

Perception and Action Augmentation for Teleoperation Assistance in Freeform Tele-manipulation

TSUNG-CHI LIN, Worcester Polytechnic Institute, Robotics Engineering

ACHYUTHAN UNNI KRISHNAN, Worcester Polytechnic Institute, Robotics Engineering

ZHI LI, Worcester Polytechnic Institute, Robotics Engineering

Teleoperation enables controlling complex robot systems remotely, providing the ability to impart human expertise from a distance. However, these interfaces can be complicated to use as it is difficult to contextualize information about robot motion in the workspace from the limited camera feedback. Thus, it is required to study the best manner in which assistance can be provided to the operator that reduces interface complexity and effort required for teleoperation. Some techniques that provide assistance to the operator while freeform teleoperating include: 1) perception augmentation, like augmented reality visual cues and additional camera angles, increasing the information available to the operator; 2) action augmentation, like assistive autonomy and control augmentation, optimized to reduce the effort required by the operator while teleoperating. In this paper we investigate: 1) which aspects of dexterous tele-manipulation require assistance; 2) the impact of perception and action augmentation in improving teleoperation performance; 3) what factors impact the usage of assistance and how to tailor these interfaces based on the operators' needs and characteristics. **The findings from this user study and resulting post-study surveys will help identify task based and user preferred perception and augmentation features for teleoperation assistance.**

CCS Concepts: • **Human-centered computing** → **Empirical studies in interaction design**; *Gestural input*; **Interaction design theory, concepts and paradigms**.

Additional Key Words and Phrases: Freeform tele-manipulation, AR visual cues, shared autonomous control

ACM Reference Format:

Tsung-Chi Lin, Achyuthan Unni Krishnan, and Zhi Li. 2022. Perception and Action Augmentation for Teleoperation Assistance in Freeform Tele-manipulation. *ACM Transactions on Human-Robot Interaction*, (2022), 37 pages.

1 INTRODUCTION

1.1 Remote Perception and Tele-action Problems in Freeform Tele-Manipulation

Problem Statement. Contemporary motion tracking interfaces (e.g., HTC Vive virtual reality system [93]) enable manipulator robots to track the natural human arm and hand motions to perform more dexterous, freeform manipulation. While human operators can efficiently and intuitively control gross manipulation (e.g., reaching to or moving an object), they may experience significant cognitive and physical workload when trying to control precise manipulation, such as carefully adjusting the robot end-effector near an object for grasping or placing. **This is usually because human operators can not acquire the necessary sensory information (e.g., visual or haptic) to**

Authors' addresses: Tsung-Chi Lin, tlin2@wpi.edu, Worcester Polytechnic Institute, Robotics Engineering, Unity Hall 200A, 27 Boynton St, Worcester, Massachusetts, 01609; Achyuthan Unni Krishnan, Worcester Polytechnic Institute, Robotics Engineering, aunnikrishnan@wpi.edu; Zhi Li, Worcester Polytechnic Institute, Robotics Engineering, zli11@wpi.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

2573-9522/2022/-ART \$15.00

<https://doi.org/>

perceive and control the remote tasks [32, 85]. For example, the operators may need the camera viewpoint from a different perspective to perceive the depth information not available in the primary camera viewpoint. They may also need proprioceptive and haptic feedback to precisely control the end-effector's motions or postures. Besides the remote perception problems, the cognitive and physical workload may also come from the difficulties in remote robot motion control. Freeform manipulation tasks typically involve both gross and precise manipulation which can be difficult to perform efficiently through interfaces designed around motion tracking. Changing the interface mapping and scaling (from human controlled inputs to robot motion outputs) based on task state or user input will be required to control the robot with efficiency and precision.

Limitations of Related Work. Related work has proposed to various approaches assist human teleoperators' remote perception and motion control. The existing solutions regarding the design of the tele-manipulation interfaces include the methods to:

- Improve the capabilities of interface to display additional sensory information (e.g., multi-camera viewpoints, haptic interfaces [12, 57]);
- Resort to an alternative sensory feedback to present the missing information, such as using augmented reality (AR) visual cues to represent the remote contact or force feedback [64];
- Delegate the part of the tele-manipulation task difficult for humans to capable and reliable autonomy [46] so that the interface only needs to present feedback on the autonomy's performance instead of the detailed sensory information.

However, related work in literature mostly compare to the *same* types of approaches to validate the effectiveness of their proposed methods. *There is still no work to compare different types of approaches, to inform how to choose among or integrate them when multiple types of approaches are available.* For example, if the tele-manipulation interfaces are capable of displaying AR visual cues, adjusting interface mapping dynamically or providing autonomy for action assistance: 1) Which method will be more effective to assist (which part of) the tele-manipulation? 2) Which method will be preferred by human operators? 3) Which operator specific factors affect the effectiveness and preferences of the teleoperation assistance features?

1.2 Division of Human and Robot Efforts in Assisted Teleoperation

Problem Statement. Another problem we are concerned with is: how to combine human operator and robot autonomy to optimally control tele-manipulation? Our insight from the related work in the literature and our prior work is that: shared autonomy to assist tele-manipulation can be more effective if its is designed to enable an appropriate **division of task and effort** between the human and robot. Such task division should allow human to have sufficient freeform control to perform the unstructured parts of the task, and allow robot autonomy to handle the structured parts of the task with desired performance (e.g., speed, accuracy, reliability).

Limitations of Related Work. In recent related work, the shared autonomy to assist remote manipulation are mostly designed to assist as much and as early as possible, based on the prediction of human intents (e.g., target object [61], expected motion trajectory [91]). These shared autonomy designs are designed to minimize control inputs and efforts, and may not always be necessary and effective to assist the operators who could prefer to have more freedom than assistance to control gross manipulation. While humans can easily perform freeform reaching motions to clearly indicate the object they intend to grasp, the most effective way to reduce human workload is to provide autonomy only to the part of the task that cause humans high cognitive or physical workload. Our prior work has developed a shared autonomy that provides autonomous actions (for grasping) to reduce human's physical workload [60]. In this paper, we will extend the assistance design

to investigate the need for robot to estimate human cognitive and physical workload to detect whether the human needs assistance, and determine whether the robot can provide the autonomy (for perception or action) useful to the human.

1.3 Overview of Research Efforts

Motivating Example. Consider a comprehensive tele-manipulation task such as workspace organization, which may involve control of reaching, grasping, moving, and placing of various kinds of objects. This task is mostly unstructured and requires freeform control, because it does not follow (and require) any procedure; What, when and how to handle each object will be decided by the user on-the-fly as the task goes. Sufficient freeform control will also allow human to improvise based on their knowledge and experience (leveraging environmental constraints, physical properties of handled objects, etc) to facilitate or enable some manipulation. To effectively assist tele-manipulation without compromising the human's control authority and freedom, the robot autonomy can provide an additional camera viewpoint from a different perspective or AR visual cues, to augment their remote perception and enable them to control the manipulation themselves. The robot autonomy can also be different interface mappings or autonomous actions that can effectively perform the task. As the human has moved the robot end-effector close enough to the target object or container to place, it is easier for the robot to infer human's intent and assist in the structured component (placing object) of the unstructured task using simple but effective autonomy.

Proposed Method. To this end, we propose systematic approaches for action and perception augmentation. The **Action Augmentation** allows humans to control the robot motions using hand pose tracking and trackpad available on the hand-held controller, for freeform or constrained motion control. It is also implemented by dynamically adjusting the scaling of operator to robot interface mapping to support both gross and precise manipulation. For **Perception Augmentation**, we provide AR visual cues to convey the visual information difficult to perceive in 2D camera (e.g., the task and robot status, interface control mode, and autonomous action affordance). We also provide a complementary camera viewpoint from a significantly different perspective, in which missing visual information, like loss of depth perception, can be more easily perceived.

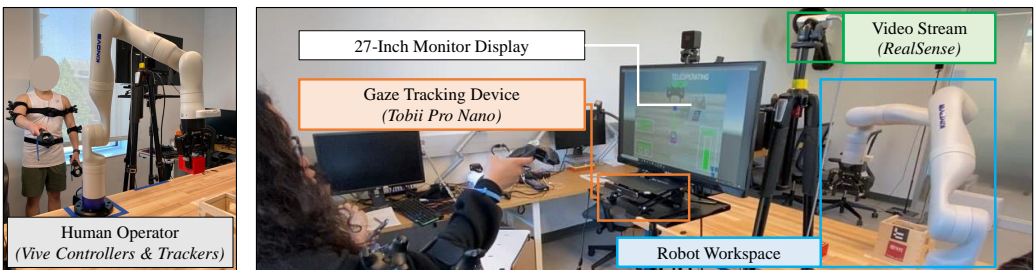


Fig. 1. Overview of the tele-manipulation system.

Implementation. We have implemented the proposed perception and action augmentation on a representative interface. Shown in Figure 1, we used the HTC Vive hand-held controller to control robot motion, and used desktop monitor to display the remote camera viewpoints and AR visual cues. Robot could provide autonomous actions (e.g., grasping and placing actions) or switch to constrained motion control using trackpad, when humans operated the robot end-effector near the target object or location to place. The implementation can be generalized to the teleoperation

system using various contemporary tele-manipulation control devices (e.g., hand-held, touch-based, wearable), display (e.g., screen-based and head-mounted visual display). Note that we only provided the autonomous actions and motion constraints needed for the pick-and-place tasks, because this work does not focus on how to predict human intents, or how to enable the autonomy.

Experiments and Results. We conducted three pilot studies to determine the design parameters of perception and action assistance. In Pilot Study I, we provided the complementary viewpoint from different perspectives according to the recommendation in the literature [26] to determine the complementary viewpoint angle and distance useful to our task and set up. In Pilot Study II, we evaluated what combinations of perception and action augmentation humans may prefer to inform the formal integration of the interface. In Pilot Study III, we validated our proposed methods for online estimation of muscle efforts using human motion trackers based on surface electromyography (sEMG) measurement. Our formal user study to validate and compare the proposed perception and action augmentation included 23 participants of diverse gender, profession and experience with technology. We use objective metrics to evaluate the task performance (e.g., completion time, types and occurrence of errors), cognitive workload (from the eye tracking data) and physical workload (from motion tracking data), to compare the effectiveness of the proposed perception and action augmentation, and their combinations. We also analyzed the survey feedback from participants to understand their preference for different perception and action augmentation. Our comprehensive user study shows that *the effectiveness of and preference for the perception and action augmentation depend on the task performance objective, the user's need for assistance, and the types of users.*

Main Contributions. The specific contributions of our work include (1) a novel shared autonomy to leverage human capabilities of freeform control and an assist-as-needed robot autonomy for effective, intuitive and ergonomic human-robot collaboration to perform tele-manipulation tasks (Section 3.2); (2) a generalizable design of AR visual cues to provide the information critical to the precision and performance of the remote manipulation (Section 3.2); (3) the integration and comparison of different types of perception and action augmentation to discover new knowledge on optimal human-robot collaboration for freeform tele-manipulation (Section 3.5); (4) a novel approach for objective physical and cognitive workload estimation based on human motion and eye tracking devices (Section 3.6); (5) a demonstration of the effectiveness and usability of the proposed interfaces and evaluation methods via a human experiment (Section 4).

2 RELATED WORK

The challenge of robot remote manipulation via freeform control has been organized in the sense of control effectiveness and effort. Our cognitive and physical workload estimation method enables the evaluation of non-neglect effort while performing the freeform tele-manipulation. We further implement the action and perception augmentation drawn from prior work (i.e., control mapping and scaling, complementary viewpoint, and augmented reality visual cues) to reduce the physical and cognitive workload.

2.1 The Cognitive and Physical Workload in Freeform Tele-manipulation

Maintaining human freedom in motion control is essential to the freeform teleoperation of unstructured, unpredictable manipulation tasks (e.g., including tele-robotic surgery [29], nursing assistance [57], manufacturing [38], hazardous material handling [92], explosive ordnance disposal [83]). Such tele-manipulation tasks are usually not feasible or error-prone for high-level robot autonomy (refer to the review on the level of autonomy [9]), and heavily depend on human knowledge, expertise and robot control dexterity. To assist freeform tele-manipulation, it is preferred to enable human to *efficiently* and *intuitively* control the remote robot and cameras,

while having some low-level robot autonomy for perception or action support to reduce human operator's *cognitive and physical workload*. Consider the various motion mapping interfaces (e.g., soft/hard exoskeletons [41], camera/IMU-based motion capture systems [58]) that can map the human body, arm and hand motions to efficiently and intuitively control the freeform motions of manipulator robots. These interfaces tend to cause non-trivial cognitive and physical workload [58], because precise control of manipulation motion or posture could be difficult without the necessary haptic or proprioceptive feedback [5, 72]. The remote visual perception problems, including the limited field of view, loss of depth information and unnatural camera viewpoint control (typically for eye-in-hand cameras), also contribute to the cognitive and physical workload. These cognitive and physical workload not only fatigues the teleoperator if they use the interfaces for hours but may also lead to work-related musculoskeletal injury for the operator who use the teleoperation interfaces on a daily basis. For the tele-manipulation tasks that are designed for manipulation dexterity rather than handling heavy payload, the “assist-as-need” action augmentation helps human to efficiently and reliably complete the manipulation actions clearly indicated by humans (e.g., by moving the end-effector close enough to the target location or object). This will be more effective than the autonomy predicting human intents (e.g. [7, 30, 46]) to assist as early and as much as possible. To effectively reduce the cognitive workload, the perception augmentation can present additional visual information using the camera viewpoint from a different perspective, or present AR visual cues to communicate the high-level task and robot states, so that human operators do not need to perceive and comprehend the low-level feedback from various sensors [28]

Table 1. Conventional and contemporary control interfaces for assisted tele-manipulation.

Type	Representative Interfaces
Conventional	
Desktop	Keyboard + mouse + point-and-click graphical user interface [16, 50, 102, 103]
Hand-Held	Xbox gamepad [55, 70, 87, 98], (Haptic) Joystick [4, 12, 23, 30, 39, 68, 69, 103], Customized teleoperation console (e.g., Da Vinci Surgeon Console [1, 22])
Contemporary	
Wearable	Arm/hand exoskeleton [10], Data glove [31, 56], Soft haptic glove [63, 71]
Hand-Held	Hand-held controllers of virtual reality systems (e.g., HTC Vive, Oculus) with trackpad and buttons, Robopuppet [27], Chopstick [49], Haptic tweezer [84], Tangible interface [14]
Motion / Gesture	Touchless: Arm/hand motion tracking (e.g., vision-based [24, 25, 90, 103], marker-based [20, 58], and IMU-based [100]), Mid-air gesture control [17, 43]; Touch-based: Touch-screen gesture control [11, 62, 101]

2.2 Action Augmentation via Interface Mapping and Scaling Design

Control Interface. Table 1 categorizes the conventional and contemporary control interfaces for assisted tele-manipulation that represent the state-of-the-arts. Compared to the conventional interfaces, the contemporary interfaces tends to: 1) improve the control *dexterity* of high degrees-of-freedom motion coordination (e.g., multi-finger coordination, hand-arm coordination), and simultaneous position and orientation control; 2) improve the *intuitiveness* of manipulation control, either by mapping natural human motions to robots, or replicating/representing the controlled robots or the manipulated objects (e.g., using 3D-printed prototypes or virtual reality). 3) improve

control *accuracy* by providing (shared) autonomy with/out haptic feedback. Considering the tele-manipulation assistance in recent related work, we also found that: while the **action support** that (partially) automates task-specific manipulation actions can improve the control accuracy and are used more for structured tasks [19, 54], **control augmentation**, such as the design of *interface mapping* and *scaling*, can better enhance the dexterity and intuitiveness, can be generalized across various interfaces, and are more used for freeform manipulation [67].

Mapping Design. While being intuitive, motion tracking interfaces are generally limited in their control efficiency. This is a result of limited accuracy of human motions, and interference of intended and unintended motions due simultaneous control of many Degrees of Freedom (DOFs). The efficiency of the controlled motions can be improved by introducing constraints in terms of virtual fixtures [76] or autonomy for teleoperation assistance (e.g. collision avoidance [73], motion guidance toward intended goal [60]). From a more general perspective, the interfaces that map gestures or point-and-click actions to autonomous robot motions or movement primitives [103] can all be considered as some kind of constraints that limit the extent to which human operator can control the robot motion freely. In addition, motion constraints can also be introduced by the separation of degrees of freedom (DOFs) in the design of interface mapping. For instance, people may use separate controllers to manipulate a 3D object's position and orientation, to avoid the interference of intended and unintended motion control [103]. Some interface hardware, such as the trackpad of hand-held controllers, the joystick of gamepads, are naturally suitable for the separation of DOFs as they can clearly distinguish the control inputs for different motion directions based on the controlled axes. For screen- or projection-based interfaces, interactive avatars such as the ring-and-arrow markers [75], the virtual handlebar [37] enable the independent control of individual DOF(s) of the manipulated (virtual) object or robot end-effector.

