Introspective Overconfidence Prediction of 6-DOF Robotic Grasp Estimators via Multimodal Feature Fusion

Furkan Kaynar[1], Kevin Rexhepi[1] and Eckehard Steinbach[1]

***Abstract*— State-of-the-art 6-DOF grasp estimators suffer from failure rates as high as 10% to 15% which makes the practical deployment of robots in human environments challenging. The failing grasp attempts typically include also grasp estimations having a high estimated grasp quality score. This means that the grasp estimators can be overconfident in estimating the quality score, leading to falsely high scores. Detecting and eliminating the overconfident grasp estimations before execution can help decreasing the grasp failure rate and increase task success. We propose a novel method based on introspective prediction of overconfident grasps using a supervisor network. Our network takes the output grasp pose of an estimator, the depth image and the point cloud at the grasp location as input; and outputs a binary overconfidence label which can be used to filter out overconfident grasp estimations. Experimental evaluation with the selected quality thresholds shows that filtering grasps with our supervisor network leads to an increase in the rate of confident grasp estimations from 21% to 57%, and from 22% to 54% on the two test sets.**

I. INTRODUCTION

Vision-based robotic grasp estimation is an advancing field, yet the existing methods are not robust enough to be fully deployed in unstructured environments like the household. State-of-the-art methods can reach a failure rate of 10% to 15% on household objects [1], [2], [3], [4], which does not meet the requirements for deployment in our daily life. To exemplify, a household robot with such failure rates could easily create damage to the humans, the object to grasp or the environment. Therefore, predicting and preventing grasp failures is of great importance for the deployment of robots in human environments.

Vision-based 6-DOF grasp estimation methods typically output predicted grasp poses and quality scores. The grasp poses are ranked according to their quality score and the ones with a high quality score are executed primarily, for they are expected to succeed. However, the robotic experiments show that even these grasp estimations with a high quality score are prone to fail [1], [2], [3], [4]. This shows that the grasp estimation methods can be overconfident about the grasps they predict, for the predicted grasp quality score is higher than it should be in reality.

Vision-based 6-DOF grasp pose estimation often consists of two main steps: perception and grasp estimation. In the perception step, typically a depth image is acquired by a depth sensor. The obtained depth images contain sensor-dependent noise and pixel regions without a valid depth value, namely "depth holes". Such imperfections by perception directly affect the subsequent grasp estimation step. Many state-of-the-art 6-DOF grasp estimation methods operate on the point cloud generated from the acquired depth image. Although the imperfections like the depth holes are visible in the depth images, they are no more prominent on the generated point cloud. This can be a reason of grasp failures induced by the perception step. For the grasp estimation step, the simulation-to-reality gap can be named as an important cause of failures. The state-of-the-art methods mostly deploy neural networks trained with synthetic data that do not truly represent the realistic sensor characteristics. Therefore, these methods are not capable of overcoming the simulation-to-reality gap and can make overconfident estimations.

To address this problem, we propose a method for detecting overconfident parallel-finger grasp estimations, which is based on the concept of introspective failure prediction [5]. Benefitting from a largescale dataset [3], we create data consisting of grasp poses and quality scores estimated by two state-of-the-art 6-DOF grasp estimators [1], [2] on the point clouds generated from the real depth images. In addition, we evaluate the estimated grasps and create ground truth evaluation scores using the force-closure metric on the corresponding ground truth 3D scenes. Thereby, we obtain the grasp estimations under depth perception imperfections, but evaluate them on the 3D meshes without the adverse effects of depth sensing. Using this data, we train a supervisor network for a selected depth camera and a grasp estimator, that learns both the sensor characteristics and the simulation-to-reality gap behaviour of the grasp estimator. The network receives the scene information and an estimated grasp pose as inputs, and outputs a binary overconfidence label. We extract features on both the point cloud and the corresponding depth image crop around the input grasp pose, aiming to detect overconfidence that may be caused by the simulation-to-reality gap and imperfect perception. Experimental evaluation showed over 80% accuracy for detecting the overconfident grasps for the selected quality thresholds in this work. After filtering out the overconfident grasps estimated by our supervisor network, we can increase the rate of grasps that are truly high quality, although some of the high quality grasps are also eliminated. We provide a more in-depth analysis in Section IV.

