Visual Perception of Real World Depth Map Resolution for Mixed Reality Rendering

Lohit Petikam Andrew Chalmers Taeyhun Rhee*

Computational Media Innovation Centre, Victoria University of Wellington, NZ
DreamFlux, NZ

(a) Reference (b) 4096×2048 (c) 426×213 (optimal) (d) 64×32 (visible artifacts)

Figure 1: Examples of compositing foreground objects (statue, teapot and armadillo) into the background (*Bedroom* scene) using its depth map. Comparison between ground truth reference rendering with high resolution 3D meshes (a), composition with high resolution depth map (4096×2048) (b), composition with perceptually optimal depth resolution (426×213) (c), and composition with insufficient depth resolution (64×32) (d) showing noticeable artifacts from poorly reconstructed geometry.

Abstract

Compositing virtual objects into photographs with known real world geometry is a common task in mixed reality (MR) applications. This geometry enables rendering of global illumination effects, such as mutual lighting, shadowing, and occlusions between the background photograph and virtual objects. Obtaining high fidelity geometric representations of the real world can be a costly procedure, and is often approximated with depth data. However, it is not clear how much fidelity the depth data should have in order to maintain high visual quality in MR rendering.

In this paper, we investigate the relationship between real world depth fidelity and visual quality in MR rendering. We do this by conducting a series of user experiments that measure how seamlessly virtual objects are blended with the background under varying depth resolutions. We independently evaluate the noticeability of multiple composition artifacts that occur with approximate depth. Perceptual thresholds in depth resolution are then obtained for each artifact. The findings can be used to inform trade-off decisions for optimising depth acquisition pipelines in MR applications.

Index Terms: Computing methodologies—Computer graphics—Graphics systems and interfaces—Mixed / augmented reality; Computing methodologies—Computer graphics—Graphics systems and interfaces—Perception

1 Introduction

The acquisition of depth information from the real world improves various aspects of mixed reality (MR), such as accurately placing virtual objects on real world surfaces, and supporting *mutual* global illumination effects. Mutual in this context refers to light transport interactions between the real world and virtual objects. Such interactions include the real world environment impacting the lighting, shadows and occlusions on the virtual objects, and vice versa. Such illumination effects improve the overall realism of MR rendering, where virtual objects can seem like they naturally belong in the real world environment. However, a reduction in depth fidelity directly impacts the quality of mutual illumination, where artifacts arise if the depth data is not a faithful reconstruction of the real world.

Obtaining high fidelity depth information from the real world can be a difficult task, involving expensive or time consuming methods such as LIDAR scanning, photogrammetry, or manual reconstruction by an artist. Alternatively, real time or consumer level depth cameras such as Microsoft Kinect can be used to reconstruct the depth information in real-time with the concession of lower fidelity. As such, there is a clear trade-off between fidelity in depth data and depth acquisition, which impacts the quality of mutual illumination. However, it remains unclear the degree to which depth fidelity impacts the quality of the rendered image as observed by the end-user.

Recent work has defined the key notion of a *seamless composition*, in which virtual objects are believably composited into the real world despite using reduced resources [3]. In order to understand how much real world depth information is required, we use the notion of a seamless composition with respect to depth fidelity to evaluate various mutual illumination effects. With each of these effects, different artifacts arise, such as protruding geometry or incorrect shadow alignment. A challenge to be addressed is that these artifacts may only violate seamlessness under different depth fidelities. Therefore, it is important to isolate each artifact and evaluate them independently.

To our knowledge, no prior work has addressed this specific problem. Closely related previous approaches have addressed visual quality of virtual geometry, such as level of detail (LOD), but do not evaluate it in the context of a seamless composition in MR rendering. Other approaches evaluate composition quality, but do not take into account depth fidelity in relation to composition directly. In this paper, we conduct a user study that evaluates depth fidelity with respect to mutual illumination. Adhering to the notion of a seamless composition, we find perceptual thresholds of depth resolution in a MR context. Through this process, we mitigate common artifacts by choosing appropriate levels of depth fidelity.

The main contributions of our paper are summarised as follows:

- We design and conduct a novel self-referencing user study to evaluate the depth fidelity required for MR rendering including mutual illumination.
- We investigate perceptual thresholds for MR rendering by defining four composition artifact types based on depth, and

*e-mail:taehyun.rhee@ecs.vuw.ac.nz

evaluating each type independently. The types include: virtual shadows cast onto the real world, real shadows cast onto virtual objects, the real world occluding virtual objects, and overlapping of real and virtual shadows.

- From our experiments we find low perceptual thresholds for depth resolution for mutual shadowing artifacts. We show that low depth fidelity is sufficient for perceptually similar composition, coarse 3D geometry modelling and estimation of depth is an alternative to accurate depth capture.

2 Related Work

2.1 MR Rendering with Depth

Seamlessly blending real and virtual worlds has been an ongoing problem in computer graphics. Debevec [8] achieved seamless rendering of virtual objects into photographs using captured illumination data from the real world, and compositing via differential rendering. Local scene geometry is used to simulate occlusion and shadow interaction between real and virtual objects. This geometry is often manually modelled to approximate real shadow receivers [38]. For complex scenes, however, these approximations are either too coarse, or it is time consuming to create an accurate geometric representation.

AR applications have used depth cameras to capture RGBD data for live reconstruction of a dynamic real scene [15,16,27]. Karsch et al. [25,26] automate scene geometry acquisition by inferring a depth map from the given photograph. This achieves realistic composition of shadows and reflections onto the scene while foreground occluders are provided by user input. In visual effects, photogrammetry is a common but resource intensive acquisition method for high quality geometry [9]. Deep compositing also incorporates depth data into visual effects pipelines but does not address mutual lighting artifacts exhibited in rendering [12].

All of these techniques trade off time, accuracy, artist effort, and hardware cost. It is not clear which method of depth acquisition is required for a given composition problem. Previous work by Jacobs et al. [24] uses a coarse geometric model of real objects in AR, and illustrates artifacts that arise. They address certain shadow artifacts by offloading computational effort towards shadow estimation rather than geometry refinement. Such work supports the hypothesis that capturing of low resolution depth could be sufficient. Likewise, measuring the noticeability of these artifacts relative to each other would be useful for these applications, which is the focus of is paper.

