Orientation Attentive Robot Grasp Synthesis

Nikolaos Gkanatsios[1], Georgia Chalvatzaki[2], Petros Maragos[1] and Jan Peters[2,3]

***Abstract*— Physical neighborhoods of grasping points in common objects may offer a wide variety of plausible grasping configurations. For a fixed center of a simple spherical object for example, there is an infinite number of valid grasping orientations. Such structures create ambiguous and discontinuous grasp maps that confuse neural regressors. We perform a thorough investigation on the challenging Jacquard dataset to show that the existing pixel-wise learning approaches are prone to box overlaps of drastically different orientations. We then introduce a novel augmented map representation that partitions the angle space into bins to allow for the co-occurrence of such orientations and observe larger accuracy margins on the ground truth grasp map reconstructions. On top of that, we build the ORientation AtteNtive Grasp synthEsis (ORANGE) framework that jointly solves a bin classification problem and a real-value regression. The grasp synthesis is attentively supervised by combining discrete and continuous estimations into a single map. We provide experimental evidence by appending ORANGE to two existing unimodal architectures and boost their performance to state-of-the-art levels on Jacquard, specifically 94.71%, over all related works, even multimodal. Code is available at `https://github.com/nickgkan/orange`.**

I. Introduction

Successful robotic grasping in unstructured environments (Fig. 1) is a long aspiration, towards the successful migration of robots, able to assess and reason about their surroundings, into human-inhabited environments, as assistants in homes, hospitals, etc. However, this includes several sub-problems to be solved, like perception, grasp planning and control. For that reason, grasping objects of different shapes, textures and sizes has been explored both in an analytical [1] and data-driven fashion [2]. In the general sense, the notion of a grasp can be parameterized by a point on the object, according to which the robot's end effector (tool) center should align, a 3-D angle with which the robot's tool should approach the grasp point and the initial configuration of the tool w.r.t. to its optimal width for performing the grasp [3].

The advantages in Deep Learning (DL), along with the introduction of low-cost RGB-D sensors, and the creation of large datasets, gave an increasing advantage to data-driven approaches for robotics perception. During the last years, such datasets have arisen for robot grasping [4], [5], containing a multitude of graspable objects usually found in human-inhabited environments that are suitable for robotic hands and grippers. Several approaches have tried to transfer methods currently achieving state-of-the-art performance in computer vision problems like object detection with bounding box regression (e.g. Faster-RCNN [6]) for detecting antipodal grasps on objects from RGB data. These approaches predict and rank thousands of grasp candidates [7]–[9], requiring much computational resources, while they are limited to static environments and precise camera calibration for performing grasps.

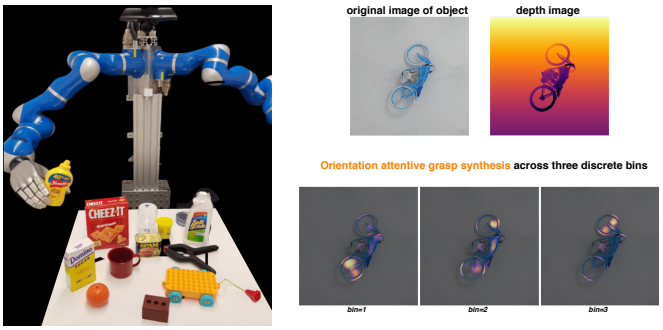

Fig. 1: **Left:** Grasping is still challenging for robots due to the high variability in the objects' morphology, especially for human-inhabited environments. **Right:** The *ORANGE* framework effectively estimates the grasping points over different orientations, thanks to its augmented grasp map representation and the orientation attentive mechanism that focuses on learning valid grasp points on the objects.

Recently, a depth-based approach attempted to confront these issues by modeling the generation of dense maps representing the possible grasping configurations of objects [10], while [11] explored the generative capabilities of a convolutional neural network (CNN), in order to estimate such maps. This pixel-wise approach of synthesizing grasp representations from depth images is rather important for real-time performance in robotics. Effectively estimating the *approach vector*, i.e. the orientation with which the robotic hand or gripper approaches, is fundamental for safe and successful grasping. The continuous orientation estimation is particularly important, especially for reactive grasp planning, either in cases when the camera is mounted on the robotic hand and changes its perspective, or when it is required to grasp moving objects. Intuitively, when humans observe an object, they argue about the object's shape and navigate their

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. #640554 (SKILLS4ROBOTS). Experimental computations have been conducted on an Nvidia DGX-1 at TU Darmstadt.

[1]School of E.C.E., National Technical University of Athens, 15773, Athens, Greece `nikos.gkanatsios93@gmail.com`, `maragos@cs.ntua.gr`

[2]Intelligent Autonomous Systems, Technische Universität Darmstadt, Hochschulstr. 10, 64289 Darmstadt, Germany `georgia@robot-learning.de`

[3]Robot Learning Group, Max-Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany `mail@jan-peters.net`

hand with appropriate orientation and opening in order to perform the grasp.

However, existing pixel-wise approaches on existing datasets suffer from ambiguities due to multiple overlapping grasping boxes with different orientations. As a result, they fail to accurately predict grasps with a significant Intersection over Union (IoU) w.r.t. the ground truth ones, and their performance is saturated into small thresholds.

