Scene-Context-Aware Indoor Object Selection and Movement in VR

Miao Wang[1,2*] Zi-Ming Ye[1] Jin-Chuan Shi[1] Yong-Liang Yang[3]

1 State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China
2 Peng Cheng Laboratory 3 University of Bath

ABSTRACT

Virtual reality (VR) applications such as interior design typically require accurate and efficient selection and movement of indoor objects. In this paper, we present an indoor object selection and movement approach by taking into account scene contexts such as object semantics and interrelations. This provides more intelligence and guidance to the interaction, and greatly enhances user experience. We evaluate our proposals by comparing them with traditional approaches in different interaction modes based on controller, head pose, and eye gaze. Extensive user studies on a variety of selection and movement tasks are conducted to validate the advantages of our approach. We demonstrate our findings via a furniture arrangement application.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction techniques—Pointing

1 INTRODUCTION

How to select and manipulate objects in a virtual environment (VE) is a fundamental problem in virtual reality (VR), and has a significant influence on human interaction experience [1, 8, 35]. Various techniques have been proposed based on different interaction modes including contact based interaction [41], remote controller based interaction [4, 40], and interactions using head pose and eye gaze [28, 39]. Despite interaction modes, research attention has been primarily paid to how to improve the interaction accuracy and speed in VE with different object settings (e.g., size, density, occlusion) [8, 24, 31].

Indoor navigation and interaction are among the most important applications in virtual and augmented reality (AR). The recent development in this area enables customized interior design based on VR/AR techniques, as witnessed by a number of interactive tools (e.g., IKEA Immerse, Wayfair Spaces) launched by IKEA, Wayfair, etc. A typical interior design process requires the user to frequently select the right objects and place them in the right positions. This poses new challenges for existing human-object interaction techniques, where the objects in the scene are simply treated as individual entities, while scene contexts such as object semantics and interrelations are not taken into account.

In this paper, we study the virtual indoor scene manipulation task in terms of object selection and movement, which have potential applications such as interior design and indoor navigation in VR. Different from previous studies [4,6,31,64] that manipulated objects in abstract or simplified scenes, we conducted experiments in high-quality virtual indoor scenes (see Figure 1). Specifically, we focus on interacting with target objects (e.g., chairs, cups) that can be selected and moved within 3D indoor scenes and investigate the *selection* and *movement* interactions using raycasting based technique, where the user selects an object by pointing to it using controller, head pose or eye gaze, and move it to another indicated position. Traveling in the VE is allowed to complete challenging tasks. For example, the user may want to take a walk and adjust his/her viewpoint to the object to complete a selection.

*Corresponding author, e-mail: miaow@buaa.edu.cn

Figure 1: Examples of virtual indoor scenes used in our interaction experiments (©ArchVizPRO Interior).

We propose to facilitate raycasting based object selection and movement by exploiting the surrounding scene contexts, including object semantics (its role in the scene) and object relationships (its relation with other objects in the scene). Based on the scene contexts, we design two experiments where the first experiment investigates the performances of controller, head pose, and eye gaze based selection techniques under different conditions regarding object size/occlusion/density, and the second experiment studies the movement methods under different moving distance/angle/occlusion conditions. Both experiments are conducted within high-quality virtual indoor scenes, with the selection and movement tasks specifically designed to meet the practical scenarios for virtual indoor object manipulation. Through extensive user studies, we demonstrated the advantages of our approach in terms of saving interaction time and enhancing user experience. We further developed a furniture rearrangement application to exemplify our findings.

Overall our work makes the following major contributions: 1) we apply contextual information to 3D selection and manipulation specifically for interior design scene in VR, and 2) we verify the effectiveness of our approach by comprehensive user experiments and a practical application.

2 RELATED WORK

2.1 Selection Techniques in VR

Selection techniques in VR are commonly divided into two types: contact based selection and raycasting (pointing) based selection [2]. The former requires the user to virtually touch an object, which is more natural, but less convenient for faraway objects. The latter allows the user to select a distant object by just pointing to but not walking to it. We will mainly discuss three primary raycasting based selection techniques with different interaction modes, as they serve as the baselines for our work.

Controller based technique (Figure 2(a)) utilizes handheld controller to select objects by emitting a ray to the scene [1]. The first object that is intersected by the ray can be selected [29]. Since it is simple and efficient, various works have been presented using this technique [15–17, 63]. However, the accuracy of selecting small or/and dense objects can be easily affected by hand movement. To solve this problem, Dang et al. [15] proposed to use cone-shaped ray with a certain range of tolerance. But the additional range requires a second confirmation [12, 21], or heuristic search [16, 48], especially for dense objects. Progressive refinements of selection with spatial

context was proposed in [3, 27]. However, without object-level context analysis of the scene, multiple steps that interactively narrow down the selected candidates were required. Lu et al. [31] made improvements based on bubbles, but the effect is not ideal for dense and occluded objects. Yu et al. [64] refined promising raycasting techniques and performed better for fully-occluded object selection.

Head pose based technique (Figure 2(b)) leverages head movement and fixation to point to object for selection, and has been applied in various scenarios in VR [13, 16, 55]. In particular, recent work on head movement for text input [62] showed its accuracy and speed for long and intensive interactions. Head pointing is well-known for its benefits of providing no submission, but its performance and usability are considered inferior to the raycasting technique. An early investigation [23] reported that the joystick points faster than the head. Lin et al. [30] compared head and hand pointing methods on large stereoscopic projection displays. The results showed that hand pointing generally performs better with lower muscle fatigue and better usability, while head pointing provides higher accuracy. Bernardos et al. [6] compared the two modes using a wall-sized projection screen. They did not find significant differences in task performance, but hand based pointing shows better user experience.

Eye gaze based technique (Figure 2(c)) relies on eye movement to select objects. Research on gaze based interaction dates back to the 1980s [9, 10]. Poupyrev et al. [38] showed that eye movement performs better than mouse movement for 2D selection. But eye movement also has problems such as high error rate and instability due to the restriction of the capturing device. Based on investigating different selection methods using eye tracking, Mackenzie [33] made an overview of several issues caused by eye trackers. Kytö et al. [28] studied eye pointing and head pointing in AR, and found that eye pointing is faster while head pointing is more accurate.

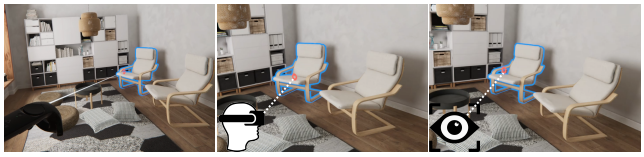
Overall, controller based technique has been widely used in commodity VR handheld devices. Head and eye selections are developing very fast and can be treated as promising complements [14]. Compared with eye selection, head selection is easier to learn and more stable. Eye selection is faster with better mobility, while its learning curve is steeper and it requires a higher workload [5, 25]. In our work, we evaluate the presented approach in all three selection modes to demonstrate its effectiveness.

2.2 Context-aware Interaction in VR

Context-aware interaction techniques for 3D data have been investigated [20, 42, 45]. For example, it has been demonstrated that semantic information, such as gravitational hierarchy [37] and clusters [18, 44, 65], can improve the performance of group selection. Besides, some work revealed that spatial information can also benefit selection techniques in 3D scenes. An early work by Bukowski et al. [11] presented a framework to help design and implement convenient 3D object manipulation methods in a 3D virtual environment, where object associations using spatial context played an important role. A series of efficient techniques [36, 46, 49, 50, 52, 53] have been proposed for 3D object creation and positioning in virtual environments by using group and contextual information. For example, Smith et al. [46] showed that with 2D user interfaces, 3D object motion and orientation can be automatically adjusted using contextual constraints. It is worth noting that our research is not the first to use contextual information in a 3D scene, and we focus on applying contextual information to pointing based 3D selection and movement techniques in VR, specifically for interior design scenes.

