Project Starline: A high-fidelity telepresence system

JASON LAWRENCE, DAN B GOLDMAN, SUPREETH ACHAR, GREGORY MAJOR BLASCOVICH, JOSEPH G. DESLOGE, TOMMY FORTES, ERIC M. GOMEZ, SASCHA HÄBERLING, HUGUES HOPPE, ANDY HUIBERS, CLAUDE KNAUS, BRIAN KUSCHAK, RICARDO MARTIN-BRUALLA, HARRIS NOVER, ANDREW IAN RUSSELL*, STEVEN M. SEITZ, and KEVIN TONG, Google Research, USA

We present a real-time bidirectional communication system that lets two people, separated by distance, experience a face-to-face conversation as if they were copresent. It is the first telepresence system that is demonstrably better than 2D videoconferencing, as measured using participant ratings (e.g., presence, attentiveness, reaction-gauging, engagement), meeting recall, and observed nonverbal behaviors (e.g., head nods, eyebrow movements). This milestone is reached by maximizing audiovisual fidelity and the sense of copresence in all design elements, including physical layout, lighting, face tracking, multi-view capture, microphone array, multi-stream compression, loudspeaker output, and lenticular display. Our system achieves key 3D audiovisual cues (stereopsis, motion parallax, and spatialized audio) and enables the full range of communication cues (eye contact, hand gestures, and body language), yet does not require special glasses or body-worn microphones/headphones. The system consists of a head-tracked autostereoscopic display, high-resolution 3D capture and rendering subsystems, and network transmission using compressed color and depth video streams. Other contributions include a novel image-based geometry fusion algorithm, free-space dereverberation, and talker localization.

CCS Concepts: • **Computing methodologies** → **Computer graphics**; **Mixed / augmented reality**; **Perception**.

Additional Key Words and Phrases: videoconferencing, telecopresence, eye contact, parallax, stereopsis, spatialized audio, 3D capture

ACM Reference Format:
Jason Lawrence, Dan B Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, Claude Knaus, Brian Kuschak, Ricardo Martin-Brualla, Harris Nover, Andrew Ian Russell, Steven M. Seitz, and Kevin Tong. 2021. Project Starline: A high-fidelity telepresence system. *ACM Trans. Graph.* 40, 6, Article 242 (December 2021), 16 pages. https://doi.org/10.1145/3478513.3480490

1 INTRODUCTION

Improvements in telecommunications have steadily increased both the fidelity and availability of synchronous communication over long-distance networks [Sterling and Shiers 2000]. Video-based systems like Skype, FaceTime, Zoom, Meet, and Teams are a recent step forward in bringing people closer together who are far apart. At the far end of this spectrum is *telepresence*, i.e., enabling remote participants to feel copresent, as if they are occupying a shared physical space [e.g., Draper et al. 1998; Gibbs et al. 1999; Kuster et al. 2012; Maimone et al. 2012; Zhang et al. 2013].

Fig. 1. Our system enables two people to communicate at a distance as if they were physically together. Users report a strong sense of presence and connection with the remote participant.

Telepresence presents tremendous opportunities to bring together the world's increasingly distributed organizations and social groups. However, achieving its full potential poses three grand challenges across multiple research areas:

(1) Capture and render a **3D audiovisual likeness** of a remote person, so realistic that one forgets it is not real.
(2) Create a **comfortable display** with retinal resolution, wide field of view, stereopsis, and motion parallax.
(3) Achieve **copresence** — the feeling that two people are together — including proximity, eye contact, and interaction.

We demonstrate a telepresence system representing a significant milestone along these different dimensions. Notably, user studies demonstrate an improved experience over traditional 2D videoconferencing.

Our unencumbered, bidirectional, 3D communication system is designed for face-to-face meetings. It renders a remote participant as if they were physically copresent, with mutual eye contact (Figure 1). We carefully design and engineer the physical layout, lighting, 3D capture, compression, rendering, display, and audio subsystems to eliminate as many hints as possible that the remote participant is not in the same room as the user.

The primary contribution of this paper is the first telepresence system that achieves measured improvements in meeting experiences and behaviors compared to 2D videoconferencing. User-study

*Now at NVIDIA.

0730-0301/2021/12-ART242
https://doi.org/10.1145/3478513.3480490

participants rated our system as significantly better at fostering various elements of communication including presence, attentiveness, reaction-gauging, eye contact, engagement, and personal connection. They also had greater meeting recall and demonstrated more nonverbal behaviors (hand gestures, head nods, and eyebrow movements) than in 2D videoconferencing.

