TerraFirma: A Low-Cost and Effective 360 VR Extension for Viewpoint Translation and Collaboration

Fengze Zhang, Ahaan Agrawal, and Voicu Popescu

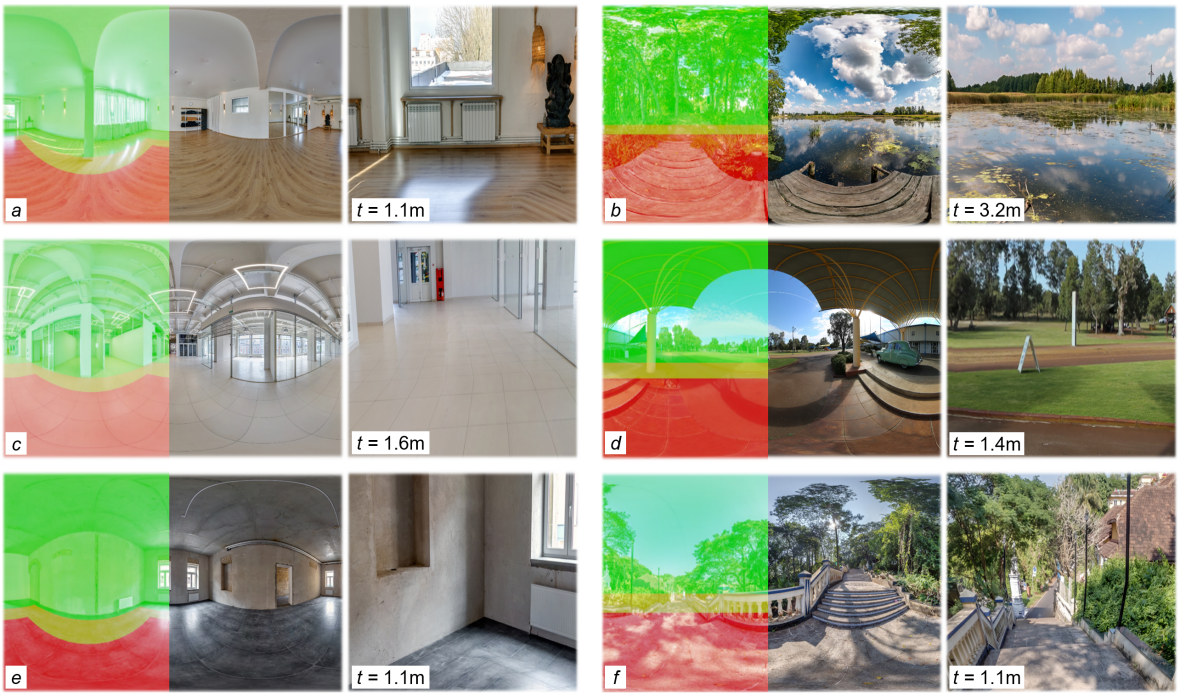


Fig. 1: Our *TerraFirma* 360 VR extension. In each of the six examples, the left image is the input equirectangular 360° image, and the right image is a sample output user frame. *TerraFirma* partitions the scene into three regions, as illustrated in the left half of the 360° images: a near (red), an intermediate (yellow), and a far (green) region. The output frame renders the near region through projective texture mapping, the far region through environment mapping, and the intermediate region through a combination of projective and environment mapping. *TerraFirma* supports user viewpoint translation—each frame gives the distance t between the user viewpoint and the 360° image viewpoint.

Abstract—This paper presents an approach for extending 360 VR to support viewpoint translation and collaboration, *without* increasing the acquisition cost. The approach models the ground under the 360° image acquisition viewpoint with a planar patch, on which the user stands and walks. The planar patch is connected to the background with visual continuity. The approach anchors the user's avatar to the ground patch, and supports collaboration, with multiple avatars standing on the patch. The approach only requires defining two trivial parameters: the approximate acquisition height, and the size of the patch. Therefore, the approach improves the immersive exploration of a real-world scene captured with a 360° image, without increasing the cost of the acquisition of the virtual environment. The approach was evaluated in a controlled user study (N = 30) with five user tasks. The results show that the approach has an advantage over conventional 360 VR in terms of anchoring the user to the ground, of user viewpoint translation, of collaboration, and of accurate integration of stationary and dynamic virtual objects into the 3D scene.

Index Terms—360 VR, viewpoint translation, collaboration, low acquisition cost, low rendering cost.

1 Introduction

Virtual reality (VR) technology can provide powerful immersive visualization not just when the virtual environment is a synthetic, imaginary world, but also when the virtual environment is a digital replica of a real-world scene. For example, virtual environments that capture real-world scenes in detail can be used for remote collaborators to pay each other effective virtual visits, for trainees to acquire real-world skills safely, and for anyone to relive vacation highlights vividly.

A major challenge impeding the immersive visualization of real-world scenes is the challenging acquisition and construction of a virtual environment that captures the real-world scene in detail. The problem has received extensive attention, from classical computer vision photogrammetry approaches [36], to structured-light [11] and time-of-flight [14] scanners, to light-field [21] and lumigraph [13] image databases, and to the recent neural radiance field [26] and 3D Gaussian splatting [18] approaches. Whereas these approaches can create quality digital replicas of real-world scenes, they have in common the challenge of high acquisition cost, which is due to the prerequisite of

- *Fengze Zhang is with Purdue University. E-mail: zhan5455@purdue.edu*
- *Ahaan Agrawal is with Purdue University. E-mail: agraw208@purdue.edu.*
- *Voicu Popescu is with Purdue University. E-mail: popescu@purdue.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

operator expertise, to high equipment cost, to long acquisition times, and/or to complex post-processing.

One approach that allows for the straightforward acquisition of a real-world scene is 360° imaging. Commodity-priced 360° still and video cameras acquire the real world in all directions, quickly, in detail, without prerequisite operator expertise, and without lengthy post-processing. Ever since QuickTime VR circa thirty years ago [9], 360° imaging has been used in a variety of applications such as campuses enticing future students, real-estate agents marketing their listings, and vacationers capturing their experiences. 360 VR, i.e., the immersive viewing of 360° images and videos in a VR headset, brings the natural view selection interface through head motions, which conveys to the user a strong sense of presence in the real-world scene.

The 360 VR ease of acquisition comes at the cost of several user experience limitations. One is the lack of depth perception—the 3D scene is assumed to be infinitely far away. Another is that the user cannot translate, with the user viewpoint confined to the acquisition viewpoint. A third limitation is that the user is floating in space, and is not anchored to the ground of the 3D scene, which is an important shortcoming for most 3D scenes, which the user is supposed to examine from ground level. A fourth limitation is that, since only one user can be at the acquisition point at one time, 360 VR has to approximate collaboration by rendering the collaborator's avatar at a position that appears spatially inconsistent with respect to the real-world scene. These 360 VR limitations are well known and they have received extensive attention from researchers. However, prior approaches for extending 360 VR, such as depth-based or neural rendering methods, come at the cost of increased acquisition, processing, or rendering complexity. Low-cost 360 VR extensions that alleviate some of the 360 VR limitations *without* increasing acquisition cost have received less attention. For this reason, we primarily restrict our comparison to conventional 360 VR to ensure a fair evaluation under matched acquisition constraints.

In this paper we present *TerraFirma*, an approach for extending 360 VR to support viewpoint translation and collaboration *without* increasing the acquisition cost (Fig. 1). *TerraFirma* models the ground under the 360° image acquisition point with a planar patch, on which the user can stand and walk. The planar patch is rendered with projective texture mapping leveraging the 360° image—a user eye ray that intersects the planar patch is looked up in the 360° image based on the direction from the acquisition point to the intersection point. A user ray that misses the patch is looked up directly based on the user ray's direction. In order to provide visual continuity between the patch and the far region, the patch is extended with an intermediate region where the ray lookup direction changes gradually from near to far. *TerraFirma* only requires two trivial parameters: the approximate acquisition height, i.e., how high above the ground the camera was when it acquired the 360° image, and the size of the patch. Therefore, *TerraFirma* improves the immersive exploration of a real-world scene captured with a 360° image, without increasing the cost of the acquisition of the virtual environment.

We evaluated *TerraFirma* in a controlled user study (N = 30) with five user tasks. The results show that *TerraFirma* has an advantage over conventional 360 VR in terms of anchoring the user and their collaborator to the ground of the scene captured by the 360° image, of allowing the user to walk on the ground, and of accurate integration of stationary and dynamic virtual objects into the 3D scene. *TerraFirma* introduces a noticeable but acceptable distortion of the scene and of the trajectory of moving virtual objects. Finally, results also show that *TerraFirma* exhibits good system usability and it does not pose cybersickness concerns. We also refer the reader to the video accompanying our submission.

