Illusory Control With Instant Virtual World Environment

Junki Aoki[1,2], Fumihiro Sasaki[1], Ryota Yamashina[1] and Ryo Kurazume[3]

***Abstract*— We proposed a teleoperation method, *illusory control (IC)*, that provides a comfortable operation experience using a seamless transition between real and pre-prepared virtual environments. Therefore, the mobile robot with IC could function solely in familiar environments. To make IC applicable in unfamiliar environments, this study proposes a novel method, *instant IC*, that eliminates the requirement for a pre-prepared virtual environment. The proposed robot system can instantly generate a virtual environment using actual 360° images of the robot in motion, utilizing instant neural graphics primitives and neural radiance fields. The 360° images allow the entire surrounding environment to be virtualized without requiring specific camera orientations. In addition, by optimizing the density of neural radiance fields using depth estimation results beforehand, the reconstruction accuracy at unknown poses can be guaranteed. Furthermore, we propose a depth scaling method based on the actual measurements obtained by LiDAR to increase the consistency of virtual and real environments. With this instant virtual environment, the proposed system enables teleoperation in unknown environments via the seamless transition between real and virtual environments. The experimental results exhibit consistent and smooth back-and-forth transitions between virtual and real space in mobile robot teleoperation.**

I. Introduction

Utilizing teleoperated mobile robots remotely operated by humans has become auspicious owing to the dwindling workforce and increasing remote work. A teleoperated robot is generally equipped with a safety system in that it autonomously detects obstacles using sensors such as LiDAR and autonomously avoids them, in addition to its functions operated remotely by a human. Accordingly, the framework that enables a human and an autonomous agent to jointly control a robot and effectively achieve specific tasks has been researched and called shared control. When a human operates a robot remotely using camera images from the robot, the robot detects obstacles; however, the human does not recognize these obstacles. Although relatively convenient, the human feels substantial stress because they tend to feel that the robot does not obey their command. This stress may be caused by the discrepancy between the intentions of the human operator and that of an autonomous agent, which may trigger a decrease in the acceptance of the robot system.

To address the aforementioned problem, we proposed a teleoperation method named *illusory control (IC)* [1] [2]. IC is a system that provides a comfortable operation experience using a seamless transition between real and virtual environment spaces. IC facilitated a safe robot operation without complications in avoiding obstacles because the operators receive feedback images that allows them the satisfaction of feeling that they are operating the robot according to their intention by switching to the virtual space without obstacles. However, IC was limited because the virtual space must be prepared beforehand. For the system to work, a provider or user of an IC system needed to visit the environment in which the robot moves beforehand, and they needed to sense the environment using a 3D scanner, etc., and post-process, such as adjusting the appearance. Consequently, IC systems could only function in familiar environments. Some applications for a teleoperated robot are difficult to visit beforehand, such as disaster responses. Therefore, the fact that the IC system could only function in known environments severely limited the applicability of IC techniques.

Here, we propose a novel method in which the virtual space does not need to be prepared beforehand. Specifically, we leverage the *instant neural graphics primitives (NGP)* [3] technique, a method that is expected to achieve convergence of *neural radiance fields (NeRF)* [4] training in a short time, for the instant construction of virtual spaces. We propose *instant IC*, a teleoperated robot system that adopts a seamless transition between the virtual space constructed by instant NGP and the real space. Furthermore, we propose a depth estimation method to increase the reconstruction accuracy at unfamiliar poses and a scaling method to fit the geometry in virtual space with the real space. These methods help increase the consistency of the appearance between the virtual and real spaces.

The contributions of this study are as follows.

- The applicability of IC in unknown environments was verified by using instant NGP that can construct a virtual space instantly.
- We verified that a prior depth estimation improves reconstruction accuracy in unknown postures.
- We verified that depth scaling based on actual measurements from LiDAR sensors can improve the geometry consistency between real and virtual space. Specifically, this is a unique challenge for ICs that use seamless transitions between real and virtual spaces.

The remainder of this paper is organized as follows. Section II presents the related work. Section III describes the background of IC and NeRF. Section IV describes the proposed method, and the experiment is described in Section V. Finally, Section VI concludes this paper. Hereafter, virtual

[1]Junki Aoki, Fumihiro Sasaki and Ryota Yamashina are with Ricoh Company, Ltd. {junki.aoki, fumihiro.fs.sasaki, ryohta.yamashina}@jp.ricoh.com

[2]Junki Aoki is with the Graduate School of Information Science and Electrical Engineering, Kyushu University.

[3]Ryo Kurazume is with the Faculty of Information Science and Electrical Engineering, Kyushu University. kurazume@ait.kyushu-u.ac.jp

space is called v-sp and the robot in v-sp is called v-bot, while real space is called r-sp and the robot in r-sp is called r-bot.

