Enhancing Perception and Immersion in Pre-Captured Environments through Learning-Based Eye Height Adaptation

Qi Feng*
Waseda University

Hubert P. H. Shum†
Durham University

Shigeo Morishima‡
Waseda Research Institute for Science and Engineering

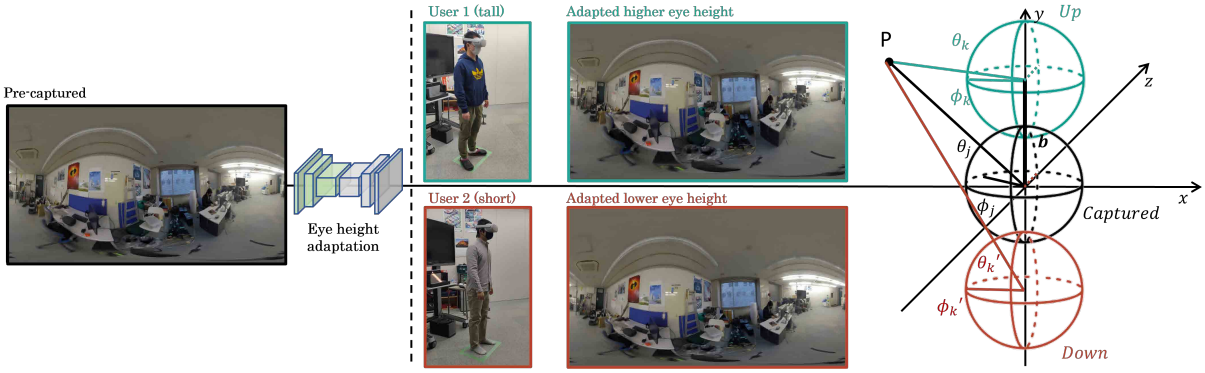

Figure 1: We propose a learning-based approach for generating novel views from pre-captured omnidirectional inputs, which can adapt to the user's eye height during playback, resulting in an improved perceptive and immersive experience.

Abstract

Pre-captured immersive environments using omnidirectional cameras provide a wide range of virtual reality applications. Previous research has shown that manipulating the eye height in egocentric virtual environments can significantly affect distance perception and immersion. However, the influence of eye height in pre-captured real environments has received less attention due to the difficulty of altering the perspective after finishing the capture process. To explore this influence, we first propose a pilot study that captures real environments with multiple eye heights and asks participants to judge the egocentric distances and immersion. If a significant influence is confirmed, an effective image-based approach to adapt pre-captured real-world environments to the user's eye height would be desirable. Motivated by the study, we propose a learning-based approach for synthesizing novel views for omnidirectional images with altered eye heights. This approach employs a multitask architecture that learns depth and semantic segmentation in two formats, and generates high-quality depth and semantic segmentation to facilitate the inpainting stage. With the improved omnidirectional-aware layered depth image, our approach synthesizes natural and realistic visuals for eye height adaptation. Quantitative and qualitative evaluation shows favorable results against state-of-the-art methods, and an extensive user study verifies improved perception and immersion for pre-captured real-world environments.

Index Terms: Computing methodologies—Computer graphics—Graphics systems and interfaces—Mixed / augmented reality; Computing methodologies—Computer graphics—Image manipulation—Image-based rendering

1 Introduction

Pre-captured immersive content for virtual reality (VR) has gained increasing attention from the commercial and research communities for its potential applications in fields such as medicine, education, entertainment, and prototyping [61]. Omnidirectional cameras capture egocentric perspectives that provide greater immersion than traditional media, fostering more effective interactions between the content and the user [12]. By directly recording the environment and simulating real-world perceptions during playback, pre-captured content offers more photo-realistic cognitive stimuli than model-based virtual environments. However, this also makes post-processing of the visuals more challenging compared to traditional model-based virtual environments.

Previous studies have shown that manipulating the visual eye height within virtual environments can significantly affect distance perception [35], and that virtual and real environments can elicit similar visual responses [12]. However, the influence of eye height on perception in pre-captured immersive content has not been studied, due to the difficulty of freely altering the eye height after the footage is captured. If a significant influence on perception and immersion is found, an effective approach to adapt the eye height of the user for existing pre-captured immersive content would be highly desirable.

Traditional image-based reconstruction methods often require specific capture setups with a sufficient number of inputs for baselines [14, 16]. Novel view synthesis typically uses multi-layered image representations combined with depth-based warping algorithms [62, 76] to replicate parallax effects, resulting in holes in occluded regions. To address missing information, recent convolutional neural network (CNN)-based approaches use light field data [26, 48], piece-wise planar images [20], or local inpainting [21]. These methods have shown great potential for manipulating the eye height in our application, but most of them are designed for pinhole cameras and do not perform well with 360-degree inputs due to irregular distortions introduced by equirectangular projection [13, 78].

In this paper, we first propose a pilot study to verify whether different eye heights have a significant influence on users' perception and immersion when viewing pre-captured real-world environments, a hypothesis that has not previously been tested. Unlike virtual environments where the eye height can be easily adjusted, we capture identical scenes at multiple eye heights under controlled conditions to optically simulate different eye levels in the real world using state-of-the-art equipment. The results of the study show an improved perception and immersion, providing the basis for the subsequent eye height adaptation system. It contributes to both future application designs and a better understanding of human perceptions.

Motivated by the pilot study, we propose a learning-based ap-

*e-mail: fengqi@ruri.waseda.jp
†e-mail: hubert.shum@durham.ac.uk
‡e-mail: shigeo@waseda.jp

proach for adapting the eye height of pre-captured immersive content. The system consists of a depth estimation stage and an inpainting stage. We first introduce a novel omnidirectional-aware multitask architecture that learns depth and semantic segmentation in two formats, enabling the network to generate high-quality depth and semantic segmentation that facilitates the inpainting stage for 360-degree input. In the inpainting stage, we improve upon the existing layered depth image (LDI) approach [60] by using the omnidirectional-aware depth and semantic segmentation information to guide the synthesis of natural and realistic textures for occluded regions, enabling eye height adaptation for pre-captured real-world environments.

With quantitative and qualitative evaluation, our experimental results shows an improved performance over existing methods, showing that the proposed method is able to generate eye height-adapted results with satisfying quality and efficiency. An extensive user study further verifies the effectiveness of our learning-based approach in improving user perception and immersion for pre-captured immersive content in VR. We believe that its application can benefit a wide range of existing pre-captured media in 360-degree format for better immersion and experience.

To summarize, our contributions are as follows:

1. We propose the first pilot study to show a significant influence of altered eye heights on perception and immersion for pre-captured immersive content with real environments;
2. We propose a two-stage approach for omnidirectional-aware eye height adaptation. A novel network estimates accurate depth and semantic segmentation, and the following inpainting stage improves the layered depth image approach with guides to synthesize high-quality visuals;
3. The implementation and the main user study validate the effectiveness of image-based eye height adaptation in improving users' perception and immersion in pre-captured real environments within VR.

2 Related Work

2.1 Perception and Immersion in Virtual Reality

Perception and immersion. In recent years, a body of VR research has sought to identify the factors that cause the distance compression effect often observed in VR. This helps resolve egocentric perceptual deficiencies and improve future VR designs. These studies typically use verbal estimates [12, 35, 73], blind walking [22, 28], and perceptual matching tasks [37] to assess distance perception. Verbal estimates require participants to report the absolute distance of a target numerically, while blind walking also involves participants' perceptual-motor skills [45, 59]. Perceptual matching tasks typically investigate the ordinal depth of multiple objects. Although different tasks have their own advantages for investigating perception in VR, verbal estimates are reported to remain consistent across a wide range of distances [29, 43] and environments, regardless of whether they are modeled virtual environments with lower visual fidelity or photo-realistic captured scenes [12], which is crucial for this work.

Egocentric perception in VR is typically influenced by human, technical, and environmental factors. In terms of human factors, physical characteristics such as gender [5], age [50], and height [50] do not significantly affect distance perception, and prior experience with VR does not improve distance estimation accuracy [50]. However, the feeling of presence, or immersion in VR, has been found to influence judgments [23, 25]. One possible reason is the poor fit of the head-mounted display (HMD) during experiments. To address this issue, we include a presence survey in our user studies to assess perception in VR and evaluate the experienced immersion for eye height adaptation.