To tackle these limitations, we present a novel *orientation-attentive* method for predicting pixel-wise grasp configurations from depth images. We revisit the notion of grasp map representation by introducing an augmented version for disentangling the different possible orientations of grasp points. We classify the grasps according to their orientations into discrete bins, while we regress their values for a continuous estimation of the grasp orientation per bin. Moreover, this orientation map acts as a bin-wise *attention mechanism* [12] over the quality map, to teach a CNN-based model to focus its attention on the true grasp points of the object. The proposed method, named *ORANGE* (ORientation AtteNtive Grasp synthEsis), is model-agnostic, since it can be combined with any CNN-based approach capable of performing segmentation, while boosting their performance in achieving accurate grasp predictions. *ORANGE* achieves state-of-the-art results on the most challenging grasping dataset [5], acquiring 94.71% using only the depth modality, against all other related methods. Knowledge from *ORANGE* can also be easily transferred and lead to significantly accurate predictions on the much smaller dataset Cornell [4], achieving 91.1% accuracy, which constitutes state-of-the-art performance for the depth-based grasp synthesis.

II. Related Work

Several works exist on grasping detection, synthesis and planning, both analytical [13] and data-driven [2], the latter being most relevant to our work. The early work of [14] addresses the problem of extracting features for grasp point detection and quality prediction. The authors of [15] propose a supervised approach from synthetic data to identify a few points in object images corresponding to good locations for grasping. Object pose learning and grasp densities are used in [16]. A new representation of grasps as rectangles, which define the 3D grasp location and orientation along with the gripper's opening and width for approaching the object, is introduced in [17].

In the era of DL the 'grasp rectangles' became the main grasp representation, employed in cascaded deep networks for detecting possible grasp candidates on one network, while the second one refined the detection [4]. This work, also, extended the *Cornell* data set, which has been extensively used by the subsequent works. Object segmentation is used in [18], to feed later the object into a CNN for detecting a grasp rectangle. Gradually, even more works have focused on performing CNN-based regression on RGB-D data for detecting grasping boxes [19]–[21]. In [22], the authors propose a CNN architecture that considers several outputs about the grasping configuration; namely not only the bounding box, but also the orientation and 'graspability' of grasp points are predicted. In the meanwhile, the convenient representation of grasp rectangles led to the creation of *Jacquard* [5] dataset. *Jacquard* is more complex, much larger in terms of the number of objects included (subset of ShapeNet [23]), and way more densely annotated, since the annotations have been generated from a simulated environment.

From that point on, research works dealing with grasp detection and synthesis are divided into those that take advantage of the successful Region Proposal Networks (RPNs) in object detection [6], and those who employ a pixel-wise approach for detecting several possible grasp points on objects [10]. The "easy" grasp representation as rectangles and the success of [6] in detecting objects in RGB images have made the application of RPNs in grasp detection natural. Such an approach is employed in [9], using a method not only for regressing the grasp bounding boxes, but also classifying them as grasps/no grasps. In parallel, [7] uses oriented anchor boxes to predict possible orientations of the grasp rectangles for the *Cornell* dataset. Following this work, [8], [24] uses two RPNs, for detecting initially multiple objects, and then define the grasping boxes per detected object. Note that [7], [8] is the first to report results also on *Jacquard* dataset using RPNs. However, these approaches employ very large DL frameworks, requiring heavy computational power and their performance in inferring grasps is rather slow [11].

On the other hand, works like [25] use a CNN framework trained to predict the robustness of candidate grasps from depth images using a large data set of synthetic point clouds. Synthetic depth data are also employed in [26], where a CNN takes as input a single depth image of an object and outputs a score for each possible grasp, taking also into account the gripper pose uncertainty. GGCNN [10], [11] is a small CNN employing dilated convolutions and is the first approach to predict pixel-wise maps, regarding the grasps configurations and their qualities. Following works propose the combination of two parallel U-Nets [27] for RGB-D data fusion [28] or they use fully-CNNs for high resolution RGB-D images [29] in order to perform pixel-wise predictions.

Our work follows the pixel-wise paradigm and further extends it with the introduction of augmented grasp maps. We leverage intuition by the analysis of the Jacquard dataset, to discretize the maps into bins of different orientations for disentangling the common issue of multiple orientations per grasp. We continue by introducing a binary graspness map accounting for the graspability of each pixel in the image. A continuous representation of the quality map, ensures better reconstruction ability of the ground truth maps and subsequently more efficient learning of the maps regression. Inspired by the RPN methods as well, we combine classification into orientation bins with the original regression problem, providing better reconstruction of the grasping areas across different orientations. We further exploit the orientation maps as an attention mechanism for the model to focus on the areas with the highest probability of acquiring a valid grasp point for all possible orientations (Fig. 1).