II. Related Works

Conventional 3D reconstruction techniques include point cloud-, voxel-, and mesh-based methods. The point cloud-based methods cannot represent object surfaces; hence, it is difficult for them to reconstruct a highly realistic 3D space on their own. The voxel-based methods are also memory inefficient and poor at reconstructing accurate geometry. The mesh-based methods can accurately reconstruct geometry and achieve highly realistic 3D space using texture. In fact, we have adopted mesh-based methods [1] or added mesh-based methods to propose methods [2] for similar v-sp and r-sp appearances using CycleGAN [5]. However, it is difficult to adapt them to IC, which attempts to circumvent v-sp preparation beforehand because pre-processing time is required to reconstruct a 3D model. In IC, considering the need to switch between r-sp and v-sp in response to obstacles with as little operator discomfort as possible, it is desirable to construct a v-sp in a few seconds.

A 2D image-based 3D reconstruction method using neural networks has been proposed [6], [7]. One approach is to estimate depth from images [8]–[10]. These images are highly realistic because they are based on actual images and perform geometrical transformations based on posture. Because these methods are image transformations based on a single image, it is difficult to reconstruct a consistent environment using images in several poses, especially for large-scale spatial restoration.

Other approaches include simultaneous localization and mapping (SLAM); in particular, dense SLAM should also be able to achieve a highly realistic reconstruction [11], [12]. These alternative approaches are expected to perform dense and highly realistic 3D reconstruction in real-time. To further improve higher realism, NeRFs have recently garnered significant attention [4]. NeRFs utilize image-posture pairs to train a neural network to perform free-viewpoint rendering. NeRFs are also being studied for the consistent and realistic reconstruction of relatively large outdoor environments [13]–[15]. Furthermore, a method combined with SLAM has also been proposed [16], [17]. In addition, a method that considers rendering on mobile devices has been proposed; hence, future developments can be expected to work on inexpensive devices [18]. Recently, a method has been proposed to increase the accuracy of reconstruction in unfamiliar postures by optimizing with prior information on the depth [19]. NeRF with 360° images has also been proposed to enable the comprehensive reconstruction of the environment from a small number of images, independent of the posture at the time of capture [20]. However, the problem with NeRF was the lack of actual learning and rendering time. Instant NGP is a technique that enables the convergence of learning in a significantly short time [3]. Instant NGP is expected to reconstruct the v-sp to accommodate instant IC's requirements. Therefore, the leveraging of instant NGP could be an effective improvement method to make IC preparation unnecessary. In addition, from the aspect of instant NGP, instant IC is an example of effective use of its technology's features.

III. Background

A. Illusory Control

This section describes the teleoperation flow of a mobile robot using the proposed IC.

First, the operator sees the camera image of the r-bot and commences the operation with the r-bot as the control target. The r-bot receives operation commands from the operator and calculates its future trajectory based on these commands. It then determines whether the calculated future trajectory will reach the obstacle or not and, if it does, switches the control target from the r-bot to the v-bot. At this point, the system moves the v-bot to the same position based on the position information of the r-bot, switches the image shown to the operator to the v-sp, and then switches the control target to the v-bot. Although the operator operates the v-bot, the r-bot moves autonomously using the position and posture of the v-bot as subgoals. If there are obstacles in r-sp at the v-bot position, the future trajectory of the v-bot is calculated, the cost on the trajectory is obtained, and the point where the cost is below a certain level is set as the subgoal. The robot system periodically acquires the position information of each of the v-bot and r-bot, and when the difference in their respective postures becomes less than a threshold value, it switches the feedback image to the operator from the v-bot to the r-bot and the control target from the v-bot to the r-bot.

Here, we have improved the part of the v-sp employed in this teleoperation flow. Specifically, we propose a method that does not require advanced preparation by immediately constructing a v-sp using image and posture information obtained while the robot is in operation.

B. NeRF

This section describes the issues addressed by the NeRF. NeRF optimizes a function $f(x, d) = (c, \sigma)$ representing a 3D scene, where x, d, c, and σ denote the 3D view position, a view direction, a color, and the density, respectively. Each parameter in NeRF is optimized using a multilayer perceptron (MLP). The rendered color $C(r)$ in some range $n - f$ on the camera ray $r = o + td$ can be defined as

$$C(r) = \int_n^f T(t)\sigma(r(t))c(r(t), d)dt,$$

$T(t) = exp(-\int_n^t \sigma(r(s))ds)$ represents the probability that the camera ray terminates at the object surface at t, starting at the neighborhood boundary n. The estimated color C is approximated as

