Fed-HANet: Federated Visual Grasping Learning for Human Robot Handovers

Ching-I Huang, Yu-Yen Huang, Jie-Xin Liu, Yu-Ting Ko, Hsueh-Cheng Wang, *Member, IEEE*, Kuang-Hsing Chiang, and Lap-Fai Yu

***Abstract*—Human-robot handover is a key capability of service robots, such as those used to perform routine logistical tasks for healthcare workers. Recent algorithms have achieved tremendous advances in object-agnostic end-to-end planar grasping with up to six degrees of freedom (DoF); however, compiling the requisite datasets is simply not feasible in many situations and many users consider the use of camera feeds invasive. This letter presents an end-to-end control system for the visual grasping of unseen objects with 6-DoF without infringing on the privacy or personal space of human counterparts. In experiments, the proposed Fed-HANet system trained using the federated learning framework achieved accuracy close to that of centralized non-privacy-preserving systems, while outperforming baseline methods that rely on fine-tuning. We also explores the use of a depth-only method and compares its performance to a state-of-the-art method, but ultimately emphasizes the importance of using RGB inputs for better grasp success. The practical applicability of the proposed system in a robotic system was assessed in a user study involving 12 participants. The dataset for training and all pretrained models are available at https://arg-nctu.github.io/projects/fed-hanet.html.**

***Index Terms*—Federated learning, human-robot interaction, service robots.**

Manuscript received 12 October 2022; accepted 3 April 2023. Date of publication 26 April 2023; date of current version 10 May 2023. This letter was recommended for publication by Associate Editor Y. Hirata and Editor H. Moon upon evaluation of the reviewers' comments. This work was supported in part by Taiwan's National Science and Technology Council under Grants 110-2221-E-A49-027-MY2, 111-NU-E-A49-001-NU, 111-2623-E-A49-007, and 112-2321-B-A49-005, in part by Qualcomm through Taiwan University Research Collaboration Project, in part by the National Yang Ming Chiao Tung University and Ministry of Education (MOE), Taiwan through Higher Education Sprout Project, and Lap-Fai Yu is supported in part by the National Science Foundation under Grant 1942531. *(Corresponding author: Hsueh-Cheng Wang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by NYCU Institutional Review Boards Application No. NYCU-REC-110-097E.

Ching-I Huang, Yu-Yen Huang, Jie-Xin Liu, Yu-Ting Ko, and Hsueh-Cheng Wang are with the National Yang Ming Chiao Tung University (NYCU), Hsinchu 30010, Taiwan (e-mail: cihuang@nctu.edu.tw; sdnd93612@gmail.com; phoebeliu1006@gmail.com; yutingk.ee11@nycu.edu.tw; hchengwang@csail.mit.edu).

Kuang-Hsing Chiang is with the Taipei Heart Institute, Taipei Medical University, Taipei 11031, Taiwan, also with the Division of Cardiology and Cardiovascular Research Center, Taipei Medical University Hospital, Taipei 110, Taiwan, also with the Department of Internal Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei 11031, Taiwan, and also with the Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 106216, Taiwan (e-mail: steve.cn@tmu.edu.tw).

Lap-Fai Yu is with the Department of Computer Science, George Mason University, Fairfax, VA 22030 USA (e-mail: craigyu@gmu.edu).

Digital Object Identifier 10.1109/LRA.2023.3270745

I. Introduction

HUMAN-ROBOT handover, referring to the transfer of an object between a human and robot, is a fundamental capability for service robots. This issue was initially addressed by having the robot remain in a stationary position, while the human places the object in the robot's gripper [1], [2], [3]. Since that time, researchers have addressed the handover of objects of various shapes and sizes using grippers of various designs [4], [5]. Researchers have also made an effort to consider the perceptions of humans involved in these interactions [4], [5]. Some researchers have sought to classify human grasp modes or hand pose estimates to generate a corresponding robot grasp pose [5], [6]. Note however that those methods require *hard-coded* associations between hand and object poses, which are impractical in most situations.

One common approach to human-to-robot handover involves hand/object detection in conjunction with algorithms for pose estimation and grasp point prediction [6], [7]. Yang et al. trained a deep neural network using PointNet++ [8] to classify point clouds around human hands into one of seven pre-defined grasp categories, each of which would trigger a specific motion plan to complete the handover [7]. Rosenberger et al. [9] expanded this work to include *object-independent* handover actions for a wide range of objects based on a YOLO-v3 object detector [10]. Their system simultaneously predicts hand and body segmentation effects to ensure that the grasping action of the robot is performed safely. Saputra et al. [11] focused on the real-time affordance detection of gripping pose using vision and depth sensors. The method also used YOLO-v3 object detector [10] and generates a topological map of the desired object with an inlier-outlier method to compute the possible gripping position. In [6], the authors addressed the challenges in handling unseen objects (occluded by the hands) using a grasp selection model based on the 6-DoF GraspNet system [12]. Success in human-to-robot handovers depends on the tracking of objects and hands; however, any action involving a human agent will introduce occlusions, which can affect the effectiveness of segmentation and pose estimation.

Recently, researchers have had considerable success in developing end-to-end algorithms for visual grasping. The objective behind these data-driven methods is the training of object-agnostic grasping policies based on learned visual features without the need for *a priori* object-specific knowledge. One project featured in the Amazon Robotics Competition used an

2377-3766 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: George Mason University. Downloaded on June 30,2023 at 02:06:35 UTC from IEEE Xplore. Restrictions apply.