Scaling Design. The motion scaling ratio of the interface affects both the control efficiency and accuracy of the tele-manipulation tasks. Scaling up the control motion can increase robot motion speed and range but may compromise the accuracy of motion control. On the other hand, scaling down the control motion will increase the motion control accuracy in the concerned small-scale workspace, but may also improve the efficiency by reducing operational errors. The scaling ratios can be fixed (commonly used by tele-robotic laparoscopic or eye surgery interfaces [40]), or vary with the user's operating speed (e.g., PRISM method [33]) or regions of operation [67]. Both fixed and varying scaling have pros and cons. Interfaces with varying scaling ratios can better adapt to the control of fast and slow motions in a large or small workspace. In contrast, interfaces with fixed scaling ratios tend to be more stable and predictable to the teleoperator. The major concern in the design of interface scaling ratio is how to achieve the trade-off of control efficiency, adaptability and predictability. Related work in the literature proposed several solutions to achieve such trade-off, which allow the user to: a) manually switch among several pre-defined scaling ratios (suggested by task experts) depending on the types of operation or size of the workspace [40]; b) manually adjust the scaling ratio as a continuous control parameter (e.g., by changing the distance between the controlling hands [37, 88]); c) autonomously adjust the scaling ratio (e.g., according to the location of operation in the workspace [25]). While these representative designs balance the performance objectives to some extent, it was also revealed that: 1) Being able to adjust scaling ratio (manually or autonomous) is useful overall but may leads to a complicated interface design hard for users to learn; 2) The manual switching of scaling ratio modes leads to more predictable interface behavior, but may increase the control efforts and mental workload; 3) Autonomously adjusting the scaling ratio reduces the control efforts and cognitive workload, but has to be carefully designed to the nature of task and the preference of users.

2.3 Visual Assistance for Perception Augmentation

Complementary Viewpoint. Human operators need to have visual information regarding the global, spatial relationship in the workspace, and the detailed, local visual information of region-of-interest critical to the precise operation. Realistic tele-manipulator robots also need to have multiple cameras dedicated to provide global and local viewpoints, or a single camera with sufficient mobility and motion of range to serve both purposes. Related literature and our prior work proposed to present both (or switch between) the global viewpoint from an onboard or standalone workspace camera, and the local viewpoint from a high-mobility, eye-in-hand camera to focus on the region of interest (e.g., using detail-in-context display [48, 86, 94]). However, for gross manipulation, humans may need to reach beyond the workspace covered by the workspace cameras. Providing a much-too-large camera FOV is not efficient for limited communication bandwidth and may compromise the resolution in visual display for the workspace of frequent manipulation operations. For precise manipulation, additional camera viewpoint from a different perspective may be necessary to confirm if the manipulation motion satisfies the task constraints in multiple degrees-of-freedom. Another problem we have to address is the dilemma to retain human's authority and freedom to control the camera while reducing the human's effort for camera control. The autonomy for dynamic camera viewpoint control and optimization can reduce human's camera control efforts, yet it may move the camera in an unexpected and unpredictable way and disrupts the operator's manipulation.

Augmented Reality Visual Support. AR visual cues can communicate very rich, detailed information using a variety of colors, shapes and displayed text. AR visual cues are preferred to assist the estimation of spatial relationship (e.g., gap estimation for driving assistance [82]), to direct and enhance visual attention (of drivers [74], and video game players [21]). More recently, the design of AR visual cues emphasizes how to intuitively communicate robot motion intent (e.g., goal and trajectory [15]) to assist the human control or supervision of more autonomous robots, and emphasizes how to enhance the depth perception [6], contextual understanding of 3D spatial relationship [95], and real-time status of robot, task and environment [97], to assist robot teleoperation. The AR display can be augmented to provide information about robot, interface, and environment [96]. The AR visual displays can also be integrated with virtual reality display of robot models [80], or presented with the haptic cues to communicate interaction force, motion constraints, or desired trajectories [36, 53, 66, 73].

3 INTERFACE AND EVALUATION DESIGN

This section will present our proposed approaches for perception and action augmentation, and their implementation on a representative tele-manipulation system and for a general-purpose pick-and-place task. We further propose methods for the estimation of eye tracking based cognitive workload, and motion tracking based physical workload to enable the evaluation of integrated interfaces in the formal user studies. **We conducted three pilot user studies to (1) provide the reasoning for complementary viewpoint design; (2) determine the effective integration of the perception and action augmentation; (3) validate the physical workload estimation with sEMG data.**

3.1 System Overview

Figure 1 shows the the tele-manipulation system developed in our prior work [59], which enables the development of the perception and action augmentation proposed for this project. The **robot platform** is a 7-DoF Kinova Gen 3 manipulator with a two fingered Robotiq gripper that can detect contact with the grasped object. Two RealSense Cameras (D435) were standing alone in the workspace for primary and complementary remote perception.

For **robot motion control**, we use an HTC Vive hand-held controller (referred as “controller” in the rest of this paper) that allows human operators to control the freeform robot motions using their natural hand motions and constrained motion using controller’s trackpad. By default (i.e., Mode 1 of Figure 2), the linear velocity of the controller will be mapped to the linear velocity of the robot. The input-to-output motion mapping ratio is 1:5 along the x-axis and 1:3 along the y- and z-axis. We locked the robot’s rotational motions because this work focuses on developing and comparing different *modalities* of teleoperation assistance instead of the capabilities of robot control. To perform a tele-manipulation task, the operator will: 1) press the menu button on the controller to send the robot to the home configuration, 2) press the menu button again to get the robot ready, and press the grip (side) button to initiate the control.

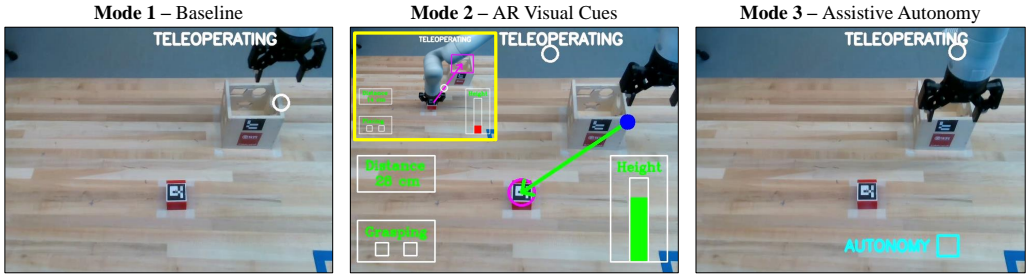


Fig. 2. Visual interfaces for baseline, AR visual cues and assistive autonomy.

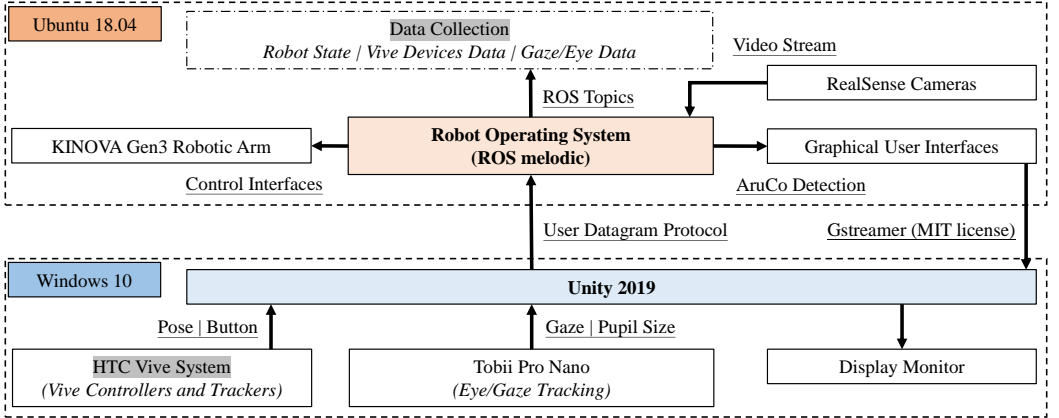


Fig. 3. System architecture.

For **visual feedback**, we used 1440×1080 pixel resolution Unity 3D window on a desktop monitor to display the remote camera video stream (at 30 Hz frame rate) and to display augmented reality visual cues (see Section 3.2 for details). By default, the graphical user interface (GUI) will only display the robot operation state. Specifically, the GUI will display: 1) “WAITING” when the tele-robotic system is ready for operation and waiting for a control command; 2) “SENDING HOME” when the operator presses a controller button to set the robot to the default pose; 3) “READY” when the robot is posed at the start position for the current task; 4) “TELEOPERATING” when the robot is being teleoperated; 5) “PAUSED” when the robot is paused by the teleoperator. Figure 3 shows the

control architecture and data communication pipeline of the tele-manipulation system. The RGB video from the remote cameras are streamed at 30 Hz frame rate. A screen-based eye tracker (Tobii Pro Nano) was attached below the monitor to track the human operator's gaze and eye movements (e.g., pupil diameter) at 60 Hz. The autonomy for perception and action can detect the ArUco tags attached on the objects, container and counter workspace [35, 79], to estimate the information for the AR visual cues and control the robot autonomous actions for precise manipulation (e.g., object grasping and placing).

3.2 Design and Implementation of AR Visual Cues and Assistive Autonomy

To assist robot remote manipulation, we implemented systematic AR visual cues and user-triggered autonomous actions as the baseline representing the common solution for remote perception and action problems. We then develop, integrate and compare different types of perception and action assistance upon the baseline AR visual support and assistive autonomy, to discover new knowledge on optimal human-robot collaboration for freeform tele-manipulation.

AR Visual Cues. Our prior work [59] has proposed four types of AR visual cues for freeform teleoperation assistance, including: 1) the **Target Locator** to indicate robot's movement direction and distance to the targeted object or goal pose; 2) the **Action Affordance** to indicate if the robot is ready to afford the action to be performed (e.g., grasping or stacking an object, with a good chance of success); 3) the **Action Confirmation** to indicate that the robot has successfully performed an appropriate action; and 4) the **Collision Alert** to alert the teleoperator if the end-effector is about to violate any environment constraints (e.g., hitting the table). Figure 2 (Mode 2) shows the implementation of these AR visual cues to assist a pick-and-place task for this work.

- The **Height** indicator shows the robot's distance to table surface. Besides the display of numerical distance, the height bar display also turns from green to red if the robot is too close (within 0.1 m) to the table.
- The **Alignment** indicator (displayed as a dot-in-circle) shows if the robot is aligned with the object to grasp or the container to place the object, in x- and y-direction. Once the blue dot moving with the robot is aligned with the pick circle displayed on the object or containers, the pink circle will changes its color to light blue to indicate the operator can reliably close or open the gripper to reliably grasp or drop the object into the container.
- The **Grasping/Placing Hint** include two square-shape that turn on and off to show whether the robot is aligned with the object or the container so that the operator can confidently close or open the gripper. It is designed to confirm the critical information conveyed in the **Alignment** and **Height** cues.
- The **Arrow with Distance** indicator shows the distance (in cm) and direction (using green and pink arrow, respectively) to show the target object to grasp or container to place.

The proposed implementation of AR visual cues is refined based on our prior design and evaluation results [59]. Specifically, we have: adjusted the **Height** indicator to be vertical instead of horizontal for a more intuitive visual display. We grouped the **Grasping/Placing Hint** into a white box and highlighted the boundary of the container to make them easier to spot at a glance. We also extended the AR support to picking-and-placing as well.

Assistive Autonomy. Shown in Figure 2 (Mode 3), we also provide autonomous actions to assist the operators to perform precise manipulation (e.g., to pick and place and object). The robot autonomy can detect the human's goal and action intents based on robot states, including the distance to the object or container, and whether the gripper is open or closed. When the gripper is open, we predict human intent to grasp the object if the robot is within the predefined distances to the center

of the object (0.05 m, 0.08 m and 0.13 m in the x-, y- and z-direction). A hint of "AUTONOMY" will be displayed to show the robot has detected human's goal and action intent by filling the box when the robot can reliably perform the action. Human therefore can press the controller's trigger to confirm the execution of the autonomous action, after which the robot will autonomously reach to grasp the object and lift to 0.2 m above the table surface. To place an object, the operator needs to move the robot to be within a predefined distance (0.08 m, 0.08 m and 0.15 m in the x-, y- and z-direction) to the center of the top of the container. Once confirmed by the human, the robot will autonomously move to the top of the container and drop the object into it reliably.

Remarks. Our proposed visual and action augmentation depend on the robot autonomy to predict human intents, determine action affordance and success, and detect and avoid collision. Here we implemented a simple design of autonomy that predicts the human's intent to grasp or place an object based on the robot states. The object detection, action affordance and collision is also simplified given that we know location and geometry of the object and the environment constraints. Note that more advanced methods to predict human intents, from human control inputs [24, 46, 81], gaze [3, 78], or their fusion [77], can be integrated with our proposed visual and action augmentation for more complex manipulation tasks. Advanced methods to detect objects and their action affordance (e.g., using Sim2real approach [47], for unknown objects [18]) can also be incorporated to enable more complicated precise manipulation and the delicate control of interaction forces. Collision in dynamic and cluttered environment can be detected using advanced method such as generalized velocity obstacles [99].

3.3 Complementary Viewpoint for Perception Augmentation

We propose to leverage an additional workspace camera to provide a complementary viewpoint in which the operator can better perceive the information missed in the primary workspace camera viewpoint. Shown in Figure 4, the GUI presents a picture-in-picture (PIP) display to embed the complementary viewpoint into the primary viewpoint. The perception augmentation in form of the complementary viewpoint can be presented always (i.e., the fixed viewpoint) or dynamically given the robot and task states (i.e., dynamic viewpoint). It can also be augmented to different interface control modes. Here we present the pilot user study (Pilot Study I) we conducted for iterative design and evaluation.

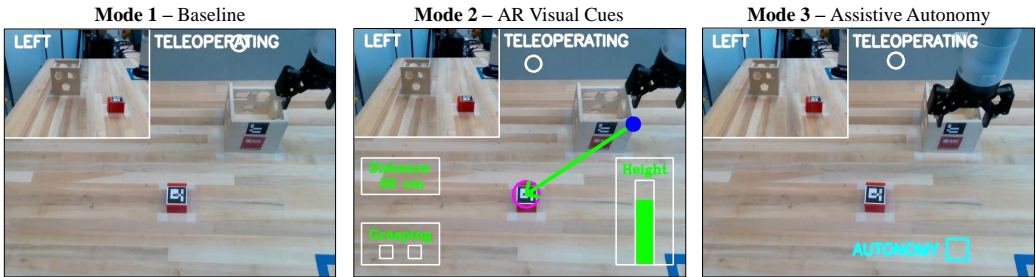


Fig. 4. Complementary viewpoints in different interface modes.

Q1 - Do we need multiple viewpoints? We conducted a pilot study with one expert participant (female, age=33, without visual or motor disability, 100+ hours experience with robot) to determine the preferred viewpoint angle and distance of the complementary viewpoint. We follow the experiment setup in the literature [26], and set up five workspace cameras (4 RealSense cameras and

1 Webcam) to observe the workspace from different perspectives (the front, back, left, right, top). Including the viewpoint from the eye-in-hand camera of the robot, we presented six viewpoint to the user and tracked her gaze fixation on each viewpoint during the pick-and-place task. Shown in Figure 5, the operator was asked to reach to grasp four blocks of different colors placed around a red cup, and place each into the cup. During the task, the participant used an HTC Vive controller to control the robot motions. **For the pilot study, the head-mounted display of the HTC Vive Pro Eye system is used to display graphical user interfaces and track human gaze.**

In Figure 5, the camera viewpoints that the human looked at are compared between different manipulation action, and compared between the manipulation of different objects. The operator's gaze fixation mostly switched between the *back view* that looks at the workspace from the operator's standing point and the viewpoint in which she could observe the object to pick up with minimal occlusion. We also found that the participant spent more time to looking the back view than any other viewpoint, which implies that we need to distinguish the primary and complementary viewpoints based on the duration of their fixation.

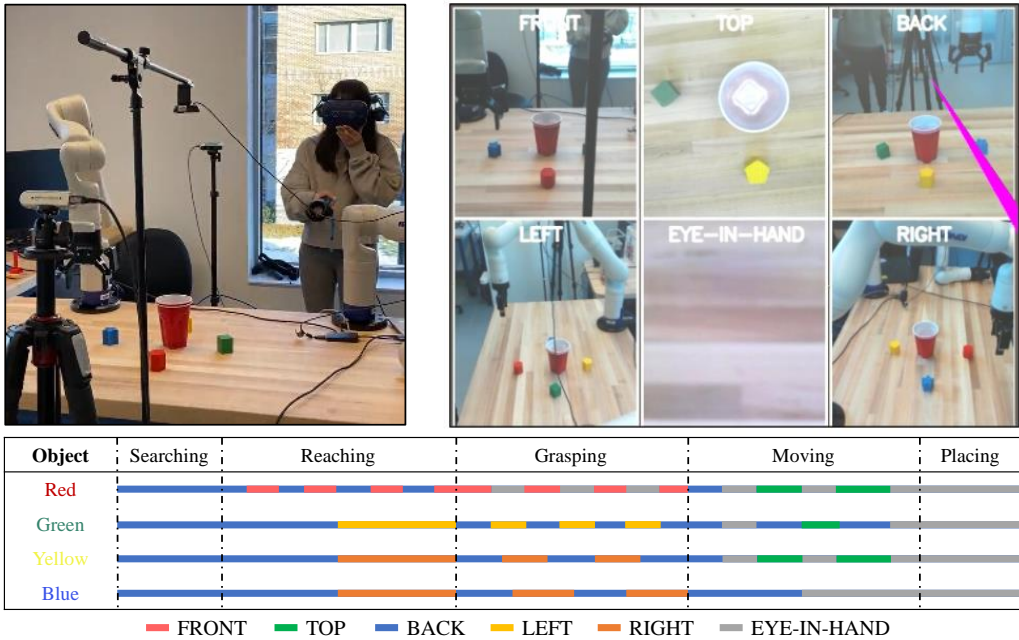


Fig. 5. Complementary viewpoint filtering.

