Revisiting Grasp Map Representation with a Focus on Orientation in Grasp Synthesis

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

***Abstract*—Innate morphological characteristics of objects may obfuscate the learning of robotic grasping. Even simple structures (e.g. cyclical) offer a wide range of plausible grasping orientations, creating ambiguities for neural regressors. We investigate and unfold multiple such conflicts on the challenging dataset Jacquard and derive a novel grasp map representation, suitable for pixel-wise synthesis. Our augmented maps disentangle co-occurent grasping orientations around the same point by partitioning the angle space into multiple bins. Subsequently, we propose the ORientation AtteNtive Grasp synthEsis (ORANGE) framework, that jointly addresses classification into bins and angle-value regression. The constructed bin-wise orientation maps further serve as an attention mechanism for areas with higher *graspness*, i.e. probability of being a true grasp point. This procedure is model-agnostic and can be embedded to any existing architecture to boost its performance. Namely, we report a new state-of-the-art 94.71% performance on Jacquard, with a simple U-Net using only depth images.**

I. Introduction

Grasping inherently different objects in unstructured environments is an essential component of the skill-set that robots shall excel in so as to be effectively integrated into human-inhabited environments [1], [2]. The problem has been explored both in an analytical [3] and data-driven fashion [4], with Deep Learning (DL) assigning an increasing advantage to the latter, powered by large datasets [5], [6] of common graspable objects, suitable for robotic hands and grippers.

Several data-driven approaches have borrowed ideas from computer vision to detect antipodal grasps on objects from RGB data [7]. These approaches predict and rank thousands of grasp candidates [8]–[10], requiring much computational resources, while they are limited to static environments and precise camera calibration. Other works rely on synthetic depth data [11] or point clouds [12] to predict the robustness of candidate grasps from depth images, possibly taking also into account the gripper pose uncertainty. Very promising have recently been the pixel-wise approaches [13], [14], that represent grasping configurations as dense maps.

*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

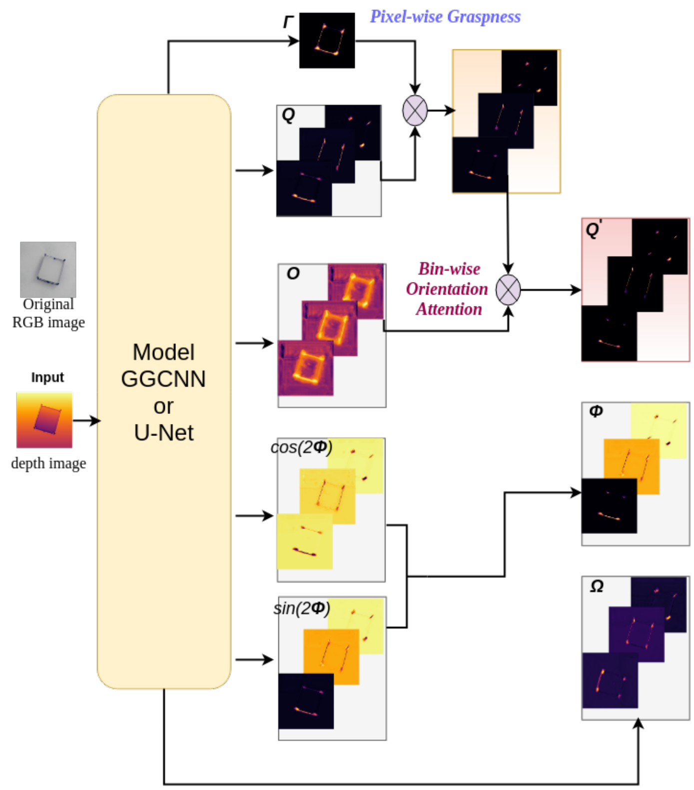

Figure 1: Overview of the ORANGE architecture. An augmented grasp map representation, that fuses continuous and discrete information, drives the transformation of the depth image into a set of grasping boxes. The discretized orientation map serves as an attention force that focuses on local maxima of graspness. Please refer to Sec. III for a thorough explanation of the symbols used in this figure.

Continuous and effective estimation of the *approach vector*, i.e. the orientation with which the robotic hand approaches the object, is fundamental to a safe and successful grasp execution, especially for reactive grasp planning, either in cases of a moving camera on the robotic arm, or when grasping moving objects. Intuitively, when humans observe an object, they reason about its shape and navigate their hand with appropriate orientation and opening in order to perform the grasp. However, even state-of-the-art pixel-wise approaches fail to model ambiguities due to multiple overlapping grasping boxes with different orientations.