Outperforming 2D videoconferencing is more challenging than it sounds, for several reasons. First, 2D video is highly realistic, whereas existing real-time 3D capture technologies are all known to suffer visual artifacts, putting them at an inherent disadvantage. Second, compared to 2D displays, most stereoscopic technologies introduce quality trade-offs such as lower resolution, tracking latency, or accommodation-vergence issues, which degrade the experience for many viewers. The fact that our system shows statistically significant user preference over standard videoconferencing despite these challenges is noteworthy.

Additional contributions in our telepresence system include:

- the first use of head-tracked audio crosstalk cancellation, creating the perception that audio originates from the remote user's mouth even as both users move,
- a rendering method that merges multiple depth and color images using an image-based formulation of geometry fusion,
- a 3D facial feature tracking subsystem that combines 2D facial landmark estimation, 3D triangulation, and double exponential filtering to yield accurate predictions at 120Hz.

Please see the accompanying video that approximates the experience of using our system.

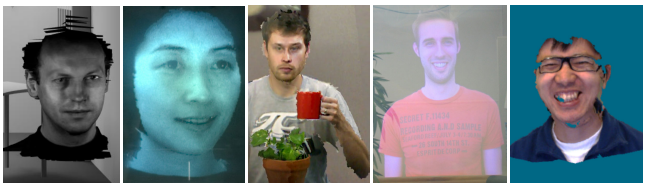
2 RELATED WORK

Videoconferencing. A number of commercial products use custom furniture and specially designed configurations of displays, cameras, microphones, and speakers to heighten the sense of sharing a common space with a remote site [e.g., Cisco Systems, Inc. 2011; DVE 2014; Hewlett-Packard 2005; Plantronics Inc. 2019; Sony 2008; Szigeti et al. 2009].

3D telepresence. Enabling a richer set of 3D depth cues (e.g., stereopsis, motion parallax, and natural scale) provides a stronger sense of immersion and copresence [Gibbs et al. 1999; Muhlbach et al. 1995]. An important goal is mutual eye gaze, a crucial nonverbal cue in human communication [Argyle and Cook 1976; Macrae et al. 2002; Muhlbach et al. 1995]. Researchers have explored telepresence systems for decades [e.g., Baker et al. 2002; De Silva et al. 1995; Dou et al. 2012; Fuchs et al. 2014; Gibbs et al. 1999; Kauff and Schreer 2002; Lanier 2001; Maimone et al. 2012; Maimone and Fuchs 2011; Majumder et al. 1999; Pejsa et al. 2016; Raskar et al. 1998; Schreer et al. 2001; Yang et al. 2002; Zhang et al. 2013]. Figure 2 shows representative images from recent works.

Jones et al. [2009] achieve both stereo and parallax depth cues along with natural eye contact by using a polarized beamsplitter, a high-frequency projector, and a fast spinning mirror to create a volumetric display. However, 3D capture is only performed for one user, so the effect of telepresence is asymmetric.

Maimone et al. [2012] perform 3D capture using 5 Kinect units. To create new stereo images for an autostereoscopic display, they rasterize each Kinect view as a triangulated depth map, then combine



[Gibbs et al. 1999] [Jones et al. 2009] [Maimone et al. 2012] [Kuster et al. 2012] [Zhang et al. 2013]

Fig. 2. Screenshots from prior telepresence research systems.

the rendered images at each pixel using a normal-based weighting of the views seeing the nearest surface. Their experiments with a single system do not demonstrate symmetric communication.

Kuster et al. [2012] realize symmetric telepresence. They perform 3D capture using a single depth sensor and transmit a video stream combining both color and depth. The use of a single depth view simplifies capture, transmission, and rendering, but provides incomplete surface coverage, resulting in disocclusion artifacts.

Zhang et al. [2013] use several IR projectors and cameras to reconstruct multiple depth images. They merge the depth maps to create a sparse 3D point cloud and transmit the point cloud along with color video streams. In contrast, our system transmits depth streams and performs geometry fusion during rendering.

Compared to these prior works, our system includes many novel elements, e.g., multiple compressed depth streams, image-based geometry fusion, high-fidelity face tracking, head-tracked lenticular display, tracker-steered audio beamforming, split-frequency audio spatialization. However, the most important aspect of our work is the significant increase in overall audiovisual fidelity, e.g., comparing Figures 2 and 13. The combined improvements in spatial resolution, color fidelity, depth accuracy, audio, and refresh rate enable our system to demonstrate for the first time an immersive telepresence experience that surpasses classical videoconferencing.

