FSG-Net: a deep learning model for semantic robot grasping through few-shot learning

FSG-Net: a Deep Learning model for Semantic Robot Grasping through Few-Shot Learning

Leonardo Barcellona[*,1,2], Alberto Bacchin[*,†,1], Alberto Gottardi[1,3], Emanuele Menegatti[1], Stefano Ghidoni[1]

***Abstract*— Robot grasping has been widely studied in the last decade. Recently, Deep Learning made possible to achieve remarkable results in grasp pose estimation, using depth and RGB images. However, only few works consider the choice of the object to grasp. Moreover, they require a huge amount of data for generalizing to unseen object categories.**

For this reason, we introduce the Few-shot Semantic Grasping task where the objective is inferring a correct grasp given only five labelled images of a target unseen object. We propose a new deep learning architecture able to solve the aforementioned problem, leveraging on a Few-shot Semantic Segmentation module. We have evaluated the proposed model both in the Graspnet dataset and in a real scenario. In Graspnet, we achieve 40,95% accuracy in the Few-shot Semantic Grasping task, outperforming baseline approaches. In the real experiments, the results confirmed the generalization ability of the network.

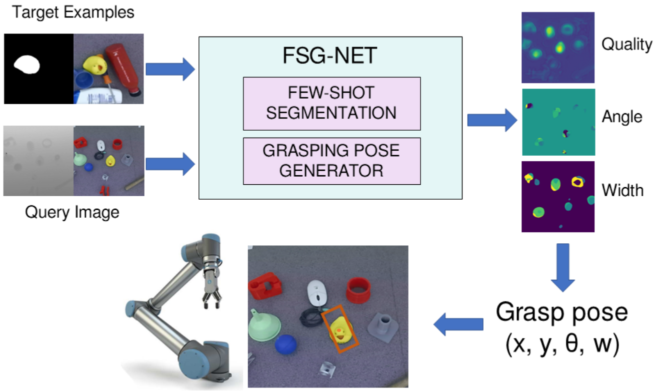


Fig. 1: The FSG pipeline. Given some examples of the object that we want to pick, FSG-Net generates a grasp pose in the image of the current scene, namely the query image.

I. INTRODUCTION

Grasping is one of the most fundamental manipulation skills for robots to interact with objects. State-of-the-art solutions already achieved impressive results on single object picking thanks to the advent of Deep Learning (DL) [1]. Since the increase in performance reached a plateau, researchers shifted the focus to the more challenging scenario of cluttered environments. Considering the latter situation, the approaches mainly rely on depth images, since they contain geometric clues of the objects, obtaining remarkable performance [2], [3].Afterwards, many authors further improved the solutions by exploiting RGB images [4], [5] or refining the grasp with the object semantic [6]. That said, what are the future perspectives in robotic grasping?

In real-world applications, exploiting semantic information is a very appreciated characteristic for grasping. From object sorting to autonomous robotic assembly, recognizing the right object to pick is crucial. Only few works [7], [8] adopt semantic knowledge and, consequently, are suitable for the aforementioned challenges. However, the methods proposed in those works do not easily generalize to unseen objects. In fact, every time a new category is required, it is necessary to expand the dataset to re-train the model.

Whereas DL approaches require huge amounts of labelled data, humans are able to recognize thousands of new and unseen objects just by few examples [9]. Starting from this observation, machine learning researchers have already introduced approaches able to reduce the data needed for training. The few-shot approaches aim to generalize to new categories or tasks providing only few examples or even one. For example, Few-shot Semantic Segmentation (FSS) models are able to segment unseen objects just by showing a small set of labelled images [10], [11], [12]. Despite the improvement in generalization brought by few-shot approaches, their usage is still limited. Furthermore, no previous studies explore the possibility to combine Few-shot Semantic Segmentation with traditional grasping pose generation to the best of our knowledge.

For these reasons, we formulate the Few-shot Semantic Grasping (FSG). The FSG is the task of generating an optimal grasp pose of an unseen target object, given only a few images labelled with semantic information of the target. We solved the task by designing a DL model, called FSG-Net, that leverage the generalization capabilities of FSS models together with a novel grasping pose generator. Figure 1 depicts the FSG pipeline and highlights the role of the proposed FSG-Net.

In detail, our contributions are:

- *A novel Deep Learning model*, called FSG-Net, that given some example images of the object to pick, computes a suitable grasping pose.
- *A pioneering use of Few-shot Semantic Segmentation* in the context of robot grasping, designing an architecture able to fuse spatial and semantic information.
- *Exhaustive experiments* on a publicly available dataset and a real setup which show the capability of the

[*] Authors equally contributed to the work.