Q2 - Which camera is preferred for the complementary viewpoint? We conducted another round of the pilot study **with the same participant** to determine the preferred camera view for a complementary viewpoint. Based on the result from Q1, we implemented a picture-in-picture multi-viewpoint display. By default, we displayed the back view camera to be the primary viewpoint and the front view camera to be the complementary viewpoint. **The complimentary viewpoint system was selected since the back view was utilized the most with different viewpoints used only when additional information was required. This implied that only one additional viewpoint to the primary viewpoint would be required.** Shown in Figure 6, the operator could also press the controller's button to switch the complementary viewpoint to be from other workspace cameras. We recorded the robot and task states, and tracked human gaze. We noticed that the operator preferred to only use the left view

camera for the complementary viewpoint, because: 1) it shows the additional objects not visible in the back view, and 2) it is less occluded by the robot arm. The participant also mentioned that manually switching the complementary viewpoint increased her cognitive workload and control efforts during a post-study interview. She also mentioned that the complementary view could be improved with a zoom functionality to provide detailed information of the task and workspace.

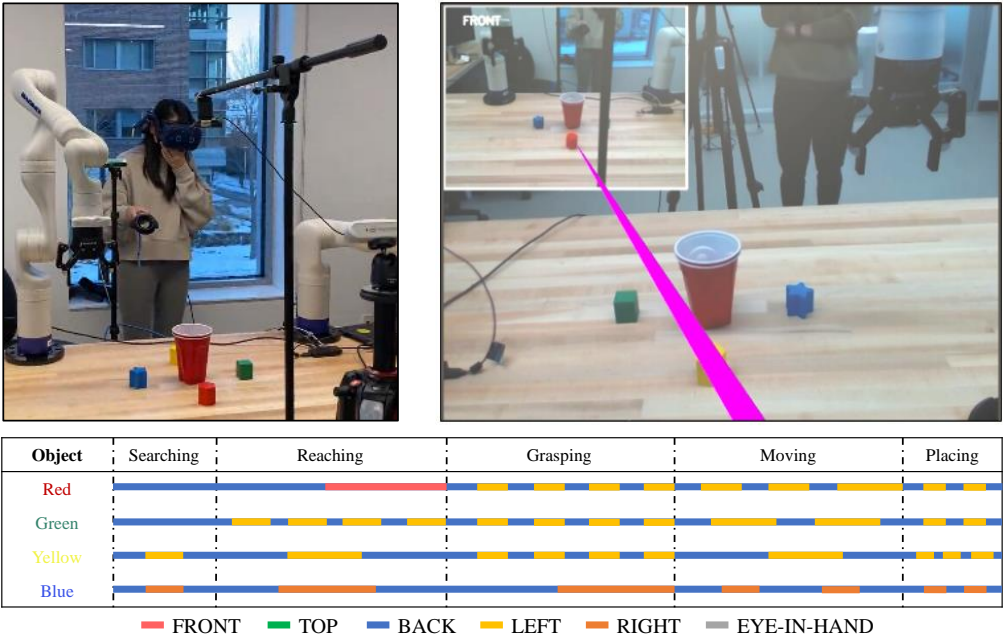


Fig. 6. Gaze fixation on each camera viewpoint.

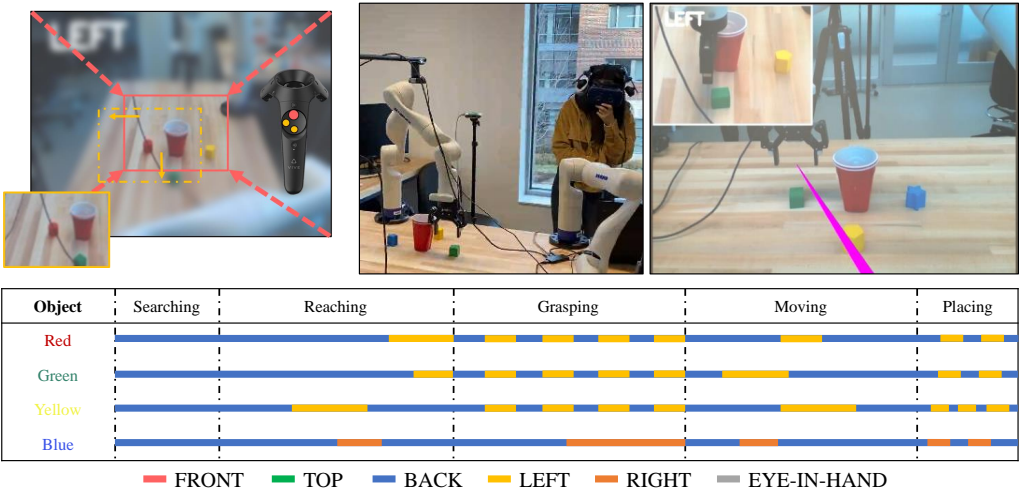


Fig. 7. Complementary viewpoint with adaptive field of view.

Q3 - Do we need to adjust the field of view for the complementary viewpoint? Based on the feedback from Q2, we enabled operator to use the controller's trackpad to control the complementary viewpoint to shift the center of the field of view (FOV), and to zoom in and out (Figure 7). We found that during the same pick-and-place task, the operator still chose the complementary viewpoint cameras in a similar way, but preferred to zoom in and shift the FOV to make the target object or container more centered and visible.

3.4 Dynamic Interface Mapping for Action Augmentation

Dynamic interface mapping, controlled by human or autonomy, enables humans to use different interface mappings or scaling ratio to effectively control precise and gross manipulation. While manually adjusting the interface mapping could be annoying and tedious, existing autonomy to adjust the interface mapping [34, 88] tend to confuse humans because they do not intuitively inform human about this change, due to which humans may find the interface inconsistent and unpredictable. Our recent work [93] shows that humans can more efficiently control precise manipulation if the interface mapping: 1) allows humans to constrain the motions to be only for position or orientation control; and 2) autonomously reduce the human-to-robot motion mapping ratio (which will reduce the robot motion speed) close to the objects and environment constraints.

For this work, we propose to improve efficiency to control precise, directional motions by: 1) allowing humans to use constrained motion control input channels (e.g., the controller's Trackpad) to control the motions of an individual DOF, and 2) reducing the scaling ratio only in the precise motion control direction.

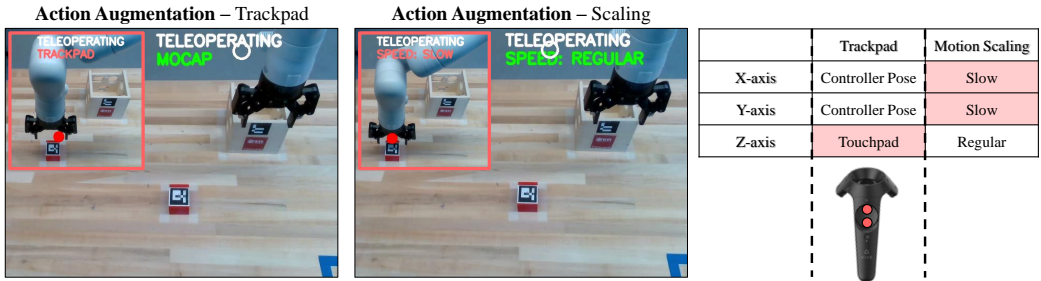


Fig. 8. Action augmentations using trackpad and motion scaling.

Figure 8 shows the implementation of our two proposed action augmentation approaches. In the “Trackpad” control mode, the human can still control the robot in the X- and Y-axis (horizontal plane motion) using natural hand motions but will control the motion in the Z-axis (vertical motion) using the controller's trackpad (to avoid collision with the table and object to be grasped). When the human controls the robot to move close to the object or containers, the corresponding AR visual cue will turn from “MOCAP” (in green) to “TRACKPAD” (in red) to inform human that the interface mapping mode has changed. The trackpad control region for the object (container) is a $0.14 \text{ m} \times 0.2 \text{ m} \times 0.6 \text{ m}$ ($0.2 \text{ m} \times 0.26 \text{ m} \times 0.4 \text{ m}$) bounding box w.r.t. to the center of the object (top of the container). In the “Scaling” mode, the interface will reduce the mapping ratio in the X- and Y-axis to allow human to precisely adjust the robot to align with the object or container, while maintaining the scaling ratio to be “Regular” in the z-axis. We define the reduced scaling region to be the same as the trackpad control region. The corresponding AR visual cue will turn from “SPEED: REGULAR” in green to “SPEED: SLOW” in red to inform the change of scaling ratio.

3.5 Integration of Perception and Action Augmentation

We conducted a pilot user study (Pilot Study II) to understand human's preferred combination of the perception and action augmentation. In total, we have 15 different experiment conditions, considering the three interface control modes with different perception and action augmentation. Our pilot study involved 8 participants (4 male and 4 female, 5 novices and 3 participants who have used the same teleoperation system before). The participants performed a single-object pick-and-place task once under every experimental condition (order of interfaces were randomized), and reported their preferred combinations of control modes and perception/action augmentation after the experiments.

Table 2. Pilot Study II: testing conditions with highlighted combinations preferred by the participants.

	Baseline (Mode 1)	AR Visual Cues (Mode 2)	Assistive Autonomy (Mode 3)
Default	1a: Single View	2a: Single View	3a: Single View
PA = Perception Augmentation	1b: Fixed PIP (PA1)	2b: Fixed PIP (PA1)	3b: Fixed PIP (PA1)
	1c: Pop-up PIP (PA2)	2c: Pop-up PIP (PA2)	3c: Pop-up PIP (PA2)
AA = Action Augmentation	1d: Trackpad (AA1)	2d: Trackpad (AA1)	3d: Trackpad (AA1)
	1e: Motion Scaling (AA2)	2e: Motion Scaling (AA2)	3e: Motion Scaling (AA2)

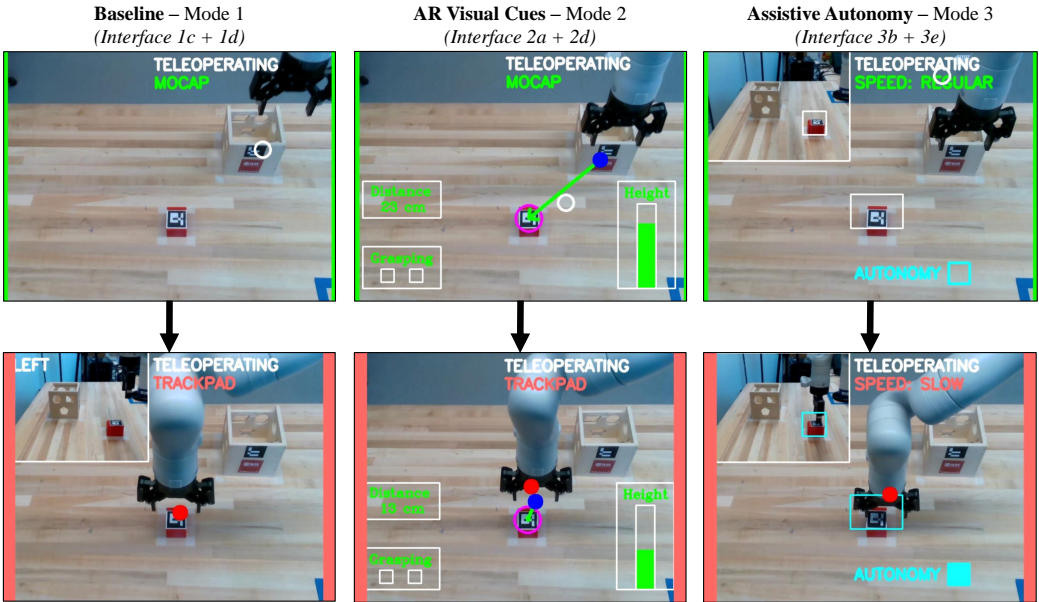


Fig. 9. Integrated interfaces of perception and action augmentation.

Table 2 highlighted the augmentation combination preferred by the majority of the participants for each mode. In Mode 1 (i.e., the baseline control mode), 5 out of 8 participants preferred to have the pop-up picture-in-picture (PIP) display of the complementary viewpoint (interface-1c) and to use trackpad control (interface-1d). Some participants commented: "...would like to have the pop-up PIP display to provide more workspace information when needed and use a trackpad to control the

robot in a single direction for precise motion". In Mode 2 when the interface can display AR visual cues, 6 out of 8 participants preferred to use the single camera view display (interface-2a) without any perception augmentation and trackpad control (interface-2d). **Participants commented:** *"...the PIP display overwhelms the user interface while the AR visual cues are available".* In mode 3 (assistive autonomy), 7 out of 8 participants preferred to use the fixed PIP display of the complementary viewpoint (interface-3b) and motion scaling (interface-3e). **As the participants commented:** *"...the fixed PIP display increases the awareness of the region where autonomy is triggered" and "...motion scaling prevents large movement that moves the robot out of the autonomy zone".* **The preferred combination of interface control modes with perception and action augmentation will be further evaluated in our formal user study.**

We further refine the interface display based on the freeform comments from the participants. Shown in Figure 9, we use sidebar in pink and green to prominently indicate the activation of action augmentation. In Mode 3, we also highlight the region to activate the autonomous actions in both the primary and complementary viewpoint. The corresponding AR visual cue (i.e., the square around the object) will be turned from white to light blue color.

3.6 Estimation of Cognitive and Physical Workload

We estimate the cognitive workload using the operator's gaze and eye movement tracked by a Tobbi Pro Nano eye tracker. We also propose a novel method to estimate the physical workload online from human motion tracking.

Estimation of Cognitive Workload (Offline). Following the methods in the literature [44, 52, 89], we will estimate cognitive workload caused by stress C_{str} , interface complexity C_{int} and task workload C_{tsk} from the operator's pupil diameter, gaze fixation and movements, and task duration. We will track difference between the operator's pupil diameter, and estimate the cognitive workload caused by stress as the difference between average pupil diameter (D_{tsk} during a task and the operator's calibrated pupil diameter D_{cal} before the task start, and will be normalized with respect to the maximum cognitive workload among all the participants, i.e., $C_{str} = \frac{D_{tsk} - D_{cal}}{\max_{p=p_1, \dots, p_n} (D_{tsk} - D_{cal})}$. Prior

literature suggests that [44, 52, 89] pupil dilates with increased workload, thus increasing the difference between the average pupil diameter during a task (D_{tsk}) and the operator's calibrated pupil diameter (D_{cal}) prior to the start of the task.

The cognitive workload caused by interface complexity C_{int} will be computed as the ratio between the average distance in pixels of the operator's gaze fixation and the center of visual display and the maximum distance in pixels (from edge to center of visual display i.e., S_{tsk} and S_{max}). Thus, the interface complexity can be calculated as, $C_{int} = S_{tsk}/S_{max}$. To compute the cognitive workload for each sub-task (e.g., picking-and-place one object), we will also estimate the cognitive workload caused by task complexity as the ratio between the time to complete a sub-task and total task completion time (namely, $C_{tsk} = T_{sub}/T_{total}$). Thus, the cognitive workload for a sub-task can be computed as the average of C_{str} , C_{int} and C_{tsk} . We also contribute the overall workload C_{task} of the entire task caused the stress and interface complexity as the average of C_{str} and C_{int} , assuming they have equal contributions.

Estimation of Physical Workload (Online). Surface Electromyography (sEMG) signals can provide more accurate measurements of the muscle efforts and physical workload than using subjective feedback (e.g., Rapid Upper Limb Assessment, namely RULA [2, 42, 65]). Our recent work has used sEMG for the *objective* but *offline* estimation of physical workload in robot teleoperation via whole-body motion mapping [58, 60]. Here we propose to learn predictive models for the online, accurate muscle effort prediction from human motion tracking data. Our prior work [58] shows

that: the muscle efforts of the anterior, lateral deltoid and bicep muscle groups, caused by shoulder flexion, abduction and elbow flexion, contributes most to the physical workload when human controls tele-manipulation using their arm and hand motions.