2.2 Perceptual Studies with Depth

Depth perception studies have measured the human visual system's (HVS) ability to make distance judgements with depth cues [21], and infer spatial relationships between virtual objects in AR [10, 11]. Depth perception has also been evaluated against a display's ability to convey depth cues [13, 33, 41]. Depth cues on more limited displays have been simulated with depth-of-field blur [2]. Other studies measure the effect of depth-of-field blur on depth perception [30]. Our studies instead focus on properties of reduced depth map data that pertain to MR.

More related to the geometry estimation problem, reduced level of detail (LOD) has been well researched for optimisation of 3D triangle meshes [4, 32]. In particular, perceptual factors such as visual acuity and gaze are often considered to obtain simplified geometry that produces visually similar renderings [31]. Corsini et al. [5] provide a survey of user studies for perceptually based error metrics for mesh distortion in both static and dynamic scenarios. These works aim to preserve perceptually important details such as texture, object silhouettes, and rendering, in fully virtual scenes. Our studies instead measure how well geometry estimation can preserve perceived lighting interaction between real and virtual objects in MR. Sugano et al. [40] show that shadows are important for believable composition, in terms of virtual object presence and spatial cues. In their perceptual studies they use static modelled geometry for shadows to be cast on the real world, but do not vary the geometry in a photorealistic composition setup. So far, no other works perceptually evaluate the effect of depth approximation in a MR setup.

2.3 Perceptually Based Rendering

Perceptually based rendering exploits limitations in the HVS by avoiding computation of rendering effects that are not perceived in the final rendering [42]. In virtual scenes, visibility approximations have been shown to be sufficient for rendering perceivably accurate indirect illumination [44]. Preliminary studies have shown that soft shadows can be rendered with blocker geometry of lower LOD than the actual blocker object [39]. While their shadows were cast from synthetic objects in a virtual scene, it supports the idea that real geometry could be approximated for soft shadowing of virtual objects in MR.

Chalmers et al. [3] showed that virtual objects can be rendered under low resolution radiance maps and maintain seamlessness in its shading and shadow. Iorns and Rhee utilise this work for perceptually optimised image based lighting with low resolution, low dynamic range 360° video for real-time MR object shading [23]. Further studies by Rhee et al. [38] use perceptually based thresholding to parameterise lights in 360° videos. With this notion of perceptually optimised illumination, we apply this concept to depth for finding optimal depth map resolutions for MR.

2.4 Perceptual Studies in Composition

In many previous perceptual experiment frameworks, composition quality is evaluated against a ground truth reference image. The Visual Equivalence [37] metric has been used to accept visually similar images, more so than the Visual Differences Predictor [6] which can still detect imperceivable differences. Křivánek et al. [28] has used this to evaluate HVS sensitivity to common global illumination rendering artifacts. MR applications have also been evaluated by ground truth comparisons [1, 25, 34].

However, often the motivation of MR is to insert objects that do not belong in the same environment as the background. Therefore, we use a framework that evaluates visual coherence and plausibility within the same image, as the ground truth reference does not always exist in practice. We adopt the idea of a self-reference perceptual experiment frameworks used by Chalmers et al. [3] and Rhee et al. [38].

3 Composition with Depth

Before the experimental setup is explained, this section reviews differential rendering composition and how we integrate depth maps in the process. We then explain and label the artifacts that arise when using approximate depth in composition.

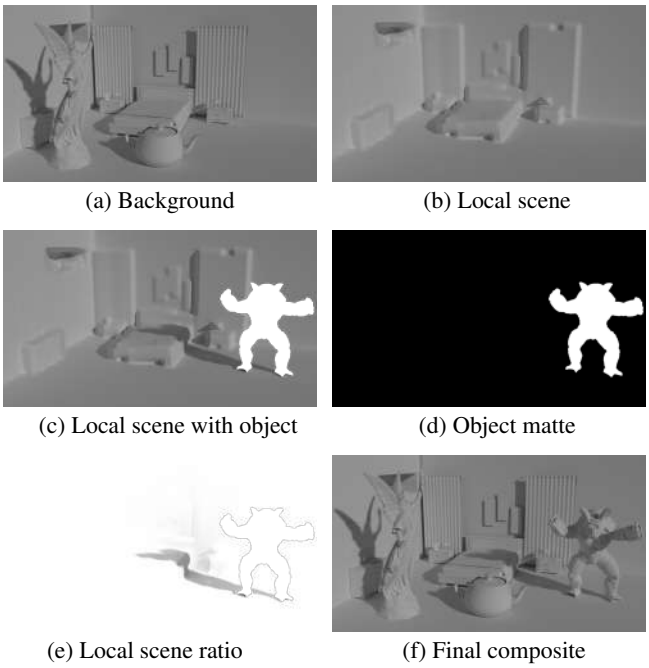
3.1 Differential Rendering

The Differential Rendering method [8] is used to composite the shadows of virtual objects into photographs. This method requires a geometry and material model of the surrounding real environment, which is referred to as the local scene. The local scene is rendered twice: once with the virtual objects and once without. The per-pixel ratio between the renderings extracts the change that the objects imparted onto the local scene (e.g. shadows, reflections, and global illumination). These changes are then composited into the background as follows:

$$I_{final} = I_{Background} \frac{LS_{Obj}}{LS_{NoObj}} \quad (1)$$

In Equation 1, I_{final} is the final composite, $I_{background}$ is the real background image, and LS_{Obj} and LS_{NoObj} are the local

scene renderings with and without objects, respectively. The ratio LS_{Obj}/LS_{NoObj} is multiplied with the background so that shadows (ratio < 1) darken the image and bounce lighting (ratio > 1) brightens it. Figure 2 shows this process. Taking the ratio between renderings and multiplying the result is a variation of differential rendering proposed by Debevec [8]. We use this instead of the more common additive method, which was found to produce noise and clipping artifacts. The local scene is used to receive their shadow detail as if cast onto the real scene. Therefore, the accuracy in the interaction of light, shadow, and occlusion between real and virtual objects is heavily dictated by how closely the local scene represents the real environment.