[†] Corresponding Author

[1] Intelligent Autonomous System Lab, Department of Information Engineering, University of Padova, 35131 Padua, Italy. `barcellona, bacchinalb, gottardial, emg, ghidoni@dei.unipd.it`

[2] Politecnico di Torino, 10138 Torino, Italy.

[3] IT+Robotics srl, 36100 Vicenza, Italy.

proposed FSG-Net to generalize to unseen objects, while retaining state-of-the-art grasping pose estimation accuracy.

The open source implementation of FSG-Net is available[1].

II. Related Works

A. Few-Shot Semantic Segmentation

In computer vision, semantic segmentation is the dense classification of the pixels of an image according to the class they represent. Almost all state-of-the-art solutions adopt DL approaches [13], [14], [15], [16]. Unet [15] is one of the most widely used deep neural networks for semantic segmentation and presents an encoder-decoder structure solving the vanishing gradient problem by concatenating the encoder features during the decoding. However, these DL models cannot generalize to unseen classes. Moreover, adding new classes requires huge amount of new labelled examples and a new training of the network. To overcome these limitations, Shaban et al. [10] proposed the Few-shot Semantic Segmentation, that segments an image, named query, given only few examples, named support. Despite the recent introduction of the topic, it caught high attention from the research community and many solutions were proposed in the literature [11], [12], [17], [18]. The most prominent approaches encode the information of the image in a discriminative vector, called prototype, matching its content with the support images. Prototyping approaches were introduced in FSS by Dong [19] after the good results obtained in few-shot classification [20]. In PFENet [11] the authors proposed a features enrichment decoder able to exploit also high-level features. In SCL [18] the authors propose a Self-Guided Module that keeps more discriminative information than the masked Global Average Pooling proposed in [17]. A very inspiring approach is Asgnet [12] that introduces a clustering approach to overcome the problem of occlusion and scale variation. The idea is to clusterize the prototypes to keep the spatial information. Despite the promising results, the applicability is still limited. In our work, we exploit FSS by means of a specific attention module (see Section IV for details) that is able to recover some segmentation errors and to increase the usability of such models. Thanks to this enhancement, we are among the first to adapt FSS models for robot grasping.

B. Robotic Grasping

Robot grasping aims to infer the correct gripper pose to stably pick up objects. Nowadays, the advent of DL overcomes previous approaches, implicitly modelling complex relationships between objects shape ad grasping poses. Redmon et al. [21] propose a CNN followed by a fully connected layer to regress the grasp pose as a single tuple of values. Hence, that work can detect only one grasp per inference. Conversely, a popular generative network, named Generative Grasping Convolutional Neural Network (GG-CNN), was proposed by Morrison et al. [2]. Starting from depth images, the authors used a multi-branch Fully Convolutional Neural Network (FCNN) that predicts for each pixel the quality of the grasp, the orientation and the width of the end-effector. This information is encoded in three dense heatmaps, directly computed by the network. In [4], authors further developed this idea in their Generative Residual Convolutional Neural Network (GR-ConvNet), using a larger model which also exploits RGB images and using the smooth L1 loss function to reduce the gradient divergence. By means of heatmaps, it is possible to predict an arbitrary number of grasps in a single inference. Inspired by [2], we also decided to encode the information in heatmaps, while using L1 loss, as [4].

All the above solutions and most of the grasping frameworks in the literature are object-agnostic, meaning that these models select the most suitable object to grasp without caring about the object type. However, many applications may require knowledge of what they are going to pick. Consequently, in [7], the authors define the problem of a robot picking up an object of a user-specified class as *semantic robotic grasping*. This introduces further complexity since it combines geometric information to infer a suitable grasping pose and semantic information to select the correct object. The same work proposes different architectures to jointly learn good 4 DoF poses and semantic labels through a self-supervised approach using many robots. Sun et al. [22] fuse the semantic segmentation from a CNN with a robust model fitting technique in order to inject semantic information into the grasping task. Training such systems is highly demanding because introducing new categories requires many efforts. In our work, we exploit the semantic segmentation to (a) introduce the class-related information enabling semantic grasping and (b) improve the estimation of the grasping pose, as previous works demonstrate [6], [23]. Differently from previous studies, we integrate into our pipeline a FSS network to tackle the problem of adding new objects with low effort. Our method can grasp objects of a specific class never seen during the training procedure just by proving few examples while delivering state-of-the-art grasping pose estimation accuracy.