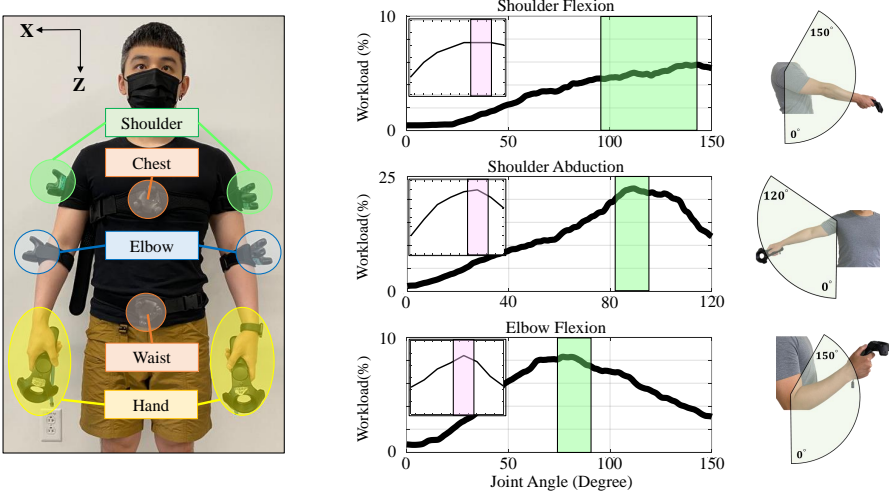


Fig. 10. Vive trackers attachment and physical workload single joint mapping and validation.

Shown in Figure 10, we thus attached 6 body trackers (Vive Tracker 3.0) to the upper arms, forearms, chest and waist of the human operator, to estimate the shoulder and elbow joint angles. Specifically, the shoulder flexion (θ_{SF}) is estimated on the **sagittal plane** as:

$$\theta_{SF} = \arccos\left(\frac{\vec{T}_{ua} \cdot \vec{g}}{\|\vec{T}_{ua}\| \|\vec{g}\|}\right) \quad (1)$$

which has \vec{T}_{ua} to be the upper arm vector estimated from shoulder and elbow trackers, and the \vec{g} to be the gravity vector, both of which are projected on the sagittal plane (i.e., the X-Y plane).

The shoulder abduction θ_{SA} is estimated on the **frontal plane** as:

$$\theta_{SA} = \arccos\left(\frac{\vec{T}_{vertical} \cdot \vec{T}_{ua}}{\|\vec{T}_{vertical}\| \|\vec{T}_{ua}\|}\right) \quad (2)$$

which has $\vec{T}_{vertical}$ to be the vector perpendicular to the vector connected two shoulder trackers, and \vec{T}_{ua} to be the vector of the upper arm formed by shoulder and elbow trackers, both of which are projected on frontal plane (i.e., the X-Z plane).

The elbow flexion θ_{EF} is estimated as

$$\theta_{EF} = \arccos\left(\frac{\vec{T}_{ua} \cdot \vec{T}_{la}}{\|\vec{T}_{ua}\| \|\vec{T}_{la}\|}\right) \quad (3)$$

which has the \vec{T}_{ua} to be the upper arm vector, and \vec{T}_{la} is the forearm vector estimated from the elbow tracker and hand-held controller positions. Both these vectors are project on the sagittal plane (i.e., the Y-Z plane). Note that: $0^\circ < \theta_{SF} < 150^\circ$, $0^\circ < \theta_{SA} < 120^\circ$ and $0^\circ < \theta_{EF} < 150^\circ$.

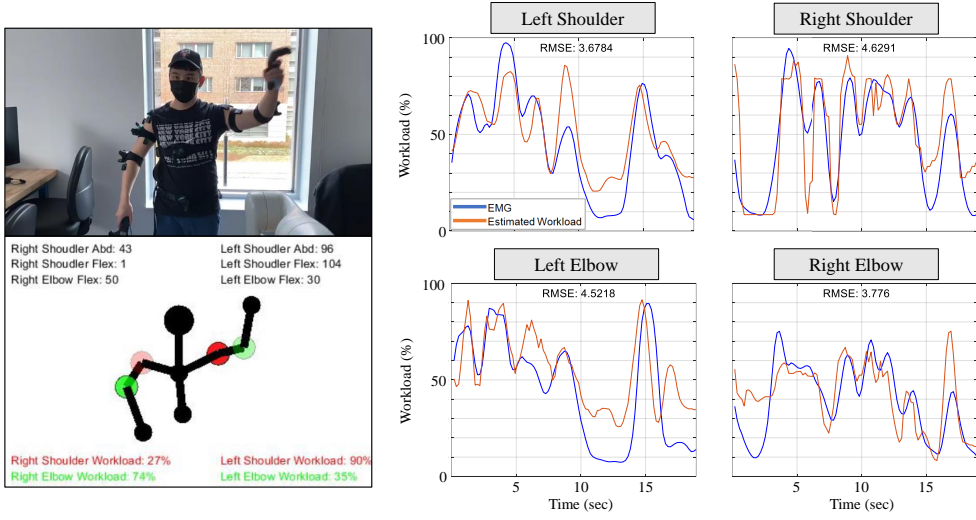


Fig. 11. Physical workload estimation via Vive trackers.

Shown in Figure 11 (Left), before tele-manipulation, we asked the human operator to perform a compound arm exercise that involves the active coordination of the anterior and lateral deltoid and the bicep muscle groups. The participants held one HTC Vive controller in each hand and move their shoulder and elbow from fully extended to fully flexed for 20 seconds at the speeds and angular velocities for typical robot control motions. We computed the joint angles of shoulder flexion, abduction and elbow flexion from the body and arm motions tracked by the HTC Vive trackers, and used the corresponding sEMG data to estimate the offline muscle efforts [58]. For the offline workload estimation, we used a band pass filter to extract the 40 Hz-700 Hz EMG signals from the wireless sEMG sensors (Delsys Trigno Avanti Sensors) attached to the anterior, lateral deltoid and bicep muscle groups. We pre-processed the data using a high pass filter (cutoff frequency 10 Hz) to remove the soft tissue artifact and offset the frequency baseline, and use a full-wave rectification then a sixth-order elliptical low pass filter (cutoff frequency 50 Hz) to remove noise and transients and develop a linear envelope of the EMG signals, following the method in the literature [45] but choose tunable parameters for our own task and data. The shoulder muscle efforts were computed using the weighted sum of the anterior and lateral deltoid (at the ratio of 3:4 based on their capabilities of force generation [51]), while the elbow efforts was calculated from the bicep flexion. The muscle efforts were computed by normalizing the processed EMG data with respect to the person's maximum voluntary contraction following the standard procedure in the literature [13]. We averaged the shoulder and elbow muscle efforts for each arm, and estimate the operator's overall physical workload as the weighted sum of muscle efforts from the dominant and non-dominant arm (at the ratio of 9:1), for the tasks that operators extensively move their dominant arms for robot motion control:

$$P_{overall} = 0.9 \times \left(\frac{P_{DS} + P_{DE}}{2} \right) + 0.1 \times \left(\frac{P_{NDS} + P_{NDE}}{2} \right) \quad (4)$$

where P_{DS} and P_{NDS} are the shoulder muscle efforts of the dominant and non-dominant arms, while the P_{DE} and P_{NDE} are the elbow muscle efforts. A set of injunctive mapping functions were learned to predict the muscle efforts based on the arm joint angles with good accuracy.

Figure 10 shows that our predictive model can estimate the sEMG-based physical workload based on the joint angles in isolation exercises, comparable to literature results [2, 42]. For compound

exercises, Figure 11 (Right) shows an example for the prediction accuracy of our simple models for one male (32 years old) and 1 female (33 years old) of functional upper extremities and normal body mass index. The root mean square errors (RMSE) between the proposed method and EMG data are 3.68, 4.52, 4.63 and 3.78 for male and 4.97, 3.81, 4.12 and 4.37 for female participant for the left shoulder, left elbow, right shoulder and right elbow.

4 USER STUDY

Research Questions. We conduct a user study to address the following research questions:

- **RQ1: What** aspects of dexterous tele-manipulation is improved while using generally preferred improvements upon freeform teleoperation?
- **RQ2: How** do the different types of augmentation impact the performance, workload and preference?
- **RQ3: When** should the teleoperators be provided with the visual and action augmentation to increase task performance and decrease operational workload?
- **RQ4: Who** should be provided with what type of augmentation for freeform teleoperation?

Experiment Setup. Figure 1 shows the tele-manipulation system used for our experiments. The participants were instructed to control the Kinova Gen 3 manipulator robot to perform a single-object pick-and-place task using an HTC Vive hand-held controller. Two RealSense D435 cameras were set up to provide the primary and complementary viewpoints, to provide the back view and left side view of the workspace, while a desktop monitor was used to display the GUI (with camera viewpoints and AR visual cues) to the operator. HTC Vive body trackers and hand-held controllers were used to track their body and arm motions for online physical workload estimation. The Tobii Pro Nano screen-based eye tracker attached under the desktop monitor display was used to track the operator's gaze and eye movements for cognitive workload estimation. *Unlike the pilot study, for this user study a screen based tracker was used because the visual interface was relayed on a computer screen as opposed to an head-mounted display.*

Participants. Our experiments include $N=23$ participants (28 ± 10 years old) diverse in gender, technological and professional experience. The participants were divided into several user groups based on the factors including:

- **Gender:** Based on gender, participants comprised of 14 male and 9 female participants.
- **Background:** The 23 participants comprised of 5 nurses and 18 users who do not have a nursing background. Participants were determined to have a nursing background if they are a nursing student or a registered nurse. *Participants with nursing background were recruited in order to incorporate our intended future users for a teleoperation platform nursing in the development stage.*
- **Proficiency:** The participants could be divided into 9 experienced and 14 inexperienced users based on their experience in having used the teleoperation system. Users were classified as experienced users if they had more than one hour of experience controlling the robot via teleoperation. They must have also have teleoperated the robot within one year to the day of their participation in the user study. The experienced users included participants of the pilot study for this user study, in addition to other experienced participants from prior *user studies for different experiments.*
- **Gaming:** The participants were divided into 16 infrequent video game players and 7 frequent video game players. Participants who spent less than 5 hours a week playing video games were classified as infrequent video game players.

- **Spatial:** Based on their spatial reasoning skills (via a spatial test from AssessmentDay [8]), the 23 participants were divided into 10 people with low spatial reasoning skills and 13 people with high spatial reasoning skills. The participants' spatial reasoning skill was evaluated using a spatial reasoning test that was part of the pre-user study survey. Participants who scored less than 60 percentile in the test were evaluated to have low spatial reasoning skills.
- **Mode Order:** The participants were also divided based on the order in which the participants used the interface modes. 8 users used the interface modes in the $1 \rightarrow 2 \rightarrow 3$ order. 7 users used the interface modes in the $2 \rightarrow 3 \rightarrow 1$ order. 8 users used the interface modes in the $3 \rightarrow 1 \rightarrow 2$ order.

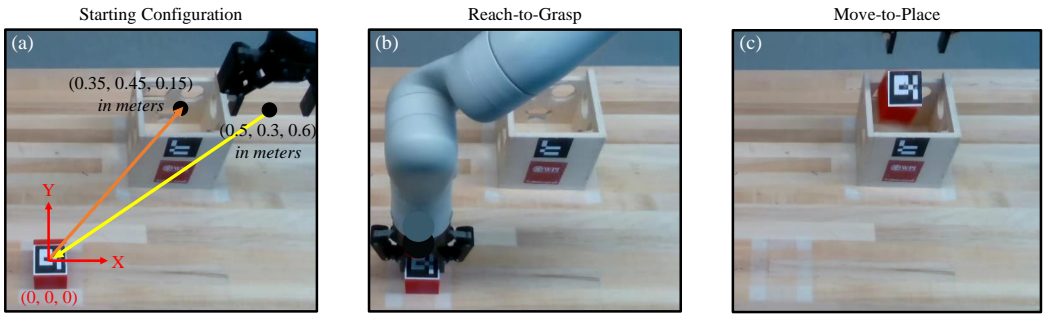


Fig. 12. Robot starting configuration and task.

Task. Our participants performed the same single-object pick-and-place task in all the trials. To focus on the comparison of different augmentation approaches, we simplify the task to be picking up and dropping a block object into the container, which does not involve the control of robot/object orientation. The robot, object and container were set to the same positions at the start of each trial (Figure 12.a). This task involves general-purpose manipulation actions and requires both gross and precise manipulation. Specifically, each task can be decomposed to four **action phases** (Figure 12.b and Figure 12.c), including: 1) **Reaching** to the object from the robot start position to within 0.3 m from the target object; 2) adjust the robot pose for **Grasping** the object; 3) **Moving** the grasped object to be 0.15 m from the container; 4) adjust the robot pose for **Placing** the object to the container. **Note that the tele-manipulation tasks could be more diverse and difficult than the single-object pick-and-place, our event-based robot autonomy and augmentation (perception and action) can still adapt to different purpose tasks (e.g., assist the object alignment in the stacking task or interact with multi-target workspace in correct order).**

Experiment Conditions. In each mode, a participant performed the task twice: 1) without any augmentation (Default), 2) with each perception augmentation (PA1 and PA2), 3) with each action augmentation (AA1 and AA2), and 4) using the preferred combination identified in the pilot study in Section 3.5. The total trials for each participant is $36 = 2 \text{ repetitions} \times 3 \text{ modes} \times (1 \text{ default} + 2 \text{ PAs} + 2 \text{ AAs} + 1 \text{ preferred integration})$. **To avoid the learning effects, participants performed in a random selection of one of the three mode orders mentioned above.**

Experiment Procedure. The experiment consists of a **training section** and **performing section**. In the training section, the experimenter explained and demonstrated how to use the default interface of selected starting mode, to perform the tele-manipulation task without any robot autonomy and interface augmentation for perception and action. The participants could practice the task (for maximally 15 min). The participants who felt confident to perform the task after practice

would perform the practiced task under aforementioned experiment conditions. **Every participant stated they felt confident in using the teleoperation interface within the allocated practice time.** We recorded the task performance (e.g., task completion time, types and occurrence of errors) during both the training and performing sections. Before the experiments, participants filled in a survey to report their experiences in video games, virtual reality environments, and spatial reasoning (via a spatial test from AssessmentDay [8]). Before the performing section, we asked the participants to look at the monitor for 30 seconds and recorded their pupil diameters, for the calibration required to estimate their cognitive workload. After completing the trials in each control mode, the participants filled in generic surveys, including NASA-TLX and System Usability Survey (SUS), and report their rating for each of the six interface conditions. After the experiment, the participants also filled in a customized questionnaire to report their preference for the control mode and interface conditions.

Data Analysis. Our data analysis considered *interface conditions* and *action phases* to be the independent variables, and considered the *task performance*, *workload*, and *user preference* to be the dependent variables. We measured the task performance objectively using the completion time, robot trajectory length, and types and occurrence of errors, for the entire task and for each action phase. We consider the physical and cognitive workload estimated using the methods in Section 3.6, and reported in NASA-TLX survey. We also consider the user's preference inferred from the gaze fixations and distributions on the interfaces and reported in the SUS and the customized surveys.

5 RESULTS

For all the comparisons, we analyzed data from all dependent variables using one-way repeated-measures analysis variance (ANOVA), including control modes, augmentations, action phases, and user groups, as a within-participants variable. All pairwise comparisons used Holm-Bonferroni correction to control for Type I error in multiple comparisons.

5.1 Effects of AR Visual Cues and Assistive Autonomy

From the comparison between different control modes (without any perception or action augmentations), we have the following results regarding the task performance, cognitive and physical workload. **As shown in Figure 13, we found that: 1) using autonomous actions can significantly reduce the occurrence of errors; 2) using AR visual cues can significantly reduce cognitive workload; 3) the overall preference of the participants for each mode was Mode 3 > Mode 2 > Mode 1.**

Task performance. The **task completion time** for the Mode 1 (baseline), Mode 2 (with AR visual cues) and Mode 3 (with autonomous actions) were on average 33.3 ± 7.7 , 30.6 ± 8.7 and 29.2 ± 4.3 seconds for all the participants. The participants completed the task faster (by 8% and 12%) with the assistive AR visual cues or autonomous actions than in Mode 1. The **total trajectory lengths** of the robot during the task for Mode 1, 2 and 3 are 1.99 ± 0.53 , 1.76 ± 0.53 and 1.78 ± 0.59 meters, respectively. The trajectory lengths were shorter in Mode 2 and 3 (by 12% and 11%) than in Mode 1. The **occurrence of errors** during the task were 4.86 ± 1.38 , 3.93 ± 0.91 and 0.15 ± 0.14 occurrences, for Mode 1, 2 and 3, respectively. **The ANOVA analysis showed no significant differences in the task completion time or the total trajectory lengths. However, post hoc comparisons showed that the occurrence of errors using Mode 3 (with autonomous actions) was significantly lower ($p < .01$) than using Mode 1 and 2, by 97% and 96%, respectively.**

Workload. The **physical workload** while using Mode 1, 2 and 3 were on average 49.3 ± 4.7 , 45.8 ± 1.2 and 49.7 ± 2.8 percent of the muscle capabilities. Mode 2 (with AR visual cues) led to lower physical workload than Mode 1 (baseline) and Mode 3 (with autonomous actions) but without significant differences. The **cognitive workloads** were on average 59.3 ± 11 , 50 ± 18.9 and $54.2 \pm$

10.9, respectively, when using Mode 1, 2 and 3. **Post hoc comparisons showed that Mode 2 (with AR visual cues) led to a significantly lower ($p < .05$) cognitive workload compared to the other Modes.**

Preference. Our post-experiment survey asked the participants "Overall, how much do you prefer to use this interface for controlling the robot on a daily basis for your work?". The participants rated their preference for each mode using the Likert scale from 1 (the least) to 5 (the most). Mode 3 was the most preferred (with the highest score 4.6 ± 0.8), while the scores for Mode 2 and 1 were 3.8 ± 1 and 2.1 ± 1.3 , respectively.