2.3 Virtual Indoor Object Movement and Arrangement

The interactive movement of indoor objects in VR consists of a series of drag-and-drop operations to move virtual objects to target positions. In contrast to object selection, object movement is less explored in VR literature. One main approach is to virtually contact and grasp objects, then make movements [26, 35]. Although human hands are capable of forming various gestures by nature, and conform to moving objects in the real world, the labor cost could be considerably high for applications that require frequent object movements such as interior design. Kang et al. [26] demonstrated that the worlds-in-miniature technique provided better usability and performance than *gaze and pinch* interaction and *direct touch and grab* interaction. In our experiment, raycasting based non-contact object movements with scene contexts are studied. Compared to traditional raycasting based methods, the influence on movement accuracy can be largely improved by involving contextual information of the scene.



(a) Controller (b) Head pose (c) Eye gaze

Figure 2: Three raycasting based interaction methods are investigated. (a) Raycasting using handheld controller. (b) Raycasting by head pose fixation. (c) Raycasting by eye gaze fixation. The user points to the target object via (a) (b) or (c) then press the trigger on the controller to complete a selection task.

In the graphics field, how to generate plausible indoor furniture arrangements has attracted attention in the last few years. The main aim is to accelerate the interior design process by automatically exploring the high dimensional layout space and providing quality arrangements for later design refinement. Early research optimizes arrangement by implementing a set of explicit design guidelines that represent not only object-object but also object-scene relationships [34, 60, 66]. Recent works leverage a data-driven approach to learn how to make arrangements from existing furniture layouts [57–59]. A detailed review in this area is beyond the scope of this paper. Interested readers may refer to [67] for a comprehensive survey. Our research focus is on how to interactively select and move indoor objects [46]. It can be used to interactively generate satisfying indoor furniture arrangements as demonstrated in the application.

3 Scene Context-aware Indoor Object Interaction

To explore how scene contexts contribute to such interactions, we make an assumption that the contextual information is available as prior knowledge, and focus on user interactions with indoor objects. This is reasonable for interior design applications where furniture types (e.g., chair, table, bed) and combinations (e.g., dining set) are usually given [54, 61]. And objects with strong relations can be easily identified by matching their properties such as shape and texture (e.g., a set of congruent chairs), or/and checking their spatial locations (e.g., alignment, support, etc.). Next, we will introduce the strategies to leverage representative indoor scene contexts, and how we integrate them into object selection and movement operations.

3.1 Indoor Scene Contexts

Depending on the functionality, an indoor scene usually exhibits fruitful contextual information provided by the constituent objects in the scene. Below we summarize the contextual information that we used in our experiments. Note that scene contexts have also been employed to improve furniture arrangement computation as in [32, 34, 66]. Our work shares a similar spirit but our goal is to improve indoor object interaction.

Object category: object category is the most basic semantic information in a scene context. In our experiments, furniture categories such as chairs, sofas, beds, and decorations such as vases are used.

Dependence relationship: object in one category might depend on object in another category to realize its functionality, see Figure 3

Figure 3: Examples of indoor virtual objects with dependence relationship and group relationship.

(Top). For example, nightstands are placed next to the bed, and vases are usually supported by a table. Without loss of generality, we denote the primary object (e.g., table) as a parent object and the smaller objects (e.g., chairs, vases) as child objects. A parent object can have more than one child object, while a child object has only one parent object to eliminate any ambiguity during selection. Such a relationship can be used to enhance the selection. For example, if the target for selection is a table, one can firstly select a chair, then switch to the table following the dependence relationship.

Group relationship: objects of the same category that are placed closely are considered as a group, see Figure 3 (Bottom). For example, congruent chairs are typically placed together next to a table, and decorative objects are placed as a group inside a cabinet. If one object is selected, the user can easily switch to other objects within the same group.

Spatial relationship: based on functionality and aesthetics considerations, objects can be spatially correlated in terms of position and orientation in the scene context. For example, chairs usually have rotational symmetry around a round table, and reflective symmetry next to a rectangular table. A bedside table is usually aligned to the bed top, etc.

3.2 Context Integration for Selection and Movement

3.2.1 Selection

As mentioned before, we evaluate three pointing based interactions based on controller, eye gaze, and head pose, respectively. Compared to contact based selection, pointing based selection also allows users to select distant objects. It performs well on isolated objects that are clearly separated from others. However, pointing based interaction may degenerate when handling complex cases of *partial occlusion* or *dense environment*. To resolve these challenging cases, many researchers have modified the original raycasting based selection [31] such as bending the ray or zooming-in the region of interests (ROI).

As a complementary to raycasting, our strategy is to allow the user to select other nearby objects with good visibility first, then switch to the target object by leveraging scene contexts as in Section 3.1. Intuitively, this strategy can reduce unnecessary and error-prone trials, avoid travelling in the scene or operating on additional interfaces (e.g., zoomed display) [12]. To this end, we propose specific strategies to enhance the selection operation for the aforementioned two challenging cases:

Partial occlusion: for the cases when a target object O_t is occluded by other objects, the user can point at a 3D position P on its parent object O_p and switch to the target O_t with a single press of the menu button on the controller, if O_t is the parent object to P, or the closest child object to P:

$$O_t = \underset{O \in \mathscr{C}(O_p)}{\arg\min} ||P_c(O) - P||_2, \tag{1}$$

where $\mathscr{C}(O_p)$ is the set of child objects of the parent object O_p, and $P_c(O)$ is the centroid of object O.

Dense environment: sometimes a target object O_t locates closely to a group of n objects $\mathscr{G}(O_t) = \{O_1, O_2, ..., O_n\}$ (including O_t). This situation can cause trouble for selection, because the user may travel heavily to a position close to O_t to make an accurate selection. We utilize the group relationship for a more flexible selection. Specifically, the user is allowed to select an object O_s from $\mathscr{G}(O_t)$ with the ray intersection point P, then use the touchpad on the controller to specify a 2D direction d, and switch back to the target object O_t as:

$$O_t = \underset{O \in \mathscr{G}(O_s)}{\arg\max} cos(Proj(P_c(O)) - Proj(P), d), \tag{2}$$

where $P_c(O)$ is the centroid of object O, $Proj(\cdot)$ is the function to project the 3D point to the camera coordinate system.

3.2.2 Movement

We proposed a strategy to handle the cases when placing a target object onto an occluded target location. In order to allow the user to focus on accurate placement rather than tedious adjustment of the object orientation, we also proposed auto orientation adjustments during placing the target object.

Orientation adjustment: we automatically adjust the yaw of an object during its movement. When moving a target object O_t, whose current 2D location projected to the floor plane is p_{O_t}, we adjust its forward direction d_{O_t} (parallel to the floor) by considering the spatial relationship to other objects within a circular region with the radius of 1 meter. Specifically, objects $\mathscr{O} = \{O_1, O_2, \ldots, O_n\}$ within the same *group* or being O_t's parent object can have an influence on d_{O_t}:

$$d_{O_t} = \sum_{O \in \mathscr{O}} \omega(O, O_t) \cdot D(O, O_t), \tag{3}$$

where $\omega(O, O_t)$ is the weight of object O for orientation adjustment considering the distance, defined as:

$$\omega(O, O_t) = \frac{\max(-\ln(||p_{O_t} - p_O||_2), 0)}{\sum_{O' \in \mathscr{O}} \max(-\ln(||p_{O_t} - p_{O'}||), 0)}, \tag{4}$$

and $D(O, O_t)$ is the direction vector of object O that contributes to the orientation adjustment:

$$D(O, O_t) = \begin{cases} d_O & O \in G(O_t) \\ -d_O & O = parent(O_t). \end{cases} \tag{5}$$

Intuitively, when O is the parent of the target object O_t, O_t should face O. On the other hand, if O and O_t are next to each other while being in the same group, the forward direction of O_t and O should also be the same. Such automatic yaw adjustments work for the *conversation* mode [34], such as chairs and table, sofa and tea table, etc., see Figure 4.