(a) Background (b) Local scene

(c) Local scene with object (d) Object matte

(e) Local scene ratio (f) Final composite

Figure 2: Compositing steps for differential rendering. (a) Background image, (b) Reconstructed local scene mesh from depth map, (c) Local scene rendered with virtual object, (d) Object matte, (e) Change in local scene: (c) divided by (b) excluding object, (f) Final Composite with object and its shadow.

Many applications acquire the local scene model via depth maps either captured by depth cameras [15, 16, 27], or estimated from the image [25, 26]. Aside from fundamental limitations such as disocclusions (holes) [7], high resolution depth maps can generate a local scene model for composition with correct shadowing and occlusion between real and virtual objects.

3.2 Local Scene Generation from 360° Depth Maps

Our studies use 360° panoramic depth maps to preserve out-of-view geometry, thus avoiding inconsistent illumination artifacts which are not being tested. Out-of-view geometry impacts illumination in the final composite even in conventional fixed-view output. 360° equirectangular depth also allows the depth map resolution parameter to be varied independently of the camera's field-of-view (FOV).

Given a 360° panoramic depth map in equirectangular format, a mesh is generated from it by creating a high density spherical mesh, centred at the origin. We set each vertex's length equal to the depth map value sampled by the vertex's direction. For low resolution depth maps we linearly interpolate the depth samples across the mesh vertices. Using a mesh mitigates artifacts such as holes from other depth image rendering methods such as point clouds.

The generated depth mesh becomes the local scene geometry in differential rendering. Figure 2b shows the geometry generated with this method using a 512×256 depth map, with visible loss of high frequency details in the original scene. This simple approach was sufficient for our experiment but other mesh generation methods such as Pajarola et al. [35] would be valid as well.

3.3 Depth Composition Artifacts

As the depth map degrades in resolution, specific artifacts are observable in the final composite and occur in different types of scenes. We choose a subset of artifacts to evaluate. Based on our preliminary pilot studies, we found that four artifact types were perceived to be most noticeable in depth based compositions. We also noticed that each artifact became noticeable at different depth resolutions. Therefore, our experiments were designed to study these four artifacts individually. Figure 3 shows how each artifact changes in appearance with varying depth resolutions. We define and explain each artifact type in the following.

Type 1 - Real objects receiving virtual shadows: Geometry reconstruction from low resolution depth maps fails to fully reconstruct high frequency details in the real objects in the background image. A virtual object's shadow received by this geometry will hence be distorted in shape, and not match the background in the composite.

Type 2 - Virtual objects receiving real shadows: Similarly to Type 1, shadows cast from the real world geometry onto virtual objects will be inconsistent when low resolution depth is used. This is due to the reconstructed real blocker being distorted, thus casting a different shadow on a virtual object.

Type 3 - Occlusion boundary: This artifact is seen when virtual objects are composited behind real objects and the occlusion boundary is misaligned with the real image. When depth data is used in the composition, depth testing can be done between virtual and real depth values. From this we obtain a visibility mask allowing the real world to occlude virtual objects in compositing. However, low resolution depth maps fail to create occlusion boundaries that exactly match those in the background image.

Type 4 - Real and virtual shadow overlap: This artifact is seen when virtual shadows over-darken the background real shadow, even though the visibility in that region has not changed. When low resolution depth maps poorly reconstruct real world shadow casters, the real shadow simulated during rendering will poorly reconstruct the same shadow region seen in the image. When a virtual object shadow also falls in this region, an artifact of double shadowing or shadow overlap will be present in the composition. The simulated real shadow dictates where virtual object shadows should not darken the background since LS_{Obj} and LS_{NoObj} are both dark. Therefore, the artifact is seen where the simulated real shadow does not fully cover the real shadow in the background. This is illustrated in Figure 4.

These four artifact types are the scope of our experiments. The next section details the design of our experiment which isolates the artifacts to analyse their noticeability independently.

4 User Experiment

The aim of our user experiment is to measure the effect of depth map resolution on seamlessness, while finding the optimal resolution thresholds. Our initial pilot studies have shown that some of the four artifact types (described in Section 3.3) are more noticeable than others. Therefore our goal was to find multiple resolution perceptual thresholds to enable applications to target different composition cases independently. The study was taken by 20 participants (five female, 15 male) between ages 20 to 52. All participants had normal or corrected to normal vision. Out of the 20 participants, 13 were either students, developers, researchers, or artists working in technical fields such as computer graphics, computer science, or computer

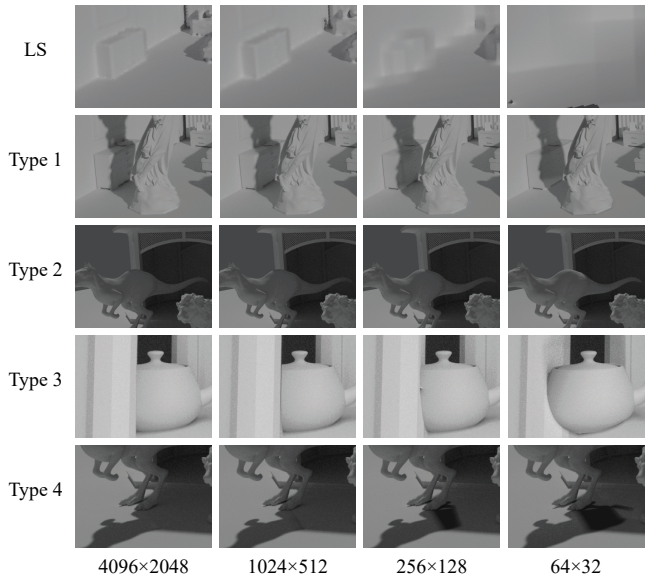


Figure 3: Artifact types and their behaviour as depth map resolution decreases. The bottom labels indicate the depth resolution used in each column. The top row (LS) shows the Type 1 scene's geometry reconstruction as an example. The decreasing depth map resolution causes the geometry to degrade, creating the artifacts in Type 1.

vision. The remaining seven had other non-technical backgrounds. We distributed the test as an online survey through Google Forms.