Technical factors also play a role in influencing egocentric perception in VR. For instance, newer hardware systems that provide a larger Field of Views (FOVs) have been shown to improve accuracy in a range of tasks [3, 28, 75]. In addition, ergonomic design can alleviate the distance compression effect [3, 24, 69, 70]. High display resolution also contributes to better visibility and improved presence and perception [12]. Stereoscopic vision provides depth cues through disparity [6], but its effectiveness diminishes for distant targets [33, 52]. For tasks beyond proximity, stereoscopic vision does not offer a clear advantage over monocular vision [5, 71].

Environmental factors, such as the realism and composition of the virtual environments, also play a role in influencing egocentric perception in VR [50, 63, 65]. Highly realistic environments improve accuracy compared to non-photorealistic renderings [65]. In non-photorealistic environments, participants consistently underestimate distances [53, 56, 59], whereas in real-world ones, estimations are often accurate [56]. We therefore hypothesize that adapting the eye height for captured real-world scenes would benefit perception and immersion in VR. Further research also indicates that compositional visual cues, such as linear perspective and ground textures, affect performance. Considering indoor and outdoor scenes play an important role in perception [1], we prepared both types of scenes to counteract compositional factors and study outdoor conditions, which are under-researched.

Influence of eye heights. The eye height is a crucial source of information for egocentric distance and depth perception [35] [12] [46]. By observing the proportion of the horizon occluded by an object, individuals can infer the height and distance of the target based on either an explicit or implicit horizon in the environment [58]. Knowing their eye level allows individuals to estimate the distance of an object based on its size. Previous research has shown that manipulations of eye height have a significant impact on perceived distance in virtual environments [35]. Increasing the virtual eye height by 50 cm increases the distance compression effect [57], while decreasing the virtual eye height does not have a significant influence on perception [27] [35]. Investigating the perception of the real world is more challenging, as it is not easy to manipulate eye height for pre-captured environments [12]. Previous research mainly focus on resolution [11], FOV [47], realism [12], and other HMD-related factors [10] when investigating perception of pre-captured environments. One study attempts manipulating height targets with a constant eye height in real world, observing a compression of distance perception similar to virtual environment studies [51]. Recent research also revealed that indoor/outdoor conditions affect distance perception in real environments [9]. In this research, we use state-of-the-art capture equipment to capture real-world environments at different eye heights and a HMD to investigate the influence of eye heights on perception.

2.2 Novel View Synthesis for Virtual Reality

Omnidirectional-aware novel view synthesis. Novel view synthesis from pre-captured images is a persistent challenge in computer vision and computer graphics. Traditionally, structure-from-motion (SfM) and multi-view stereo are applied to a collection of images to estimate point clouds and camera extrinsic through geometric models [54]. However, it usually required sufficient baselines for multiple viewpoints [39], and the computation is quite expensive to generate and represent the entire scene with detailed meshes. To address this challenge, researchers have proposed several representation methods that allow for the generation of novel views without the need for a complete 3D model. These methods include multi-plane images (MPI) [48, 62], layered depth images [36, 60], and light fields [26, 34]. However, each of these methods has its own limitations. For example, MPI representation is lightweight and can capture specular surfaces, but its discretized representation can lead to suboptimal performance for sloped surfaces [66]. In addition, MPI representation with predetermined layer structures can suffer from abrupt layer changes across discontinuities in depth, leading to inferior preserved locality. LDI representation, on the other hand, allows for arbitrary depth complexity with great efficiency thanks to

its sparsity. Recent work has proposed storing connectivity information between layers in LDI representation [16, 17, 36], which allows for the breakdown of the global inpainting problem into sub-areas that can be solved iteratively. This representation is also well-suited for omnidirectional images due to its extremely large field of view.

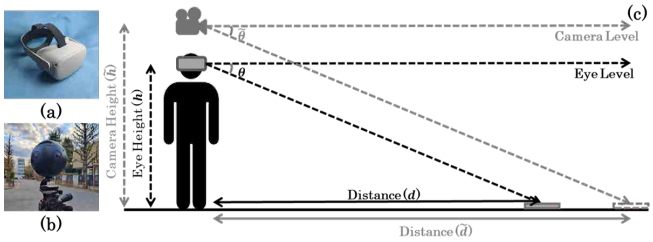

Figure 2: The configuration of the pilot study: (a) HMD (Meta Quest 2). (b) Omnidirectional camera (Insta360 Pro 2). (c) Study design.

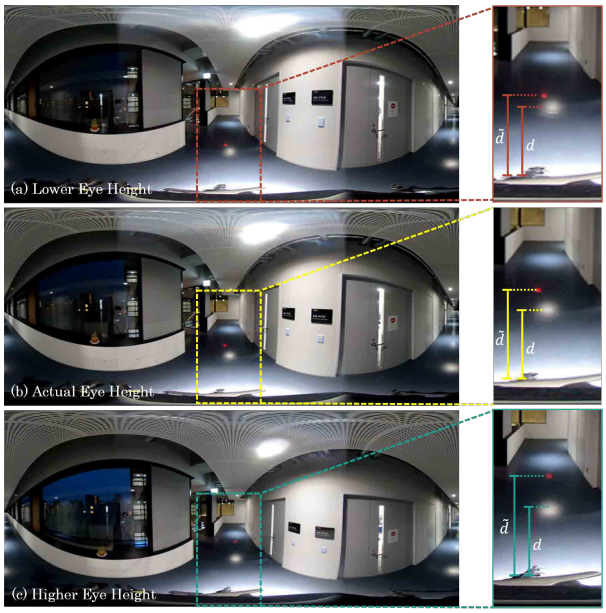

Figure 3: An example of a pre-captured immersive environment with different eye heights. The distances to the white spot (d) and the angle of declination to the red spot (θ) are kept constant across each condition. As demonstrated, the different eye heights result in discrepancies in the perception of $\tilde{d}$ in virtual environments.

Recently, CNN-based approaches have been used to generate views from sparse light field data [26], piecewise planar images [41], and neural radiance field (NeRF) [49]. Light-field photography allows for photorealistic rendering of novel views, but generally requires dense capture of the scene to achieve good results. Piecewise planar image-based approaches enable view synthesis from monocular image input, but are restricted to specific scenarios [41]. Pathdreamer [30] uses 2D image-to-image translation to synthesize novel views for omnidirectional RGB-D images, but it focuses on a specific domain (indoor scenarios) with limited 3D consistency. NeRF [49] shows great potential with high-quality and photorealistic views for complex scenes, but it requires large amounts of consistent input images with known relative positions. Later research has proposed ways to alleviate the requirement of input [74] and to reduce computational cost [18] [8]. For omnidirectional input, OmniNeRF [19] learns from omnidirectional RGB-D images and shows good performance for our application. However, it relies on neighborhood interpolation to complete occluded regions, which can result in visual artifacts. Combining with additional weakness of fine details and lengthy per-frame training, we find image-based methods to be a more practical choice for this application.

Image inpainting. Image inpainting is the process of filling in missing regions of an image with plausible content. Traditional methods for inpainting include example-based methods, which transfer the texture of other regions to the missing pixels using non-parametric patch synthesis [7, 20] or Markov Random Fields to propagate from the boundaries [31]. More recent approaches have employed convolutional neural networks (CNNs) to predict semantically meaningful results by learning from large training datasets [21]. Further improvements to network architectures have been proposed to handle irregularly-shaped holes [42] with diffusion models [44]. Two-stage approaches have also been developed that predict the structure of missing areas before completing them contextually [44]. For omnidirectional images, cubemap projection is used to represent the spherical nature of the inputs [13]. Data-driven image completion for 360-degree images has shown promising results when combined with OmniNeRF [19], but is limited to indoor scenarios due to limited training data. In this paper, we extend the LDI-based approach with cubemap representation and adopt the two-stage approach [60] for inpainting missing regions. By breaking down the large field of view into non-distorted local regions, we can solve the inpainting problem using a standard CNN.