III. FSG Formulation

This section gives a mathematical definition of the FSG task. Before that, we introduce two strictly related tasks: the grasp inference and the FSS. Their definitions are essential to formulate the FSG problem.

Grasp inference. In our work, we define a 4 DoF grasp pose similarly to [2]. A grasp is defined as the tuple

$$gr = (x, y, \theta, w)\,, \tag{1}$$

where x and y are the coordinates of the center of the grasp, θ is the orientation of the gripper along the axis perpendicular to the image plane in radians, w is the width of the gripper suitable to pick up the object without collisions.

Few-shot Semantic Segmentation. We define a support set as $\mathcal{S} = \{(I_s, M_s)_i\,,\, i = 1, ..., k\}$, that is a set of k tuples of images I_s and the segmentation masks M_s of a certain class. Given a query image I_q and a support set $\mathcal{S}$, the FSS

[1] https://github.com/leobarcellona/FSG-Net

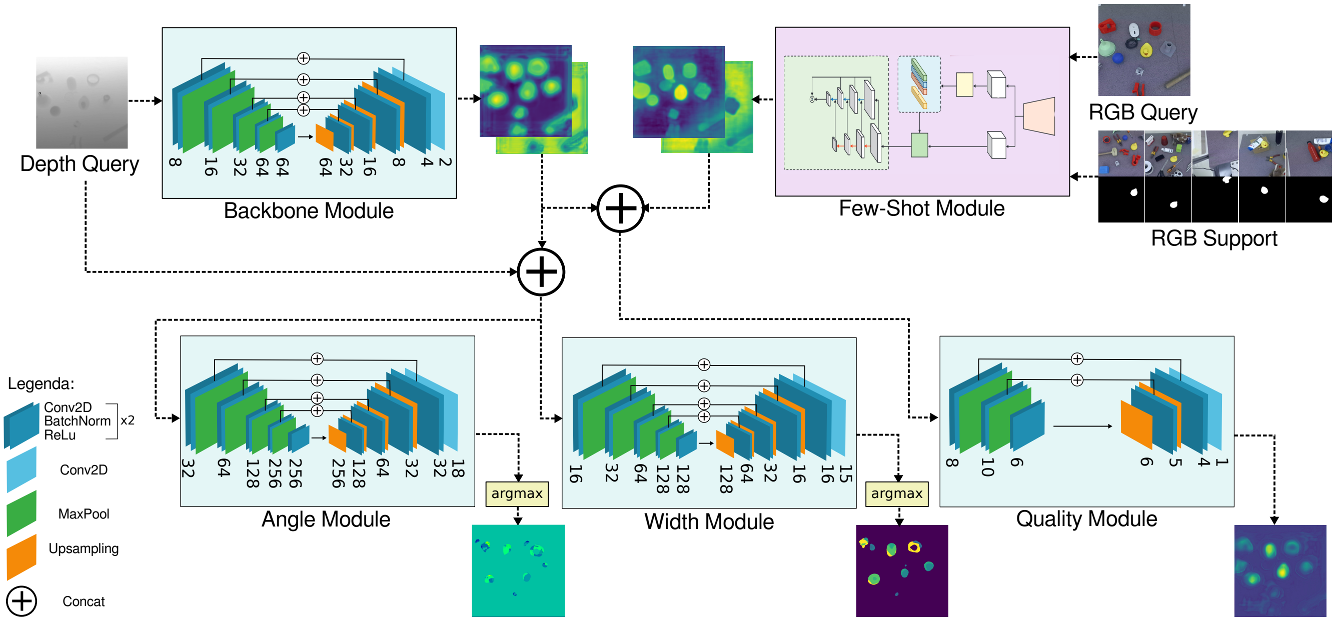


Fig. 2: A detailed scheme of the proposed architecture. The *Few-Shot Module* (upper-right corner) extracts semantic information from RGB images. The *Backbone module* (upper-left corner) is in charge to output low-level spatial features. The three heads of the model (bottom) are used for the final prediction of A, W and Q. The light-blue boxes high-lights self-designed sub-networks, based on Unet [15].

task estimates the segmentation mask M_q of I_q for the class identified by $\mathcal{S}$. In this setting, we define the training set as $\mathcal{D}_{train} = \{(I_q, M_q, S)_i \ , \ i = 1, ..., n\}$ and the test set as $\mathcal{D}_{test} = \{(I_q, M_q, S)_i \ , \ i = 1, ..., m\}$. The two sets have no class in common, formally $\text{cls}(\mathcal{D}_{train}) \cap \text{cls}(\mathcal{D}_{test}) = \{\emptyset\}$, where $\text{cls}(\cdot)$ return the list of classes in a set.