2 Prior work

Here, we discuss related work on monocular 360° depth estimation, motion parallax and 6DoF viewpoint synthesis, the use of 360° media for collaborative VR applications, and multiperspective images.

Monocular 360 depth estimation. One approach to enabling free-viewpoint navigation from panoramic imagery is to recover depth and approximate scene geometry from a single input. Early methods explored equirectangular and cubemap projections, as well as hybrid fusion strategies [39]. Later efforts introduced cube-based field representations and multi-plane image prediction to improve spatial consistency across panoramic depth maps [7]. More recent systems combine monocular depth estimation with Gaussian splatting and inpainting to generate navigable 3D photos from a single panorama [30]. While these methods demonstrate the feasibility of transforming a single panoramic capture into an immersive 3D scene, they remain computationally expensive and struggle with handling occlusions, depth ambiguity, and boundary artifacts.

Motion parallax and 6DoF viewpoint synthesis. Rather than explicitly estimating depth, an alternative line of research focuses on conveying 3D structure through motion parallax and limited viewpoint translation. Many approaches achieve this by reconstructing simplified scene representations, such as layered depth images, or by directly interpolating between the captured camera poses. For example, motion parallax panoramas can be generated from constrained video trajectories [4], while layered multi-sphere or mesh-based techniques offer more accurate motion parallax and disparity cues for immersive playback [1, 34]. More lightweight pipelines aim to approximate layered representations from casually captured videos [37]. A more closely related line of work bypasses explicit geometry altogether by learning implicit image-based representations. For instance, Parallax360 [24] uses a novel curve-based fitting method for disparity motion fields from a grid of images, enabling real-time viewpoint synthesis and smooth transitions without building a 3D mesh. While effective, this approach still relies on a dense capture protocol (72x8 images), which limits its practicality. Our method proposes a novel, low-cost algorithm for viewpoint translation that aims to mitigate these specific limitations.

360° media for collaborative VR applications. Beyond single-user immersion, panoramic imagery has also been studied in collaborative contexts. Early research into volumetric telepresence established the value of capturing and transmitting live environments for remote collaboration [12, 28]. This has evolved into systems that integrate panoramic video with real-time 3D reconstructions, demonstrating the specific value of combining live 360° environments with neural representations [15]. The potential of this technology is particularly evident in educational settings. Studies have shown the enhanced learning benefits of 360° media [3], which are amplified in multi-user social VR field trips that create shared educational experiences, even with limited interactivity [16].

Multiperspective images. *TerraFirma* essentially attempts to turn a 360° image into a multiperspective image, with a dynamic viewpoint inside the near region, a fixed viewpoint in the far region, and a viewpoint that morphs in between the two in the intermediate region. Throughout the years, multiperspective images have been investigated by artists motivated to break free from the constraints of single perspective (e.g., cubist painters), by computer vision researchers motivated to capture a real-world scene more completely from multiple vanishing points [33], and by scientific and information visualization researchers motivated by conveying a complex dataset more completely in a single image [40]. Near-far partitioning schemes have been used for virtual environment complexity management, where the far region geometry is pre-rendered to a cubemap and rendered as a backdrop to the near region geometry [29]. Our work focuses on the inverse problem, on aiming to recreate a functional 3D virtual environment from a 360° image, without increasing acquisition cost.

3 *TerraFirma* 360 VR Extension

Our goal is to extend 360 VR to support viewpoint translation and collaboration, while preserving the 360 VR fundamental advantages of simple acquisition and output visual quality. We first discuss the design concerns we aim to meet and how we plan to do so (Sec. 3.1), and then describe our *TerraFirma* method in detail (Sec. 3.2).

3.1 Design concerns and approach overview

User grounding. In 360 VR the user view region is reduced to a single 3D point, i.e., the acquisition viewpoint, which is also the center of the 360° panorama. There is simply no space for the user to exist. To

support immersion, presence, and embodiment, a 360 VR application can choose to render the user's avatar, but the avatar will appear to float. Specifically, the ground appears to be infinitely far away, and the avatar feet dangle without making contact with the ground. We address this concern by modeling the ground under the acquisition viewpoint with a planar patch, colored using the 360° image through projective texture mapping. The planar patch is, of course, just a proxy for the geometry of the ground, but, in many cases, this proxy provides an acceptable approximation. The patch allows planting the avatar's feet firmly on the ground.

Viewpoint translation support. 360 VR does not support user viewpoint translation, not even to enforce the interpupilary distance, as needed for stereo viewing. Adding depth perception to 360 VR is not as simple as resorting to two 360° images acquired from adjacent locations, since, as the user pans their head, the interpupilary baseline rotates away from the fixed acquisition baseline. The ground patch described above is 3D geometry, which provides depth cues for the region close to the user. Furthermore, the user can translate their viewpoint freely above the ground patch.

Visual quality. One of the great advantages of 360 VR is that it depicts complex real world scenes with great fidelity and realism. 360 VR delivers on the greatest promise of image-based rendering, that of caching visual quality during acquisition to pass it on to the output image directly, without the challenges and computational expense of complex rendering algorithms that require detailed knowledge of geometry and of surface reflectance properties. 360 VR will support any effect that can be captured by the camera, by simply resampling the 360° image with a homography to compute the output frame. If the 360° image has at least twice the output frame's resolution as required by Nyquist's Law, the visual quality of the output frame is flawless.

Any extension of 360 VR would be well served to preserve this visual quality. We note that even if the RGB acquisition camera is replaced with an RGBD depth camera, the depth channel per pixel is not sufficient to transform the 360° image into a 3D scene that can be rendered to produce high quality frames. Paradoxically, although depth per pixel allows for viewpoint translation, any viewpoint translation reveals scene surfaces that were not visible from the acquisition viewpoint and are therefore missing. These highly objectionable disocclusion artifacts have to be addressed with additional acquisition viewpoints, which not only greatly complicates acquisition, but also brings the challenge of merging overlapping and inconsistent data.

Our approach preserves quality over the ground patch, i.e., through projective texture mapping, and away from the ground patch, i.e., through environment mapping. However, the challenge that has to be solved is that of visual continuity between the patch and its surroundings, where the viewpoint switches abruptly from that of the dynamic user viewpoint, to the fixed acquisition viewpoint.

Seamless integration of virtual objects. 360 VR applications should be able to go beyond allowing the user to contemplate the real-world scene passively. For example, a simulation and training VR application might want to use the 360° image as a realistic backdrop to a virtual workspace with which the trainee interacts. For this, the extension of 360 VR should allow for the seamless integration of stationary and dynamic virtual objects. The ground patch allows placing virtual objects in the 3D space surrounding the user, including in contact with the ground. The virtual objects can also be rendered into the 360° image, from the acquisition viewpoint. Like discussed above in the context of the quality of the visualization of the real world scene, the challenge is to be able to integrate virtual objects that extend–or that move–from the planar patch into the background.

Collaboration support. Effective collaboration in VR requires that the collaborators be able to communicate effectively. For this, they need to see each other or each other's avatars, at a relative position that matches their relative position in the physical world hosting the VR application, such that they can turn to each other intuitively, guided by direction of the sound of their voices, and such that they can move safely within the space. In conventional 360 VR the avatars of the collaborators can be rendered at the correct relative position, but the avatars float in the scene. The collaborators feel like they are *observing* the same real-world scene, for example from adjacent locations in a 360° movie theater, and not that they are *collaborating* in the same scene. The planar patch provides a 3D space where the collaborators are collocated and integrated in the real-world scene, in support of collaboration. Another prerequisite for effective collaboration in VR is the ability to cross-reference elements of the virtual environment.

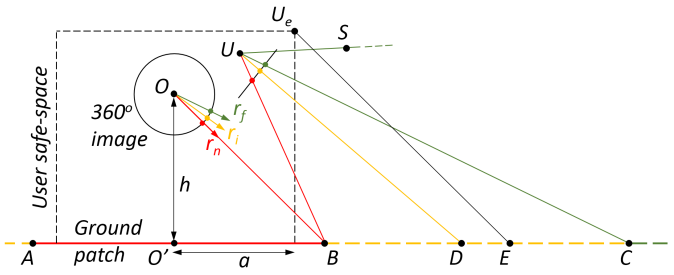


Fig. 2: *TerraFirma* overview. The 2D drawing shows the vertical plane through the acquisition viewpoint O and the user viewpoint U. The ground is modeled with a planar patch AB. The user frame is set from the 360° image through projective texture mapping (e.g., eye ray UB), through environment mapping (e.g., UC and US), and through a morph between projective texture mapping and environment mapping (e.g., UD).