To tackle these limitations, we present a novel *orientation-attentive* method for predicting pixel-wise grasp configurations from depth images. We revisit the grasp map representation by introducing an augmented version for resolving orientations' conflicts. We classify the grasps into discrete orientation bins and regress their values for a continuous estimation of the orientation per bin. This orientation map acts as a bin-wise *attention mechanism* [15] over the quality map, to guide the model's focus on the true grasp points of the object. The proposed method, named *ORANGE* (ORientation AtteNtive Grasp synthEsis) (Fig. 1), is model-agnostic; it can be combined with any approach capable of performing segmentation, boosting its performance in generating accurate grasp predictions. *ORANGE* surpasses all related methods on Jacquard [6] using only the depth modality.

II. Problem Statement

Grasp synthesis refers to finding the optimal grasp configuration $\mathbf{g} = \{x, y, z, \phi, w, q\}$, containing the grasp center $\{x, y, z\}$ to which the robotic hand 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. For a (depth) image $\mathbf{I}$, grasp synthesis is the problem of finding the *grasp map* [14]:

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

where Φ, Ω, Q are each of them a map in $\mathbb{R}^{H\times W}$, containing the pixel-wise values of ϕ, w, q respectively. $\mathbf{G}$ 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 $\mathbf{g}^* = \arg\max_{\mathbf{Q}} \mathbf{G}$.

III. Revisiting Grasp Map Representation

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 [14], [16], [17] are prone to discontinuities that cause saturated performance (Fig. 2). Motivated by such observations on the challenging Jacquard dataset [6], we introduce an augmented grasp map representation that fuels both the continuous grasping orientation regression problem and a discrete classification problem.

The Jacquard Dataset: Jacquard is currently one of the most diverse and densely annotated grasping datasets with 54000 images and 1.1 million grasp annotations. Grasps are represented as rectangles with given center, angle, width (gripper's opening) and height (jaws' size). 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, leaving it as a free variable to be arbitrarily chosen during evaluation.

The authors of [14] tackle these challenges by generating pixel-wise quality, angle and width maps, 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

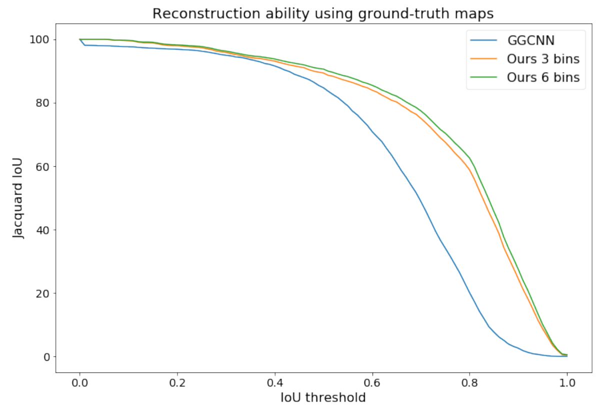

Figure 2: IoU score across per threshold for three ground-truth maps: GGCNN, ours with 3 orientation bins and 6 bins. The performance of the proposed maps saturates smoothly towards larger thresholds, demonstrating a more robust representation of the annotations.

angles will be overwritten 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, a hypothetical regressor that perfectly predicts the evaluation ground-truth maps fails to reconstruct the annotated bounding boxes and scores only $\sim 96.2\%$ using the Jaccard (IoU) index at the 0.25 threshold, while its performance degrades rapidly towards larger thresholds (Fig. 2). Not surprisingly, this performance is not invariant to shuffling the order we access the annotations.

Focusing on Orientation: To tackle the above challenges, we partition the angle values into N bins, 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 that we do not discretize the angles' values; we instead place them inside the corresponding bins. For the remaining overlaps, we pick the value corresponding to 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. 3). We find that this is significant for the trained networks to learn to maximize the quality value on the grasp points.

One remaining issue is the multiple instances of the same grasp centers and angles using different jaw sizes. We pick the smallest size, closer to the boundaries of the objects' shape. Intuitively, the annotated quality map gives a rough estimate of the segmentation mask (Fig. 3), information important for extracting grasp regions [14]. During evaluation, we adopt the half jaw size presented in [14] to be directly comparable. Although having to estimate such a parameter hurts performance,