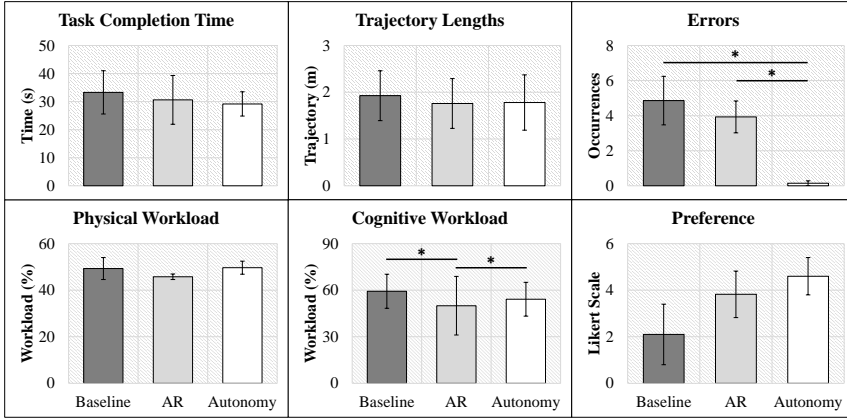


Fig. 13. Comparison of task performance, workload, and preference between control modes.

5.2 Effects of Various Perception and Action Augmentations

From the comparison between different augmentations (perception, action, and integration), we have the following results regarding the task performance and workload. Table 3 compares the performance and workload between the interfaces for each mode. The green (red) color indicates the best (worst) case among all the augmentation interfaces for each mode. We found that: 1) *using Fixed PIP (PA 1) can improve the performance of task completion time and total trajectory lengths*; 2) *using Trackpad (AA 1) can reduce the occurrence of errors*; 3) *using the Integrated interface can reduce the cognitive workload*. Table 6 compares the subjective feedback from NASA-TLX and SUS forms between the interfaces for each mode. We found that: 1) *using Trackpad (AA 1) result in higher mental and physical workload*; 2) *using Scaling (AA 2) result in higher overall workload and lower SUS score*. Moreover, the analysis of gaze fixation indicates that: 1) *participants tend to check on the PIP more for all perception augmentations in Mode 1*; 2) *the perception augmentations reduce the usage of the AR visual cues in Mode 2*.

5.2.1 Task Performance and Workload. Based on Table 3, we conducted multiple post hoc comparisons: 1) between the best and worst case for each mode, and 2) of the best and worst case across three modes (Figure 14). The significant differences we found include:

Task completion time. In Mode 1, the interface with *Fixed PIP* outperforms the interface with *Trackpad*, with $p < .05$; In Mode 2, the interface with *Pop-up PIP* outperforms the interface with *Trackpad*, with $p < .05$; In Mode 3, the interface with *Fixed PIP* outperforms the interface with *Trackpad*, with $p < .01$; **The ANOVA analysis showed no significant difference for the comparisons of the best and worst cases across three modes.**

Total trajectory length. The ANOVA analysis showed no significant difference for the comparisons between the best and worst interface in each mode as well as of the best and worst cases across three modes.

Occurrence of Errors. In Mode 1, the interface with *Trackpad* outperforms the interface with *Scaling*, with $p < .01$; In Mode 2, the interface with *Trackpad* outperforms the interface with *Fixed PIP*, with $p < .01$; In Mode 3, the *Default* interface outperforms the interface with Preferred Integration of PA and AA, with $p < .01$; For the best interfaces in each mode, the ANOVA analysis showed a significant difference between modes and the post hoc comparisons indicated the *Default* in Mode 3 outperforms the *Trackpad* in both Mode 1 and 2, with $p < .05$. For the worst interfaces in each mode, the ANOVA analysis showed a significant difference between modes and the post hoc comparisons indicated the *Integrated* in Mode 3 outperforms the *Scaling* in Mode 1 and *Fixed PIP* in Mode 2, with both $p < .05$.

Physical Workload. In Mode 2, the *Default* interface outperforms the interface with *Trackpad*, with $p < .05$; The ANOVA analysis showed no significant difference for the comparisons of the best and worst cases across three modes.

Cognitive Workload. In Mode 1, the interface with preferred *Integration* of PA and AA outperforms the *Default* interface, with $p < .01$; In Mode 3, the interface with preferred *Integration* of PA and AA outperforms the *Default* interface, with $p < .01$; For the worst interfaces in each mode, the ANOVA analysis showed a significant difference between modes and the post hoc comparisons indicated the *Default* in Mode 2 outperforms the *Default* in both Mode 1 and 3, with both $p < .05$.

Table 3. Task performance and overall workload for all interfaces with mean and standard deviation. The green (red) color indicates the best (worst) case among all the augmentation interfaces for each mode.

	Default (Single View)	Perception Augmentation (PA)		Action Augmentation (AA)		Integrated (PA + AA)
		<i>Fixed PIP</i>	<i>Pop-up PIP</i>	<i>Trackpad</i>	<i>Scaling</i>	
<i>Time (s)</i>						
Mode 1	33.3 (7.7)	28.4 (5.8)	28.8 (8.6)	36.6 (4.8)	34.5 (2.3)	36.2 (7.9)
Mode 2	30.6 (8.7)	31.4 (7.8)	28.9 (6)	40.6 (9.7)	36.5 (3.1)	34.6 (9.3)
Mode 3	29.2 (4.3)	27 (4)	28.8 (4.7)	36.7 (3.4)	32.1 (7.1)	31.7 (3.1)
<i>Trajectory (m)</i>						
Mode 1	1.99 (0.53)	1.75 (0.39)	1.75 (0.49)	1.74 (0.3)	1.83 (0.1)	1.76 (0.31)
Mode 2	1.76 (0.53)	1.58 (0.31)	1.62 (0.48)	1.73 (0.21)	1.6 (0.19)	1.61 (0.37)
Mode 3	1.78 (0.59)	1.56 (0.37)	1.64 (0.52)	1.59 (0.21)	1.64 (0.32)	1.65 (0.37)
<i>Error (num.)</i>						
Mode 1	4.86 (1.38)	5.3 (1.77)	4.61 (2.1)	3.14 (0.85)	5.63 (1.39)	3.7 (1.26)
Mode 2	3.93 (0.91)	4.39 (1.42)	4.29 (2.06)	1.93 (0.79)	3.96 (1.44)	2.01 (0.42)
Mode 3	0.15 (0.14)	0.54 (0.42)	1.27 (1.09)	1.78 (1.59)	1.59 (1.23)	2.33 (1.22)
<i>Physical (%)</i>						
Mode 1	49.3 (4.7)	48.2 (2.3)	48.6 (1.8)	51 (1.3)	49 (1.7)	50.6 (2.9)
Mode 2	45.8 (1.2)	48.6 (2.3)	51 (2.2)	52.5 (2.4)	50 (2.3)	51 (1.8)
Mode 3	49.7 (2.8)	50.5 (2.5)	51 (3.1)	51.1 (1.4)	50 (2)	48 (2.5)
<i>Cognitive (%)</i>						
Mode 1	59.3 (11)	50.9 (12.8)	51.5 (11.7)	55.5 (10.5)	55.3 (11.1)	47.6 (12.4)
Mode 2	50 (18.9)	46 (16.8)	45.1 (18.1)	47.8 (18.1)	49.3 (13)	48.7 (15.1)
Mode 3	54.2 (10.9)	48.9 (12.7)	45.4 (19.3)	50.2 (16)	48.2 (15.1)	43.3 (13.7)

*Mode 1: baseline | Mode 2: AR visual cues | Mode 3: assistive autonomy

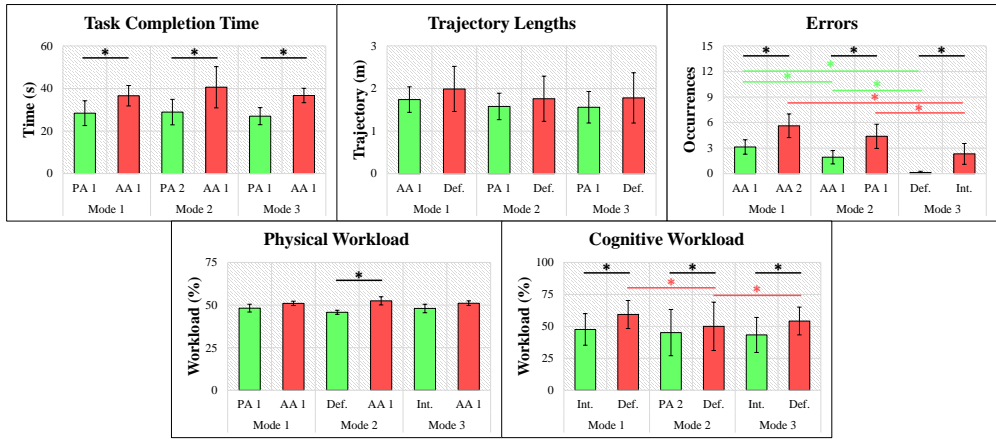


Fig. 14. Comparison of task performance and workload for augmentations.

5.2.2 Usage of Complementary View and AR Visual Cues. Table 4 shows the duration of the gaze fixation on the complementary viewpoint (measured by the percentage of task completion time). Note that we only count the gaze fixation longer than 0.1 sec on a particular interface feature. We found that with perception augmentations including PA 1 (Fixed PIP), PA 2 (Pop-up PIP), and Integrated, the time that participants spent looking at the complementary viewpoint is longer (yet not significantly) in Mode 1 than in the other two modes.

Table 4. Duration of gaze fixation on the complementary viewpoints (wr.r.t task completion time).

	PA 1	PA 2	Integrated
Mode 1	29 (16)	17 (9)	21 (13)
Mode 2	16 (13)	10 (9)	–
Mode 3	17 (16)	10 (11)	12 (15)

Table 5. Duration of gaze fixation on the AR visual cues (wr.r.t task completion time).

	AR Visual Cues				
	Object AR	Box AR	Height Bar	Hint Boxes	Distance
Default	33 (12)	26 (12)	2 (4)	0.8 (1)	0.3 (1)
PA 1	30 (12)	19 (11)	1 (3)	0.3 (0.9)	0.2 (0.6)
PA 2	32 (13)	22 (13)	0.2 (0.5)	0.2 (0.5)	0.8 (1)
AA 1	34 (12)	25 (13)	2 (3)	0.2 (0.6)	0.2 (0.6)
AA 2	31 (11)	28 (13)	0.7 (1)	0.5 (2)	0.3 (2)
Integrated	36 (12)	25 (13)	0.3 (1)	0.4 (2)	0.6 (2)

Table 5 compares the gaze fixation duration on the different AR visual cues (measured by the percentage of task completion time). It shows that the height bar, the hint boxes and distance box were the least used, because the participants only glanced at them to find out the action affordance and confirmation. On the other hand, the object and box AR features were much more used. This is because the participants need to look at them to control the continuously performed reaching and moving motions, and to precisely adjust the robot position for grasping or placements. Table 5 also shows that the perception and action augmentations may change the participants' reliance

and usage of the AR visual cues. **Post hoc comparisons showed that PA 1 (Fixed PIP) significantly reduced the use of the “object” AR visual cue ($p < 0.05$), compared to AA 1 (Trackpad) or the integrated interface.** However, the perception and action augmentations have no significant impacts on the use of the “box” cue.

Table 6. NASA-TLX and SUS subjective feedback. The green (red) color indicates the best (worst) case among all the augmentation interfaces for each mode.

	Default (Single View)	Perception Augmentation (PA)		Action Augmentation (AA)		Integrated (PA + AA)
		Fixed PIP	Pop-up PIP	Trackpad	Scaling	
<i>Mental Demand (NASA-TLX)</i>						
Mode 1	41 (14)	34 (2)	33 (3)	55 (11)	54 (8)	53 (17)
Mode 2	31 (2)	33 (3)	32 (10)	49 (17)	52 (13)	48 (12)
Mode 3	24 (2)	25 (3)	27 (5)	58 (11)	47 (5)	42 (10)
<i>Physical Demand (NASA-TLX)</i>						
Mode 1	39 (4)	33 (11)	32 (11)	48 (10)	41 (8)	42 (11)
Mode 2	28 (11)	30 (11)	30 (12)	51 (7)	42 (15)	42 (15)
Mode 3	40 (9)	25 (4)	27 (4)	45 (11)	40 (11)	23 (5)
<i>Overall Workload (NASA-TLX)</i>						
Mode 1	39 (9)	33 (3)	31 (4)	50 (11)	54 (7)	48 (7)
Mode 2	30 (3)	32 (2)	32 (7)	48 (6)	53 (10)	45 (7)
Mode 3	42 (12)	26 (3)	28 (4)	50 (13)	46 (8)	24 (4)
<i>SUS</i>						
Mode 1	73 (2)	81 (3)	82 (2)	51 (13)	49 (7)	55 (12)
Mode 2	81 (1)	76 (4)	76 (3)	57 (14)	56 (10)	61 (15)
Mode 3	84 (3)	81 (2)	77 (5)	50 (10)	61 (6)	65 (10)

*Mode 1: baseline | Mode 2: AR visual cues | Mode 3: assistive autonomy

5.2.3 Subjective Feedback. Table 6 compares the reported mental and physical workload from the NASA-TLX survey between the augmentation interfaces for each mode, and the reported usability from the SUS. We also compare the overall NASA-TLX score, using the coefficients of: 5 for mental demand, 4 for physical demand, 0 for temporal demand, 2 for performance, 3 for effort, and 1 for frustration. The weighting coefficients were generated by choosing from a series of pairs of rating scale factors that were deemed to be important based on the official instructions. Similar to Table 3, the green (red) color indicates the best (worst) case for each mode. **We conducted multiple post hoc comparisons between the best and worst case for each mode.** Here are the significant differences we found from the comparisons:

Mental Demand. In Mode 1, the interface with PA 1 (*Fixed PIP*) significantly outperforms the interface with AA 1 (*Trackpad*), with $p < .01$; In Mode 2, the *Default* interface significantly outperforms the interface with AA 2 (*Scaling*), with $p < .05$; In Mode 3, the *Default* interface significantly outperforms the interface with AA 1 (*Trackpad*), with $p < .01$.

Physical Demand. In Mode 2, the *Default* interface significantly outperforms the interface with AA 1 (*Trackpad*), with $p < .05$; In Mode 3, the *Integrated* Interface significantly outperforms the interface with AA 1 (*Trackpad*), with $p < .01$;

Overall Workload. In Mode 1, the interface with PA 2 (*Pop-up PIP*) significantly outperforms the interface with AA 2 (*Scaling*), with $p < .01$; In Mode 2, the *Default* interface significantly

outperform the interface with AA 2 (*Scaling*), with $p < .01$; In Mode 3, the *integrated* interface significantly outperforms the interface with AA 1 (*Trackpad*), with $p < .05$;

SUS. In Mode 1, the interface with PA 2 (*Pop-up PIP*) significantly outperforms the interface with AA 2 (*Scaling*), with $p < .01$; In Mode 2, the *Default* interface significantly outperforms the interface with AA 2 (*Scaling*), with $p < .05$; In Mode 3, the *Default* interface significantly outperforms the interface with AA 1 (*Trackpad*), with $p < .01$;

5.3 Effects on Different Action Phases

We further analysis the interface modes and perception/action augmentation on the performance and workload for the different action phases defined in Section 4. **From the comparison between different action phases in each mode, we have the following results regarding the task performance, cognitive and physical workload. As shown in Figure 15, we averaged the data from all augmentation interfaces (default, PAs, AAs, integrated) for each action phase (i.e., reaching, grasping, moving, and placing) in each mode (i.e., baseline, AR visual cues, and assistive autonomy). The ANOVA analysis and multiple post hoc comparisons showed that: 1) the action phase of grasping takes a significantly ($p < .01$) longer time than the reaching phase in all modes; 2) the action phase of placing results in a significantly ($p < .01$) higher physical workload in all modes; 3) the action phase of grasping and placing results in a significantly ($p < .05$) higher cognitive workload than the reaching and moving phases respectively in all modes.**

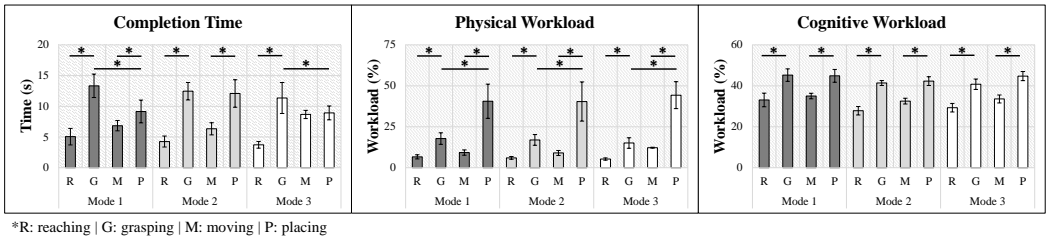


Fig. 15. Comparison of completion time and workload between action phases for each mode.