$$\hat{C}(r) = \sum_{i=1}^{N} T_i(1 - exp(-\sigma_i\delta_i))c_i,$$

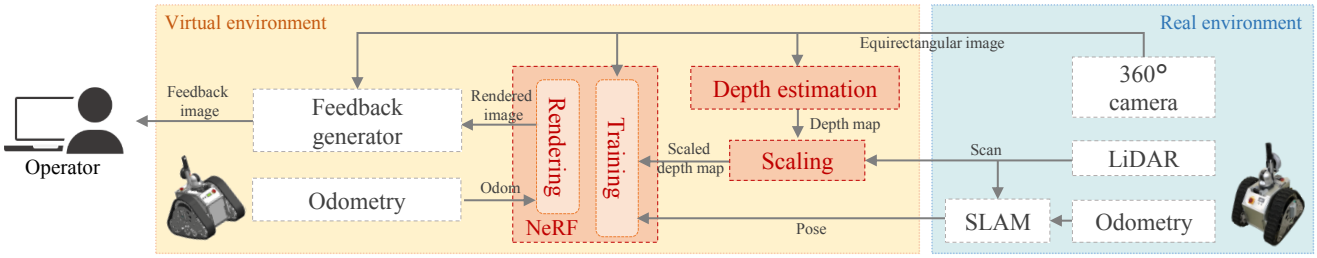

Fig. 1. Image processing flow in instant IC.

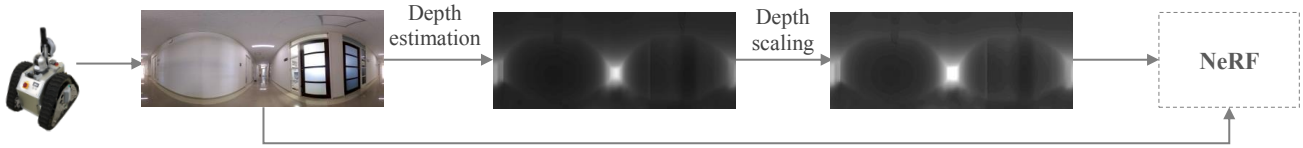

Fig. 2. Excerpt of the processing part of the image. In this figure, the depth images are adjusted brighter than the actual images to improve legibility.

where $T_i = exp(-\sum_{j=1}^{i-1} \sigma_j \delta j)$ and $\delta_i = t_{i+1} - t_i$ represents the distance between two consecutive samples on the ray r.

IV. Method

The data flow of the proposed method is presented at Fig. 1. The image data process provided to NeRF is presented at Fig. 2. The data flow presented here is only the part related to the construction and rendering of the v-sp; for the module structure of the entire robot, kindly refer to [2].

A. Instant Training and Rendering

The process of constructing a v-sp using NeRF can be divided into two main parts. The first is a learning process in which r-sp image data obtained from the robot under operation are collected and fed to NeRF. The second process involves inputting the posture of the robot on the simulator to NeRF using the learning results and rendering the images seen from that posture.

The NeRF technology employed was instant NGP [3]. This method was adopted because it is expected to achieve learning convergence in a short time, and real-time rendering at approximately 10 fps is feasible by adjusting the rendering quality. Our previous method required advanced preparation of the v-sp, thereby resulting in a significant time difference between the current time and the time required to create the v-sp. The proposed method can absorb temporal alterations in the environment, except for dynamic obstacles such as humans, because the temporal difference between the r-sp and v-sp at the time of creation can be maintained from a few seconds to a few dozen seconds.

First, we describe the process of acquiring data from the robot under operation and creating data to be fed to NeRF. Two types of data are acquired from the robot during operation: image and robot-posture information. This information is periodically sent to the computer that constructs the v-sp using the communication protocol of the ROS (Robot Operating System). Upon receiving this information, the computer preprocesses the depth estimation and depth scaling described below and then creates and stores the data to be given to NeRF.

The image information used to train NeRF is a 360° image transformed into a total of six perspective images at 90° horizontally and 90° vertically. By employing 360° images, the v-sp can be constructed such that the robot's field of view is not limited by its orientation during data acquisition. When the robot moves a certain distance, the training images are replaced with the latest ones. The posture information used to train NeRF is the robot's posture on the map frame. Specifically, a map frame was constructed while performing SLAM using LiDAR and the robot's wheel odometry, and the posture information for this map frame was adopted. The Gmapping algorithm was employed for SLAM.

Next, the rendering process was simultaneously conducted with instant NGP training. The input was the posture information obtained from the robot on the simulator operated by the operator. The operator was presented with the results rendered by instant NGP as feedback during the operation of the v-bot.