4.1 Stimuli

Users were shown images of virtual objects composited into fully synthetic scenes, rather than real photographs with 360° depth. This is partly due to the lack of available datasets with 360° panoramas with both high dynamic range (HDR) radiance (as required for photorealistic rendering [8]) and corresponding accurate depth. Furthermore, this would ensure that depth resolution effects are evaluated independently from illumination accuracy. It also allowed us to control camera placement to ensure the artifacts were clearly exhibited in the composites for a meaningful user response. Although 360° depth data was used, users were shown fixed-view composites to direct focus towards the region of interest with virtual objects and artifacts.

To simulate MR rendering with depth, we convert the synthetic scenes' 3D mesh geometry into a depth map by rendering their depth to an image. Composites were created using the method described in Section 3, where equirectangular depth maps were rendered along with the background image. The synthetic depth meshes were then generated from the depth maps and were used as the local scene for each composite. With known illumination and camera projection used to render the background, the virtual objects were then rendered under the same conditions with only the local scene model changing with depth map resolution.

All images were rendered at 960×540 resolution. The 360° depth map resolutions were chosen to be 64×32, 128×64, 256×128, 512×256, 1024×512, 2048×1024, 4096×2048. This choice of range was informed by preliminary pilot studies. These depth maps were re-rendered from the original 3D scene (as in Figure 1a), at the specified resolutions rather than downscaled from a high resolution depth map. This introduces aliasing in the depth maps but avoids filtering to preserve accurate depth measurements. It also simulates capturing from low resolution depth cameras that cannot filter across multiple samples.

4.1.1 Scenes

To measure each artifact type independently we used three test scenes in the stimuli. The scenes cover each mutual illumination situation that exhibits the type being evaluated. We refer to the renderings of these scenes as the *backgrounds* in the compositions. The objects placed in them that exhibit the artifacts are referred to as the *foreground objects*. Figure 5 shows each scene and illustrates the artifacts in the stimuli shown to participants. We have kept the scenes simple to exhibit only the relevant artifacts, while avoiding the influence of factors that were not being tested. For example, all scenes were entirely diffuse to avoid reflection inaccuracies in the virtual objects. Similarly, all materials were the same colour to avoid sensitivity variation with different colours.

The *Bedroom* scene (Figure 5a), which showed Type 1 artifacts, has background cuboid geometry and walls, with the foreground objects Lucy, Teapot, and Armadillo. The shadows of the foreground objects are cast on various surfaces such as the floor, bed, and walls. The requirement of this scene's design was to ensure that shadows did not solely fall on trivial surfaces such as large planes which do not exhibit Type 1 artifacts as depth resolution decreases.

The *Gazebo* scene (Figure 5b) showed both Type 2 and Type 4 artifacts by having a large background Gazebo whose shadow partially falls on the foreground objects Killeroo, Lion, and Dragon. This arrangement provided Type 2 shadows falling on objects, and Type 4 shadow overlap artifacts where the object and Gazebo shadows met. To test these two artifacts independently, we masked out the Type 2 artifact in one set of stimuli images, and masked out Type 4 in the other set.

The *Sponza* scene (Figure 5c) had teapots occluded behind pillars in the environment to show Type 3 artifacts. During rendering indirect light would normally bounce behind the pillars, and illuminate the back walls. After generating the depth mesh, the indirect light would be blocked due to depth disocclusions (cannot store pillar and the wall behind in a single depth map). Hence, objects placed behind the pillars would be dark in the composition. We therefore needed to replicate this in the scene by placing geometry behind pillars, before rendering the background. This prevents indirect light passing behind pillars, thus matching the reconstruction. This enabled the *Sponza* scene to eliminate other illumination artifacts so that the HVS response only depended on the Type 3 artifact.

4.1.2 Objects

The *Bedroom* shadow scene used tall objects (Lucy and Armadillo) under a light with low elevation in order to cast long shadows. These interact with the background across a sufficiently visible area to clearly present any composition artifacts in the stimuli. The teapot is an exception being much shorter but was needed to keep the Armadillo shadow visible in the image. The *Gazebo* scene objects needed to be long enough to be partially in and out of shadow. The Gazebo casts a shadow onto the objects and using long objects ensures that the shadow edge is clearly visible. Similarly, the overlap in the object shadow and Gazebo shadow must be visible on the ground. The light direction was adjusted from behind the Gazebo such that it shadowed the objects in this way. Different objects were used in *Bedroom* and *Gazebo* scenes to prevent participants relying on pattern matching of artifacts to identify the composited objects. The *Sponza* scene used the same teapot objects but they were placed behind different occluders for artifact variation.

To fairly test shadow artifact noticeability, all objects and their shadows were equally visible from the camera's perspective. Similarly, although all composites had the same resolution, the relative distance and scale between objects and camera were consistent within scenes. This imposed restrictions on the type and placement of objects such as the teapot exception in the *Bedroom* scene. Therefore, any object could be used as long as they exhibit the artifacts being evaluated. Objects within *Bedroom* and *Gazebo* had consis-