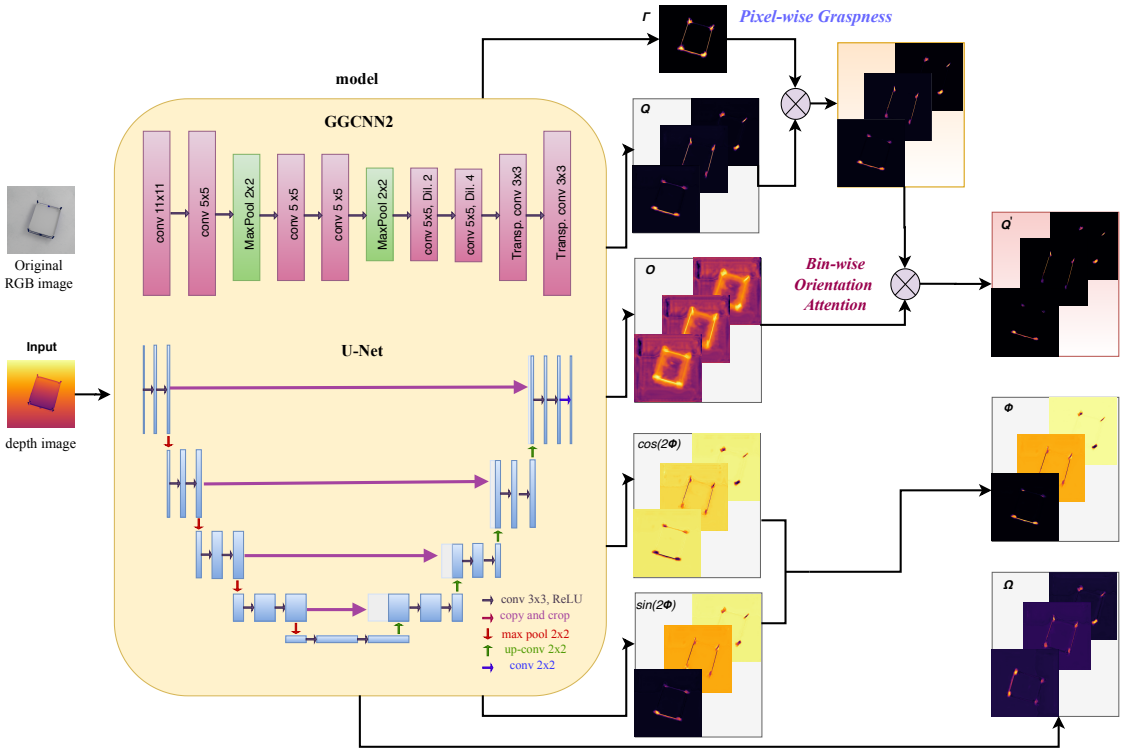

Fig. 2: Overview of the ORANGE architecture. An augmented grasp map representation, that fuses continuous and discrete information, drives the transformation of a depth image into a set of grasping boxes. The discretized orientation map serves as an attention force that focuses on local maxima.

III. Problem Statement

Given an object's RGB and/or depth image, grasp synthesis consists of finding the optimal grasp configuration $\mathbf{g} = \{x, y, z, \phi, w, q\}$, containing the grasp pose $\{x, y, z\}$ in the Cartesian space, to which the center of the robot's hand or gripper should be aligned, the orientation ϕ around the z axis and the required fingers' or jaws' opening (width) w. A quality measure q characterizes the success of the respective grasp configuration.

Particularly for grasp synthesis from depth images $\mathbf{I} \in \mathbb{R}^{HxW}$ of height H and width W, we aim to estimate the grasp $\bar{\mathbf{g}} = \{u, v, \bar{\phi}, \bar{w}, q\}$, where $\{u, v\}$ correspond to the pixel coordinates in image space, $\bar{\phi}$ is the orientation relative to the camera reference frame, and $\bar{w}$ is the respective grasp width in the image space. With known intrinsic camera parameters and the mapping between image and robot/world frame, grasp synthesis can be expressed as the problem of finding the *grasp map* [11]:

$$\mathbf{G} = \{\Phi, \Omega, Q\} \in \mathbb{R}^{3 \times H \times W} \tag{1}$$

where Φ, Ω, Q are each of them a map in $\mathbb{R}^{H \times W}$, containing the values of $\bar{\phi}, \bar{w}, q$ respectively $\forall \{u, v\} \in \mathbf{I}$. $\mathbf{G}$ can be approximated through a learnt mapping $\mathbf{I} \xrightarrow{\hat{f}_\theta} \mathbf{G}$ using a deep neural network (θ being its weights). The best visible grasp configuration can now be estimated as $\bar{\mathbf{g}}^* = \max_{\mathbf{Q}} \mathbf{G}$.

IV. Method

Real-world objects with peculiar morphology can be grasped in multiple angles even around nearby physical points. As a result, the constructed grasp maps of pixel-wise learning approaches [11], [28], [29] are prone to discontinuities that cause saturated performance. Motivated by such observations on the challenging *Jacquard* dataset [5], we introduce an augmented grasp map representation that fuels both the continuous grasping orientation estimation, commonly treated as a regression problem, and a discrete classification problem. We show that by discretizing angle maps we are able to disentangle cases of multiple possible orientations, even when the annotated bounding boxes are highly overlapping. More importantly, the binarized orientation map is used as an attention mechanism on the quality map, shifting the network's focus towards real grasping centers. Moreover, the proposed method is model agnostic and can be appended to any state-of-the-art architecture (Fig. 2) to boost its performance. We first present an extensive analysis on *Jacquard*, as well as previous approaches on grasp map estimation, before we discuss our specific design choices for tackling the problem of robotic grasp synthesis from depth images.

A. Revisiting Grasp Map Representation on Jacquard

Jacquard is currently one of the most diverse and densely annotated grasping datasets with 54000 images and 1.1 million grasp annotations, also considering different jaw sizes. Other options include Dex-Net [25], that contains millions of images of 1500 object classes and a single grasp annotation per image and *Cornell* [4], with only 885 images for 240 objects and 8019 grasp human-labelled annotations. Smaller dataset sizes and little varying annotations make

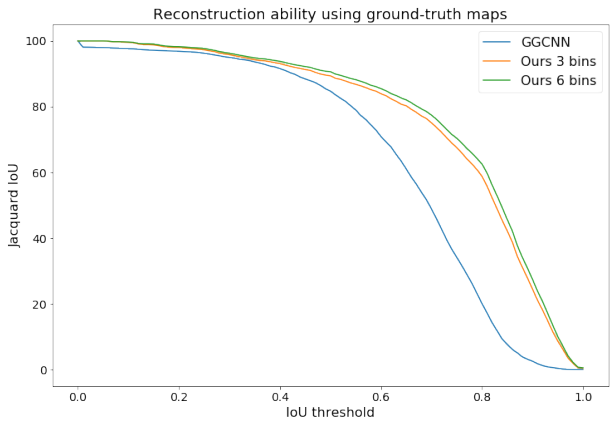

Fig. 3: IoU score across all different thresholds using three different ground-truth maps: GGCNN, ours with 3 orientation bins and ours with 6 bins. The proposed grasp maps saturate smoothly towards larger thresholds and lead to a more robust representation of the annotations.

