Quality-focused Active Adversarial Policy for Safe Grasping in Human-Robot Interaction

Chenghao Li, *Graduate Student Member, IEEE*, Razvan Beuran, *Senior Member, IEEE*, and Nak Young Chong, *Senior Member, IEEE*

***Abstract*—Vision-guided robot grasping methods based on Deep Neural Networks (DNNs) have achieved remarkable success in handling unknown objects, attributable to their powerful generalizability. However, these methods with this generalizability tend to recognize the human hand and its adjacent objects as graspable targets, compromising safety during Human-Robot Interaction (HRI). In this work, we propose the Quality-focused Active Adversarial Policy (QFAAP) to solve this problem. Specifically, the first part is the Adversarial Quality Patch (AQP), wherein we design the adversarial quality patch loss and leverage the grasp dataset to optimize a patch with high quality scores. Next, we construct the Projected Quality Gradient Descent (PQGD) and integrate it with the AQP, which contains only the hand region within each real-time frame, endowing the AQP with fast adaptability to the human hand shape. Through AQP and PQGD, the hand can be actively adversarial with the surrounding objects, lowering their quality scores. Therefore, further setting the quality score of the hand to zero will reduce the grasping priority of both the hand and its adjacent objects, enabling the robot to grasp other objects away from the hand without emergency stops. We conduct extensive experiments on the benchmark datasets and a cobot, showing the effectiveness of QFAAP. Our code and demo videos are available in the supplementary items.**

***Note to Practitioners*—This work is inspired by adversarial attacks but from a completely different perspective: exploring the benign aspects of adversarial attacks to address the safety problem of DNNs-based grasping in cluttered HRI scenarios. Specifically, we aim to enable the robot to grasp objects away from the human hand and its adjacent objects without triggering emergency stops. This is realized by designing benign, quality score-based adversarial examples with shape adaptability to alter the grasping sequence, thereby avoiding collision risks between the robot and human hand during grasping. Our approach presents innovative solutions for future research on benign adversarial attacks in real-world robot grasping and offers practical insights for the engineering implementation of safe robot grasping systems based on DNNs.**

***Index Terms*—Robot grasping, human-robot interaction, grasp quality score, deep learning, adversarial attack.**

I. INTRODUCTION

VISION-guided robot grasping is one of the critical capabilities for HRI [1], aimed at helping humans improve work efficiency in the service and manufacturing domain. Traditional visual grasping methods typically construct a grasp database based on three-dimensional (3D) object models, incorporating performance metrics derived from geometric and physical properties [1], [2] and employing stochastic sampling to account for grasping uncertainty [3]. However, these methods are inherently limited by their reliance on known 3D object models, rendering them ineffective when applied to novel objects. To address this limitation, recent studies [4], [5] have introduced an alternative paradigm that leverages DNNs [6]–[10] to train function approximators. These approximators predict grasp candidates directly from images, utilizing datasets comprising empirical grasp successes and failures, thereby enabling efficient generalization to previously unseen objects at substantially lower cost. However, these methods with this generalizability may also recognize the human hand and its adjacent objects as graspable targets, compromising safety during HRI. Particularly in multi-object or cluttered scenes, collisions between the robot and the human can occur in two situations. The first is when the robot directly recognizes the human hand and attempts to grasp it, resulting in a collision. The second is when the robot identifies an object adjacent to the human hand, and while grasping this object, the gripper opens to a specific width and consequently collides with the hand (as shown in Fig. 1). Therefore, given the growing trend of large-scale deployment of DNNs-based visual grasping systems in HRI scenarios, it is important to address the safety concerns in both situations to avoid workplace injuries and accidents.

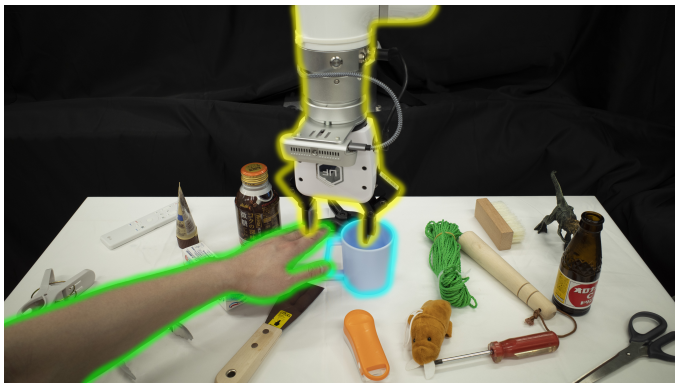

Fig. 1. An example of a cluttered HRI scenario: the robot mistakenly identifies the human hand or adjacent objects as graspable targets for autonomous grasping, causing harm to the human. We highlight the robot, the human hand, and the target object using yellow, green, and blue borders, respectively.

Some methods assist robots in avoiding collisions with human hands and enabling interaction by segmenting human

This work was supported by JSPS KAKENHI Grant Number JP23K03756 and the Asian Office of Aerospace Research and Development under Grant/Cooperative Agreement Award No. FA2386-22-1-4042.

The authors are with the School of Information Science, Japan Advanced Institute of Science and Technology, 1-1, Asahidai, Nomi, 923-1292, Ishikawa, Japan (e-mail: chenghao.li@jaist.ac.jp; nakyoung@jaist.ac.jp).

hands or estimating their pose or motion, as exemplified in Robot-to-Human Handover (R2H) [11] and Human-to-Robot Handover (H2R) [12]–[15]. Although these methods are effective in helping robots avoid human hands during handover, most are limited to the handover problem between humans and robots in simple single-object scenarios. However, in real-world HRI contexts, cluttered scenes are more general, and human hands typically appear within the grasping view of the robot rather than only during handover. For instance, in collaborative sorting, services, and household assistance, ensuring that robots can execute grasping operations while simultaneously avoiding both the human hand and nearby objects within the grasping view is critical. Specifically, consider a scenario in which a robot and a human jointly clean a cluttered table: the robot executes grasping operations, while the human receives the grasped objects and transfers them to a storage bin located outside the robot's workspace. From a robot-centric viewpoint, when the robot prepares for the next grasp and detects a human hand appearing in its camera view to receive an object, failure to avoid the hand or nearby objects may potentially lead to human injury. Therefore, different from the problem that the above handover works focused on, this paper will emphasize the grasping safety problem of how to enable robots to autonomously avoid the human hand and objects close to the hand during grasping without emergency stops in cluttered HRI scenarios, which is a new and more challenging problem in DNNs-based visual grasping.

How to address this problem? A straightforward engineering approach is first detecting the human hand mask, then applying dilation to expand the mask, and setting the grasp quality scores within the expanded mask to zero, thereby enabling the robot to avoid both the human hand and adjacent objects during grasping. However, our experiments reveal that this method substantially reduces the workspace of the robot because a large dilation radius is necessary for it to be effective, which means that the invalid workspace will include areas that will not result in colliding with the hand. An alternative approach is to use a decay function after the dilation process to gradually reduce the grasp quality score based on the distance between the original mask and the expanded mask, thereby preserving most of the workspace of the robot. Nevertheless, our experimental results show that this method requires manually set heuristic parameters, which are rigid and less adaptable to variations in hand pose. Therefore, addressing the problem of avoiding grasping human hands and nearby objects in cluttered HRI scenarios through the adaptive optimization policy should be more appropriate.

Inspired by adversarial attacks [16]–[18], which leverage the interpretability flaws of DNNs to optimize perturbations that interfere with model predictions, we investigate from a novel perspective: whether adversarial attacks can be used as benign adversarial perturbations to interfere with the grasp quality score, thereby dynamically adjusting the grasping sequence of the robot to actively avoid the human hand and objects adjacent to it. Therefore, based on this new perspective, the method we aim to design differs significantly from common adversarial attacks. Firstly, most adversarial attack methods focus on how to attack the model. In contrast, our goal is not to attack or defend [19] but to address the safety issue in DNNs-based visual grasping within HRI scenarios through controllable perturbations. Secondly, our method emphasizes actively perturbing the grasp quality score to alter the grasping priority of human hands and their neighboring objects, thereby guiding the robot to avoid grasping them. In contrast, common adversarial attacks primarily aim to degrade detection accuracy [20], [21], cause misclassification [22], [23] or mislocalization [21], and evade detection [24]–[26]. Finally, since human hands can appear with arbitrary postures to perform tasks in various HRI scenarios, the perturbation we want to design must conform closely to the shape of the hand at a fast speed, keeping the hand away from the robot gripper. This is much more difficult than other adversarial attacks [16], [20], [21] that apply perturbations with fixed shapes or extend to other specific shapes through complicated processes and high costs [25]–[27].

Along these lines, this paper proposes the Quality-focused Active Adversarial Policy (QFAAP), which first optimizes an Adversarial Quality Patch (AQP) with high quality scores by the adversarial quality patch loss and the grasp dataset. Next, integrate AQP that contains only the hand region within each real-time frame with the Projected Quality Gradient Descent (PQGD), ensuring AQP has fast adaptability to the human hand shape. By applying AQP and PQGD, the hand can actively interfere with nearby objects, reducing their quality score. Further, setting the quality score of the hand to zero will simultaneously lower the grasping priority of both the hand and surrounding objects, enabling the robot to actively avoid them while grasping without emergency stops.

A summary of the contributions in this work is as follows:

1) We reveal a new and more challenging problem in DNNs-based visual grasping: how to enable robots to simultaneously and adaptively avoid human hands and nearby objects without emergency stops during grasping in clutter. Addressing this problem is critical for achieving safe grasping in broad HRI scenarios.
2) We propose the QFAAP, the first comprehensive safe grasping policy based on benign adversarial perturbations. QFAAP enables fast adaptive and controllable perturbations that alter grasping priorities, ensuring that the human hand and its neighboring objects are deprioritized while preserving the model's original grasping ability. This policy highlights how adversarial attacks can be transformed into safety-enhancing mechanisms, offering both theoretical insights and practical guidelines for the development of safe robot grasping systems.
3) We release our code publicly available to facilitate reproducibility and to foster further research on adversarially enhanced safe grasping in cluttered HRI scenarios.

This paper is organized into the following sections. Section II (Related Work) reviews vision-guided robot grasping and adversarial attacks. Section III (Proposed Method) provides an overview of QFAAP, detailing its two components (AQP and PQGD), and discusses how QFAAP is implemented in robot grasping. Section IV (Experiments) validates the effectiveness of our method in benchmark datasets and real-world grasping

scenarios. Finally, Section V (Conclusion) summarizes the work of this paper and provides prospects for future research.

II. Related Work

A. Vision-guided Robot Grasping

While many grasping frameworks exist, this work focuses explicitly on vision-guided 4-Degree-of-Freedom (4-DOF) grasping using a parallel-jaw gripper, which can be broadly categorized into traditional methods and DNNs-based methods. Traditional grasping methods are founded on mathematical and physical models that characterize object geometry, kinematics, and dynamics [1]–[3]. These methods typically assume the availability of a detailed 3D model of the object, which is leveraged to compute stable grasp configurations. For instance, Gallegos *et al.* [28] optimized grasping strategies by utilizing predefined contact points on known 3D object models. Similarly, Pokorny *et al.* [29] introduced the concept of grasping spaces, enabling the mapping of objects to these spaces for grasp synthesis. While these approaches exhibit robustness in structured environments, their applicability is inherently constrained by the prerequisite of complete 3D object models, and they are often unavailable in unstructured environments containing novel objects. This limitation underscores the need for more flexible grasping strategies to handle object uncertainty in unstructured environments.

DNNs-based visual grasping methods demonstrate strong generalization capabilities to novel objects by employing function approximators trained on extensive datasets to predict the grasp success probability from images. Consequently, datasets play a pivotal role in these methods. A notable human-labeled dataset is the Cornell Grasping Dataset [30], which comprises approximately 1,000 RGB-D images and has been widely utilized for training grasping models in single-object scenarios [31]–[37]. The Dex-Net series [4], [38]–[41] introduced a large-scale synthetic dataset that integrates various cluttered environments to acquire cluttered grasping capabilities, significantly advancing the field of visual grasping. Similarly, GraspNet [5], [42], [43] constructed a real-world dataset encompassing one billion grasp labels and nearly 100,000 images across 190 densely cluttered scenes and support both 4-DOF and 6-Degree-of-Freedom (6-DOF) grasping, which further improves the grasping ability for unknown objects in cluttered scenarios.

Although the aforementioned DNNs-based methods demonstrate strong generalization capabilities for unknown objects in unstructured environments, they emphasize grasp generalization while neglecting grasp safety. Specifically, these methods with this generalizability will also recognize human hands and adjacent objects as graspable targets, compromising safety during HRI.

B. Adversarial Attacks

Since Szegedy *et al.* [44] first identified adversarial examples, extensive research has been conducted to expose the vulnerability of DNNs. These efforts generally fall into two categories: single-image adversarial attacks and image-agnostic attacks (adversarial patch attacks). Single-image adversarial attacks achieve their attacks by maximizing the discriminative loss of the model to generate global perturbations that cover the entire image. Goodfellow *et al.* [16] designed a Fast Gradient Sign Method (FGSM) to produce strong perturbations based on investigating the model's linear nature. Wang *et al.* [45] and Madry *et al.* [22] further broke the one-step generation of perturbation in FGSM into iterative generation and proposed I-FGSM and Projected Gradient Descent (PGD) attack. Although the single-image adversarial attacks can rapidly attack image classification models, causing them to produce misclassification results, they were limited to one specific image and entire image regions, which means each new image requires re-optimization. Thus, this limitation highlights the need for more flexible methods to attack arbitrary images and any local regions within an image.

Adversarial patch attacks, characterized by their locality and image-agnostic nature, effectively compromise object detection models with localization properties. For instance, Liu *et al.* [46] designed DPatch to attack widely used object detectors, degrading their detection accuracy and thereby causing mislocalization or misclassification. Later, Lee *et al.* [47] investigated failure cases of DPatch and subsequently introduced the Robust DPatch. Beyond causing mislocalization or misclassification, some studies focused on evading detection, preventing detectors from recognizing objects occluded by adversarial patches, as explored in [21], [24]. Later works, such as [25]–[27], extended adversarial patches by replicating them into adversarial clothing, enabling more flexible evasion across different viewing angles. However, this replication-based extension is costly and typically limited to the fold variations of clothes.