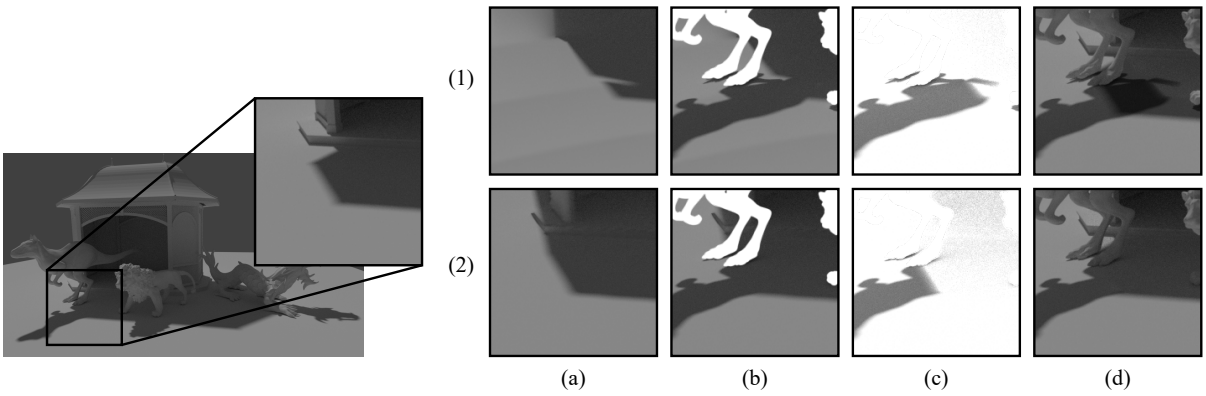
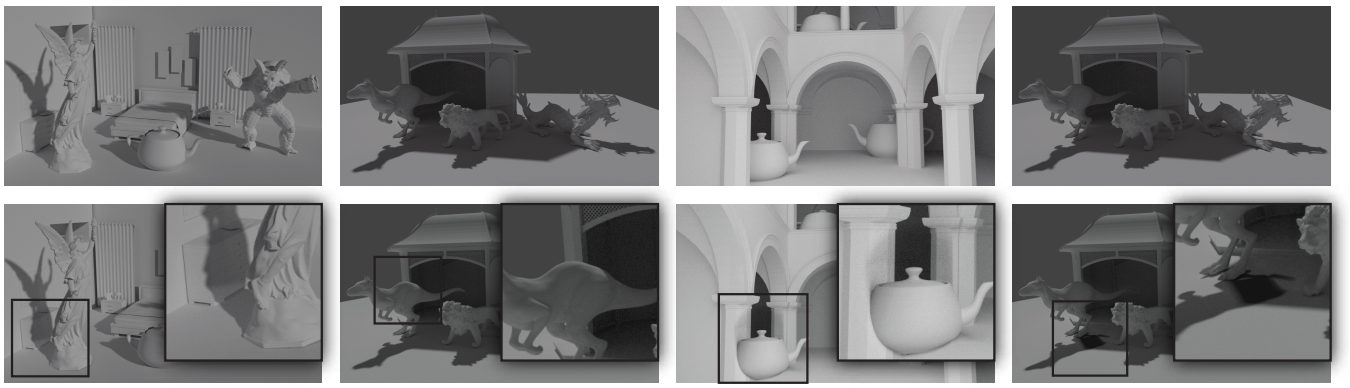


Figure 4: How the Type 4 shadow overlap artifact occurs in differential rendering. The left image and close-up is the background image. Column (a) is the local scene, (b) is the local scene with objects, (c) is the error, and (d) is the final composite. Row (1) shows the process using a 64×32 resolution depth map, while row (2) uses 4096×2048.



(a) Type 1 - Bedroom (b) Type 2 - Gazebo (c) Type 3 - Sponza (d) Type 4 - Gazebo

Figure 5: Top row: ground truth renderings of the scenes used in our experiment. Bottom row: example stimuli with the left most object being the inserted object in each scene. Type 2 (c) and Type 4 (d) share the same scene but with different artifacts masked out in the composition.

tent distances away from the camera, to test artifacts independently from distance. While *Sponza* had objects at different distances, their screen-space sizes were equalised. Plausible camera viewpoints were chosen such that these requirements were met and all objects and their artifacts were visible.

4.1.3 Illumination

Soft shadows were found to be too faint around composition artifacts and were thus unable to stimulate the desired user response across the shadow artifact types. It has also been shown that composition artifacts are much less noticeable under soft lighting compared to hard lighting [3]. Therefore, hard directional lighting was used to create sharp shadows that clearly exhibited the artifacts caused by shadow composition.

Constant colour environment lighting was present in all scenes. This was used to ensure shadows were not too dark. This was required for perceiving the Type 3 shadow overlap artifact because no overlap can be seen when both shadows are pure black.

The *Sponza* scene, however, used only the ambient occlusion rendering pass for lighting. This was to eliminate any hard shadow artifacts from directed lighting and object illumination artifacts from unreconstructed depth. This ensured that this test only evaluated Type 3 artifacts instead of the other shadow artifacts, while keeping the images realistic.

4.2 Procedure

Measuring visual equivalence with a reference can assess the perceived accuracy to which composition matches a ground truth rendering. Instead, we aim to measure the perceived seamlessness of composition into a given background image. Users were shown compositions of foreground objects in the previously described backgrounds, where one of the three foreground objects was composited with the depth map instead of rendered with the environment. Users were then asked which of the three was the composited object. For each object they were asked "How noticeable was the insertion artifact?" with the Likert item ratings "1. Slightly noticeable", "2. Moderately noticeable", "3. Very noticeable", and "4. Extremely noticeable". If the user instead chose a real object, the default noticeability of 0 was given. Users were forced to choose a composited object even when it was too difficult to identify. This means that responses where users picked the correct object by chance, with noticeability of 1, will introduce a bias towards the artifact being noticeable in the results. However, this forced choice does prevent users from frequently relying on a default option, improving the usability of the data. In total, 84 images were shown to each participant (7 depth resolutions × 3 objects × 4 artifact types).

5 Results and Analysis

We firstly average noticeability values from user responses across objects, at each 360° depth map resolution and for each type. In Figure

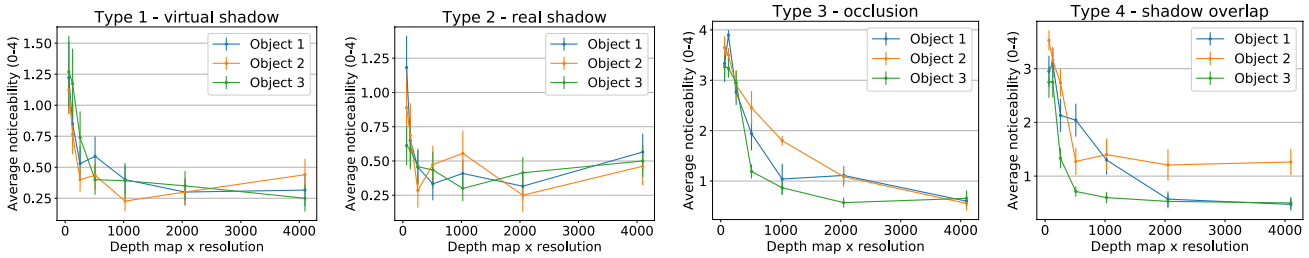


Figure 6: Experiment results of artifact noticeability across all types and objects, against depth map width (height is half width). Lower noticeability values are better. Objects 1, 2 and 3 are the left, middle and right objects, respectively, in each Type's scene.