Telepresence using HMDs. The benefits of virtual- and augmented-reality head-mounted displays [Maimone et al. 2013; Orts-Escolano et al. 2016; Wei et al. 2019] include a more immersive experience and a more portable, affordable device. The main difficulty is to obtain a high-quality real-time 3D capture of the user's face while it is hidden behind the headset [Chu et al. 2020; Frueh et al. 2017; Lombardi et al. 2018, 2019; Richard et al. 2021; Wei et al. 2019]. Current work aiming for photorealistic quality involves precaptured user data, unlike in our system.

Gaze redirection. Several techniques improve eye contact with faces in conventional 2D video by digitally altering their perceived gaze direction [Criminisi et al. 2003; Ganin et al. 2016; He et al. 2019; Kononenko and Lempitsky 2015; Wolf et al. 2010; Yang and Zhang 2002]. Our system achieves mutual eye gaze by accurately reproducing the 3D appearance of each user as seen from the other's vantage point, without requiring special processing of eye regions.

Immersive audio in teleconferencing. Spatialized audio in multiperson remote meetings often involves widely distributed microphones and loudspeakers [Plantronics Inc. 2019]. Zhang et al. [2013] incorporate 3D immersive audio using just two loudspeakers, as in our system. Although they mention the possibility of head-tracked

audio rendering and crosstalk cancellation, their system uses a simpler spatialization approach based on gain-and-delay panning. Our system uses talker-tracked microphone-array beamforming for enhanced audio capture, and it uses talker/listener-tracked virtual spatialization with listener-tracked binaural crosstalk cancellation to improve realism.

Autostereoscopic display. Several stereo display technologies show a different image to each eye without requiring glasses [Chen et al. 2014; Dodgson 2005; Wetzstein et al. 2012]. The lenticular display used in our system places a lens array at a precise distance in front of a 2D display [Borner et al. 2000; Matusik and Pfister 2004]. The lens array is similar to a parallax barrier, revealing a different subset of the display pixels to each eye, but the lenses are more optically efficient. A lenticular display can be combined with active head-tracking to *steer* the stereo images to a single user's eyes during head motion. This is accomplished by adjusting the interlaced mapping from the stereo images to the underlying 2D display as a function of the eye locations [Boev et al. 2008; Jurk and de la Barré 2014].

3 HIGH-LEVEL DESIGN

Design goals. Our overriding objective is unencumbered telepresence, i.e., recreating the appearance and sound of a remote user with sufficient quality to enable all conversational cues, while retaining the simplicity of just sitting down and talking with a person in real life. We identify the following requirements:

- Life-size depiction at high resolution, high framerate, and with accurate color;
- Stereopsis and parallax, with left and right views rendered from continuously moving viewpoints with low latency;
- Symmetric video experience, enabling eye contact;
- Symmetric audio experience, with speech perceived to emanate from the virtual participant's mouth;
- Absence of HMD, glasses, tracking fiducials, headphones, or lapel microphones;
- Comfortable use for typical meeting durations.

Design choices. We considered both sitting and standing poses for participants, and selected a **seated** configuration to enable more comfortable conversations. Guided by proxemics work [Hall 1963], we chose a nominal eye-to-eye distance of 1.25 m, just above the boundary between personal and social space, to facilitate a range of social and business interactions[1].

Our choice to pursue a screen-based system is motivated in part by the significant weight and discomfort associated with most current AR and VR headsets. It also eliminates the difficulties of capturing a face through a headset [Wei et al. 2019]. Moreover, it dovetails with our quality objectives, as most widely available VR headsets have an angular resolution less than 20 pixels per degree, and no currently available AR headset has sufficient field of view to span the width and height of a seated human torso. An available technology that meets our combined acuity and field-of-view goals is a head-tracked **autostereoscopic** display based on a 65-inch 8K panel with 33.1M full-color pixels updating at 60 Hz. For a typical adult inter-pupil distance and an eye-to-display distance of 1.25 m, the lens array presents each eye a separate subset of the display pixels (≈5M pixels of each red, green, and blue primary), resulting in an approximate angular resolution of 45 pixels per degree.

Head-tracked autostereoscopic displays can suffer from left-right visual crosstalk, tracking latency, and vergence-accommodation conflict. The impact of these deficiencies increases with **disparity**, which in turn increases as the 3D content is rendered further from the display plane [Perlin et al. 2000]. We mitigate these issues by positioning the virtual space of the remote user such that their face — the typical focus of conversation — lies near the display plane.

Another concern is the abrupt loss of stereo at the display edges. Although 65-inch diagonal panels can comfortably display both the torso and head of most subjects, the torso and hands are clipped at the bottom of the display, giving the impression that a closer object (e.g., hand) is occluded by a more distant object (the display bezel). Such **depth conflicts** can be disorienting or even uncomfortable, pulling participants out of the illusion of presence. As a solution, we place a "middle wall" 0.59 m in front of the display to block the user's view of the display bottom. (We assume a user seated 1.25 m from the display, with seated height less than the 95th percentile or 97 cm.) The wall induces the illusion that the hands and seated legs of a remote user may exist just behind it, thereby avoiding contradictory visual cues.