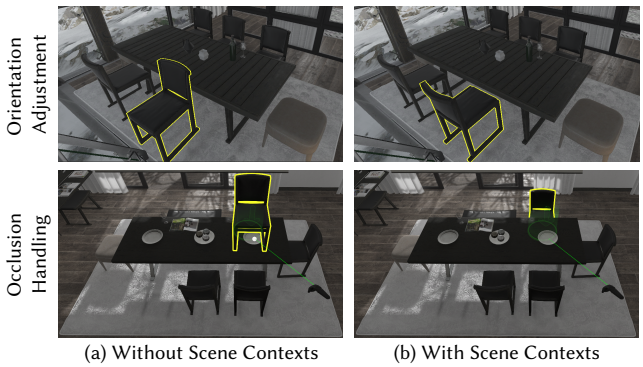


Figure 4: Scene contexts for automatic movement refinement: orientation adjustment and occlusion handling (a) without and (b) with scene contexts.

Occluded target position: it could happen that the target position P_t for placing target object (O_t) is occluded by some other objects. When selecting an occluded position, the raycasting technique will always point to the first object that is intersected with the ray. Inspired by [51,64], we allow the ray to penetrate all supportive objects (including the floor) $\mathscr{O} = \{O_1, O_2, \ldots, O_n\}$ along the ray with intersection points $\mathscr{P} = \{P_i(O_1), P_i(O_2), \cdots, P_i(O_n)\}$, and locate on the nearest object O' that is legal to support O_t, at the point P'_t:

$$P'_t = \underset{O \in \mathscr{O}}{\arg\min} ||P_i(O) - P_t|| \qquad \text{if} \quad Legal(O, O_t), \tag{6}$$

where $Legal(O, O_t)$ is a Boolean function that returns *true* if O is legal to support O_t based on contextual knowledge. Figure 4 shows an illustration of our improvement.

Brief summary. In this section, we have proposed refinements to raycasting based selection techniques, that can simply relocate selected objects to nearby targets based on the *dependence* or *group* relationship, by a single press of a button on the controller. To simplify the movement procedure after selecting a target object, automatic orientation adjustment and occlusion handling are computed, with tedious and laborious operations avoided.

4 Experiment Overview

We have presented the indoor object selection and movement problem that we aim to investigate, and how to leverage indoor contexts in object selection and movement interactions. We designed two experiments to evaluate the proposed approach in virtual indoor scenes. To mimic practical scenarios for interior design applications, participants were allowed to travel. However, if the participant failed some trials and chose to travel, the completion time would increase accordingly, which can reflect the user performance and difficulty of the tasks. In the first experiment, we conducted a user study to evaluate raycasting based selection methods and the proposed refinements (3 baselines + 3 refinements). Selection tasks under various conditions were tested, with measures such as the selection time, error count, and walking distances recorded. Subjective questionnaires were collected as further evidence for comparison.

The second experiment compared the performance of the automatic refinements for the controller, eye gaze, and head pose interactions on movement tasks. A pilot study revealed that without our refinement, the manual tuning of object orientation and handling occluded target placement location significantly increased the processing time and walking distances. Thus we only evaluated the refined methods in this study. We clarify that while simple 3D docking tasks were used for evaluating 3D object manipulation techniques in the literature [7,47,56], our goal is to study how the scene contexts facilitate indoor object movement techniques on various angle/distance/occlusion conditions in a practical indoor scene.

We further demonstrated the usage of the interaction methods via an application, where the user employed the best-ranked methods in the studies to rearrange furniture objects, with a sequence of selection and movement interactions.

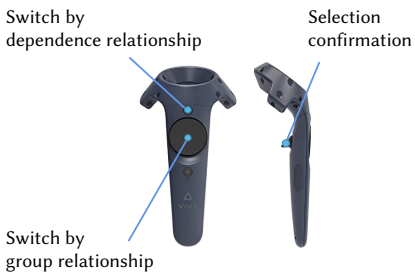


Figure 5: Key mapping of the controller.

5 Experiment 1: Selection Study

We studied the performance of controller, head pose, and eye gaze based methods, as well as corresponding refinements with contextual information integration, for virtual indoor object selection.

5.1 Participants and Apparatus

We recruited 12 participants (4F/8M, Mean age = 21, SD = 1.76) from a university in this study. 8 participants had prior VR experience. The experiment duration ranged from 30 minutes to 40 minutes. The apparatus included an HTC Vive Pro Eye with a Vive controller, and a desktop PC (Core i7 9700K, GTX 2080Ti Graphics card, 32 GB RAM) with Microsoft Windows 10. We used Unity 3D v2018.3.5f1 software for implementation. The virtual scenes were from Unity Asset Store - *ArchVizPRO Interior Vol. 4-6*.

5.2 Selection Methods

Six pointing based METHODS were investigated, where the first three were baseline methods, and the last three were our refinements:

Controller: In this method, a ray is emitted from the top of the controller. The first object hit by the ray is assumed to be selected. The selection is confirmed if the user presses the trigger on the controller (see Figure 5). To select target objects, the user can move the handle or/and travel in the scene.

HeadPose: The user selects objects by turning the head. In this method, an invisible ray is emitted from the forehead instead of the controller, to the virtual scene, with the intersection point indicated by a cursor. The user confirms the selection by pressing the trigger on the controller.

EyeGaze: This method requires capturing the eye gaze. We used an HTC Vive Pro Eye HMD that was able to track eye gaze, and the API provided by Sranipal to obtain this information. An invisible ray is emitted from the eye, and the direction is determined by the eye gaze. The user needs to look at the object for localization, and uses the controller trigger to confirm the choice.

Controller+: In this method, the integration of scene contexts is implemented as explained in Section 3.2.1 on top of *Controller*. The user can press the menu button or the touchpad to switch selected object based on the dependence/group relationships (see Figure 3 and Figure 5).

HeadPose+: Similar to *Controller+*, a simple press of the controller button can activate the semantic integration.

EyeGaze+: This method is the same as *HeadPose+*, except the pointing is based on eye gaze.

5.3 Design and Procedure

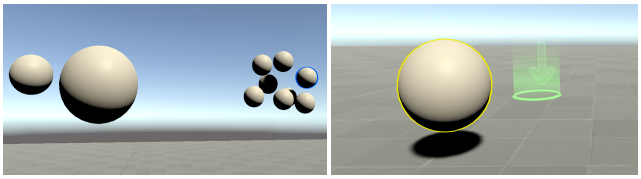
An indoor scene usually contains objects with different properties and distributions. Following Fitts's law [19] and evidence from previous work, we chose five representative conditions of indoor objects for evaluation (see Figure 6). When starting a trial, the participant's initial location and the target object was specified, where the object was guaranteed visible from the current view.

- ***Non-occluded large object***: an object which is close to the participant or with a large size, and is not occluded by other objects.
- ***Non-occluded small object***: an object which is far from the participant or with a small size, and is not occluded by other objects.
- ***Partially occluded object***: an object that is partially occluded by other objects.
- ***Non-occluded object in dense environment***: a non-occluded object with distractors around.
- ***Partially occluded object in dense environment***: an object that is partially occluded by other distractive objects around.

For each condition, we created 3 variations with different target objects. Based on a within-subject experimental design, each participant completed 6 methods $\times$ 5 conditions $\times$ 3 variations $\times$ 3 repetitions = 270 trials.

(a) (b) (c) (d) (e)

Figure 6: Conditions for selection task. (a) Non-occluded large object. (b) Non-occluded small object. (c) Partially occluded object. (d) Non-occluded object in dense environment. (e) Partially occluded object in dense environment. The target objects for selection are highlighted with blue silhouettes. Zoom-in views are shown at top-right corners.