Overall, the aforementioned single-image adversarial and adversarial patch attacks have demonstrated effectiveness, but how to transform these attacks into controllable benign adversarial to address safety concerns in DNNs-based grasping remains unexplored. Moreover, another important yet underexplored direction is how to actively manipulate the grasp quality score in DNNs-based grasping to alter the grasping priority of the robot. Finally, rapidly achieving shape adaptability for adversarial perturbations at minimal cost is critical and practical in robot grasping, which often needs to deal with objects with different shapes. So, in this work, we leverage the advantages of single-image adversarial and adversarial patch attacks, and propose a novel active adversarial method with rapid human hand shape adaptability by manipulating the grasp quality score, which aims to address the safety problem of DNNs-based grasping in the HRI process.

III. Proposed Method

In this section, we will first make an overview of QFAAP. Then, a comprehensive description of two important modules (AQP and PQGD) will be provided. Finally, we will explain how to deploy QFAAP to improve visual grasping safety in cluttered HRI scenarios.

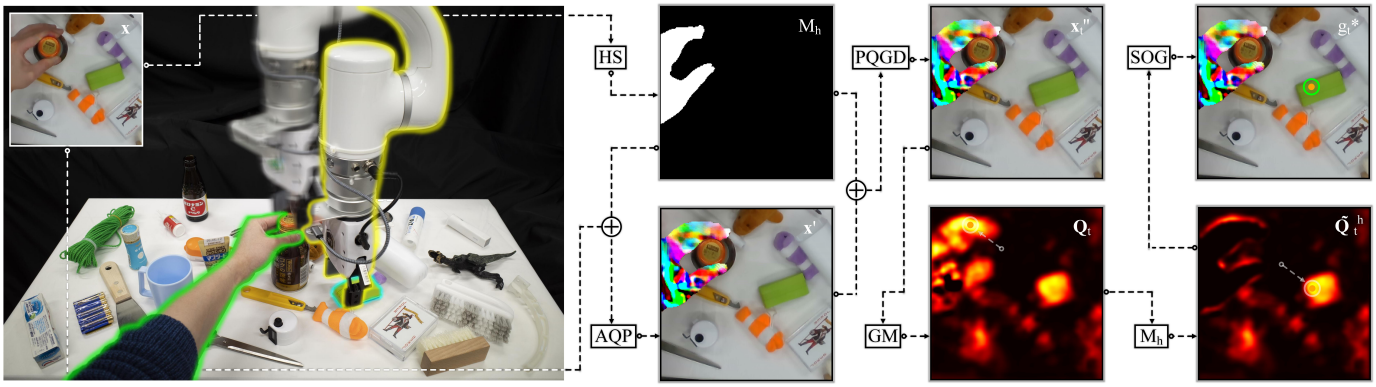

Fig. 2. Pipeline of QFAAP: Firstly, the original RGB frame $\mathbf{x}$ is captured by the depth camera, and a hand segmentation algorithm (HS) is applied to obtain the hand mask $\mathcal{M}_h$, as shown in the subfigure on the far left (first column) and the top row of the second column. Next, the optimized AQP is incorporated into $\mathbf{x}$ while preserving only the hand region, generating $\mathbf{x}'$, as shown in the bottom row of the second column. In the third stage, PQGD is applied to $\mathbf{x}'$ with $\mathcal{M}_h$ to rapidly endorse the shape adaptability of AQP, producing $\mathbf{x}_t''$, as shown in the top row of the third column. In the fourth stage, $\mathbf{x}_t''$ is fed into the grasping model (GM) to obtain the quality map $\mathbf{Q}_t$, followed by getting the quality map $\tilde{\mathbf{Q}}_t^h$ outside the hand region by $\mathcal{M}_h$, as shown in the bottom rows of the third and fourth column. Finally, selecting the optimal grasp (SOG) g_t^* (emphasized by the green circle and orange dot) with the maximum quality score (emphasized by the orange dot, translucent white circle, and translucent white dotted arrow) within $\tilde{\mathbf{Q}}_t^h$, as shown in the top row of the fourth column. The above process can effectively shift the initial hazardous grasp (the robot is emphasized as a blurred version) located near the hand (emphasized by the green border) toward a safer grasp (the object being grasped and the robot are emphasized with the blue and yellow borders), as shown in the first column.

A. Overview of QFAAP

We propose the Quality-focused Active Adversarial Policy (QFAAP) to enhance the safety of DNNs-based visual grasping in cluttered HRI scenarios. QFAAP consists of two key modules: the Adversarial Quality Patch (AQP) and Projected Quality Gradient Descent (PQGD). The AQP is optimized by the adversarial quality patch loss and grasp dataset, ensuring adversarial effectiveness against the quality score of any image. The PQGD can be integrated with AQP, which contains only the hand region within each real-time frame, endowing AQP with fast human hand shape adaptability. By applying AQP and PQGD, the hand can actively perturb nearby objects to reduce their quality score in the model prediction process. Further, setting the quality score of the hand to zero will simultaneously lower the grasping priority of both the hand and surrounding objects, enabling the robot to actively avoid them while grasping without emergency stops in cluttered HRI scenarios. The pipeline of the QFAAP framework is illustrated in Fig. 2.

NOMENCLATURE

α	Empirical parameter for $\mathcal{L}_q^p$.
β	First empirical parameter for $\mathcal{L}_{aqp}$.
δ_{aqp}	Learning rate for AQP.
δ_{model}	Learning rate for grasping model.
δ_{pqgd}	Learning rate for PQGD.
ϵ	Projection restriction parameter of PQGD.
γ	Second empirical parameter for $\mathcal{L}_{aqp}$.
$\hat{q}_i(n)$	Quality score at n of a label related x_i.
$\mathbf{p}_t$	AQP.
$\mathbf{Q}_t$	Quality map of $\mathbf{x}_t''$.
$\mathbf{Q}_t^h$	Quality map inside the hand area of $\mathbf{x}_t''$.
$\mathbf{w}_t$	Weights of grasping model.
$\mathbf{x}$	RGB image of the real-time frame.
$\mathbf{x}'$	RGB image of the real-time frame after adding AQP within the hand area.
$\mathbf{x}_t''$	RGB image of the real-time frame after adding AQP and PQGD within the hand area.
$\tilde{\mathcal{Q}}_i^p$	Quality map outside the AQP area of x_i.
$\mathcal{L}_\theta$	Loss of angle for grasping model.
$\mathcal{L}_d$	Loss of difference for AQP.
$\mathcal{L}_q$	Loss of quality for grasping model.
$\mathcal{L}_q(n)$	Loss of quality at n for grasping model.
$\mathcal{L}_q^p$	Loss of quality for AQP.
$\mathcal{L}_w$	Loss of width for grasping model.
$\mathcal{L}_{aqp}$	Loss of total for AQP.
$\mathcal{L}_{model}$	Loss of total for grasping model.
$\mathcal{L}_{pqgd}$	Loss of total for PQGD.
$\mathcal{L}_{tv}$	Loss of total variation for AQP.
$\mathcal{M}_h$	Mask of hand.
$\mathcal{Q}_i$	Quality map of x_i.
$\mathcal{Q}_i^p$	Quality map inside the AQP area of x_i.
$\mathcal{R}_q$	Quality rate evaluation for AQP.
$\tilde{\mathbf{Q}}_t^h$	Quality map outside the hand area of $\mathbf{x}_t''$.
i_t, j_t	Quality score location of $\tilde{\mathbf{Q}}_t^h$, $i_t \neq j_t$.
j^p, k^p	Pixel location of AQP, $j^p \neq k^p$.
j_i^p, k_i^p	Pixel location of scaled AQP in x_i, $j_i^p \neq k_i^p$.
$q_i(n)$	Quality score at n of x_i.
x_i	Sample RGB image within a batch.
p_i	Scaled AQP related to x_i.

B. Adversarial Quality Patch (AQP)

The DNNs-based visual grasping model typically first defines the grasp configuration [48], which is composed of parameters $(j^g, k^g, w^g, h^g, \theta^g)$ forming a rotated box in the image coordinate system, and this box is denoted by the grasp candidate g_i. Here, (j^g, k^g) represents the center position of

the box, w^g and h^g denote the width and height of the box, and θ^g represents the angle of the box relative to the horizontal direction. Accordingly, in the robot coordinate system, the grasp and its corresponding parameters are defined as $\mathcal{G}_i$ and $(I^g, J^g, Z^g, W^g, \Theta^g)$ (the coordinate transformation from g_i to $\mathcal{G}_i$ is explained in Section. III-D). Then, based on the grasp configuration in the image coordinate system, corresponding objective loss functions are designed, such as the quality loss $\mathcal{L}_q$ associated with (j^g, k^g), the width loss $\mathcal{L}_w$ associated with w^g, and the angle loss $\mathcal{L}_\theta$ associated with θ^g. Assuming that for an image sample x_i within one batch (batch size is B), the predicted and labeled quality scores at position n of x_i are denoted as $q_i(n)$ and $\hat{q}_i(n)$. The quality loss at n of x_i for the model can be defined as Eq. 1.

$$\mathcal{L}_q(n) = \begin{cases} 0.5[q_i(n) - \hat{q}_i(n)]^2, & \text{if } |q_i(n) - \hat{q}_i(n)| < 1 \\ |q_i(n) - \hat{q}_i(n)| - 0.5, & \text{otherwise} \end{cases} \tag{1}$$

By computing the average $\mathcal{L}_q(n)$ across all positions N, the complete quality loss for the model can be given by Eq. 2.

$$\mathcal{L}_q = \frac{1}{N} \sum_{n=1}^{N} L_q(n) \tag{2}$$

The losses $\mathcal{L}_w$ and $\mathcal{L}_\theta$ follow the same computation as $\mathcal{L}_q$, consistent with the formulations in Eq. 1 and Eq. 2. By summing these losses, the total loss for the model can be shown as Eq. 3.

$$\mathcal{L}_{model} = \mathcal{L}_q + \mathcal{L}_\theta + \mathcal{L}_w \tag{3}$$

Finally, $\mathcal{L}_{model}$ can be used for model training, where the model weights are optimized via gradient descent. The weight update process is expressed as Eq. 4. Here, $\mathbf{w}_t$ and $\mathbf{w}_{t-1}$ represent the model weights at time steps t and $t-1$, respectively, while the derivative of $\mathcal{L}_{model}$ with respect to $\mathbf{w}_{t-1}$ denotes the gradient and δ_{model} is the learning rate of the model. Notably, during training, the quality score within the central one-third region of the grasp label is set to 1 (Maximum), while all other positions are set to 0 (Minimum). This design encourages the model to focus more on learning features in these key regions, thereby increasing the predicted quality score when encountering similar features during inference. Therefore, the quality score is of utmost importance, as it not only determines the grasping position parameters and other parameters corresponding to it, but also dictates the grasping priority, with a higher quality score indicating a higher priority in the grasping sequence.

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \delta_{model} \frac{\partial \mathcal{L}_{model}}{\partial \mathbf{w}_{t-1}} \tag{4}$$

The AQP is also optimized from the perspective of the quality score. However, unlike optimizing the grasping model, we aim for AQP to optimize in the direction of increasing the quality score rather than minimizing the difference between the predicted quality score and the labeled quality score. Therefore, we first initialize AQP following a uniform distribution, with the same shape as the input image of the model. In optimization, the AQP will be randomly scaled to be applied to the image sample.

Next, we define the quality loss of AQP ($\mathcal{L}_q^p$). Let the quality map predicted by the frozen grasping model within the AQP area of x_i be represented as $\mathcal{Q}_i^p$. The quality loss $\mathcal{L}_q^p$ is then defined as in Eq. 5, where $\mathbb{E}(\mathcal{Q}_i^p)$ and $\text{Var}(\mathcal{Q}_i^p)$ denote the mean and variance of $\mathcal{Q}_i^p$, respectively. The α is an empirical parameter that controls the influence of variance on $\mathcal{L}_q^p$. This loss can be minimized using a gradient descent algorithm by continuously decreasing the negative value (increasing in the negative direction) of $\mathbb{E}(\mathcal{Q}_i^p)$, thereby enhancing the quality score of AQP. So, this can be regarded as the reverse operation of a gradient descent algorithm, achieving gradient ascent to optimize AQP. Additionally, reducing $\text{Var}(\mathcal{Q}_i^p)$ ensures a more stable increase in the quality score.

$$\mathcal{L}_q^p = \frac{1}{B} \sum_{i=1}^{B} [-\mathbb{E}(\mathcal{Q}_i^p) + \alpha \text{Var}(\mathcal{Q}_i^p)] \tag{5}$$

In this step, we employ the same total variation loss $\mathcal{L}_{tv}$ from [24] to mitigate noise introduced during AQP optimization, ensuring a smoother optimization, as shown in Eq. 6. Here, $\mathbf{p}_t(j^p, k^p)$ represents the pixel value of AQP ($\mathbf{p}_t$) at location (j^p, k^p), W and H are the width and height of $\mathbf{p}_t$. This loss is computed as the mean of the Euclidean distance between all adjacent pixel values within AQP.

$$\mathcal{L}_{tv} = \frac{1}{H \times W} \sum_{j^p=1}^{H} \sum_{k^p=1}^{W} \|\mathbf{p}_t(j^p, k^p)\|_2 \tag{6}$$

To further reinforce the optimization of the quality score for AQP, we introduce the difference loss $\mathcal{L}_d$. Let the quality map predicted by the frozen grasping model outside the AQP area of x_i be denoted as $\tilde{\mathcal{Q}}_i^p$. The $\mathcal{L}_d$ is defined as in Eq. 7. This loss can strengthen AQP by letting $\min \mathcal{Q}_i^p$ approach $\max \tilde{\mathcal{Q}}_i^p$. Consequently, AQP will be optimized so that the model predicts a higher quality score for AQP than for other objects in the scene. Thereby, the AQP can effectively interfere with the quality scores of other objects.

$$\mathcal{L}_d = \frac{1}{B} \sum_{i=1}^{B} \left| \min \mathcal{Q}_i^p - \max \tilde{\mathcal{Q}}_i^p \right| \tag{7}$$

