4D Attention: Comprehensive Framework for Spatio-Temporal Gaze Mapping

Shuji Oishi[1], Kenji Koide[1], Masashi Yokozuka[1], and Atsuhiko Banno[1]

Abstract—This study presents a framework for capturing human attention in the spatio-temporal domain using eye-tracking glasses. Attention mapping is a key technology for human perceptual activity analysis or Human-Robot Interaction (HRI) to support human visual cognition; however, measuring human attention in dynamic environments is challenging owing to the difficulty in localizing the subject and dealing with moving objects. To address this, we present a comprehensive framework, *4D Attention*, for unified gaze mapping onto static and dynamic objects. Specifically, we estimate the glasses pose by leveraging a loose coupling of direct visual localization and Inertial Measurement Unit (IMU) values. Further, by installing reconstruction components into our framework, dynamic objects not captured in the 3D environment map are instantiated based on the input images. Finally, a scene rendering component synthesizes a first-person view with identification (ID) textures and performs direct 2D-3D gaze association. Quantitative evaluations showed the effectiveness of our framework. Additionally, we demonstrated the applications of 4D Attention through experiments in real situations[1].

Index Terms—Localization, Visual Tracking, Intention Recognition, Multi-Modal Perception for HRI

I. INTRODUCTION

> "*The eyes which are the windows of the soul.*"
> — Plato (427 BC - 347 BC)

E YE movements are crucial but implicit cues for determining people's attention. Gaze estimation enables the study of visual perception mechanisms in humans, and has been used in many fields, such as action recognition [1], situation awareness estimation [2], and driver attention analysis [3]. It is also a non-verbal communication method, and thus, it can be applied to shared autonomy [4] or teleoperation [5] in the context of Human-Robot Interaction (HRI).

Recent studies have enabled human attention mapping in 3D environments using mobile eye-tracking glasses [6] [7]. Most approaches compute a 3D gaze by extending a measured 2D gaze vector from a camera pose estimated by visual

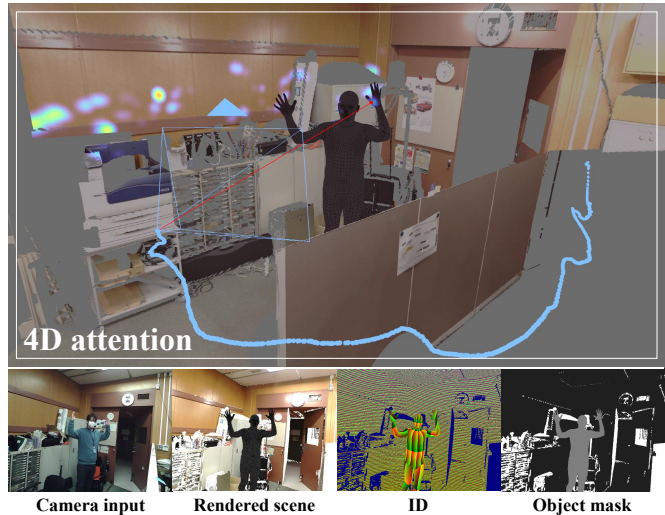[1]See the attached video or `https://youtu.be/NG2V3Qs_L8g` for what 4D Attention can offer.

Fig. 1. 4D attention analysis in a dynamic environment. Given first-person view with the subject's 2D gaze, it projects human attention onto the static 3D map and dynamic object models employing visual localization, rapid intersection search, and instance object reconstruction.

Camera input   Rendered scene   ID   Object mask

localization or motion capture systems in a pre-built static 3D map. They are assumed to operate in static environments; however, the real world is a place of constant change, with objects appearing and disappearing from the scenes. Human attention analysis in both spatial and temporal domains is still an open problem, which when solved will help determine human behavior in the real world.

To address this issue, we propose a comprehensive framework for 4D attention mapping (see Fig.1). The main contributions of this study are three-fold:

- A new framework, *4D Attention*, is proposed for capturing human attention to static and dynamic objects by assembling 6-DoF camera localization, rapid gaze projection, and instant dynamic object reconstruction. Human attention is accumulated on each 3D mesh model, which makes gaze mapping much more meaningful, for example, the semantic analysis of perceptual activities rather than generating cluttered 3D gaze point clouds.
- The framework is designed so that scene rendering plays a central role. This makes the entire system simple and does not require additional map or object model representations for localization and attention mapping. Additionally, it facilitates a unified attention-mapping procedure regardless of the target objects.
- We examined the accuracy and precision of our method using a moving target board whose ground truth position
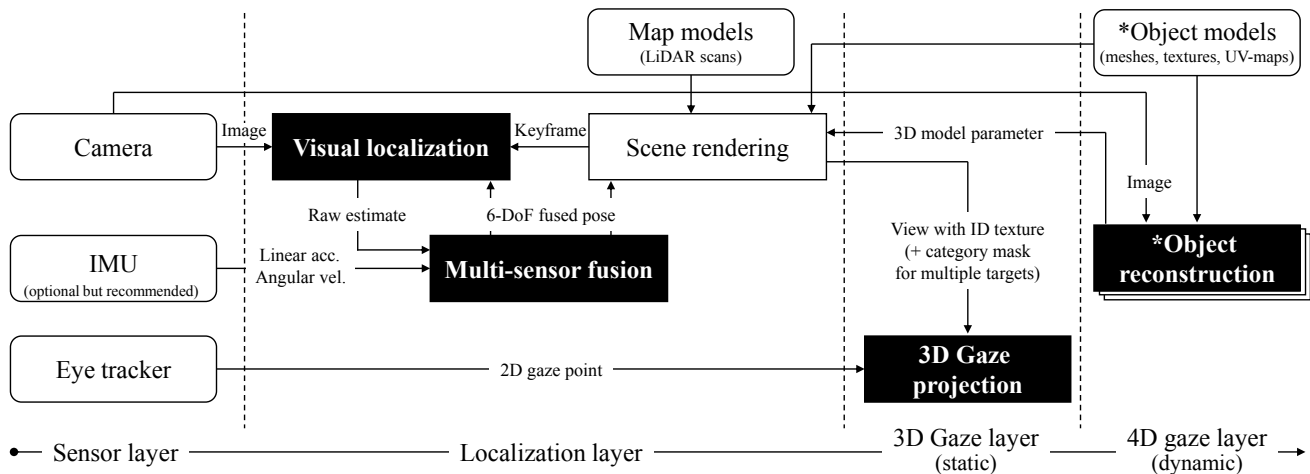
Fig. 2. Overview of 4D Attention: In the localization layer, we compute the pose of the eye-tracker with C* [8] based on images from a scene camera. Fusing IMU data with the raw pose estimate can further boost and stabilize visual localization [9]. In the 3D gaze layer, the intersection of a gaze ray and the environment map is calculated using the direct 2D-3D gaze association via scene rendering with an ID texture. The 4D gaze layer incorporates any object reconstruction components into the framework to instantiate dynamic objects, which facilitates the analysis of spatio-temporal human attention in the real world.