(a) (b)

Figure 7: Training scenes of the experiments. (a) The training scene of the selection study. The target object is highlighted with a blue contour. (b) The training scene of the movement study. The target object is highlighted with a yellow contour and the target placement position is indicated with a green arrow.

The experiment was approved by the ethics committee of the university. After welcoming the participant into our lab, the experimenter instructed the participant about the aim and procedure of the study. The participant was aware that he/she was encouraged to complete the tasks as fast and accurately as possible. She/he then signed an informed consent form and filled a demographic form.

Before formal tests, the participant practiced each method in a simple training scene as shown in Figure 7(a). Although being simplified, the scene contains all 5 conditions as in the formal test. For each practice trial, a random target sphere was highlighted with a blue contour, the user then made a selection using the corresponding testing method. The participant was aware that contextual information from indoor scenes will replace the simple ones.

In the formal tests, the order of the selection methods was counterbalanced using a Latin Square approach, and the order of condition/variation was randomized in three repetition loops. In each task, the target object for selection was highlighted using a blue contour. The participant was asked to successfully select the correct object unless exceeding a 30-second time limit. During the experiment, we recorded the selection time and the travelling distance for each trial. After experiencing each selection method, the participant was asked to first complete the UEQ-S [43] and NASA-TLX [22] questionnaires, then take a 2-minute break (for the next test if exists). The experiment took about 40 minutes.

After the experiment, the participant was asked to rank the methods based on his/her overall preference. In the end, we conducted a short interview to receive feedback on the overall experiment. A voucher of 10$ was given to the participant for acknowledgment.

5.4 Results

We have collected 3240 data points (12 participants $\times$ 270 trials), where the outliers above three standard deviations from the mean (75 trials, 2.3%) were removed to analyze the selection time and travelling distance. The data were normally distributed, and we performed a repeated-measures ANOVA and pairwise comparisons with Bonferroni adjustment to analyze method performance.

A Mauchly's Test was used to test the sphericity. Once violated, the degrees of freedom produced by repeated-measures ANOVA were then adjusted using Greenhouse-Geisser correction. Figure 8 shows the results of the selection time. We note that the travelling distance values were mostly less than 0.1 meters, except when objects were partially occluded. Considering very small travelling distances would be less meaningful, we only analyze the travelling distance on partially occluded objects. The travelling distance results were provided in the supplementary file.

Non-occluded Large Object METHOD had a significant effect on selection time ($F(3.76,131.62) = 6.573, p < 0.001, \eta_p^2 = 0.158$). *Controller+* was on average the fastest, and was significantly faster than *HeadPose+* ($-0.39\,s, p = 0.003$).

Non-occluded Small Object METHOD had a significant effect on selection time ($F(3.79,132.72) = 6.573, p < 0.001, \eta_p^2 = 0.179$). *Controller+* was the fastest, and was significantly faster than *HeadPose* ($-0.67\,s, p < 0.001$), *EyeGaze* ($-0.65\,s, p = 0.002$) and *HeadPose+* ($-0.50\,s, p = 0.001$).

Partially Occluded Object A significant effect of METHOD was found on selection time ($F(5,175) = 43.602, p < 0.001, \eta_p^2 = 0.555$). *EyeGaze+* was significantly faster than *HeadPose* ($-2.27\,s, p < 0.001$), *EyeGaze* ($-2.43\,s, p < 0.001$) and *Controller* ($-1.83\,s, p < 0.001$). There were significant improvements when using scene contexts, with *HeadPose+* faster than *HeadPose* ($-2.00\,s, p < 0.001$), *EyeGaze+* faster than *EyeGaze* ($-2.43\,s, p < 0.001$), and *Controller+* faster than *Controller* ($-1.75\,s, p < 0.001$). A significant effect of METHOD was found on travelling distances ($F(2.15,75.14) = 131.883, p < 0.001, \eta_p^2 = 0.790$). *EyeGaze+* had the lowest travelling distance and was significantly lower than baseline methods ($-0.396\,m$ for *HeadPose*, $-0.437\,m$ for *EyeGaze*, $-0.420\,m$ for *Controller*, $p < 0.001$ for all). Post-hoc analysis indicated that *HeadPose+* was significantly smaller than *HeadPose* ($-0.392\,m, p < 0.001$), *EyeGaze+* was significantly smaller than *EyeGaze* ($-0.437\,m, p < 0.001$), and *Controller+* was significantly smaller than *Controller* ($-0.417\,m, p < 0.001$).

Non-occluded Object in Dense Environment METHOD had a significant effect on selection time ($F(3.53,123.68) = 9.603, p < 0.001, \eta_p^2 = 0.215$). The fastest method was *EyeGaze+*, and it was significantly faster than *HeadPose* ($-0.82\,s, p < 0.001$) and *HeadPose+* ($-0.59\,s, p = 0.001$).

Partially Occluded Object in Dense Environment METHOD had a significant effect on selection time ($F(3.52,123.33) = 21.042, p < 0.001, \eta_p^2 = 0.375$). *Controller+* was the fastest and was significantly faster than *Controller* ($-1.77\,s, p < 0.001$), *HeadPose* ($-2.52\,s, p < 0.001$) and *EyeGaze* ($-2.49\,s, p < 0.001$). There was significant differences between *EyeGaze+* and *EyeGaze* ($-2.38\,s, p < 0.001$), *HeadPose+* and *HeadPose* ($-2.03\,s, p < 0.001$). METHOD has a main effect on travelling distance ($F(2.94,102.73) = 32.552, p < 0.001, \eta_p^2 = 0.482$). Post-hoc analysis indicated that the travelling distances for all techniques with refinements were statistically smaller ($p < 0.001$ for all) than the baselines. However, the effect sizes were very small, with distance differences being $-0.068\,m$, $-0.051\,m$, and $-0.072\,m$ for head pose, eye gaze and controller based methods respectively.

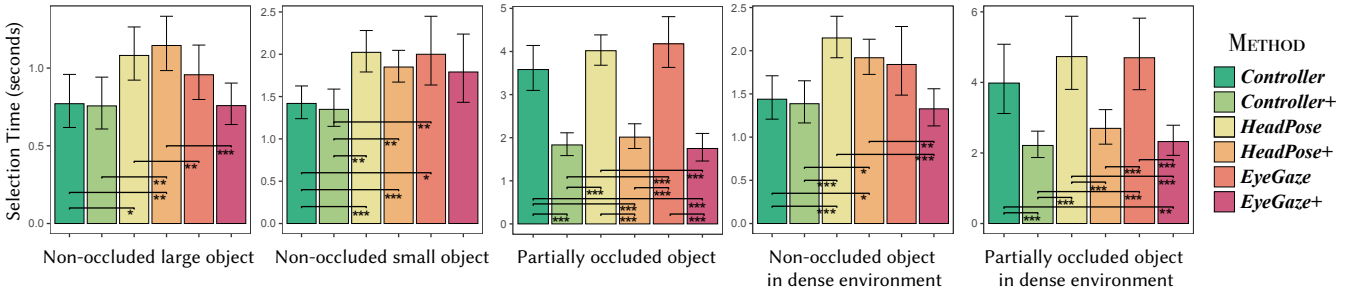


Figure 8: Selection time for methods regarding the interaction conditions with non-occluded large object, non-occluded small object, partially occluded object, non-occluded object in dense environment, and partially occluded object in dense environment. Error bars indicate the 95% confidence interval. Significance codes: $***p < .001$, $**p < .01$ and $*p < .05$.

Table 1: UEQ-S results show the pragmatic rating, hedonic rating, and overall rating of methods in the selection study.">avg." represents "above average", "<avg." stands for "below average" and "exc." means "excellent".