Finally, we combine the three aforementioned losses with two additional empirically determined parameters, β and γ, controlling $\mathcal{L}_{tv}$ and $\mathcal{L}_d$, respectively, to obtain the total loss of AQP ($\mathcal{L}_{aqp}$), as defined in Eq. 8. Similarly, we optimize AQP by minimizing this loss using the gradient descent algorithm with Adam optimizer [49], as shown in Eq. 9. Here, $\mathbf{p}_t$ and $\mathbf{p}_{t-1}$ represent AQP at time steps t and $t-1$, respectively, while the derivative of $\mathcal{L}_{aqp}$ with respect to $\mathbf{p}_{t-1}$ denotes the gradient, and δ_{aqp} is the learning rate of AQP. Since the optimization process is based on the entire grasp dataset, the optimized AQP can be effective on any image.

$$\mathcal{L}_{aqp} = \mathcal{L}_q^p + \beta \mathcal{L}_{tv} + \gamma \mathcal{L}_d \tag{8}$$

$$\mathbf{p}_t = \mathbf{p}_{t-1} - \delta_{aqp} \frac{\partial \mathcal{L}_{aqp}}{\partial \mathbf{p}_{t-1}} \tag{9}$$

Following the optimized AQP ($\mathbf{p}_t$), we define an evaluation method to assess the quality score level of AQP in one testing batch. Let j_i^p, k_i^p denote the pixel position of the scaled AQP in x_i, and let W_i^p and H_i^p as the width and height of the scaled AQP. We compute the ratio $\mathcal{R}_q$ as the proportion of pixels within all AQP regions across a batch where the quality score $\mathcal{Q}_i^p(j_i^p, k_i^p)$ exceeds 0.5, relative to the total number of pixels (N^p) in all sample image, as shown in Eq. 10. Here, $\mathbb{1}$ means the indicator function. After defining $\mathcal{R}_q$, we compute the average $\mathcal{R}_q$ for each batch to evaluate the quality score level of AQP across the entire test set, which is denoted by Quality Accuracy (Q-ACC) and will be used in the Experiments section.

$$\mathcal{R}_q = \frac{1}{N^p} \sum_{i=1}^{B} \left\{ \sum_{j_i^p=1}^{H_i^p} \sum_{k_i^p=1}^{W_i^p} \mathbb{1}[\mathcal{Q}_i^p(j_i^p, k_i^p) > 0.5] \right\} \tag{10}$$

C. Projected Quality Gradient Descent (PQGD)

The PGD [22] is typically used to attack classification models by inducing misclassification, with the attack targeting the entire region of a single image. In contrast, PQGD primarily focuses on specific local regions within a single image and emphasizes quality score optimization like AQP. Since PQGD, like PGD, exhibits fast optimization properties, it can be employed to further enhance the quality score of local regions in AQP, thereby rapidly endowing AQP with shape adaptability.

Let $\mathbf{x}$ denote a real-time RGB frame from a depth camera, and let $\mathcal{M}_h$ represent the mask of the hand associated with $\mathbf{x}$, obtained using the upper limb segmentation algorithm [50]. We first define $\mathbf{x}'$ as the RGB frame after adding AQP (the same size as $\mathbf{x}$) within the hand area, as shown in Eq. 11.

$$\mathbf{x}' = \mathbf{x}(1 - \mathcal{M}_h) + \mathbf{p}_t \mathcal{M}_h \tag{11}$$

Then, let the RGB frame after adding both AQP and PQGD within the hand area be denoted as $\mathbf{x}_t''$. We define the loss of PQGD as $\mathcal{L}_{pqgd}$, as shown in Eq. 12, where $\mathbf{Q}_t^h$ represents the quality map inside the hand area of $\mathbf{x}_t''$.

$$\mathcal{L}_{pqgd} = -\mathbb{E}(\mathbf{Q}_{t-1}^h) \tag{12}$$

Finally, we leverage $\mathcal{L}_{pqgd}$ and the hand mask $\mathcal{M}_h$ to rapidly optimize the AQP within the hand region of $\mathbf{x}_t''$, as shown in Eq. 13. Here, sgn represents the sign function, which is used to compute the direction of the derivative of $\mathcal{L}_{pqgd}$ with respect to $\mathbf{x}_{t-1}''$, thereby accelerating optimization. The parameter δ_{pqgd} represents the learning rate of PQGD. The parameter ϵ, similar to ϵ in PGD [22], denotes the projection restriction parameter of PQGD, which constrains $\mathbf{x}_t''$ from deviating excessively from $\mathbf{x}'$ during optimization. This ensures that the additional PQGD perturbation only slightly alters the pixel values of AQP (such that the modification remains nearly imperceptible to the human eye), thereby preserving the effectiveness of the original AQP. It is important to emphasize that the optimization process is guided by $\mathcal{M}_h$ to operate solely within the hand region, endowing AQP with the adaptability to the human hand shape, which constitutes the most critical aspect of PQGD optimization.

$$\mathbf{x}_t'' = \left\{ \prod_{\mathbf{x}',\epsilon} [\mathbf{x}_{t-1}'' - \mathrm{sgn}(\delta_{pqgd} \frac{\partial \mathcal{L}_{pqgd}}{\partial \mathbf{x}_{t-1}''})] \right\} \mathcal{M}_h + \mathbf{x}'(1 - \mathcal{M}_h) \tag{13}$$

D. Active Adversarial for Robot Grasping

This part explains how QFAAP is applied to robot grasping to manipulate the quality score, enabling the robot to avoid grasping human hands and nearby objects. In this work [51], Li *et al.* observed an intriguing property and empirically confirmed it through extensive real experiments that moving a specific object in a cluttered scenario can dynamically alter the quality score of this scenario. Specifically, if this object has a higher quality score, it can perturb objects with lower quality scores when the distance between them is very close (approximately 0.5–1 cm), leading to a further reduction in their quality scores. Moreover, as this object with the high quality score approaches, the quality scores of the affected objects will gradually decrease, and when they come into contact, the quality scores of these objects may drop sharply to zero. Notably, this phenomenon only occurs between adjacent objects; if the objects are far apart, no interference will happen, and their quality scores will remain unchanged. Thus, we are motivated to explore whether this property can be leveraged to enhance grasping safety in cluttered HRI scenarios.

QFAAP follows the property observed by [51], processing the features within the human hand to increase its quality score using AQP and PQGD. Consequently, the human hand can be directly regarded as a benign adversarial perturbation that is actively against adjacent objects in any posture, thereby suppressing their quality scores. After the interference, the quality score within the human hand will be set to zero, reducing the grasping priority of both the hand and its adjacent objects. In other words, the manipulation of the quality score by QFAAP is entirely controllable and does not affect the original performance of the grasping model.

First, we use $\mathcal{M}_h$ to process $\mathbf{Q}_t$ from Section III-C, setting the quality score within the hand region to zero. This results in a quality map outside the hand area of $\mathbf{x}_t''$, denoted as $\tilde{\mathbf{Q}}_t^h$. The robot then uses the perturbed $\tilde{\mathbf{Q}}_t^h$ as a reference and selects the object (away from the human hand and its adjacent objects) corresponding to the highest quality score in $\tilde{\mathbf{Q}}_t^h$ as the optimal grasping target. This process is defined in Eq. 14. Here, (i_t^*, j_t^*) corresponds to the previously defined grasp candidate position parameters (j^g, k^g), with the distinction that (i_t^*, j_t^*) represents the optimal grasping position after QFAAP perturbation (where t is to emphasize the influence of QFAAP). Furthermore, based on (i_t^*, j_t^*), other optimal grasping parameters w_t^*, h_t^*, and θ_t^* can be determined, forming the optimal grasp g_t^*.

$$(i_t^*, j_t^*) = \arg\max_{(i_t, j_t) \in (H,W)} \tilde{\mathbf{Q}}_t^h(i_t, j_t) \tag{14}$$

Next, g_t^* needs to undergo the following transformations to complete the grasping. Since h_t^* is used only for visual

representation and not in the conversion process, we denote the transferred optimal grasp in the robot end effector coordinate systems as $\mathcal{G}_t^*(I_t^*, J_t^*, Z_t^*, W_t^*, \Theta_t^*)$, which corresponds to the previously defined $\mathcal{G}_i(I^g, J^g, Z^g, W^g, \Theta^g)$, t and $*$ are intended to emphasize the impact of QFAAP and optimal grasp. Here, (I_t^*, J_t^*, Z_t^*) represents the grasp position in the robot end effector coordinate system, W_t^* is the opening stroke of the parallel jaw gripper, and Θ_t^* is the rotation angle of the gripper relative to the Z axis. The conversion process is divided into three parts. The first part involves converting (i_t^*, j_t^*): using depth information (d) and the camera's intrinsic parameters (f_x, f_y for focal lengths and c_x, c_y for the image center coordinates), we convert (i_t^*, j_t^*) from the image coordinate system to the camera coordinate system $(i_{ct}^*, j_{ct}^*, z_{ct}^*)$, as shown in Eq. 15.

$$\begin{bmatrix} i_{ct}^* \\ j_{ct}^* \\ z_{ct}^* \end{bmatrix} = \begin{bmatrix} f_x^{-1} & 0 & -c_x f_x^{-1} \\ 0 & f_y^{-1} & -c_y f_y^{-1} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} i_t^* \\ j_t^* \\ 1 \end{bmatrix} d \tag{15}$$

The first part is followed by converting $(i_{ct}^*, j_{ct}^*, z_{ct}^*)$ (denoted by p_{ct}^*) to the robot end effector coordinate system (I_t^*, J_t^*, Z_t^*) (denoted by $\mathcal{P}_t^*$) conducting off-line hand-eye calibration, as shown in Eq. 16, where the rotation and translation parts are denoted by $\mathbf{R}$ and $\mathbf{T}$, and $\mathbf{0}_{1\times3}$ represents a 1×3 zero matrix.

$$\begin{bmatrix} \mathcal{P}_t^* \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ \mathbf{0}_{1\times3} & 1 \end{bmatrix} \begin{bmatrix} p_{ct}^* \\ 1 \end{bmatrix} \tag{16}$$

The final part involves the conversion between the gripper stroke W_t^* and rotation Θ_t^* relative to the grasp box's width w_t^*, and rotation θ_t^*, which can be manually adjusted because of their linear relationship.

After a series of conversions, the final grasp pose $(I_t^*, J_t^*, Z_t^*, \Theta_t^*, \Theta_{xt}^*, \Theta_{yt}^*)$ in the robot end effector coordinate system can be obtained, where Θ_{xt}^* and Θ_{yt}^* represent the constant rotations relative to the X-axis and the Y-axis. Therefore, the gripper can be moved to the target pose using inverse kinematics and its stroke is kept to the width W_t^*, thus achieving the avoidance of human hands and adjacent objects without emergency stops. The pseudocode of QFAAP is shown in Algorithm 1.

IV. Experiments

In this section, we validate the effectiveness of our proposed method through extensive experiments. In the benchmark datasets experiment, we first test the performance of AQP optimized by different grasping models. Then, we validate the generalization ability of AQP trained by one dataset across other datasets. Finally, we add PQGD to AQP to analyze the effectiveness of PQGD, as well as explore the impact of the iteration number on PQGD. In the real-world experiment, we first make the distance-based quantitative analysis related to the property of grasp quality score suppression by QFAAP in single-object scenarios. This is followed by comparing the detection performance of QFAAP with the original methods and the engineering-based methods in single-object scenarios. Later, we further compare QFAAP with these methods and the version of QFAAP without PQGD on a cobot in different HRI scenarios, including single-object HRI scenarios, mid-clutter HRI scenarios, and high-clutter HRI scenarios with multi-hand interference. Finally, we arranged the HRI user study to evaluate QFAAP from the user perspective.

A. Experimental Settings

1) Setting for QFAAP: We employ the Cornell Grasp Dataset [30], Jacquard Grasp dataset [52], and OCID Grasp Dataset [53]. The Cornell Grasp Dataset and the Jacquard Grasp datasets are single-object RGB-D datasets, while the OCID is a cluttered RGB-D dataset. Cornell comprises 885 RGB-D images with a resolution of 640×480, 240 different real objects, and 5k annotations. Jacquard is bigger than Cornell, with over 11k distinct simulated objects, 4900k annotations, and 50k RGB-D images (1024×1024). OCID [54], designed to evaluate semantic segmentation methods in complex scenarios, provides diverse settings, including objects, backgrounds, lighting conditions, and so on. Therefore, we utilized an improved version from [53] for the grasping model, consisting of over 1.7k RGB-D images (640×480) and 75k annotations.