Fig. 1. Our approach allows human robot handovers without privacy concerns, which is well-suited for service robots.

TABLE I
COMPARISONS OF HANDOVER AND VISUAL GRASPING METHODS

Literature (et al.)	6-DoF Grasping	Object Independent	Dynamic Objects	Application to H2R handover	Open Dataset or Software	Data Privacy Considerations
Rosenberger [9]	✓	✓	✗	✓	△	-
Yang [6]	✓	✓	✗	✓	✗	-
Yang [7]	✓	✗	✓	✓	✗	-
Cini [5]	✓	✗	✓	✓	✗	-
Nemlekar [23]	△	✗	✗	✓	✗	-
Costanzo [3]	✓	✗	✓	✓	✗	-
GGCNN [24]	✗	✓	✓	-	✓	-
ConvNet [13]	✗	✓	✗	-	✓	-
Song [17]	✓	✓	✓	-	✓	-
Ours	✓	✓	✓	✓	✓	✓

△ : *partial – including*

end-to-end network to generate affordance maps for suction devices or two-finger parallel grippers in heavily cluttered scenes [13]. DexNet models [12] use large simulation-only datasets pertaining to thousands of objects to train end-to-end CNN networks. Some visual grasping algorithms focus on planar grasping [13], [14], wherein the grasping point (x, y) and the θ of a two-finger gripper are generated from an input image. Note however that human-to-robot handovers are performed within a 3D space, and therefore require grasping actions with six DoFs.

A number of studies have addressed the problem of grasping unknown objects using movements with 6-DoFs [15], [16], [17], [18]. In [17], the authors presented an overview of crucial issues related to the collection of training data for learning-based algorithms. It is possible to increase the size of a dataset by including datapoints from simulations; however, that approach is unable to deal with many thorny issues associated with visual perception. For example, most previous studies conducted in a lab setting use a plain background, which is not generalizable to other environments [19]. Overall, collecting a sufficient amount of training data and annotations in a variety of environments is infeasible in most practical situations, particularly in light of the fact that the robots must operate in close proximity to humans, such that robot cameras could be viewed as an intrusion on the privacy of the human user.

Federated learning (FL) is one approach to safeguarding the personal space and privacy of human participants [20]. This involves exchanging model weights among distributed clients, rather than collecting a large amount of imagery data directly, thereby making it possible to train models to grasp a variety of objects in a variety of situations. FL has been used in the analysis of medical images [21] and domain transfer for semantic segmentation in autonomous driving [22]; however, it has not previously been used for visual grasping or handover tasks. This letter makes the following contributions:

- *Federated training scheme for handover visual grasping (Fed-HANet).* This is the first ever study to use federated learning to facilitate the training of a handover system for service robot applications, as demonstrated in Fig. 1.
- *Practical evaluation of handover system.* The practicality of Fed-HANet was assessed in a user study involving 12 participants.
- *Open access toolkit applicable to federated learning for human-robot handover operations.* We collected and manually labeled a dataset (with subsets) for handover operations. The dataset and pretrained weights are available as open access resources for download.

II. RELATED WORK

Handovers tasks have been addressed in the context of human-robot interactions, computer vision, and visual grasping by robots. As shown in Table I, **6-DoF grasping** is preferable to planar grasping for handover tasks, **closed-loop control** is well-suited to dynamic situations, and **object-independence** makes it possible to handle a variety of objects.

In this section, we examine a number of visual grasping algorithms of relevance to the proposed scheme. Readers are referred to [5], [9] for other aspects of human-robot handovers.

The adoption of deep learning techniques has greatly advanced research into visual grasping. Most previous research has focused on object-agnostic grasping. Data-driven methods avoid dependence on specific objects and many studies have compiled real-world datasets for training. One pioneering study [14] used a self-supervised robot to collect a dataset related to 50 k instances of grasping and trained a deep neural network classifier to predict the success of grasp attempts. In [13], researchers presented a category agnostic algorithm to map RGB-D images to a pixel-level probability affordance map. In [25] the authors sought to compile a more diverse dataset by including images captured in homes instead of a lab setting. In [17], a low-cost hardware interface was used to collect instances of human grasping in diverse environments for the training of a 6-DoF closed-loop algorithm via reinforcement learning.

Another approach to augmenting datasets involves simulations. DexNet [12], [26], [27] is an end-to-end robotic manipulator that includes code and algorithms for use in generating datasets of synthetic point clouds based on a dataset of 6.7 million datapoints generated entirely through simulation. Note however that datasets generated entirely via simulations make it difficult to incorporate data related to human intentions.

In [13], the authors implemented an end-to-end affordance prediction method, which is similar to an affordance map in

Authorized licensed use limited to: George Mason University. Downloaded on June 30,2023 at 02:06:35 UTC from IEEE Xplore. Restrictions apply.

that it evaluates the probability of success in grasping an object based on each pixel in an image rather than the object itself. This eliminates the need to detect or identify the object and then determine the optimal grasping point for that object. Note that the purpose of our study was not to introduce any new model architectures in the field of logistic pick-and-place tasks. Instead, the main contribution of this work lies in the development of a federated learning-based training scheme for visual grasping. As described in Section III, this iterative training process makes it possible to continuously improve the handover affordance map, which in turn makes it possible to formulate an end-to-end control system for service robots without impinging on the privacy of human users.