[1] Technical University of Munich / School of Computation, Information and Technology / Department of Computer Engineering / Chair of Media Technology / Munich Institute of Robotics and Machine Intelligence (MIRMI) - Munich, Germany furkan.kaynar@tum.de

This work has been funded by the Lighthouse Initiative Geriatronics by StMWi Bayern (Project X, grant no. 5140951) and LongLeif GaPa gGmbH (Project Y, grant no. 5140953).

II. RELATED WORK

Various methodologies have been proposed for vision-based 6-DOF grasp estimation. A typical 2-step framework includes the grasp pose sampling and then the grasp evaluation steps [2]. The grasp evaluation can be performed either analytically, by simulation, or in a data-driven way via machine learning. More recently, grasp estimators based on end-to-end neural network training were proposed [1]. These methods receive the scene information, usually a point cloud, as input, and give the estimated grasp poses and the quality scores as outputs. Since they do not have separate steps for grasp sampling and evaluation, these methods cannot be used to evaluate given grasps.

In all vision-based grasp estimation methods, the accuracy of the estimated grasp quality score is closely related to the perception step, which is often not considered and can lead to grasp failures. Additionally, out-of-distribution samples can be a cause of grasp failures. The bottleneck of the system can be caused by different elements for each method, which may be hard to determine.

To address this issue, introspective failure prediction approaches can be used without the need for detecting the exact source of the failure. Rather than analyzing each modular part of a system, these approaches take the entire system as a black box and train machine learning models to detect the failure cases. Introspective failure prediction has been applied successfully to other problems like autonomous driving [6]. We follow this idea for grasp failure estimation by defining the failures as "overconfidence", standing for a high estimated grasp quality score and a low analytically computed ground truth score.

III. METHODOLOGY

A. Dataset Creation

Our aim is to detect whether a grasp pose with an estimated high quality score is indeed high quality or rather an overconfident estimation that may be likely to fail. To train our supervisor network, we need grasp poses and their quality scores estimated on real depth images, and in addition, the ground truth grasp quality scores that are not biased by the perception step.

The GraspNet-1Billion dataset [3] is appropriate for this, it includes nearly 100,000 RGB-D images, along with the 3D meshes and 6-DOF poses of the objects in the viewed scenes. With this data, it is possible to generate grasp estimations on the real depth images and evaluate them on the ground truth 3D scenes. We use 100 scenes of the GraspNet1Billion dataset to create our grasp overconfidence dataset. The procedure for creating the overconfidence dataset is as follows:

- **Grasp estimations.** We select depth images of the GraspNet1Billion dataset, and use object segmentation masks to determine the region of interest for single objects.
- We deploy a state-of-the-art grasp estimator and estimate 6-DOF grasp poses on the point clouds generated from the segmented depth images.
- Non-maximum suppression is applied to eliminate grasps having a similar pose.
- **Grasp evaluations.** We generate the 3D scenes using the object meshes and their ground truth object poses given in the GraspNet1Billion dataset.
- For each estimated grasp, we determine which object it is grasping in the 3D scene by checking the closest object point in the middle of the gripper fingers.
- We determine if the grasp area inside the gripper is empty or the gripper is in collision with the point cloud. In these cases, we label the grasp as empty or colliding.
- If not empty or colliding, we evaluate the grasp pose on the corresponding object mesh by using the force-closure metric as in [3]. We test different friction coefficients and obtain a grasp stability score.

We create 4 different grasp confidence datasets each with a different grasp estimator-sensor combination. The selected state-of-the-art estimators are Contact-GraspNet and 6-DOF GraspNet. The depth images acquired by Intel RealSense 435 or Kinect 4 Azure are used in each combination. In total we generated and evaluated nearly 5 million 6-DOF grasp poses. An analysis of the created data is given in the next section.