Algorithm 1 Quality-focused Active Adversarial Policy

Input: Training sample x_i, realtime RGB frame $\mathbf{x}$ acquired sequentially from the video stream
Output: Optimal grasp in the robot end effector coordinate system $\mathcal{G}_t^*$
// Adversarial Quality Patch: Using sample x_i from grasp dataset $\mathbb{D}$, and solve Eq. 9 to optimize AQP.
for $x_i \in \mathbb{D}$ **do**
  $\mathbf{p}_t \leftarrow \mathcal{L}_{aqp}, \delta_{aqp}, x_i$
end for
// Projected Quality Gradient Descent : First, $\mathbf{p}_t$ is added to the hand region by $\mathcal{M}_h$, generating $\mathbf{x}'$. Then, shape-adaptive optimization of AQP is performed by solving Eq. 13, yielding $\mathbf{x}''$. Finally, $\mathbf{x}_t''$ is fed into the grasping model to obtain $\mathbf{Q}_t$, along with the quality map $\tilde{\mathbf{Q}}_t^h$ outside the hand region after guided by $\mathcal{M}_h$.
$\mathbf{x}' \leftarrow \mathbf{x}, \mathbf{p}_t, \mathcal{M}_h$
$\mathbf{x}'' \leftarrow \mathbf{x}'_{t-1}, \mathbf{x}', \mathcal{L}_{pqgd}, \delta_{pqgd}, \mathcal{M}_h, \varepsilon$
$\mathbf{Q}_t \leftarrow \mathbf{x}_t''$
$\tilde{\mathbf{Q}}_t^h \leftarrow \mathbf{Q}_t, \mathcal{M}_h$
// Active Adversarial for Robot Grasping: First, based on $\tilde{\mathbf{Q}}_t^h$, the grasp position (i_t^*, j_t^*) corresponding to the maximum quality score is computed. Then, the remaining grasp parameters are obtained using (i_t^*, j_t^*) to form the optimal grasp g_t^*. Finally, g_t^* is transformed into the optimal grasp $\mathcal{G}_t^*$ in the robot end effector coordinate system by solve Eq. 15, Eq. 16.
for $(i_t, j_t) \in (H, W)$ **do**
  $(i_t^*, j_t^*) \leftarrow \arg\max \tilde{\mathbf{Q}}_t^h(i_t, j_t)$
end for
$g_t^* \leftarrow (i_t^*, j_t^*), w_t^*, h_t^*, \theta_t^*$
$\mathcal{G}_t^* \leftarrow g_t^*$
return $\mathcal{G}_t^*$

We train these DNNs-based grasping models in advance, thus leveraging them for the optimization of AQP: GG-CNN [31], GG-CNN2 [32], GR-ConvNet [34], FCG-Net [35], SE-ResUNet [33], and TF-Grasp [55]. GR-ConvNet, FCG-Net, SE-ResUNet, and TF-Grasp support RGB images as input, while GG-CNN and GG-CNN2 accept Depth information. In our experiments, we extend GG-CNN and GG-CNN2 to handle RGB inputs by adjusting the number of input channels. These models were trained on a single NVIDIA RTX 4090 GPU with 24 GB of memory. The computer system is Ubuntu 22.04, and the deep learning framework is PyTorch 2.3.1 with CUDA 12.1. We follow the same image-wise setting in GR-ConvNet [34], randomly shuffling the entire dataset, selecting 90% for training and 10% for testing before training. During training stage, the data will be uniformly cropped to 224×224 (GG-CNN and GG-CNN2 are 300×300), the total number of epochs for training is set to 50, the learning rate δ_{model} is fixed to 0.001, batch size B is set to 8, and data augmentation (random zoom and random rotation) is applied (except Jacquard Grasp dataset). Finally, we employ the same rectangle (box) metric from [48] to assess the model performance, denoted as Original Accuracy (O-Acc). According to this metric, a predicted grasp by the grasping model is considered valid when it satisfies two conditions: the Intersection over Union score between the ground truth and predicted grasp rectangles is over 25%, and the offset between the orientation of the ground truth rectangle and that of the predicted grasp rectangle is less than 30°.

For the optimization of AQP, we use the same device, system, and training parameters as the grasping model. Differently, we first initialize an AQP with a uniform distribution of size 224×224 (300×300 for GG-CNN and GG-CNN2). Next, during each iteration, we apply a random scale (ranging from 0.1 to 1 of the original size) to the AQP and paste it onto a random position of the training sample. We set α, β, and γ in $\mathcal{L}_q^p$ and $\mathcal{L}_{aqp}$ to 0.1, 0.1, and 0.5, respectively. The initial learning rate δ_{aqp} is set to 0.03 (decreasing by a factor of ten at the 30th and 40th epochs). It is important to note that since AQP does not need to be printed in the real world, as required by adversarial patch attacks, no additional data augmentation operations for AQP are used. Finally, we evaluate the performance of the AQP on the test set using the previously defined Q-ACC.

For the operation of PQGD, since it only processes real-time RGB frames, we only need to set the following parameters: the iteration number N^i is set to 1, the learning rate δ_{pqgd} is fixed at 0.008, and ϵ is set to 8/255. In addition, we use the pre-trained model from [50] for real-time hand segmentation to guide the PQGD optimization. Finally, since PQGD is based on AQP, we use the same Q-ACC to evaluate the performance of PQGD.

2) Setting for Robot Grasping: Our robot grasping system and part of the experimental objects are illustrated in Fig. 3. For the grasping system, we adopt an eye-in-hand grasping architecture, where the camera is fixed on the robot, and the field of view faces downward. For the experimental objects, we collected 40 novel objects that are not included in the training dataset. We define the following evaluation criteria to assess the effectiveness of our method in the real world, including the success rate of detecting optimal grasps that do not occur on the hand or its adjacent objects (ND-ACC) and the collision

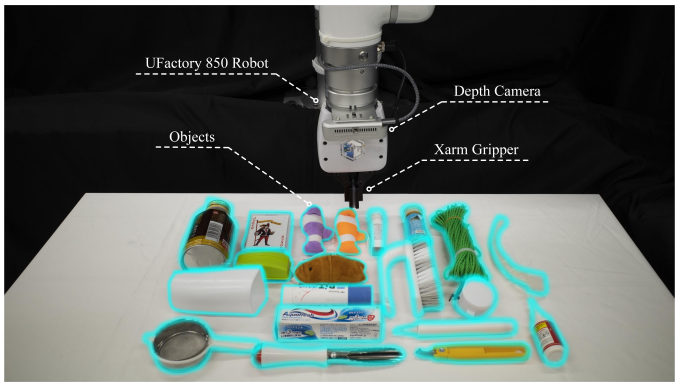

Fig. 3. Experimental setup of robot grasping: primarily consisting of an Intel RealSense D435 depth camera, a UFactory 850 robot, a UFactory xArm gripper, and part of the experimental objects (emphasized by blue borders).

TABLE I
RESULTS OF AQP ON THE CORNELL GRASP DATASET

Methods	O-ACC (%)	Q-ACC (%)	Runtime (s)
GG-CNN	87.6	**99.4**	**0.003**
GG-CNN2	92.1	71.4	**0.003**
GR-Convnet	96.6	94.2	0.005
FCG-Net	96.6	97.4	0.009
SE-ResUNet	95.5	90.4	0.013
TF-Grasp	**96.8**	27.0	0.008

TABLE II
RESULTS OF AQP ON THE OCID GRASP DATASET

Methods	O-ACC (%)	Q-ACC (%)	Runtime (s)
GG-CNN	18.6	96.9	**0.003**
GG-CNN2	44.6	90.0	**0.003**
GR-Convnet	**53.7**	93.9	0.006
FCG-Net	52.5	91.1	0.008
SE-ResUNet	46.3	**98.5**	0.014
TF-Grasp	26.0	94.1	0.007

TABLE III
RESULTS OF AQP ON THE JACQUARD GRASP DATASET

Methods	O-ACC (%)	Q-ACC (%)	Runtime (s)
GG-CNN	83.7	74.8	**0.004**
GG-CNN2	86.0	71.5	**0.004**
GR-Convnet	91.8	70.9	0.007
FCG-Net	86.3	79.3	0.011
SE-ResUNet	85.5	**82.3**	0.017
TF-Grasp	**93.6**	51.3	0.013

TABLE IV
RESULTS OF AQP GENERALIZABILITY ACROSS DIFFERENT DATASETS

Methods	C → O Q-ACC (%)	C → J Q-ACC (%)	O → C Q-ACC (%)	O → J Q-ACC (%)	J → C Q-ACC (%)	J → O Q-ACC (%)
GG-CNN	78.1 (↓ 21.3)	76.6 (↓ 22.8)	92.8 (↓ 4.1)	86.4 (↓ 10.5)	67.4 (↓ 7.4)	65.2 (↓ 9.6)
GG-CNN2	22.4 (↓ 49.0)	40.9 (↓ 30.5)	82.9 (↓ 7.1)	78.0 (↓ 12.0)	65.5 (↓ 6.0)	51.5 (↓ 20.0)
GR-Convnet	89.3 (↓ 4.9)	89.8 (↓ 4.4)	71.2 (↓ 22.7)	82.3 (↓ 11.6)	55.3 (↓ 15.6)	51.1 (↓ 19.8)
FCG-Net	77.2 (↓ 20.2)	88.3 (↓ 9.1)	84.6 (↓ 6.5)	86.0 (↓ 5.1)	66.0 (↓ 13.3)	59.2 (↓ 20.1)
SE-ResUNet	86.2 (↓ 4.2)	87.5 (↓ 2.9)	98.3 (↓ 0.2)	**98.6** (↑ 0.1)	80.0 (↓ 2.3)	71.6 (↓ 10.7)
TF-Grasp	19.3 (↓ 7.7)	23.3 (↓ 3.7)	88.4 (↓ 5.7)	87.8 (↓ 6.3)	46.0 (↓ 5.3)	29.8 (↓ 21.5)

TABLE V
RESULTS OF AQP&PQGD ON THE CORNELL GRASP DATASET

Methods	O-ACC (%)	Q-ACC (%)	Runtime (s)
GG-CNN	87.6	**99.5**	**0.011**
GG-CNN2	92.1	72.3	0.016
GR-Convnet	96.6	94.9	0.031
FCG-Net	96.6	97.6	0.042
SE-ResUNet	95.5	91.4	0.056
TF-Grasp	**96.8**	31.3	0.038

TABLE VI
RESULTS OF AQP&PQGD ON THE OCID GRASP DATASET

Methods	O-ACC (%)	Q-ACC (%)	Runtime (s)
GG-CNN	18.6	97.6	**0.012**
GG-CNN2	44.6	93.0	0.017
GR-Convnet	**53.7**	94.9	0.031
FCG-Net	52.5	92.4	0.044
SE-ResUNet	46.3	**98.7**	0.058
TF-Grasp	26.0	94.7	0.033

TABLE VII
RESULTS OF AQP&PQGD ON THE JACQUARD GRASP DATASET

Methods	O-ACC (%)	Q-ACC (%)	Runtime (s)
GG-CNN	83.7	76.0	**0.017**
GG-CNN2	86.0	74.6	0.023
GR-Convnet	91.8	73.4	0.037
FCG-Net	86.3	82.2	0.052
SE-ResUNet	85.5	**84.2**	0.069
TF-Grasp	**93.6**	57.1	0.069

rate of the robot to the hand during the grasping process (CH-Rate). For the safety evaluation setting, the hand will enter the camera view and remain static before conducting grasping in HRI scenarios (the dynamic evaluation part is shown in APPENDIX-B).

Specifically, the hand will approach an object with the highest grasp quality score (we know the location of the highest grasp quality score in advance) in the camera view, and the distance between the hand and this object remains within 0.5 cm, without making physical contact. We define the object with this distance to the human hand as the adjacent object. This setting allows us to evaluate the effectiveness of our method under extremely challenging conditions. If the human hand is capable of reducing the highest grasp quality score in the scene, then it may also reduce all other grasp quality scores in the same manner, which ensures that the presence of the human hand at any location within the scene remains safe. Finally, it is important to emphasize that we will compare methods that may cause injury to the human in the grasping. Therefore, we fix the robot at a safe height (other predicted position parameters by the grasping model remain unchanged) and then slowly move the robot to the actual height during each grasping.

B. Effectiveness of AQP

We employ the same experimental setting of AQP and grasping model discussed in Section IV-A1, with the corresponding results presented in Table I (optimized using the Cornell Grasp dataset), Table II (optimized using the OCID Grasp dataset), and Table III (optimized using the Jacquard Grasp dataset). To ensure consistency and avoid confusion, we refer to some results reported in the original papers, such as the O-Acc of GR-ConvNet [34] and TF-Grasp [55] trained on the Cornell and Jacquard Grasp datasets. In Table I, AQP optimized by most models achieve a Q-AAC exceeding 90%, except for those optimized by GG-CNN2, which attains 71.4%, and TF-Grasp, which records 27.0%. In Table II, AQP optimized by all models exhibits a Q-AAC above 90%. In Table III, despite being optimized using a large-scale dataset (with extensive test images for testing), AQP optimized by most models still surpass 70%, except for those optimized by TF-Grasp, which gets 51.3%.

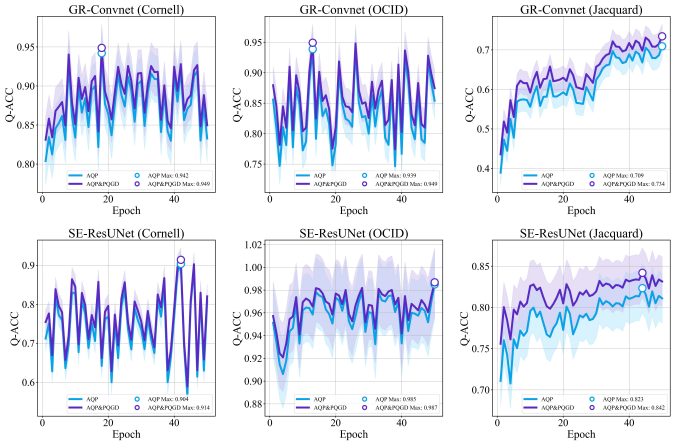

Fig. 4. Line graphs showing the effectiveness of PQGD across all epochs, including its impact on the AQP optimized by GR-ConvNet and three different datasets, as well as the AQP optimized by SE-ResUNet and three different datasets. Here, the AQP and AQP&PQGD are represented by blue and purple lines, and we also use blue and purple dots to emphasize their corresponding maximum quality score across all epochs.

The above analyses indicate that AQP optimized across different datasets and models is effective. Furthermore, AQP optimized using cluttered datasets demonstrates superior performance compared to single-object datasets, providing a solid foundation for the subsequent application of QFAAP in cluttered grasping scenarios. Finally, we visualize the quality performance of AQP across these datasets in the first two rows of Fig. 5, Fig. 6, and Fig. 7. As illustrated in this figure, although the highest quality scores are not located on AQP in columns 3 and 5-8 of Fig. 6, as well as columns 1, 2, and 5 of Fig. 7, most highest scores are concentrated on AQP, further demonstrating the effectiveness of AQP in manipulating the quality score.