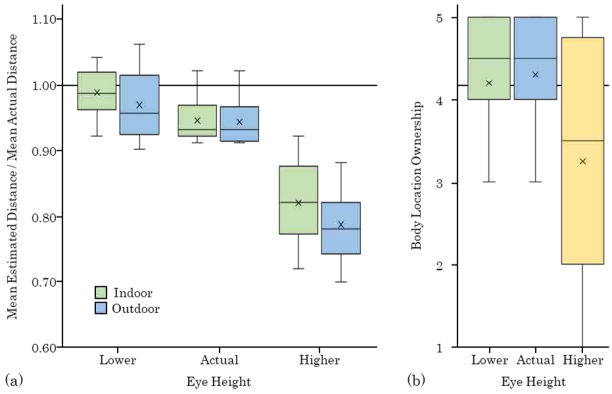

Figure 4: Pilot study results in distance perception and immersion.

3 Pilot Study: Eye Height in Pre-Captured Real Environments

Study concept. The aim of this pilot study is to examine the potential influence of eye height on perception and immersion in pre-captured immersive environments. It has been suggested that the more closer the visuals in a virtual environment match real-world experience, the greater the sense of presence in that environment [12]. While the influence of altered eye height in model-based virtual environment has been verified [35], it has yet to be confirmed for pre-captured real environments.

Design and tools. We conducted experiments that consisted of two stages: capture and playback. During the capture stage, we used a state-of-the-art omnidirectional camera to record real-world environments with high visual fidelity. In the playback stage, participants used a HMD to view the pre-captured environments in an immersive way. The HMD used in the experiment was the Oculus Quest 2, which had a display resolution of 1832×1920 per eye and a refresh rate of 90 Hz. The effective horizontal and vertical FOV of the HMD was $90^\circ \pm 5^\circ$ and $93^\circ \pm 5^\circ$, respectively. This difference was due to the varying distances between participants' eyes and the lens of the HMD (e.g., whether they were wearing glasses or not). The interpupillary distance used during the experiments was $6.5cm$, as suggested by previous studies [64]. During playback, we used a basic omnidirectional video playback function in the HMD,

which allowed for 3-degrees-of-freedom interactions between the participants and the environment.

To assess perception, participants were asked to provide verbal estimates of the distance of a truncated cone-shaped target with a height of 10*cm* in multiple pre-captured environments with varying eye heights. The dimensions of the target were not disclosed to participants in order to prevent the use of prior experience in determining distance [12]. To assess immersion, we incorporated presence and embodiment questionnaires [72] and modified it with a 5-point Likert scale ranging from "totally agree" to "totally disagree". It assesses participants' subjective experiences of body ownership in the virtual environment (e.g., whether they felt as if their own body was located where the virtual body was seen to be).

In our study, the captured real environments include two outdoor scenes (a park, a road) and two indoor scenes (a room, a corridor). For each environment, we captured a range of perspectives at varying eye heights from the lowest (140*cm*) to the highest (190*cm*) with 1*cm* increment based on the average eye height (165*cm*), revealed by previous research [35]. We then determine a lower eye height ($-25cm$) and a higher eye height ($+25cm$) based on the actual eye heights of each participant in the experiment. Fig. 2 shows the setup used during the pilot study, and Fig. 3 shows the captured real environments used in the experiment. Since stereoscopic vision does not significantly differ from monocular vision for perception beyond proximity, and to streamline the capture process for various height conditions, the captured environments were in monoscopic format. To prevent participants from using the movement of the target to guess the distance of the target, we prepared isolated clips of the real environments with the target positioned at fixed distances of 4*m*, 5*m*, and 6*m* for each condition.

Procedure. After providing instructions to the participants, they were equipped with a HMD and underwent a brief calibration process. To familiarize themselves with the HMD, participants were given the opportunity to try it out before the start of the experiment. Each experiment consisted of a sequence of 36 distance estimation trials, during which participants were asked to view a pre-recorded virtual environment through the HMD and estimate the distance of a target object within the scene. The virtual environments consisted of four different settings (i.e, a park, a road, a room, and a corridor), and the target distances and virtual eye heights were varied across trials. To ensure that participants understood the task and to collect their distance estimates, the experimenter communicated with them verbally throughout the experiment. Participants did not receive feedback on the accuracy of their estimations during the experiment. Upon completion of the distance estimation sequence, participants removed the HMD and were instructed to complete the accompanying questionnaires and provide their consent.

Results and findings. We recruited 20 volunteers from the university, and the participants consisted of 14 males and 6 females. The sample was relatively homogeneous with an average age of 23.7 years old ($SD = 2.93$, 19-28 years old) and an average height of 168.9 cm ($SD = 9.72$, 154-183 cm). We set the level of significance to $\alpha = 0.05$ and the power of the test to 1 - $\beta = 0.8$.

No estimate given by the participants was removed from the analysis for being three standard deviations apart from the mean estimate. The analysis was conducted with a repeated-measures ANOVA with distance as the within-subjects factor, eye height and environment as between-subjects variables, and estimated distances as the dependent measure. Confirming our hypothesis, the influence of eye height on distance perception was significant (see Fig. 4). Across the lower eye height (M_{indoor} = .98, SE_{indoor} = .035, $M_{outdoor}$ = .97, $SE_{outdoor}$ = .047), actual eye height (M_{indoor} = .94, SE_{indoor} = .033, $M_{outdoor}$ = .94, $SE_{outdoor}$ = .035), and higher eye height (M_{indoor} = .82, SE_{indoor} = .060, $M_{outdoor}$ = .79, $SE_{outdoor}$ = .050), estimated distances varied significantly ($p < .001$). With Fisher's LSD tests, we found that the estimates given by the participants were different from the higher eye height significantly ($p < .001$), while a significant difference between the estimates under lower eye height and actual eye height was absent ($p = .25$). Furthermore, a similar effect was verified for body location ownership: a significant difference between the higher eye height and the other two conditions ($p < .001$), but no significant difference between the actual and the lower eye height ($p = .85$). The actual and lower height conditions showed very high responses of ownership, 4.3 and 4.2, on the 5-point Likert scales.

4 Learning-based Eye Height Adaptation

Motivated by the findings of the pilot study, we propose a novel learning-based approach for synthesizing novel views with varying eye heights from omnidirectional images. The pipeline is shown in Fig. 5). The system consists of two stages: depth estimation and image inpainting. After receiving the color input, the depth estimation stage uses a novel multitask network to simultaneously provides depth information and semantic segmentation guides from RGB input. In the inpainting stage, we improve existing LDI representation to enable high-quality inpainting results for omnidirectional images. By leveraging the semantic guides (see Fig. 7), we generate visually convincing results at occluded regions. Finally, we merge the inpainted results back into the original LDI representation to render novel views with altered eye heights.

4.1 Depth estimation and semantic segmenatation

We present a novel multitask architecture that simultaneously learns depth and semantic segmentation in two different formats (Fig. 6). By simultaneously learning a semantic segmentation task in paral-

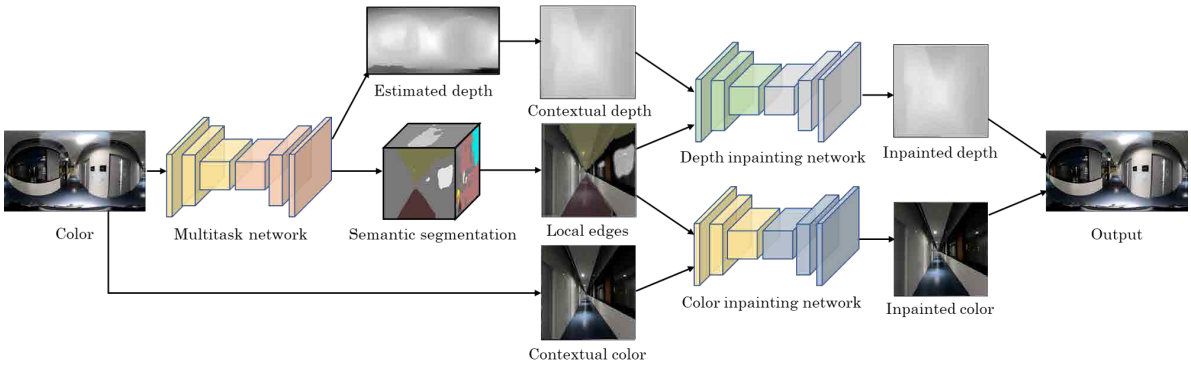

Figure 5: The overview of the proposed eye height adaptation approach. The proposed eye height adaptation approach is a multitask architecture that leverages both equirectangular and cubemap projections to predict depth and semantic segmentation. The architecture is omnidirectional-aware and uses semantic segmentation as guiding information to complete the layered depth and color images using inpainting networks. The final step involves synthesizing natural and realistic visuals that are adapted to different eye heights for improved perception and immersion.