Few-shot Semantic Grasping. From the above definitions, given a query-support tuple $(I_q, \mathcal{S})$, we define the Few-Shot Semantic Grasping as the task of inferring the optimal grasp

$$gr_s^* = (x^*, y^*, \theta^*, w^*) \quad s.t. \ \ \text{cls}(\{I_q(x,y)\}) = \text{cls}(\mathcal{S}) \tag{2}$$

where $I_q \in \mathbb{R}^{4\times H\times W}$ is the RGB-D query image, $I_s \in \mathbb{R}^{k\times 3\times H\times W}$ are the RGB support images and $M_s \in \mathbb{R}^{k\times H\times W}$ are the binary segmentation masks.

Inspired by generative approaches like [2], [4], we formulate a solution for the FSG task as the prediction of three heatmaps $(Q, A, W) \in \mathbb{R}^{3\times H\times W}$ which respectively estimate, for each pixel, the grasp quality (i.e. probability of successfully grasp an object of class C_i), the grasping angle θ and the gripper width w. Under this setting, we can extract the following grasping pose:

$$gr_s = (x, y, A(x,y), W(x,y)) \ \ , \quad (x,y) = \arg\max_{x,y} Q \tag{3}$$

The aim of the learning procedure is to provide a good estimation $gr_s \simeq gr_s^*$.

IV. Approach

In this section, we present a novel DL model to grasp a selected object in a clutter, given I_q as an RGB-D image of the scene and a support set $\mathcal{S}$ with cardinality $k = 5$. The model is the composition of 5 modules, namely the *Few-Shot Module*, the *Backbone*, the *Angle Module*, the *Width Module* and the *Quality Module*. For the *Few-Shot Module*, we selected a model from the literature [12], while the other modules are derived from the Unet [15] encoder-decoder structure. The main building block consists in the repeated application of 3x3 convolutions, each followed by Batch-Normalization and ReLU activation. The building block is followed by a MaxPooling operation during the encoding pass and an UpSampling operation in the decoding pass. Shortcut connections are provided to avoid the vanishing gradient problem. While the four modules have the same structure, there are some slight differences in the dimensions that are detailed in Figure 2. The *Backbone*, the *Angle Module* and the *Width Module* are composed by 5 building blocks, while the *Quality Module* by 3. This difference comes from experimental observations. Also, we noticed that the *Angle Module* benefits from a higher number of convolutional filters compared to the *Width Module*. The next sections will provide more details about each module.

A. Few-Shot Module

The objective of the module is solving the FSS task, namely segmenting an object from an RGB image given only the support $\mathcal{S}$ that is composed by k=5 image-mask tuples. The output consists of two heatmaps encoding the confidence grade of the network in segmenting a pixel as target object or as background. Segmentation was not used only for choosing the target object, but also to deliver useful geometric clues, such as the shape of the object, in order to improve performance. The modular structure of the entire model allows using any FSS networks as long as they share the same input and output. After an empirical comparison of different models, detailed in Section V, we selected Asgnet [12] for the *Few-shot Module*.

B. Backbone Module

The *Backbone Module* is an Unet architecture with the sequence of filters reported in Figure 2. It uses the depth image to extract two attention maps with double functionality: encoding the graspable positions and solving the well-known problem of class imbalance [24], which is caused by a dominant extension of the background class that deteriorates the performance in the training phase. The attention maps solve class imbalance by implementing a Spatial Attention Mechanism (SAM) [25], [26], in order to let the *Angle, Width and Quality Module* focus on relevant areas of the image, i.e. areas taken up by graspable objects. Usually, SAM implicitly learns some attention maps based on data distribution [25], [26]; such maps are then fused through element-wise multiplication. Differently, we force the *Backbone Module* to learn attention maps with an explicit meaning, concatenating them with the input of *Angle, Width and Quality Module* and letting these modules learn how to exploit them, similarly to [27].