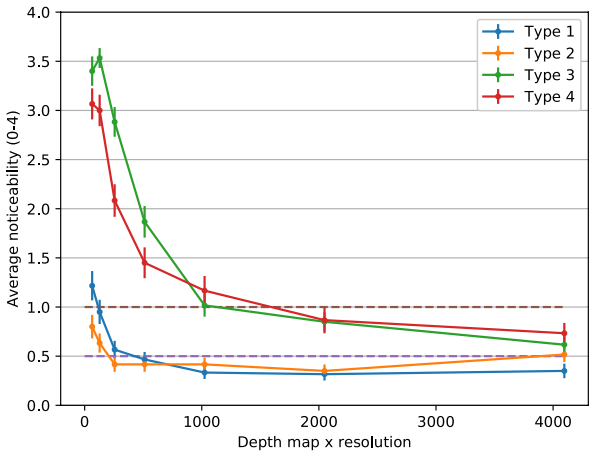


Figure 7: Average artifact noticeability against depth map width, for all types. The dotted line at noticeability = 0.5 indicates where the perceptual thresholds are determined. The dotted line at noticeability = 1.0 indicates where artifacts become *slightly noticeable*. Lower noticeability values are better.

6 we plot the relationship between average artifact noticeability and depth resolution. Specifically, we plot noticeability against the depth map width, as each depth map's height was half its width. The error bars show the standard error of the mean (SEM) indicating relative error size between points. These plots indicate that noticeability decreases with resolution, and is overall consistent across objects and scenes, as expected. It also verifies that the choice of scenes, objects, and placement were suitable for showing the expected behaviour of higher noticeability at lower resolutions.

To find meaningful perceptual thresholds for each artifact type, we average noticeability values across objects and plot this variation with depth resolution in Figure 7. The error bars show each point's SEM. The dashed line at noticeability = 0.5 is halfway between the artifact being noticeable (1.0) and it being not noticeable (0.0). Average noticeability values above 0.5 can distinguish that over 50% of those artifacts were noticeable, instead of not noticeable (users guessing). From this line we obtain the perceptual resolution thresholds for Types 1 and 2, though there are none for Types 3 and 4 within the tested range. As mentioned, users guessing the correct object by chance will bias the results towards higher noticeabilities. Therefore, we also plot a weaker noticeability line at 1.0 to account for this. The thresholds measured at this line indicate where artifacts become *slightly noticeable*, corresponding to the response option in the survey. We also analyse inflection points of each curve. These points indicate the resolution at which the noticeability starts to increase the most. The perceptual thresholds and inflection points are given in Table 1. Analysis methods such as ANOVA and t-test calculation are not suited for our experiment design as we measured gradual change rather than distinct groups with which to measure differences in means.

Table 1: Table of depth map resolution perceptual thresholds (PT), *slightly noticeable* thresholds (SN), and inflection points (IP) from experimental results.

	Type 1	Type 2	Type 3	Type 4
PT	426×213	206×103	-	-
SN	116×58	-	1126×563	1596×798
IP	250×125	228×114	354×177	336×168

The inflection points for Type 1 and 2 closely correspond to their perceptual thresholds, further validating the position of the thresholds. While there are no perceptual thresholds for Types 3 and 4, within the measured range, the *slightly noticeable* thresholds can be found for Types 3 and 4, but not Type 1. All inflection points indicate the optimal design trade-off for quality and resolution for each type.

A level of noise is seen across all plots, causing some artifacts at high resolutions to be more noticeable than at lower resolutions. We attribute this to depth map aliasing as the resolution is reduced without filtering (for accurate depth). Each aliased version sampled the depth map differently, such that artifacts are coincidentally hidden at lower resolutions. The averaged plot in Figure 7 reduces the noise but it is still seen between the first two points of Type 3, and across Type 2.

6 Applications

From our results we provide a number of perceptual thresholds that can be used by artists to understand which areas in the real background require high quality geometry recreation, and which require coarse modelling. Developers can use this understanding to balance computational resources to trade-off performance and quality as required. Researchers can also use the results to develop new perceptually optimised depth acquisition algorithms or depth cameras specifically for MR.

One practical application for the Type 1 result is that, since its perceptual threshold is low, it only requires coarse recreation of surrounding geometry which saves resources such as memory and computation. Coarse box model estimation of geometry [14, 18, 45] is also sufficient as the *Bedroom* test scene composition shows that can seamlessness is achieved without fine details on large background objects and walls.

Figure 8 shows the Type 1 perceptual threshold being used in practice, in a MR rendering of virtual objects into a photorealistic, rendered background image. This simulates compositing into a real photograph with a captured 360° depth map. We illustrate

(a) Composite 4096×2048 (b) Composite 426×213 (c) Mesh 4096×2048 (d) Mesh 426×213

Figure 8: Example application of the Type 1 perceptual threshold in MR rendering into a photorealistic background. The glossy teapot, statue and white chair, are the inserted objects. Compositions with a 4096×2048 depth map (a) and a perceptually optimal 426×213 depth map (b). Reconstructed geometry for the composites are also shown: meshes with 4096×2048 (c) and 426×213 (d) depth maps. Top row: composition into a background scene with continuous depth (couch object). Bottom row: fail case of composition into discontinuous depth (car object with artifacts around wheels in both (b) and (c)).

that composition using a depth map at our perceptually optimal threshold yields visually similar results to using a high resolution depth map. Figure 8 also shows a limitation in which artifacts arise from discontinuous depth (car object in the figure with sharp depth transitions around the wheels), which were not accounted for in our study. In these cases, areas with smooth depth (car bonnet) are visually similar, but artifacts at discontinuities are visible even at high depth resolutions.