C. Generalizability of AQP

In this part, we also adopt the same experimental setting for the AQP and grasping model as discussed in Section IV-A1. The results are presented in Table IV, where (C $\rightarrow$ O) denotes that the AQP is trained on the Cornell Grasp dataset and tested on the OCID Grasp dataset; other notations follow a similar convention. From this table, although the Q-ACC of the AQP

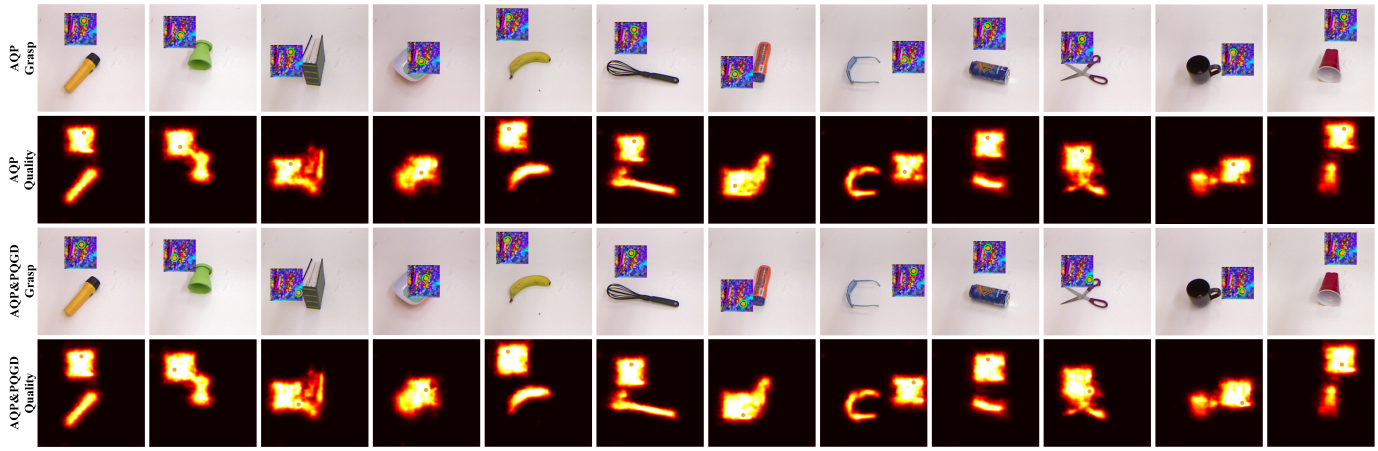

Fig. 5. Quality score visualization of AQP (first two rows) before and after adding PQGD (last two rows). Here, the GGCNN2 and the Cornell Grasp dataset are used to optimize the AQP. And AQP is scaled to 0.3 of the original size (the same size as the image).

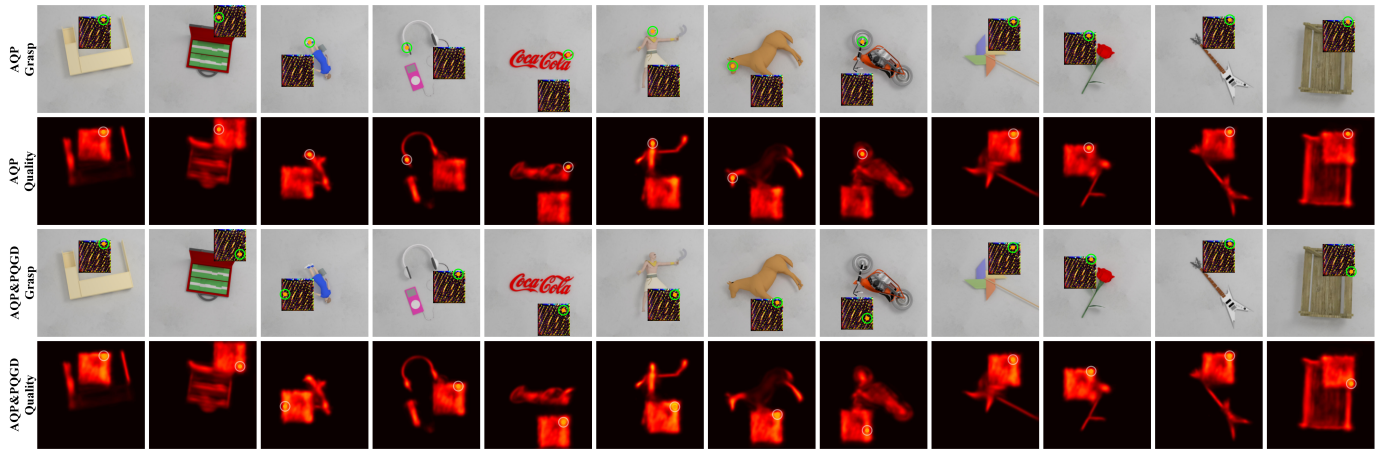

Fig. 6. The meaning of each row is consistent with Fig. 5. Here, the SE-ResUNet and the Jacquard Grasp dataset are used to optimize the AQP. And AQP is scaled to 0.3 of the original size (the same size as the image).

Fig. 7. The meaning of each row is consistent with Figs. 5 and 6. Here, the GR-ConvNet and the OCID Grasp dataset are used to optimize the AQP. And AQP is scaled to 0.3 of the original size (the same size as the image).

trained on a specific dataset generally decreases when tested on different datasets, most of them still maintain a Q-ACC above 60%. In particular, most of the AQP trained on the OCID Grasp dataset, which contains cluttered scenes, even still achieves high Q-ACC (above 80%) when tested on other datasets. For example, the AQP trained using SE-ResUNet and the OCID dataset even achieves an increased Q-ACC of 98.6% on the Jacquard Grasp dataset. These results demonstrate that the AQP exhibits a certain generalizability across different datasets, with training on cluttered-scene datasets leading to

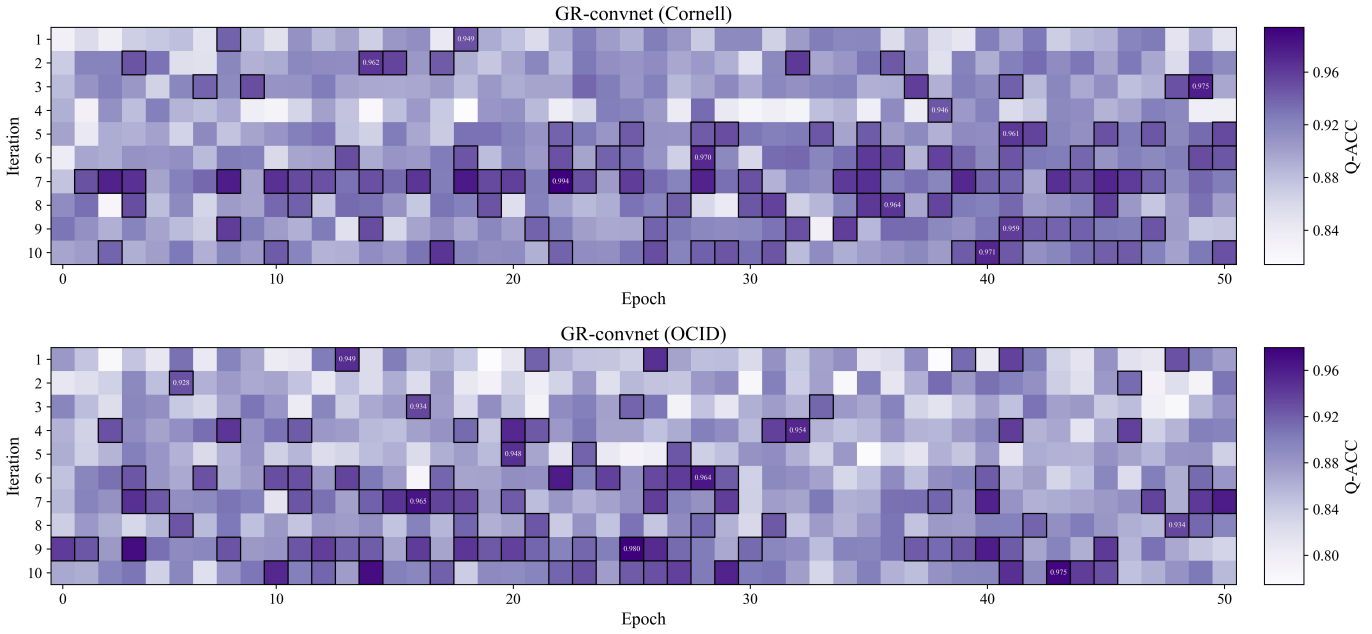

Fig. 8. Heatmap showing the impact of the iteration number N^i on PQGD across all epochs. Here, the AQP is optimized by GR-ConvNet on the Cornell Grasp dataset (upper sub-figure) and the OCID Grasp dataset (lower sub-figure). In addition, the maximum quality score for each row is printed in white numbers for emphasis.

more robust performance.

D. Effectiveness of PQGD

We validate PQGD by applying it to the AQP optimized in Section IV-B and employing the experimental settings of PQGD discussed in Section IV-A1. In addition, the iteration number N^i is set to 1 in this part. The experimental results are presented in Table V (for the Cornell Grasp dataset), Table VI (for the OCID Grasp dataset), and Table VII (for the Jacquard Grasp dataset). By comparing these tables with their corresponding Table I, Table II, and Table III, it can be observed that PQGD consistently improves the quality score of the AQP optimized by all models and datasets, with a more pronounced effect on the Jacquard Grasp dataset, resulting in an overall quality score improvement of approximately 2%. Although the prediction speed (all running on one NVIDIA RTX 4090 GPU) decreases with adding PQGD, it remains real-time performance. This reduction has no impact on the efficiency of robot grasping, as the movement time of the robot is significantly longer than the prediction time of the grasping model in practice. Therefore, we enable AQP to rapidly acquire the human hand shape adaptability at a low cost. Additionally, we show the effectiveness of PQGD across all epochs in Fig. 4, including its impact on the AQP optimized by GR-ConvNet and three different datasets, as well as the AQP optimized by SE-ResUNet and three different datasets. As illustrated in this figure, it is evident that PQGD remains effective throughout all epochs. Since we applied only a random scale to AQP without additional augmentations, the quality score exhibits fluctuations on the smaller Cornell Grasp and OCID Grasp datasets due to overfitting. However, this issue is eliminated for the larger Jacquard Grasp dataset. Overall, this fluctuation does not impact the subsequent deployment of our QFAAP, as our objective is not to attack the model but to ensure the achievement of a high quality score. We also visualize the quality performance of AQP after adding PQGD across these datasets in the last two rows of Fig. 5, 6, and 7. As shown in these figures, all of the mean quality scores within the AQP can be further improved after adding PQGD. In addition, all cases where the highest quality scores were originally outside the AQP (e.g., columns 3–8 in Fig. 6 and columns 1, 2, and 5 in Fig. 7) are corrected after adding PQGD, with the highest quality scores shifting into the AQP; this demonstrates that PQGD can further enhance the highest quality scores within the AQP to some extent. Overall, the PQGD proves effective across different datasets, laying a foundation for subsequent grasping experiments to improve HRI safety by suppressing low-quality scores through high-quality scores with adaptability.

E. Impact of Iteration Number on PQGD

This part primarily investigates the impact of the iteration number N^i on PQGD. We conduct experiments using the AQP optimized by GR-ConvNet on the Cornell Grasp dataset and the OCID Grasp dataset, with the iteration number N^i ranging from 1 to 10. Other experimental settings remain the same as in Section IV-A1. The results are presented in Table VIII, which shows that the optimal number of iterations for PQGD is around 7 for the Cornell Grasp dataset and around 9 for the OCID Grasp dataset. Overall, different numbers of iterations consistently lead to an improvement in Q-ACC. Additionally, we visualize the effect of the number of iterations N^i on PQGD across all epochs in Fig. 8. In the upper part of the figure (Cornell Grasp dataset), it can be observed

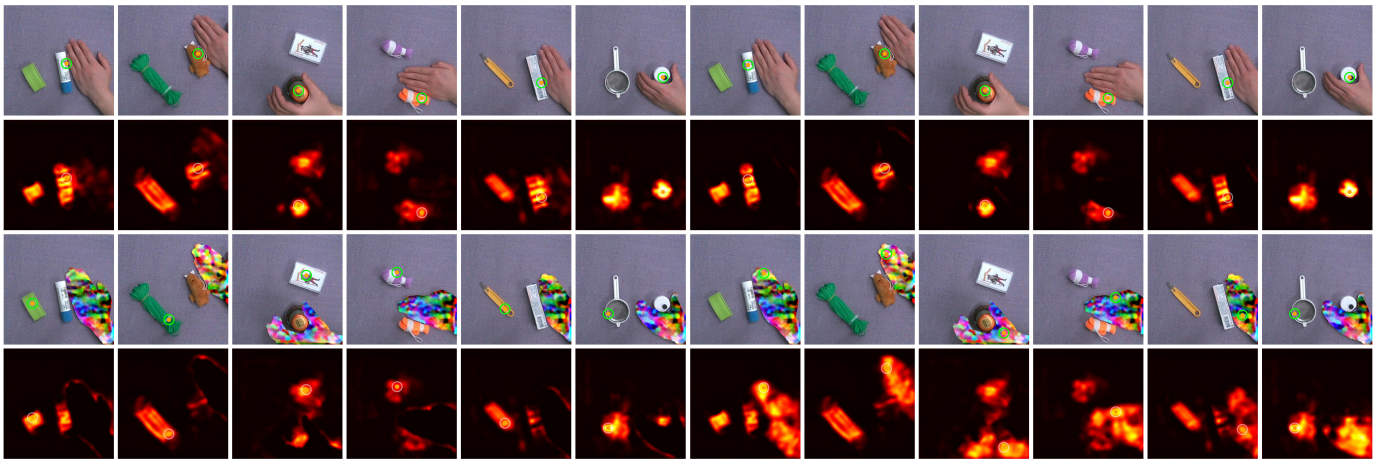

Fig. 9. Visualization of optimal grasp and quality map for Original (first two rows of the first to sixth columns), Original-SZ (first two rows of the seventh to twelfth columns), QFAAP (last two rows of the first to sixth columns), and QFAAP-NSZ (last two rows of the seventh to twelfth columns).

