their experiments were limited to a small set of objects, and their algorithm proved computationally intensive. More recently, Mahler et al. adopted a cloud-based robotics approach to significantly reduce the number of samples required for robust grasp planning. Additionally, Johns et al. generated training data using physics simulations and depth image simulations with 3-D object meshes to develop a grasp scoring method

resilient to gripper pose uncertainty.

Jeremy et al. proposed a grasp point detection technique with a commendable precision of 92%, albeit limited to cloth towels and unsuitable as a general-purpose grasp detection method. Another approach for grasp pose detection was introduced by Gualtieri et al., tailored for removing objects from densely clustered environments. However, their evaluation was confined to a small set of objects using a research robot.

In contrast to previous works employing AlexNet for feature extraction, as seen in prior studies, we utilize ResNet-152, the current state-of-the-art deep convolutional neural network. Additionally, we introduce a multi-modal model that leverages features from both RGB and Depth images to predict grasp configurations.

3. PROBLEM FORMULATION

The task of robotic grasp detection involves determining a suitable grasp configuration g for a given object image I. A grasp configuration g is represented in five dimensions as:

$$g = f(x, y, h, w, \theta) \quad (1)$$

where (x, y) represents the center of the grasp rectangle, h denotes the height of the parallel plates, w indicates the maximum distance between the parallel plates, and θ signifies the orientation of the grasp rectangle relative to the horizontal axis. Typically, h and w remain fixed parameters for a specific robot's end-of-arm tooling (EOAT). Fig. 1 illustrates this representation.

Our focus lies on planar grasps, as highlighted by Lenz et al., who revealed the possibility of projecting a five-dimensional grasp configuration back to a seven-dimensional configuration for practical implementation on a real robot. To tackle this grasp detection challenge, we adopt a unique approach, elaborated in section IV.

4. APPROACH

Deep Convolutional Neural Networks (DCNNs) have proven to be highly effective in solving detection and classification tasks within computer vision, surpassing previous state-of-the-art techniques. In our study, we utilize DCNNs to both detect target objects within images and predict optimal grasp configurations. Unlike previous methods which relied on a two-step approach involving repetitive classification on small image patches, we propose a more streamlined single-step prediction technique. By feeding the entire RGB-D image directly into the DCNN, we are able to make grasp predictions efficiently, reducing computational overhead.

In theory, deeper DCNN architectures should offer superior performance due to their increased representational capacity. However, conventional optimization methods like stochastic gradient descent (SGD) face challenges when training ultra-deep networks. Contrary to theoretical
expectations, deeper networks often exhibit higher training errors, indicating the difficulty in optimizing them effectively using SGD. To address this issue, we adopt residual layers inspired by ResNet, which redefine the mapping function between layers, aiding in more efficient training.

Similar to previous studies, we assume that each input image contains only one graspable object, and our goal is to predict a single grasp configuration for that object. This assumption allows us to make global grasp predictions by analyzing the entire image. However, it's important to acknowledge that this assumption may not always hold true, especially in scenarios beyond controlled experimental conditions. In such cases, a model would need to first segment the image into regions, each containing only one object,

before making grasp predictions.

5. ARCHITECTURE

Our model represents a significant departure from previous approaches (e.g., [4], [5], [30]) by employing a notably deeper architecture. Rather than relying on the eight-layer AlexNet, we opt for ResNet-152, a sophisticated deep residual model, to tackle the grasp detection problem. ResNet introduces the notion of residual learning to overcome the challenge of learning identity mapping. By integrating skip connections that circumvent certain layers, each giving rise to a residual block, ResNet adapts the conventional feed-forward CNN framework. Within each residual block, convolution layers predict a residual that is subsequently added to the input of the block. The fundamental idea is to permit only the identity of the input feature to traverse the skip connection, while bypassing the convolution and non-linear activation layers in the kth residual block. Illustrated in Figure 3 is an example of a residual block with skip connections. Formally, the residual block is expressed as:

$$Hk = F(Hk-1, Wk) + Hk-1$$

where, $Hk-1$ is the input to the residual block, Hk is the output of the block, and Wk are the weights learned for the mapping of function F. To gain a deeper understanding of the ResNet architecture, readers are advised to consult [31].