In designing the remote-to-local geometry mappings, it is important to ensure **mutual eye contact**. Let S_1, S_2 denote the spaces of two users U_1, U_2. User U_1 sees a local virtual representation $T_{21}(U_2)$ where $T_{21} : S_2 \mapsto S_1$ is a rigid transformation. Similarly, user U_2 sees the representation $T_{12}(U_1)$. The virtual remote user $T_{21}(U_2)$ should appear to look directly at U_1. However, the gaze of U_2 is directed to $T_{12}(U_1)$. Eye contact is satisfied iff $T_{12} = T_{21}^{-1}$. (If T is parameterized as a rotoreflection R and translation vector t, these must satisfy $R_{12} = R_{21}^{-1}$ and $t_{21} = -R_{21}t_{12}$.) We also desire each transform to provide a level view, equalize seat heights, and position the remote face near the display plane. Many configurations satisfy all properties (Figure 3), including reflection and 180° rotation about the eye-to-display midpoint. Our system supports both these modes. Because people's features are subtly asymmetric, and moreover any text appearing on objects or clothing is obviously asymmetric, we prefer to avoid reflection and therefore choose 180° rotation by default.

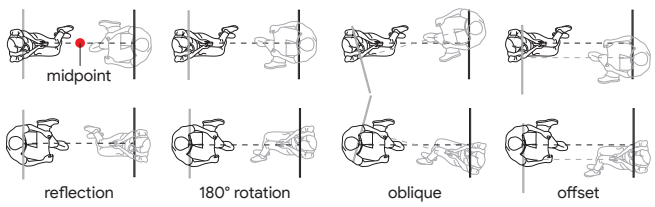


Fig. 3. Examples of geometric maps between system endpoints, showing the pair of real users (black) and their virtual counterparts (gray).

[1]Although this boundary is culturally-dependent, we chose a value appropriate for our North American user study participants.

4 SYSTEM IMPLEMENTATION

As shown in Figure 4, our system comprises two main structures: a *display unit* housing a display, cameras, speakers, microphones, illuminators, and computer, and a *backlight unit* housing an infrared backlight and also serving as a bench seat. Both units contain white LED strips angled toward the walls and ceiling to produce soft bounce lighting (Section 4.1).

The capture subsystem consists of three synchronized stereo RGBD capture pods: two above the display, and one in the "middle wall" below the display. The lower pod includes an extra color camera, zoomed into the subject's face (Section 4.3). Four monochrome tracking cameras, two above the display and one on each side, capture high-speed wide-angle images for real-time 3D localization of the eyes, ears, and mouth (Section 4.4). Figure 5 illustrates the arrangement of our capture and display components. Details of the system components are provided in Appendix A.

The four color and three depth streams from the RGBD capture pods are compressed on the GPU and transmitted alongside tracked 3D face points using WebRTC (Section 4.5).

On the receiving side, the three depth streams are rendered from the viewer's left and right eye locations using a novel "image-based fusion" raycasting algorithm. The four color texture streams are projected onto the fused surface and blended using weights determined from smoothed surface normals (Section 4.6).

The audio capture subsystem uses four cardioid microphones and the tracked mouth position for beamforming to reduce extraneous sounds and echo. The display subsystem uses the tracked talker mouth and listener ears together with a head-related transfer function (HRTF) model to generate a spatialized binaural audio output and then uses the tracked listener ears for loudspeaker-based delivery of this output (Section 4.7).

Computations are performed on a Lenovo P920 PC with two PCIe expanders, dedicated USB 3.0 controllers for the cameras, and four NVIDIA GPUs (two Quadro RTX 6000 and two Titan RTX). All video processing in the system is performed at 60 Hz, except the face tracking at 120 Hz and the NIR stereo pattern capture at 180 Hz. Figure 6 illustrates the data flow between components in a pair of sending and receiving endstations.

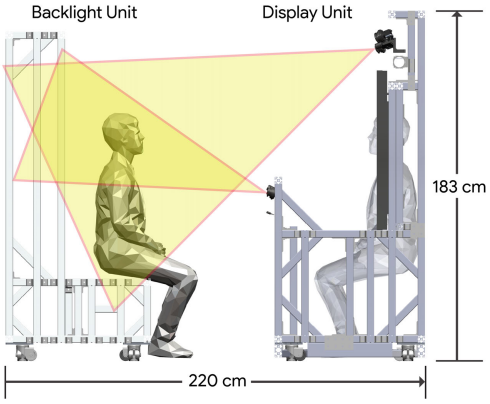


Fig. 4. Side-elevation view of our prototype system, illustrating the relative placement of the user, cameras, display, and virtual remote participant.