III. Federated Training Scheme for Handover Visual Grasping

Federated learning makes it possible to train a shared prediction model collaboratively at each local client without the need for direct access to client data from a server.

Our work addresses the following research questions:

- How does the performance of models trained using the federated learning scheme compare with those trained using other methods?
- How does the proposed method compare with existing methods when evaluated using a variety of datasets?

A. Federated Training Scheme

This study employed the open-source federated learning framework, (Flower [28]), which has been widely adopted in the research community. Flower employed binary serialization format Remote Procedure Call (gRPC) streams and numerous cross-platform clients to enhance efficiency and scalability. We prepared three workstations (i7-7700 CPU and RTX 2060 GPU) for the respective hosting of training data subsets as clients. A server (i7-8750H CPU and RTX 2070 GPU) was used to aggregate model weights using the FedAvg algorithm [20]. The Adam optimizer was used to train the model for each client, and binary cross-entropy was used as the loss function (BCEWithLogitsLoss in PyTorch) with a batch size of 5, a fixed learning rate of 10^{-3}, and momentum of 0.99. The model was trained in PyTorch running on a PC with an Intel Core i5-9400F and NVIDIA RTX2080. Training was completed in roughly one hour.

We designated N-epoch individual training runs for each client, after which the trained model weights were transmitted to a server within a specified time frame. Upon completion of the aggregation process, the weights were returned to each individual client. Note that each weight aggregation cycle is referred to as a *round*, and the server proceeds to the next round only after the weights from all three clients (i.e., under all three conditions) have been aggregated. Our network was initially trained for 50 epochs without the federated learning pipeline. To assess the impact of the number of rounds and epochs per round on performance, we trained the network model using the federated learning pipeline under two different conditions:

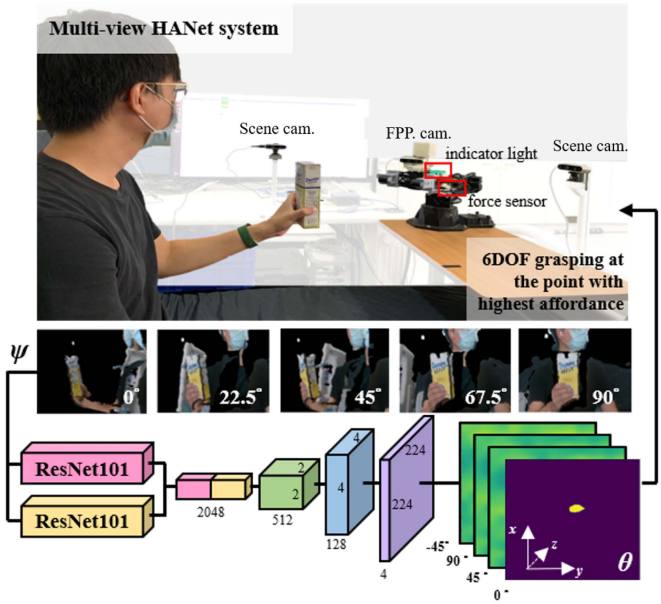


Fig. 2. Proposed HANet produced a 6-DoF prediction of orientations θ and the directions of grasping attempt ψ from multiple views.

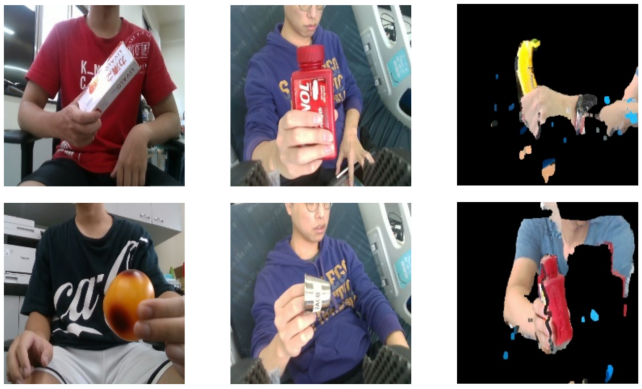
- *Fed-HANet (5E):* With 5 epochs per round and 10 rounds of training.
- *Fed-HANet (10E):* With 10 epochs per round and 5 rounds of training.

B. Network Architecture

In the current study, we developed an affordance prediction method using an architecture similar to that of ConvNet [13]. The proposed model architecture comprises two ResNet-101 networks, the respective inputs of which are RGB images and corresponding depth data captured by the same RGB-D camera (see Fig. 2). Captured depth data undergoes pre-processing, such that all depths exceeding 75 cm are assigned the same value (i.e., a flat surface) to reduce noise and thereby enhance prediction accuracy. We concatenate the two ResNet-101 outputs (RGB and depth), followed by three additional spatial convolution layers to merge the corresponding features. We then spatially up sample the outputs bilinearly and apply a softmax function to output two pixel-wise layers (non-graspable and graspable) to represent the inferred affordances. The fully convolutional networks (FCNs) then output four affordance maps. Note that our network is a modification of the system proposed in [13] with several fundamental differences. Our aim in designing the proposed Fed-HANet architecture was to enhance the efficiency of multi-view inference by implementing specific modifications, as shown in Fig. 2.

- *Orientation for First Person Perspective (FPP).* ConvNet [13] uses multiple inferences to estimate the optimal orientation for planar grasping; however, this process is slow. The proposed model is able to generate four affordance maps representing four candidate grasp orientations θ $(0°, 45°, 90°, and -45°)$. using as an input a single RGB-D image from a camera mounted on the robot arm.