B. Prior Depth Estimation

If training and rendering are performed by simply providing NeRF with the aforementioned image and posture information, the rendering accuracy for unknown postures will be extremely low. Because the concept of instant IC is to circumvent advance preparation, image information for postures unfamiliar to the robot cannot be obtained beforehand. Therefore, images and postures obtained from the current robot can be employed; however, images and postures at future assumed positions cannot be adopted.

To address this problem, DS-NeRF [19] presented an approach that can render good quality images from a small number of images by providing prior information on the depth. Based on this approach, we attempted to provide prior information on depth. Specifically, NeRF datasets were provided images with pre-estimated depths using the SliceNet

[10] algorithm, which can perform direct depth estimation on 360° images (equirectangular). SliceNet was selected because it can achieve depth estimation for 360° images and exhibits the fastest depth estimation time among any of them. Because SliceNet is a supervised learning technique, it is necessary to construct and train a dataset in an environment where depth information is available beforehand. However, this method adopts only the pre-trained model published by the authors of SliceNet, and no additional training, such as fine-tuning, was performed in the current experimental environment.

C. Depth scaling

Depth scaling was conducted on the pre-estimated depth images using the actual LiDAR measurements. Because the depth pre-estimation employed with SliceNet is based on a pre-trained model, adopting it as it is in the actual operating environment will trigger a discrepancy in the scale of depth. Because IC is required to switch seamlessly between r-sp and v-sp, any discrepancy between the scale of the actual environment and that of the model will cause feedback as if the posture is significantly off when the image is switched between r-sp and v-sp. Therefore, we scaled the pre-estimated depth image using the measured values from a range sensor such as LiDAR. Specifically, we scaled the depth image, such that the maximum value of the point cloud information obtained from the range sensor matches the maximum value of the depth image.

The scaled depth D_{ic} is expressed as

$$D_{ic} = \frac{d_{max_l}}{d_{max_s}} D_s$$

where d_{max_l} and d_{max_s} represent the max depth values from the range sensor scan and depth map D_s estimated by SliceNet, respectively. Note that d_{max_l} and d_{max_s} denote the maximum values at the 99.5% confidence interval accounting for outliers.

This is expected to reduce the posture shift when the image switches between r-sp and v-sp.

V. Experiment

Instant construction and teleoperation experiments of a v-sp with instant IC were conducted using an actual mobile robot. The mobile robot was a crawler robot manufactured by Ricoh. Ricoh Theta Z1 and Velodyne LiDAR VLP-16 were utilized as sensing devices mounted on the crawler robot. A desktop computer with an Intel Core i9 CPU and NVIDIA RTX 4090 GPU was employed for the v-sp construction and teleoperation client.

We aimed to answer the following six questions:

- Can instant IC ensure equivalent reconstruction accuracy when compared to conventional IC, which requires pre-preparation of the v-sp?
- How much pre-preparation time is required to build a v-sp using instant IC?
- Do the prior depth estimation and depth scaling contribute to ensuring the appearance consistency between v-sp and r-sp?
- Can instant IC allow the teleoperation of a mobile robot using a transition between v-sp and r-sp?
- Are there any differences in impressions of the system when compared to instant IC and conventional IC?
- Can instant IC ensure reconstruction accuracy in multiple indoor environments?

A. Reconstruction Accuracy

In this experiment, we compared the reconstruction accuracy between conventional IC [1] [2] and instant IC. Conventional IC requires a v-sp to be constructed beforehand using a 3D scanner and the CycleGAN to adjust appearance. Instant IC is the proposed method.

For the instant IC, a v-sp was constructed using the proposed method based on 360° images obtained from the robot. Only one image obtained at the robot's initial position and the robot's posture at that time were used for the images. We evaluated the degree to which the reconstruction accuracy transitions when the robot moves away from the point where the last image was acquired.

Fig. 3 illustrates the transitions of the peak signal to noise ratio (PSNR) and structural index similarity (SSIM) when the robot moved straight ahead from its initial position. Fig. 4 illustrates the transitions of PSNR and SSIM when the robot curves forwarding in the corridor in Fig. 9. Each movement trajectory is presented on the left side of Fig. 3 and 4. The qualitative results of the rendering comparing IC and instant IC are also presented in Fig. 7.

Focusing on SSIM, it infers that instant IC scores higher than conventional IC in the interval of approximately 2 m from the initial position in the straight and approximately 1–2 m for the curve.

In other words, for a movement range of approximately 2 m, IC can be applied with high reconstruction accuracy by constructing a v-sp based on one-shot images using instant IC. In addition, the curving case demonstrates that the v-sp construction with 360° images allows the reconstruction accuracy to be maintained even if there is movement in the turning direction.