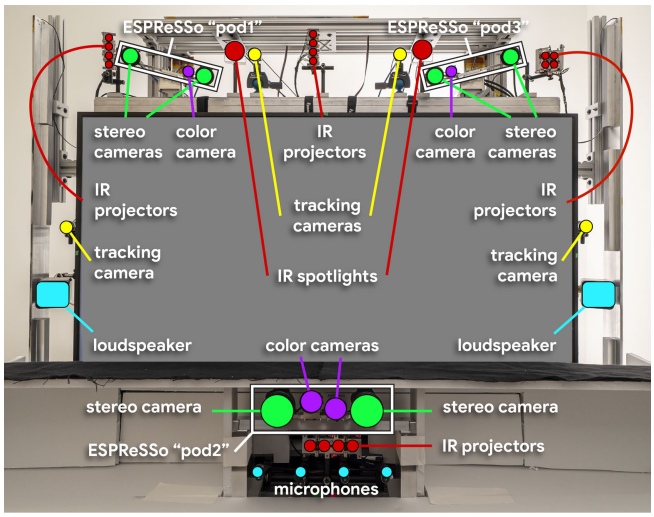


Fig. 5. A front view of the display unit with the three capture pods, the NIR projectors, tracking cameras, loudspeakers, and microphones.

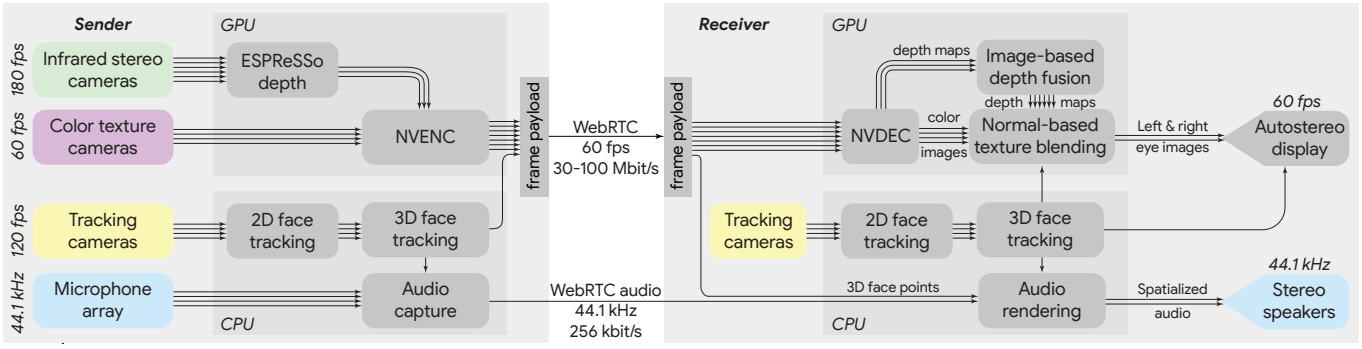


Fig. 6. Overview of the data flow in our system, illustrating how the main processing components are mapped to the GPU and CPU. All video processing in the system is performed at 60 Hz, except the face tracking at 120 Hz and the infrared stereo pattern capture at 180 Hz.

Fig. 7. Left: Our system prototype, showing the LED lights on the sides of the display and backlight units that create bounce lighting on the adjacent walls. Right: Illumination of a subject in our system.

4.1 Lighting

To render the captured surface at novel viewpoints, our image-based rendering approach does not reconstruct illumination or reflectance models. Instead, it simply interpolates the textures from the four color cameras. A drawback is that surfaces with non-Lambertian (e.g., specular) reflectance are rendered incorrectly under non-diffuse lighting. To mitigate this, we create a soft lighting environment using indirect "bounce" sources. On the sides and back of the display and backlight units, white LED light strips illuminate the surrounding walls, producing a pleasing diffuse source that minimizes sharp highlights. This spread-out light is also more comfortable for the user than direct illumination by the bright LEDs.

At the same time, it is important to maintain some illumination nonuniformity. We find that completely uniform incident lighting makes faces and other 3D shapes look flat and artificial, hindering the other 3D cues in the system. Photographers and cinematographers refer to the contrast between the fully lit and shadowed sides of a subject as the "lighting ratio" [Warren 2003]. To retain a sense of dimensionality on the subjects, we use stronger intensities on one side of the display unit adjacent to a nearby wall, producing a lighting ratio of approximately 2:1 (Figure 7).