was measured by a total station. Additional experiments for monitoring human attention in the real world demonstrated the capability of analyzing human attention in static and dynamic targets including maps, household items, and people, during the free movement of the subject.

II. RELATED WORK

Eye movement patterns: Eye movements imply visual perception activities. Several approaches have inferred or determined perceptual activities based on the observations from electrooculography (EOG). Bulling *et al.* [10] developed a pioneering system that classifies several activities from eye movement patterns by utilizing machine learning. Ishimaru *et al.* [11] also determined daily activities including typing, reading, eating, and talking, using signals from EOG glasses. This approach allows us to identify the current activity of a subject without complex settings, and can be applied to HCI to provide relevant services.

2D contextual analysis: However, human beings live in a context. Visual perception activities are not independent of the surrounding environment; in fact, they are induced by "attractive" objects in the scene. Eye-tracking and gaze overlay on 2D camera views make it possible to determine the focus of the subject, as in [12]. For semantic human attention analysis in natural environments, Fritz and Paletta [13] introduced object recognition in mobile eye tracking using local image descriptors. A similar approach can be observed in [14], which identifies objects fixated by the subject for a museum guide. [15] further progressed toward online object-of-interest recognition using a hierarchical visual feature representation.

3D gaze mapping: For the holistic estimation of human attention, recent techniques have attempted to obtain fixations in the real 3D world leaving the image plane. [16] and [2] extended 2D gaze mapping by combining it with a motion capture system to track the pose of gaze glasses, which

enables the measurement of the 3D point of interest. [17] built a similar system relying on visual markers for monocular camera tracking and 3D gaze analysis. However, they require a complex setup of multiple sensors, making the measurement area small and unscalable to large environments. Thus, several approaches compute the 3D gaze by localizing an agile monocular camera using visual localization or structure-from-motion. [6] was the pioneering work, and was followed by promising techniques such as [7], [18] where they estimated camera poses using visual features and projected 3D gaze information onto the pre-built 3D environment map.

Toward attention analysis in the real world: 3D gaze mapping facilitates the analysis of human attention regardless of the scale of the environment; however, they still operate only in the static environment. Attention analysis in dynamic situations is still an open problem; it is necessary to address the *spatio-temporal* attention analysis to truly comprehend perceptual activities in the real world.

III. PROPOSED METHOD

A. System overview

In this study, we propose a comprehensive framework to capture *4D human attention*, which is attention in the spatial and temporal domains in dynamic environments. A schematic overview of the proposed system is depicted in Fig.2. Obtaining 4D human attention from eye-tracking glasses with a scene camera has three main problems that need to be solved: robust camera localization, rapid 3D gaze mapping, and instant processing of dynamic objects.

Principally, 4D attention mapping is performed by projecting a first-person 2D human gaze onto a 3D environment map (static) and moving objects (dynamic). It first requires accurate and stable 6-DoF camera localization even in dynamic environments, which means that appearance of the pre-built 3D map and current view can be significantly changed. Additionally, given the camera pose, the system has to compute
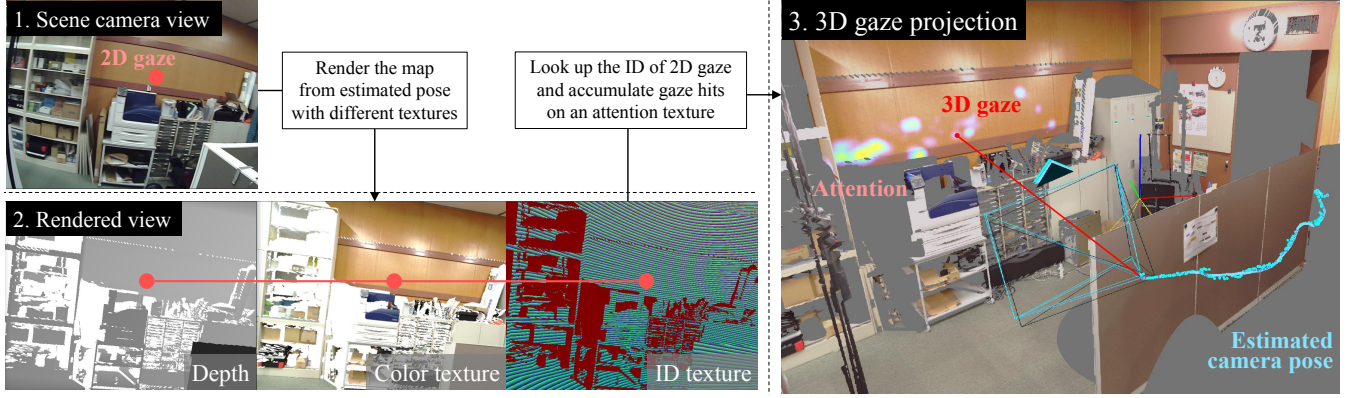
Fig. 3. ID texture mapping for direct 2D and 3D gaze association: The scene rendering component synthesizes an image from the estimated camera pose to simulate the first-person view. Different textures help to comprehend the scene. Here, we attach an ID texture to the 3D environment map to directly look up the surface with which the input 2D gaze intersects. Gaze hits are accumulated on an attention texture [19], which simplifies the management of human attention information.

the intersection of the gaze ray and target object surface in real-time to record the 3D distribution of the subject's interest. Furthermore, dynamic objects such as humans or daily objects should not stay in the same position, but should rather change their poses. Therefore, they cannot be captured in the 3D map in advance; instead, they should be processed on the fly.