This experiment was a v-sp construction using only images obtained in the initial posture of the robot to understand the capabilities of instant IC. In an actual teleoperation system with IC, images are acquired, and the v-sp is reconstructed every time the robot moves a certain distance; hence, the reconstruction accuracy does not continue to decrease as the robot moves, as illustrated in Fig. 3 and 4.

B. Training Time

The time required to learn NeRF was evaluated using the proposed method. Fig. 5 presents the transition of PSNR and SSIM when the v-sp construction using the proposed method commenced with a single 360° image of the robot in its initial posture and drawn in its initial posture.

For both PSNR and SSIM, the scores converge at approximately 3–4 s after learning commences. In other words, if a learning time of approximately 3–4 s is provided beforehand, a v-sp with high reconstruction accuracy can be fed back

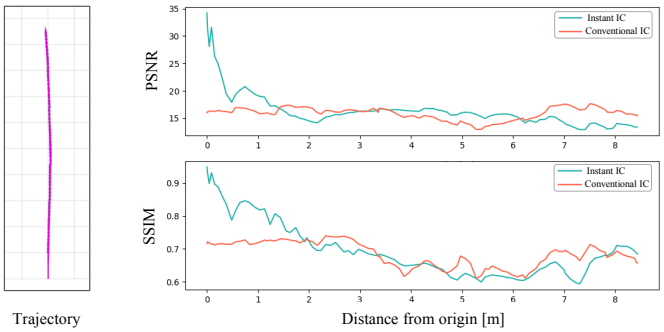

Fig. 3. Evaluation results of the PSNR and SSIM in the straight trajectory. Instant IC is the proposed method. Conventional IC is method which requires pre-preparation of the v-sp.

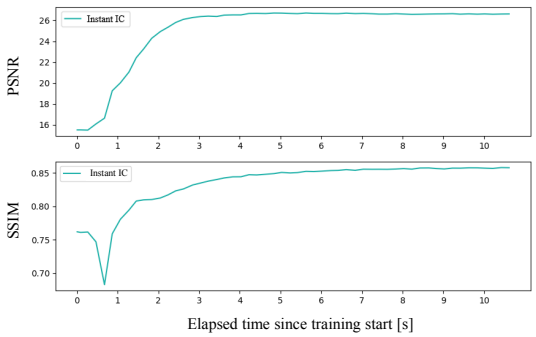

Fig. 5. Evaluation results of the training time of NeRF.

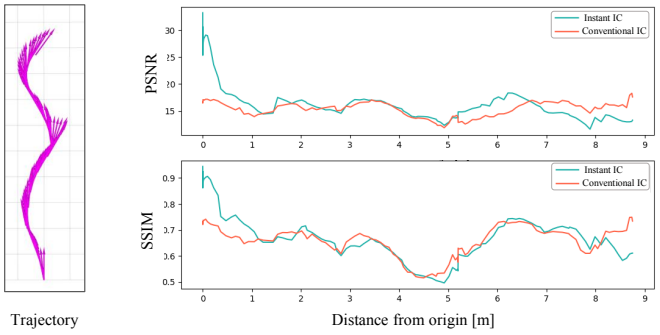

Fig. 4. Evaluation results of the PSNR and SSIM in the curved trajectory.

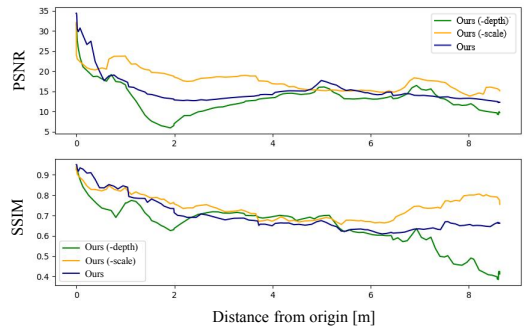

Fig. 6. Evaluation results of the ablation study in the straight trajectory. Ours (-depth) is the proposed method without the prior depth estimation. Ours (-scale) is the proposed method without the depth scaling. Ours is the proposed method with the prior depth estimation and the depth scaling.

to the operator. Given the teleoperation by IC, the operator will operate the system while seeing images in r-sp until they encounter an obstacle; hence, it is hypothesized that a situation may not emerge so much where the operator is made to wait for the v-sp to be constructed.