Parameter	GGCNN [11]	Our method	Benefit
Discretize angle		✓	Less overlaps
Map dimensions	2-d	3-d	Less overlaps
Quality map	Binary	Non-binary	Accurate centering
Postprocessing	Gaussian filtering	None	Faster
Picked jaws' size	All	Minimum	Improved segmentation
Handle ovelaps	Overwrite	Keep minimum	Rigid map spaces
Max. IoU@0.25	96.24	97.32	Better reconstruction
Max. IoU@0.30	94.96	95.83	Better reconstruction
Max. IoU@0.50	84.72	89.38	Better reconstruction

TABLE I: Comparison of design choices between the proposed method and prior literature [11] concerning grasp map construction. Our real-valued maps resolve ambiguities due to overlaps, leading to better reconstruction ability.

the performance of DL methods questionable in terms of generalization to different images and object configurations, therefore making *Jacquard* an obvious dataset choice.

Jacquard represents grasps as rectangles with given center, angle, width (gripper's opening) and height (jaws' size). Nonetheless, the annotations are simulated and not human-labeled, resulting into multiple overlapping boxes considering all possible grasp orientations per grasp point and many different jaw sizes. To make matters worse, box annotations are invariant to the jaws size and for a given center and angle, the opening is independent of all the possible jaw sizes - indeed, only 3 out of the 54000 images contain a box deviating from this norm. Therefore, the jaw size is a free variable to be arbitrarily chosen during evaluation. As we show in Fig. 3, even the ground-truth grasp maps cannot score perfectly without knowing the jaws' size.

The authors of [11] tackle these challenges by generating pixel-wise quality, angle and width maps (see sec. III), by iterating over the annotated boxes and stacking binary maps, equal to the value of interest inside the box and zero elsewhere. Since the quality map is a binary map, the result of such stacking is indifferent to the order of the boxes and equivalent to iterating only on the boxes with the maximum jaws' size. For angle and width maps however, overlapping boxes with different centers and angles will be overwritten

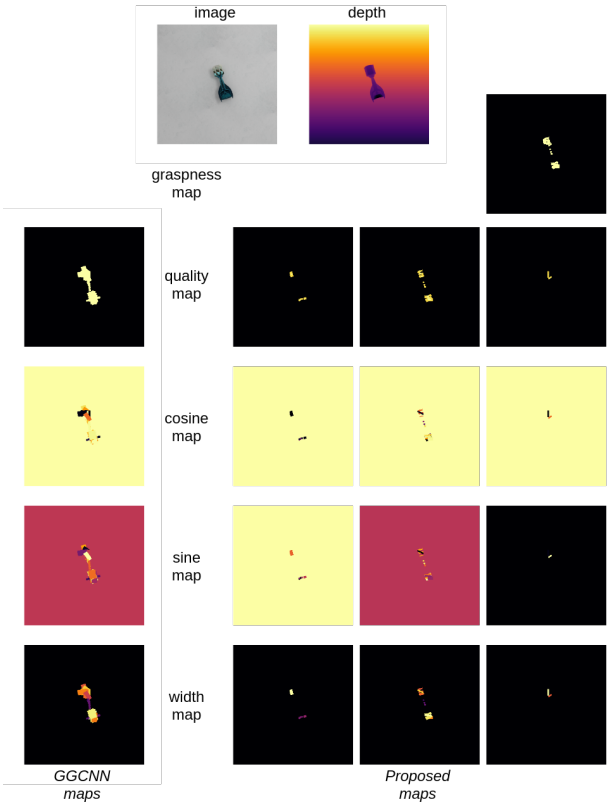

Fig. 4: Comparison of the target representation for GGCNN (left column) and the proposed 3-bin method (right 3 columns). GGCNN maps suffer from highly overlapping boxes that lead to discontinuities, while their binary quality map is a dense region that lies further than the object's boundaries. Contrary to that, our maps are sparse and clear from overlaps, while the quality maps contain rigid areas with a well-defined maximum. Our "graspness" map roughly approximates the object's segmentation mask.

by the box that appears later in the annotations, leading to discontinuities. Lastly, a binary quality map does not ensure a valid maximum: all non-center points inside an annotated box are maxima as well, and have equal probability of being selected as a grasp center.

Due to all these choices w.r.t. the annotations handling, an ideal regressor that perfectly predicts the ground-truth maps during evaluation is not able to reconstruct the annotated bounding boxes. We measure that such a model scores only $\sim 96.2\%$ accuracy using the Jaccard (IoU) index at the 0.25 threshold, while its performance degrades rapidly towards larger thresholds (Fig. 3). Lastly, not surprisingly, if we shuffle the order we access the annotations, this hypothetical network's performance changes.

B. Grasp Maps with Discretized Orientation

To tackle the aforementioned challenges, we adapt ideas from recent works on pixel-wise grasp synthesis [11], [28], [29] and partition the angle values into N bins, so as to minimize the overlaps of annotated boxes. Since we are dealing with antipodal grasps, it is sufficient to predict an angle in the range of $\{-\pi/2, \pi/2\}$. We, thus, proceed to construct 3-dimensional maps of size $H \times W \times N$, where each bin corresponds to a range of $180/N$ degrees. Note,

however, that we do not discretize the angles' values: we instead place them inside the corresponding bins. For the remaining overlaps, we pick the value with the smallest angle, ensuring that the network is trained on a valid ground-truth angle value, instead of some statistics of multiple values (e.g. mean or median), while remaining invariant to the order of the annotations.