Table 7. Time of each action phase for all interfaces.

	Default	PA1	PA2	AA1	AA2	Integrated
Mode 1: Baseline						
Reaching	7.2 (3.4)	5.5 (1.3)	6.1 (1.8)	4.4 (1.1)	3.4 (0.5)	3.7 (0.6)
Grasping	14.2 (3.4)	11.5 (2.5)	10.6 (3.8)	16.2 (1.1)	12.8 (2.6)	14.7 (2.3)
Moving	6.6 (2.8)	6.4 (2.7)	5.7 (1.8)	7.5 (3)	6.5 (1.4)	8.3 (3.6)
Placing	7.8 (0.6)	7.2 (1.6)	7.8 (2.7)	9.2 (2.1)	12.6 (4.3)	10.3 (4.5)
Mode 2: AR Visual Cues						
Reaching	5.5 (1.4)	4.8 (0.7)	3.9 (0.9)	4.9 (0.2)	2.9 (0.4)	3.5 (0.9)
Grasping	9.7 (2.1)	12.6 (3.5)	11.9 (3.5)	14.3 (2.8)	13.2 (1.3)	13 (3.2)
Moving	6.1 (2.5)	5 (0.5)	5.2 (1)	7.6 (2.9)	7.3 (1.8)	6.8 (2.4)
Placing	10.7 (5.6)	10.4 (5.9)	9.1 (2.9)	15.5 (7.3)	14.2 (2.3)	12.6 (6.3)
Mode 3: Assistive Autonomy						
Reaching	4.9 (1.1)	3.4 (0.3)	3.6 (0.8)	3.6 (0.2)	3.6 (0.5)	3.2 (0.4)
Grasping	9.5 (2.4)	8.6 (1.6)	9.6 (1.7)	16.1 (2.2)	11.7 (4)	12.6 (2.5)
Moving	7.6 (1.1)	8.7 (2.2)	9.3 (3.3)	8.2 (1.6)	9.6 (4.1)	8.6 (1.6)
Placing	8.6 (1.2)	8 (1.9)	7.1 (1.3)	10.4 (2.1)	9.7 (2.2)	9.7 (1.1)

*PA: perception augmentation | AA: action augmentation

Task completion time. Table 7 shows the task completion time of each action phase for all the interfaces of all the 3 modes. Post hoc comparisons showed that: The *grasping* action takes significantly longer time than the *reaching* action, for all interfaces and all the modes, with $p < .01$. The *placing* action also takes significantly less time than the *moving* action for all the interfaces of Mode 1 and 2, but not for Mode 3, with $p < .01$. We noticed that the interface that takes the least time for the **grasping** action is different for each mode. Specifically, it is the interface with PA 2 (*Pop-up PIP*) in Mode 1, the *Default* interface in Mode 2, and the interface with PA 1 (*Fixed PIP*) in Mode 3. The interface with PA 1 (*Fixed PIP*) in Mode 3 takes the least time to grasp, across all the modes and interfaces. We also noticed that the interface that takes the least time for the **placing** action is different for each mode. Specifically, it is the interface with PA 1 (*Fixed PIP*) in Mode 1, the interface with PA 2 (*scaling*) for Mode 2 and 3. The interface with PA 2 in Mode 2 takes the least time to place, across all the modes and interfaces. Note that the interfaces that take the least time for the grasping and placing action are highlighted in green in Table 7.

Physical Workload. Table 8 shows the physical workload of each action phase for all the interfaces of all the 3 modes. Post hoc comparisons showed that: The *grasping* action has significantly higher physical workload than the *reaching* action, for all interfaces and all the modes, with $p < .05$. The *placing* action also has significantly higher physical workload than the *moving* action for all the interfaces of all the modes, with $p < .01$. We noticed that the interface that has the least physical workload for the **grasping** action is different for each mode. Specifically, it is the interface with PA 2 (*Pop-up PIP*) in Mode 1, the *Default* interface in Mode 2, and the interface with PA 1 (*Fixed PIP*) in Mode 3. The interface with PA 1 (*Fixed PIP*) in Mode 3 takes the least time to grasp, across all the modes and interfaces. However, the interface that has the least physical workload for the **placing** action is the *Default* interface for all the three modes. The *Default* interface in Mode 2 has the least physical workload to place, across all the modes and interfaces. Note that the interfaces that has the least physical workload for the grasping and placing action are highlighted in green in Table 8.

Table 8. Physical workload of each action for all interfaces.

	Default	PA1	PA2	AA1	AA2	Integrated
Mode 1: Baseline						
Reaching	8.5 (0.6)	7 (0.6)	7.9 (0.6)	5.9 (0.6)	4.9 (0.5)	5.5 (0.1)
Grasping	16.7 (0.9)	15 (1.4)	13.3 (1.4)	23.9 (3.6)	17.2 (3.7)	20.6 (3.1)
Moving	8.4 (1.2)	7.7 (0.8)	7.4 (1)	10.8 (2.3)	9 (0.6)	11.9 (2.6)
Placing	29.4 (4.9)	33 (5.2)	36.7 (2.4)	34.6 (2.8)	57.1 (5.1)	52.7 (2.8)
Mode 2: AR Visual Cues						
Reaching	6.4 (0.3)	6.5 (0.6)	5.8 (0.9)	7.3 (1.2)	4.8 (1.3)	5 (1)
Grasping	11.9 (1.3)	15.7 (0.2)	13.9 (1.2)	20.2 (1.3)	20 (2.1)	20 (3.7)
Moving	7.3 (1.1)	7.4 (0.7)	8 (0.7)	11.3 (2.5)	10.1 (0.4)	9.8 (1.3)
Placing	17.4 (1.4)	35.4 (5.3)	42 (4.8)	42.6 (2.9)	53.7 (1.4)	51.3 (2.8)
Mode 3: Assistive Autonomy						
Reaching	6.3 (0.6)	5.1 (0.4)	4.6 (0.6)	6.1 (1.6)	5.2 (0.9)	4.3 (0.7)
Grasping	12.8 (1.1)	11.9 (2)	12.8 (1.3)	21.4 (1.5)	15.6 (2.4)	15.9 (1.4)
Moving	12 (0.4)	11.9 (0.4)	12.5 (1.3)	11.9 (0.1)	12.5 (2)	12.4 (1.1)
Placing	32.6 (2)	38.9 (6.1)	40.8 (4.6)	43.6 (3.5)	54.8 (5.7)	55 (5.1)

*PA: perception augmentation | AA: action augmentation

Cognitive Workload. Table 9 shows the cognitive workload of each action phase for all the interfaces of all the 3 modes. Post hoc comparisons showed that: The *grasping* action has significantly higher cognitive workload than the *reaching* action, for all interfaces and all the modes, with $p < .01$.

The *placing* action also has significantly higher cognitive workload than the *moving* action for all the interfaces of all the modes, with $p < .01$. We noticed that the interface that has the least cognitive workload for the **grasping** action is different for each mode. Specifically, it is the integrated interface in Mode 1, the AA 2 (scaling) interface in Mode 2, and the integrated interface in Mode 3. The integrated interface in Mode 3 takes the least cognitive workload to grasp, across all the modes and interfaces. In terms of *placing*, PA 2 (pop-up PIP) caused the least cognitive workload for Mode 1 and 2, and integrated interface caused the least cognitive workload for Mode 3. The PA 2 (pop-up PIP) interface in Mode 2 caused the least cognitive workload to place, across all the modes and interfaces. Note that the interfaces that has the least cognitive workload for the grasping and placing action are highlighted in green in Table 9.

Table 9. Cognitive workload of each action for all interfaces.

	Default	PA1	PA2	AA1	AA2	Integrated
Mode 1: Baseline						
Reaching	37 (8.7)	34.7 (9.6)	37 (8.3)	31.3 (8.4)	29.7 (6.5)	28.7 (8.5)
Grasping	49.7 (7)	43.3 (7.6)	42.5 (9)	49 (7.9)	44.8 (8.4)	42.1 (8.4)
Moving	37.6 (10.6)	34.3 (10.4)	35.7 (7.9)	35 (7.4)	34.1 (9.2)	33.4 (10.7)
Placing	48 (9.8)	42.9 (9.9)	41.7 (8.2)	45.1 (8.7)	49.8 (9.8)	41.7 (8.6)
Mode 2: AR Visual Cues						
Reaching	31.3 (12.2)	29.2 (11.4)	25.4 (13.3)	28.1 (10.7)	25.5 (10)	27.4 (8.1)
Grasping	42.2 (12.5)	41.6 (10.3)	39.9 (10.4)	41.3 (11.3)	39.9 (9.3)	43.3 (10.7)
Moving	33.3 (11.1)	30.5 (11.7)	31.8 (11.2)	32.4 (12.3)	35 (8.1)	32.1 (10)
Placing	43.2 (11.2)	39.9 (11.4)	39.6 (11.3)	42.1 (12.2)	45.9 (9.2)	43 (10.7)
Mode 3: Assistive Autonomy						
Reaching	33.3 (8)	29 (8.8)	28.6 (11.8)	30.2 (9.7)	28 (8.4)	26.6 (8.2)
Grasping	44 (7.3)	40.3 (8.6)	38 (10.6)	44.1 (9.5)	40.5 (8.3)	37.7 (8.4)
Moving	37.2 (7.6)	34.4 (9.1)	33.6 (12.6)	31.7 (10.1)	33.1 (10.6)	31.4 (10.1)
Placing	47.7 (8.3)	45.7 (9.1)	42.4 (12.5)	46 (10.7)	45.4 (11.3)	41.2 (10.3)

*PA: perception augmentation | AA: action augmentation

5.4 Effects of Other Human Factors

We further analyze effects of several human factors, including gender, background and experience with technology, on the performance and usage of the interfaces. The participants were divided into groups based on conditions as defined in Section 4. From the comparison between different user groups, we have the following results regarding the task performance in terms of the completion time. As shown in Figure 16, the ANOVA analysis and multiple post hoc comparisons showed that: 1) *users' background impacts the task performance with more augmentation interfaces and control modes*; 2) *using assistive autonomy can mitigate the gap between different user groups in task performance*.

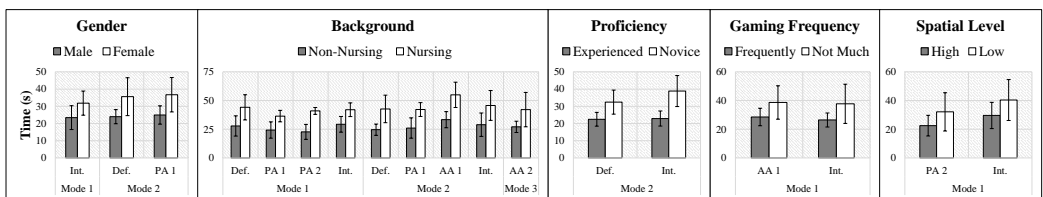


Fig. 16. Indication of the significant differences in the comparison of completion time between user groups for each mode.

Table 10 compares the task completion time between user groups of different gender, background, proficiency with the tele-manipulation interface, gaming frequency, spatial reasoning level and mode order. We highlighted the identified significant differences, which indicate that the male, non-nursing, experienced users, frequent video game players, and high-level spatial reasoning users took less time to complete the task.

Table 10. Comparison between groups: completion time.

	Mode 1						Mode 2						Mode 3					
	Def.	PA1	PA2	AA1	AA2	Int.	Def.	PA1	PA2	AA1	AA2	Int.	Def.	PA1	PA2	AA1	AA2	Int.
G	.53	.30	.09	.40	.35	< .05	< .05	< .05	.31	.25	.50	.12	.45	.33	.37	.31	.28	.87
B	< .05	< .05	< .01	.14	.50	< .01	< .01	< .05	.10	< .05	.32	< .05	.06	.07	.09	.15	< .05	.19
P	.50	.19	.09	.08	.56	.08	< .05	.06	.06	.07	.91	< .01	.09	.06	.07	.11	.23	.55
F	.63	.17	.10	< .05	.85	< .05	.11	.15	.61	.25	.14	.15	.24	.08	.27	.18	.28	.17
S	.35	.36	< .05	.17	.52	< .05	.13	.10	.36	.06	.47	.38	.14	.32	.16	.92	.21	.87
M	.97	.37	.28	.57	.83	.73	.64	.89	.75	.38	.79	.62	.67	.52	.95	.98	.48	.69

*PA: perception augmentation | AA: action augmentation

*G: gender | B: background | P: proficiency | F: gaming frequency | S: spatial reasoning level | M: mode order

Table 11. Comparison between user groups: use of complementary view.

	Mode 1		Mode 2		Mode 3		Integrated	Integrated
	Fixed	Pop-up	Fixed	Pop-up	Fixed	Pop-up	(Mode 1)	(Mode 3)
Gender	< .05	.08	.21	.14	.12	.13	< .05	.47
Background	< .05	< .01	.46	.90	.68	.10	.06	.21
Proficiency	.58	.51	.36	.96	.89	.65	.43	.78
Gaming	.71	.63	.80	.45	.60	.64	.80	.22
Spatial	.17	.21	.21	.14	.41	.10	< .01	< .05
Mode Order	.35	.22	.12	.09	.33	.60	.82	.08

Table 11 compares the use of the complementary viewpoint between user groups, using the duration of gaze fixation on the complementary viewpoint with respect to the total task completion time. We found significant differences between the male and female user groups, when using the interface with PA 1 (*Fixed PIP*) and using the *integrated* interface in Mode 1. For both interfaces, the male participants used the complementary viewpoints more than the female participants by 13% and 8.7% on average, respectively. We also noticed that: in Mode 1, both PA 1 and PA 2 led to significant difference in the use of complementary viewpoint between the users with and without nursing professional experience, with $p < .05$ and $p < .01$, respectively. Participants without nursing experience or training used the complementary viewpoint 13.4% more on average for the interface with PA 1 (*Fixed PIP*), and 9.8% more on average for the interface with PA 2 (*Pop-up PIP*). Moreover, we found significant difference between user groups of different spatial reasoning skills in the use of complementary viewpoint in Mode 1 and Mode 3 when using the integrated interface, with $p < .01$ and $p < .05$. In Mode 1 and Mode 3, when using the integrated interface, participants of lower spatial reasoning skills use the complementary viewpoint more by 10.3% and 9.2% on average.

Table 12 highlights the significant difference between user groups by the use of the “object” and “box” AR visual cues. In Mode 2, we found significant differences ($p < .05$) between users of different **background** in the use of the “object” cue when using the interface with AA 1 (*Trackpad*), and in the use of the “box” cue when using the *Default* interface. Specifically, the users without nursing experience used the “object” cue more by 8% when using the interface with AA 1, while the users with nursing experience use the “box” cue more by 10.3% when using the *Default* interface. Regarding the factor of **proficiency**, we found that: in Mode 2, the participants with

prior experience of robot teleoperation used the “object” cue significantly more (with $p < .05$) than the other group by 7.9%, when using the interface with PA 1 (*Fixed PIP*). We also found that the high-proficiency group used the “object” cue significantly more (with $p < .05$ and $p < .01$) than the other group by 8.8% and 9.0%, when using the interface with AA 1 (*trackpad*) and AA 2 (*scaling*). Regarding the effects of **gaming experience**, we found that frequent video game players used the “object” cue significantly more ($p < .01$) than the other user group when using the interface with AA 1 (*Trackpad*), by 10.3% on average. We also found that frequent video game players used the “object” cue significantly more ($p < .05$) than the other user group when using the *integrated* interface, by 7.3% on average. Regarding the factor of **spatial reasoning skills**, we found that the participants of better spatial reasoning skills used the “object” cue significantly more ($p < .05$) than the other users when using the interface with AA 1 (*Trackpad*), by 7.0% on average. We also found that the participants of better spatial reasoning skills used the “box” cue significantly less ($p < .05$) than the other users when using the interface with AA 1 (*Trackpad*), by 9.1% on average). Regarding the effects of **mode orders**, we found that the users performed the mode in $3 \rightarrow 1 \rightarrow 2$ order used the “box” cue significantly more ($p < .05$) than in $2 \rightarrow 3 \rightarrow 1$ order when using the *Default* interface, by 12.3% on average. We also found that the participants used the mode in $1 \rightarrow 2 \rightarrow 3$ order used the “box” cue significantly more ($p < .05$) than in $2 \rightarrow 3 \rightarrow 1$ order when using the interface with PA 1 (*Fixed PIP*), by 10.8% on average.



Table 12. Comparison between user groups: use of object and box AR visual cues.

	Def.	PA1	PA2	AA1	AA2	Int.	Def.	PA1	PA2	AA1	AA2	Int.
	<i>Object AR</i>						<i>Box AR</i>					
G	.18	.09	.56	.55	.17	.15	.16	.29	.21	.66	.99	.56
B	.42	.64	.98	< .05	.05	.23	< .05	.45	.42	.19	.95	.44
P	.08	< .05	.15	< .05	< .01	.40	.59	.99	.23	.59	.73	.35
F	.74	.17	.86	< .01	.22	< .05	.67	.92	.91	.69	.12	.85
S	.82	.68	.37	< .05	.07	.33	.50	.73	.77	< .05	.75	.76
M	.63	.85	.13	.64	.62	.34	< .05	< .05	.09	.53	.96	.34

*G: gender | B: background | P: proficiency | F: frequency of gaming | S: spatial | M: mode order

Table 13. Comparison between user groups: use of height bar, hint boxes and distance AR visual cues.