C. Ablation Study

The ablation experiment was conducted to evaluate the effects of a prior depth estimation (Ours (-depth)) and depth scaling (Ours (-scale)). Fig. 6 illustrates the evolution of PSNR and SSIM when moving straight down the corridor from the initial position where the image that functions as the training data was acquired. It appears that Ours (-scale) has the best score in all intervals, although there are a few fluctuations in the scores. However, when observing the rendering results, it appears that without scaling, the actual geometry of the environment is not correctly represented. Fig. 8 qualitatively presents the results of rendering via the proposed method using the posture from which the ground truth image was obtained as input. Ours (-depth) has already passed through the mobile environment in ③ and ④. Furthermore, Ours (-scale) has roughly reached the end of the mobile environment at ④. Although there is room for further improvement in the reconstruction accuracy, compared to the above two, Ours appears to be able to reflect geometric information in r-sp.

D. Teleoperation Experiment

Fig 9 presents the results of a comprehensive teleoperation experiment using instant IC. A white obstacle, which is difficult to recognize from the image, is placed in front of the robot in its initial position. When the robot approaches an obstacle, the feedback image to the operator switches to that of the v-sp, and the operator's operation target switches to the v-bot (Fig. 9 ②). In Fig. 9 ②-④, the operator operates the v-bot while the robot avoids obstacles by autonomous movements while seeing the images in the v-sp. When the difference in posture between the v-bot and r-bot falls below a certain value, the feedback image to the operator switches to that of the r-sp, and the robot operated by the operator switches to the r-bot (Fig. 9 ⑤). The aforementioned operation flow is the same as that for conventional ICs; however, with instant IC, it was verified that this flow could be achieved without preparing a 3D model beforehand. In particular, some roughness emerged in the reconstruction accuracy of areas that were not obtained from the image data, such as the door visible on the right side of the corridor. Improving the reality of these areas is an issue for the future.

Fig. 7. Qualitative results of the reconstruction accuracy. In conventional IC, the v-sp is constructed by 3D scanner (FARO Focus3D). ①–④ show time-series changes, with the younger numbers indicating earlier times.

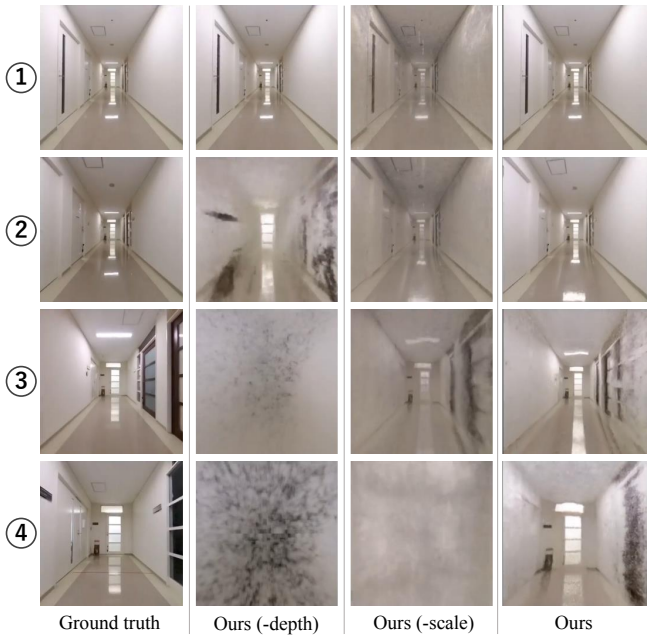

Fig. 8. Qualitative results obtained from the ablation study. ①–④ show time-series changes, with the younger numbers indicating earlier times.

E. User Study

We conducted a user study to verify whether there were differences in impressions of the system operating the robot between conventional and instant IC. The five participants were university researchers specializing in robotics. They participated in the study regardless of their remote control skills and were not compensated. The experiments were approved by the Ethics Review Committee of the Graduate School of Information Science and Electrical Engineering, Kyushu University.

We employed qualitative semi-structured interviews to assess the impressions of the system in terms of realism and usability. This approach was chosen because concerns regarding realism and ease of use differ among individuals, making it challenging to identify these aspects solely through standardized multiple-choice questions. Participants were instructed to remotely control the robot in the environment, as illustrated in Fig. 9, navigating around two obstacles using the controller while viewing images displayed on the desktop application via both conventional IC and instant IC. The visual effect on V-bot deceleration, as described in [2], was not implemented in order to evaluate the pure impression of the v-sp appearance. To create variations in the timing of the transition to v-sp, obstacles were placed in different positions for each participant. This allowed for diverse movement distances from the training image acquisition position and the elapsed time since the training's start, thus providing variations in reconstruction accuracy when transitioning to v-sp during the instant IC experiment. Participants were required to experience at least one set of transitions between r-sp and v-sp in each trial. The experimental order of conventional IC and instant IC was alternated for each participant. During the trial, they were not informed whether the conventional or proposed method was being used. After allowing participants to freely experience teleoperation using both conventional IC and instant IC for approximately 2-3 minutes each, they were asked to respond to the following questions:

- **Appearance**: Which did you find the more realistic appearance, conventional IC or instant IC?
- **Usability**: Which did you find easier to operate, conventional IC or instant IC?