In this section, we describe the major components of the framework shown in Fig.2 that are assembled to address these issues and capture 4D attention in the real world.

B. Localization

1) Monocular camera localization: Visual localization is used to infer the pose of an agile monocular camera in a given 3D map. It can be categorized as either indirect methods via feature point matching, or direct methods via appearance comparison. Although major 3D gaze mapping methods [7] [18] rely on indirect methods to estimate the camera pose, they require the construction and maintenance of an extra feature point 3D map for localization. As will be explained later in Section III-C, the subject's gaze is projected and accumulated on the dense 3D environment map (or dynamic object models); thus, the requirement doubles the map building cost. It also incurs other problems such as a 7-DoF exact alignment (including scale) between the environment and feature point maps.

Therefore, for a simple and straightforward system, we employ a direct localization method, specifically C* [8], which facilitates the localization of the agile monocular camera only with the colored 3D environment map. It leverages the information-theoretic cost, the Normalized Information Distance (NID), to directly evaluate the appearance similarity between the current camera view and 3D map. It achieves high robustness to large appearance changes owing to lighting conditions, dynamic obstacles, or different sensor properties [8], and results in minimal effort in map management.

Given the current view I_t, C* estimates the camera pose $\mathbf{T}_W^t$ in the world coordinate system W via $\mathbb{SE}(3)$ local

tracking against a synthetic key frame I_k rendered at a known pose $\mathbf{T}_W^k$:

$$\hat{\mathbf{T}}_k^t = \arg\min_{\mathbf{T}_k^t} \ \delta I_{NID}\left(I_t, I_k, \mathbf{T}_k^t\right),$$
$$\mathbf{T}_W^t = \hat{\mathbf{T}}_k^t \circ \mathbf{T}_W^k. \tag{1}$$

C* reduces the localization problem to alternate local tracking and occasional key frame rendering for efficiency, which leads to 6-DoF real-time localization regardless of the 3D map scale.

The NID metric between the current frame I_t and key frame I_k is given as follows:

$$\delta I_{NID}\left(I_t, I_k, \mathbf{T}_k^t\right) \equiv \frac{H_{t,k}\left(\mathbf{T}_k^t\right) - I_{t,k}\left(\mathbf{T}_k^t\right)}{H_{t,k}\left(\mathbf{T}_k^t\right)} \tag{2}$$

where $H_{t,k}$ and $I_{t,k}$ denote the joint entropy and mutual information calculated based on the color co-occurrence in I_t and I_k, respectively. To determine the most likely relative pose $\mathbf{T}_k^t$, gradient-based optimization is performed. Specifically, starting from the given initial guess or previously estimated pose, the BFGS is employed to iteratively solve Eq.1 according to the Jacobian of the NID as follows:

$$^{(i+1)}\mathbf{T}_k^t = {}^{(i)}\mathbf{T}_k^t - \alpha B_k^{-1}\frac{d\delta I_{NID}}{d^{(i)}\mathbf{T}_k^t},$$
$$\frac{d\delta I_{NID}}{d\mathbf{T}_k^t} = \frac{\left(\frac{dH_{t,k}}{d\mathbf{T}_k^t} - \frac{dI_{t,k}}{d\mathbf{T}_k^t}\right)H_{t,k} - \left(H_{t,k} - I_{t;k}\right)\frac{dH_{t,k}}{d\mathbf{T}_k^t}}{H_{t,k}^2}. \tag{3}$$

2) Visual-Inertial integration for rapid head and eye movement tracking: C* is capable of providing reliable camera poses at several tens of hertz. To track the rapid head movements of the subjects, we further fuse the localization results and measurements from an Inertial Measurement Unit (IMU) calibrated to the camera in a loosely coupled manner [9]. The framework allows us to achieve more than several hundreds of hertz estimation rates according to the IMU rates. Simultaneously, it significantly stabilizes visual localization by forming a closed loop that feeds the output pose into the localizer as the next initial guess of the optimization. Localization boosting and stabilization are suitable for real-time gaze projection, as described in the following section.

C. 3D gaze projection onto the environment map

Given the camera pose (subject's head pose) and gaze position on the 2D image, the 3D human gaze can be recovered by generating a 3D ray beginning from the camera through the gaze point. To determine the fixation point, the intersection of the gaze ray and target object must be calculated.

Ray casting can be computationally expensive for real-time operation. Therefore, Paletta *et al.* [7] pre-computed a hierarchical map representation, specifically, an Oriented Bounding Box Tree (OBB-Tree), and traversed the tree to rapidly find the intersection. In [20] and [21], the authors estimated the 3D gaze point by first applying Delaunay triangulation to the feature point map, detecting the triangular plane that includes the 2D gaze, and finally investing the sub-mesh 3D gaze point into the world coordinate system from the triangle vertices. Although these methods work efficiently, they require pre-computation to build certain data structures for 3D gaze mapping, and their resolutions significantly affect the balance between the runtime computation cost and mapping accuracy. Furthermore, when dealing with dynamic objects that are not included in the pre-built 3D environment map, a more flexible scheme that does not require the construction of the data structure each time is preferable.

Thus, for a unified framework of human gaze projection, we propose ID texture mapping as depicted in Fig.3. Texture mapping is a very popular method for attaching a highly detailed appearance to a geometric model that provides realistic rendering images. Given a 3D mesh model, its texture image, and per-vertex UV coordinates, we can generate a textured 3D model with GPU acceleration. Any texture images are available in texture mapping; therefore, we attach a 32-bit integer texture that contains an unique ID of each pixel in its position, for example, $p(x, y) = y * width + x$, for gaze projection. Specifically, we determine the pixels that are currently observable by rendering the 3D map from the camera pose with the ID texture, and directly find the 3D gaze point by accessing the pixel corresponding to the 2D gaze point.