We introduce two distinct architectures for robotic grasp prediction: the uni-modal grasp predictor and the multi-modal grasp predictor. The uni-modal grasp predictor is a 2D grasp predictor that exclusively utilizes single-modality information (e.g., RGB) from the input image to anticipate the grasp configuration. Conversely, the multi-modal grasp predictor is a 3-D Grasp Predictor that harnesses multi-modal data (e.g., RGB and Depth). In the ensuing subsections, we delve into the intricacies of these two architectures.

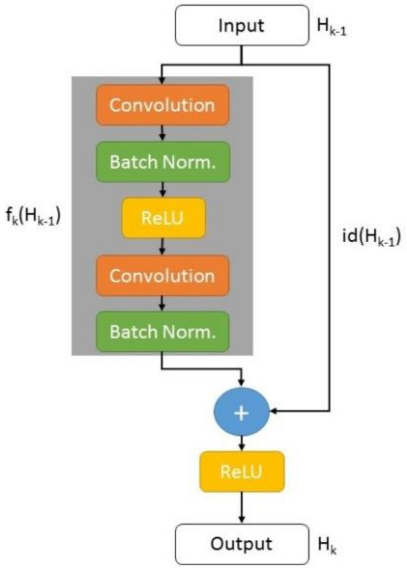


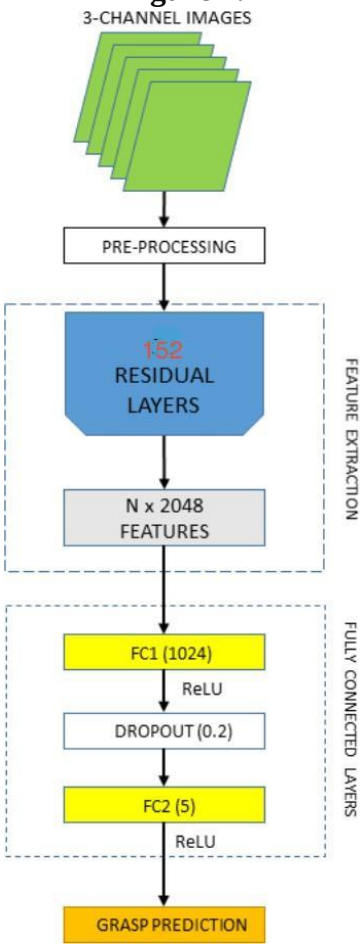
Figure 3.

Uni-Modal Grasp Detector

Large-scale image classification datasets are typically limited to RGB images, leading us to pre-train our deep convolutional neural networks using only three channels. Our uni-modal grasp predictor model is specifically engineered to discern grasps utilizing just three channels (RGB or RGD) from the raw image data. Figure 4 illustrates the intricate architecture of our uni-modal grasp predictor. To extract features from the RGB channels of the image, we employ a pre-trained ResNet-152 model from

ImageNet. As a baseline, we utilize a linear SVM classifier to predict the grasp configuration based on the features extracted from the last hidden layer of ResNet-152. Within our uni-modal grasp predictor, we replace the last fully connected layer of ResNet-152 with two fully
connected layers featuring rectified linear unit (ReLU) activation functions. Additionally, a dropout layer is introduced after the first fully connected layer to mitigate overfitting. Training optimization is accomplished through SGD, utilizing mean squared error (MSE) as our loss function.
The 3-channel image is then fed into the uni-modal grasp predictor, leveraging residual convolutional layers to extract features from the input image. The final fully connected layer serves as the output layer, predicting the grasp configuration for the object depicted in the image. During training, the weights of convolutional layers in ResNet-152 remain fixed, while only the weights of the last two fully connected layers are fine-tuned. We initialize the weights of these last two layers using the Xavier weight initialization technique.

Figure 4.