METHOD	Pragmatic	Hedonic	Overall
Controller	1.36 (>avg.)	0.88 (<avg.)	1.12 (>avg.)
HeadPose	1.40 (>avg.)	1.34 (>avg.)	1.37 (>avg.)
EyeGaze	1.21 (>avg.)	1.61 (good)	1.41 (>avg.)
Controller+	1.89 (exc.)	1.68 (good)	1.76 (good)
HeadPose+	1.80 (exc.)	1.76 (good)	1.78 (good)
EyeGaze+	1.99 (exc.)	1.88 (good)	1.94 (exc.)

Table 2: Subjective ranking results of the selection and movement studies. The average ranking with standard error ("Ranking"), and the times each method was on the first or second place ("#1/2") are listed.

METHOD	Selection		Movement	
	Ranking	#1/2	Ranking	#1/2
Controller	4.08 (1.04)	0/1	-	-
HeadPose	5.00 (0.71)	0/0	-	-
EyeGaze	5.50 (0.76)	0/0	-	-
Controller+	2.00 (0.71)	3/6	1.42 (0.64)	8/6
HeadPose+	2.25 (1.09)	5/0	2.25 (0.72)	5/5
EyeGaze+	2.17 (1.21)	4/5	2.33 (0.75)	2/4

Subjective Results We summarized the subjective results of UEQ-S in Table 1 and the average ranking in Table 2. The UEQ-S results revealed that *EyeGaze+* was most favored, while the subjective ranking indicated the highest ranking of *Controller+*. NASA-TLX results concluded that the refinements were better than the baselines, see supplementary material.

The open comments were mostly focused on *Controller+*, *EyeGaze+* and *HeadPose+*. Most participants ($N = 11, 91.67\%$) felt *EyeGaze+* and *Controller+* were "performing very well when selecting large object". Some participants ($N = 4, 33.33\%$) thought *EyeGaze+* was hard to select objects in dense environment. *HeadPose+* was commented "I felt tired after completing a lot of tasks" ($N = 3, 25.00\%$). In summary, the comments on improved methods were more positive than those on original ones.

Most of the participants ($N = 10, 83.33\%$) confirmed that the integration of contextual information improved the baseline methods well. Almost all participants ($N = 11, 91.67\%$) thought scene context was helpful when selecting occluded objects. However, when selecting non-occluded large objects, some ($N = 9, 75.00\%$) thought scene context only slightly improved the performance.

5.5 Discussion

The experimental results showed that the improved methods generally performed better than the original ones. The performance difference is particularly significant for complex tasks such as selecting a partially occluded target in dense environment.

For non-occluded large objects, *Controller+* performed best, followed by *EyeGaze+* and *Controller*. Other selection methods were similar. As the selection task is not difficult in this case, *Controller+* was more flexible.

In the case of non-occluded small objects, *Controller+* performed best. This was because the accuracy of eye tracking was limited, and head tracking only allowed slow rotation. Selecting small objects usually doubled the time than large objects, or even worse.

For partially occluded objects, the improved methods preformed better by halving the selection time. The performance of *EyeGaze+* and *Controller+* were almost the same, and *EyeGaze+* was slightly better. Travelling distances were also shortened, showing that the dependence relationship of occluded objects really helped.

Regarding non-occluded objects in dense environment, the results were similar to small objects. This was due to the fact that selecting these objects was basically the same as selecting small objects, except that there were more surrounding objects. *EyeGaze+* and *Controller+* took less time than others, which demonstrated that the group relationship was helpful.

Finally, for partially occluded objects in dense environment, our refinements also performed better than baselines. *Controller+* was the best, followed by *EyeGaze+*. This was because occlusion caused by distractors made the selection task challenging. It was easy to switch selected objects based on dependence or group relationships.

The NASA-TLX and UEQ-S results also showed similar statistics. Participants were more satisfied when using *Controller+*. *EyeGaze* and *HeadPose* were not convenient to select objects under some complex conditions. Participants preferred the improved methods with scene context as it was convenient to understand and use, making the selection more efficient. In summary, *Controller+* had the lowest selection time, and *EyeGaze+* had the lowest travelling distance. The refinements were better than the baselines when dealing with occluded objects or objects with distractive surroundings.

6 EXPERIMENT 2: MOVEMENT STUDY

Apart from selection, we also studied indoor object movement using controller, head pose, and eye gaze based methods with the help of scene contexts.

6.1 Participants and Apparatus

We recruited 15 participants (4F/11M, Mean age = 21, SD = 1.68) from a university. 10 participants had prior VR experience. The experiment duration ranged from 20 minutes to 30 minutes. The apparatus for implementation and experiment is the same as before.

6.2 Movement Methods

Three context-aware indoor object movement METHODS were investigated as follows:

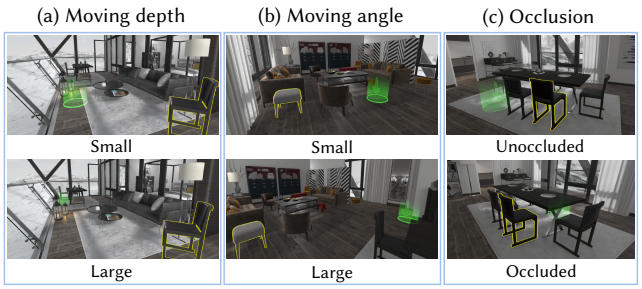


Figure 9: Conditions for movement task. (a) moving depth (small and large). (b) moving angle (small and large). (c) occlusion of target position (unoccluded and occluded). Objects highlighted with yellow silhouette indicate the target object for moving, and green circles with an arrow indicate its target position.

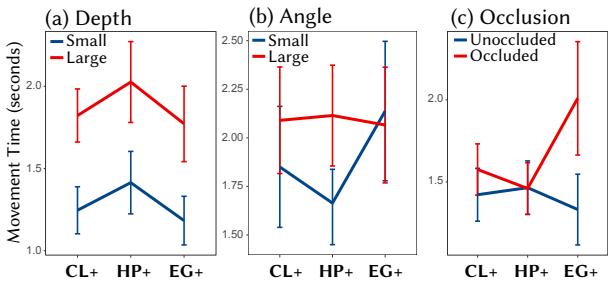


Figure 10: The movement time of methods on different levels of (a) moving depth, (b) moving angle, and (c) occlusion of the target position. "CL+" represents *Controller+*, "HP+" represents *HeadPose+*, and "EG+" stands for *EyeGaze+*. Error bars indicate the 95% confidence interval.

Controller+: The user needs to select the target object first. Then by moving the controller, the object will be moved to the position where the ray emitted from the controller intersects the scene. The user can also travel in the scene to better adjust the intersection position. The movement is successful when the user put down the object in the vicinity of the target position via pressing the trigger.

HeadPose+: Same as the previous method, the user needs to select the target object via head-based selection. The object will then be moved to the indicated position of the ray from the head, until the same condition of success is met.

EyeGaze+: This method is the same as *HeadPose+*, except that raycasting based selection and movement is determined by eye gaze instead of head pose.

Remarks. To focus on the movement test and facilitate user interaction, the above methods do not require any extra operation (e.g., switch objects) from the user. Also, the user does not need to laboriously fine-tune object movement, especially when the target position is occluded by other objects in the scene. This is because the position and orientation of the object are already pre-computed based on scene contexts when applicable (e.g., rotational/reflective symmetry of chairs w.r.t. the table, sofa, and wall alignment). Besides, to eliminate the influence of selection before movement, the target object is prepared to be easy for the user to select. There is no distraction or occlusion. To ensure a fair comparison on object movement between different methods, we only counted the movement time once the target object was picked up, until a successful placement on the target position.