Authorized licensed use limited to: George Mason University. Downloaded on June 30,2023 at 02:06:35 UTC from IEEE Xplore. Restrictions apply.



(a) HA-Office (b) HA-MedBed (c) HA-MultiView

Fig. 3. Training datasets with different scenarios.

The affordance maps describe the probability of success in grasping an object with respect to each pixel in an image. The point on the image with the highest affordance score corresponds to the optimal location at which to grasp the object. This location was set as the target point for the effector on the end of the robot arm. The affordance maps and depth information are used to generate a 6-DoF pose for the pre-grasp position (set 4 cm from the predicted handover point) based on the length of the robot fingers.

- *Multi-view (MV).* We sought to resolve the bin-picking problem by adopting the ConvNet approach [13], which involves utilizing multiple grasp primitives with interchangeable end-effectors (e.g., two-finger or vacuum grippers), based on a single top-down view. The robot arm in this study was equipped with a two-finger gripper rather than a vacuum gripper. The size of the gap between the fingers limited the size of object that could be gripped; however even large objects could potentially be picked up if they were gripped along their narrowest side. Our use of two cameras to capture the surrounding environment in conjunction with the camera on the robot arm provided multiple views of the object, thereby making it possible to identify the narrow side of the object from which it should be grasped. We also synthesized two additional scenes (at ψ 22.5° and 67.5° relative to the object) using point clouds obtained from the three cameras at 0° (overall scene), 45° (robot arm), and 90° (overall scene) relative to the object. This involves calibrating and then fusing raw RGB-D streams from the two scene cameras to form a dense point cloud. Note that specific pre-defined views were selected for the synthesis of 2D images as model inputs. This strategy allowed the camera on the robot arm to provide an optimal viewpoint without the need to physically move the camera into position.

C. Handover Datasets

1) Training Datasets: We compiled a series of handover datasets, with a diversity of backgrounds, objects, and viewing angles, to enable federated learning. As shown in Fig. 3 and Table II, the proposed handover dataset included three subsets: (1) *HA-Office*, (2) *HA-MedBed*, and (3) *HA-MultiView*. *HA-Office* and *HA-MedBed* comprised RGB-D images (640 × 480) of an object held in either the right or left hand, as recorded using an Intel RealSense D435 RGBD camera. For the sake of reproducibility, most of the objects used in this study were adopted from YCB object set [29]. Nonetheless, we also included several additional objects, including household objects and labeled bottles applicable to nursing. The imaging data were collected in a variety of scenes with ambient light from a variety of angles.

TABLE II
SUMMARY OF TRAINING AND TESTING DATASETS

	Source	N of Img.	Avg. Obj.	Hand
Training		Total: 1,075		
HA-Office	Ours	584	1	Y
HA-MedBed	Ours	174	1	Y
HA-MultiView	Ours	317	1	Y
Testing				
YCB-Obj	[29]	92	4.4	N
ARC-Cluttered	[13]	77	3.4	N
HA-Upright	Ours	200	1	Y
HA-Rotated	Ours	240	1	Y

We compiled a series of handover datasets featuring a diversity of backgrounds, objects, and view angles for use in federated learning. As shown in Fig. 3 and Table II, the proposed handover dataset included three subsets: (1) *HA-Office*, (2) *HA-MedBed*, and (3) *HA-MultiView*. *HA-Office* and *HA-MedBed* comprised RGB-D images (640 × 480) of an object held in either the right or left hand, as recorded using an Intel RealSense D435 RGBD camera. For the sake of reproducibility, most of the objects used in this study were adopted from the YCB object set [29]. Nonetheless, we also included several additional objects, including household objects and labeled bottles applicable to nursing. The imaging data were collected in a variety of scenes with ambient light from a variety of angles.

HA-MultiView comprised images captured in a multi-camera setup, in which the 3D point cloud from cameras 1 and 3 were re-projected from arbitrary perspectives, similar to the height map in [13]. Note that some of the re-projected RGB-D images in *HA-MultiView*contained null values, which were filled in with black pixels. Fig. 3(c) presents sample images from the datasets. In preparing the training data, we captured images of objects held in a human hand and labeled the graspable regions of the object with their orientation and direction. This made it possible to train the model to simultaneously predict the direction ψ (i.e., the direction the robot arm relative to the user as it approaches the object) and orientation θ (i.e., the angle of the gripper relative to the object as it closes the gripper assembly) within a simulation space with 6-DoFs.

All images were manually labeled using the open-source annotation tool, LabelMe [30]. Note that for each image, we labelled several line segments indicating the location and orientation that would allow a two-finger gripper to grasp the object without touching the human hand. The output is a densely labeled pixel-wise map (640 × 480 px) with each pixel value normalized to between 0 and 1 in the form of a heat map. The dataset included a total of 1,075 RGB-D images with annotations for a total of 8,734 possible grasp configurations.

Authorized licensed use limited to: George Mason University. Downloaded on June 30,2023 at 02:06:35 UTC from IEEE Xplore. Restrictions apply.

2) Testing Datasets: We included *HA-Rotated* and additional 3 test datasets for the assessments. Two of which are open-source datasets pertaining to object grasping, while the other two were collected in this study Assessments were performed using *HA-Rotated* and three additional test datasets. Note that two datasets (YCB-Obj and ARC-Cluttered) were open-source, while the other two (HA-Upright and HA-Rotated) were compiled in the current study (see Table II). Sample images of those datasets can be found in the supplementary materials.