To overcome the information loss on the construction of binary maps, we create soft quality maps that contain ones on the exact positions of the centers of the boxes, while their values degrade moving towards the boxes' edges (Fig. 4). We find that this is significant for the trained networks to learn to maximize the quality value on the grasp points. Such networks saturate more smoothly in performance towards greater threshold values (Fig. 3), while the exact prediction of a center counterbalances the necessity of strong Gaussian filtering [11] and consequently reduces post-processing time.

One remaining issue is the multiple instances of the same grasp centers and angles using different jaw sizes. In such cases, the boxes with larger jaws' size will overwrite the smaller boxes, thus making it sensible that we pick a single size. We experiment with both the minimum and the maximum size and pick the smallest, as it is the one closer to the boundaries of the objects' shape. Intuitively, the annotated quality map gives a rough estimate of the objects segmentation mask, which is an information important for extracting grasp regions, as also noted in previous works [11]. As for the jaw size during evaluation, we adopt the half jaw size presented in [11] to be directly comparable. Although having to estimate such a parameter hurts performance, our approach still achieves large reconstruction ability. Table I briefly summarizes a comparison between our grasp maps and prior works.

We reformulate Eq. (1) under the constraint of N orientation bins:

$$\mathbf{G} = \{\Phi, \Omega, Q, O, \Gamma\} \in \mathbb{R}^{(4\times N)+1\times H\times W} \tag{2}$$

where:

- $\Phi \in \mathbb{R}^{N\times H\times W}$ is the angle map. For facilitating learning, we adopt the angle encoding suggested by [11], [30] into the cosine, sine components that lie in the range of $[-1,1]$. Since the antipodal grasps are symmetrical around $\pm\frac{\pi}{2}$, we employ the sub-maps for $cos(2\Phi_i)$ and $sin(2\Phi_i)$ $\forall\Phi_i$ with $i \in N$ bins. The angle maps are then computed as: $\Phi = \frac{1}{2}\arctan\frac{sin(2\Phi)}{cos(2\Phi)}$.
- $\Omega \in \mathbb{R}^{N\times H\times W}$ represents the gripper's width map.
- $Q \in \mathbb{R}^{N\times H\times W}$, is a real-valued quality map, where '1' indicates a grasp point with maximum visible quality.
- $O \in \mathbb{R}^{N\times H\times W}$ is a binary orientation map where '1' indicates a filled angle bin in the respective position.
- $\Gamma \in \mathbb{R}^{1\times H\times W}$ is the pixel-wise 'graspness' map. This binary map contains '1s' only in the annotated grasp points of the object w.r.t. the image $\mathbf{I}$, and helps assessing the graspability of the pixels, i.e. the probability of representing grasp points of the real world.

An example of the constructed grasping maps, as well as a comparison with those of [11] can be seen in Fig. 4.

C. ORANGE: Orientation-attentive grasp synthesis

The proposed framework, *ORANGE* is depicted in Fig. 2. *ORANGE* is model-agnostic; it suffices to employ any CNN-based model that has the capacity to segment regions of interest. Assuming such a model, an initial depth image is processed to output an augmented grasp map $\mathbf{G}$, as defined in (2). Φ, Ω, Q, O and Γ are combined to reconstruct the center, angle and width information.

Training: Each map is separately supervised: we minimize the Mean Square Error (MSE) of the real-valued Q, $cos(2\Phi)$, $sin(2\Phi)$ and Ω and their respective ground-truths, and we force a Binary Cross-Entropy loss (BCE) on O and Γ. Next, we employ an attentive loss that directly minimizes the MSE between $Q*O$ (element-wise multiplication) and the ground-truth quality map. This attention mechanism drives the network's focus over regions of the feature map that correspond to filled bins and thus regions nearby a valid grasp center. We found it useful to scale the MSE losses by multiplying them with the number of bins N. The total objective function takes the form:

$$\begin{aligned}\mathscr{L} = {} & \mathscr{L}_{BCE}(O) + \mathscr{L}_{BCE}(\Gamma) \\ & + N * \{\mathscr{L}_{MSE}(Q) + \mathscr{L}_{MSE}(cos(2\Phi)) + \mathscr{L}_{MSE}(sin(2\Phi)) \\ & + \mathscr{L}_{MSE}(\Omega) + \mathscr{L}_{MSE}(Q*O)\}\end{aligned} \tag{3}$$

Inference: First, Q and Γ are multiplied to obtain a graspness-refined quality map. This can be viewed as a pixel-wise prior regularization, where Γ is the prior probability of a pixel to be a grasping point and Q is the posterior, measuring its grasping quality. This product is multiplied by O to filter out values in empty bins, resulting in the final quality map, $Q*\Gamma*O$. Fig 5 shows the intermediate effects of the quality map refinement on real images. Finally, we choose the optimum grasping center as the global maximum of the quality map and retrieve the respective values of Φ and Ω to reconstruct a grasping box.

Models employed by ORANGE: We embed *ORANGE* to two off-the-shelf architectures, GGCNN2 [11] and the larger U-Net [27], as depicted in Fig. 1. GGCNN2 [11], an improved version of GGCNN [10], is characterized by a minimal architecture employing dilated convolutional layers, previously used in semantic segmentation tasks. U-Net is commonly used for image segmentation and consists of a contracting path capturing context and a symmetric expanding one for localization. While these architectures are totally different in terms of capacity, we show that both can perform significantly better when trained under *ORANGE* framework.

V. Experiments & Discussion

We validate *ORANGE*'s effectiveness on Jacquard and Cornell datasets, following the standard 90/10% split for training and testing respectively and no data augmentation. The input depth images are resized from their initial size to

320×320, to allow for higher training speed, but without much information loss due to resizing. We train the whole network end-to-end using initial learning rate 0.002, that decays exponentially and weight decay 0.0001. We have employed an early stopping strategy for training, to prevent model overfitting. On Cornell especially, we 'warm-start' the network [31] by initializing it with pretrained weights on Jacquard and then perform training without any data augmentation.