6.3 Design and Procedure

As shown in Figure 9, for each trial, the user needs to move the target object to the target position. Both targets are highlighted so that the user clearly understands the task. We chose three experimental variables to evaluate indoor object movement as follows:

- ***Moving depth*** (DEPTH): the distance between the initial position and the target position in depth, which is mainly related to the elevation of the ray pointing to the scene.
- ***Moving angle*** (ANGLE): the horizontal change between the initial position and the target position, which is mainly related to the azimuth of the ray pointing to the scene. Once the target position is out of the current view, an arrow is shown to indicate the moving direction.
- ***Occlusion of target position*** (OCCLUSION): the target position can often be occluded by other objects in the scene (e.g., move chair behind table, move corner table behind sofa, etc.), which mainly affects the intersection between the ray and the scene.

Under each variable, there were 2 cases (small/large, small/large, unoccluded/occluded) in 2 scenarios. With 3 repetitions, each participant completed 3 methods $\times$ 3 variables $\times$ 2 cases $\times$ 2 scenarios $\times$ 3 repetitions = 108 trials. Similar to the selection study, the movement study consists of three phases: training, experiment, and interview. The participant first familiarized himself/herself with the test environment and procedure (see Figure 7(b)), and filled in the relevant information. Then in the formal tests, the participant practiced each technique with different parameter settings in random order. During the experiment, we recorded the time for movement and the travelling distance from the participant. After the experiment, the participant was asked to complete a subjective questionnaire and attend a short interview. A voucher of 5$ was given to the participant for acknowledgment.

6.4 Results

To analyze the movement time and travelling distance, we discarded the outliers of 47 trails (4.35%). Figure 10 shows the movement time for tested methods (travelling distance results are provided in the supplementary file). We used a repeated-measures ANOVA (with Greenhouse-Geisser correction) for effect analysis and Bonferroni-adjusted pairwise comparisons for post-hoc analysis. Because we are aiming at investigating the difference between methods, we analyze the interaction effects and main effects related to METHOD.

Movement Time No significant METHOD $\times$ DEPTH effect was found ($F(2,58) = 0.063, p = 0.939, \eta_p^2 = 0.002$). METHOD had a significant main effect on the movement time ($F(2,58) = 7.649, p = 0.001, \eta_p^2 = 0.209$). *EyeGaze+* was the fastest but only significantly faster than *HeadPose+* ($-0.24\,s, p = 0.004$).

A significant interaction effect of METHOD $\times$ ANGLE was found ($F(1.43,41.46) = 7.110, p = 0.005, \eta_p^2 = 0.197$). METHOD had a significant effect only under small angle condition ($F(2,58) = 6.929, p = 0.002$). *HeadPose+* performed best, significantly faster than *EyeGaze+* ($-0.49\,s, p = 0.003$).

There was a significant interaction effect on movement time of METHOD $\times$ OCCLUSION ($F(1.33,38.64) = 12.774, p < 0.001, \eta_p^2 = 0.306$). METHOD had a significant effect on movement time when moving an object onto occluded target position ($F(1.23,35.59) = 6.67, p = 0.01$). *HeadPose+* was the fastest, and being significantly faster than *EyeGaze+* ($-0.55\,s, p = 0.016$).

Travelling Distance Although a statistically significant interaction effect of METHOD $\times$ ANGLE was found ($F(2,58) = 8.047, p = 0.001, \eta_p^2 = 0.217$), the travelling distance differences were very small (less than $0.04\,m$), thus resulting in a very small effect size.

Subjective Results As in the selection study, we collected subjective responses to UEQ-S and NASA-TLX questionnaires. Table 3 lists the UEQ-S results (see supplementary file for the NASA-TLX results). The rankings of the methods are given in Table 2.

Most participants ($N = 11, 73.33\%$) felt *HeadPose+* "can accurately move the object to its destination". However, half participants ($N = 8, 53.33\%$) complained about *HeadPose+* that "I felt tired

Table 3: The UEQ-S results of the movement study. ">avg." means "above average" and "exc." stands for "excellent".

METHOD	Pragmatic	Hedonic	Overall
Controller+	1.96 (exc.)	1.45 (>avg.)	1.71 (good)
HeadPose+	1.81 (exc.)	1.70 (good)	1.75 (good)
EyeGaze+	1.70 (good)	1.78 (good)	1.74 (good)

after completing a lot of tasks". Besides, participants thought *EyeGaze+* was disappointing in moving objects to target position that was occluded ($N = 12, 80\%$), with large angle ($N = 13, 86.67\%$), and with long distance ($N = 12, 80\%$), but felt that *HeadPose+* and *Controller+* performed very well in general. The feedback on *HeadPose+* was the most positive except that it was tiring.

6.5 Discussion

The experimental results showed that by taking into account scene contexts, users can focus on selecting an object and moving it to the target position without tedious fine-tuning on object location and orientation. Moving depth had a similar impact on all methods. As it increased, the movement time of all three techniques also increased significantly. There seems to be a linear correlation in-between. *Controller+* and *HeadPose+* performed best here. Moving angle influenced *HeadPose+* a lot, but not for *EyeGaze+* and *Controller+*. One potential reason is that *HeadPose+* is more tiring, while *EyeGaze+* and *Controller+* are more flexible without rotating the head by a large angle. Besides, scene contexts also benefit the situation when the target position is occluded, which is comparable to cases with no occlusion. This is because the target object can only be placed at where it can be supported in the right context. For example, a chair should be placed on the floor (behind a table) but not on the table. It also showed that *Controller+* and *HeadPose+* are more stable than *EyeGaze+*, as it is difficult to gaze at the occluded target position in *EyeGaze+*.

7 APPLICATION

The selection and movement studies showed that the refined methods with scene contexts generally performed well. We further implemented a furniture rearrangement application based on the selection and movement interactions, as a preliminary demonstration of our approach. Compared to the user studies, the application reflected the performance of tested methods in a more complex scenario.

In this VR application, a set of furniture objects were initially placed outside a virtual room, the task for the user was to select and move each furniture object into the room sequentially, following the target layout that was displayed as a picture on the wall (see Figure 11 and the supplementary video). Scene contexts that existed in this application includes: 1) dependence relationship: sofa and hanger, chair and footrest; 2) group relationship: candles, potted plants; and 3) spatial relationship: desk and sofa, desk and chair. We conducted a small-scale user experience experiment with three participants (all male). Each participant was initiated with his location outside the room. The task was to select, move and rearrange the furniture object as fast as possible. The participant was allowed to decide the order of objects for selection and movement. The average time for completing the furniture rearrangement task using *Controller+*, *HeadPose+* and *EyeGaze+* was 94.5 seconds, 127.4 seconds and 91.8 seconds, while the average travelling distances were 1.8 meters, 1.7 meters and 3.1 meters, respectively.

Participants overall preferred *Controller+* most, because using the controller, they were able to select and move the target object more flexibly and frequently in the overall application about selection and movement. We clarify that if the user prefers a specific object orientation, he/she can easily make an adjustment after the movement operation. However, we did not observe such intention and feedback from users during the experiment.

(a) Before rearrangement (b) After rearrangement

Figure 11: Virtual furniture rearrangement application. The user selects furniture objects outside a room and moves them into the room following the target layout illustrated as a picture on the wall. (a) Furniture objects before rearrangement; (b) rearranged furniture objects.

8 LIMITATIONS AND DISCUSSION

Our work has limitations. First, we only used controller buttons to access the scene context. Notwithstanding it was a natural way to blend context with traditional selection methods, other approaches can be further investigated, such as using voices or gloves, or visualizing the context via user interfaces (e.g., zoomed-in region, overhead view, etc.). Second, in our experiment, only typical contexts of indoor scenes were used in our selection and movement tasks. For a totally new scene with an unfamiliar context, the user may need time to learn how to use the scene context. Third, we acknowledge that scene context was not fully exploited, some more global contextual information, such as the overall balance of furniture objects, the functional relationships between objects (sofa in front of TV set), can be further explored in the future. Meanwhile, the tested tasks involved only two levels of hierarchy. Nevertheless, we clarify that our refinements are compatible with multiple levels of hierarchy.