- *YCB-Obj:* The 92 images in this dataset were extracted from the first frame of the videos provided in [31]. This dataset features 21 objects placed on a table. The original video did not include ground truth examples of grasping affordance; therefore, we manually verified the performance of each method.
- *ARC-Cluttered:* The 77 images in this dataset were used in the 2017 Amazon Robotic Challenge (ARC) [9]. This dataset features several objects in cluttered scenes (an average of 3.4 objects per scene). The test dataset includes human-annotated two-finger grasping labels.
- *HA-Upright:* These 200 images of a handover scenario respectively featured one object held upright in the hand. We selected five of the objects in [9] based on whether they would fit in the gripper of the robot used in this study.
- *HA-Rotated:* We collected an additional 240 images (including 10 YCB objects), none of which appeared in our training dataset. We were confident that all of the objects could be grasped by our robot, as they were smaller than 6 cm along at least one side. The dataset included images of objects that were occluded to various degrees by the human hand in which they were held. Each object was held by a person lying on a medical bed, while the camera was held at various orientations ($[0, 45, 90, -45]$, unit: deg.) with various degrees of occlusion $[>40\%, <40\%]$ and viewing ranges ($[10, 20, 30, 40]$, unit: cm).

D. Training Scheme Evaluation

1) Effectiveness of Federated Learning: We compared the performance of models trained using federated learning with those of models trained using non-federated methods, as shown in Fig. 4. When applied to the *HA-Rotated* test set, the accuracy of *Fed-HANet (5E)* and *Fed-HANet (10E)* approached that of the non-privacy-preserving (centralized) method after several training rounds. The accuracy of *Fed-HANet (5E)* converged more quickly than did *Fed-HANet (10E)*, due perhaps to the overfitting of *Fed-HANet (10E)* to local datasets. These results demonstrate that federated learning can achieve a high degree of accuracy without the need to share raw data pertaining to each client.

We trained a non-privacy-preserving *HANet* model using a dataset comprising three sub-datasets collected under three conditions. The first subset was trained from scratch, while the second and third subsets were trained using pre-trained weights, resulting in a total of six training sequences. The accuracy of the *fine-tuning* sequences ranged from 22.92% (*Finetune-2-1-3:* subset 2, 1, and 3) to 83.33% (*Finetune-2-3-1:* subset 2, 3, and 1). The performance of the model was significantly affected by the order in which fine-tuning was applied to the training subsets. Nonetheless, *Fed-HANet* significantly outperformed the fine-tuning method.

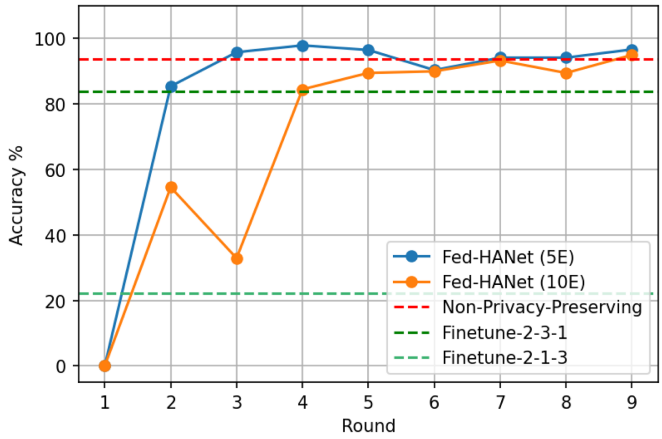


Fig. 4. Training scheme evaluations.

2) Depth-Only Inputs: We employed a baseline model called Fed-HANet-Depth, trained solely using depth images, to investigate the efficacy of using depth-only sensors in medical applications for remote service robots. The success rate of grasp attempts varied depending on the test dataset (*HA-Rotated*), with Fed-HANet-Depth achieving lower success rates (63%) compared to Fed-HANet (94%), which used RGB inputs in addition to depth. These results emphasize the importance of using RGB inputs for better accuracy and grasp success. Additionally, in the Supplementary Materials, we provided inference results from handover examples executed using FedHANet, FedHANet-Depth, and state-of-the-art Grasp Quality Convolutional Neural Networks (GQ-CNNs), demonstrating that the use of depth-only inputs led to erroneous target point predictions, causing the robot to grasp the human's arm or fingers.

E. Comparisons to Existing Methods

The performance of the proposed system was compared with that of existing systems in terms of grasp success/failure. We also sought to elucidate the reasons for the failures. Each method was tasked with identifying the optimal grasp point (within a 3D simulation space), defined as the point at which the object could be stably grasped without touching the fingers of the human user.

Three existing methods were adopted as baselines for comparison:

- *DOPE* [31]: This state-of-the-art object pose estimation approach was used as a baseline. *DOPE* is constrained by the need for pre-trained models of known objects. In the current study, the *DOPE* baseline was combined with pre-defined poses related to the grasping of known YCB objects. As shown in Table III, *DOPE* was effective in estimating the pose of objects in *YCB-Obj*; however, it was not nearly as effective when applied to *HA-Upright* or *HA-Rotated*, due to interference caused by hand occlusions. Nonetheless, *DOPE* performed satisfactorily when

Authorized licensed use limited to: George Mason University. Downloaded on June 30,2023 at 02:06:35 UTC from IEEE Xplore. Restrictions apply.