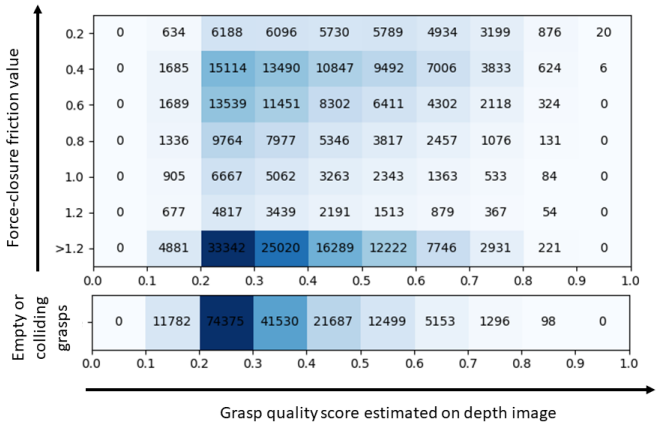
B. Cross-Score Analysis and Overconfidence Labeling

The overconfidence dataset includes two kinds of grasp quality scores for each grasp pose:

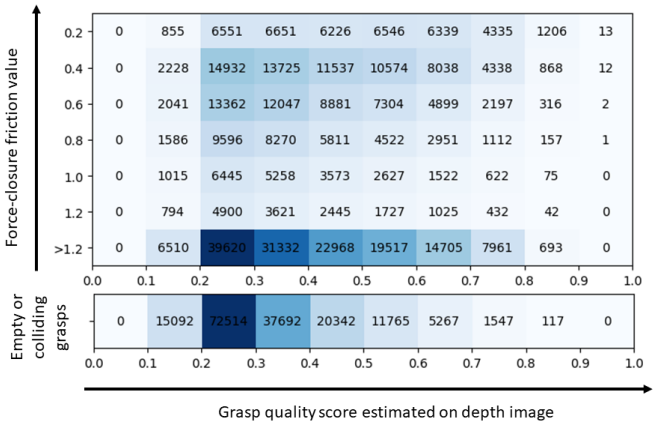
- Estimated grasp quality score on the real depth image. Note that this is data-driven for the selected grasp estimators and shows estimator-specific behaviour.
- Force-closure score computed analytically on the ground truth 3D scene based on different friction coefficients [3].

The grasp score distribution of the created datasets is shown in Fig. 1. For each matrix, the horizontal axis shows the estimated grasp quality scores, the right direction indicates a higher stability estimation. The vertical axis indicates the force-closure friction values obtained by evaluation on the 3D ground truth scenes, the upper direction indicates more stable grasps. The last row of each matrix indicates empty or colliding grasps.

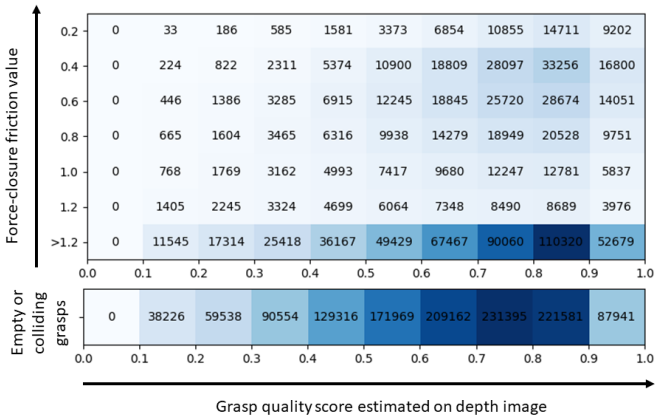
The first row in Fig. 1 (subfigures a. and b.) shows the score distribution of the grasps generated by Contact-GraspNet [1] and the second row (subfigures c. and d.) shows the distribution obtained with 6-DOF GraspNet [2]. Left matrices (subfigures a. and c.) show grasps estimated using the RealSense images, and the right matrices (subfigures b. and d.) using the Kinect images. The estimator-specific behaviour is visible as the difference between the first and second rows, although both methods have been originally trained with similarly generated synthetic data. The comparison of the left and right matrices within the rows suggests that the type of the camera does not affect the grasp score distribution behaviour as much as the grasp estimator choice. It is worth noting that many of the grasps generated by 6-DOF GraspNet are labeled as colliding, because this method is trained with isolated objects and does not consider collisions, however the scenes we use are cluttered.