4.2 Calibration

Stereo capture, 3D face tracking, and rendering all require precise knowledge of the camera geometries. We calibrate the cameras by minimizing reprojection error [Zhang 2000] over images of an approximately planar target [Calibu Contributors 2014], while simultaneously estimating the target's non-planar warp. This provides the camera intrinsics and their relative extrinsics, but not the absolute camera positions relative to the display. All cameras are front-facing and do not directly see the display, so we show a calibration pattern on the display and use a hand-held mirror to reflect the pattern into the cameras' fields of view. Given a set of such mirror images, we use the approach of Hesch et al. [2008] to solve for the relative transform between the display and the set of calibrated cameras.

We color-calibrate the system's RGB cameras by adjusting each camera's gain, color correction (3×3) matrix, and gamma to make a standard color target [McCamy et al. 1976] match its reference color values under the D65 illuminant, thereby neutralizing the effects of room lighting. The display is color-calibrated to make an image captured under the D65 illuminant look like it is captured under the local room's lighting condition (intensity and color). This color calibration regimen ensures that the system automatically corrects for small differences in lighting between the two users' locations.

4.3 Capture

Our goal is to render novel images of each user as they should appear from the other user's left and right eyes. Obviously, if we could place cameras precisely at these eye positions, the capture would be trivial. Unfortunately, this is infeasible because (1) these positions would lie near the center of the display (thus occluding or being occluded by it) and (2) users are free to move in all 3 dimensions. Emerging see-through display technologies could potentially solve part of this problem. However, transparent *autostereoscopic* displays do not yet exist, and in any case would not address viewer motion.

Thus, we place the capture sensors around the periphery of the display. Because the display subtends a large angle to the local user, the capture viewpoints are distant from the eye locations we need to render. To account for this large parallax, we reconstruct geometric approximations of the user, using a combination of visible and near-infrared (NIR) global-shutter image sensors.

The sensors are arranged into three capture *pods*, two above the display and one in the wall below it (Figure 5). The upper pods have a good view of hand gestures and sides of the head and torso, while the lower pod has a good view of the neck, face, and chin. The pod configuration spans a sufficiently large volume (width 1.4 m, height 1.0 m, depth 0.9 m) to capture the head, torso, arms, and hands of a seated user conversing and gesturing naturally.

For our task, commercially available RGBD sensors have insufficient resolution and inadequate reconstruction quality near depth discontinuities, so we design our own capture pods around a high-quality, real-time, active-pattern spacetime stereo algorithm [Nover et al. 2018]. We use the same resolution, framerate, and algorithm as that published technique, but different IR illumination conditions as explained below.

Each pod has a 1600×1200 RGB camera for texture and a pair of monochrome 1280×1024 NIR cameras for stereo. Pods create depth maps at 60 Hz by incorporating information from overlapping time windows of 5 NIR image pairs. In four of these image pairs, the scene is illuminated by different NIR dot patterns (created by 16 diffractive optical element projectors arranged in 4 banks around the display, $\lambda \approx 825$ nm). The fifth image pair, called the guide image pair (see inset), is captured under patternless NIR illumination; its main purpose is to provide a signal for guided filtering in the stereo computation. This guide-image NIR illumination is designed independently from the visible illumination. It involves 3 types of NIR lights ($\lambda \approx 850$ nm). The first is diffuse NIR bounce lighting. The second is a pair of NIR spotlights, which cast shadows at depth discontinuities, thereby helping the guided filter delineate different surfaces. And third, lights inside the backlight unit illuminate the back wall behind the user, so that bright pixels in the guide image can be more robustly classified as background, similar to flash matting [Sun et al. 2006].

With our large capture volume, the 2M-pixel RGB images captured by the pods are sufficient for the torso but cannot resolve finer details on the face. We mitigate this by adding to the bottom *pod2* a second RGB camera with a shorter focal length, framed around the seated user's head. Thus, in total, we capture color images from four viewpoints and depth maps from three.

4.4 3D face tracking

Precise 3D tracking of user facial features is crucial for several system components. The eye locations determine stereo viewpoints for rendering, and are used in the autostereoscopic display to steer the left and right views towards the corresponding eyes. The mouth position enables beamforming in audio capture. And, both the mouth and ear locations contribute to spatialized audio rendering and crosstalk cancellation.

We employ four synchronized 1280×1024 monochrome cameras operating at 120 Hz, with filters to block NIR light. For each captured image, we detect the face and locate 34 facial landmarks [FaceDetector 2019]. We determine the 2D locations of five features (eyes, mouth, and ears) as weighted combinations of nearby landmarks. For each feature found in at least two of the four tracking cameras, we use triangulation to obtain its 3D position.