Low acquisition and rendering cost. A primary design concern is that the extension of 360 VR to provide the desired functionality enumerated above should preserve the low acquisition and rendering cost of conventional 360 VR. Acquisition should not go beyond the conventional acquisition of 360° still images or video using conventional panoramic cameras. The conversion of the 360° imagery into eloquent virtual environments that allow the user to experience the captured real world scene should be straightforward. Furthermore, rendering the user output frame should be tractable on the thinnest of VR clients, such as all-in-one VR headsets (e.g., Meta's Quest 3). Modeling the ground with a single planar patch satisfies these requirements. The only few required parameters, such as how high above the ground the camera was when it acquired the 360° imagery and the geometry and size of the patch can be set by any user, without prerequisite VR, graphics, vision, or computer science knowledge. The challenge is to connect the ground patch to the background automatically, i.e., without manual modeling effort, and without increasing computational cost during rendering.

3.2 Method and Implementation

We first give a high level description of our *TerraFirma* method (Sec. 3.2.1) and then describe its implementation (Sec. 3.2.2).

3.2.1 Method

Fig. 2 gives an overview of our method. The ground under the acquisition viewpoint O is modeled with a horizontal planar patch AB. A user frame can have up to three regions, as is the case shown in the figure: a near region, an intermediate region, and a far region. The near region is the part of the frame where the user sees the planar patch. This region is rendered through projective texture mapping. For example, user ray UB is set by looking up the 360° image along ray OB, which has direction r_n. The intermediate region is the part of the frame where the user sees just beyond the planar patch, i.e., where the user sees an extension (here BC) of the ground patch. The role of the intermediate region is to connect the near and the far regions seamlessly, as described shortly. The far region is the rest of the frame, i.e., the part of the frame where the user sees neither the ground patch nor its extension. This region is rendered through environment mapping. For example, user ray UC is set by looking up the 360° image along direction r_f, which is the direction of UC. Similarly, a user ray US that misses the ground plane altogether is looked up along its direction.

Maintaining visual continuity. The intermediate region has to be rendered in a way that connects the near and far regions seamlessly. Consider user ray UD. The ray intersects the ground plane at D, which is in the extension BC of the planar patch AB. Based on the position of

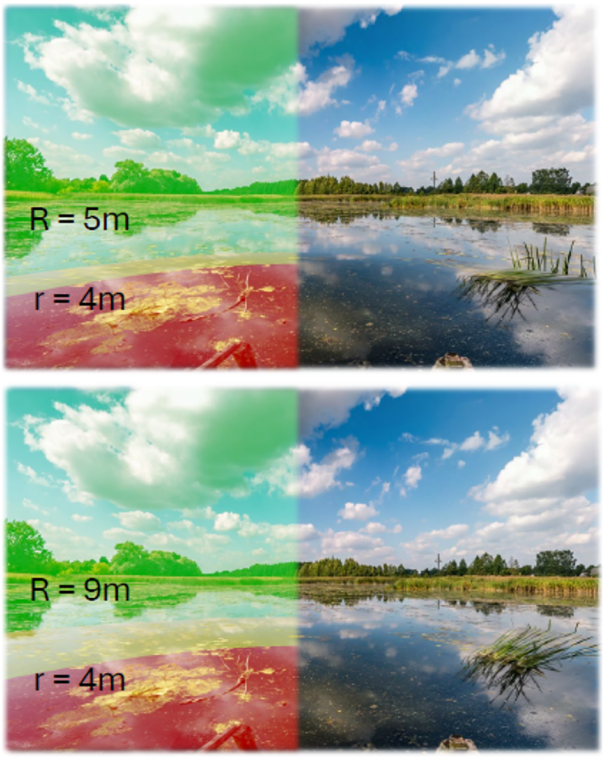


Fig. 3: Output frames for two ground patch extensions. The near region of radius r = 4 m is extended by 1 and by 5 m. The narrow extension distorts the floating grass to the right which crosses from the near to the far region. The larger intermediate region avoids the distortion.

D, rendering has to shift gradually from projective texture mapping to environment mapping. In other words, the rendering viewpoint has to switch gradually from the user viewpoint U to the acquisition viewpoint O. The closer D to B, the more ray UD should be set like a ray that sees the planar patch directly, and the closer to C, the more it should be set like a ray that misses the planar patch and its extension altogether. This is accomplished by setting user ray UD by looking up the 360° image along direction r_i that is a linear blend of the projective texture mapping ray r_p and the environment mapping ray r_e, as shown in Eq. 1.

$$\begin{aligned} r_p &= (D-O)/\|D-O\| \\ r_e &= (D-U)/\|D-U\| \\ w &= \|D-B\|/\|C-B\| \\ r_i &= (1-w)r_p + wr_e \end{aligned} \tag{1}$$

Parameter Selection. For a symmetrical ground patch, e.g., a disk or a square, that is centered under the acquisition point O, the user has to specify only two parameters: the height h of O above the ground, and the maximum distance a the user should be allowed to translate away from the ground plane projection O' of O. h can be measured at acquisition or approximated based on the 360° image, based on familiar objects on or close to the ground. a should be chosen based on the real-world scene, commensurate to the amount of empty space around O'. a and the maximum possible user height define a cylindrical—if the ground patch is a disk—or cubical user safe-space. The actual ground patch radius $O'B$ is set to be larger than a, e.g., to $1.2a$, to not allow the user to get all the way to the boundary of the planar patch.

The size of the ground patch extension, i.e., the length of segment BC in Fig. 2, has to be set based on two competing considerations. One is to avoid that the 360° image fold upon itself in the output frame. In other words, the patch extension should be rendered with its own segment of the 360° image, and not reuse a part of the 360° image that was used to projectively texture map the ground patch. For this,

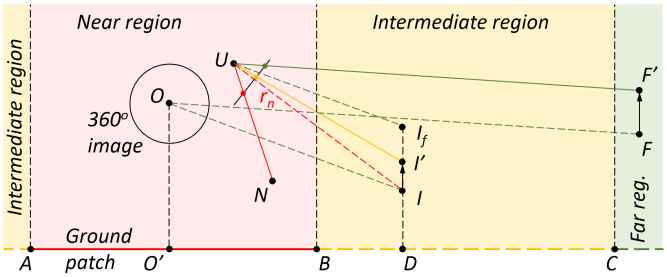


Fig. 4: Vertex projection for virtual object integration. Near-region points (e.g., N) use standard projection; far-region points (e.g., F) are shifted by the environment-map direction; intermediate points (e.g., I) interpolate between the two.

if r_t is the direction of vector $C-U$, the ray (O, r_t) should intersect the ground plane at, or to the right of B. Considering all possible user viewpoints U in the user safe-space, the top-right corner of the user safe-space is the extremal user viewpoint U_e that defines the left-most possible position E of C. If C is chosen at E and the user viewpoint is at U_e, the intermediate region is rendered with none of the pixels of the 360° image. To avoid that the 360° image be stretched over the intermediate region, C should be well to the right of the critical point E. The second consideration is to limit the distortion in the user frame, which is inherent over the intermediate region where the viewpoint switches from the user viewpoint to the acquisition viewpoint. Setting C infinitely far away will assign all of the southern hemisphere of the 360° image to the intermediate region, except for the south pole region used for the ground patch. This provides the largest region over which to dilute the distortion, but it also limits the distortion-free far region to the northern hemisphere of the 360° image. Fig. 3 shows the effects of various ground patch extension sizes, for the same ground patch size.

Virtual object integration. Integrating a virtual object into a real-world scene requires achieving a virtual / real agreement in terms of appearance and in terms of geometry. An important prerequisite for appearance agreement is modeling the dominant light sources in the real-world scene. One approach is to do so interactively, for example, by clicking on the Sun in the equirectangular 360° image in Fig. 1*(f)* to determine the direction of its light.

Achieving geometric agreement requires that the virtual object geometry be projected the same way the real world scene is, with the same approximations, in order to achieve consistency. The ground patch and its extension partition the 3D scene into three regions, as shown in Fig. 4: a near region, above the ground patch, an intermediate region, above the extension of the ground patch, and a far region, everywhere else. A 3D point in the near region, e.g., N in Fig. 4, is projected conventionally. A 3D point in the far region, e.g., F, has to first be shifted to a new position F' such that when projected from the user viewpoint U it lands in the user frame where it would land if it were looked up from the environment map. F is shifted vertically to keep it at the same depth. The amount of the shift is determined such that UF' be parallel to OF.