In addition to the simple setup and direct 2D-3D gaze association, the framework offers other benefits with the use of different types of textures. For example, by preparing another texture filled with zero and counting gaze hits, attention accumulation can be easily managed on a 2D image similar to the *attention texture* proposed in [19]. Additionally, overlaying a texture with an object class or semantics on the ID texture enables the semantic understanding of the subject's perceptual activities [18] in a unified pipeline.

ID texture mapping provides a simple yet efficient way of projecting the human gaze onto any geometric model, which is not limited to the map data. In the next section, we extend this framework to dynamic objects for 4D attention mapping.

D. Dynamic object handling for 4D attention mapping

Objects that do not exist in the map building phase cannot be stored in the 3D environment map, which means that the map data should only record static objects. However, many dynamic objects such as humans or household items are observed in daily life, and they seem to have "illegally" appeared in the
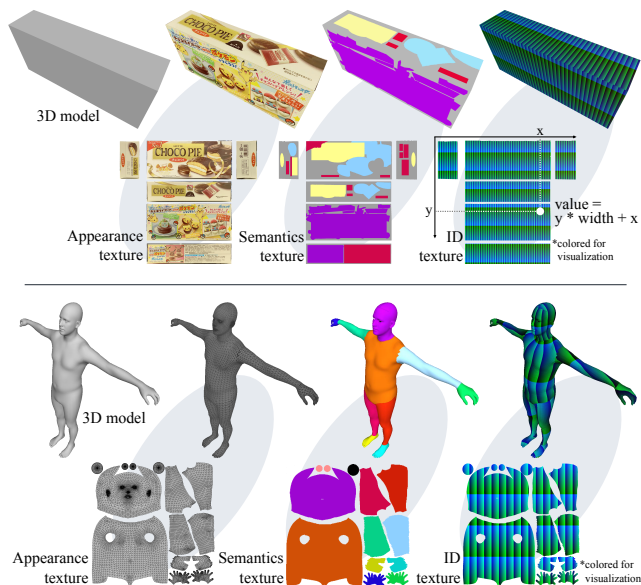

Fig. 4. Texturing dynamic objects: Attention mapping onto dynamic objects is performed in the same way as the case of 3D environment maps by employing the corresponding ID textures. Notably, attaching different textures, for example, a semantic texture, helps to determine perceptual activities in the spatio-temporal domain.

static 3D map. The temporal gap between the mapping and runtime phases causes the absence or presence of dynamic objects, which leads to incorrect gaze projection.

Most conventional works only focus on static scenes and have no choice but to ignore dynamic objects. To analyze human gaze in dynamic scenes, Fathaliyan *et al.* [1] proposed a 3D gaze tracking method that relies on a marker-based motion capture system installed in a small space. It inquires the motion capture tabletop objects' poses in a moment and computes the intersections between the object models and gaze vector; however, the settings are costly and the model does not scale to larger environments. For wearable 3D gaze acquisition outside the laboratory, Qodseya *et al.* [23] and Hausamann *et al.* [24] developed eye-trackers equipped with depth sensors. They overlay 2D gaze points on the depth image and directly reconstruct the 3D human gaze. However, the scheme is highly sensitive to depth noise and the maximum measurement range. Moreover, the 3D gaze information is represented as cluttered 3D point clouds, which makes gaze analysis less meaningful than accumulation on model surfaces.

To address this, we enable the framework to install additional components of object reconstruction for instantiating dynamic objects not captured in the 3D environment map. The recent development of object recognition and tracking techniques has facilitated the determination of full 3D shapes of target objects from monocular images on the fly. Here, we exploit two methods to handle rigid and non-rigid objects, specifically household items and human models, respectively, for 4D attention mapping. Notably, any desired components that estimate the poses and 3D shapes of specific objects can be incorporated as explained below.

1) Household item models (Rigid objects): We introduce a pose detection and tracking method [25] into our system.

(a) Evaluation 1: Static target × Walking around   (b) Evaluation 2: Dynamic target × Standing still   (c) Evaluation 3: Dyamic target × Following
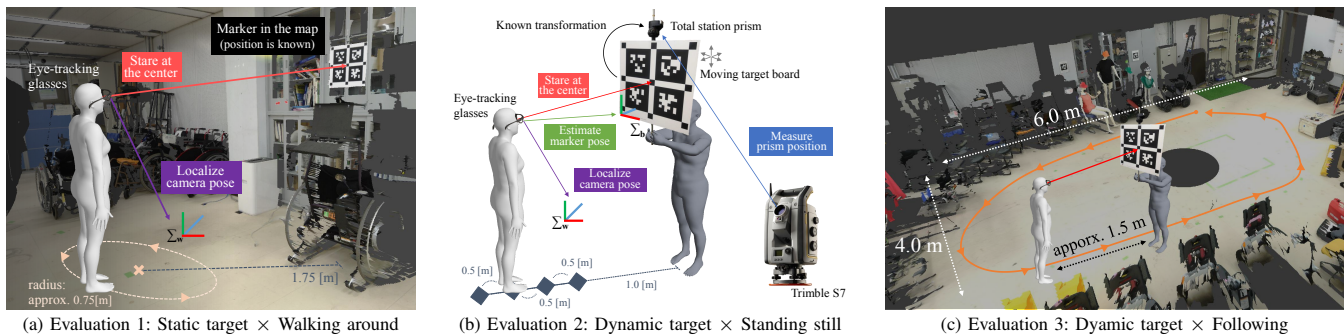
Fig. 5. Overview of the quantitative evaluation: AprilTag [22] was used as a target object to evaluate our attention mapping. Our framework generated successive 3D gaze points by finding gaze intersections while the subject stared at its center. The target board and subject changed their states: (Static or dynamic), and (walking around, standing still, or following), respectively, to demonstrate the robustness of the framework to scene dynamics. Notably, AprilTag was embedded in the 3D map in Evaluation 1, whereas it was reconstructed on the fly in Evaluations 2 and 3.