	Def.	PA1	PA2	AA1	AA2	Int.	Def.	PA1	PA2	AA1	AA2	Int.	Def.	PA1	PA2	AA1	AA2	Int.
	<i>Height Bar</i>						<i>Hint Boxes</i>						<i>Distance</i>					
G																		
B																		
P																		
F																		
S																		
M																		

*Highlighted: male (G); nurse (B); novice (P); low gaming (F); bad spatial (S); Order 1  and Order 2  (M)

Remarks: Table 13 shows the trend in the use of the height bar, hint box and distance between user groups. The table presents a comparison of the total number of frames for each AR cue used while performing the task across all the participants in the user group. We highlighted the user group that used these features more, measured by the total number of frames.

6 DISCUSSION

6.1 Summary of Novelty and Contributions

Our prior work developed shared autonomy to assist the remote perception and action control in freeform tele-manipulation. Specifically, we have leveraged AR visual cues [59], interface mapping [93] and autonomous actions for precise manipulation (e.g., grasping) [60], to effectively

reduce the human's workload while improving the control efficiency. The work in this paper further proposed new approaches for perception and action augmentation, to achieve more efficient and effortless freeform tele-manipulation motion control. **Moreover, we proposed a novel method for objective physical and cognitive workload estimation based on human motion and eye tracking devices, to accurately evaluate human comfort while performing remote manipulation.** To address the research questions (listed in Section 4), we conducted a comprehensive user study to evaluate and compare various integration of tele-manipulation assistance, and discovered new knowledge about their impacts on performance, workload and the preference of the users groups that differ in various human factors. The main findings include:

- (1) Based on the comparison between different control modes (direct manual, AR visual cues, and assistive autonomy without any augmentations), using **AR visual cues** can significantly reduce cognitive workload, while using **assistive autonomy** can significantly reduce the occurrence of errors.
- (2) Based on the comparison between different augmentations for all control modes, using **Fixed PIP (PA 1)** can improve the performance of completion time and trajectory lengths, using **Trackpad (AA 1)** can reduce the occurrence of errors but result in higher mental and physical workload, while using the **Integrated (preferred PA+AA)** can reduce the cognitive workload.
- (3) Based on the comparison between different action phases (reaching, grasping, moving, and placing) in the object pick-and-place task, using **perception augmentations** can reduce the completion time for grasping action (takes the longest time), and using **Integrated (preferred PA+AA)** can reduce the cognitive workload for both grasping and placing actions (with higher workload than reaching and moving).
- (4) Based on the comparison between user groups, using **assistive autonomy control mode with perception augmentations and Trackpad** can mitigate the gap in task performance between different user groups.

6.2 The Effective Integration of Perception and Action Assistance is Task-Dependent

Related work in literature has developed various augmented reality interfaces and shared autonomy for action supports, in order to assist remote perception and robot motion control [6, 15, 95], yet there is still no comprehensive comparisons to evaluate their individual and integrated impacts on task performance, workload and user preference. To fill this gap, we first conducted a user study to compare interfaces that only have the perception assistance (of AR visual cues) *or* the action assistance (of autonomous actions), to the baseline interface without any assistance. Our results show that both the perception and action assistance can effectively reduce the cognitive workload. Moreover, the assistive action can effectively reduce the occurrence of errors.

We further evaluated to what extent the additional perception and augmentations proposed in this work can further assist the tele-manipulation. Our results show that the effectiveness of perception and action assistance (and their integration) depend on the task performance objective. Specifically, we found that: for the tasks that need to be completed as fast as possible, the integration of Fixed PIP display and autonomous actions is the most effective, because it led to the least task completion time and motion efficiency (measured by the total robot trajectory length). As the participants commented: “... *the complementary viewpoint helped me to clearly understand the relationship between robot and target so that can move faster in the right direction to active the autonomy feature*”. For the task that emphasizes the reliability and precision of motion control, the interface that only provides autonomous actions for assistance turns out to be most effective, because the operators can focus more on the use of autonomous actions and will not be distracted by other interface

augmentation designed to enhance task efficiency or to reduce the workload. However, when the human's workload and comfort is prioritized in the tele-manipulation, it is more effective to provide only the AR visual cues and assistive autonomy integrated with Fixed PIP interfaces, as they can effectively reduce human's physical and cognitive workload, respectively. **These findings suggest the extension of the task- or goal-dependent perception and augment assistance design, which will be intelligent to not only provide suitable augmentation but activate the assistance based on the online estimation of human comfort.** From the user preference and SUS, we also found that, humans strongly prefer to use the autonomy to assist their remote perception and action, but *only if they are reliable*. Most of the participants commented that: "*if the robot autonomy is reliable, I would like to use it on daily bases*". Particularly, the nursing participants "*... would like the robot to be as autonomous as possible because we do not have more bandwidth to control the robot during nursing duty*". **However, the robot autonomy may not be consistently reliable due to the perception and action uncertainty of the robots, as well as the complexity of the manipulation tasks. It is still unclear how to adjust the level and type of robot assistance if the reliability of the robot autonomy may vary. Our future work therefore will further investigate the impacts of unreliable autonomy in human-robot collaboration for robot remote manipulation.**

6.3 Assist-as-Needed to Different Tasks and User Groups

Effects on Different Action Phases. The work in this paper also compared the effects of tele-manipulation assistance on different action phases during the tasks. While our comparison is limited to the action phases of a general-purpose pick-and-place task, the findings may still imply a design guideline generally applicable to freeform tele-manipulation. The pick-and-place task we performed has four action phases, namely, the reaching, grasping, moving, and placing. Our results show that: the grasping and placing, which require more precise motion control, took significantly longer time for human control and need more perception and action assistance. Most participants found: "*... it was a relief that the robot can take over the grasping and placing part of the task*". We also found that the precise manipulation actions cause significantly higher physical and cognitive workload than gross manipulation (i.e., reaching and moving). Especially, the physical workload for the placing action is also significantly higher than the grasping action. Our results show that the most effective interface to assist the *grasping* action needs to integrate the AR visual cues, fixed PIP display, and assistive autonomous actions, while the most effective interface to assist the *placing* action only need the AR visual cues. In terms of cognitive workload, the interface with assistive autonomy integrated with fixed PIP and scaling motion significantly reduced the workload while grasping. Moreover, the interface with AR visual cues, and pop-up PIP display of the complementary viewpoint can significantly reduce the cognitive workload of the placing action. The participants in the post-study survey commented that: "*it will be great if the assistive autonomy can integrate the certain AR visual cues to gain more information*", which is similar to the design we developed in the integrated interface for Mode 3. **This insight suggests the extension of the research to investigate what type of AR visual cues should be displayed to enable humans to seamlessly utilize various levels of robot autonomy for remote manipulation.**

Effects on Different User Groups. We also find significant differences regarding the impacts of perception and action assistance on different user groups. Regarding the *task performance*, we found that: while the assistive autonomous actions can effectively reduce the task completion time for almost all the comparisons between user groups, the AR visual cues are only significantly more effectively for some user groups (of specific gender, background and proficiency). We also found that the female and nursing participants were less capable to use the complementary viewpoints when the interface does not provide AR visual cues or assistive autonomous actions. The action

augmentations proposed in this work, including the “Trackpad” and “Scaling” to improve the efficiency and accuracy of motion control, do not effectively reduce the task completions time for most of the users. However, when these action augmentations were provided, we found significant differences between user groups in their use of the AR visual cues during grasping. This finding implies that for some user groups, the action augmentation only reduced their task completion time to limited extent (for the task we studied), because they can better leverage the AR visual cues. We also found that significant differences from most of the between-user-group comparisons (in the task performance and in the use of the the complementary viewpoint), when participants used interface that the integrated some perception and action augmentation proposed in this work (i.e., pop-up PIP complementary viewpoint display and the Trackpad control, as preferred by the pilot study user). This finding implies, when the robot autonomy like the AR visual cues and assistive autonomous action are not reliable, some user groups can benefits from some perception or action augmentation. Our future work therefore will investigate how to adjust the integration of the perception and action assistance for different users.

6.4 Limitations

We are aware of several limitations of the work in this paper. Our user study can be solidified by recruiting more participants of nursing profession (e.g., nursing students, faculty, and practitioners), and consider their professional experiences (e.g., grade of students, teaching years as faculty, hours of nursing services) as the factors in our data analysis. More end-users as participants will facilitate the human-in-the-loop design ranging from the relevant task setup to robot assistance development (e.g., testing of the feasibility of the nursing tasks [57]). We will further collaborate with the hospital and nursing school to not only work with the nursing profession but deploy the robot to work in a (simulated) hospital room.

The task in our experimental study is limited to a general-purpose pick-and-place task with four simple manipulation actions. Even though our robot autonomy and augmentation (perception and action) might adapt to different purpose tasks, it is unclear if our findings can be generalized to other manipulation actions, tasks, and specialized manipulation for nursing assistance tasks. The tele-manipulation task is also limited because it only involved freeform position control. It is unclear whether the proposed perception and action assistance can be effective in manipulation tasks that emphasize orientation control or 3D object pose control. A natural extension of our implementation is to further map the angular velocity as an input to control the robot's orientation. Such an approach could enable robots to perform a wider range of the remote manipulation tasks such as tracking the path, aligning, and stacking. Future work could implement similar perception and action assistance proposed in this paper for assisting the orientation control.

7 CONCLUSION

This paper conducted a comprehensive evaluation of individual and integrated approaches to assist the remote perception and motion, when humans control a 7 DOF manipulator (using natural hand motions) to perform a single-object pick-and-place task. We analyzed the performance, workload and human preference when using the interfaces that integrated different perception and action assistance, and discovered new knowledge about how to effectively provide task- and use-dependent tele-manipulation assistance.

Specifically, we have discovered the effective integration of perception assistance (e.g., AR visual cues, display of complementary viewpoint) and action assistance (e.g., assistive autonomous actions, directional precision motion control using trackpad and larger motion mapping scaling) vary with the task performance objectives, manipulation actions, and the human factors of the user groups.

REFERENCES

- [1] [n.d.]. Da Vinci Surgeon Console. <https://www.intuitive.com/en-us/products-and-services/da-vinci/systems>. Accessed: 2022-04-28.
- [2] David C Ackland, Sasha Roshan-Zamir, Martin Richardson, and Marcus G Pandey. 2011. Muscle and joint-contact loading at the glenohumeral joint after reverse total shoulder arthroplasty. *Journal of Orthopaedic Research* 29, 12 (2011), 1850–1858.
- [3] Henny Admoni and Siddhartha Srinivasa. 2016. Predicting user intent through eye gaze for shared autonomy. In *2016 AAAI Fall Symposium Series*.
- [4] Marco Aggravi, Daniel AL Estima, Alexandre Krupa, Sarthak Misra, and Claudio Pacchierotti. 2021. Haptic teleoperation of flexible needles combining 3D ultrasound guidance and needle tip force feedback. *IEEE Robotics and Automation Letters* 6, 3 (2021), 4859–4866.
- [5] Luis Almeida, Paulo Menezes, and Jorge Dias. 2020. Interface transparency issues in teleoperation. *Applied Sciences* 10, 18 (2020), 6232.
- [6] Stephanie Arévalo Arboleda, Tim Dierks, Franziska Rücker, and Jens Gerken. 2021. Exploring the visual space to improve depth perception in robot teleoperation using augmented reality: the role of distance and target’s pose in time, success, and certainty. In *IFIP Conference on Human-Computer Interaction*. Springer, 522–543.
- [7] Reuben M Aronson, Thiago Santini, Thomas C Kübler, Enkelejd Kasneci, Siddhartha Srinivasa, and Henny Admoni. 2018. Eye-hand behavior in human-robot shared manipulation. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 4–13.
- [8] assessmentday. 2022. *Spatial Reasoning Free Test*. <https://www.assessmentday.com/free/spatial-reasoning/>
- [9] Jenay M Beer, Arthur D Fisk, and Wendy A Rogers. 2014. Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of human-robot interaction* 3, 2 (2014), 74.
- [10] Hadi Beik-Mohammadi, Matthias Kerzel, Benedikt Pleintinger, Thomas Hulin, Philipp Reisich, Annika Schmidt, Aaron Pereira, Stefan Wermter, and Neal Y Lii. 2020. Model mediated teleoperation with a hand-arm exoskeleton in long time delays using reinforcement learning. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 713–720.
- [11] Stephen Bier, Rui Li, and Weitian Wang. 2020. A full-dimensional robot teleoperation platform. In *2020 11th International Conference on Mechanical and Aerospace Engineering (ICMAE)*. IEEE, 186–191.
- [12] Henri Boessenkool, David A Abbink, Cock JM Heemskerk, Frans CT van der Helm, and Jeroen GW Wildenbeest. 2012. A task-specific analysis of the benefit of haptic shared control during telemanipulation. *IEEE Transactions on Haptics* 6, 1 (2012), 2–12.
- [13] Craig E Boettcher, Karen A Ginn, and Ian Cathers. 2008. Standard maximum isometric voluntary contraction tests for normalizing shoulder muscle EMG. *Journal of orthopaedic research* 26, 12 (2008), 1591–1597.
- [14] Evren Bozgeyikli and Lal Lila Bozgeyikli. 2021. Evaluating Object Manipulation Interaction Techniques in Mixed Reality: Tangible User Interfaces and Gesture. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 778–787.
- [15] Connor Brooks and Daniel Szafr. 2020. Visualization of intended assistance for acceptance of shared control. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 11425–11430.
- [16] Maria E Cabrera, Kavi Dey, Kavita Krishnaswamy, Tapomayukh Bhattacharjee, and Maya Cakmak. [n.d.]. Cursor-based Robot Tele-manipulation through 2D-to-SE2 Interfaces. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4230–4237.
- [17] Simon Chamorro, Jack Collier, and François Grondin. 2021. Neural Network Based Lidar Gesture Recognition for Realtime Robot Teleoperation. In *2021 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 98–103.
- [18] Fu-Jen Chu, Ruinian Xu, Landan Seguin, and Patricio A Vela. 2019. Toward affordance detection and ranking on novel objects for real-world robotic manipulation. *IEEE Robotics and Automation Letters* 4, 4 (2019), 4070–4077.
- [19] Didier Crestani, Karen Godary-Dejean, and Lionel Lapierre. 2015. Enhancing fault tolerance of autonomous mobile robots. *Robotics and Autonomous Systems* 68 (2015), 140–155.
- [20] D Dajles, F Siles, et al. 2018. Teleoperation of a humanoid robot using an optical motion capture system. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOB)*. IEEE, 1–8.
- [21] Kody R Dillman, Terrance Tin Hoi Mok, Anthony Tang, Lora Oehlberg, and Alex Mitchell. 2018. A visual interaction cue framework from video game environments for augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [22] Simon DiMaio, Mike Hanuschik, and Usha Kreaden. 2011. The da Vinci surgical system. In *Surgical robotics*. Springer, 199–217.
- [23] Mark Draelos, Brenton Keller, Cynthia Toth, Anthony Kuo, Kris Hauser, and Joseph Izatt. 2017. Teleoperating robots from arbitrary viewpoints in surgical contexts. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2549–2555.