TABLE VIII
THE IMPACT OF DIFFERENT ITERATION NUMBERS OF PQGD ON Q-ACC

Iteration Number N^i	1	2	3	4	5	6	7	8	9	10
Cornell Q-ACC (%)	94.9	96.2	97.5	94.6	96.1	97.0	**99.4**	96.4	95.9	97.1
OCID Q-ACC (%)	94.9	92.8	93.4	95.4	94.8	96.4	96.5	93.4	**98.0**	97.5

TABLE IX
DETECTION RESULTS BETWEEN QFAAP AND ORIGINAL METHODS

Object Pairs	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Overall (%)	Runtime (s)
Original ND-ACC	1/10	0/10	2/10	0/10	1/10	1/10	1/10	1/10	3/10	3/10	13	0.0069
Original-SZ ND-ACC	1/10	0/10	3/10	0/10	1/10	2/10	1/10	1/10	3/10	3/10	15	0.0087
QFAAP ND-ACC	7/10	9/10	9/10	10/10	8/10	9/10	10/10	8/10	8/10	10/10	**88**	0.0759

TABLE X
DETECTION RESULTS BETWEEN QFAAP AND ENGINEERING METHODS

Object Pairs	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Overall (%)
Original-DSZ ND-ACC	3/10	4/10	5/10	4/10	4/10	5/10	3/10	4/10	4/10	4/10	40
Original-Decay ND-ACC	5/10	6/10	8/10	6/10	4/10	6/10	6/10	6/10	4/10	7/10	58
QFAAP ND-ACC	9/10	10/10	10/10	9/10	8/10	8/10	9/10	9/10	8/10	9/10	**89**

that when the iteration number is 7, the high quality scores (darker purple blocks) are more densely distributed across all epochs compared to other iteration numbers, indicating greater stability. Similarly, in the lower part of the figure (OCID Grasp dataset), the high-quality scores are most densely concentrated when the iteration number is 9. Therefore, the observation from this figure aligns well with the statements discussed in Table VIII.

F. Effectiveness of QFAAP in Real World

1) Detection Comparison with Original and Engineering Methods: Here, we compare the detection performance of QFAAP with original and engineering methods in single-object scenarios. First, we select 20 objects from the experimental objects and group them into ten pairs. To assess these methods, the hand approaches an object with the highest quality score within each object pair ten times, where the object positions and human hand postures are randomly adjusted in each trial. The comparison methods are divided into two groups.

The first group is original methods, including Original (the original grasping model) and Original-SZ (a variant of the grasping model where the quality score of the hand region is set to zero). The second group is engineering methods, including Original-DSZ (enhanced version of Original-SZ with the zeroed area dilation) and Original-Decay (enhanced version of Original-DSZ with the distance-based linear decay). For Original-DSZ, the dilation size is set to 10 pixels since a larger size will reduce the workspace of the robot. For Original-Decay, the dilation size is set to 15 pixels, and a distance-based linear decay factor ranging from 0 to 0.8 is applied to the quality score of the region between the boundary of the original area and the boundary of the dilated area, that is the closer to the boundary of the original area, the lower the

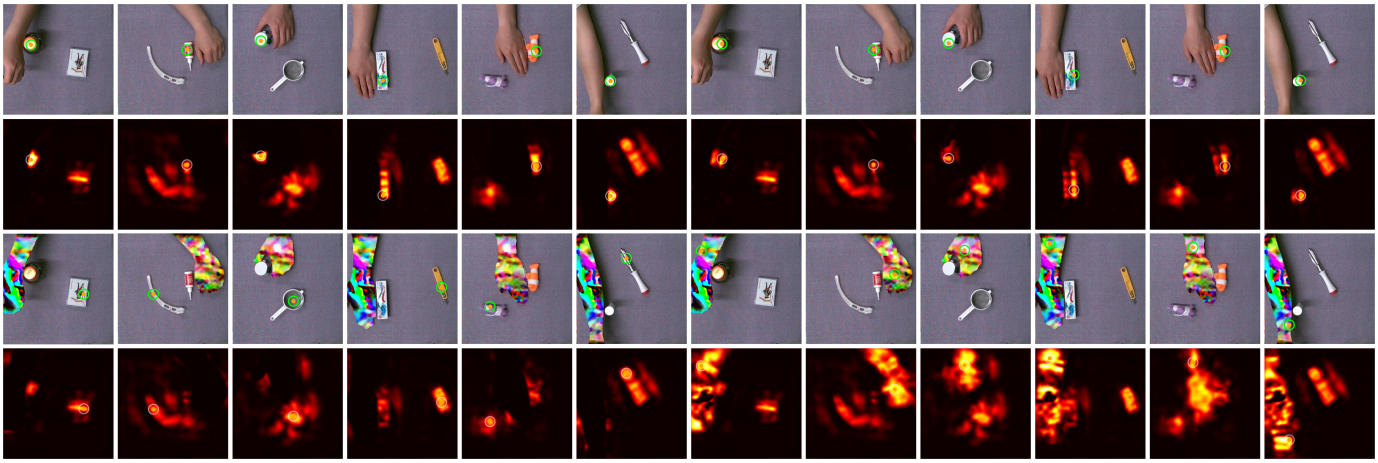

Fig. 10. Visualization of optimal grasp and quality map for Original-DSZ (first two rows of the first to sixth columns), Original-Decay (first two rows of the seventh to twelfth columns), QFAAP (last two rows of the first to sixth columns), and QFAAP-NSZ (last two rows of the seventh to twelfth columns).

TABLE XI
DISTANCE-BASED DETECTION RESUTLS FOR QFAAP

Objects	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20
QFAAP (2.0 cm) ND-ACC	0/5	2/5	1/5	1/5	2/5	0/5	1/5	2/5	1/5	0/5	3/5	1/5	0/5	0/5	1/5	0/5	0/5	0/5	0/5	0/5
QFAAP (1.0 cm) ND-ACC	1/5	3/5	2/5	3/5	4/5	0/5	2/5	4/5	3/5	0/5	3/5	0/5	0/5	1/5	3/5	2/5	0/5	4/5	0/5	2/5
QFAAP (0.5 cm) ND-ACC	3/5	3/5	3/5	5/5	4/5	2/5	4/5	5/5	4/5	3/5	4/5	5/5	3/5	1/5	5/5	5/5	3/5	4/5	3/5	5/5
QFAAP (0.0 cm) ND-ACC	5/5	5/5	4/5	5/5	5/5	5/5	5/5	5/5	4/5	5/5	5/5	5/5	5/5	4/5	5/5	5/5	3/5	4/5	5/5	5/5
Objects	B21	B22	P23	P24	P25	P26	P27	P28	P29	P30	B31	B32	B33	B34	B35	B36	B37	B38	B39	Overall (%)
QFAAP (2.0 cm) ND-ACC	0/5	4/5	0/5	2/5	0/5	2/5	0/5	0/5	3/5	2/5	0/5	0/5	3/5	2/5	0/5	0/5	1/5	1/5	0/5	17.4
QFAAP (1.0 cm) ND-ACC	3/5	4/5	3/5	3/5	3/5	5/5	2/5	4/5	5/5	3/5	3/5	3/5	5/5	3/5	1/5	3/5	1/5	2/5	0/5	47.6
QFAAP (0.5 cm) ND-ACC	5/5	5/5	4/5	5/5	5/5	5/5	3/5	5/5	5/5	5/5	5/5	4/5	5/5	5/5	2/5	5/5	3/5	4/5	4/5	81.0
QFAAP (0.0 cm) ND-ACC	5/5	5/5	5/5	5/5	5/5	5/5	3/5	5/5	5/5	5/5	5/5	4/5	5/5	5/5	4/5	5/5	3/5	5/5	4/5	**93.3**

TABLE XII
THE IMPACT OF PQGD ON QFAAP IN REAL GRASPING

Object Pairs	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Overall (%)
QFAAP-without PQGD CH-Rate	2/10	3/10	2/10	1/10	2/10	3/10	1/10	2/10	1/10	3/10	20
QFAAP CH-Rate	1/10	2/10	2/10	0/10	1/10	1/10	0/10	2/10	1/10	2/10	**12**

quality score will be. For our method QFAAP, we use the AQP optimized by GR-ConvNet and OCID Grasp dataset and set the iteration number N^i to 5 for PQGD. All other experimental settings about QFAAP are consistent with Section IV-A2.

The results between our methods and original methods are presented in Table IX. Our method significantly outperforms both Original and Original-SZ, over 70% ND-ACC, which means that it can noticeably enhance the safety performance of the grasping model in single-object scenarios. For the runtime (all running on one NVIDIA RTX 3090 Ti GPU), although QFAPP is lower than other methods due to the incorporation of the hand segmentation algorithm, it still gets 0.0759 s per frame, which satisfies the real-time requirement in real-world grasping. Then, we show the results between our methods and engineering methods in Table X. Our method still surpass them by a large margin, over 30% ND-ACC. This demonstrates the superiority of the shape adaptability of QFAAP compared with Original-Decay, and the better performance of QFAAP in enhancing safety without influencing the workspace of the robot compared with Original-DSZ. We also visualize some of our results in Fig. 9 and Fig. 10, including the optimal grasp and quality map for Original, Original-SZ, Original-DSZ, Original-Decay, QFAAP, and QFAAP-NSZ (a variant of the QFAAP where the quality score of the hand region is not set to zero). As shown in these figures, compared with other methods, our method can always shift the highest quality score to the object away from the human hand by decreasing the quality score of the object near the human hand, no matter the different scenarios and hand poses. It should be noted that QFAAP-NSZ is only to emphasize the strength of the quality score for QFAAP and is not included in the experimental tables. Finally, the few failure cases of QFAAP primarily result from situations where the object approached by the human hand still maintains a higher quality score than the other object. In future work, we will enhance our optimization methods to strengthen QFAAP.

2) The Impact of Distance on the Effectiveness of QFAAP: We conduct extensive distance-based quantitative experiments to explore the quality suppression behavior of QFAPP, using an experimental setup similar to that in Section IV-F1. Specifically, one object among the 40 experimental objects is selected as the non-target object, while the remaining 39

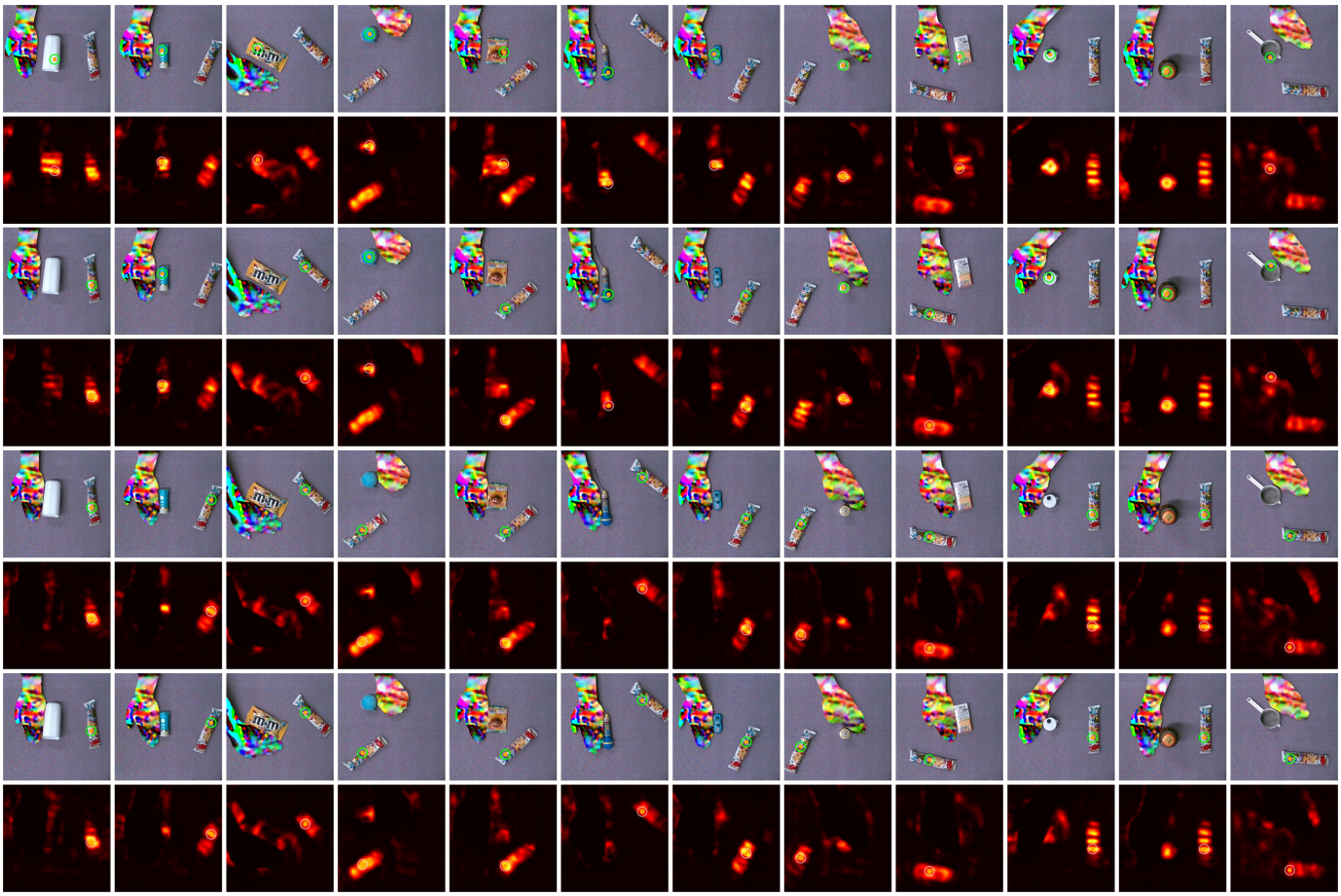

Fig. 11. Visualization of optimal grasp and quality map for QFAPP with distance 2 cm (first two rows), distance 1 cm (third and fourth rows), distance 0.5 cm (fifth and sixth rows), and distance 0 cm (last two rows).

objects are treated as target objects to be approached by the human hand, forming 39 object pairs in total. For each of the 39 pairs, we perform five trials of hand-approaching experiments. In each trial, the position of the object pair and the posture of the hand are first randomly changed. Then, the hand gradually approaches the target object with the same posture until contact is made. During this approaching process, we record the changes in the highest grasp quality score at distances of 2 cm, 1 cm, 0.5 cm, and upon contact. The results are shown in Table XI, where noticeable suppression begins to occur at a distance of 1 cm, when the ND-ACC reaches 47.6%. Subsequently, the ND-ACC increases to 81.0% at 0.5 cm and further to 93.3% upon contact with the target object. These results strongly demonstrate that our distance-based quantitative analysis aligns well with the property in [51], namely, that the quality score suppression of QFAPP becomes effective when the hand is within 0.5–1 cm of the target object and reaches its maximum effect at contact.