Given the mesh models and textures of the target objects, it facilitates the recovery of the 6-DoF poses of hundreds of objects in real-time through the proposed scene simulation with SIFT features. The acquired information is sent to the same process as the 3D environment maps described in Section III-C; By attaching an ID texture to each model (Fig.4) and rendering it at the estimated 6-DoF pose, we can easily associate the 2D human gaze with the object model surface. Notably, Multiple Render Targets (MRT) on OpenGL are used to create an integer mask image that helps to distinguish the categories and individuals captured in the rendered view (see the bottom right of Fig.1). In the following experiments, an 8-bit integer mask was rendered in addition to the ID image in the MRT manner to distinguish up to 256 objects belonging to three categories: map, object, and human.

2) Human models (Non-rigid objects): The human model is a representative example of non-rigid objects that are important for analyzing perceptual activity in the real world. Humans change their postures unlike rigid objects; therefore, the reconstruction includes non-rigid deformation, making it more complicated than just detecting 6-DoF poses. In this research, we use the state-of-the-art method, FrankMocap [26], to instantiate humans in a 3D environment map. It fits a statistical body model SMPL-X [27] to each person captured in the input image and provides their shape and pose parameters. The renderer in our framework subscribes the parameters to reconstruct the human models on-demand and examines whether the 3D human gaze hits the surfaces as in the rigid objects.

IV. EXPERIMENTS

A. Setup

In this section, we verify the capability of the proposed framework to recover 4D human attention in dynamic environments. We first quantitatively evaluated the accuracy and precision of the recovered gaze points using a dynamic target marker, followed by demonstrations in real situations.

To build 3D environment maps, we used LiDAR, Focus3D (FARO Technologies, Inc.), which enabled us to capture dense and colored 3D point clouds. A panoramic spherical image can be generated by arranging each vertex color; we used

it as a texture of the 3D map while thinning out some vertices to save GPU memory usage. Notably, our method only assumes that colored or textured 3D models are available for localization and gaze mapping, and thus it also operates on 3D geometric models reconstructed with different sensors, for example, RGB-D SLAM [28], similar to [7].

The rendering and localization components rely on GPU parallelization; a GeForce GTX2080 performed the computations in all the experiments. We also used a wearable eye tracker, Tobii Pro Glasses 3 (Tobii Technology, Inc.) to capture first-person views with the subject's 2D gaze information and IMU data.

B. Performance evaluation

To evaluate the proposed attention mapping, AprilTag [22], which provides reliable 6-DoF marker poses, was employed as shown in Fig.5, whereas the subject was changing the relative positions and its states. We asked the subject to stare at the center of the target board (0.24×0.24 [m]) wearing the eye-tracker, and our method generated the corresponding 3D gaze points. In Evaluation 1, the board was embedded in the 3D map; thus, we calculated the Absolute Position Error (APE) between the generated 3D gaze points and the center of the board. In Evaluations 2 and 3, the ground truth trajectories of the agile target board were obtained by tracking a total station prism attached to the board with the known relative transformation using a Trimble S7 (Trinble Navigation, Limited.). Subsequently, we synchronized the pairs of trajectories based on system timestamps to evaluate the Absolute Trajectory Error (ATE) [29] with a least-squares transformation estimation [30], in addition to APE. Notably, the 3D trajectory comparison computes a rigid transformation that minimizes the positional errors between the two point clouds. The minimization process cancels the systematic bias underlying the framework, which is caused by reasons such as eye-camera miscalibration. Therefore, the ATE is approximately equivalent to the precision of our framework, whereas the APE is equivalent to the accuracy.

Table I and Figure 6 present the evaluation results. —

Evaluation 1: We demonstrated the performance of our framework in a static scene to compare it with the most

TABLE I
ERRORS OF 3D GAZE POINTS IN THE QUANTITATIVE EVALUATION.

No.	object state	subject state	distance [m] from board	APE [m] ($\simeq$ inaccuracy)	ATE [m] ($\simeq$ imprecision)
1	static	walking around	approx. 1.0 - 2.5	0.034 ± 0.015 0.115 ± 0.021†	- -
2	dynamic	standing still	1.0	0.028 ± 0.016	0.020 ± 0.012
			1.5	0.034 ± 0.012	0.017 ± 0.011
			2.0	0.049 ± 0.019	0.034 ± 0.017
			2.5	0.070 ± 0.025	0.034 ± 0.018
3	dynamic	following	approx. 1.5	0.046 ± 0.0092	0.024 ± 0.014

†: Errors of 3D gaze points generated by [7] (our implementation) as a baseline.

relevant work [7] as a baseline. Specifically, we implemented [7] whose localizer was replaced with state-of-the-art indirect visual localization [31] for a comparison in the same 3D map retaining the concept of the method. Compared with [7], 4D attention achieved high accuracy of 3D gaze mapping benefitting from the rendering-centerd framework such as direct localization and ID texture mapping, which suppress the systematic error.

Evaluation 2: The subject watched the center of the moving target standing at four different positions to evaluate the influence of proximity following the evaluations in previous studies [7] [18]. Overall, although the APE (inaccuracy) increased proportionally with the distance from the target board, the framework successfully suppressed the increase in the ATE (imprecision).

Evaluation 3: The subject walked around a 4×6 [m] space to follow the moving target board approximately 1.5 [m] behind while watching the center. Notably, the subject and the person to follow held an assistant rope to maintain their distance. Although the proposed framework slightly increased the APE and ATE owing to the necessity of the 6-DoF and instant object reconstruction in a complicated situation, it successfully facilitated valid attention mapping even in highly dynamic environments.
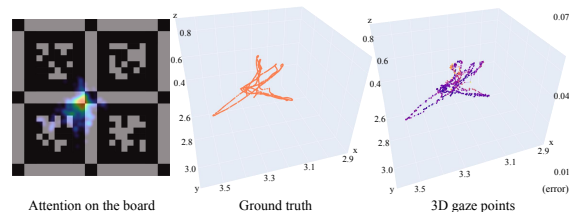
C. Applications

To further evaluate our method, we performed attention mapping in three realistic situations as shown in Fig. 7. Figure 8 picks up "attractive" models in each case, in which accumulated human gaze is highlighted. 4D Attention robustly estimated the subject's poses and 3D gaze directions, and simultaneously projected human gaze onto the static and dynamic targets. This facilitates the analysis of human intention or semantic understanding of the subject's perceptual activities in the real world.