In this paper, we presented an indoor object selection and movement approach by further considering its surrounding scene context, including object semantics and object interrelationships. We employed our approach on three major object selection and movement methods based on handheld controller, head pose, and eye gaze. To compare different methods and validate the effectiveness of our approach, we performed indoor object selection and movement studies under various conditions on object size/occlusion/distraction, and movement distance/angle/occlusion. The selection study showed that *Controller+* and *EyeGaze+* generally performed better than the other methods. *Controller+* was mostly favored by participants The movement study showed that *Controller+* performed best in general. In summary, scene contexts are confirmed to be effective in reducing the interaction time, especially in complex interaction tasks with occluded object and dense environment.

To investigate how the selection and movement techniques perform in complex interaction scenarios, we proposed to let the users freely combine and apply selection and movement interactions in a furniture arrangement application. Results showed that although *HeadPose+* and *EyeGaze+* performed well in the selection and movement studies, *Controller+* was more stable in the overall performance, with higher user preference.

There are still many issues worthy of further discussion and research, such as raycasting stability in long distance, intuitive semantics for understanding, etc. Besides, we expect a higher accuracy of eye gaze to further improve the performance of *EyeGaze+*, with the development of eye tracking technology. Finally, we believe that scene context will benefit human-object interaction not only for raycasting techniques, but also for gesture and haptics based methods in the future.

ACKNOWLEDGMENTS

The authors wish to thank the anonymous reviewers for their helpful advices. This work was supported by the National Natural Science Foundation of China (Project Number: 61902012 and 61932003) and Baidu academic collaboration program. Yong-Liang Yang was supported by RCUK grant CAMERA (EP/M023281/1, EP/T014865/1), and a gift from Adobe.

REFERENCES

[1] F. Argelaguet and C. Andujar. A survey of 3D object selection techniques for virtual environments. *Computers & Graphics*, 37(3):121 – 136, 2013.

[2] F. Argelaguet and C. Andujar. A survey of 3d object selection techniques for virtual environments. *Computers & Graphics*, 37(3):121–136, 2013.

[3] F. Bacim, R. Kopper, and D. A. Bowman. Design and evaluation of 3d selection techniques based on progressive refinement. *International Journal of Human-Computer Studies*, 71(7):785 – 802, 2013.

[4] M. Baloup, T. Pietrzak, and G. Casiez. Raycursor: A 3d pointing facilitation technique based on raycasting. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

[5] R. Bates and H. O. Istance. Why are eye mice unpopular? a detailed comparison of head and eye controlled assistive technology pointing devices. *Universal Access in the Information Society*, 2(3):280–290, 2003.

[6] A. M. Bernardos, D. Gómez, and J. R. Casar. A comparison of head pose and deictic pointing interaction methods for smart environments. *International Journal of Human-Computer Interaction*, 32(4):325–351, 2016.

[7] L. Besançon, P. Issartel, M. Ammi, and T. Isenberg. Mouse, tactile, and tangible input for 3d manipulation. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 4727–4740, 2017.

[8] L. Besançon, M. Sereno, L. Yu, M. Ammi, and T. Isenberg. Hybrid touch/tangible spatial 3d data selection. In *Computer Graphics Forum*, vol. 38, pp. 553–567. Wiley Online Library, 2019.

[9] R. A. Bolt. Gaze-orchestrated dynamic windows. *ACM SIGGRAPH Computer Graphics*, 15(3):109–119, 1981.

[10] R. A. Bolt. Eyes at the interface. In *Proceedings of the 1982 conference on Human factors in computing systems*, pp. 360–362, 1982.

[11] R. W. Bukowski and C. H. Séquin. Object associations: a simple and practical approach to virtual 3d manipulation. In *Proceedings of the 1995 symposium on Interactive 3D graphics*, pp. 131–ff, 1995.

[12] D. L. Chen, R. Balakrishnan, and T. Grossman. Disambiguation techniques for freehand object manipulations in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 285–292, 2020.

[13] R. M. Clifford, N. M. B. Tuanquin, and R. W. Lindeman. Jedi forceextension: Telekinesis as a virtual reality interaction metaphor. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 239–240, 2017.

[14] N. Cournia, J. D. Smith, and A. T. Duchowski. Gaze-vs. hand-based pointing in virtual environments. In *CHI'03 extended abstracts on Human factors in computing systems*, pp. 772–773, 2003.

[15] N.-T. Dang, H.-H. L. H.-H. Le, and M. Tavanti. Visualization and interaction on flight trajectory in a 3d stereoscopic environment. In *Digital Avionics Systems Conference, 2003. DASC'03. The 22nd*, vol. 2, pp. 9–A, 2003.

[16] G. De Haan, M. Koutek, and F. H. Post. Intenselect: Using dynamic object rating for assisting 3d object selection. In *Ipt/egve*, pp. 201–209. Citeseer, 2005.

[17] H. G. Debarba, J. G. Grandi, A. Maciel, L. Nedel, and R. Boulic. Disambiguation canvas: a precise selection technique for virtual environments. In *IFIP Conference on Human-Computer Interaction*, pp. 388–405. Springer, 2013.

[18] H. Dehmeshki and W. Stuerzlinger. Intelligent mouse-based object group selection. In *International Symposium on Smart Graphics*, pp. 33–44. Springer, 2008.

[19] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6):381, 1954.

[20] M. Gosele, W. Stürzlinger, and U. O. Ulm. Semantic constraints for scene manipulation. In *in Proceedings Spring Conference in Computer Graphics'99 (Budmerice, Slovak Republic*, pp. 140–146, 1999.

[21] T. Grossman and R. Balakrishnan. The design and evaluation of selection techniques for 3d volumetric displays. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, pp. 3–12, 2006.

[22] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, pp. 904–908, 2006.

[23] R. J. Jagacinski and D. L. Monk. Fitts' law in two dimensions with hand and head movements movements. *Journal of motor behavior*, 17(1):77–95, 1985.

[24] S. Jalaliniya, D. Mardanbegi, and T. Pederson. Magic pointing for eyewear computers. In *Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pp. 155–158, 2015.

[25] S. Jalaliniya, D. Mardanbeigi, T. Pederson, and D. W. Hansen. Head and eye movement as pointing modalities for eyewear computers. In *2014 11th International Conference on Wearable and Implantable Body Sensor Networks Workshops*, pp. 50–53, 2014.

[26] H. J. Kang, J.-h. Shin, and K. Ponto. A comparative analysis of 3d user interaction: How to move virtual objects in mixed reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 275–284, 2020.

[27] R. Kopper, F. Bacim, and D. A. Bowman. Rapid and accurate 3d selection by progressive refinement. In *2011 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 67–74, 2011.

[28] M. Kytö, B. Ens, T. Piumsomboon, G. A. Lee, and M. Billinghurst. Pinpointing: Precise head-and eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2018.

[29] J. J. LaViola Jr, E. Kruijff, R. P. McMahan, D. Bowman, and I. P. Poupyrev. *3D user interfaces: theory and practice*. Addison-Wesley Professional, 2017.

[30] C. J. Lin, S.-H. Ho, and Y.-J. Chen. An investigation of pointing postures in a 3d stereoscopic environment. *Applied ergonomics*, 48:154–163, 2015.

[31] Y. Lu, C. Yu, and Y. Shi. Investigating bubble mechanism for ray-casting to improve 3d target acquisition in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 35–43.

[32] R. Ma, A. G. Patil, M. Fisher, M. Li, S. Pirk, B.-S. Hua, S.-K. Yeung, X. Tong, L. Guibas, and H. Zhang. Language-driven synthesis of 3d scenes from scene databases. *ACM Transactions on Graphics (TOG)*, 37(6):1–16, 2018.