- [24] Anca D Dragan and Siddhartha S Srinivasa. 2013. A policy-blending formalism for shared control. *The International Journal of Robotics Research* 32, 7 (2013), 790–805.
- [25] Guanglong Du, Gengcheng Yao, Chunquan Li, and Peter X Liu. 2019. Natural human–robot interface using adaptive tracking system with the unscented Kalman filter. *IEEE Transactions on Human-Machine Systems* 50, 1 (2019), 42–54.
- [26] Jan Dufek, Xuesu Xiao, and Robin R Murphy. 2021. Best Viewpoints for External Robots or Sensors Assisting Other Robots. *IEEE Transactions on Human-Machine Systems* 51, 4 (2021), 324–334.
- [27] Anna Eilering, Giulia Franchi, and Kris Hauser. 2014. Robopuppet: Low-cost, 3d printed miniatures for teleoperating full-size robots. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1248–1254.
- [28] Mica R Endsley. 2017. Toward a theory of situation awareness in dynamic systems. In *Situational awareness*. Routledge, 9–42.
- [29] Chadrick R Evans, Melissa G Medina, and Anthony Michael Dwyer. 2018. Telemedicine and telerobotics: from science fiction to reality. *Updates in surgery* 70, 3 (2018), 357–362.
- [30] Marco Ewerton, Oleg Arenz, and Jan Peters. 2020. Assisted teleoperation in changing environments with a mixture of virtual guides. *Advanced Robotics* 34, 18 (2020), 1157–1170.
- [31] Bin Fang, Di Guo, Fuchun Sun, Huaping Liu, and Yupei Wu. 2015. A robotic hand-arm teleoperation system using human arm/hand with a novel data glove. In *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2483–2488.
- [32] Manuel Ferre, Rafael Aracil, Juan M Bogado, and Roque J Salterén. 2004. Improving force feedback perception using low bandwidth teleoperation devices. In *Proceedings of EuroHaptics Conference EH'2004*.
- [33] Scott Frees, G Drew Kessler, and Edwin Kay. 2007. PRISM interaction for enhancing control in immersive virtual environments. *ACM Transactions on Computer-Human Interaction (TOCHI)* 14, 1 (2007), 2–es.
- [34] Xiao Gao, Joao Silvério, Emmanuel Pignat, Sylvain Calinon, Miao Li, and Xiaohui Xiao. 2021. Motion mappings for continuous bilateral teleoperation. *IEEE Robotics and Automation Letters* 6, 3 (2021), 5048–5055.
- [35] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Rafael Medina-Carnicer. 2016. Generation of fiducial marker dictionaries using mixed integer linear programming. *Pattern recognition* 51 (2016), 481–491.
- [36] Vicent Gírbés-Juan, Vinicius Schettino, Yiannis Demiris, and Josep Tornero. 2020. Haptic and visual feedback assistance for dual-arm robot teleoperation in surface conditioning tasks. *IEEE Transactions on Haptics* 14, 1 (2020), 44–56.
- [37] P Christopher Gloumeau, Wolfgang Stuerzlinger, and JungHyun Han. 2020. PinNPivot: Object Manipulation Using Pins in Immersive Virtual Environments. *IEEE transactions on visualization and computer graphics* 27, 4 (2020), 2488–2494.
- [38] Claudia González, J Ernesto Solanes, Adolfo Munoz, Luis Gracia, Vicent Gírbés-Juan, and Josep Tornero. 2021. Advanced teleoperation and control system for industrial robots based on augmented virtuality and haptic feedback. *Journal of Manufacturing Systems* 59 (2021), 283–298.
- [39] Weston B Griffin, William R Provancher, and Mark R Cutkosky. 2003. Feedback strategies for shared control in dexterous telemanipulation. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, Vol. 3. IEEE, 2791–2796.
- [40] Jing Guo, Chao Liu, and Philippe Poignet. 2019. A scaled bilateral teleoperation system for robotic-assisted surgery with time delay. *Journal of Intelligent & Robotic Systems* 95, 1 (2019), 165–192.
- [41] Sami Haddadin, Lars Johannsmeier, and Fernando Díaz Ledezma. 2018. Tactile robots as a central embodiment of the tactile Internet. *Proc. IEEE* 107, 2 (2018), 471–487.
- [42] Rena Hale, Daniel Dorman, and Roger V Gonzalez. 2011. Individual muscle force parameters and fiber operating ranges for elbow flexion–extension and forearm pronation–supination. *Journal of biomechanics* 44, 4 (2011), 650–656.
- [43] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. 2020. Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9164–9170.
- [44] Eckhard H Hess and James M Polt. 1964. Pupil size in relation to mental activity during simple problem-solving. *Science* 143, 3611 (1964), 1190–1192.
- [45] Paul W Hodges and Bang H Bui. 1996. A comparison of computer-based methods for the determination of onset of muscle contraction using electromyography. *Electroencephalography and Clinical Neurophysiology/Electromyography and Motor Control* 101, 6 (1996), 511–519.
- [46] Shervin Javdani, Henny Admoni, Stefania Pellegrinelli, Siddhartha S Srinivasa, and J Andrew Bagnell. 2018. Shared autonomy via hindsight optimization for teleoperation and teaming. *The International Journal of Robotics Research* 37, 7 (2018), 717–742.
- [47] Abhishek Kadian, Joanne Truong, Aaron Gokaslan, Alexander Clegg, Erik Wijmans, Stefan Lee, Manolis Savva, Sonia Chernova, and Dhruv Batra. 2020. Sim2real predictivity: Does evaluation in simulation predict real-world

- performance? *IEEE Robotics and Automation Letters* 5, 4 (2020), 6670–6677.
- [48] Mitsuhiro Kamezaki, Junjie Yang, Hiroyasu Iwata, and Shigeki Sugano. 2016. Visibility enhancement using autonomous multicamera controls with situational role assignment for teleoperated work machines. *Journal of Field Robotics* 33, 6 (2016), 802–824.
 - [49] Liyiming Ke, Ajinkya Kamat, Jingqiang Wang, Tapomayukh Bhattacharjee, Christoforos Mavrogiannis, and Sidhartha S Srinivasa. 2020. Telemanipulation with chopsticks: Analyzing human factors in user demonstrations. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 11539–11546.
 - [50] David Kent, Carl Saldanha, and Sonia Chernova. 2017. A comparison of remote robot teleoperation interfaces for general object manipulation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. 371–379.
 - [51] George M Kontakis, Konstantinos Steriopoulos, John Damilakis, and Emmanouel Michalodimitrakis. 1999. The position of the axillary nerve in the deltoid muscle: A cadaveric study. *Acta orthopaedica Scandinavica* 70, 1 (1999), 9–11.
 - [52] Krzysztof Krejtz, Andrew T Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. 2018. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLoS one* 13, 9 (2018), e0203629.
 - [53] Roel J Kuiper, Dennis JF Heck, Irene A Kuling, and David A Abbink. 2016. Evaluation of haptic and visual cues for repulsive or attractive guidance in nonholonomic steering tasks. *IEEE Transactions on Human-Machine Systems* 46, 5 (2016), 672–683.
 - [54] Jin Sol Lee, Youngjib Ham, Hangu Park, and Jeonghee Kim. 2022. Challenges, tasks, and opportunities in teleoperation of excavator toward human-in-the-loop construction automation. *Automation in Construction* 135 (2022), 104119.
 - [55] Bennie Lewis and Gita Sukthankar. 2011. Two hands are better than one: Assisting users with multi-robot manipulation tasks. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2590–2595.
 - [56] Min Li, Yueyan Zhuo, Jiazhou Chen, Bo He, Guanghua Xu, Jun Xie, Xingang Zhao, and Wei Yao. 2020. Design and performance characterization of a soft robot hand with fingertip haptic feedback for teleoperation. *Advanced Robotics* 34, 23 (2020), 1491–1505.
 - [57] Zhi Li, Peter Moran, Qingyuan Dong, Ryan J Shaw, and Kris Hauser. 2017. Development of a tele-nursing mobile manipulator for remote care-giving in quarantine areas. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3581–3586.
 - [58] Tsung-Chi Lin, Achyuthan Unni Krishnan, and Zhi Li. 2019. Physical fatigue analysis of assistive robot teleoperation via whole-body motion mapping. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2240–2245.
 - [59] Tsung-Chi Lin, Achyuthan Unni Krishnan, and Zhi Li. 2022. Comparison of Haptic and Augmented Reality Visual Cues for Assisting Tele-manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 9309–9316.
 - [60] Tsung-Chi Lin, Achyuthan Unni Krishnan, and Zhi Li. 2022. Intuitive, Efficient and Ergonomic Tele-Nursing Robot Interfaces: Design Evaluation and Evolution. *ACM Transactions on Human-Robot Interaction* (2022).
 - [61] Dylan P Losey, Craig G McDonald, Edoardo Battaglia, and Marcia K O'Malley. 2018. A review of intent detection, arbitration, and communication aspects of shared control for physical human-robot interaction. *Applied Mechanics Reviews* 70, 1 (2018).
 - [62] Rute Luz, José Corujeira, Laurent Grisoni, Frédéric Giraud, Jose Luis Silva, and Rodrigo Ventura. 2019. On the use of haptic tablets for UGV teleoperation in unstructured environments: System design and evaluation. *IEEE Access* 7 (2019), 95443–95454.
 - [63] Matteo Macchini, Thomas Havy, Antoine Weber, Fabrizio Schiano, and Dario Floreano. 2020. Hand-worn haptic interface for drone teleoperation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 10212–10218.
 - [64] Zhanat Makhataeva and Huseyin Atakan Varol. 2020. Augmented reality for robotics: A review. *Robotics* 9, 2 (2020), 21.
 - [65] Lynn McAtamney and E Nigel Corlett. 1993. RULA: a survey method for the investigation of work-related upper limb disorders. *Applied ergonomics* 24, 2 (1993), 91–99.
 - [66] Leonardo Meli, Claudio Pacchierotti, Gionata Salvietti, Francesco Chinello, Maurizio Maisto, Alessandro De Luca, and Domenico Prattichizzo. 2018. Combining wearable finger haptics and augmented reality: User evaluation using an external camera and the microsoft hololens. *IEEE Robotics and Automation Letters* 3, 4 (2018), 4297–4304.
 - [67] Daniel Mendes, Fabio Marco Caputo, Andrea Giachetti, Alfredo Ferreira, and Joaquim Jorge. 2019. A survey on 3d virtual object manipulation: From the desktop to immersive virtual environments. In *Computer graphics forum*, Vol. 38. Wiley Online Library, 21–45.
 - [68] Davide Nicolis, Marco Palumbo, Andrea Maria Zanchettin, and Paolo Rocco. 2018. Occlusion-free visual servoing for the shared autonomy teleoperation of dual-arm robots. *IEEE Robotics and Automation Letters* 3, 2 (2018), 796–803.

- [69] Allison M Okamura. 2004. Methods for haptic feedback in teleoperated robot-assisted surgery. *Industrial Robot: An International Journal* (2004).
- [70] Andrew L Orekhov, Caroline B Black, John Till, Scotty Chung, and D Caleb Rucker. 2016. Analysis and validation of a teleoperated surgical parallel continuum manipulator. *IEEE Robotics and Automation Letters* 1, 2 (2016), 828–835.
- [71] Sungman Park, Yeongtae Jung, and Joonbum Bae. 2016. A tele-operation interface with a motion capture system and a haptic glove. In *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*. IEEE, 544–549.
- [72] Soran Parsa, Horia A Maior, Alex Reeve Elliott Thumwood, Max L Wilson, Marc Hanheide, and Amir Ghalamzan Esfahani. 2022. The Impact of Motion Scaling and Haptic Guidance on Operators' Workload and Performance in Teleoperation. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.
- [73] Nicolò Pedemonte, Firas Abi-Farraj, and Paolo Robuffo Giordano. 2017. Visual-based shared control for remote telemanipulation with integral haptic feedback. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5342–5349.
- [74] Minh Tien Phan, Indira Thouvenin, and Vincent Frémont. 2016. Enhancing the driver awareness of pedestrian using augmented reality cues. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 1298–1304.
- [75] Will Pryor, Balazs P Vagvolgyi, Anton Deguet, Simon Leonard, Louis L Whitcomb, and Peter Kazanzides. 2020. Interactive Planning and Supervised Execution for High-Risk, High-Latency Teleoperation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1857–1864.
- [76] Camilo Perez Quintero, Masood Dehghan, Oscar Ramirez, Marcelo H Ang, and Martin Jagersand. 2017. Flexible virtual fixture interface for path specification in tele-manipulation. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5363–5368.
- [77] Harish Chaandar Ravichandar, Avnish Kumar, and Ashwin Dani. 2018. Gaze and motion information fusion for human intention inference. *International Journal of Intelligent Robotics and Applications* 2, 2 (2018), 136–148.
- [78] Yosef Razin and Karen Feigh. 2017. Learning to predict intent from gaze during robotic hand-eye coordination. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- [79] Francisco J Romero-Ramirez, Rafael Muñoz-Salinas, and Rafael Medina-Carnicer. 2018. Speeded up detection of squared fiducial markers. *Image and vision Computing* 76 (2018), 38–47.
- [80] Eric Rosen, David Whitney, Elizabeth Phillips, Gary Chien, James Tompkin, George Konidaris, and Stefanie Tellex. 2019. Communicating and controlling robot arm motion intent through mixed-reality head-mounted displays. *The International Journal of Robotics Research* 38, 12-13 (2019), 1513–1526.
- [81] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. 2020. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research* 39, 8 (2020), 895–935.
- [82] Michelle L Rusch, Mark C Schall Jr, John D Lee, Jeffrey D Dawson, and Matthew Rizzo. 2014. Augmented reality cues to assist older drivers with gap estimation for left-turns. *Accident Analysis & Prevention* 71 (2014), 210–221.
- [83] Dongseok Ryu, Chang-Soon Hwang, Sungchul Kang, Munsang Kim, and Jae-Bok Song. 2005. Wearable haptic-based multi-modal teleoperation of field mobile manipulator for explosive ordnance disposal. In *IEEE International Safety, Security and Rescue Robotics, Workshop, 2005*. IEEE, 75–80.
- [84] Sophia Sakr, Thomas Daunizeau, David Reversat, Stéphane Régnier, and Sinan Haliyo. 2018. An ungrounded master device for tele-microassembly. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1–9.
- [85] Philipp Schleer, Philipp Kaiser, Sergey Drobinsky, and Klaus Radermacher. 2020. Augmentation of haptic feedback for teleoperated robotic surgery. *International Journal of Computer Assisted Radiology and Surgery* 15, 3 (2020), 515–529.
- [86] Stela H Seo, Daniel J Rea, Joel Wiebe, and James E Young. 2017. Monocle: interactive detail-in-context using two pan-and-tilt cameras to improve teleoperation effectiveness. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 962–967.
- [87] J Ernesto Solanes, Adolfo Muñoz, Luis Gracia, Ana Martí, Vicent Gírbés-Juan, and Josep Tornero. 2020. Teleoperation of industrial robot manipulators based on augmented reality. *The International Journal of Advanced Manufacturing Technology* 111, 3 (2020), 1077–1097.
- [88] Peng Song, Wooi Boon Goh, William Hutama, Chi-Wing Fu, and Xiaopei Liu. 2012. A handle bar metaphor for virtual object manipulation with mid-air interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1297–1306.
- [89] Alexis D Souchet, Stéphanie Philippe, Domitile Lourdeaux, and Laure Leroy. 2022. Measuring visual fatigue and cognitive load via eye tracking while learning with virtual reality head-mounted displays: A review. *International Journal of Human-Computer Interaction* 38, 9 (2022), 801–824.
- [90] Todor Stoyanov, Robert Krug, Andrey Kiselev, Da Sun, and Amy Loutfi. 2018. Assisted telemanipulation: A stack-of-tasks approach to remote manipulator control. In *2018 IEEE/RSJ International Conference on Intelligent Robots and*

- Systems (IROS)*. IEEE, 1–9.
- [91] Ajay Kumar Tanwani and Sylvain Calinon. 2017. A generative model for intention recognition and manipulation assistance in teleoperation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 43–50.
 - [92] Ozan Tokatli, Pragna Das, Radhika Nath, Luigi Pangione, Alessandro Altobelli, Guy Burroughes, Emil T Jonasson, Matthew F Turner, and Robert Skilton. 2021. Robot-assisted glovebox teleoperation for nuclear industry. *Robotics* 10, 3 (2021), 85.
 - [93] Achyuthan Unni Krishnan, Tsung-Chi Lin, and Zhi Li. 2022. Design Interface Mapping for Efficient Free-form Tele-manipulation. In *to appear in the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE.
 - [94] Alexandra Valiton, Hannah Baez, Naomi Harrison, Justine Roy, and Zhi Li. 2021. Active Telepresence Assistance for Supervisory Control: A User Study with a Multi-Camera Tele-Nursing Robot. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE.
 - [95] Michael Walker, Zhaozhong Chen, Matthew Whitlock, David Blair, Danielle Albers Szafr, Christoffer Heckman, and Daniel Szafr. 2021. A mixed reality supervision and telepresence interface for outdoor field robotics. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2345–2352.
 - [96] Michael Walker, Hooman Hedayati, Jennifer Lee, and Daniel Szafr. 2018. Communicating robot motion intent with augmented reality. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 316–324.
 - [97] Michael E Walker, Hooman Hedayati, and Daniel Szafr. 2019. Robot teleoperation with augmented reality virtual surrogates. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 202–210.
 - [98] Samuel S White, Keion W Bisland, Michael C Collins, and Zhi Li. 2020. Design of a High-Level Teleoperation Interface Resilient to the Effects of Unreliable Robot Autonomy. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 11519–11524.
 - [99] David Wilkie, Jur Van Den Berg, and Dinesh Manocha. 2009. Generalized velocity obstacles. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 5573–5578.
 - [100] Yuqiang Wu, Pietro Balatti, Marta Lorenzini, Fei Zhao, Wansoo Kim, and Arash Ajoudani. 2019. A teleoperation interface for loco-manipulation control of mobile collaborative robotic assistant. *IEEE Robotics and Automation Letters* 4, 4 (2019), 3593–3600.
 - [101] Liang Yang, Yong Chen, Zhi Liu, Kairui Chen, and Zixuan Zhang. 2019. Adaptive fuzzy control for teleoperation system with uncertain kinematics and dynamics. *International Journal of Control, Automation and Systems* 17, 5 (2019), 1158–1166.
 - [102] Erkang You and Kris Hauser. 2012. Assisted teleoperation strategies for aggressively controlling a robot arm with 2d input. In *Robotics: science and systems*, Vol. 7. MIT Press USA, 354.
 - [103] Tian Zhou, Maria E Cabrera, Juan P Wachs, Thomas Low, and Chandru Sundaram. 2016. A comparative study for telerobotic surgery using free hand gestures. *Journal of Human-Robot Interaction* 5, 2 (2016), 1–28.