Case 1: As described in Sec.III-C, attaching different types of textures onto the models makes it possible to access various properties of the models, for example, semantics (see Fig. 8(a)). We easily understand which body parts the subject was focusing on (face and hands, in this case).
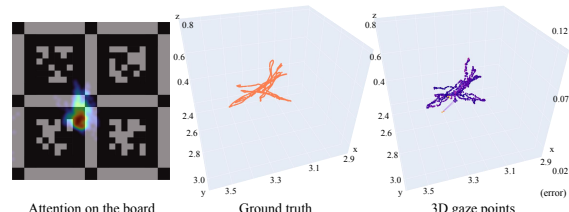
Case 2: Instance object reconstruction allows us to observe human attention in highly dynamic situations, for example, object manipulation. In case 2, after pouring hot water into the mug, the subject picked up freebies and took one. By accumulating gaze information on the models, we may acquire cues to determine the reason for the subject's choice (Fig. 8(b)).
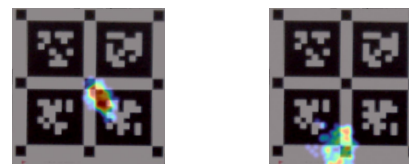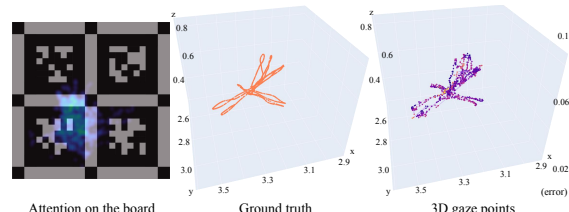

Attention on the board embedded in the 3D map
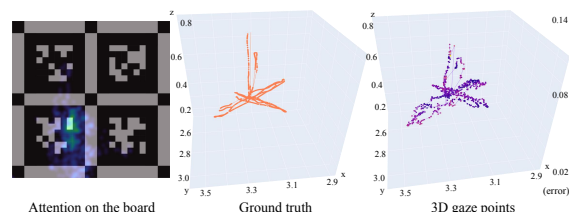
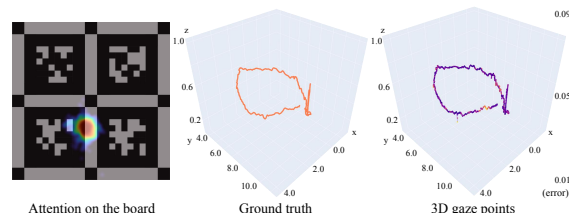(a) Evaluation 1: 4D Attention (left) and [7] (right)

Attention on the board   Ground truth   3D gaze points

(b) Evaluation 2 (1 m)

Attention on the board   Ground truth   3D gaze points

(c) Evaluation 2 (1.5 m)

Attention on the board   Ground truth   3D gaze points

(d) Evaluation 2 (2 m)

Attention on the board   Ground truth   3D gaze points

(e) Evaluation 2 (2.5 m)

Attention on the board   Ground truth   3D gaze points

(f) Evaluation 3

Fig. 6. Quantitative evaluation results: The 3D gaze points obtained in each situation were compared with the ground truth (see also Table I). The proposed framework overwhelmed the competitive method [7] and achieved high-precision 4D gaze mapping in every case. However, the bias was clearly observed in the gaze accumulation, and the accuracy proportionally decreased as the distance from the target board increased. The results imply that our framework is capable of providing stable gaze projection onto dynamic objects, and strict gaze-camera calibration of eye-tracking glasses may improve the accuracy cancelling the systematic error.

Case 3: We simulated a more realistic situation: The subject walked to a vending machine passing by a person and bought a drink from it. Our method successfully provided the trajectory,

first-person view

4D attention

(a) Case 1: Observe physical actions of a person

first-person view

4D attention

(b) Case 2: Take a Coffee break

first-person view

4D attention

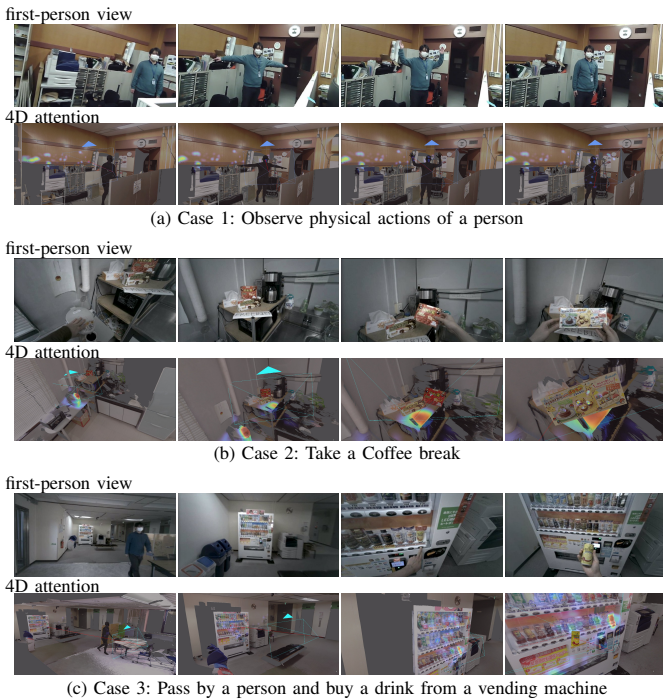(c) Case 3: Pass by a person and buy a drink from a vending machine

Fig. 7. First-person views and attention accumulation in different situations. 4D attention successfully localized the subject and simultaneously detected moving objects. Spatio-temporal human attention was accumulated on the target models according to the subject's observations.