A 3D point in the intermediate region, e.g., I, has to be shifted an amount in between the zero shift of the near region, i.e., I, and the full shift of the far region, i.e., I_f. I_f is computed similarly to F', such that UI_f be parallel to OI. The amount of the shift is determined by how close point I is to the boundary between the near and intermediate regions, as shown in Eq. 2. Fig. 5 shows a virtual object moving consistently in the *TerraFirma* virtual environment.

$$\begin{aligned} w &= \|D-B\|/\|C-B\| \\ I' &= I + (I_f - I)w \end{aligned} \tag{2}$$

Collaboration support. Two or more collaborators can share the ground patch, and, as long as their tracking data is shared, they can communicate with one another naturally, seeing each other's avatars at

Fig. 5: Moving virtual object, i.e., a robot vacuum cleaner, integrated into the real-world scene. The robot is shown here at several time steps to illustrate its trajectory with a single image.

relative positions congruent with the collaborators' relative positions in the physical world. Effective collaboration also requires that the collaborators cross-reference elements of the virtual environment, say by pointing at them with a virtual laser paradigm [20]. The *TerraFirma* provides a consistent, shared 3D space that allows cross-referencing objects in the near, intermediate, and far regions, as shown in Fig. 6. The laser pointer is drawn as a line segment starting at the user's hand, extending in the direction in which the user is pointing, and having a large length (i.e., 100 m). Both the near and far laser segment endpoints are projected conventionally–the near point endpoint is always in the near region, and the far endpoint is sufficiently far, due to the large laser segment length, for the translational offset between the acquisition viewpoint and the user viewpoint not to matter.

3.2.2 Implementation

We have implemented our *TerraFirma* approach in Unity 3D [27], version 2022.3.48f1, and we have deployed the virtual environment on a Meta Quest 3 [25] all-in-one VR headset.

Virtual environment. The virtual environment has three components: (1) an environment map, (2) the extended ground patch geometry, and (3) the virtual objects.

1. Environment map. The 360° image, typically loaded from an equirectangular image, is used to build an environment map, implemented as a cubemap. The environment map is rendered as background.

2. Extended ground patch. The extended ground patch is modeled with a rectangle. The rectangle is rendered with a fragment shader that colors it according to the method described above and illustrated in Fig. 2. The fragment shader runs on all pixels touched by the rectangle. The fragment shader provides the world position of the 3D intersection point P between the eye ray and the ground plane. When a circular ground patch is desired, it is implemented by discarding fragments whose 3D point is outside the circle. If the length d of the segment $O'P$ is less than that of $O'B$, the fragment is inside the near region and its color is set by looking up the cubemap along the direction of segment OP; otherwise, if d is less than the length of segment $O'C$, the fragment is inside the intermediate region, and its color is set by looking up the cubemap along the morphed direction r_i from Eq. 1;

Algorithm 1 *TerraFirma* fragment shader pseudocode

Require: user pixel q, eye position U, ground plane G, blend α, cubemap sampler $CM(\cdot)$, the cubemap acquisition point O, the projection of the acquisition point on the ground plane O', the boundary point of the near region B, the boundary point of the far region C

$U_q \leftarrow$ view ray from U through pixel q
$(hit, P) \leftarrow \text{INTERSECT}(G, U_q)$ ▷ $P = G \cap U_q$
$d \leftarrow \|\overrightarrow{O'P}\|$ ▷ Euclidean distance from P to O'
if $d < \|\overrightarrow{O'B}\|$ **then** ▷ P_1 case
    $q.\text{color} \leftarrow CM(\overrightarrow{OP})$ ▷ lookup along ray OP
else if $\neg hit$ **or** $d > \|\overrightarrow{O'C}\|$ **then** ▷ P_2 case
    $q.\text{color} \leftarrow CM(\overrightarrow{UP})$ ▷ lookup along ray UP
else ▷ P_3 case
    $r_0 \leftarrow \text{UNITVECTOR}(\overrightarrow{UP})$
    $r_1 \leftarrow \text{UNITVECTOR}(\overrightarrow{OP})$
    $\alpha \leftarrow \|\overrightarrow{O'C}\| - d$
    $w \leftarrow 1 - \frac{\alpha}{\|\overrightarrow{O'C}\| - \|\overrightarrow{O'B}\|}$
    $r_\alpha \leftarrow r_0 \cdot w + r_1 \cdot (1 - w)$
    $q.\text{color} \leftarrow CM(r_\alpha)$ ▷ lookup along ray r_α
end if

otherwise, the fragment is in the far region and its color is set by looking up the cubemap along the direction of segment UP, where U is the user viewpoint. The pseudocode for the fragment shader is shown in Algorithm 1.

3. Virtual objects. Virtual objects are rendered with a vertex shader that implements the vertex shift. First, the shader determines the region to which the vertex belongs. Then near region vertices are projected conventionally. Intermediate and far region vertices are first shifted as described above, and then projected conventionally.

User avatar and collaboration. We utilized the Meta Avatar SDK to provide virtual representations for the users. A set of predefined virtual avatars with hand tracking [8, 19] is used in the application. In addition to the high quality models, an important feature provided by the SDK is the synthesized animations for legs, which helps users feel grounded in the virtual environment. We use Photon Fusion2 to implement the shared and networked *TerraFirma* for multiple users. To accommodate both the first-person perspective and the third-person perspective, we hide the mesh of the head to avoid seeing through the inside of the body for each user's local view, while showing the fully tracked avatars on the remote ends. In addition, to improve the online experience, the feature of lip synchronization of the avatars is applied to enhance the sense of communication. Fig. 7 illustrates the collaboration support provided by *TerraFirma*.

3.3 Analytical evaluation

Distortion. Fig. 8 visualizes the *TerraFirma* distortion in a scene with a checkerboard ground plane, for scenarios corresponding to our indoor and outdoor scenes. Fig. 9 plots the intermediate region distortion for the central ground floor line starting at the user and moving away from

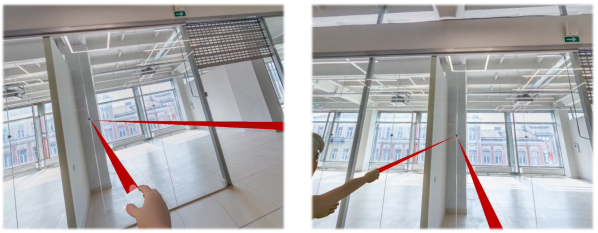


Fig. 6: Consistent virtual laser pointing during collaboration in the *TerraFirma* virtual environment: (left) frame of the first collaborator, (right) frame of the second collaborator.

Fig. 7: Collaboration in the *TerraFirma* virtual environment: photo of the two collaborators in the physical space (a), frame of left (b) and right (c) collaborators.

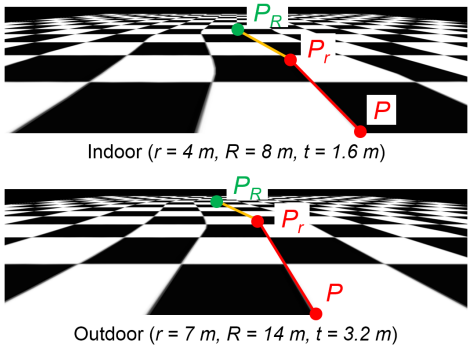


Fig. 8: User frames illustrating the *TerraFirma* distortion over the intermediate region, i.e., from P_r (end of near region and start of intermediate region) to P_R (end of intermediate region and start of far region).

the user, i.e., PP_rP_R in Fig. 8. The distortion is quantified as the angle between the intermediate region direction of the line and the correct, near-region direction. The left side plot shows that the distortion starts at 0 and then increases as the viewpoint translates away from the center of the 360° image. As expected, the distortion is smaller for the outdoor scene where the intermediate region is both larger as well as farther from the user. The right side plots show the distortion as a function of the distance to the user, for points A and B on the left side plots. The distortion is nearly linear, i.e., scene lines project to nearly straight lines over the intermediate region. In conclusion, *TerraFirma* connects the near and far regions with approximately piece-wise linear continuity.