Minimizing tracking latency is critical. In particular, lag in the eye points used to steer the images out of the autostereoscopic display causes crosstalk between the perceived left and right views. Large lag can even cause stereo reversal, resulting in substantial user discomfort. By examining the interval between the trigger signal of the tracking camera and the response of a photodetector in front the display, we measure a tracking latency of approximately 33 ms.

We mitigate this latency by extrapolating the 3D positions of the tracked features. Such extrapolation amplifies noise, which would result in render viewpoint jitter, so we apply double exponential smoothing, i.e., filtering both estimated velocity and position [Wikipedia 2021]. When the user is stationary, small positional fluctuations persist even after double exponential smoothing. We remove this small noise using a "change band" hysteresis filter. This time-domain filter holds the output constant whenever the input lies within a small band of values. When the input moves above or below this band, the filter output switches to the raw input value, and the "change band" is moved up or down accordingly to track with the input. When used with a very small band, small fluctuations in static input are removed while moving inputs are largely unaffected.

4.5 Compression and transmission

Our goal is to transmit a colored 3D representation bidirectionally between distant systems while maintaining high fidelity and acceptable bitrate. Given data with such high resolution and framerate, real-time compression exploiting temporal coherence is not currently possible with common 3D representations like textured meshes, point clouds, or occupancy volumes.

Instead of creating and sending a merged 3D representation of the captured user, we transmit the multiple color images and stereo-reconstructed depth maps using traditional *video compression*, and delay their "fusion" until the rendering (Section 4.6) of the left and right eye views in the receiving client.

Table 1. The transmission bitrates (in Mbit/s) for each of the 7 video streams vary significantly based on the user's appearance and motion magnitude.

	Depth			Color				
User state	pod1	pod2	pod3	pod1	pod2	pod3	zoom	Total
User wearing a shirt with uniform color:								
stationary	7.7	7.9	8.1	1.4	0.9	1.1	0.9	28.0
speaking	7.7	7.7	8.4	1.7	2.1	1.6	2.3	31.5
moving hands	10.2	14.4	11.7	3.0	3.7	3.1	1.9	47.8
moving arms	14.6	15.4	16.8	4.0	4.6	4.5	3.1	63.1
User wearing a shirt with high-frequency texture:								
stationary	6.9	7.8	7.9	2.4	4.4	3.6	1.5	34.5
speaking	7.2	8.1	8.3	5.0	9.7	6.9	2.9	48.0
moving hands	11.2	13.3	13.3	7.0	15.5	11.8	3.9	76.0
moving arms	15.0	16.4	18.3	11.5	20.8	16.1	4.2	102.3

By using video compression, we are able to exploit the highly optimized video encoders and decoders found in modern GPUs. Specifically, we use the NVENC/NVDEC units of the four NVIDIA GPUs. These have sufficient throughput to process the 4 color streams and 3 depth streams at full resolution and 60 Hz framerate. Both the color and depth streams are encoded using the H.265 codec with YUV420 chroma subsampling. The color streams use 8 bits per channel. The depth streams use 10 bits per channel, with the depth data stored in the Y luminance channel and the UV chroma channels set to 512 (gray). We reduce encoding and decoding latency by omitting bidirectionally encoded (B) frames.

The stereo reconstruction (Section 4.3) yields floating-point depth values. To reduce quantization artifacts when converting to the 10-bit video channel, we linearly rescale each pixel's depth according to the [min, max] depth interval of the pixel's ray through the workspace volume. (We found that encoding reciprocal depth was not necessary, due to our shallow capture volume.)

For each frame, we gather the encoded video packets from all 7 video streams (as well as the tracked face points) into a single data payload, and transmit it using WebRTC [Johnston and Burnett 2012]. In the rare case of a transmission timeout, we reinitialize by sending intra (I) frames for all 7 video streams.

We find that an acceptable level of visual quality is obtained by setting the codec quantization parameter (QP) to 14 for depth data and to 22 for color. The ablation experiment in Figure 16 (Appendix E) shows that 10-bit H.265 compression of the depth images does not significantly affect the quality of the final renderings. As shown in Table 1, the resulting transmission bandwidth varies from about 30 to 100 Mbit/s depending on the texture detail in the user's clothes and the magnitude of their gestures. Although this bitrate range is higher than for traditional 2D videoconferencing, it is already feasible in enterprise networks and increasingly so at home. Variable bitrate coding could be used to regulate bandwidth if warranted.