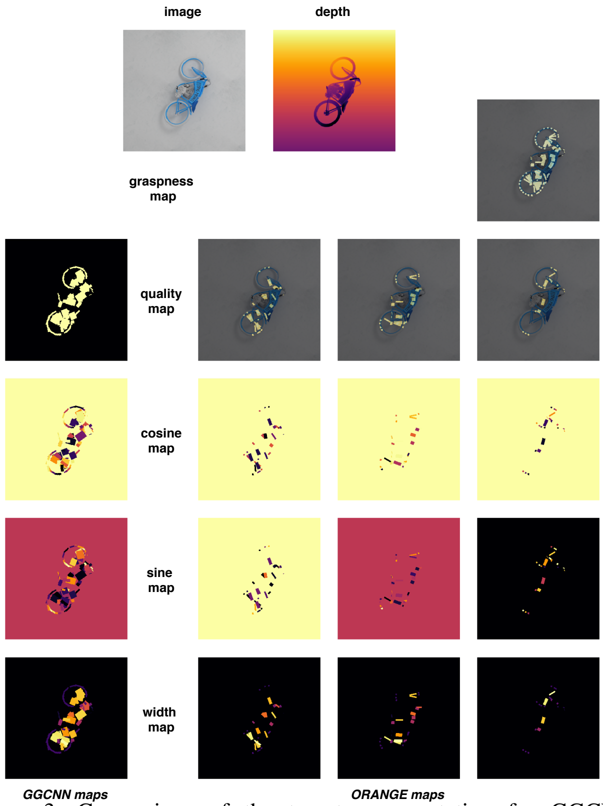

Figure 3: Comparison of the target representation for GGCNN (left column) and the proposed 3-bin method (right 3 columns). GGCNN maps suffer from highly overlapping boxes that lead to discontinuities, while their binary quality map is a dense region that lies further than the object's boundaries. Contrary 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.

our approach still achieves large reconstruction ability.

We thus reformulate Eq. (1) to consider 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. To facilitate learning, we adopt the angle encoding suggested by [14], [18] 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.

ORANGE: Orientation-attentive grasp synthesis: The proposed framework, *ORANGE* is depicted in Fig. 1. *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. We employ two off-the-shelf architectures, GGCNN2 [14] and the larger U-Net [19], both able of performing segmentation. While these architectures have totally different capacity, we show that both can perform significantly better when trained under the *ORANGE* framework.

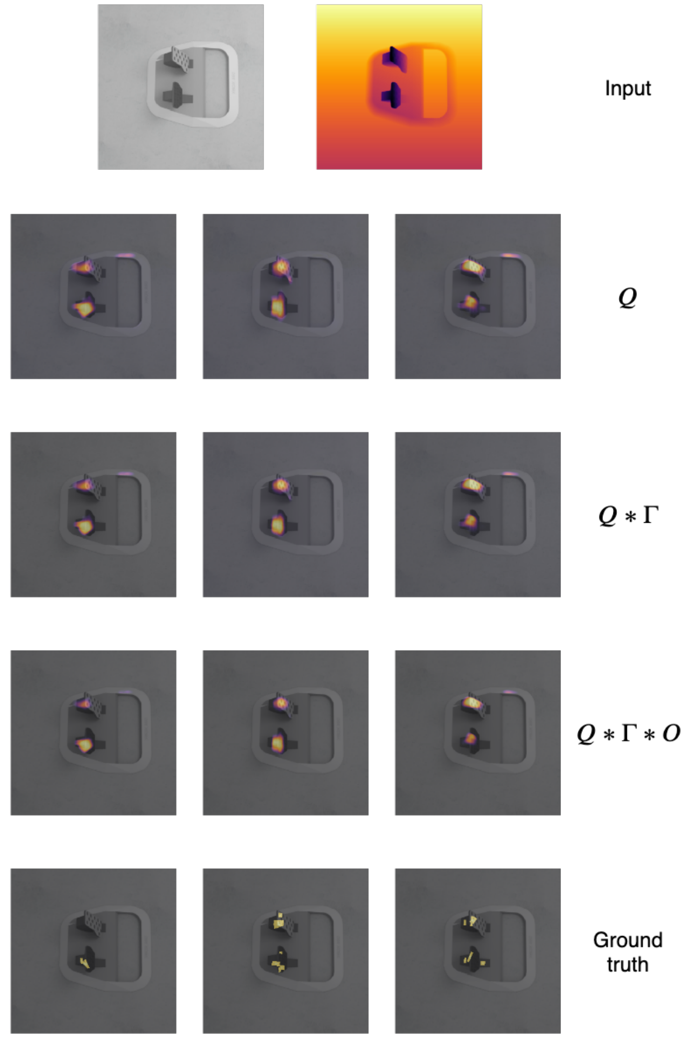

Figure 4: A closer look to quality map reconstruction. The regressed Q is noisy, but multiplication by Γ smooths the quality map pixel-wise, while O filters outliers bin-wise. The final estimation is much clearer and closer to the ground-truth.