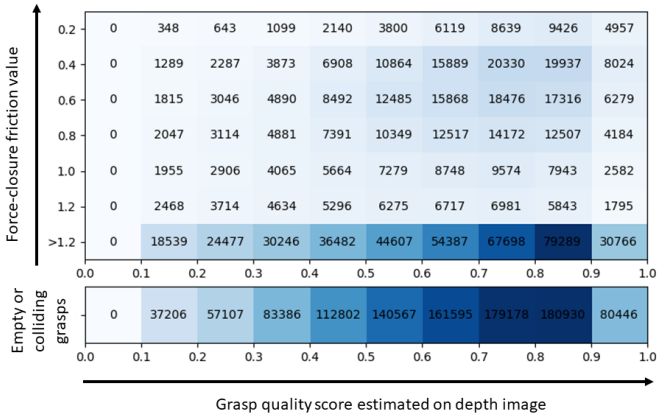
(a) Contact-GraspNet [1] estimations on RealSense depth images



(b) Contact-GraspNet [1] estimations on Kinect depth images



(c) 6-DOF GraspNet [2] estimations on RealSense depth images



(d) 6-DOF GraspNet [2] estimations on Kinect depth images

Fig. 1: The grasp score distribution of the created datasets. The darker color indicates a higher number of grasps.

Overconfidence labeling. If both the estimated and the evaluation scores are high, we deduce that the grasp is indeed high quality and the grasp estimator is rightly confident. On the other hand, if the estimated score is high but the ground truth score is low, we assume that the grasp estimator is falsely overconfident and we label these grasps accordingly. We do not label the grasp estimations with a predicted low quality score (left region of each matrix), because they are not creating an overconfidence case. These grasps already have low predicted scores and therefore they are not executed by the robot in principle.

Grasp quality thresholds. To create binary labels for overconfidence, we select the subset of all grasps that have an estimated quality score higher than 0.4 (right region in the figures). In this subset, we label the grasps as overconfident if the corresponding force-closure score requires a higher friction value than 0.4 (lower right region). The rest of the subset (upper right region) is labeled as correctly confident. Note that, although we use the same threshold value (0.4) for both scores, the metrics are different from each other and there can be other valid choices as well.

C. Supervisor Network Architecture

The structure of our network is given in Fig.2. 6-DOF grasp evaluation requires 3D information processing, and this is typically performed using point clouds. We process the point cloud crop at the grasp location with the PointNet-based upper branch. Since sensor imperfections can be more evident in the depth image domain, we additionally process the corresponding depth image crop at the grasp location in a CNN-based branch. We fuse the CNN features and the corresponding point cloud features densely by concatenation, as shown in the middle of Fig.2. The general architecture of our network is inspired by DenseFusion [7], however the DenseFusion network has a CNN branch processing RGB instead of the depth image to estimate the 6-DOF object poses. We replace the CNN branch by the depth processing CNN branch of another network [8] from the surface normal estimation literature.

IV. EXPERIMENTS

Datasets. We train and test our supervisor network on two of the created datasets: RealSense-Contact-GraspNet and Kinect-Contact-GraspNet. We use the grasp poses estimated on the first 85 scenes of GraspNet1Billion dataset for training, and the poses estimated on the next 15 scenes for testing. Some statistics on the used datasets are given in Table I.

Inference. Each input sample to the network consists of the local point cloud at the grasp location, the crop of depth image around the corresponding pixel location, and

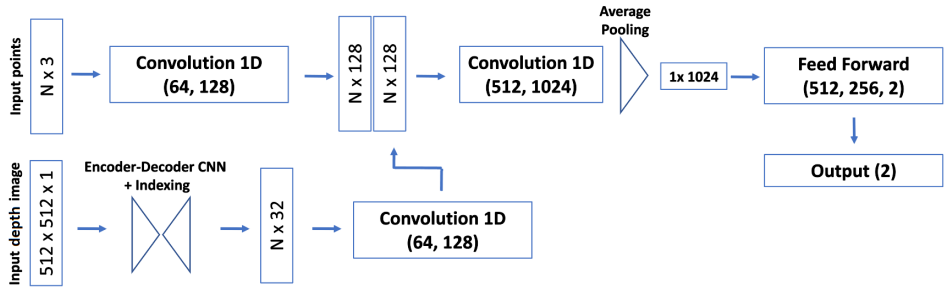


Fig. 2: Supervisor Network Architecture. The point cloud features and depth image features are fused in the middle.