Multi-Modal Grasp Detector

We propose the introduction of a multi-modal grasp predictor, drawing inspiration from the RGB-D object recognition approach pioneered by Schwarz et al. [32]. This innovative approach utilizes multi-modal (RGB-D) information extracted from raw images to predict grasp configurations. To facilitate this, the raw RGB-D images are transformed into two distinct images: a conventional RGB image and a depth image converted into a 3-channel format, akin to grayscale to RGB conversion. Subsequently, these two 3-

channel images are separately inputted into independent pre-trained ResNet-152 models, where the ResNet-152 layers function as feature extractors for both images. Analogous to the uni-modal grasp predictor, features are extracted from the penultimate layer of both ResNet-152 networks.

These extracted features are then normalized using L2-normalization and concatenated, forming input for a shallow convolutional neural network consisting of three fully connected layers with ReLU activation functions. To combat overfitting, dropout layers are incorporated after the first and second fully connected layers. Similar to the uni-modal model, optimization is achieved through SGD, employing MSE as the loss function. Figure 5 illustrates the comprehensive architecture of our proposed multi-modal grasp predictor.

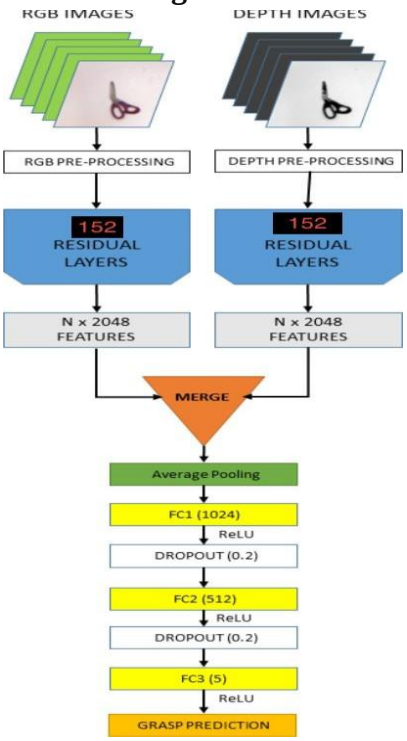
By employing two DCNNs in parallel, our model efficiently extracts features from both RGB and depth images, allowing it to learn multimodal features from RGB-D datasets. The weights of the two DCNNs are initialized using pre-trained

ResNet-50 models, while the weights of the shallow network are initialized using Xavier weight initialization. During training, these weights are

fine-tuned to enhance performance.

Additionally, while we propose these advancements, it's important to note that the implementation of these methodologies will be pursued in future work due to time constraints. We acknowledge that further exploration and experimentation are needed to validate the effectiveness of these proposed approaches.

Figure 5.



6. EXPERIMENTS

DATASETS

For evaluating our architecture and comparing it with existing methods, we conducted experiments using the widely used Cornell Grasp Dataset, comprising 885 images featuring 240 distinct objects. Each image is annotated with multiple grasp rectangles, categorized as either successful (positive) or failed (negative),

specifically tailored for parallel plate grippers. In total, the dataset contains 8019 labeled grasps, including 5110 positive and 2909 negative instances. Figure 6 provides a visual representation of the ground truth grasps using the rectangular metric for this dataset.

Figure 6.

Consistent with prior research, we employed five-fold cross-validation for all our experiments. The dataset was partitioned in two distinct manners:

1. **Image-wise split:** In this approach, all the images in the dataset were randomly divided into five folds. This methodology allows us to assess the network's ability to generalize to objects encountered in different positions and orientations.
2. **Object-wise split:** Here, the object instances were randomly divided into folds, ensuring that all images of a particular object were placed within the same validation set. This enables us to evaluate the network's generalization capability to objects it has not encountered previously.

DATA PREPROCESSING

Before inputting the data into the DCNN, we conducted a limited amount of data pre-processing. The input to our DCNN consists of a patch centered around the grasp point, which is extracted from a training image. This patch is resized to dimensions of 224 × 224, aligning with the input image size required by the ResNet-152 model. Additionally, the depth image undergoes rescaling to ensure itsvalues fall within the range of 0 to 255. Notably, certain pixels in the depth image may contain NaN (Not a Number) values, typically occurring due to occlusion in the original stereo image. To address this, we replaced these NaN-valued pixels with zeros to facilitate further analysis.