C. Angle and Width

The *Angle Module* and the *Width Module* share a similar structure. Both are fed with the concatenation of the depth query image with the *Backbone Module* to implement the aforementioned attention mechanism. We decided to use only depth images since it contains enough geometric information to predict a good grasping pose, as demonstrated in [2], [3]. Previous studies [28], [29] indicate that directly applying regression to complex heatmaps, like A and W, is more difficult than classification. Therefore, we quantized angle and width values to formulate the prediction of A and W as a classification problem. Following [28] suggestion, rotation angles around the gripper axis have been encoded between $[-\frac{\pi}{2}, +\frac{\pi}{2}]$ rad and discretized in 18 bins of size 10°. Similarly, gripper width values are encoded between $[0, 150]$ mm and discretized in 15 bins. Consequently, the output of the modules is respectively composed by a set of 18 and 15 activation maps, modelling the probability that each pixel belongs to a certain bin and mimicking what is commonly seen in semantic segmentation networks [30]. The final heatmaps are computed through a depth-wise $\arg\max$ operation. The internal design of the network is the same for both *Angle Module* and the *Width Module*, except for the number of layers because the A prediction is more complex due to the higher number of categories.

D. Quality Module

The *Quality Module* fuses semantic and spatial information, respectively from the *Few-Shot Module* and the *Backbone Module*, in order to estimate the grasp quality Q. The higher the value of Q the higher the probability of grasping the object. Since Q is highly related to the grasp position and the target object, the concatenation of the backbone and few-shot heatmaps is fed into this module. Unlike the previously described modules, we decided to formulate the inference of Q as a regression problem, obtaining a single-channel continuous heatmap as output. The continuity of Q removes the ambiguities while extracting the grasping position (x, y) as formulated in Eq. 3 and leads to a more accurate estimation of the final grasping pose. Moreover, the prediction of Q is less demanding and can be solved through regression, as shown in previous studies [4], [5] and also by our empirical observations. For this reason, we designed a lighter architecture for the *Quality Module*.

TABLE I: Class id of each split in Graspnet [31].

Split 0	Split 1	Split 2	Split 3
0-2-5-7	14-17-18-20	27-29-30-38	52-58-60-61
8-9-11	21-22-26	41-48-51	62-63-66

V. EXPERIMENTS

A. Dataset

To train and evaluate the proposed model, we used the Graspnet dataset [28]. It contains images captured both with a Kinect Azure and a RealSense D435 of 88 different objects placed in 190 small cluttered groups, bringing 97,280 images in total. Each image is densely annotated with 7 DoF grasp poses and segmentation masks. The dataset comes with 4 predefined splits: “train”, “test-seen”, “test-similar” and “test-novel”. Each split contains different scenes. The splits “train” and “test-seen” share the same objects but with a different disposition, while “test-similar” and “test-novel” contain completely new objects compared to the “train” and “test-seen” splits. This division simplifies benchmarking grasps of unseen objects.

In section III, we defined a 4 DoF grasp in Eq. 1. Since the Graspnet dataset contains 7 DoF poses, we need to reduce them to our formulation. To do so, we consider only grasp poses oriented in the same direction of the camera with a tolerance of ± 0.1 rad. We now have to generate the ground truth heatmaps Q_{GT}, A_{GT} and W_{GT} in order to train our model, following a procedure similar to [2]. From each selected grasp pose, we generate an oriented bounding box $BB = (x, y, \theta, h, w)$. Starting from empty heatmaps, each bounding box contributes to a patch of size $\frac{h}{6} \times \frac{w}{2}$ centered in (x, y) filled θ value for A_{GT} and w value for W_{GT}. For quality heatmap Q_{GT}, we filled each patch with value 1 if BB belongs to the target object and 0.25 for BB belonging to other objects. Since the *Quality Module* is trained through smooth L1 loss, we want to give less weight to errors in choosing the wrong object than the background. In this way, we give more emphasis on semantic grasping during the training process, retaining the overall grasping capability. We decided to use only Kinect images to match the real setup (Section VI-C).

Given the modularity of our architecture, we design a step-by-step training procedure which is different from the end-to-end approach used in many works. Indeed, Glasmachers in

[32] demonstrated that end-to-end learning may be not effective for training neural network models composed of multiple non-trivial modules. Since our model fits this definition, the training procedure has been split as detailed below.

Few-Shot Module fine-tuning. The datasets used by Few-shot Semantic Segmentation models are mainly PASCAL-5^i [10] and COCO-20^i [33], while the standard metric is the Intersection over Union (IoU) of the new classes [10]. Despite that FSS models are able to generalize to new classes, the performance considerably decreases when applied to a new dataset. For example, when trained on COCO-20^i and tested on PASCAL-5^i, there is a performance drop [34], [35]. Moreover, Graspnet contains more specific classes compared to PASCAL-5^i or COCO-20^i. For these reasons, we decided to fine-tune the *Few-Shot Module* before using it.