(a) Human model in Case 1: Appearance (left) and semantics (right)

(b) Electric pot and snacks in Case 2

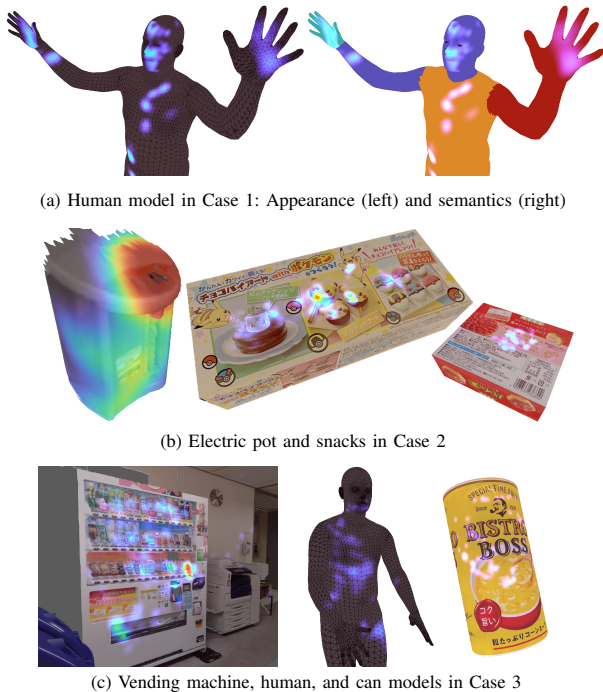(c) Vending machine, human, and can models in Case 3

Fig. 8. Attractive models in each case with accumulated attention

and attention to the static and dynamic objects of the subject (Fig. 8(c)), which helps in determining human behavior in the spatio-temporal domain.

V. DISCUSSION

In this section, we discuss the contributions, limitations, and practicality of the proposed method. According to Table II, which comprehensively compares the characteristics of different works, our framework is distinguished from other competitive methods in several aspects, for example, various targets, real-time operation, and easy setup on a simple 3D map. In particular, the rendering-centered framework provides significant benefits to direct localization and gaze projection via ID texture mapping, which leads to high accuracy of attention mapping as demonstrated in the evaluations.

Map-based methods, however, require a denser 3D map for accurate localization and attention mapping, which can also be a limitation of 4D Attention. Large 3D map reconstruction and rendering can restrict the application of the method to certain scenes. Fortunately, 3D reconstruction technologies, such as SLAM with LiDAR [32] or RGB-D cameras [28], have evolved and are widely available. Techniques such as view frustum culling [33] also help in rendering large 3D maps for real-time processing for further applications in indoor and outdoor environments.

Moreover, as demonstrated in Section IV-C, learning-based shape inference, for example, [26] [34], enables attention mapping to unknown dynamic objects by reconstructing target shapes on the fly. This also allows easier setup to free us from 3D modeling of specific objects, and strengthens our framework toward various usages.

VI. CONCLUSIONS

We developed a novel gaze-mapping framework to capture human attention in the real world. The experiments demonstrated that the combination of robust camera localization, unified attention mapping, and instant object reconstruction enables access to 4D human attention.

The proposed system is capable of providing a series of human head poses (trajectory) and simultaneous gaze targets; thus, it would be applicable in action recognition, for example, skill-level evaluation in humanitude tender-care [35]. It also allows us to incorporate any desired components of instance object reconstruction into the framework, which facilitates attention analysis to specific objects and is helpful for gaze-based target selection in dynamic scenes [36]. Additionally, gaze accumulation on 3D models with multiple textures enables semantic analysis of human behavior.

REFERENCES

[1] A. Haji Fathaliyan, X. Wang, and V. J. Santos, "Exploiting Three-Dimensional Gaze Tracking for Action Recognition During Bimanual Manipulation to Enhance Human-Robot Collaboration," *Frontiers in Robotics and AI*, vol. 5, p. 25, 2018.

[2] A. Dini, C. Murko, S. Yahyanejad, U. Augsdörfer, M. Hofbaur, and L. Paletta, "Measurement and prediction of situation awareness in human-robot interaction based on a framework of probabilistic attention," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 4354–4361.

[3] Y. Maekawa, N. Akai, T. Hirayama, L. Y. Morales, D. Deguchi, Y. Kawanishi, I. Ide, and H. Murase, "An analysis of how driver experience affects eye-gaze behavior for robotic wheelchair operation," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 4443–4451.

[4] H. Admoni and S. Srinivasa, "Predicting User Intent Through Eye Gaze for Shared Autonomy," in *Proceedings of the AAAI Fall Symposium Series:Shared Autonomy in Research and Practice (AAAI Fall Symposium)*, 2016, pp. 298–303.

TABLE II
COMPARISON OF GAZE MAPPING METHODS

method	target		scalable	Real-time	sensors except eye tracker	map	localization	attention mapping
	static map	dynamic objects						
[1]		✓ †		✓	Motion capture	-	Motion capture	Ray casting
[2]		✓ †	✓	✓	Motion capture	-	Motion capture	Ray casting (Sphere approx.)
[3]	✓		✓		LiDAR & Motion capture & IMU	3D point cloud	AMCL & Motion capture	Exhaustive examination
[7]	✓		✓	✓	RGB camera	Color meshes & feature points‡	Indirect visual localization	OBB-Tree
[17]		✓ †		✓	RGB camera (& Kinect)	-	Visual markers	Ray casting (Box approx.)
[18]	✓		✓		RGB camera	Color meshes & feature points‡	Structure-from-Motion	Ray casting
[21]	✓		✓		RGB camera	[Simultaneously built]	Mult-View Stereo & Geometry	Projection onto Delaunay trinagles
[23]	✓		✓		Stereo camera	[Simultaneously built]	RGB-D SLAM	3D cluttered points from the depth
Proposed	✓	✓ (rigid&non-rigid)	✓	✓	RGB camera (& IMU)	Color meshes	Direct visual localization (C*)	ID texture mapping

†: Optical or visual marker(s) should be associated with each object for pose tracking.
‡: Construction of an extra feature-point map that is strictly aligned to the 3D map is required for localization.

[5] J. D. Webb, S. Li, and X. Zhang, "Using visuomotor tendencies to increase control performance in teleoperation," in *2016 American Control Conference (ACC)*, 2016, pp. 7110–7116.

[6] S. M. Munn and J. B. Pelz, "3D Point-of-Regard, Position and Head Orientation from a Portable Monocular Video-Based Eye Tracker," in *Proceedings of the 2008 Symposium on Eye Tracking Research and Applications*, 2008, pp. 181–188.