PRE-TRAINING

In scenarios where domain-specific data is scarce, as observed in the Cornell grasp dataset, pre-training becomes imperative. Hence, we initiate pre-training of the ResNet-152 model on ImageNet. Our assumption is that the majority of the learned filters are not exclusively tailored to the ImageNet dataset, with only the upper layers demonstrating specificity towards classifying 1000 categories.

Throughout this pre-training phase, the DCNN acquires universal visual features by fine-tuning millions of parameters. Subsequently, we extract features fromI the last layer and feed them into our shallow convolutional neural network. It's noteworthy that the ImageNet dataset exclusively comprises RGB images, thereby leading the DCNN to primarily learn RGB features.

TRAINING

To train and validate our models, we utilized the Keras deep learning library, a Python-based framework built on top of Theano. The experiments were conducted using a CUDA-enabled NVIDIA GeForce GTX 645 GPU paired with an Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz. While GPUs are not yet standard components of robotic systems, their popularity is on the rise in vision-based robotic applications due to their enhanced computational capabilities.

Our training process was bifurcated into two stages: initially, only the shallow network underwent training, followed by training the complete network end-to-end. For training our uni-modal grasp predictor, we employed SGD optimization with the following hyperparameters during the first stage: learning rate = 0.001, decay = 1e-6, momentum = 0.9, mini-batch size = 32, and a maximum of 30 epochs. Conversely, for the multi-modal grasp predictor, the following hyperparameters were utilized during the first stage: learning rate = 0.0006, decay = 1e-6, momentum = 0.9, mini-batch size = 32, and a maximum of 50 epochs. During the second phase of fine-tuning the network, we adopted a significantly lower learning rate and stabilized it if the training loss failed to decrease.

EVALUATION

Previous studies have employed two distinct performance metrics to evaluate grasps on the Cornell grasp dataset: the rectangle metric and the point metric. The point grasp metric involves comparing the distance between the center point of the predicted grasp and the center points of all ground truth grasps. However, the specific threshold values used to determine grasp success were not disclosed in prior research. Additionally, this metric overlooks the grasp angle, which is a crucial parameter in grasping tasks.

On the other hand, the rectangle grasp metric evaluates the complete grasp rectangle, deeming a grasp to be successful if the disparity between the predicted grasp angle and the ground truth grasp angle is less than 30 degrees, and the Jaccard similarity coefficient of the predicted grasp and ground truth grasp exceeds 25%. The Jaccard similarity coefficient, also known as the Jaccard index, quantifies the similarity between the predicted grasp and the ground truth grasp. It is calculated as:

$$J(\hat{\theta}, \theta) = \frac{|\hat{\theta} \cap \theta|}{|\hat{\theta} \cup \theta|}$$

Where θ represents the predicted grasp and θ denotes the ground truth grasp. Since the rectangle metric demonstrates better discrimination between 'good' and 'bad' grasps, we opt to utilize this metric for our experiments. Across all our models, we select the best-scoring grasp rectangle using the rectangle metric for grasp prediction.

7. RESULTS

Table 1 provides a comparative analysis of our findings against previous research regarding the accuracy of grasp detection using the rectangle metric on the Cornell RGB-D grasp dataset. Across the spectrum, both of our models exhibit superior performance compared to existing robotic grasp detection algorithms, excelling in both accuracy and speed. Notably, the results from previous studies are based on their self-

reported scores. To evaluate the generalization capability of our network, assessments were conducted using both image-wise split and object-wise split methodologies.