Graspnet was not designed for FSS task, so we organized it to be compliant with the definition of FSS (Section III). We used the "train" and the "test-seen" splits for train and validation respectively. The classes inside both splits were subdivided into four sub-splits to apply the standard k-fold validation of FSS [10]. The classes of each split are reported in Table I. Since the modular structure of our model allow us to insert many different FSS architectures, we compared three different state of the art solutions: PFENet [18], SCL-PFENet [17] and Asgnet [12]. We chose Asgnet, as reported in Section IV-A, because it achieved the best results with an IoU of 0.459 on Graspnet.

Backbone, Angle and Width Module. Modules that are independent from semantic information are trained using the "train" split and validated using the "test-seen" split. Firstly, we train the *Backbone Module* using a Cross-Entropy Loss. The ground truth is obtained from Q_{GT} by ceiling 0.25 to 1. After training, the weights of the backbone are frozen. In this way, *Quality, Angle and Width Modules* training process does not override the knowledge previously learned. Finally, *Angle and Width Modules* are trained using the Cross-Entropy loss [36].

Quality Module. The input of the module is the concatenation of backbone attention maps and the FSS heatmaps. Since the FSS model has been previously fine-tuned on some classes of the "train-seen" split, the performance on these classes may be overfitted. To avoid this bias, we trained the *Quality Module* only with classes of the "train" split still unused. Additionally, we included the classes of the validation split used in the *Few-Shot Module* fine-tuning, namely the split 3 of Table I. As training loss, we used the smooth L1 function defined in [4] to avoid exploding gradients. Also, we observed a better accuracy in estimating gr_s compared to the Cross-Entropy loss.

VI. Results

A. Training Procedure

The section shows the results of the proposed FSG-Net on the Graspnet dataset and in a robotic workcell where a robot grasp objects from a clutter placed on a table. To measure the system performance we employ the grasping accuracy, namely the number of correct grasps divided by the number of grasps. Since we want to catch a specific object we also distinguish between successful grasps of the target object and successful grasps of a wrong object.

TABLE II: Results of our network and two state of the art generative model from the literature on "test-similar" and "test-novel" split of Graspnet. The column represent the accuracy in object selection (A_{Sem}), semantic grasp (A_{SemGR}) and class-agnostic grasp (A_{AgnGR}).

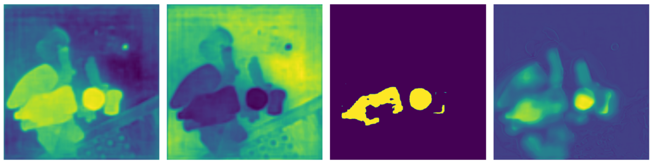
Model	Similar		
	A_{Sem}	A_{SemGR}	A_{AgnGR}
GGCNN [2] + FSS	41.83	30.24	89.19
GR-ConvNet [4] + FSS	34.57	23.62	69.76
FSG-Net Backbone + FSS	32.95	19.45	69.91
FSG-Net	51.54	43.51	93.88
Model	Novel		
	A_{Sem}	A_{SemGR}	A_{AgnGR}
GGCNN [2] + FSS	43.32	34.09	87.13
GR-ConvNet [4] + FSS	40.05	28.04	73.74
FSG-Net Backbone + FSS	30.42	18.77	67.95
FSG-Net	49.44	40.95	94.42

B. Results on Graspnet

After training and validating the model on the "train" and "test-seen" splits we evaluated it on the"test-unseen" and the "test-novel". Since these two folds contain new classes not seen during training, we used them to test the generalization capability. We adopted the same metric of [6] for evaluating a grasp. Let gr_p the grasp obtained by the network and gr_{gt} the ground-truth. We consider gr_p correct if the following conditions are satisfied: $\frac{gr_p \cap gr_{gt}}{gr_p \cup gr_{gt}} > 0.25$ and $|angle(gr_p) - angle(gr_{gt})| < 30$.

We define a correct semantic grasp as the event in which both conditions hold and the right object is chosen. When the two conditions hold but the selected object is not the right target, a correct class-agnostic grasp is obtained. We also reported the results relaxing the first condition by lowering the threshold to 0, in order to evaluate the ability of the network to correctly locate the right target, without taking into account the grasp success. The percentage of correct semantic grasps overall is the semantic grasp accuracy (A_{SemGR}). Similarly, we define class-agnostic grasp accuracy (A_{AgnGR}). The ratio between correctly located objects and the total is instead the semantic accuracy (A_{Sem}). We report the results in Table II comparing our approach with other models which use heatmaps to encode grasping poses. We retrained these networks on Graspnet since their authors exploited other datasets. To create a baseline for Few-shot Semantic Grasping performance evaluation, we apply directly the segmentation mask from the *Few-Shot Module* to the heatmaps from these networks. Our model is able to achieve 43.51% and 40.95% accuracy in the *similar* and *novel* splits. Thanks to the *Few-Shot Module* we are able to achieve almost the same results in both splits.