TABLE III
ESTIMATION OF GRASP POINT OUTCOMES VIA DIFFERENT APPROACH BY HANET TESTING DATASET

	DOPE	Ro. [9]	ConvNet	Fed-HANet
YCB-Obj	56.5%	27.1%	21.7%	56.5%
ARC-Cluttered	-	38.9%	98.7%	77.8%
HA-Upright	47.5%	80.5%	0.0%	89.0%
HA-Rotated	37.5%	39.1%	0.8%	97.9%

the fingers were placed on the sides of the object. Note that *DOPE* was of limited benefit when applied to the *ARC-Cluttered* database containing unknown objects.

- *ConvNet* [13]: Note that this is the unmodified framework on which the network in this study was based. We employed an FCN network architecture with a ResNet-101 backbone with weights pre-trained by ImageNet. *ConvNet* with pre-trained weights performed well when applied to *ARC-Cluttered* (98.7%); however, it often led to contact with human hands when applied to *HA-Upright* and *HA-Rotated*, thereby indicating the need to fine-tune the dataset specifically for handover scenarios. *ConvNet* performed poorly when applied to *YCB-Obj*, indicating that it is limited to planar grasping and scenarios with a uniform background.
- *Ro.* [9]: This state-of-the-art method (considered a strong baseline) integrates multiple modules to process input images for object detection, hand and body segmentation, and grasp selection for use in identifying the pixels that are most likely to coincide with the optimal grasp point. This method performed well when applied to *HA-Upright* (80.5%); however, it was not nearly as effective when applied to *HA-Rotated*, which is ill-suited to objects held in arbitrary orientations. It also proved ineffective when applied to object-only datasets, due to its use of the YOLO-v3 object detector [10], which is prone to false positives when dealing with a complex background or situations involving occlusion.

The proposed *Fed-HANet* system consistently outperformed all baseline methods when applied to the *YCB-Obj*, *HA-Upright*, or *HA-Rotated* test datasets. These results demonstrate the efficacy of the proposed method in preventing contact with human fingers when handing over objects. These results also demonstrate that an end-to-end method can be more reliable than discrete predictions pertaining to the position of hands and objects in situations involving occlusion. Note however that the accuracy dropped to below that of *ConvNet* (68.8%) when applied to the ARC-Cluttered dataset, which may be the result of trade-offs during model optimization for handover scenarios.

IV. HANDOVER USER STUDY

The practicality of the proposed Fed-HANet was evaluated in a robot-human handover scenario involving 12 participants (ages 20-30) with no prior experience using such systems. The objects were selected from the YCB object set. We also evaluated multi-view performance by including six box-shaped objects that were graspable from only one side. This user study was conducted with the approval of the NYCU (NYCU-REC-110-097E)

A. System Overview

The system proposed in the current study included a low-cost manipulator (Trossen Robotics ViperX 300) equipped with a first-person perspective (FPP) RGB-D camera (Intel RealSense D435) in conjunction with two RGB-D *scene* cameras, respectively facing the front and left sides of the user (see Fig. 2). To ensure safety in this study, a force sensor (Robotiq FT300) was installed at the last joint of the arm behind the gripper. This provides a fail-safe mechanism in case of issues with recognition or user movement, allowing for prompt detection and intervention to prevent potential harm. All hardware modules were connected to a computer (Intel NUC) via USB cables to enable the processing of three image streams as well as tactile force feedback and control commands. An additional GPU-equipped computer was connected via intranet to genrate Fed-HANet predictions. We equipped the robot with indicators (green, orange, and red lights) to make the user aware of the current state of the robot. When the robot arm was ready to move, the indicator would turn green. As the robot arm approached the user (i.e., while moving), the indicator would turn red.

B. Methods

We selected the Fed-HANet model for this assessment, due to its effectiveness (exceeding that of all other methods) in avoiding human contact (fingers or arms) during handover scenarios. Trials were run in *FPP* and *MV* modes. The objective in all trials was to determine direction ψ and orientation θ. We defined five candidate directions, including 0° (perpendicular to the user), 22.5°, 45°, and 67.5° to 90° (facing the user). We also defined four candidate orientations (0°, 45°, 90°, and 135°). Fed-HANet inferred the predicted grasp pose based on depth data and then computed the desired trajectory by an inverse kinematic (IK) module in conjunction with the MoveIt [32] path planning node. We implemented a state machine using the ROS Smach package to control arm movements, as illustrated in the Supplementary Materials.

- *First Person Perspective (FPP) mode:* We implemented a closed-loop design to handle dynamic situations in which the hands of our user are prone to movement during the approach stage. The average inference time per frame was 0.031 seconds, which was fast enough to enable operations at 10 frames per second (fps). The use of only one camera (attached to the robot arm) necessitated moving the robot arm to five positions from which to capture images corresponding to the five candidate directions.
- *Multi-View (MV) mode:* We utilized the two scene cameras calibrated to merge the point clouds. Unlike FPP mode, Fed-HANet inference operations were performed for the five views without moving the robot arm. Inference operations were completed in roughly 0.16 seconds.

In *MV* mode, the need to compute multiple inferences from multiple views greatly increased the computation time. Note also that the design of the joints used in this low-cost robot arm occasionally skewed the trajectories during closed-loop processing. We sought to mitigate this issue by using only open-loop processing in both *FPP* and *MV* modes, based on

Authorized licensed use limited to: George Mason University. Downloaded on June 30,2023 at 02:06:35 UTC from IEEE Xplore. Restrictions apply.

the assumption that the user would keep the object stationary after the Fed-HANet inference phase.