Author	Algorithm	Accuracy %	
		Image-Wise Split	Object- Wise Split
Jiang et.al.	Fast Search	60.5	58.3
Lenz et.al.	SAE, struct. reg. two stage	73.9	75.6
Redmon et.al.	AlexNet, MultiGrasp	88	87.1
Wang et.al.	Two-Stage Closed Loop	85.3	
Asif et.al.	STEM-CaRFs	88.2	87.5
Ours	**Uni-Modal Grasp Predictor**		
	RESNET-152 SVM (Baseline)	85.23	84.77
	RESNET-152 ReLu	89.26	88.67
	Multi-Modal Grasp Predictor		
	RESNET-152 linear-SVM (Baseline)	88.72	88.14
	RESNET-152 ReLu RGB-D	91.83	91.43

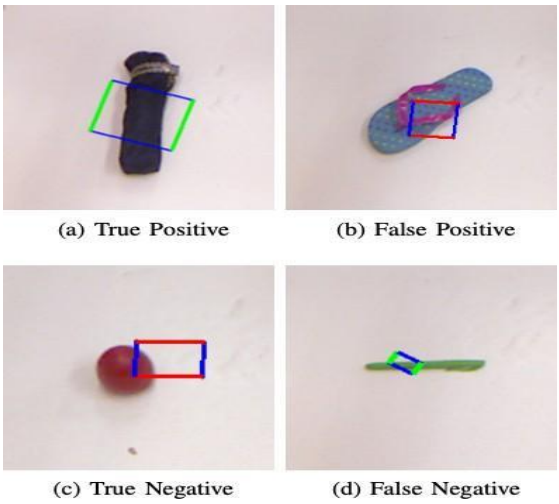
Table 1

We present the outcomes of various iterations of uni-modal and multi-modal grasp predictors achieved by altering the information inputted into the channels. In the RGB version of the uni-modal grasp predictor, solely the RGB channels of the input image are employed. The baseline uni-modal grasp predictor model attained an accuracy of 85.23%.Subsequently, our uni-modal grasp predictor utilizing RGB data achieved an accuracy of 89.26%. The baseline multi-modal grasp predictor, leveraging RGB-D data, achieved an accuracy of 88.72%, establishing a new performance baseline for RGB-D robotic grasp detection. Surpassing this, our multi-modal grasp predictor achieved an accuracy of 91.83%, thereby setting a new state-of-the-art performance for RGB-D robotic grasp detection. Figure 7 presents an accuracy comparison of all proposed models in this study using 5-fold cross-validation. Overall, our multi-modal grasp predictor demonstrated superior performance with the Cornell grasp dataset.

Figure 7 showcases examples of predicted graspability using the modified multi-modal grasp predictor. A green box indicates a successful predicted grasp, while a red box indicates an unsuccessful prediction. The instances of false negatives depicted in Figure 7b and 7d represent incorrect predictions. In Figure 7b, we attribute the model's failure to understand the depth features of the slipper strap, which are crucial for the grippers to grasp the slipper. Conversely, in Figure 7d, the model struggled to comprehend the orientation of the grasp rectangle with respect to the object.

Despite encountering challenging instances like these, the model demonstrated high accuracy in predicting the graspability of various object types.

Figure 7.



(a) True Positive (b) False Positive

(c) True Negative (d) False Negative

8. CONCLUSION

This paper introduces a novel multi-modal robotic grasp detection system designed to predict the graspability of unfamiliar objects for a parallel plate robotic gripper using RGB-D images. Additionally, we propose a uni-modal model utilizing RGB data exclusively. Our study showcases the efficacy of employing DCNNs in parallel to extract features from multi-modal inputs, enabling accurate grasp configuration prediction for objects. The integration of deep residual layers has been instrumental in extracting enhanced features from input images, subsequently utilized by the fully connected layers to output grasp configurations. Notably, our models have significantly improved upon the state-of-the-art performance on the Cornell Grasping Dataset while operating at real-time speeds.

Moving forward, we aim to explore transfer learning concepts to leverage the trained model on the grasp dataset for executing grasps using a physical robot. Furthermore, within an industrial context, we anticipate achieving even higher detection accuracy, thereby rendering grasp detection for pick and place tasks robust across diverse shapes and sizes of parts.