Frame rate. We leveraged the OVR Metrics Tool bundled with the Meta Quest Developer Hub to monitor the frame rate for our Quest 3 headset in standalone mode (i.e., w/o connection to a workstation). We recorded sessions in both traditional 360 VR and *TerraFirma*. The average frame rate was 72.03 Hz for 360 VR and 71.97 Hz for *TerraFirma*. This shows that *TerraFirma* does not bring a significant computational cost and that the VR application can still run at the 72 Hz refresh rate of our headset.

Comparison to light-weight depth-based approaches. The goal of *TerraFirma* is to support user viewpoint translation without increasing the acquisition cost. User viewpoint translation can be supported if the 2D environment captured and modeled with a 360° image is upgraded to a 3D environment. Emerging methods, such as DA^2 [23], promise to enhance a 360° image with per pixel depth automatically, without complicating acquisition. DA^2 did not produce a usable depth map for our outdoor scene Fig. 1f. When depth extraction succeeds, the 360° image pixels can be unprojected to 3D points and reprojected to novel user frames, as shown in Fig. 10, without the intermediate region distortions of *TerraFirma*. However, depth based approaches suffer from disocclusion error artifacts (see purple shadow of column in Fig. 10), from distortions where depth is inaccurate (see ceiling beam), and from a lower-quality point-based reconstruction of the output frame.

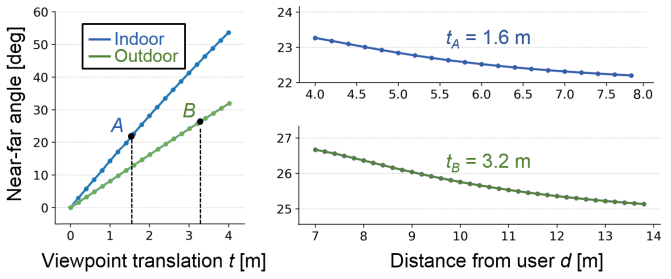


Fig. 9: Average intermediate region distortion as a function of viewpoint translation (left), and as a function of the distance from the user (right). The distortion is quantified as the angle between segment PP_r and P_rP_R

.

Fig. 10: User frame from a depth-based approach [23] for the indoor environment shown in Fig. 1a.

.

Furthermore, depth extraction results in hundreds of thousands of points which consume much of the rendering budget of thin VR clients such as all-in-one headsets, precluding using the real world environment as a realistic but inexpensive backdrop for mixed reality applications whose narratives focus on the interaction with complex synthetic objects.

4 User Study

We conducted a user study (N = 30) with the approval of Purdue University's Institutional Review Board (protocol number IRB 2025-572). Participants completed a series of five tasks during which they were exposed to the features of *TerraFirma*.

Research hypotheses. We formulate the following research hypotheses based on our *TerraFirma* design aims and based on the known limitations of conventional 360 VR.

RH1: *TerraFirma* allows giving the user a stronger sense of standing and walking on the ground of the scene captured by the 360^o image, compared to conventional 360 VR.

RH2: *TerraFirma* introduces a noticeable distortion of the scene captured by the 360^o image.

RH3: *TerraFirma* allows placing stationary and moving virtual objects more convincingly on the ground of the scene captured by the 360^o image, compared to conventional 360 VR.

RH4: *TerraFirma* introduces a noticeable distortion of the trajectory of moving virtual objects.

RH5: *TerraFirma* supports collaboration by giving the user a stronger sense that their collaborator stands on the ground of the scene captured by the 360^o image, and by allowing the user to follow their collaborator's pointing direction more easily, compared to conventional 360 VR.

RH6: *TerraFirma* exhibits good system usability.

RH7: *TerraFirma* provides a greater sense of presence in the virtual environment, compared to conventional 360 VR.

RH8: *TerraFirma* does not pose cybersickness concerns.

4.1 Methods

Participants. We recruited N = 30 students from our campus. 73% were undergraduate, and 27% were graduate students with an average age of 21.9 years. 53% were male, 40% were female, and 7% selected "other" or preferred not to answer. Regarding familiarity with virtual reality (VR) headset technology, 10% of participants indicated that they had used a VR headset once, 30% fewer than five times, 33% more than five times, and 27% frequently.

Conditions. One controlled study had two conditions: an experimental condition (EC), in which participants used our *TerraFirma* extension of 360 VR, and a control condition (CC), in which participants used conventional 360 VR.

Study design. We used a within-subjects design where each participant experienced both the control and the experimental conditions.

Tasks. Participants were asked to perform five tasks, designed to investigate the potential strengths of *TerraFirma* compared to conventional 360 VR. *TASK 0* prompted participants to look and walk around the scene for 20 seconds and familiarize themselves with the scene and their VR avatar. *TASK 1* showed four spheres of different colors around the participant forming a square with a side of two meters. The participant was prompted to walk to each sphere in a given order, thus encouraging them to move around the scene and to observe the scene as they moved. *TASK 2* directed participants to point with a virtual laser at an object, i.e., a robotic vacuum cleaner or a wicker basket, positioned in the scene, then again after walking towards the object. *TASK 3* had participants view a dynamic object moving through the scene in a straight line, crossing between the near, intermediate, and far regions. Finally, *TASK* 4 had the participant join a virtual scene together with a collaborator, impersonated by a researcher, where the participant had to follow the pointing direction of the researcher. Each task was performed in two real-world scenes captured by a 360^o image: an outdoor and an indoor scene (Fig. 1 *(b)* and *(c)*).

Data collection. We investigated our research hypotheses through standard and custom user experience questionnaires.

Custom user experience questionnaires. After each task, participants answered several questions in the VR environment [10, 31]. The answers were provided on a five-point Likert scale, i.e., "strongly disagree", "disagree", "neither agree nor disagree", "agree", "strongly agree", scored with integers from 1 to 5. Some questions are phrased negatively to avoid mechanical answers. The scores of the negatively phrased questions are flipped, i.e., a score of a becomes $6-a$, for a higher score to always be desirable. We indicate below the negatively phrased questions by appending the word (negative) to the question, which was not seen by the participants.

The questions for TASK 0 are: **Q01:** I felt like I was standing on the ground. **Q02:** I felt like I was floating (negative). **Q03:** I felt like I was in a movie theater with a wraparound screen. **Q04:** I felt like I was in a 3D scene. The questions for TASK 1 are: **Q11:** I felt like I was actually walking in the scene captured by the 360° image. **Q12:** Although I was moving from one sphere to the next, I felt like I was not really moving in the scene captured by the 360° image (negative). **Q13:** As I was walking, I noticed that the scene captured by the 360° image was distorted (negative). **Q14:** I found the distortion of the scene captured by the 360° image highly objectionable (negative).

The questions for TASK 2 are: **Q21 (Indoor):** As I walked, the robot stayed in the same place on the floor. **Q21 (Outdoor):** As I walked, the basket stayed in the same place on the dock. The questions for TASK 3 are: **Q31:** The object appeared to move on the ground (negative). **Q32:** The object appeared to move on a straight line. The questions for TASK 4 are: **Q41:** I felt like my collaborator and I were standing on the ground. **Q42:** I felt like my collaborator and I were floating (negative). **Q43:** I felt like it was easy to follow the pointing direction of my collaborator.

The questions were selected to gauge the participants' subjective perception of strengths and weakness of *TerraFirma* and of conventional 360 VR. Specifically, the questions do not avoid any potential shortcomings of either method, for example drawing the participants' attention and requiring their input regarding the distortion of the scene visible in the intermediate region, and the perceived grounded presence of both the avatar and the objects within the scenes.

Standard questionnaires. After each condition, participants also completed the following questionnaires: the System Usability Scale (SUS [6]), Igroup's Presence Questionnaire (IPQ [32]), and the Simulator Sickness Questionnaire (SSQ [17]). In addition to these questionnaires, participants were also invited to share their opinion on the two conditions through a free-form, open-ended answer.

Data analysis. We analyze the data through descriptive statistics, provided through boxplots and tables, and through inferential statistics, leveraging standard statistical tests. We assess data normality with the Shapiro-Wilk test [35]. Virtually none of the data was normally distributed, so we analyze all data with the non-parametric Wilcoxon's signed rank test [41]. The test results are given through the Wilcoxon's z and through the significance value p. Effect sizes are estimated using Wilcoxon's r, which are then used to estimate the statistical power $1-\beta$ of the tests given our N = 30 participants. We chose to enroll N = 30 participants in our study so we can detect large effects with good statistical power, i.e., above 0.80.

The standard questionnaire answers are analyzed with the customary procedures prescribed by their authors and by followup research [2, 5, 6, 17, 22, 32, 38].