The framework was trained in a NVIDIA-DGX-1 station, while the inference was executed on a PC running Ubuntu 18.04 with a 3.6 GHz Intel Core i7-6850 CPU and NVIDIA GeForce GTX 1080 graphics card. The code for *ORANGE* is available online to allow for reproducibility of results at `https://github.com/nickgkan/orange`.

Following prior literature, we adopt IoU@0.25, as our evaluation metric. We also report results over greater thresholds, namely 0.30 and 0.50, a practice that is in general not followed by most of the related works on the same topic. Continuing, we compare our proposed method with all known works for *Jacquard* and *Cornell* datasets. Note that, all related works, except for [11], have employed multimodal data (i.e. RGB, RGB-D or RGD) for predicting grasps. We will show how taking advantage of the orientation discretization and the attentive treatment over the quality grasp maps delivers impressive results using a unimodal channel.

A. Ablation study

For motivating the design choices in the augmented grasp map representation and the benefits of the individual components in the *ORANGE* framework (Fig. 2), we report an ablation study in Table II. To allow for maximum comparability between the different configurations, we employ as base model for *ORANGE* the U-Net architecture, as it has the capacity to learn the multiple grasp representations we have introduced in Section IV-B.

Inspecting the different combinations of the individual components, as described in Section IV-C, we can see that the full proposed model, with the pixel-wise graspness and the bin-wise orientation attention, performs better when using 3 bins compared to 6 for the lowest accuracy threshold, since discretizing the angle space into N bins means N regressions for the model to learn and N classes to identify. In particular for the angle range of $\{-\pi/2, \pi/2\}$ in the antipodal grasps, the $N = 6$ discretization, divides into bins of 30^o range, i.e. there are smaller differences in the appearances among neighboring orientations, while it requires 25 regressions. The higher the number of bins the more difficult and confusing it is for the network to learn to disentangle the multiple grasping orientations.

The application of the pixel-wise graspness Γ on the quality maps Q has an evident benefit on the model for the low accuracy thresholds, while the effectiveness of the model degrades on the 0.50 threshold. The graspness loss focuses locally on the best grasp points and restricts the exploration of the feature space, thus decreasing the grasp box area.

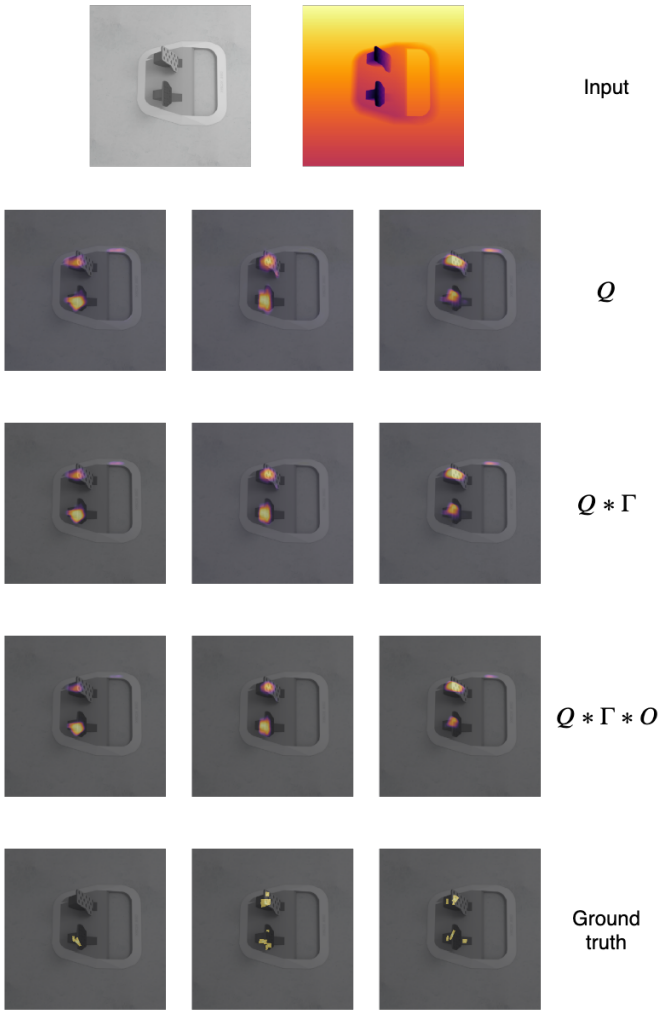

Fig. 5: Intermediate results when reconstructing the quality map. Q alone is noisy since it is the output of a regression problem. Multiplication by Γ smooths the quality map pixel-wise, while O filters outliers bin-wise. The final estimation is much clearer and closer to the ground-truth.

The selection of the jaw size during the construction of the ground truth maps also affects the performance of *ORANGE* over all accuracy thresholds. As we have already discussed in Section IV-A, the jaw size is a feature indifferent of the object. Our intuition is that, the min jaw size produces bounding boxes closer to the boundaries of the object, and thus can be easier to be segmented by U-Net.