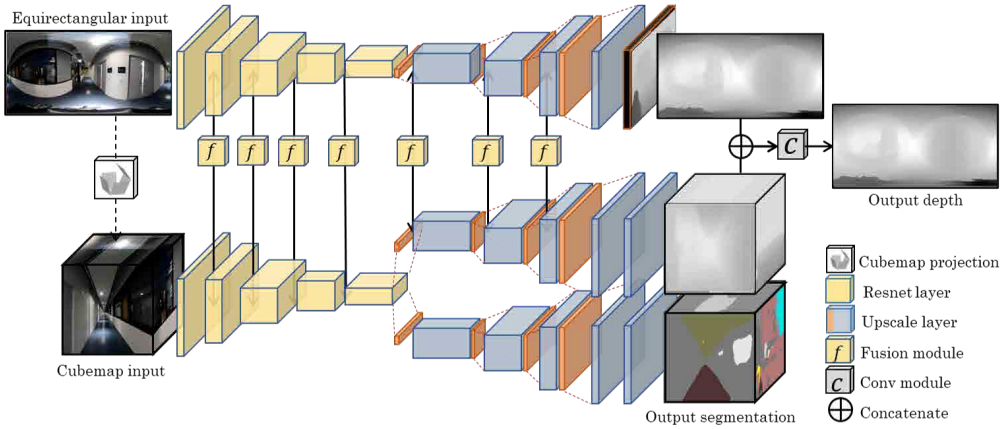

Figure 6: The overview of the proposed multitask network. Depth estimation: the top and middle branches learn to predict the depth of the same scene with both equirectangular projection and cubemap projection, leading to sharper boundaries for local objects while maintaining consistent and smooth prediction for the entire scene with the extreme FOV. Semantic segmentation: the middle and bottom branches with cubemap projection jointly learn to reveal the scene layout with a smaller FOV, providing valuable guides and facilitating the inpainting stage.

lel with depth estimation, our model not only improves boundary estimation for local objects, but also facilitates the subsequent image inpainting stage by providing useful semantic guides. This approach offers a promising solution for synthesizing novel views from omnidirectional images.

Multitask architecture. To obtain accurate depth maps for omnidirectional images, we propose regressing dense global depth estimation from a single view equirectangular image in two different projections: equirectangular and cubemap. For the equirectangular input, the architecture has an encoder-decoder structure that progressively down-projects and up-projects to the original size. The advantage of directly learning depth estimation on the entire 360-degree input is that low spatial frequencies better represent global features such as structures and shapes in the scene. Coarse perception of the scene can be further exploited by a smoothness loss function to ensure that the learned depth is consistent and uniform. However, the disadvantage of projecting a sphere onto a flat 2D plane is the strong distortion introduced by uneven pixel densities. Distortion is stronger for sparse pixels near the poles and less prominent at the equator [77]. To address this issue, we use rectangular filters with varying sizes at the first convolution layer along the vertical axis of the input equirectangular image. The encoder of this branch shares the same structure as ResNet-50, while the decoder consists of four up-projection blocks [32].

For the cubemap input, we first project the spherical image onto a cube to obtain cubemap faces. The use of cubemap projection in depth estimation is motivated by the desire to reduce distortion in the input image and provide higher spatial frequencies for improved shape and boundary detection of local objects. When directly using equirectangular images to learn depth estimation, details of local objects with steep gradient changes are usually omitted during the training process. By learning the depth estimation from both the equirectangular and cubemap projections, our model is able to learn complementary features from the same input. To encourage feature sharing and balance between the two branches, we use a fusion process and spherical padding [68] to connect the cube faces. This allows our model to adapt to weights trained for pinhole cameras and improve learning accuracy and efficiency. The final depth estimation is generated by projecting the output from the cubemap back onto the equirectangular projection and applying a convolution module.

We use m_p and m_f to represent feature maps from the equirectangular and cubemap branches, respectively. These maps are reprojected to $\hat{m_f}$ in equirectangular format and $\hat{m_p}$ in cubemap projection and fed into the next layer of the respective branches. In this layer, the reprojected maps are passed through a convolution layer (C) and added to the original feature maps. The result, $m_p + C(\hat{m_f})$, is then passed to the next layer of the equirectangular branch, while $m_f + C(\hat{m_p})$ is passed to the cubemap branch. This fusion process enables our model to learn complementary features from the two projections, improving the accuracy of the depth estimation.

To enhance the detection of depth discontinuities and improve the performance of the inpainting stage, our multitask architecture learns semantic segmentation from the cubemap input. We use the same encoder for both depth estimation and semantic segmentation to improve boundary recognition. Additionally, we train a separate decoder to generate semantic segmentation in the cubemap format. With a similar FOV of perspective images, the branch utilizes abundant training data and pre-trained weights to improve the accuracy and efficiency of the entire training process of the proposed network.

Loss Functions. In our model, we use supervised loss constraints for both the depth estimation and semantic segmentation tasks. For the depth estimation task, we use the inverse Huber loss function, which is defined in [32], as the optimizing objective $L_{Berhu}(d_i, \hat{d_i})$:

$$L_{Berhu}(d_i, \hat{d_i}) = \begin{cases} |d_i - \hat{d_i}| & |d_i - \hat{d_i}| \geq c \\ \frac{(d_i - \hat{d_i})^2 + c^2}{2c} & |d_i - \hat{d_i}| > c \end{cases} \quad (1)$$

where d_i is the ground truth depth of the ith pixel, and $\hat{d_i}$ is the predicted depth of the ith pixel, and $c = \max(|d_i - \hat{d_i}|)/5$. The loss function L_{Seg} for semantic segmentation is a cross-entropy loss between the estimated segmentation $\hat{s_i}$ and the result predicted with a pre-train network S. Combined, the total loss can be defined as:

$$L_{Total} = L_{Berhu}(d_i, \hat{d_i}) + L_{Seg}(s_i, \hat{s_i}) \quad (2)$$

4.2 Context-aware inpainting

In the inpainting stage, we improve upon the existing LDI approach [60] by incorporating guidance from the omnidirectional-aware depth and semantic segmentation information. Our method can therefore synthesize realistic textures for occluded regions, leading to more natural and realistic images. This is especially beneficial for pre-captured environments, where the ability to adapt to different eye heights is crucial for creating a convincing experience.

Backbone. LDIs are effective ways to synthesize novel viewpoints from color and depth information. LDIs can handle arbitrary numbers of layers, which allows them to represent complex scenes. Each LDI pixel contains color information, a corresponding depth value, and pointers to horizontal and vertical neighboring pixels. In the case of depth discontinuities, LDI pixels will have zero neighboring pixels in the relevant cardinal direction.

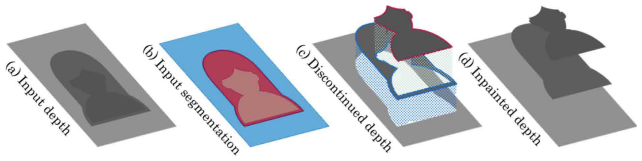

Figure 7: Illustration of the proposed process of LDI. For fully connected input depth image (a), we first separate the input according to segmentation information (b), resulting in foreground (red) and background regions (gray). After missing pixels are inpainted and no holes are left for each layer, results are then merged into the final LDI.

To use an LDI (Fig. 7), we first initialize it with a single layer that is fully connected in all directions. We then utilize the estimated semantic segmentation information and the previously-generated depth map to identify depth discontinuities and group them into simple and connected local edges. Subsequently, we disconnect LDI pixels across these edges and apply inpainting networks to fill in the occluded regions for both the depth map and color image. After inpainting, we merge the synthesized pixels back into the LDI until all the local edges have been resolved.

Multi-layer inpainting. Our goal is to use LDIs to inpaint occluded regions in an omnidirectional image and render views from a different eye height. This will allow us to synthesize novel views that closely resemble the real environment from the specified viewpoint.