Procedure. We scheduled individual sessions with each user in a large research space with a cleared floor area 3 meters by 3 meters wide to accommodate uninterrupted movement during the study. Participants underwent a verbal pre-screening, filled out the consent form and demographics questionnaire, completed the set of five tasks for one condition for both the outdoor and the indoor scene, removed the headset and filled out questionnaires administered electronically through the researcher's laptop, completed the set of five tasks for the second condition for both scenes, and finally removed the headset again and filled out the questionnaires a second time. The total participant involvement took sixty minutes, and participants were compensated with a gift card of a value equivalent to 30 USD.

4.2 Results and discussion

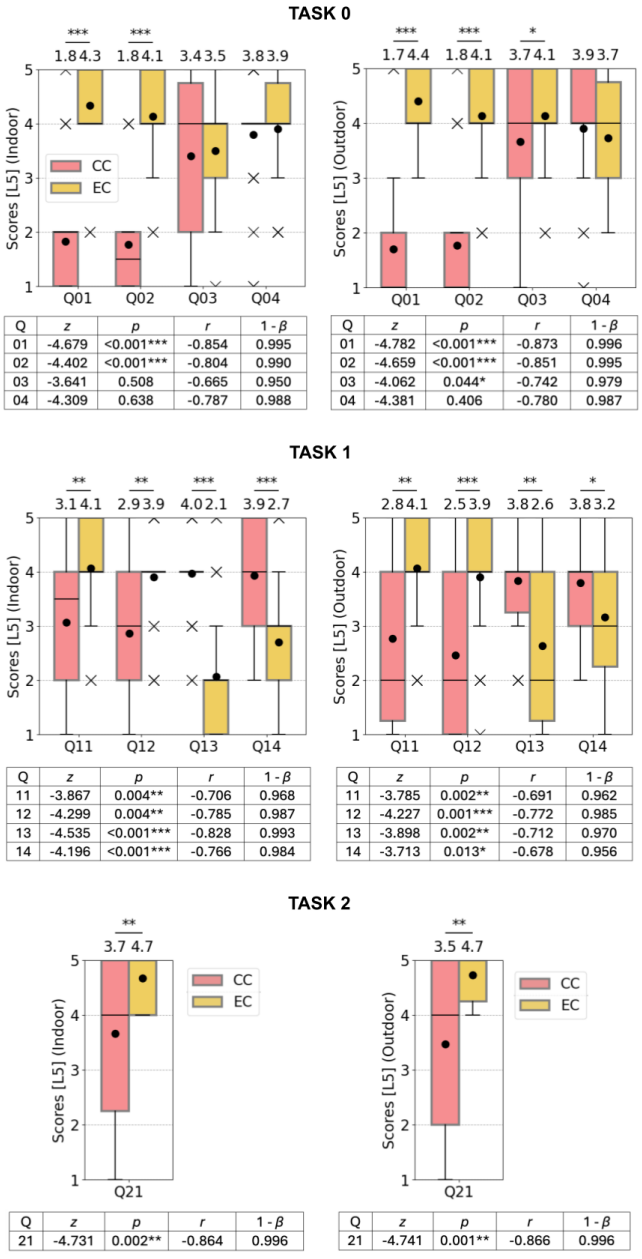
Custom user preference questionnaire. The descriptive and inferential statistics of the answers to the custom user preference questionnaire are given in Fig. 11 for the first three tasks, and in Fig. 12 for the last two tasks.

TASK 0. EC has a significant advantage over CC for Q01 (standing on the ground) and for Q02 ([not] floating), for both the Indoor and Outdoor scenes. The ground patch anchors the user to the ground effectively. EC has an advantage over CC for Q03 (in a 360 movie theater), but only for Outdoor. Both approaches score highly on this metric, which fails to differentiate them. Both approaches score highly on Q04 (in a 3D scene), which suggests that the addition of the ground patch did not enhance the 3D effect of the scene in the EC compared to the CC. Alternatively, it could suggest that the question was too ambiguous and the meaning of "in a 3D scene" was not clear, as both conditions show a 3D scene, albeit a 2D panorama of a 3D scene. The ability to translate the viewpoint or not did not make participants judge the scene as more or less 3D. The only part of the scene that provided any disparity between the user's left and right eye images was the planar patch, which is not enough to label conventional 360 VR as 2D.

TASK 1. EC has a significant advantage over CC for Q11 (felt like actually walking) and for Q12 (actually moving when walking from sphere to sphere). This confirms that for EC, participants perceived "traction" on the ground patch, changing their position in the scene, compared to "pretend walking" in CC. Q13 and Q14 ask about the [lack of] scene distortion and its acceptability. As expected, participants did not find any distortion for CC. For EC, participants noticed the distortion (CC median answer 2.0 for Q13 for both scenes) and were neutral regarding its objectionability (EC median answer 3.0 for Q14 for both scenes). These results indicate that the distortion exceeds detectability but not acceptability thresholds.

TASK 2. EC has a significant advantage over CC for the single question of whether the virtual object stayed in the same place in the scene as the participant walked, i.e., robot on the floor for Indoor and basket on the dock for Outdoor. The mean for EC is 4.7, which leaves no doubt that participants felt the virtual object was anchored to the real-world scene.

TASK 3. EC has a significant advantage over CC for Q31 (object moved on the ground) for the vacuum cleaning robot in the Indoor scene. Due to a data collection technical error we failed to record the answers for Q31 in the Outdoor scene. EC has a significant disadvantage over CC for Q32 (object moved on a straight line), for both scenes. For CC, the object moves in a straight line "through the air", without connection to the ground. For EC, the object always moves on the ground, on a straight line in the far region and then on a straight line in the near region. The two linear trajectories are slightly misaligned, and they are connected with C^0 continuity over the intermediate region. The



Q	z	p	r	1 - β
01	-4.679	<0.001***	-0.854	0.995
02	-4.402	<0.001***	-0.804	0.990
03	-3.641	0.508	-0.665	0.950
04	-4.309	0.638	-0.787	0.988

Q	z	p	r	1 - β
01	-4.782	<0.001***	-0.873	0.996
02	-4.659	<0.001***	-0.851	0.995
03	-4.062	0.044*	-0.742	0.979
04	-4.381	0.406	-0.780	0.987

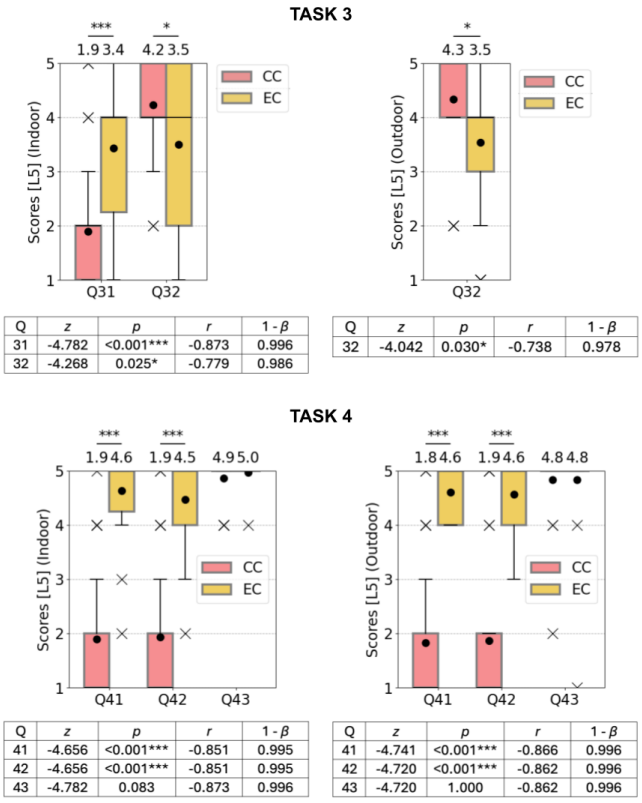
Q	z	p	r	1 - β
11	-3.867	0.004**	-0.706	0.968
12	-4.299	0.004**	-0.785	0.987
13	-4.535	<0.001***	-0.828	0.993
14	-4.196	<0.001***	-0.766	0.984

Q	z	p	r	1 - β
11	-3.785	0.002**	-0.691	0.962
12	-4.227	0.001***	-0.772	0.985
13	-3.898	0.002**	-0.712	0.970
14	-3.713	0.013*	-0.678	0.956

Q	z	p	r	1 - β
21	-4.731	0.002**	-0.864	0.996

Q	z	p	r	1 - β
21	-4.741	0.001**	-0.866	0.996

Fig. 11: Descriptive (graph) and inferential (table) statistics of the answers to the custom user preference questionnaire for the first three tasks and the two scenes (Indoor left, Outdoor right), over the two conditions.