Using the aforementioned questions as a starting point, the interviews proceeded by inquiring why participants felt the way they did. The user study's results showed that for appearance, instant IC received three votes while conventional IC garnered two. In terms of usability, instant IC and conventional IC each received two votes, with one participant considering them equivalent. Below are some comments gathered from the interviews:

- Conventional IC appeared to exhibit a different coloring than the r-sp, whereas the instant IC appeared more realistic.
- In the instant IC, the large noise in the image was observed, whereas in the conventional IC, the boundaries of doors and walls could be clearly identified. Consequently, it was easier to recognize robot's position within the image using the conventional IC.
- Both conventional IC and instant IC have an appearance that is easy to operate.
- In conventional IC, the obstacles present in r-sp disappeared in v-sp, while in instant IC, the obstacles were carried over from r-sp to v-sp, providing a sense of

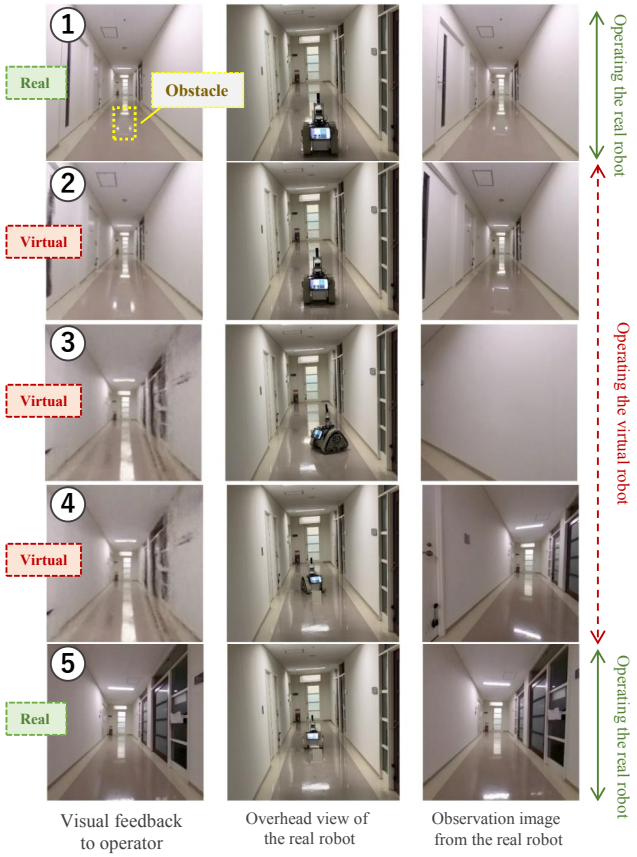

Fig. 9. Experimental results of the teleoperation by instant IC. ①–⑤ show time-series changes, with the younger numbers indicating earlier times.

Fig. 10. Examples of transitions and not-so-accurate transitions with relatively accurate reconstruction of v-sp in the user study.

consistency.

In instant IC, some participants appeared to feel that the appearance was incomplete. Examples of transitions and less accurate transitions with relatively precise v-sp reconstructions are shown in Fig. 10. Others believed that instant IC seemed more realistic when the operation in v-sp commenced at a location close to where the training data were collected. As demonstrated in the learning time experiment, instant IC required a training time of 3-4 seconds. Ideally, learning should converge at the point of transition to v-sp, so the design acquires data at fixed movement intervals; however, depending on the timing of the transition to v-sp, it may not provide sufficient realism. Moreover, even with adequate training time, when the operation began in v-sp at a distance from where the training data were obtained, participants seemed to perceive degradation, such as blurring, from the r-sp image.

Instant IC can accommodate changes in the environment, such as the emergence of obstacles, because the virtual space was created only several seconds prior. In this regard, some participants felt that instant IC operated more consistently. Adapting to environmental changes in a virtual space is an essential aspect of IC. As previously mentioned, as long as realism can be enhanced, instant IC can offer a more consistent operation.

F. Qualitative results in multiple indoor environments

In this experiment, the reconstruction was evaluated in various indoor settings. Images were captured at the robot's initial position and utilized for training. After sufficient training, the posture reconstruction was then qualitatively assessed as the robot navigated through an indoor environment. The results of the experiments conducted in two indoor environments are displayed in Fig. 11.