To identify the regions that require inpainting, it is essential to accurately identify depth discontinuities in the input image. Traditional methods employ thresholding algorithms, which result in blurred boundaries across multiple pixels [60]. In addition, generating photo-realistic output from existing LDIs requires precise pairing of depth and color images, which can be challenging. Although CNNs have been utilized to address this issue for single-image inputs, models designed for images captured with pinhole cameras often produce suboptimal results when synthesizing depth information, resulting in inconsistent novel views with artifacts [2]. To overcome these challenges and improve the accuracy of the generated LDIs, we employ the proposed omnidirectional-aware multitask network that generates guiding semantic segmentation and detects discontinuities in the estimated depth maps (Fig. 7 (c)). We create a binary mask, labeling depth discontinuities as 1 and the remaining pixels as 0. We then merge adjacent discontinuities into linked local edges and use connected component analysis to prevent merging across discontinuities. Finally, we exclude local edges with less than ten pixels in length to obtain the regions for the inpainting process.

To perform inpainting of both depth and color information, we utilize a standard encoder-decoder architecture with U-Net and partial convolution for the depth inpainting network, as proposed in [42]. The color inpainting network is structured similarly, with the same number of layers. The depth inpainting network takes the contextual depth and local edges as input, while the color inpainting network takes the contextual color image and local edges as input. The training objectives for each network are as follows:

$$L_{Depth} = \frac{1}{N}||S \odot (d_i, \tilde{d_i})|| \tag{3}$$

$$L_{Color} = \alpha(\frac{1}{N}||S \odot (c_i, \tilde{c_i})||) + \beta L_{Perceptual} \tag{4}$$

where d_i is the ground truth depth of the ith pixel, $\tilde{d_i}$ is the inpainted depth of the ith pixel, c_i is the ground truth color of the ith pixel, and $\tilde{c}i$ is the inpainted color of the ith pixel. S is a binary mask that describes the contextual region, $\odot$ denotes the Hadamard product, and $L_{Perceptual}$ is the loss function for the color inpainting task. It is obtained using the output of layers from a pre-trained VGG-16 model . The color inpainting network is trained on COCO-2017 [40], while the depth inpainting network is trained on MegaDepth [38].

Eye height adaptation. To modify the viewpoint of a pre-captured environment with a different eye height, we use a vertical geometric model to reproject the LDIs to the desired viewpoint, as depicted in Fig. 1. Specifically, we apply the vertical spherical model introduced in [77] to transform the original view j to the altered view k with a vertical disparity d. The transformation is done in polar coordinates, where each point p at (x,y,z) in Cartesian coordinate is represented by its longitude ϕ and latitude θ. The radial distance r (i.e., depth value) of a point is given by $\sqrt{x^2+y^2+z^2}$, and the vertical distance is defined as $\delta = (\phi_j - \phi_k, \theta_j - \theta_k)$. As we only need to adapt the eye height, we only consider vertical disparity $d = (0, dy, 0)$, and the disparity is reduced to $\delta = (\frac{\partial \phi}{\partial y}, \frac{\partial \theta}{\partial y})$.

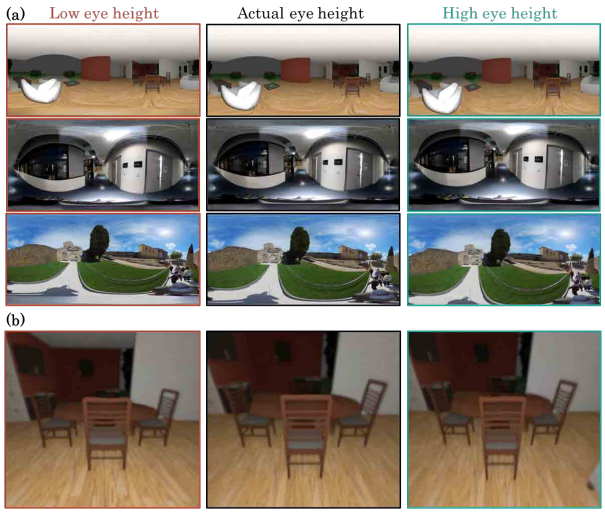

Figure 8: Results of the proposed eye height adaptation approach. (a) shows the omnidirectional visuals when adapted to a higher eye height ($+25cm$) and a lower eye height ($-25cm$). (b) shows perspective visuals when viewed with a smaller FOV (i.e., HMD).

To render a target view $\hat{k}$ from the source input j, each pixel $p = (\phi, \theta)$ on the equirectangular image is a function of the vertical disparity d and the radial distance r. Since we already have the depth and color information from LDIs, we can compute the target frame $\hat{k}$ with a function:

$$\hat{k}(p) = \Gamma_{j \to \hat{k}}(\tilde{d}, d_{j \to k}, j(p)) \tag{5}$$

5 Experimental Results

5.1 Implementation Details

We have implemented our proposed multitask network using the PyTorch framework , and trained it on a single Nvidia RTX 2080Ti graphics card, using data from the Depth360 dataset [13]. During training, we employed the Adam optimizer with a learning rate of 3e-4, and used a batch size of 1 due to graphics memory constraints. Our equirectangular branch was initialized with Xavier initialization [15], while the cubemap branch was initialized with ImageNet pretrained weights. Our approach takes around $150ms$ to predict depth maps and semantic segmentation for a single equirectangular image. For the contextual inpainting networks, we trained the depth inpainting network on MegaDepth [38] for 5 epochs, while the color inpainting network was trained on the MS-COCO [40] for 10 epochs. We used $\alpha = 1$ and $\beta = 0.05$ as the parameters for the inpainting process.

5.2 Qualitative Evaluation

We qualitatively evaluates our proposed eye height adaptation method for pre-captured immersive content. Fig. 8 displays the visual results obtained when our method was tested on unseen equirectangular images with both indoor and outdoor settings. As shown in Fig. 8 (b), our method successfully generates new perspectives with varying eye heights. To showcase the effectiveness and accuracy of

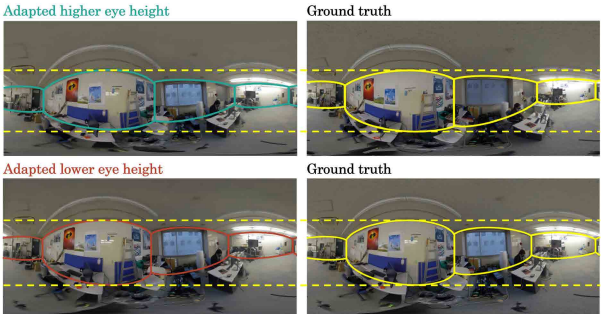

Figure 9: Comparisons of scene structures between adapted views and captured ground truth showed that the generated visuals at both higher ($+25cm$) and lower ($-25cm$) adapted eye heights matched the ground truth scene structures.

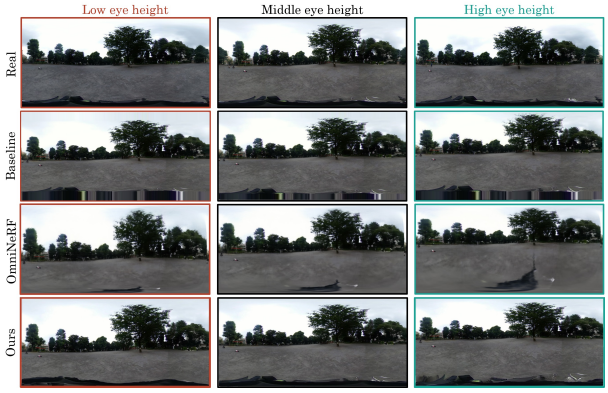

Figure 10: Qualitative comparisons against a baseline method of parallax mapping that completes the occluded regions with bilinear interpolation, and a state-of-the-art NeRF-based method designed for novel view synthesis from omnidirectional input.

our method, we captured the ground-truth at respective eye heights using a 360-degree camera. As demonstrated in Fig. 9, the adapted views from different eye heights match the ground-truth with high accuracy, demonstrating the efficacy of our approach.