To assess the importance of the *Quality Module*, a specific test was run after removing the *Quality Module* from the



(a) (b) (c) (d)

Fig. 3: The heatmaps - (a) foreground and (b) background - from the *Few-Shot Module* and (c) the corresponding segmentation mask. The output of the *Quality Module* (d), recovers the target object (the ball) even if the segmentation partially fails.

proposed architecture. To compute the grasping pose gr_s, we apply directly the few-shot segmentation mask to the *Backbone Module* output. We refer to this ablation as *FSG-Net Backbone + FSS*. The results demonstrate that the recovery ability of the *Quality Module* is non-negligible and is due to the ability of the module to recover wrong segmentation masks using the heatmaps. For example, in figure 3 the *Few-Shot Module* over-segments the mask, but the *Quality Module* is still able to recover the right target (the ball). To evaluate the pure grasping ability of our network, we have to consider the correct object grasp metric. Also in this metric, we outperform the other approaches based on heatmaps, since *GGCNN [2] + FSS* was able only to correctly grasp the 34.09% of the objects, while *GR-ConvNe [4] + FSS* even less. We suppose that the semantic information can help to better refine the grasping pose. Similarly, the use of the *Backbone Module* output may help to improve the estimation of A and W by focusing the modules on the areas containing good grasps with high probability. The superimposition of these factors leads to a better accuracy in grasp pose estimation.

C. Results in Real-world

In this section, we validate our approach in a real scenario with static cluttered objects. All the experiments were performed using a 6-DoFs Universal Robots UR5 manipulator equipped with Onrobot's RG6 gripper and Microsoft Azure Kinect camera. The camera is mounted on the wrist of the robot. This setup is shown in Fig. 4. The *FSG-Net* computations are performed on a NVIDIA GeForce GTX 1050 Ti Mobile. We used a set of 20 novel objects, depicted in Figure 4, of various sizes and shapes. We also collected and labelled the 5 images to use as support for each object. We perform 10 trials and, for each trial, 10 random objects were placed on the table. The system selects each of the 10 objects one per time, extracts five support images with current RGB-D query images and provides them as input to the *FSG-Net*. As described in previous sections, the *FSG-Net* returns the grasp gr_p. Finally, this information is reprojected in 3D and forwarded to the robot that performs the pick-up. In order to make each grasp comparable, the object taken is put back on the table.

In the real environment, the A_{Sem} that we achieved is 38.38%, while the A_{SemGR} is 32.32%. In general, our approach is able to complete the gripping of an object with an A_{AgnGR} of 69.70%. The performance reduction compared to what was obtained on the Graspnet dataset is due to the limitations encountered in the real environment, the first consisting in the different lighting conditions. Secondly, the quality of the FSS output can be affected by the position of the camera. Another limitation was caused by the camera depth measurement. Indeed, we observed that in the furthest part of the depth image, data was affected by noise. This effect caused 7 grasp failures. Overall, the performance assessment described so far leads us to the following conclusions: i) the proposed model demonstrated its superior performance over other systems proposed in the literature on standard benchmarks, like the Graspnet dataset; ii) our model is ready for real-world applications, even though a number of additional factors shall be considered in such scenario. Our future research activity will target this new challenge.

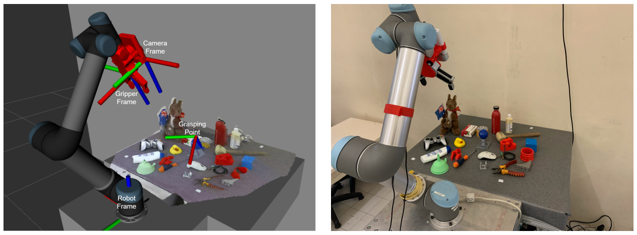


Fig. 4: The setup and the objects used during the real experiments.

VII. CONCLUSIONS

In this paper, we proposed *FSG-Net*, a Deep Learning model for semantic grasping using few-shot segmentation. It takes in input an RGB-D image and five examples of a target unseen object and infers the grasping pose. Thanks to the *Quality Module*, which exploits spatial and semantic information, the model is able to recover the right grasping position even with wrong segmentation masks. Our proposal achieves promising results in the Graspnet dataset compared with similar approaches in the literature. The experiments in the real setup confirmed the generalizability of our network taking into account a slight decrease in performance.