[7] L. Paletta, K. Santner, G. Fritz, and H. Mayer, "3D recovery of human gaze in natural environments," in *Intelligent Robots and Computer Vision XXX: Algorithms and Techniques*. SPIE, 2013.

[8] S. Oishi, Y. Kawamata, M. Yokozuka, K. Koide, A. Banno, and J. Miura, "C*: Cross-modal simultaneous tracking and rendering for 6-dof monocular camera localization beyond modalities," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5229–5236, 2020.

[9] S. Lynen, M. W. Achtelik, S. Weiss, M. Chli, and R. Siegwart, "A robust and modular multi-sensor fusion approach applied to mav navigation," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 3923–3929.

[10] A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster, "Eye Movement Analysis for Activity Recognition Using Electrooculography," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 741–753, 2011.

[11] S. Ishimaru, K. Kunze, Y. Uema, K. Kise, M. Inami, and K. Tanaka, "Smarter Eyewear: Using Commercial EOG Glasses for Activity Recognition," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 2014, pp. 239–242.

[12] J. Pelz, T. Kinsman, and K. Evans, "Analyzing complex gaze behavior in the natural world," in *Electronic Imaging*, 2011.

[13] G. Fritz and L. Paletta, "Semantic analysis of human visual attention in mobile eye tracking applications," in *2010 IEEE International Conference on Image Processing*, 2010, pp. 4565–4568.

[14] T. Toyama, T. Kieninger, F. Shafait, and A. Dengel, "Gaze Guided Object Recognition Using a Head-Mounted Eye Tracker," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, 2012, pp. 91–98.

[15] K. Harmening and T. Pfeiffer, "Location-based Online Identification of Objects in the Centre of Visual Attention using Eye Tracking," in *Proceedings of the 1st International Workshop on Solutions for Automatic Gaze Data Analysis (SAGA2013)*, 2013.

[16] T. Pfeiffer, "Measuring and Visualizing Attention in Space with 3D Attention Volumes," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, 2012, pp. 29–36.

[17] T. Pfeiffer, P. Renner, and N. Pfeiffer-Leßmann, "EyeSee3D 2.0: Model-Based Real-Time Analysis of Mobile Eye-Tracking in Static and Dynamic Three-Dimensional Scenes," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research and Applications*, 2016, pp. 189–196.

[18] K. Hagihara, K. Taniguchi, I. Abibouraguimane, Y. Itoh, K. Higuchi, J. Otsuka, M. Sugimoto, and Y. Sato, "Object-Wise 3D Gaze Mapping in Physical Workspace," in *Proceedings of the 9th Augmented Human International Conference*, 2018.

[19] T. Pfeiffer and C. Memili, "Model-Based Real-Time Visualization of Realistic Three-Dimensional Heat Maps for Mobile Eye Tracking and Eye Tracking in Virtual Reality," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research and Applications*, 2016, pp. 95–102.

[20] K. Takemura, K. Takahashi, J. Takamatsu, and T. Ogasawara, "Estimating 3-D Point-of-Regard in a Real Environment Using a Head-Mounted Eye-Tracking System," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 4, pp. 531–536, 2014.

[21] R. Matsumoto and K. Takemura, "Semantic 3D Gaze Mapping for Estimating Focused Objects," in *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2019.

[22] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2011, pp. 3400–3407.

[23] M. Qodseya, M. Sanzari, V. Ntouskos, and F. Pirri, "A3D: A Device for Studying Gaze in 3D," in *ECCV Workshops (1)*, 2016, pp. 572–588.

[24] P. Hausamann, C. Sinnott, and P. R. MacNeilage, "Positional Head-Eye Tracking Outside the Lab: An Open-Source Solution," in *ACM Symposium on Eye Tracking Research and Applications*, 2020.

[25] K. Pauwels, L. Rubio, and E. Ros, "Real-Time Pose Detection and Tracking of Hundreds of Objects," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 12, pp. 2200–2214, 2016.

[26] Y. Rong, T. Shiratori, and H. Joo, "FrankMocap: Fast Monocular 3D Hand and Body Motion Capture by Regression and Integration," *arXiv preprint arXiv:2008.08324*, 2020.

[27] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10975–10985.

[28] J. H. Lee, H. Ha, Y. Dong, X. Tong, and M. H. Kim, "TextureFusion: High-Quality Texture Acquisition for Real-Time RGB-D Scanning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[29] Z. Zhang and D. Scaramuzza, "A Tutorial on Quantitative Trajectory Evaluation for Visual(-Inertial) Odometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 7244–7251.

[30] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, 1991.

[31] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM," *IEEE Transactions on Robotics*, pp. 1–17, 2021.

[32] M. Yokozuka, K. Koide, S. Oishi, and A. Banno, "LiTAMIN2: Ultra Light LiDAR-based SLAM using Geometric Approximation applied with KL-Divergence," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 0–0.

[33] U. Assarsson and T. Möller, "Optimized View Frustum Culling Algorithms for Bounding Boxes," *Journal of Graphics Tools*, vol. 5, pp. 9–22, 2000.

[34] Manhardt, Fabian and Wang, Gu and Busam, Benjamin and Nickel, Manuel and Meier, Sven and Minciullo, Luca and Ji, Xiangyang and Navab, Nassir, "CPS++: Improving Class-level 6D Pose and Shape EstimationFrom Monocular Images With Self-Supervised Learning," *arXiv preprint arXiv:2003.05848v3*, 2020.

[35] A. Nakazawa, Y. Mitsuzumi, Y. Watanabe, R. Kurazume, S. Yoshikawa, and M. Honda, "First-person Video Analysis for Evaluating Skill Level in the Humanitude Tender-Care Technique," *Journal of Intelligent & Robotic Systems*, vol. 98, 04 2020.

[36] R. Chacón-Quesada and Y. Demiris, "Augmented Reality Control of Smart Wheelchair Using Eye-Gaze-Enabled Selection of Affordances," in *IEEE/RSJ International Conference on Intelligent Robots and Systems IROS'18 workshop*, 2018.