C. Human-to-Robot Handover Workflow

The participants were tasked with holding objects in arbitrary directions (ψ) and orientations (θ) to be grasped by the robot. The robot arm was programmed to take the object only if the participant's hand was still, similar to the user studies in [6], [7], [9]. The participants were allowed to hold the objects by pinching them using two fingers, grasping them between multiple fingers, or balancing them on an open palm. Each of the participants performed the handover task twice for each of the 16 objects in *FPP* and *MV* modes, which resulted in 768 trials.

The object should be reachable as long as it is within 50 to 60 cm from the initial 'home' arm position. The reachable grasping points are visualized in the Supplementary Materials. When the end effector reached a point 4 cm in front of the target handover point, the effector was slowed down for its approach toward the target. This slowing-down movement was intended to reduce the anxiety felt by the individual interacting with a robot arm. In some situations, where depth information was missing due to light reflective or light absorptive materials or situations where the illumination was less than ideal, the system attempted to move the robot arm 2 cm toward the object. After the end effector reached the target point and the gripper closed, the unit returned to its home position.

To mimic the behavior typical of patients on a hospital bed, we opted not to require that participants hold the objects within a certain zone. In situations where the target point fell outside the range of the arm robot, the arm was moved to the point closest to the predicted pose. At this point, the system waited for the user to place the object between the two fingers of the end effector. Note that during the *handover*, the indicator turned orange to inform the user that the object must be placed in any area between the two fingers, excluding the fingertips. We set a threshold value for the force torque sensor, which if exceeded, would trigger the gripper to close and the arm to return to the home position. This so-called *hand-in* mode is discussed separately below.

D. Human-to-Robot Handover Results

1) Metrics: The system was evaluated in terms of effectiveness (mean success rate in completing the task) and efficiency (mean time to completion). After each round, the participants also filled out a 7-point Likert questionnaire evaluating the system. The participants were asked to choose the number that best represented their level of agreement or disagreement with each aspect of the system.

2) MV vs.FPP Approaches: In terms of completion time, the proposed *MV* method (mean: 24.8s; std: 2.1s) outperformed the baseline *FPP* method (mean: 48.2s; std: 3.1s). Our results revealed that the multiple view approach made it easier to identify the graspable side of the objects, particularly when dealing with the six objects that were graspable from only one side. The handover success rate was as follows: Baseline *FPP* method (96.5%; std: 7.2%) and *MV* method (mean: 86.9%; std: 8.4%). Note that the difference did not meet the level of significance.

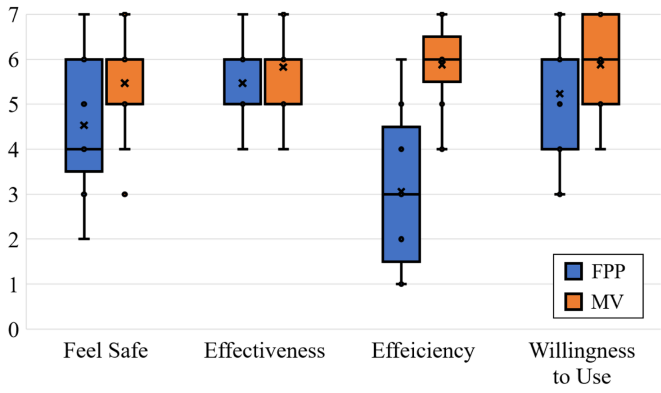


Fig. 5. Questionnaire analysis. The 7 points on the scale range from 1 (strongly disagree; negative feedback) to 7 (strongly agree; positive feedback).

Note also that the robot arm did not pinch the participants' fingers in any of the trials.

3) Exceptional Cases – Hand-In Mode: In some trials, the participant held the object too far from the robot arm, such that the task had to be completed in *hand-in* mode, which involved the robot arm moving to a point close to the user and waiting for the user to place the object between the two gripper fingers, the force of which triggered the gripper to close. The incidence of *hand-in* mode was similar between the two methods: *FPP* (29.9%) and *MV* (25.7%). The success rate in *hand-in* mode was also similar: *FPP* (96.3%) and *MV* (98.3%).

E. Questionnaire Analysis

The questionnaire was meant to obtain subjective feedback pertaining to the performance of the proposed system. The results are summarized in Fig 5.

- *Q1: Feeling of Safety.* The participants were asked if they were worried about being injured by the robot. The system received high ratings for safety, regardless of the operating mode. We presume that this favorable assessment can be attributed to the accuracy of the model and speed at which the robot arm moved while in the vicinity of the participants.
- *Q2: Reactivity.* TThe participants rated the system on its ability to detect and react appropriately to their movements, which presumably affected their confidence in the capability of the robot arm. High ratings were obtained for both modes.
- *Q3: Efficiency.* Participants rated the efficiency of the system in terms of their satisfaction with the speed of the robot arm. The responses were less positive here (close to 1). In fact, the participants reported feeling impatient with the the entire operation. Overall, the *MV* method outperformed *FPP*.
- *Q4: Willingness to use.* Participants rated their willingness to use the proposed handover system in a real-world setting. Both methods received high ratings.

Authorized licensed use limited to: George Mason University. Downloaded on June 30,2023 at 02:06:35 UTC from IEEE Xplore. Restrictions apply.