In order to reduce the domain gap between real and synthetic results observed in the real experiments, we plan to extend our model to predict 7 DoF grasping poses. Moreover, the segmentation mask predicted by the few-shot model is still limiting the effectiveness of our model, despite the recovery capability introduced by the *Quality Module*. Therefore, we plan to redesign the *Few-shot Module* to increase performance in object segmentation.

ACKNOWLEDGMENT

Part of this work was supported by MIUR (Italian Minister for Education) under the initiative "PON Ricerca e Innovazione 2014 - 2020", CUP C95F21007870007.

References

[1] K. Kleeberger, R. Bormann, W. Kraus, and M. Huber, "A survey on learning-based robotic grasping," *Current Robotics Reports*, vol. 1, p. 239–249, 12 2020.

[2] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 183–201, 2020. [Online]. Available: https://doi.org/10.1177/0278364919859066

[3] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," 2017. [Online]. Available: https://arxiv.org/abs/1703.09312

[4] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 9626–9633.

[5] M. Gou, H.-S. Fang, Z. Zhu, S. Xu, C. Wang, and C. Lu, "Rgb matters: Learning 7-dof grasp poses on monocular rgbd images," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 459–13 466.

[6] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 452–13 458.

[7] E. Jang, S. Vijayanarasimhan, P. Pastor, J. Ibarz, and S. Levine, "End-to-end learning of semantic grasping," 2017. [Online]. Available: https://arxiv.org/abs/1707.01932

[8] S. Iqbal, J. Tremblay, A. Campbell, K. Leung, T. To, J. Cheng, E. Leitch, D. McKay, and S. Birchfield, "Toward sim-to-real directional semantic grasping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 7247–7253.

[9] I. Biederman, "Recognition-by-components: a theory of human image understanding." *Psychological review*, vol. 94, no. 2, p. 115, 1987.

[10] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," in *BMVC*, 2017.

[11] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[12] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8334–8343.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[17] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *IEEE transactions on cybernetics*, vol. 50, no. 9, pp. 3855–3865, 2020.

[18] B. Zhang, J. Xiao, and T. Qin, "Self-guided and cross-guided learning for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8312–8321.

[19] N. Dong and E. P. Xing, "Few-shot semantic segmentation with prototype learning." in *BMVC*, vol. 3, no. 4, 2018.

[20] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[21] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2015, 12 2014.

[22] G.-J. Sun and H.-Y. Lin, "Robotic grasping using semantic segmentation and primitive geometric model based 3d pose estimation," in *2020 IEEE/SICE International Symposium on System Integration (SII)*, 2020, pp. 337–342.

[23] M. Dong, S. Wei, J. Yin, and X. Yu, "Real-world semantic grasp detection based on attention mechanism," 2021. [Online]. Available: https://arxiv.org/abs/2111.10522

[24] M. S. Hossain, J. M. Betts, and A. P. Paplinski, "Dual focal loss to address class imbalance in semantic segmentation," *Neurocomputing*, vol. 462, pp. 69–87, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221011310

[25] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6298–6306.

[26] Z. Zhong, Z. Q. Lin, R. Bidart, X. Hu, I. B. Daya, Z. Li, W.-S. Zheng, J. Li, and A. Wong, "Squeeze-and-attention networks for semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 062–13 071.

[27] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3285–3294.

[28] 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.

[29] F.-J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3355–3362, 2018.

[30] I. Ulku and E. Akagündüz, "A survey on deep learning-based architectures for semantic segmentation on 2d images," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2032924, 2022. [Online]. Available: https://doi.org/10.1080/08839514.2022.2032924

[31] 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.

[32] T. Glasmachers, "Limits of end-to-end learning," in *Proceedings of the Ninth Asian Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M.-L. Zhang and Y.-K. Noh, Eds., vol. 77. Yonsei University, Seoul, Republic of Korea: PMLR, 15–17 Nov 2017, pp. 17–32. [Online]. Available: https://proceedings.mlr.press/v77/glasmachers17a.html

[33] K. Nguyen and S. Todorovic, "Feature weighting and boosting for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 622–631.

[34] M. Boudiaf, H. Kervadec, Z. I. Masud, P. Piantanida, I. Ben Ayed, and J. Dolz, "Few-shot segmentation without meta-learning: A good transductive inference is all you need?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 979–13 988.

[35] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6941–6952.

[36] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, "A comprehensive survey of loss functions in machine learning," *Annals of Data Science*, vol. 9, 04 2022.