Q	z	p	r	1 - β
31	-4.782	<0.001***	-0.873	0.996
32	-4.268	0.025*	-0.779	0.986

Q	z	p	r	1 - β
32	-4.042	0.030*	-0.738	0.978

Q	z	p	r	1 - β
41	-4.656	<0.001***	-0.851	0.995
42	-4.656	<0.001***	-0.851	0.995
43	-4.782	0.083	-0.873	0.996

Q	z	p	r	1 - β
41	-4.741	<0.001***	-0.866	0.996
42	-4.720	<0.001***	-0.862	0.996
43	-4.720	1.000	-0.862	0.996

Fig. 12: Descriptive (graph) and inferential (table) statistics of the answers to the custom user preference questionnaire for the last two tasks and the two scenes (Indoor left, Outdoor right), over the two conditions.

non-linearity of the trajectory was not judged as severe, with an average score of 3.5 and a median of 4.0.

TASK 4. In the collaborative scenario, EC had a significant advantage over CC for Q41 (participant and collaborator standing on the ground), and for Q42 (participant and collaborator not floating), for both scenes. Both conditions scored very high for Q43 (easy to follow pointing direction), i.e., above 4.8, in both scenes. Indeed, telling where the collaborator points is not hard in CC, what would have been hard is to reference a virtual object based on a real-world element in its proximity. Each of the two collaborators would see a different real-world element in the proximity of the same virtual object, making cross-referencing the virtual object challenging.

For all the comparisons, the effect sizes are large, i.e., $r > 0.5$, and our N = 30 participants are sufficient for ample statistical power, i.e., $1 - \beta > 0.97$ in all cases.

Standard questionnaires. The standard questionnaire results are shown in Fig. 13.

SUS. Both conditions have high SUS scores: 79.3 for CC, which corresponds to a letter grade of A- and an adjective rating of "Excellent", and 80.8 for EC, for a letter grade of A and an adjective rating of "Excellent". The difference between CC and EC is not significant. Conventional 360 VR is an easy to use immersive visualization of a real-world 3D scene. The viewpoint translation and user grounding provided by EC over CC only translate in a modest increase of the SUS score for our task.

IPQ. EC has slightly higher presence scores on all three sub-scales, but the differences are not significant. The biggest contributor to the sense of presence is the 360° image, which both approaches have, and which dwarfs the contribution of the translation and grounding that only EC has. Larger studies are needed to investigate this smaller effect.

SSQ. The SSQ scores are low for both approaches, as for VR total scores below 20 confirm that there are no cybersickness concerns [5]. Tasks that would force participants to walk extensively in a virtual world that floats above the real world, and to engage cognitively with the spatial relationship between the real and virtual worlds, might give our method the opportunity to avoid concerning levels of cybersickness in conventional 360 VR.

Participant free-form feedback. The majority of participants (60%) preferred our method due to a feeling of standing on the ground in the virtual scene compared to a strong sense of floating in the control condition. 39% of these participants were particularly frightened by the feeling of floating, displaying significantly more hesitation in their movements within the virtual environment and expressing unease. For this reason, the feeling of being grounded dramatically increased their comfort within the environment, as they felt it was safer to move around. Two users also stated that the lack of viewpoint translation in the control condition made them dizzy. Some users (20%), on the other hand, stated that they didn't mind the perceived feeling of floating, providing reasons such as it allowed for a better view of the surrounding

scene, or it simply didn't affect their experience.

One major drawback to our method is the distortion visible in the scene. 33% of participants found the distortion extremely noticeable and a major hindrance to fully enjoying the VR experience. Particularly, they felt it disrupted the realism of the environment. However, users were split between whether they preferred the indoor or outdoor scene in the experimental condition due to the distortion. For those who expressed a preference for the indoor scene, they felt the distortion of natural elements, e.g., underwater grass, significantly ruined the sense of realism. For those who preferred the outdoor scene, they felt the visible warping of man-made elements, e.g., the edges of indoor floor tiles, was more unrealistic. Conversely, a few users (10%) felt that despite the distortion, the stronger sense of immersion in our method made the scene feel more realistic.

In terms of collaboration, nearly all users expressed a preference for our method for its ability to have multiple avatars stand on the same planar patch. Many stated it felt more like they were truly interacting with the researcher and experiencing the scene together.

4.3 Summary of findings

Our results lend support to our seven research hypotheses as follows. The results for Q01 and Q02 (standing on ground and [not] floating), and Q11 and Q12 (actually walking in the scene) provide support for **RH1**. The results for Q13 (noticed distortion) provide support for **RH2**. Furthermore, the results for Q14 show that participants are neutral regarding the objectionability of the distortion. The results for Q21 and Q31 show that when the participant walked, virtual objects on the ground appeared to stay in one place, and that when the object moved, it appeared to do so while remaining on the ground, supporting **RH3**. Q32 shows that participants noticed the non-linear trajectory of moving objects, supporting **RH4**. Q41 and Q42 support the first part of the collaboration research hypothesis **RH5**, i.e., the participant perceived their collaborator as better grounded in EC than in CC, but Q43 does not support the second part of **RH5**, i.e., following the collaborator's pointing direction was just as easy in CC as in EC. We conclude that the results only bring partial support for **RH5**. SUS and SSQ results support **RH6** and **RH8**. However, the IPQ results do not support **RH7**– *TerraFirma* does not surpass conventional 360 VR in terms of sense of presence. Although the means are higher for each of the three IPQ subscales, the differences are not significant, pointing to the need for larger studies, or for tasks more focused on the extensions brought by *TerraFirma*, in order to discern between the two approaches.

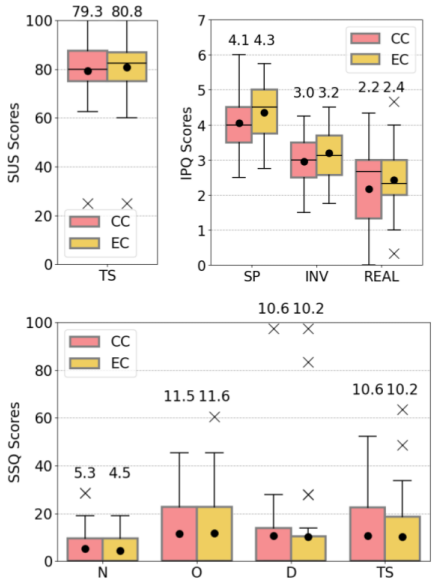


Fig. 13: Subjective feedback to the standard questionnaires.

5 Conclusions. Limitations. Future Work

We have presented *TerraFirma*, an approach for extending the single viewpoint of conventional 360 VR to a view region, which supports viewpoint translation, grounding the user and their collaborator, as well as integrating stationary and dynamic virtual objects consistently in the scene captured by the 360° image. A user study with N = 30 participants confirms these advantages over conventional 360 VR. Our method preserves 360 VR's essential advantage of acquisition and rendering simplicity.

One limitation of *TerraFirma* is the distortion inherent to switching viewpoints from the dynamic user viewpoint to the fixed acquisition viewpoint. Our user study confirms that the distortion is noticeable. We rely on a linear blend between the two viewpoints, and future work could examine more effective ways of diluting the viewpoint change. Our implementation is by-and-large scene independent, but future work could devise more complex ground patch perimeters that route the near to intermediate region line through the scene regions that hide the viewpoint transition more effectively.

The fixed viewpoint of the far region makes it devoid of motion parallax as the user viewpoint translates. Thus the intermediate region act like a rubber band connecting the moving near region to the fixed far region. The artifact is more visible in the context of non-immersive visualization where the view is more likely to change with a pure translation, e.g., under the control of a traditional keyboard or mouse interface. In VR, the user view changes are never pure translations, and any residual rotation will reduce the saliency of the lack of parallax in the far region. Nonetheless, future work should examine injecting artificial rotational changes in the far region, against the viewpoint translation direction, to simulate the missing motion parallax.

Another limitation of our work is that it focuses exclusively on the *geometric* integration of static and dynamic virtual objects into the real world scene, and ignores *appearance* integration, which makes the virtual objects somewhat dissonant with their real world surroundings. Therefore, another direction of future work is to improve the *appearance* integration of virtual objects. One of the challenges that has to be overcome is porting lighting and shading computation to our context of a viewpoint that is fluid over the intermediate region.