V. Conclusion

This study adopted federated learning to tackle privacy concerns in amassing training databases for service robot applications. To the best of our knowledge, this is the first application of federated learning in tasks that involve visual grasping. The effectiveness of the proposed system was compared with non-federated baselines, depth-only methods, and state-of-the-art methods. The practicality of the proposed system was assessed in a user study using a low-cost robot arm with inexperienced users. This method holds potential for use in various human-robot collaboration scenarios.

Acknowledgment

The authors would like to thank Po-Jui Huang for providing assistance with the baseline evaluations.

References

[1] A. Edsinger and C. C. Kemp, "Human-robot interaction for cooperative manipulation: Handing objects to one another," in *Proc. 16th IEEE Int. Symp. Robot Hum. Interactive Commun.*, 2007, pp. 1167–1172.

[2] J. Aleotti, V. Micelli, and S. Caselli, "Comfortable robot to human object hand-over," in *Proc. 21st IEEE Int. Symp. Robot Hum. Interactive Commun.*, 2012, pp. 771–776.

[3] M. Costanzo, G. De Maria, and C. Natale, "Handover control for human-robot and robot-robot collaboration," *Front. Robot. AI*, vol. 8, 2021, Art. no. 672995.

[4] C.-M. Huang, M. Cakmak, and B. Mutlu, "Adaptive coordination strategies for human-robot handovers," *Robot.: Sci. Syst.*, vol. 11, pp. 1–10, 2015.

[5] F. Cini, V. Ortenzi, P. Corke, and M. Controzzi, "On the choice of grasp type and location when handing over an object," *Sci. Robot.*, vol. 4, no. 27, 2019, Art. no. eaau9757.

[6] W. Yang, C. Paxton, A. Mousavian, Y.-W. Chao, M. Cakmak, and D. Fox, "Reactive human-to-robot handovers of arbitrary objects," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 3118–3124.

[7] W. Yang, C. Paxton, M. Cakmak, and D. Fox, "Human grasp classification for reactive human-to-robot handovers," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 11123–11130.

[8] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5105–5114.

[9] 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, Jan. 2021.

[10] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," in *Computer Vision and Pattern Recognition*. vol. 1804. Berlin/Heidelberg, Germany: Springer, 2018, p. 1–6.

[11] A. A. Saputra, C. W. Hong, and N. Kubota, "Real-time grasp affordance detection of unknown object for robot-human interaction," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 2019, pp. 3093–3098.

[12] J. Mahler et al., "Learning ambidextrous robot grasping policies," *Sci. Robot.*, vol. 4, no. 26, 2019, Art. no. eaau4984.

[13] A. Zeng et al., "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," *Int. J. Robot. Res.*, vol. 4, pp. 690–705, 2019. [Online]. Available: https://doi.org/10.1177/0278364919868017

[14] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2016, pp. 3406–3413.

[15] A. Mousavian, C. Eppner, and D. Fox, "6-DoF GraspNet: Variational grasp generation for object manipulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2901–2910.

[16] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox, "6-DoF grasping for target-driven object manipulation in clutter," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 6232–6238.

[17] S. Song, A. Zeng, J. Lee, and T. Funkhouser, "Grasping in the wild: Learning 6DoF closed-loop grasping from low-cost demonstrations," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4978–4985, Jul. 2020.

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

[19] R. Julian, B. Swanson, G. S. Sukhatme, S. Levine, C. Finn, and K. Hausman, "Never stop learning: The effectiveness of fine-tuning in robotic reinforcement learning," in *Proc. Conf. Robot Learn.*, 2020, pp. 2120–2136.

[20] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[21] L. Li, Y. Fan, M. Tse, and K.-Y. Lin, "A review of applications in federated learning," *Comput. Ind. Eng.*, vol. 149, 2020, Art. no. 106854.

[22] L. Fantauzzo et al., "Feddrive: Generalizing federated learning to semantic segmentation in autonomous driving," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 11504–11511.

[23] H. Nemlekar, D. Dutia, and Z. Li, "Object transfer point estimation for fluent human-robot handovers," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 2627–2633.

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

[25] A. Gupta, A. Murali, D. P. Gandhi, and L. Pinto, "Robot learning in homes: Improving generalization and reducing dataset bias," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9112–9122.

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

[27] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg, "Dex-Net 3.0: Computing robust robot suction grasp targets in point clouds using a new analytic model and deep learning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 5620–5627.

[28] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, and N. D. Lane, "Flower: A friendly federated learning research framework," 2020, *arXiv:2007.14390*.

[29] B. Çalli, A. Singh, A. Walsman, S.S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB object and model set: Towards common benchmarks for manipulation research," in *Proc. Int. Conf. Adv. Robot.*, 2015, pp. 510–517.

[30] K. Wada, "LABELME: Image polygonal annotation with python," 2016. [Online]. Available: https://github.com/wkentaro/labelme

[31] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," in *Proc. Conf. Robot Learn.*, 2018, pp. 306–316.

[32] S. Chitta, I. Sucan, and S. Cousins, "Moveit! [ros topics]," *IEEE Robot. Automat. Mag.*, vol. 19, no. 1, pp. 18–19, Jan. 2012.

Authorized licensed use limited to: George Mason University. Downloaded on June 30,2023 at 02:06:35 UTC from IEEE Xplore. Restrictions apply.