An important decision choice was the usage or not of binary values for the quality maps in the ground truth data synthesis, a method used in [11]. As we can inspect from Table II, using binary maps in *ORANGE* produces 4% less accurate grasp predictions w.r.t. to our approach. However, the use of binary ground truth maps gives better results for the 0.50 threshold. Binary maps do not penalize the neighboring pixels of a grasp center and can thus return a nearby pixel as the center; this is not correct but the IoU metric will not penalize the response. This greedy solution creates higher confusion about which pixel point is actually

Network	Design Choices									Accuracy Threshold		
	regression	graspness	bin class.	attention	binary map	max jaw size	min jaw size	$N=3$	$N=6$	0.25	0.30	0.50
U-Net [27]	✓	✓	✓	✓			✓	✓		**94.71**	**92.65**	70.37
	✓	✓	✓	✓			✓		✓	91.51	89.07	71.05
	✓		✓	✓			✓	✓		92.34	90.44	77.65
	✓	✓	✓				✓	✓		93.36	90.90	70.95
	✓	✓	✓	✓		✓		✓		94.11	91.83	68.97
	✓	✓	✓	✓	✓		✓	✓		91.75	90.27	**79.92**
	✓				✓					89.85	88.13	76.14
GGCNN2 [11]	✓	✓	✓	✓			✓	✓		**88.92**	**85.94**	**67.18**
	✓		✓	✓			✓	✓		87.88	85.52	63.34
	✓				✓					85.23	82.67	62.68

TABLE II: Ablation study over different design choices for both *ORANGE* implementations with U-Net and GGCNN2. For each instantiation, the accuracy (%) is reported over different thresholds of the Jaccard index.

a grasp point. However, *ORANGE* seems to mitigate this confusion achieving an accuracy of 91.75% (although 3% lower than using our approach, Sec. IV-B), while a U-Net implemented as suggested in [11], succeeds a 89.85% at the 0.25 accuracy metric threshold.

Concluding this discussion about the design choices in *ORANGE*, we provide the accuracy scores for the GGCNN2 in *ORANGE*, which we achieve to improve from 85.23% in the original implementation into 88.92% using the orientation attentive method and the new grasp map representation. This result can only confirm the characterization of *ORANGE* as model-agnostic, since it boosts the performance of each embedded CNN-based architecture.

B. Comparative results w.r.t. literature

After having validated the proposed framework over several different design combinations, we compare the best configuration with the results found in literature. Table III presents the comparative results for the *Jacquard* dataset. As we have already discussed, Jacquard is a relatively new and contains complex annotations to be easily handled. This is the reason why only a few works exist employing this data-set. In the previous section, we have already established the superior behavior of the *ORANGE* framework with the GGCNN2 backbone. Inspecting the results of Table IV, it is evident that the *ORANGE* implementation with a U-Net model achieved state-of-the-art performance compared to all existing approaches that employ RGB or RGD data. Namely, we achieve 94.7% accuracy using depth data, while the previous higher accuracy score was 93.6% for multi-modal RGD data [8].

These impressive results rely on two factors that we have established with the proposed method: Firstly, the annotation handling of the Jacquard data set and the augmented grasp map representation. Secondly, the bin-wise attention of the orientation estimation over the quality maps, has benefited the disentanglement of the overlapping annotations due to multiple orientations per grasp point, leading to better learning. We expect even larger margins of accuracy if we also use the RGB channel, however this is beyond the scope of our work that focuses on improving the grasp map representation.

Subsequently, we wanted to check the performance of *ORANGE* for the Cornell data-set. Cornell data set is rather small, it contains hand annotations and therefore many errors. Usually, all implementations in literature use massive data augmentation to achieve a good learning performance. However, such results are not easy generalisable to the real-world. We, on the other hand, wanted to evaluate the transferability of our network to a new data-set. To achieve this we use the technique of 'warm-starting' [31], a method usually used for initializing the weights of a pretrained network, while not allowing for catastrophic forgetting of the previously learned representations. We trained only on 90% of the Cornell data set, without using any data augmentation, and achieved a rather high testing accuracy.

methods	modality	Accuracy (%)
Morisson et al. [11]	D	85.2
Depierre et al. [5]	RGB-D	74.2
Zhou et al. [7]	RGB	91.8
Zhou et al. [7]	RGD	92.8
Zhang et al. [8]	RGB	90.4
Zhang et al. [8]	RGD	93.6
ORANGE with GGCNN2 (ours)	D	**88.9**
ORANGE with U-Net (ours)	D	**94.7**

TABLE III: Comparative results for the *Jacquard* dataset.

Table IV presents the comparative results for the Cornell data set. Specifically, the depth-based *ORANGE* approach achieves 91.1% with U-Net and 87.5% with GGCNN2 implementation. This accuracy score constitutes state-of-the-art performance for the depth based methods. GGCNN2 [11], for example achieves 78.6% when train/tested on Cornell, which *ORANGE* improves by 9%. It is natural to expect an amelioration of our results, with data augmentation, however, this is not reported in this work, as our main focus is to stress the benefits of the *ORANGE* framework and its generalization abilities.

VI. Conclusions & Future work

Performing robust robotic grasping remains still a challenging problems. Various object morphologies, in terms of sizes and shapes, correspond to various plausible grasping points. Robotic vision wishes to tackle this issue by employing deep neural networks on large datasets of robot-related tasks. However, current datasets on object grasping are either prohibitively small and hand-annotated, like Cornell, or are satisfactorily large and simulated automatically, like

methods	modality	Accuracy (%)
Morisson et al. [11]	D	78.6
Depierre et al. [5] trained on *Jacquard*	RGB-D	81.92
Depierre et al. [5] trained on *Cornell*	RGB-D	86.88
Zhou et al. [7]	RGB	97.7
Zhang et al. [8]	RGB	93.6
Zhang et al. [8]	RGD	92.3
Guo et al. [22]	RGB	93.2
Chu et al. [9]	RGB	94.4
Chu et al. [9]	RGB-D	96.0
Wang et al. [29]	RGB-D	94.4
ORANGE with GGCNN2 (ours)	D	**87.5**
ORANGE with U-Net (ours)	D	**91.1**

TABLE IV: Comparative results for the *Cornell* data set. *ORANGE* was trained on Cornell with weights 'warm-started' from the Jacquard dataset. All other works, unless stated otherwise, were trained/tested on Cornell using data augmentation.