More comprehensive evaluation for *TerraFirma* is needed as well. Our study relied on subjective user assessments and does not include objective performance metrics or quantitative measures of geometric or visual error introduced by *TerraFirma*. While the user study provides strong evidence that participants perceive improvements in grounding, viewpoint translation, and object integration compared to conventional 360 VR, future work should introduce controlled tasks such as distance estimation, object alignment, or spatial judgment tasks to quantify the impact of *TerraFirma* on spatial understanding and perception. A revised questionnaire with proper 'catch' or 'foil' questions should also be designed to avoid meta-level response bias.

The current *TerraFirma* implementation can handle environments with uneven terrain, by approximating them with a planar patch (e.g., Fig. 1f). Future work could examine extending *TerraFirma* to model uneven terrain with higher fidelity, generalizing the patch geometry. *Terrafirma* can also be extended to handle 360° videos for dynamic scenes. From an implementation standpoint, *TerraFirma* can readily handle videos, by simply using the current 360° video frame as opposed to the same 360° still image. We speculate that the results might be acceptable as long as the dynamic real-world objects stay in the far region. The future work challenge is to provide quality output frames when dynamic real-world objects get close to the acquisition viewpoint, entering the near region. A promising approach is to model such dynamic objects with vertical video sprites, and to inpaint the gaps left in the static background using earlier frames.

Acknowledgments

We thank Javier Hurtado and the members of the Purdue Computer Science Extended Reality Laboratory (XR Lab) for their help. This material is based upon work supported by the United States National Science Foundation under Awards No. 2212200, 2219842, 2318657, 2309564, 2506783, and 2417510.

REFERENCES

[1] B. Attal, S. Ling, A. Gokaslan, C. Richardt, and J. Tompkin. Matryodshka: Real-time 6dof video view synthesis using multi-sphere images, 2020. 2

[2] A. Bangor, P. Kortum, and J. Miller. Determining what individual sus scores mean: adding an adjective rating scale. *J. Usability Studies*, 4(3):114–123, May 2009. 7

[3] K. Batra, Z. Zhang, S. Yang, A. Agrawal, Y. Gu, B. Benes, A. Magana, and V. Popescu. Xrxl: A system for immersive visualization in large lectures. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 370–380, 2025. doi: 10.1109/VR59515.2025.00061 2

[4] T. Bertel, N. D. F. Campbell, and C. Richardt. Megaparallax: Casual 360° panoramas with motion parallax. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1828–1835, 2019. doi: 10.1109/TVCG.2019.2898799 2

[5] P. Bimberg, T. Weissker, and A. Kulik. On the usage of the simulator sickness questionnaire for virtual reality research. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 464–467, 2020. doi: 10.1109/VRW50115.2020.00098 7, 8

[6] J. Brooke. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996. 7

[7] W. Chang, H. Ai, T. Zhang, and L. Wang. Cube360: Learning cubic field representation for monocular 360 depth estimation for virtual reality, 2024. 2

[8] Y. Che and Y. Qi. Detection-guided 3d hand tracking for mobile ar applications. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 386–392. IEEE, 2021. 5

[9] S. E. Chen. Quicktime vr: An image-based approach to virtual environment navigation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pp. 29–38, 1995. 2

[10] M. Feick, N. Kleer, A. Tang, and A. Krüger. The virtual reality questionnaire toolkit. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp. 68–69, 2020. 7

[11] J. Geng. Structured-light 3d surface imaging: a tutorial. *Advances in optics and photonics*, 3(2):128–160, 2011. 1

[12] D. Gilbert, A. Bose, T. W. Kuhlen, and T. Weissker. Pascal - a collaboration technique between non-collocated avatars in large collaborative virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 31(5):3525–3535, 2025. doi: 10.1109/TVCG.2025.3549175 2

[13] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 453–464. 2023. 1

[14] R. Horaud, M. Hansard, G. Evangelidis, and C. Ménier. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine vision and applications*, 27(7):1005–1020, 2016. 1

[15] X. Huang, M. Yin, Z. Xia, and R. Xiao. Virtualnexus: Enhancing 360-degree video ar/vr collaboration with environment cutouts and virtual replicas. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3654777.3676377 2

[16] S. R. Kalvakolu, H. Söbke, J. Baalsrud Hauge, and E. Kraft. *Combining 360° Spaces and Social VR*, p. 375–380. Springer Nature Switzerland, Dec. 2024. doi: 10.1007/978-3-031-78269-5_38 2

[17] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993. 7

[18] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[19] C. Khundam, V. Vorachart, P. Preeyawongsakul, W. Hosap, and F. Noël. A comparative study of interaction time and usability of using controllers and hand tracking in virtual reality training. In *Informatics*, vol. 8, p. 60. MDPI, 2021. 5

[20] S. Lee, J. Seo, G. J. Kim, and C.-M. Park. Evaluation of pointing techniques for ray casting selection in virtual environments. In *Third international conference on virtual reality and its application in industry*, vol. 4756, pp. 38–44. SPIE, 2003. 5

[21] M. Levoy and P. Hanrahan. Light field rendering. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 441–452. 2023. 1

[22] J. R. Lewis and J. Sauro. Item benchmarks for the system usability scale. *Journal of User Experience*, 13:158–167, 2018. 7

[23] H. Li, W. Zheng, J. He, Y. Liu, X. Lin, X. Yang, Y.-C. Chen, and C. Guo. Da ^{2}: Depth anything in any direction. *arXiv preprint arXiv:2509.26618*, 2025. 6

[24] B. Luo, F. Xu, C. Richardt, and J.-H. Yong. Parallax360: Stereoscopic 360° scene representation for head-motion parallax. *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1545–1553, 2018. doi: 10.1109/TVCG.2018.2794071 2

[25] Meta. Quest 3 Mixed Reality Headset. `https://www.meta.com/quest/quest-3/`. Accessed: 2024-10-02. 5

[26] 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. 1

[27] V. T. Nguyen and T. Dang. Setting up virtual reality and augmented reality learning environment in unity. In *2017 IEEE International symposium on mixed and augmented reality (ISMAR-Adjunct)*, pp. 315–320. IEEE, 2017. 5

[28] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, V. Tankovich, C. Loop, Q. Cai, P. A. Chou, S. Mennicken, J. Valentin, V. Pradeep, S. Wang, S. B. Kang, P. Kohli, Y. Lutchyn, C. Keskin, and S. Izadi. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, p. 741–754. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2984511.2984517 2

[29] V. Popescu, S. H. Lee, A. S. Choi, and S. Fahmy. Complex virtual environments on thin vr systems through continuous near-far partitioning. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 35–43, 2022. doi: 10.1109/ISMAR55827.2022.00017 2

[30] M. Rey-Area and C. Richardt. 360° 3D Photos from a Single 360° Input Image . *IEEE Transactions on Visualization & Computer Graphics*, 31(05):2426–2434, May 2025. doi: 10.1109/TVCG.2025.3549538 2

[31] S. Safikhani, M. Holly, A. Kainz, and J. Pirker. The influence of in-vr questionnaire design on the user experience. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, VRST '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3489849.3489884 7

[32] T. Schubert, F. Friedmann, and H. Regenbrecht. The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments*, 10(3):266–281, 2001. 7

[33] S. M. Seitz and J. Kim. Multiperspective imaging. *IEEE Computer Graphics and Applications*, 23(6):16–19, 2003. 2

[34] A. Serrano, I. Kim, Z. Chen, S. DiVerdi, D. Gutierrez, A. Hertzmann, and B. Masia. Motion parallax for 360° rgbd video. *IEEE Transactions on Visualization and Computer Graphics*, 2019. 2

[35] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965. 7

[36] C. C. Slama, C. Theurer, and S. W. Henriksen. *Manual of photogrammetry*. Number Ed. 4. 1980. 1

[37] H. Sun and S. Zollmann. Towards casually captured 6dof vr videos. In *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 176–179, 2022. 2

[38] T. Q. Tran, T. Langlotz, J. Young, T. W. Schubert, and H. Regenbrecht. Classifying presence scores: Insights and analysis from two decades of the igroup presence questionnaire (ipq). *ACM Trans. Comput.-Hum. Interact.*, 31(5), Nov. 2024. doi: 10.1145/3689046 7

[39] 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 (CVPR)*, June 2020. 2

[40] Z. Wen, W. Zeng, L. Weng, Y. Liu, M. Xu, and W. Chen. Effects of view layout on situated analytics for multiple-view representations in immersive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):440–450, 2022. 2

[41] F. Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*, pp. 196–202. Springer, 1992. 7