We further visualize the changes in quality scores and optimal grasps for different distances in Fig. 11. When the distance is 2 cm (first two rows: optimal grasps in the first row, corresponding quality scores in the second row), almost no suppression effect is observed. At a distance of 1 cm (third and fourth rows), some suppression occurs, though failures are still observed in columns 2, 4, 6, 8, 10, 11, and 12. When the distance is reduced to 0.5 cm (fifth and sixth rows), or contact is made (last two rows), the highest quality score is consistently shifted away from the object near the human hand. Notably, in the first and ninth columns of the last two rows (contact case), the quality scores of adjacent target objects are nearly suppressed to zero.

3) Grasping Comparison with the Version of QFAAP without PQGD: In this part, we evaluate the influence of PQGD on QFAPP in a real robot grasping system. We follow the same experimental setting in Section IV-F1 and Section IV-A2. Specifically, we perform 10 grasps for each object pair where the object positions and human hand postures are randomly adjusted in each grasp. The experimental results are presented in Table XII. The CH-Rate of QFAAP without PQGD reaches 20%. After integrating PQGD, the CH-Rate decreases to 12%, demonstrating that PQGD can also effectively enhance the fast adaptability to the human hand shape in real-world grasping scenarios. Finally, we showcase the grasping performance with and without PQGD in our demo videos.

4) Grasping Comparison with Original and Engineering Methods: We use a similar experimental setting as in Section IV-F1 and Section IV-A2 for this part. Specifically, we first select 10 objects from the experimental objects to create 10 mid-clutter grasping scenes and perform 10 grasps (the hand

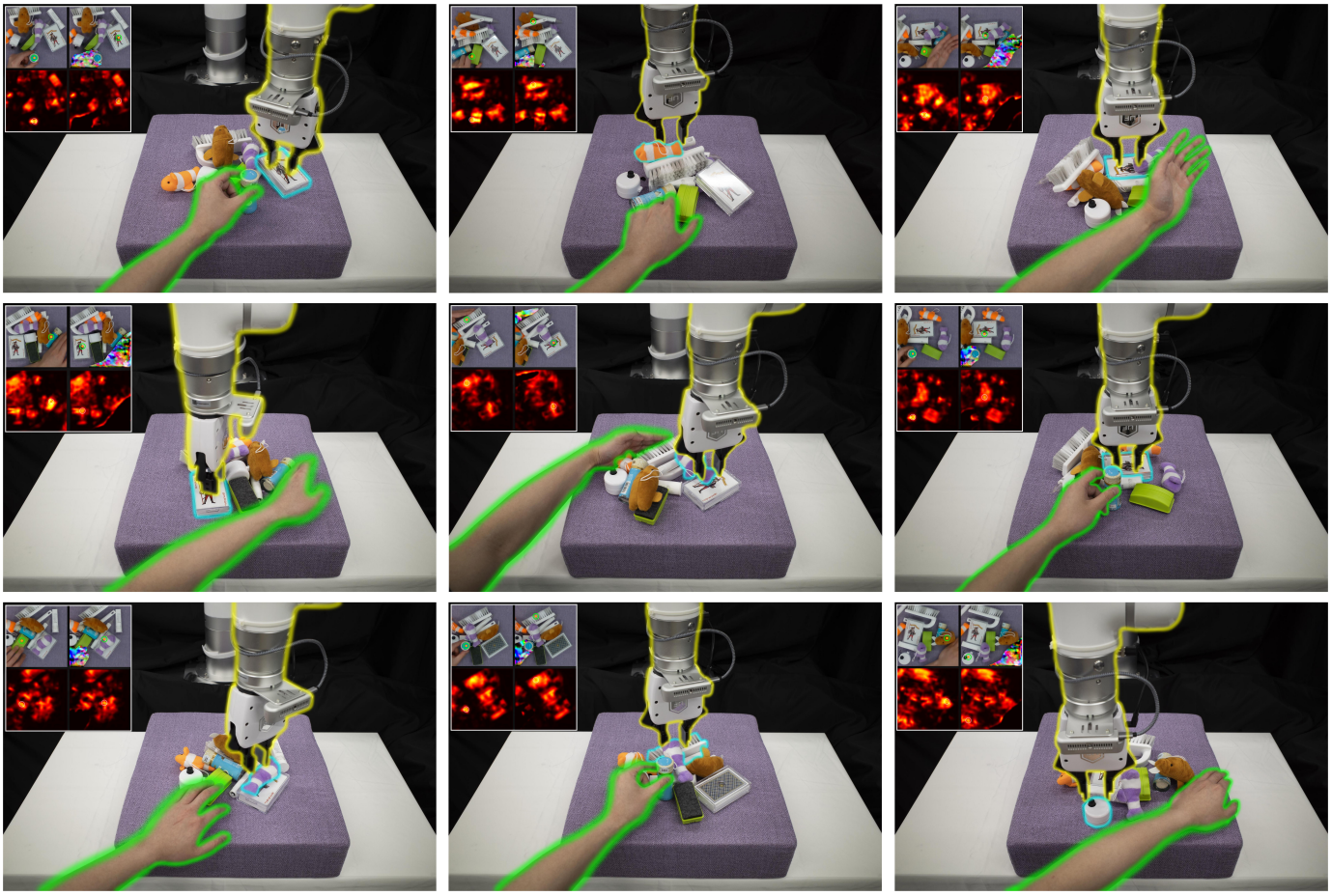

Fig. 12. Grasping in mid-clutter scenarios. We use yellow, green, and blue borders to highlight the robot, the human hand, and the objects being grasped in each subfigure. In addition, we added the optimal grasp and quality map for QFAAP (left) and the original grasping model (right) to each subfigure.

TABLE XIII
GRASPING RESULTS BETWEEN QFAAP AND ORIGINAL METHODS

Scenarios	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Overall (%)
Original CH-Rate	8/10	6/10	6/10	7/10	8/10	4/10	5/10	6/10	7/10	5/10	62
Original-SZ CH-Rate	6/10	6/10	6/10	6/10	7/10	4/10	5/10	6/10	7/10	5/10	58
QFAAP CH-Rate	2/10	1/10	2/10	3/10	2/10	0/10	2/10	1/10	2/10	1/10	**16**

pose will be changed in each grasping) for each scene to compare QFAAP with the original method. Then, similarly, we select 30 objects from the experimental objects to create 5 high-clutter grasping scenes and perform 30 grasps with multi-hand interference for each scene to compare QFAAP with the engineering method.

The experimental results between QFAAP and original methods are shown in Table XIII, where our method consistently outperforms both the Original and Original-SZ methods, achieving a notably low CH-Rate of 16%. This result demonstrates the effectiveness of QFAAP in enhancing the safety of the HRI in mid-clutter grasping scenarios. We also visualize some grasping results of QFAAP in Fig. 12. Compared with the original grasping model, our method can effectively shift the robot to grasp the object away from the human hand.

The experimental results between QFAAP and engineering methods are shown in Table XIV, where our method also outperforms both the Original-DSZ and Original-Decay methods, achieving a promising CH-Rate of 13.3%, which is more than 20% lower than them. This result demonstrates the superiority of QFAAP in enhancing the safety of the HRI in high-clutter grasping scenarios with multi-hand interference. We also visualize some grasping results of QFAAP in Fig. 13. The first row shows normal grasping without hand interference. The second row shows the result without our method under multiple hand interferences, where the robot easily collides

TABLE XIV
GRASPING RESULTS BETWEEN QFAAP AND ENGINEERING METHODS

Scenarios	S1	S2	S3	S4	S5	Overall (%)
Original-DSZ CH-Rate	18/30	15/30	13/30	16/30	14/30	50.7
Original-Decay CH-Rate	12/30	11/30	10/30	11/30	11/30	36.7
QFAAP CH-Rate	4/30	7/30	3/30	3/30	3/30	**13.3**

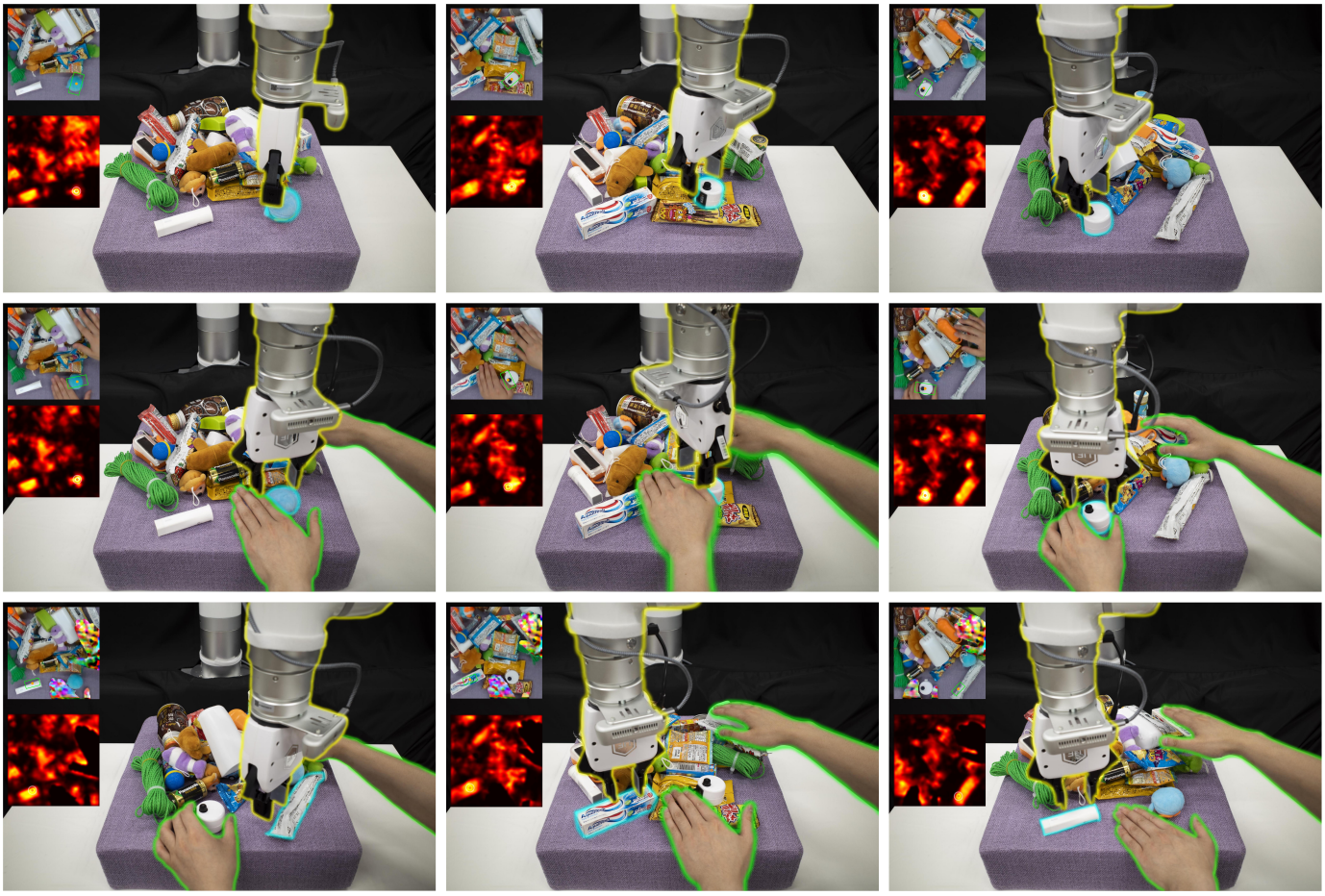

Fig. 13. Grasping in high-clutter scenarios with bimanual interference. The first row shows normal grasping without hand interference. The second and third rows show the grasping without and with our method under bimanual interference. We use yellow, green, and blue borders to highlight the robot, the human hands, and the objects being grasped in each subfigure. In addition, we added the optimal grasp and quality map to each subfigure.

TABLE XV
RESULTS OF HRI USER STUDY

Participants	ADCS		ADHP		UPS		RRF		UOS	
	Original	QFAAP	Original	QFAAP	Original	QFAAP	Original	QFAAP	Original	QFAAP
Participant 1	2	4 (↑ 2)	1	5 (↑ 4)	2	5 (↑ 3)	4	4 (−)	2	4 (↑ 2)
Participant 1	1	4 (↑ 3)	1	4 (↑ 3)	2	5 (↑ 3)	5	4 (↓ 1)	2	4 (↑ 2)
Participant 3	2	4 (↑ 2)	1	5 (↑ 4)	3	5 (↑ 2)	4	4 (−)	2	5 (↑ 3)
Participant 4	1	4 (↑ 3)	1	5 (↑ 4)	0	4 (↑ 4)	4	4 (−)	1	4 (↑ 3)
Participant 5	3	4 (↑ 1)	1	5 (↑ 4)	1	5 (↑ 4)	4	3 (↓ 1)	2	5 (↑ 3)
Average	1.8	**4** (↑ 2.2)	1	**4.8** (↑ 3.8)	1.6	**4.8** (↑ 3.2)	**4.2**	3.8 (↓ 0.4)	1.8	**4.4** (↑ 2.6)

with the human hands. The third row presents the result using our method under the same multi-hand interference scenario, where the robot successfully avoids all hands and nearby objects during grasping. Finally, the reasons for the failure cases of QFAAP in these scenarios remain consistent with those in Section IV-F1.

5) HRI User Study: In this part, we conduct the HRI user study to evaluate the safety of our proposed method from the users' perspective. To minimize safety risks during the experiments while ensuring the depth of this study, we strictly limit the number of participants to five, all of whom are researchers with professional knowledge in robotics. And we have obtained their approvals. We define the following user-centered evaluation metrics: ADCS (the adaptability of the method to different clutter scenarios), ADHP (the adaptability of the method to different hand poses), UPS (the user perceived safety), RRF (the robot response fluency), and UOS (the user overall satisfaction). Each metric is rated on a five-point integer scale (1 to 5).

During the experiment, users compare the performance of QFAAP with the Original method under a similar setting as described in Section IV-F4. Specifically, each user performs five interactions with each method in a high-clutter scene containing 30 objects, testing whether the robot can grasp objects while avoiding human hands and their neighboring objects. In addition, users are allowed to choose single-hand

or multi-hand configurations freely for each interaction, and the hand poses and object positions are varied across all trials.