9. REFERENCES

1. A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, 2008.
2. Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pp. 3304–3311, 2011.
3. M. Ciocarlie, K. Hsiao, E. G. Jones, S. Chitta, R. B. Rusu, and I. A. S ¸ucan, "Towards reliable grasping and manipulation in household environments," in *Experimental Robotics*, pp. 241–252, 2014.
4. I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," The International Journal of Robotics Research, vol. 34, no. 4- 5, pp. 705–724,2015.
5. J. Redmon and A. Angelova, "Real-time grasp detection using convo- lutional neural networks," in 2015 IEEE International Conference on Robotics and Automation (ICRA), pp. 1316–1322, May 2012 Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 05 2015.
6. A.Krizhevsky,I.Sutskever,andG.E.Hinton,"Imagenetcl assification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, pp. 1097–1105, 2012.

7. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.
8. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.
9. I.Sutskever,O.Vinyals,andQ.V.Le,"Sequencetosequen celearning with neural networks," in Advances in neural information processing systems, pp. 3104–3112, 2014.
10. K. Cho, B. Van Merrie ̈nboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.
11. R.Socher,E.H.Huang,J.Pennin,C.D.Manning,andA.Y.N g,"Dy- namic pooling and unfolding recursive autoencoders for paraphrase detection," in Advances in Neural Information Processing Systems, pp. 801–809, 2011.
12. E. Johns and G.-Z. Yang, "Generative methods for long-term place recognition in dynamic scenes," International Journal of Computer Vision, vol. 106, no. 3, pp. 297–314, 2014
13. N. Su ̈nderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on, pp. 4297–4304, 2015.
14. S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," Journal of Machine Learning Research, vol. 17, no. 39, pp. 1–40, 2016.
15. E. Johns, S. Leutenegger, and A. J. Davison, "Pairwise decomposition of image sequences for active multi-view recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3813–3822, June 2016.
16. S. Kumra and F. Sahin, "Dual flexible 7 dof arm robot learns like a child to dance using q-learning," in *IEEE System of ystems Engineering Conference (SoSE)*, pp. 292–297, 2015
17. C. Chen, A. Seff, A. Kornhauser, and J. Xiao, Deepdriving: Learning affordance for direct perception in autonomous driving," in *Pro- ceedings of the IEEE International Conference on Computer Vision*, pp. 2722–2730, 2015.
18. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.
19. J.Yosinski,J.Clune,Y.Bengio,andH.Lipson,"Howtransf erableare tures in deep neural networks?," in *Advances in neural information processing systems*,pp. 3320–3328, 2014.
20. J.BohgandD.Kragic,"Learninggraspingpointswithsha pecontext," *Robotics and Autonomous Systems*, vol. 58, no. 4, pp. 362–377, 2010.
21. M. Krainin, B. Curless, and D. Fox, "Autonomous generation of com- plete 3d object models using next best view manipulation planning," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pp. 5031–5037, 2011.
22. Q. V. Le, D. Kamm, A. F. Kara, and A. Y. Ng, "Learning to grasp objects with multiple contact points," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pp. 5062–5069, May 2010.
23. D. Song, K. Huebner, V. Kyrki, and D. Kragic, "Learning task constraints for robot grasping using graphical models," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International*

Conference on, pp. 1579–1585, Oct 2010.

24. J. Yu, K. Weng, G. Liang, and G. Xie, "A vision-based robotic grasping system using deep learning for 3d object recognition and pose estimation," in *Robotics and Biomimetics (ROBIO), 2013 IEEE International Conference on*, pp. 1175–1180, 2013.
25. J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Krger, J. Kuffner, and K. Goldberg, "Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1957–1964, May 2016.
26. E. Johns, S. Leutenegger, and A. J. Davison, "Deep learning a grasp function for grasping under gripper pose uncertainty," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4461–4468, Oct 2016.
27. J.Maitin-Shepard,M.Cusumano-Towner,J.Lei,andP.A bbeel,"Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding," in *2010 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2308–2315, 2010.