Jacquard, leading to hundreds of grasp point annotations per object, overlaps and multiple grasping orientations. These challenges cannot be mitigated by neural regressors. After a thorough analysis over the challenging Jacquard, we introduced an augmented grasp map representation, that discritizes all grasp orientations into bins, thus transforming the problem of grasp synthesis into an orientation classification problem combined with bin-wise real-value regression.

Leveraging on the new grasp map representation, we built *ORANGE*, an orientation attentive mechanism for grasp synthesis. This attention mechanism employs the outcome of the bin-wise orientation map, and acts over the quality map, in order to draw the network's focus on areas with higher probability of containing a grasping center. *ORANGE* is, also, model-agnostic, as it can be combined with any CNN-based architecture capable of performing mask segmentation, boosting their performance significantly. We reported extensive experimental results, which justified the effectiveness of *ORANGE*, that achieves state-of-the-art performance, 94,71% on Jacquard using only the depth modality.

In the future, we will investigate ways of performing soft attention for extracting useful features across all maps. We also aim to identify good quality grasps on objects, by reasoning about the objects' shape, texture and category.

REFERENCES

[1] A. Sahbani, S. El-Khoury, and P. Bidaud, "An overview of 3d object grasp synthesis algorithms," *RAS*, vol. 60, no. 3, 2012, Autonomous Grasping.

[2] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesisa survey," *IEEE Trans. on Robotics*, vol. 30, no. 2, 2014.

[3] S. Ekvall and D. Kragic, "Learning and evaluation of the approach vector for automatic grasp generation and planning," in *IEEE Int'l Conf. on Robotics and Automation*, 2007.

[4] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *IJRR*, vol. 34, no. 4-5.

[5] A. Depierre, E. Dellandra, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *IEEE Int'l Conf. on Intelligent Robots and Systems*, 2018.

[6] Shaoqing R., Kaiming H., Ross G., and Jian S., "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*.

[7] X. Zhou, X. Lan, H. Zhang, Z. Tian, Y. Zhang, and N. Zheng, "Fully convolutional grasp detection network with oriented anchor box," in *IEEE Int'l Conf. on Intelligent Robots and Systems*, Oct 2018.

[8] H. Zhang, X. Lan, S. Bai, X. Zhou, Z. Tian, and N. Zheng, "Roi-based robotic grasp detection for object overlapping scenes," in *IEEE Int'l Conf. on Intelligent Robots and Systems*, 2019.

[9] F. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robotics & Automation Letters (R-AL)*, vol. 3, no. 4, Oct 2018.

[10] D. Morrison, P. Corke, and J. Leitner, "Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach," in *RSS*, 2018.

[11] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *IJRR*, vol. 39, no. 2-3.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing*. 2017.

[13] K. B. Shimoga, "Robot grasp synthesis algorithms: A survey," *IJRR*, 1996.

[14] I. Kamon, T. Flash, and S. Edelman, "Learning to grasp using visual information," in *IEEE Int'l Conf. on Robotics and Automation*, 1996, vol. 3.

[15] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *IJRR*, vol. 27, no. 2.

[16] R. Detry, E. Baseski, M. Popovic, Y. Touati, N. Kruger, O. Kroemer, J. Peters, and J. Piater, "Learning object-specific grasp affordance densities," in *ICDL*, 2009.

[17] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *IEEE Int'l Conf. on Robotics and Automation*, 2011.

[18] Z. Wang, Z. Li, B. Wang, and H. Liu, "Robot grasp detection using multimodal deep convolutional neural networks," *Advances in Mechanical Engineering*, vol. 8, no. 9, pp. 1687814016668077, 2016.

[19] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *IEEE Int'l Conf. on Robotics and Automation*, 2015.

[20] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," in *IEEE Int'l Conf. on Intelligent Robots and Systems*, 2017.

[21] J. Watson, J. Hughes, and F. Iida, "Real-world, real-time robotic grasping with convolutional neural networks," in *Towards Autonomous Robotic Systems*, Cham, 2017, Springer International Publishing.

[22] D. Guo, F. Sun, H. Liu, T. Kong, B. Fang, and N. Xi, "A hybrid deep architecture for robotic grasp detection," in *IEEE Int'l Conf. on Robotics and Automation*, May 2017, pp. 1609–1614.

[23] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An information-rich 3d model repository," 2015.

[24] H. Zhang, X. Lan, S. Bai, L. Wan, C. Yang, and N. Zheng, "A multi-task convolutional neural network for autonomous robotic grasping in object stacking scenes," in *IEEE Int'l Conf. on Intelligent Robots and Systems*, 2019.

[25] 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.

[26] E. Johns, S. Leutenegger, and A. J. Davison, "Deep learning a grasp function for grasping under gripper pose uncertainty," in *IEEE Int'l Conf. on Intelligent Robots and Systems*, 2016.

[27] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351 of *LNCS*.

[28] Y. Song, J. Wen, Y. Fei, and C. Yu, "Deep robotic prediction with hierarchical rgb-d fusion," 2019.

[29] S. Wang, X. Jiang, J. Zhao, X. Wang, W. Zhou, and Y. Liu, "Efficient fully convolution neural network for generating pixel wise robotic grasps with high resolution images," in *IEEE Int'l Conf. on Robotics and Biomimetics*, Dec 2019.

[30] K. Hara, R. Vemulapalli, and R. Chellappa, "Designing deep convolutional neural networks for continuous object orientation estimation," 2017.

[31] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *CoRR*, vol. abs/1612.00796, 2016.