In addition, we conducted a comparative analysis of our method against a baseline approach and a state-of-the-art NeRF-based method called OmniNeRF [19]. The baseline approach utilizes parallax mapping based on the new viewpoint and the depth estimated by our multitask network, and then employs bilinear interpolation for occluded pixels to complete the view. The results of the experiment are presented in Fig. 10 and Fig. 11. Our proposed method generates visually compelling results with sharper details without the need for re-training for each scene over prolonged periods. On the other hand, although NeRF is capable of synthesizing images at arbitrary resolution through its implicit formation, the experiment showed low visual fidelity due to sparse sampling in a single omnidirectional image. Nonetheless, it is worth noting that NeRF methods excel in rendering specular surfaces and reflections with its ray tracing capability, while image-based methods encounter challenges in achieving comparable realism in this aspect.

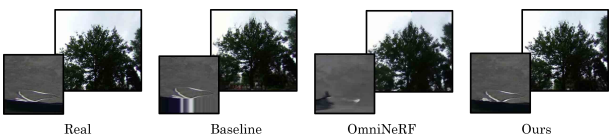

Figure 11: Close-up views for the output with adapted eye heights. Compared to other approaches, the proposed method generates more natural and clearer visuals for both local regions and image boundaries with strong distortion.

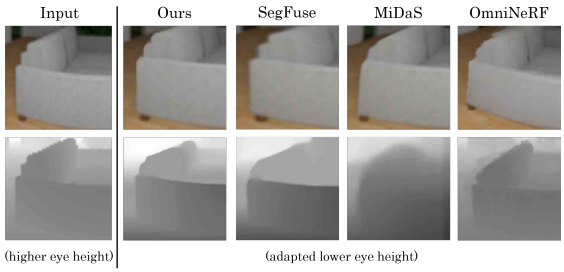

Figure 12: Comparisons between different depth estimation approaches. For the same input, we swap the proposed multitask network to different depth estimation models for eye height adaptation.When decreasing the eye height by $25cm$, the proposed method shows the sharpest edges for local objects and the most consistent depth estimation.

5.3 Quantitative Evaluation

In this section, we assess the accuracy of our proposed approach in adapting input RGB images to different eye heights for omnidirectional images and compare it with state-of-the-art methods. To mitigate potential lighting and scene configuration variations when capturing the same real-world scene with different eye heights, we utilized SynDepth360 [13], a small-scale synthetic omnidirectional dataset with 3D models, to generate novel perspectives with virtual camera placements altered for evaluating different methods. We employed SSIM, PSNR, and LPIPS metrics [60] to quantify output quality as shown in Table 1.

Run-time efficiency. We evaluate the processing time and memory requirements from inputting the image to render adapted views in Table 2. For our method, it means passing the input through the feed-forward models (Fig. 5); for NeRF, it means the system augments the single input and learn volumetric representation/rendering. In addition to greatly improved run-time efficiency, our LDI-based method is memory efficient, making it a viable option for future applications on commercial HMDs with independent processors.

Network evaluation. We aim to evaluate the performance of our proposed multitask network for accurate depth estimation, which is crucial for generating the LDI required for rendering adapted views. To assess our approach, we adopt commonly used depth prediction metrics from literature to evaluate robustness and accuracy against state-of-the-art techniques. Table 3 presents the results of our proposed method alongside those of SegFuse [13] and BiFuse [68], which are omnidirectional-aware methods, as well as perspective-based methods MiDaS (v2.1) [55] and MegaDepth [38]. All models were trained using their released code bases for 20 epochs. We further demonstrate the results for adapted eye height with swapped depth estimation components in Fig. 12. While our multitask network outperforms all other methods in predicting the layout of indoor omnidirectional images, the depth estimation learned through cubemap projection can suffer from boundary issues in outdoor settings, as previously pointed out in [13]. Despite this limitation, our method still shows comparable accuracy to SegFuse in outdoor settings. Overall, our method achieves high accuracy in depth estimation and effectively achieves natural and accurate eye height adaptation.

5.4 User Study

The goal of this study was to examine the effectiveness of the proposed eye height adaptation approach for pre-captured immersive environments. These experiments explore the influence of manipulated eye heights on perception and immersion using verbal estimations of egocentric distances and body location ownership questionnaires.

Participants and apparatus. In this study, we recruited a sample of 22 participants from the university, including 15 males and 7 females. The average age of the participants was 23.6 years old ($SD = 3.1$), and the average height was 166.9 cm ($SD = 9.17$). All

Table 1: **Quantitative comparison** on the SynDepth360 dataset [13].

Method	*SSIM* ↑	*PSNR* ↑	*LPIPS* ↓
Baseline	0.8347	23.74	0.0982
OmniNeRF [19]	0.8528	26.74	0.0824
Ours	**0.8776**	**27.19**	**0.0736**

Table 2: **Average computation time and memory requirement per image.** Every approach is evaluated using Nvidia RTX 2080Ti and Intel i7-7800X, using input with a resolution of 1024 × 512.

Method	*Computation time* $[s]$	*Memory cost* $[MB]$
OmniNeRF [19]	$> 250,000$	13.6
Ours	**0.450**	**1.42**

Table 3: **Quantitative results** of depth estimation on Matterport3D dataset [4], SunCG dataset [67], and Depth360 dataset [13].

Method	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
The Matterport3D dataset [4] (Real domain, indoor settings.)					
BiFuse [68]	0.6259	0.1134	84.52%	93.19%	96.32%
SegFuse [13]	0.6029	0.1100	84.60%	93.14%	96.28%
MiDaS [55]	0.7641	0.1420	77.05%	88.94%	95.77%
MegaDepth [38]	0.7845	0.1502	69.50%	87.94%	94.21%
Ours	**0.5880**	**0.0986**	**85.17%**	**93.45%**	**96.85%**
The SunCG dataset [67] (Synthetic domain, indoor settings.)					
BiFuse [68]	0.2596	0.0443	**95.90%**	98.23%	99.07%
SegFuse [13]	0.2540	0.0427	95.32%	98.34%	99.10%
MiDaS [55]	0.3244	0.0730	89.90%	96.62%	95.44%
MegaDepth [38]	0.4041	0.0845	84.06%	93.92%	94.85%
Ours	**0.2490**	**0.0425**	95.57%	**98.45%**	**99.26%**
The Depth360 dataset [13] (Real domain, outdoor settings.)					
BiFuse [68]	5.0725	0.8316	40.13%	59.17%	67.92%
SegFuse [13]	**4.0442**	**0.7777**	**82.26%**	**91.35%**	**94.22%**
MiDaS [55]	6.0132	0.9574	52.26%	58.73%	65.31%
MegaDepth [38]	6.7320	0.9863	48.33%	61.75%	64.82%
Ours	4.0544	0.7965	81.07%	90.89%	94.18%

participants had normal visual acuity and were comfortable using HMDs. The experiment was implemented in Unity3D and played back wirelessly on an Meta Quest 2 HMD using Air Link. The system used an Nvidia RTX 2080 Ti GPU, an Intel i7-7800X CPU, and 32GB RAM. The HMD had a refresh rate of 90 HZ and a field of view of $90^\circ \pm 5^\circ$ horizontally and $93^\circ \pm 5^\circ$ vertically, based on individual participant fitting. We used the same questionnaires as in the pilot study to assess perceived immersion and the same metric to compare the discrepancy between estimated and actual distance.

Design and procedure. In our main study, we used a within-subjects design to manipulate eye height as the independent variable. Unlike the pilot study, we used the HMD's tracking capability to obtain the participant's actual eye height. We used the lowest capture (140 cm) as the input to the pipeline, and rendered monoscopic views from 141 cm to 190 cm. During the study, we prepare adapted, low ($-25cm$), and high ($+25cm$) eye heights for each user. We tested this method's effectiveness by providing four similar environments to the pilot study, two indoor and two outdoor scenarios, while minimizing the impact on distance estimation accuracy.

During the experiment, the experimenter first provided instructions to each participant and answered any questions until the participant had a clear understanding of the task. Before equipping the HMD, the experimenter showed the participant a one-meter ruler to ensure their understanding of distance. After loading each condition, participants had time to freely explore the environment until they were ready to verbally estimate the distance to the target. This was crucial for the post-experiment presence questionnaire, which assessed subjective body ownership and immersion. On average, each participant completed the experiment in approximately 20 minutes.