[33] I. S. MacKenzie. Evaluating eye tracking systems for computer input. In *Gaze interaction and applications of eye tracking: Advances in assistive technologies*, pp. 205–225. IGI Global, 2012.

[34] P. Merrell, E. Schkufza, Z. Li, M. Agrawala, and V. Koltun. Interactive furniture layout using interior design guidelines. *ACM transactions on graphics (TOG)*, 30(4):1–10, 2011.

[35] M. R. Mine, F. P. Brooks Jr, and C. H. Sequin. Moving objects in space: exploiting proprioception in virtual-environment interaction. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pp. 19–26, 1997.

[36] J.-Y. Oh and W. Stuerzlinger. Moving objects with 2d input devices in cad systems and desktop virtual environments. In *Proceedings of Graphics Interface 2005*, pp. 195–202, 2005.

[37] J.-Y. Oh, W. Stuerzlinger, and D. Dadgari. Group selection techniques for efficient 3d modeling. In *3D User Interfaces (3DUI'06)*, pp. 95–102. IEEE, 2006.

[38] I. Poupyrev, M. Billinghurst, S. Weghorst, and T. Ichikawa. The go-go interaction technique: non-linear mapping for direct manipulation in vr. In *Proceedings of the 9th annual ACM symposium on User interface software and technology*, pp. 79–80, 1996.

[39] Y. Y. Qian and R. J. Teather. The eyes don't have it: an empirical comparison of head-based and eye-based selection in virtual reality. In *Proceedings of the 5th Symposium on Spatial User Interaction*, pp. 91–98, 2017.

[40] H. Ro, S. Chae, I. Kim, J. Byun, Y. Yang, Y. Park, and T. Han. A dynamic depth-variable ray-casting interface for object manipulation in ar environments. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2873–2878, 2017.

[41] M. S. M. Y. Sait, S. P. Sargunam, D. T. Han, and E. D. Ragan. Physical hand interaction for controlling multiple virtual objects in virtual reality. In *Proceedings of the 3rd International Workshop on Interactive and Spatial Computing*, pp. 64–74, 2018.

[42] T. Salzman, S. Stachniak, and W. Stürzlinger. Unconstrained vs. constrained 3d scene manipulation. In *IFIP International Conference on Engineering for Human-Computer Interaction*, pp. 207–219. Springer, 2001.

[43] M. Schrepp, A. Hinderks, and J. Thomaschewski. Design and evaluation of a short version of the user experience questionnaire (ueq-s). *IJIMAI*, 4(6):103–108, 2017.

[44] G. Shan, M. Xie, Y. Gao, X. Chi, et al. Interactive visual exploration of halos in large-scale cosmology simulation. *Journal of Visualization*, 17(3):145–156, 2014.

[45] G. Smith and W. Stuerzlinger. On the Utility of Semantic Constraints. In B. Froehlich, J. Deisinger, and H.-J. Bullinger, eds., *Eurographics Workshop on Virtual Environments*. The Eurographics Association, 2001. doi: 10.2312/EGVE/EGVE01/041-050

[46] G. Smith, W. Stuerzlinger, and T. Salzman. 3d scene manipulation with 2d devices and constraints. In *Proceedings of the Graphics Interface 2001 Conference, June 7-9 2001, Ottawa, Ontario, Canada*, pp. 135–142, June 2001.

[47] M. Speicher, F. Daiber, G.-L. Kiefer, and A. Krüger. Exploring task performance and user's preference of mid-air hand interaction in a 3d docking task experiment. In *Proceedings of the 5th symposium on spatial user interaction*, pp. 160–160, 2017.

[48] A. Steed. Towards a general model for selection in virtual environments. In *3D user interfaces (3DUI'06)*, pp. 103–110, 2006.

[49] W. Stuerzlinger and G. Smith. Efficient manipulation of object groups in virtual environments. In *Proceedings IEEE Virtual Reality 2002*, pp. 251–258. IEEE, 2002.

[50] J. Sun and W. Stuerzlinger. Extended sliding in virtual reality. In *25th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–5, 2019.

[51] J. Sun and W. Stuerzlinger. Selecting and sliding hidden objects in 3d desktop environments. In *Proceedings of Graphics Interface 2019*, GI 2019. Canadian Information Processing Society, 2019. doi: 10.20380/GI2019.08

[52] J. Sun, W. Stuerzlinger, and B. E. Riecke. Comparing input methods and cursors for 3d positioning with head-mounted displays. In *Proceedings of the 15th ACM Symposium on Applied Perception*, pp. 1–8, 2018.

[53] J. Sun, W. Stuerzlinger, and D. Shuralyov. Shift-sliding and depth-pop for 3d positioning. In *Proceedings of the 2016 Symposium on Spatial User Interaction*, pp. 69–78, 2016.

[54] R. J. Teather and W. Stuerzlinger. Guidelines for 3d positioning techniques. In *Proceedings of the 2007 conference on Future Play*, pp. 61–68, 2007.

[55] E. Tse, M. Hancock, and S. Greenberg. Speech-filtered bubble ray: improving target acquisition on display walls. In *Proceedings of the 9th international conference on Multimodal interfaces*, pp. 307–314, 2007.

[56] V. Vuibert, W. Stuerzlinger, and J. R. Cooperstock. Evaluation of docking task performance using mid-air interaction techniques. In *Proceedings of the 3rd ACM Symposium on Spatial User Interaction*, pp. 44–52, 2015.

[57] K. Wang, Y.-A. Lin, B. Weissmann, M. Savva, A. X. Chang, and D. Ritchie. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. 38(4), 2019.

[58] K. Wang, M. Savva, A. X. Chang, and D. Ritchie. Deep convolutional priors for indoor scene synthesis. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

[59] M. Wang, X.-Q. Lyu, Y.-J. Li, and F.-L. Zhang. VR content creation and exploration with deep learning: A survey. *Computational Visual Media*, 6(1):3–28, 2020.

[60] T. Weiss, A. Litteneker, N. Duncan, M. Nakada, C. Jiang, L.-F. Yu, and D. Terzopoulos. Fast and scalable position-based layout synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 25(12):3231–3243, 2018.

[61] K. Xu, J. Stewart, and E. Fiume. Constraint-based automatic placement for scene composition. In *Proceedings of the Graphics Interface 2002 Conference, May 27-29, 2002, Calgary, Alberta, Canada*, pp. 25–34, May 2002.

[62] W. Xu, H.-N. Liang, Y. Zhao, T. Zhang, D. Yu, and D. Monteiro. RingText: Dwell-free and hands-free text entry for mobile head-mounted displays using head motions. *IEEE transactions on visualization and computer graphics*, 25(5):1991–2001, 2019.

[63] D. Yu, H.-N. Liang, X. Lu, K. Fan, and B. Ens. Modeling endpoint distribution of pointing selection tasks in virtual reality environments. *ACM Trans. Graph.*, 38(6), Nov. 2019.

[64] D. Yu, Q. Zhou, J. Newn, T. Dingler, E. Velloso, and J. Goncalves. Fully-occluded target selection in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2020.

[65] L. Yu, K. Efstathiou, P. Isenberg, and T. Isenberg. Cast: Effective and efficient user interaction for context-aware selection in 3d particle clouds. *IEEE transactions on visualization and computer graphics*, 22(1):886–895, 2015.

[66] L. F. Yu, S. K. Yeung, C. K. Tang, D. Terzopoulos, T. F. Chan, and S. J. Osher. Make it home: automatic optimization of furniture arrangement. *ACM Transactions on Graphics (TOG)-Proceedings of ACM SIGGRAPH 2011, v. 30,(4), July 2011, article no. 86*, 30(4), 2011.

[67] S.-H. Zhang, S.-K. Zhang, Y. Liang, and P. Hall. A survey of 3D indoor scene synthesis. *Journal of Computer Science and Technology*, 34(3):594, 2019.