Type 3 occlusion artifacts are much more noticeable so we suggest that an occlusion mask or layer is created/detected to mask out virtual objects behind the real world. This is easier to create than a mesh, which must also be manually aligned with the background. The mask can remain in the equirectangular domain for simple compositing.

Types 2 and 4 present a more interesting method as they are both caused by real world shadow casters but have vastly different noticeability. The low noticeability of Type 2 indicates that the blocker geometry can be low fidelity, but this will in turn increase the Type 4 artifact severity. Therefore, we suggest creating/detecting a shadow region matte to mask the shadow overlap. This combination trades off optimised geometry fidelity for requiring a shadow matte input. Both steps combined are often less resource intensive than creating or detecting high fidelity geometry to solve both problems. The coarse blocker geometry could be automatically detected via segmentation and stored as layers [20, 36]. Similarly, shadow detection has also been automated in previous research [17, 29].

We measured artifact noticeability in relation to depth in a full FOV equirectangular format. If a standard, lower FOV depth map is used then depth resolutions even lower than our perceptual threshold may also suffice.

7 Conclusion and Future Work

The aim of this paper was to study the required level of fidelity in background geometry for seamless composition. We achieved this by studying the perceptual effects of depth resolution reduction on the noticeability of mutual illumination artifacts. We formalised four artifact definitions and tested these individually. From the results we obtained perceptual thresholds and optimal trade-off points for multiple MR cases. We finally used these findings to demonstrate where resources can be optimised and where more care must be taken.

Disocclusions in depth capture may lead to discontinuous depth, meaning real shadows cannot be fully reconstructed where depth behind occluders is unknown. However, our studies were limited to continuous depth to avoid such artifacts, so further study is needed to evaluate this case. While our studies evaluated artifacts from depth resolution reduction, we assumed the depth was still accurate. Captured or estimated depth may be both inaccurate and low resolution, so depth accuracy limits for perceived seamlessness requires further studies. Our results could be refined with more participants and repeating the study with a focussed resolution range would find finer perceptual thresholds. Having tested on fully diffuse scenes, we cannot infer how depth resolution affects perceived reflections on virtual objects from the real world. Measuring perceptual depth fidelity thresholds required for indirect illumination and colour bleeding would also be useful. The experiment setup could be extended to measure how multiple lights and lighting complexity affect seamless shadow composition.

Having measured perceptual thresholds in a 360° depth format, our findings can be applied to new 360° depth acquisition [22,43,45] techniques which can enhance MR experiences for mutual rendering effects in 360° content [19]. Further evaluation in dynamic MR scenes would increase the practical use of our results. Composition of animated objects into a video background with dynamic depth could be used in our method. Using animated virtual objects moving in and out of artifact regions could measure if artifacts are more pronounced in dynamic scenes. Further studies using head tracking hardware in a 360° environment could measure how artifact perception is affected while immersed. It is also worth exploring new depth acquisition algorithms with perceptual optimisation in mind.

Acknowledgments

This research was supported in part by the *HDI4D* project funded by *MBIE* and *Entrepreneurial University Programme* funded by *TEC* in New Zealand.

References

[1] M. Borg, M. M. Paprocki, and C. B. Madsen. Perceptual evaluation of photo-realism in real-time 3d augmented reality. In *Proc. of Computer Graphics Theory and Applications, 2014*, pp. 1–10. IEEE, 2014.

[2] K. Carnegie and T. Rhee. Reducing visual discomfort with hmds using dynamic depth of field. *IEEE computer graphics and applications*, 35(5):34–41, 2015.

[3] A. Chalmers, J. J. Choi, and T. Rhee. Perceptually optimised illumination for seamless composites. *Pacific Graphics, The Eurographics Association*, 2014.

[4] J. H. Clark. Hierarchical geometric models for visible surface algorithms. *Communications of the ACM*, 19(10):547–554, 1976.

[5] M. Corsini, M.-C. Larabi, G. Lavoué, O. Petřík, L. Váša, and K. Wang. Perceptual metrics for static and dynamic triangle meshes. In *Computer Graphics Forum*, vol. 32, pp. 101–125. Wiley Online Library, 2013.

[6] S. J. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, vol. 1666, pp. 2–16. International Society for Optics and Photonics, 1992.

[7] I. Daribo and B. Pesquet-Popescu. Depth-aided image inpainting for novel view synthesis. In *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*, pp. 167–170. IEEE, 2010.

[8] P. Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Proceedings of Computer Graphics and Interactive Techniques*, SIGGRAPH '98, pp. 189–198. ACM, 1998.

[9] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 11–20. ACM, 1996.

[10] A. Dey, A. Cunningham, and C. Sandor. Evaluating depth perception of photorealistic mixed reality visualizations for occluded objects in outdoor environments. In *Proceedings of the 17th ACM Symposium on Virtual Reality Software and Technology*, pp. 211–218. ACM, 2010.

[11] A. Dey, G. Jarvis, C. Sandor, and G. Reitmayr. Tablet versus phone: Depth perception in handheld augmented reality. In *Mixed and Augmented Reality (ISMAR), 2012 IEEE International Symposium on*, pp. 187–196. IEEE, 2012.

[12] T. Duff. Deep compositing using lie algebras. *ACM Transactions on Graphics (TOG)*, 36(3):26, 2017.

[13] D. Dunn, C. Tippets, K. Torell, P. Kellnhofer, K. Akşit, P. Didyk, K. Myszkowski, D. Luebke, and H. Fuchs. Wide field of view varifocal near-eye display using see-through deformable membrane mirrors. *IEEE Transactions on Visualization and Computer Graphics*, 23(4):1322–1331, 2017.

[14] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pp. 2366–2374, 2014.

[15] L. Gruber, T. Langlotz, P. Sen, T. Höherer, and D. Schmalstieg. Efficient and robust radiance transfer for probeless photorealistic augmented reality. In *2014 IEEE Virtual Reality (VR)*, pp. 15–20. IEEE, 2014.

[16] L. Gruber, J. Ventura, and D. Schmalstieg. Image-space illumination for augmented reality in dynamic environments. In *2015 IEEE Virtual Reality (VR)*, pp. 127–134, March 2015.