Results and general discussion. After conducting an analysis using a repeated-measures ANOVA with manipulated eye height (low, adapted, or high) and environment (indoor or outdoor) as the between-subjects factor, and estimated distances as the dependent measure, we confirmed the similar result to the pilot study that the adapted eye height significantly influenced distance perception (refer to Fig. 13). The estimated distances varied significantly for low eye height (M_{indoor} = .99, SE_{indoor} = .049, $M_{outdoor}$ = .96, $SE_{outdoor}$ = .046), adapted eye height (M_{indoor} = .95, SE_{indoor} = .039, $M_{outdoor}$ = .94, $SE_{outdoor}$ = .039), and high eye height (M_{indoor} = .81, SE_{indoor} = .066, $M_{outdoor}$ = .78, $SE_{outdoor}$ = .055), with the effect being statistically significant ($p < .001$). The outdoor conditions show significant ($p < .001$) distance underestimation when compared to indoor conditions. This aligns with the findings from previous research [46]. Regarding body location ownership, our findings show that the adapted eye height conditions resulted in improved immersion responses with an average of 3.14 compared to low eye height ($M = 2.81$) and high eye height ($M = 2.22$). Although a significant influence was still observed when the manipulated eye height was higher than the user's actual eye height ($p < .001$), no significant effect on immersion was confirmed when the manipulated eye height was lower than the actual eye height of the user, consistent with both the pilot study and previous research. This finding can be explained by the abundance of daily actions people perform to lower their eye height, while actions to increase eye height are less common. In summary, our approach effectively improves distance perception and provides better immersion with adapted eye height when the environment is captured with a higher declination angle.

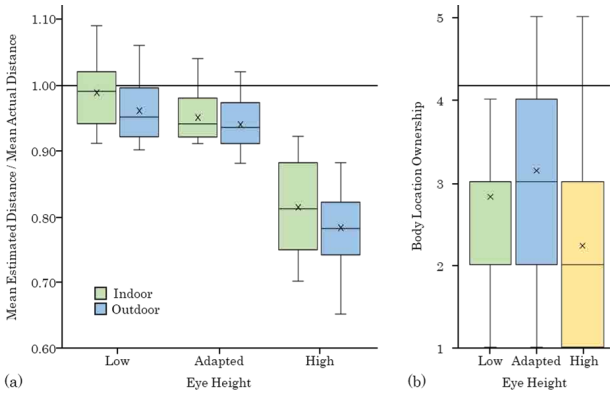

Figure 13: User study results in distance perception and immersion.

6 Conclusion and Future Work

In this paper, we present a learning-based eye height adaptation method that generates corrected views from pre-captured immersive environments based on the user's actual eye height during playback. We first conducted a pilot study to confirm the hypothesis that different eye heights significantly influence distance perception and immersion. Subsequently, we propose a learning-based approach with a novel multitask architecture that learns to predict depth and semantic segmentation for omnidirectional images with high accuracy. By utilizing layered depth image representation and image inpainting to generate views with altered eye heights, our approach efficiently synthesizes natural-looking visuals. Evaluation against state-of-the-art approaches demonstrates the effectiveness of our proposed method, while a user study confirms the improvements in perception and immersion. Future work will explore networks with better capability to generate high-resolution results for mixed reality, as well as the incorporation of efficient NeRF algorithms to improve accuracy in environments with specular surfaces. Finally, while this method can be applied to highly dynamic scenes on a per-frame basis, exploiting temporal and geometric information in videos is a another promising direction to ensure visual consistency.

Acknowledgements This research is supported in part by JSPS KAKENHI (ref: JP21H05054) and the EPSRC NorthFutures project (ref: EP/X031012/1).

REFERENCES

[1] J. Andre and S. Rogers. Using verbal and blind-walking distance estimates to investigate the two visual systems hypothesis. *Perception & Psychophysics*, 68(3):353–361, 2006.

[2] B. Attal, S. Ling, A. Gokaslan, C. Richardt, and J. Tompkin. Matryodshka: Real-time 6dof video view synthesis using multi-sphere images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020*, pp. 441–459. Springer, 2020.

[3] L. E. Buck, M. K. Young, and B. Bodenheimer. A comparison of distance estimation in hmd-based virtual environments with different hmd-based conditions. *ACM Transactions on Applied Perception (TAP)*, 15(3):1–15, 2018.

[4] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[5] S. H. Creem-Regehr, P. Willemsen, A. A. Gooch, and W. B. Thompson. The influence of restricted viewing conditions on egocentric distance perception: Implications for real and virtual indoor environments. *Perception*, 34(2):191–204, 2005.

[6] J. E. Cutting and P. M. Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of space and motion*, pp. 69–117. Elsevier, 1995.

[7] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012.

[8] N. Deng, Z. He, J. Ye, B. Duinkharjav, P. Chakravarthula, X. Yang, and Q. Sun. Fov-nerf: Foveated neural radiance fields for virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3854–3864, 2022.

[9] J. M. Dukes, J. F. Norman, and C. D. Shartzer. Visual distance perception indoors, outdoors, and in the dark. *Vision Research*, 194:107992, 2022.

[10] F. El Jamiy and R. Marsh. Distance estimation in virtual reality and augmented reality: A survey. In *2019 IEEE International Conference on Electro Information Technology (EIT)*, pp. 063–068. IEEE, 2019.

[11] F. El Jamiy and R. Marsh. Survey on depth perception in head mounted displays: distance estimation in virtual reality, augmented reality, and mixed reality. *IET Image Processing*, 13(5):707–712, 2019.

[12] I. T. Feldstein, F. M. Kölsch, and R. Konrad. Egocentric distance perception: A comparative study investigating differences between real and virtual environments. *Perception*, 49(9):940–967, 2020.

[13] Q. Feng, H. P. Shum, and S. Morishima. 360 depth estimation in the wild-the depth360 dataset and the segfuse network. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 664–673. IEEE, 2022.

[14] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5515–5524, 2016.

[15] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

[16] P. Hedman, S. Alsisan, R. Szeliski, and J. Kopf. Casual 3d photography. *ACM Transactions on Graphics (TOG)*, 36(6):1–15, 2017.

[17] P. Hedman and J. Kopf. Instant 3d photography. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018.

[18] Y. Hong, B. Peng, H. Xiao, L. Liu, and J. Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20374–20384, 2022.

[19] C.-Y. Hsu, C. Sun, and H.-T. Chen. Moving in a 360 world: Synthesizing panoramic parallaxes from a single panorama. *arXiv preprint arXiv:2106.10859*, 2021.

[20] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf. Image completion using planar structure guidance. *ACM Transactions on graphics (TOG)*, 33(4):1–10, 2014.

[21] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

[22] J. A. Jones, D. M. Krum, and M. T. Bolas. Vertical field-of-view extension and walking characteristics in head-worn virtual environments. *ACM Transactions on Applied Perception (TAP)*, 14(2):1–17, 2016.

[23] J. A. Jones, J. E. Swan, G. Singh, and S. R. Ellis. Peripheral visual information and its effect on distance judgments in virtual and augmented environments. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*, pp. 29–36, 2011.

[24] J. A. Jones, J. E. Swan, G. Singh, E. Kolstad, and S. R. Ellis. The effects of virtual reality, augmented reality, and motion parallax on egocentric depth perception. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*, pp. 9–14, 2008.

[25] J. A. Jones, J. E. Swan, G. Singh, S. Reddy, K. Moser, C. Hua, and S. R. Ellis. Improvements in visually directed walking in virtual environments cannot be explained by changes in gait alone. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 11–16, 2012.

[26] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–10, 2016.

[27] J. W. Kelly. Distance perception in virtual reality: A meta-analysis of the effect of head-mounted display characteristics. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[28] J. W. Kelly, L. A. Cherep, and Z. D. Siegel. Perceived space in the htc vive. *ACM Transactions on Applied Perception (TAP)*, 15(1):1–16, 2017.

[29] E. Klein, J. E. Swan, G. S. Schmidt, M. A. Livingston, and O. G. Staadt. Measurement protocols for medium-field distance perception in large-screen immersive displays. In *2009 IEEE virtual reality conference*, pp. 107–113. IEEE, 2009.

[30] J. Y. Koh, H. Lee, Y. Yang, J. Baldridge, and P. Anderson. Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14738–14748, 2021.