In relatively small and simple indoor environments, such as corridors, the proposed method can successfully replicate real spaces. However, in some cases, it does not perform well in open indoor environments. Depth estimation and scaling struggle in settings where windows are closely spaced, as depicted in the left column of Fig. 11, and in complex environments with no windows but featuring intersecting corridors, as seen in the right column of Fig. 11. Enhancing estimation accuracy to ensure that the method functions correctly in all indoor environments without prior preparation remains a future challenge.

VI. Conclusion

Here, we proposed a novel method, instant IC, to eliminate the need for the advanced preparation of v-sp in IC. Specifically, we employed instant NGP, a method that is

Fig. 11. Qualitative results of instant IC in multiple indoor environments. The example on the left is of an indoor environment with a view of the exterior beyond the window. The example on the right is of an intricate environment with no windows but with intersecting corridors. ①-⑤ show time-series changes, with the younger numbers indicating earlier times.

expected to learn NeRF in real-time, to construct the v-sp instantaneously. We also proposed a method to improve the rendering accuracy in unfamiliar postures by performing prior depth estimation on images, including a method to improve the agreement with actual geometry by performing depth scaling to the r-sp using measured values from LiDAR. Using these proposed methods, we constructed an instant IC system that switched between the v-sp created by instant NGP and images in the r-sp, evaluated the reconstruction accuracy and training time, verified its operation using a crawler robot, and conducted the user study.

Building a technique to interpolate the appearance of areas with missing data is a future task. Possible approaches to solving this issue include using a generative model, such as inpainting, for interpolation [21], [22], ignoring whether it matches the actual visibility, or planning autonomous movement such that it actively acquires areas where data are insufficient [23].

ACKNOWLEDGMENT

This research was partially supported by JSPS KAKENHI Grant Number JP21K18701 and JP20H00230. Additionally, this research is the result of joint research by Ricoh Company, Ltd. and Kyushu University.

REFERENCES

[1] J. Aoki, R. Yamashina, and R. Kurazume, "Teleoperation method by illusion of human intention and time," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 2021, pp. 482–487.

[2] J. Aoki, F. Sasaki, R. Yamashina, and R. Kurazume, "Teleoperation by seamless transitions in real and virtual world environments," *Robotics and Autonomous Systems*, vol. 164, p. 104405, 2023.

[3] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics*, vol. 41, no. 4, 2022.

[4] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

[6] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[8] R. Hu, N. Ravi, A. C. Berg, and D. Pathak, "Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image," *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12 508–12 517, 2021.

[9] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," 2023. [Online]. Available: https://arxiv.org/abs/2302.12288

[10] G. Pintore, M. Agus, E. Almansa, J. Schneider, and E. Gobbetti, "SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 11 536–11 545.

[11] K. M. Jatavallabhula, G. Iyer, and L. Paull, "∇slam: Dense slam meets automatic differentiation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 2130–2137.

[12] Z. Teed and J. Deng, "DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras," *Advances in neural information processing systems*, 2021.

[13] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. Srinivasan, J. T. Barron, and H. Kretzschmar, "Block-NeRF: Scalable large scene neural view synthesis," *arXiv*, 2022.

[14] H. Turki, D. Ramanan, and M. Satyanarayanan, "Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 12 922–12 931.

[15] K. Rematas, A. Liu, P. P. Srinivasan, J. T. Barron, A. Tagliasacchi, T. Funkhouser, and V. Ferrari, "Urban radiance fields," *CVPR*, 2022.

[16] E. Sucar, S. Liu, J. Ortiz, and A. Davison, "iMAP: Implicit mapping and positioning in real-time," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[17] A. Rosinol, J. J. Leonard, and L. Carlone, "Nerf-slam: Real-time dense monocular slam with neural radiance fields," 2022.

[18] Z. Chen, T. Funkhouser, P. Hedman, and A. Tagliasacchi, "Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures," *arXiv preprint arXiv:2208.00277*, 2022.

[19] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan, "Depth-supervised NeRF: Fewer views and faster training for free," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[20] K. Gu, T. Maugey, S. Knorr, and C. Guillemot, "Omni-nerf: Neural radiance field from 360° image captures," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 1–6.

[21] R. Fridman, A. Abecasis, Y. Kasten, and T. Dekel, "Scenescape: Text-driven consistent scene generation," 2023.

[22] E. R. Chan, K. Nagano, M. A. Chan, A. W. Bergman, J. J. Park, A. Levy, M. Aittala, S. D. Mello, T. Karras, and G. Wetzstein, "GeNVS: Generative novel view synthesis with 3D-aware diffusion models," in *arXiv*, 2023.

[23] H. Zhan, J. Zheng, Y. Xu, I. Reid, and H. Rezatofighi, "Activermap: Radiance field for active mapping and planning," 2022.