TABLE I: Statistics on the datasets used for training and test

Dataset (created using Contact-GraspNet)	Total nr. of grasps	Overconfident grasps %	Confident grasps %
Kinect - Training set	185,314	71.25 %	28.75 %
Kinect - Test set	31,843	78.76 %	21.24 %
RealSense - Training set	156,128	69.78 %	30.22 %
RealSense - Test set	23,244	77.77 %	22.23 %

the estimated grasp pose given by the 6-DOF grasp estimator. The local point cloud is transformed to the grasp pose frame, inherently encoding the grasp pose into the input, as done in the literature. The output of our network is a binary label showing the overconfidence.

Implementation details. We take the 1000 points around the grasp location, and use a crop of 512x512 by the depth image. We train our neural network on an NVIDIA GeForce GTX 1080 with 8 GB memory that allows a batch size of 4. We use Adam optimizer and Negative Log Likelihood Loss function with class-wise normalization weights to counteract data imbalance in the datasets. The learning rate is initiated at 0.0001 and is reduced at 30th epoch by a factor of 0.7.

Results. Tables II and III show the experimental evaluation results after training for 32 Epochs. The results indicate that the supervisor network successfully detects the overconfident grasps with high accuracy, precision and recall values. For confident grasps, we observe that the precision and recall values are lower. Considering the data imbalance (see Table I), this indicates that some of the confident grasps are mislabeled as overconfident. Overall, we can deduce that the supervisor network is oversensitive and eliminates some of the good grasps as well. However, this ensures that the remaining grasps have a considerably lower overconfidence rate, therefore the remaining estimations are expected to lead to higher grasp execution success. Tables I and III show the confidence rates before and after filtering the grasp estimations. After filtering, we see a boost in the rate of confident grasp estimations from 21% to 57% on the Kinect test set, and from 22% to 54% on the RealSense test set.

V. CONCLUSION

We proposed a novel method for introspective prediction of overconfident grasp estimations to avoid grasp failures. We created datasets using real depth data and state-of-the-art grasp estimators. Experimental results indicate a successful elimination of overconfident grasps, at the cost of losing some of the confident grasps. The future work can include extension of the datasets, improvements in the neural network architecture for better accuracy and generalization, and robotic experiments for empirical evaluation.

TABLE II: Quantitative evaluation of overconfidence estimation for grasps in the Kinect and RealSense test sets

Type of grasps	Accuracy	Precision	Recall	F1-score
Kinect test set:				
Overconfident grasps	0.82	0.89	0.88	0.88
Confident grasps	0.82	0.57	0.61	0.59
RealSense test set:				
Overconfident grasps	0.80	0.89	0.85	0.87
Confident grasps	0.80	0.54	0.63	0.58

TABLE III: The rates after filtering out overconfident grasps

Dataset	Nr. of remaining grasps	Overconfident grasps %	Confident grasps %
Kinect - Test set	7,266	42.93 %	57.07 %
RealSense - Test set	6,034	46.42 %	53.58 %

REFERENCES

[1] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 438–13 444.

[2] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2901–2910.

[3] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[4] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *arXiv preprint arXiv:2212.08333*, 2022.

[5] S. Daftry, S. Zeng, J. A. Bagnell, and M. Hebert, "Introspective perception: Learning to predict failures in vision systems," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1743–1750.

[6] C. B. Kuhn, M. Hofbauer, G. Petrovic, and E. Steinbach, "Introspective failure prediction for autonomous driving using late fusion of state and camera information," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 4445–4459, 2020.

[7] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3343–3352.

[8] J. Zeng, Y. Tong, Y. Huang, Q. Yan, W. Sun, J. Chen, and Y. Wang, "Deep surface normal estimation with hierarchical rgb-d fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6153–6162.