The user feedback results are summarized in Table XV. As shown in the table, users consistently rated QFAAP significantly higher than the Original method across nearly all evaluation metrics. For example, (4 vs. 1.8) in the average of ADCS, (4.8 vs. 1) in the average of ADHP, (4.8 vs. 1.6) in the average of UPS, and (4.4 vs. 1.8) in the average of UOS. Although users reported a slightly lower RRF for QFAAP compared to the Original method (a difference of 0.4), they perceived the difference as minor and confirmed that QFAAP is capable of conducting HRI in real-time. Finally, after completing the experiments, all participants expressed that QFAAP demonstrated strong safety and adaptability, and were willing to deploy this method. We further illustrate several examples of the HRI process in Fig. 14, where it can be seen that regardless of the number of hands, hand poses, or scene variations, the robot consistently performs grasps while avoiding human hands and their nearby objects. More HRI processes are recorded in the demo videos.

Fig. 14. Examples of HRI user study. We use yellow, green, and blue borders to highlight the robot, the human hand, and the objects being grasped.

V. Conclusion

In this paper, we proposed the Quality-focused Active Adversarial Policy (QFAAP), which first optimized an Adversarial Quality Patch (AQP) with high quality scores using the adversarial quality patch loss and a grasp dataset. Then, the Projected Quality Gradient Descent (PQGD) was introduced to optimize AQP further, endowing it with the adaptability to the human hand shape. By leveraging AQP and PQGD, the hand itself can be an active perturbation source against nearby objects, reducing their quality scores. Further setting the quality score of the hand to zero will reduce the grasping priority of both the hand and its adjacent objects, enabling the robot to avoid them without emergency stops for autonomous grasping. We conducted extensive experiments on the benchmark datasets and a cobot, showing that QFAAP can improve the safety of robot grasping both in single-object and cluttered HRI scenarios.

Future work can be divided into two major parts. The first part can focus on addressing the issues highlighted in Section IV-F to enhance the method proposed in this paper. The second part involves exploring how to extend QFAAP to incorporate multimodal properties, which can then be utilized to address the backdoor attack problem proposed in [51].

Appendix

In this appendix, we first conduct additional grasping experiments to assess whether our method affects the original grasping performance of the model. Moreover, we validate the effectiveness of our method in scenarios involving hand dynamic interference.

A. Grasping with and without Hand Interference

Since QFAAP employs the same grasping model as the original method, it only modifies the output grasp quality scores when hand interference is present, thereby altering the grasping sequence. In the absence of such interference, it remains consistent with the original method. Therefore, in this section, we directly compare grasping performance with and without human-hand interference to verify whether our

Fig. 15. Cases of the DRD process under hand dynamic interference. Each row corresponds to one case, and the images in each row respectively illustrate the initial approach of the robot to the target object, the first deviation of the robot after interference, the re-approach (return) of the robot to the target object after the hand departs, and the second deviation of the robot after the second interference. Yellow, green, and blue borders are also used to highlight the robot, the human hand, and the target objects, respectively.

TABLE XVI
Grasping results with and without hand Interference

Scenarios	S1	S2	S3	S4	S5	GS-Rate (%)
without hand	24/30	24/30	24/30	25/30	25/30	**81.3**
with hand	23/30	24/30	24/30	24/30	25/30	80.0

TABLE XVII
Grasping results with hand dynamic Interference

Scenarios	S1	S2	S3	S4	S5	DRD-Rate (%)
QFAAP-without PQGD	22/30	22/30	23/30	23/30	22/30	74.7
QFAAP	24/30	24/30	25/30	27/30	26/30	**84.0**

method affects the original performance of the grasping model. Specifically, we use a similar experimental setting as in Section IV-F4, selecting 30 objects from the experimental objects to create 5 high-clutter grasping scenes and perform 30 grasps with or without multi-hand interference for each scene. In addition, we use the same Grasping Success Rate (GS-Rate) as an evaluation metric from [56], which is calculated by dividing the total number of successful grasps by the total number of grasp attempts across five scenes.

The grasping results with and without hand interference are presented in Table XVI. The GS-Rate is 81.3% without hand interference and 80.0% with hand interference, and the total number of successful grasps with hand interference is only two shy of that without interference. These results indicate that the grasping performance remains nearly identical in both cases, suggesting that our method has almost no impact on the original performance of the grasping model while ensuring grasping safety.

B. Grasping with Hand Dynamic Interference

To more comprehensively validate the effectiveness of QFAAP, particularly its real-time reactive capability, we conduct additional grasping experiments under hand dynamic interference in this section. We adopt a similar experimental setting as in Section IV-F4, also selecting 30 objects from the experimental set to create 5 high-clutter grasping scenes and performing 30 grasp attempts for each scene, while randomly introducing unimanual or bimanual dynamic interference during each grasping. We reproduce the closed-loop control method from [32] and integrate it with QFAAP, endowing QFAAP with the reactive capability to counteract hand dynamic interference. Specifically, during each grasping, when the robot tends to move toward the target object, we quickly introduce hand interference by approaching the target object. After the interference, the robot will move away from the human hand and its adjacent objects (First Deviation). Once the extent of deviation becomes large, we quickly remove the hand, and the robot will resume moving toward the target object (Return). Similarly, after the extent of the move toward the target object becomes large, we again rapidly introduce hand interference and keep the hand in place until the robot deviates away from the hand and its neighboring objects (Second Deviation) and completes the safe grasping. We employ the Deviation–Return–Deviation Rate (DRD-Rate) as the evaluation metric, and a trial is considered successful if the robot completes the entire Deviation–Return–Deviation process.

The experimental results are shown in Table XVII, the DRD-Rate of QFAAP without PQGD reaches 74.7%. After integrating PQGD, the DRD-Rate dramatically increases to 84.0%, demonstrating the effectiveness of our method in hand dynamic interference scenarios, and also validates that PQGD can more obviously enhance the fast adaptability to the human hand shape in hand dynamic interference scenarios compared with hand static interference in Section IV-F3. Finally, we show two cases of the DRD process in Fig. 15, where it can be seen that the robot consistently avoids the human hand and its nearby objects with hand dynamic interference. More DRD processes are recorded in the demo videos.

REFERENCES

[1] R. M. Murray, Z. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*. Boca Raton, FL, USA: CRC Press, 2017.

[2] D. Prattichizzo and J. C. Trinkle, "Grasping," in *Springer Handbook of Robotics*, Berlin, Germany: Springer 2008.

[3] B. Kehoe, A. Matsukawa, S. Candido, J. Kuffner, and K. Goldberg, "Cloud-based robot grasping with the google object recognition engine," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2013, pp. 4263–4270.

[4] J. Mahler et al., "Learning ambidextrous robot grasping policies," *Sci. Robot.*, vol. 4, no. 26, pp. 1–12, 2019.

[5] H. S. Fang, M. Gou, C. Wang, and C. Lu, "Robust grasping across diverse sensor qualities: The GraspNet-1Billion dataset," *Int. J. Robot. Res.*, vol. 42, no. 12, pp. 1094–1103, 2023.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Conf. Neural Informat. Process. Syst.*, 2017, pp. 6000–6010.

[7] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] T. Brown, b. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and S. Agarwal, "Language models are few-shot learners," in *Proc. Conf. Neural Informat. Process. Syst.*, 2020, pp. 1877–1901.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and J. Uszkoreit, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:*2010.11929.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580-587.

[11] C. Meng, T. Zhang, and T. l. Lam, "Fast and comfortable interactive robot-to-human object handover," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 3701-3706.

[12] S. Christen, L. Feng, W. Yang, Y.-W. Chao, O. Hilliges, and J. Song, "SynH2R:Synthesizing hand-object motions for learning human-to-robot handovers," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2024, pp. 3168–3175.

[13] H. Duan, P. Wang, Y. Yang, D. Li, W. Wei, Y. Luo, and G. Deng, "Reactive Human-to-Robot Dexterous Handovers for Anthropomorphic Hand," *IEEE Trans. Robot.*, vol. 41, pp. 742 - 761, 2024.

[14] Z. Wang, J. Chen, Z. Chen, P. Xie, R. Chen, and L. Yi, "GenH2R: Learning generalizable human-to-robot handover via scalable simulation, demonstration, and imitation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16362–16372.

[15] P. Rosenberger et al., "Object-independent human-to-robot handovers using real time robotic vision," *IEEE Robot. Automat. Lett.*, vol. 6, no. 1, pp. 17–23, 2021.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations.*, 2015.

[17] T. Long, Q. Gao, L. Xu, and Z. Zhou, "A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions," *Comput. Secur.*, vol. 121, pp. 102847, 2022.

[18] J. Wang et al., "PISA: Pixel skipping-based attentional black-box adversarial attack," *Comput. Secur.*, vol. 121, pp. 102947, 2022.

[19] Z. Wang, F. Nie, H. Wang, H. Huang, and F. Wang, "Toward robust discriminative projections learning against adversarial patch attacks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 12, pp. 18784 - 18798, 2024.

[20] G. Li, Y. Xu, J. Ding, and G.-S. Xia, "Towards generic and controllable attacks against object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024.

[21] K.-H. Chow et al., "Adversarial objectness gradient attacks in real-time object detection systems," in *IEEE Int. Conf. Trust, Privacy Secur. Intell. Syst.*, 2020, pp. 263–272.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Representations.*, 2018.

[23] X. Cui, A. Aparcedo, Y. K. Jang, and S.-N. Lim, "On the robustness of large multimodal models against image adversarial attacks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 24625–24634.

[24] S. Thys, W. Van Ranst, and T. Goedeme, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops.*, 2019, pp. 49–55.

[25] Z. Hu, S. Huang, X. Zhu, F. Sun, B. Zhang, and X. Hu, "Adversarial texture for fooling person detectors in the physical world," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13307–13316.

[26] K. Xu et al., "Adversarial T-shirt! Evading person detectors in a physical world," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 665–681.

[27] Z. Hu, W. Chu, X. Zhu, H. Zhang, B. Zhang, and X. Hu, "Physically realizable natural-looking clothing textures evade person detectors via 3D modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 16975–16984.

[28] C. Rosales, J. M. Porta, and L. Ros, "Grasp optimization under specific contact constraints," *IEEE Trans. Robot.*, vol. 29, no. 3, pp. 746–757, 2013.

[29] F. T. Pokorny, K. Hang, and D. Kragic, "Grasp moduli spaces," in *Proc. Robot.: Sci. Syst.*, 2013.

[30] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vol. 34, no. 4–5, pp. 705–724, 2015.

[31] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," in *Proc. Robot.: Sci. Syst.*, 2018.

[32] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *Int. J. Robot. Res.*, vol. 39, no. 2-3, pp. 183–201, 2020.

[33] S. Yu, D.-H. Zhai, Y. Xia, H. Wu, and J. Liao, "SE-ResUNet: A novel robotic grasp detection method," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 5238–5245, 2022.

[34] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 9626–9633.

[35] M. Shan, J. Zhang, H. Zhu, C. Li, and F. Tian, "Grasp Detection Algorithm Based on CPS-ResNet," in *Proc. IEEE Int. Conf. Image Process. Comput. Vis. Mach. Learn.*, 2022, pp. 501-506.

[36] H. Cao, G. Chen, Z. Li, Q. Feng, J. Lin, and A. Knoll, "Efficient grasp detection network with Gaussian-based grasp representation for robotic manipulation,"*IEEE/ASME Trans. Mech.*, vol. 28, no. 3, pp. 1384–1394, 2022.

[37] S. Yu, D.-H. Zhai, and Y. Xia, "CGNet: Robotic grasp detection in heavily cluttered scenes," *IEEE/ASME Trans. Mech.*, vol. 28, no. 2, pp. 884–894, 2023.

[38] J. Mahler et al., "Dex-Net 1.0: A cloud-based network of 3D objects for robust grasp planning using a multi-armed bandit model with correlated rewards," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2016, pp. 1957–1964.

[39] J. Mahler et al., "Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Proc. Robot.: Sci. Syst.*, 2017.

[40] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg, "Dex-Net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 5620–5627.

[41] J. Mahler and K. Goldberg, "Learning deep policies for robot bin picking by simulating robust grasping sequences," in *Conf. Robot Learn.*, 2017, pp. 515–524.

[42] H. S. Fang, C. Wang, M. Gou, and C. Lu, "GraspNet-1billion: A large scale benchmark for general object grasping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11444–11453.

[43] H. S. Fang et al., "AnyGrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Trans. Robot.*, vol. 39, no. 5, pp. 3929–3945, 2023.

[44] C. Szegedy et al., "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Representations.*, 2014.

[45] J. Wang, "Adversarial examples in physical world," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 4925–4926.

[46] X. Liu, H. Yang, Z. Liu, L. Song, Y. Chen, and H. Li, "DPATCH: An adversarial patch attack on object detectors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 1–8.

[47] M. Lee and Z. Kolter, "On physical adversarial patches for object detection," 2019, *arXiv:* 1906.11897.

[48] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2011, pp. 3304–3311.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:* 1412.6980.

[50] M. Gruosso, N. Capece, and U. Erra, "Egocentric upper limb segmentation in unconstrained real-life scenarios," *Virtual Reality.*, vol. 27, pp. 3421–3433, 2023.

[51] C. Li, Z. Gao, and N. Y. Chong, "Shortcut-enhanced Multimodal Backdoor Attack in Vision-guided Robot Grasping," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 18629-18645, 2025.

[52] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 3511–3516.

[53] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from RGB," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 13452–13458.

[54] M. Suchi, T. Patten, and M. Vincze, "EasyLabel: A semi-automatic pixel-wise object annotation tool for creating robotic RGB-D datasets," in *Proc. IEEE Conf. Robot. Automat.*, 2019, pp. 6678–6684.

[55] S. Wang, Z. Zhou, and Z. Kan, "When transformer meets robotic grasping: Exploits context for efficient grasp detection," *IEEE Robot. Automat. Lett.*, vol. 7, no. 3, pp. 8170–8177, 2022.

[56] C. Li, N. Y. Chong, "Monozone-Centric Instance Grasping Policy in Large-Scale Dense Clutter," *IEEE/ASME Trans. Mech.*, Early Access, 2025.