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

	Design Choices									Threshold	
Network	regression	graspness	bin class.	attention	binary map	max jaw size	min jaw size	$N=3$	$N=6$	0.25	0.30
U-Net [19]	✓	✓	✓	✓			✓	✓		**94.71**	**92.65**
	✓	✓	✓	✓			✓		✓	91.51	89.07
	✓		✓	✓			✓	✓		92.34	90.44
	✓	✓	✓				✓	✓		93.36	90.90
	✓	✓	✓	✓		✓		✓		94.11	91.83
	✓	✓	✓	✓	✓		✓	✓		91.75	90.27
	✓				✓					89.85	88.13
GGCNN2 [14]	✓	✓	✓	✓			✓	✓		**88.92**	**85.94**
	✓		✓	✓			✓	✓		87.88	85.52
	✓				✓					85.23	82.67

Table I: Ablation study over different design choices for both *ORANGE* implementations with U-Net and GGCNN2.

objective function takes the form:

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

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, $Q * \Gamma * O$. 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.

IV. Experiments & Discussion

We validate *ORANGE* on Jacquard, following the standard 90/10% split without any data augmentation. Depth images are resized into 320×320, to speed training up. Following prior literature, we report IoU@0.25 and 0.30.

Ablation study: We inspect the different combinations of the individual components of *ORANGE* in Table I, employing U-Net as the base model, as it has more capacity to absorb the multiple grasp representations. Our 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 lower 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, making it more difficult to disentangle the multiple grasping orientations.

Moreover, the application of the pixel-wise graspness Γ on the quality maps Q has an evident benefit on the model, since it focuses on the most prominent grasp points and restricts the exploration of the feature space, thus decreasing the grasp box area and more precisely localizing the grasp center (Fig. 4).

The selection of the jaw size during the construction of the ground truth maps also affects the performance of *ORANGE* over both thresholds, confirming that picking the minimum leads to more accurate predictions, as it produces bounding boxes closer to the object's boundaries.

methods	modality	IoU@0.25 (%)
Morrison et al. [14]	D	85.2
Depierre et al. [6]	RGB-D	74.2
Zhou et al. [8]	RGB	91.8
Zhou et al. [8]	RGD	92.8
Zhang et al. [9]	RGB	90.4
Zhang et al. [9]	RGD	93.6
ORANGE with GGCNN2 (ours)	D	**88.9**
ORANGE with U-Net (ours)	D	**94.7**

Table II: Comparative results on the Jacquard dataset.

An important decision choice is whether to use binary values for the quality maps in the ground truth data synthesis [14]. Using binary maps in *ORANGE* produces 4% less accurate grasp predictions w.r.t. to our approach. However, *ORANGE* achieves an IoU@0.25 of 91.75% (although 3% lower than using our approach), while a U-Net implemented as suggested in [14], succeeds a 89.85% at the 0.25 IoU threshold.

Lastly, we also improve GGCNN2 from 85.23% in the original implementation into 88.92% using *ORANGE*, confirming *ORANGE*'s model-agnostic character.

Comparing to previous works: Pure depth-based *ORANGE* outperforms all existing approaches on Jacquard (Table II) to achieve a new state-of-the-art of 94.7% IoU@0.25, improving by an absolute 1.1% the main competitor [9], that uses multi-modal RGD data. Both the augmented grasp map representation and the bin-wise attention of the orientation estimation over the quality maps, are major factors of performance. We expect even better performance if we also use the RGB channel, however this is beyond the scope of our work that focuses on improving the grasp map representation.

V. Conclusions & Future work

We discuss and address the problem of multiple orientations per grasping point on the Jacquard dataset. Our method, *ORANGE*, jointly solves an angle-bin classification and real-value angle regression, while exploiting the former to guide a graspness attention mechanism over the grasp quality map. Extensive experimental results justify the effectiveness of *ORANGE* components, that achieves state-of-the-art performance using only the depth modality. An interesting future direction is to jointly reason about the objects' grasping points, shape and category. The quality of the generated grasps can also be ranked in an adversarial setting, while interacting with real objects, to learn to identify task-related grasp points.

REFERENCES

[1] J. Xu, A. Bhardwaj, G. Sun, T. Aykut, N. Alt, M. Karimi, and E. Steinbach, "Learning-based modular task-oriented grasp stability assessment," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 3468–3475.

[2] Kuan Fang, Yuke Zhu, Animesh Garg, Andrey Kurenkov, Viraj Mehta, Li Fei-Fei, and Silvio Savarese, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 202–216, 2020.

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

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

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

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

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

[17] 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.

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

[19] 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*.