[17] R. Guo, Q. Dai, and D. Hoiem. Single-image shadow detection and removal using paired regions. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 2033–2040. IEEE, 2011.

[18] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *Computer vision, 2009 IEEE 12th international conference on*, pp. 1849–1856. IEEE, 2009.

[19] P. Hedman, S. Alsisan, R. Szeliski, and J. Kopf. Casual 3D Photography. In *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, vol. 36, pp. 234:1–234:15. ACM, 2017.

[20] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. *ACM transactions on graphics (TOG)*, 24(3):577–584, 2005.

[21] I. P. Howard. *Seeing in depth, Vol. 1: Basic mechanisms.* University of Toronto Press, 2002.

[22] J. Huang, Z. Chen, D. Ceylan, and H. Jin. 6-dof vr videos with a single 360-camera. In *Virtual Reality, 2017 IEEE*, pp. 37–44. IEEE, 2017.

[23] T. Iorns and T. Rhee. Real-time image based lighting for 360-degree panoramic video. In *Revised Selected Papers of the PSIVT 2015 Workshops on Image and Video Technology - LNCS, Volume 9555*, pp. 139–151, 2016.

[24] K. Jacobs, J.-D. Nahmias, C. Angus, A. Reche, C. Loscos, and A. Steed. Automatic generation of consistent shadows for augmented reality. In *Proceedings of Graphics Interface 2005*, pp. 113–120. Canadian Human-Computer Communications Society, 2005.

[25] K. Karsch, V. Hedau, D. Forsyth, and D. Hoiem. Rendering synthetic objects into legacy photographs. In *ACM Transactions on Graphics (TOG)*, vol. 30, p. 157. ACM, 2011.

[26] K. Karsch, K. Sunkavalli, S. Hadap, N. Carr, H. Jin, R. Fonte, M. Sittig, and D. Forsyth. Automatic scene inference for 3d object compositing. *ACM Trans. Graph.*, 33(3):32:1–32:15, June 2014.

[27] R. Koch, I. Schiller, B. Bartczak, F. Kellner, and K. Köser. Mixin3d: 3d mixed reality with tof-camera. In *Dyn3D*, pp. 126–141. Springer, 2009.

[28] J. Křivánek, J. A. Ferwerda, and K. Bala. Effects of global illumination approximations on material appearance. In *ACM Transactions on Graphics (TOG)*, vol. 29, p. 112. ACM, 2010.

[29] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan. Detecting ground shadows in outdoor consumer photographs. In *European conference on computer vision*, pp. 322–335. Springer, Berlin, Heidelberg, 2010.

[30] E. Langbehn, T. Raupp, G. Bruder, F. Steinicke, B. Bolte, and M. Lappe. Visual blur in immersive virtual environments: does depth of field or motion blur affect distance and speed estimation? In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, pp. 241–250. ACM, 2016.

[31] D. Luebke and B. Hallen. Perceptually driven simplification for interactive rendering. In *Rendering Techniques 2001*, pp. 223–234. Springer, 2001.

[32] D. P. Luebke. *Level of detail for 3D graphics*. Morgan Kaufmann, 2003.

[33] R. Messing and F. H. Durgin. Distance perception and the visual horizon in head-mounted displays. *ACM Transactions on Applied Perception (TAP)*, 2(3):234–250, 2005.

[34] G. Nakano, I. Kitahara, and Y. Ohta. Generating perceptually-correct shadows for mixed reality. In *Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pp. 173–174. IEEE Computer Society, 2008.

[35] R. Pajarola, M. Sainz, and Y. Meng. *Depth-mesh objects: Fast depth-image meshing and warping*. Information and Computer Science, University of California, Irvine, 2003.

[36] B. Peng, L. Zhang, and D. Zhang. A survey of graph theoretical approaches to image segmentation. *Pattern Recognition*, 46(3):1020–1038, 2013.

[37] G. Ramanarayanan, J. Ferwerda, B. Walter, and K. Bala. Visual equivalence: towards a new standard for image fidelity. In *ACM Transactions on Graphics (TOG)*, vol. 26, p. 76. ACM, 2007.

[38] T. Rhee, L. Petikam, B. Allen, and A. Chalmers. Mr360: Mixed reality rendering for 360 panoramic videos. *IEEE Transactions on Visualization and Computer Graphics*, 23(4):1379–1388, 2017.

[39] M. Sattler, R. Sarlette, T. Mücken, and R. Klein. Exploitation of human shadow perception for fast shadow rendering. In *Proceedings of the 2nd symposium on Applied perception in graphics and visualization*, pp. 131–134. ACM, 2005.

[40] N. Sugano, H. Kato, and K. Tachibana. The effects of shadow representation of virtual objects in augmented reality. In *Mixed and Augmented Reality, 2003. Proceedings. The Second IEEE and ACM International Symposium on*, pp. 76–83. IEEE, 2003.

[41] J. Swan, M. A. Livingston, H. S. Smallman, D. Brown, Y. Baillot, J. L. Gabbard, and D. Hix. A perceptual matching technique for depth judgments in optical, see-through augmented reality. In *Virtual Reality Conference, 2006*, pp. 19–26. IEEE, 2006.

[42] M. Weier, M. Stengel, T. Roth, P. Didyk, E. Eisemann, M. Eisemann, S. Grogorick, A. Hinkenjann, E. Kruijff, M. Magnor, et al. Perception-driven accelerated rendering. In *Computer Graphics Forum*, vol. 36, pp. 611–643. Wiley Online Library, 2017.

[43] J. Xu, B. Stenger, T. Kerola, and T. Tung. Pano2cad: Room layout from a single panorama image. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pp. 354–362. IEEE, 2017.

[44] I. Yu, A. Cox, M. H. Kim, T. Ritschel, T. Grosch, C. Dachsbacher, and J. Kautz. Perceptual influence of approximate visibility in indirect illumination. *ACM Transactions on Applied Perception*, 6(4):24, 2009.

[45] Y. Zhang, S. Song, P. Tan, and J. Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *European Conference on Computer Vision*, pp. 668–686. Springer, 2014.