[31] N. Komodakis and G. Tziritas. Image completion using efficient belief propagation via priority scheduling and dynamic pruning. *IEEE Transactions on Image Processing*, 16(11):2649–2661, 2007.

[32] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pp. 239–248. IEEE, 2016.

[33] J. S. Lappin, A. L. Shelton, and J. J. Rieser. Environmental context influences visually perceived distance. *Perception & psychophysics*, 68(4):571–581, 2006.

[34] M. Levoy and P. Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 31–42, 1996.

[35] M. Leyrer, S. A. Linkenauger, H. H. Bülthoff, U. Kloos, and B. Mohler. The influence of eye height and avatars on egocentric distance estimates in immersive virtual environments. In *Proceedings of the ACM SIGGRAPH symposium on applied perception in graphics and visualization*, pp. 67–74, 2011.

[36] J. Li, Z. Feng, Q. She, H. Ding, C. Wang, and G. H. Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12578–12588, 2021.

[37] Z. Li, J. Phillips, and F. H. Durgin. The underestimation of egocentric distance: Evidence from frontal matching tasks. *Attention, Perception, & Psychophysics*, 73(7):2205–2217, 2011.

[38] Z. Li and N. Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2041–2050, 2018.

[39] K.-E. Lin, Z. Xu, B. Mildenhall, P. P. Srinivasan, Y. Hold-Geoffroy, S. DiVerdi, Q. Sun, K. Sunkavalli, and R. Ramamoorthi. Deep multi depth panoramas for view synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII*, pp. 328–344. Springer, 2020.

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*,

pp. 740–755. Springer, 2014.

[41] C. Liu, J. Yang, D. Ceylan, E. Yumer, and Y. Furukawa. Planenet: Piecewise planar reconstruction from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2579–2588, 2018.

[42] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 85–100, 2018.

[43] J. M. Loomis and J. W. Philbeck. Measuring spatial perception with spatial updating and action. In *Embodiment, ego-space, and action*, pp. 17–60. Psychology Press, 2008.

[44] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

[45] P. Maruhn, S. Schneider, and K. Bengler. Measuring egocentric distance perception in virtual reality: Influence of methodologies, locomotion and translation gains. *PloS one*, 14(10):e0224651, 2019.

[46] S. Masnadi, K. Pfeil, J.-V. T. Sera-Josef, and J. LaViola. Effects of field of view on egocentric distance perception in virtual reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–10, 2022.

[47] S. Masnadi, K. P. Pfeil, J.-V. T. Sera-Josef, and J. J. LaViola. Field of view effect on distance perception in virtual reality. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 542–543. IEEE, 2021.

[48] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

[49] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[50] A. Murgia and P. M. Sharkey. Estimation of distances in virtual environments using size constancy. *International Journal of Virtual Reality*, 8(1):67–74, 2009.

[51] T. L. Ooi, B. Wu, and Z. J. He. Distance determined by the angular declination below the horizon. *Nature*, 414(6860):197–200, 2001.

[52] S. Palmisano, B. Gillam, D. G. Govan, R. S. Allison, and J. M. Harris. Stereoscopic perception of real depths at large distances. *Journal of vision*, 10(6):19–19, 2010.

[53] A. Peer and K. Ponto. Evaluating perceived distance measures in room-scale spaces using consumer-grade head mounted displays. In *2017 IEEE symposium on 3d user interfaces (3dui)*, pp. 83–86. IEEE, 2017.

[54] E. Penner and L. Zhang. Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics (TOG)*, 36(6):1–11, 2017.

[55] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019.

[56] R. S. Renner, B. M. Velichkovsky, and J. R. Helmert. The perception of egocentric distances in virtual environments-a review. *ACM Computing Surveys (CSUR)*, 46(2):1–40, 2013.

[57] B. Ries, V. Interrante, M. Kaeding, and L. Anderson. The effect of self-embodiment on distance perception in immersive virtual environments. In *Proceedings of the 2008 ACM symposium on Virtual reality software and technology*, pp. 167–170, 2008.

[58] S. Rothe, B. Kegeles, and H. Hussmann. Camera heights in cinematic virtual reality: How viewers perceive mismatches between camera and eye height. In *Proceedings of the 2019 acm international conference on interactive experiences for tv and online video*, pp. 25–34, 2019.

[59] S. Schneider, P. Maruhn, and K. Bengler. Locomotion, non-isometric mapping and distance perception in virtual reality. In *Proceedings of the 2018 10th International Conference on Computer and Automation Engineering*, pp. 22–26, 2018.

[60] M.-L. Shih, S.-Y. Su, J. Kopf, and J.-B. Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8028–8038, 2020.

[61] R. Shimamura, Q. Feng, Y. Koyama, T. Nakatsuka, S. Fukayama, M. Hamasaki, M. Goto, and S. Morishima. Audio–visual object removal in 360-degree videos. *The Visual Computer*, 36(10):2117–2128, 2020.

[62] P. P. Srinivasan, R. Tucker, J. T. Barron, R. Ramamoorthi, R. Ng, and N. Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 175–184, 2019.

[63] R. T. Surdick, E. T. Davis, R. A. King, and L. F. Hodges. The perception of distance in simulated visual displays: A comparison of the effectiveness and accuracy of multiple depth cues across viewing distances. *Presence: Teleoperators & Virtual Environments*, 6(5):513–531, 1997.

[64] W. B. Thompson, P. Willemsen, A. A. Gooch, S. H. Creem-Regehr, J. M. Loomis, and A. C. Beall. Does the quality of the computer graphics matter when judging distances in visually immersive environments? *Presence*, 13(5):560–571, 2004.

[65] K. Vaziri, P. Liu, S. Aseeri, and V. Interrante. Impact of visual and experiential realism on distance perception in vr using a custom video see-through system. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 1–8, 2017.

[66] J. Waidhofer, R. Gadgil, A. Dickson, S. Zollmann, and J. Ventura. Panosynthvr: Toward light-weight 360-degree view synthesis from a single panoramic input. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 584–592. IEEE, 2022.

[67] F.-E. Wang, H.-N. Hu, H.-T. Cheng, J.-T. Lin, S.-T. Yang, M.-L. Shih, H.-K. Chu, and M. Sun. Self-supervised learning of depth and camera motion from 360*circ* videos. In *Asian Conference on Computer Vision*, pp. 53–68. Springer, 2018.

[68] F.-E. Wang, Y.-H. Yeh, M. Sun, W.-C. Chiu, and Y.-H. Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 462–471, 2020.

[69] P. Willemsen, M. B. Colton, S. H. Creem-Regehr, and W. B. Thompson. The effects of head-mounted display mechanics on distance judgments in virtual environments. In *Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization*, pp. 35–38, 2004.

[70] P. Willemsen, M. B. Colton, S. H. Creem-Regehr, and W. B. Thompson. The effects of head-mounted display mechanical properties and field of view on distance judgments in virtual environments. *ACM Transactions on Applied Perception (TAP)*, 6(2):1–14, 2009.

[71] P. Willemsen, A. A. Gooch, W. B. Thompson, and S. H. Creem-Regehr. Effects of stereo viewing conditions on distance perception in virtual environments. *Presence: Teleoperators and Virtual Environments*, 17(1):91–101, 2008.

[72] B. G. Witmer, C. J. Jerome, and M. J. Singer. The factor structure of the presence questionnaire. *Presence: Teleoperators & Virtual Environments*, 14(3):298–312, 2005.

[73] B. G. Witmer and P. B. Kline. Judging perceived and traversed distance in virtual environments. *Presence*, 7(2):144–167, 1998.

[74] D. Xu, Y. Jiang, P. Wang, Z. Fan, H. Shi, and Z. Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pp. 736–753. Springer, 2022.

[75] M. K. Young, G. B. Gaylor, S. M. Andrus, and B. Bodenheimer. A comparison of two cost-differentiated virtual reality systems for perception and action tasks. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 83–90, 2014.

[76] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[77] N. Zioulis, A. Karakottas, D. Zarpalas, F. Alvarez, and P. Daras. Spherical view synthesis for self-supervised 360 depth estimation. In *2019 International Conference on 3D Vision (3DV)*, pp. 690–699. IEEE, 2019.

[78] N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 448–465, 2018.