4.6 Rendering

On the receiving client, after decompressing the 3 depth maps and 4 color images, we render novel left and right perspective views of the virtual remote user from the eye locations of the local user. Our rendering approach processes each frame independently, without temporal history. It consists of three steps:

(1) for each of the 4 color cameras, compute a shadow map using **raycasting** by finding for each ray the first intersection with a surface fused from the input depth maps,
(2) for each of the 2 user views (left and right eye), compute an output depth map using the same **raycasting** algorithm, and
(3) for each output depth map point, compute a **weighted color blend** of the images determined visible by the shadow maps computed in step 1.

Raycasting for geometry fusion. Because the raycasting of fused depth images is invoked a total of 6 times each frame, it is a time-critical component in our system. We present a new method that is 6.7 times faster than prior work.

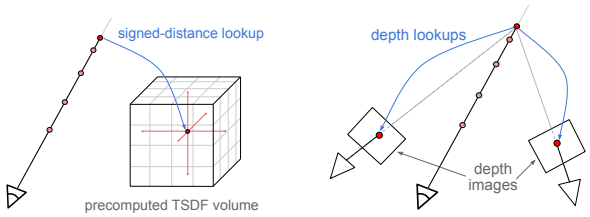
We define the fused surface using a truncated signed-distance function (TSDF) [Curless and Levoy 1996]. The traditional approach is to first accumulate the distance function contributed by each depth view into a **volumetric** grid (e.g., at 1024^3 resolution), weighting the contribution of each depth pixel based on the magnitude of the depth gradient. To render a view, one marches along rays in the voxel grid (Figure 8a) [Hadwiger et al. 2005], sampling the signed-distance until finding a root. Niessner et al. [2013] and Chen et al. [2013] describe several acceleration strategies. In our system, the small number of input depth views and output raycast views at each time step reduces the efficacy of precomputation strategies.

Our contribution is twofold: (1) a fast rasterization scheme to determine search bounds along the rays, and (2) an image-based fusion algorithm that avoids creating a voxel grid.

First, to avoid marching across the entire voxel grid, we compute conservative lower and upper bounds of the distance along each ray by **rasterizing** the input depth views (using point splatting) to a lower-resolution version of the output view, applying 2D min/max filters, and dilating the resulting range bounds by a small margin.

Second, we eliminate voxel storage by using an **image-based** approach that fuses the TSDF on the fly while marching along the rays (Figure 8b). At any point p along a ray, for each depth image $j \in \{1, 2, 3\}$, we transform p into the view coordinates of the depth camera image to sample the stored depth d_j as well as a fusion weight w_j (described later). (If p lies outside the depth camera view frustum, we set $w_j = 0$, so that the depth image is ignored.) We subtract d_j from the z coordinate of the camera-space point to obtain a signed-distance value s_j. Note that s_j is positive if and only if the point p lies in front of the frontmost surface visible from the depth camera. The fused truncated signed-distance is the weighted sum $s = \sum_j w_j \text{clamp}(s_j, -T, T)$ where T is the TSDF truncation distance (2 cm). Similarly to Curless and Levoy [1996], we set $w_j = 0$ for samples where $s_j < -T$ to prevent overcarving. We advance along the ray, within the depth interval computed in the rasterization step, using a step of size $0.8 \cdot s$ until s changes sign. We then perform three steps of bisection search to improve the accuracy of the root. The computation is parallelized across rays using CUDA. This image-based fusion scheme is faster and requires less memory because it reads cached 2D textures and avoids creating a voxel grid. It can be viewed as a generalization of relief texture mapping [Oliveira et al. 2000] to multiple input depth images.

Due to noise in stereo estimation, we find it beneficial to down-weight depth image values in regions where they have high variance.



(a) Volumetric fusion (b) Image-based fusion

Fig. 8. With traditional volumetric fusion, raycasting iteratively samples a precomputed TSDF voxel grid. Our image-based approach instead evaluates the fused signed distance on-the-fly by projectively sampling the input depth images and taking a weighted average.

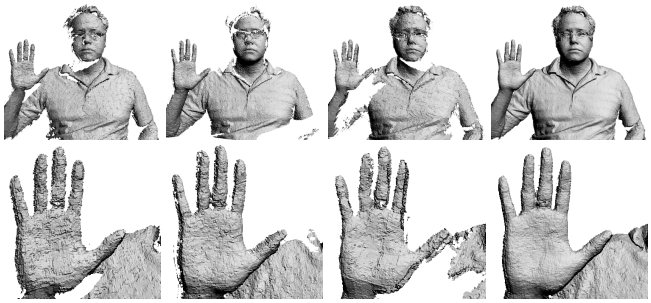


Fig. 9. Contribution of each stereo depth image and resulting fused surface.

Within each depth image j, we assign each pixel i the fusion weight $w_j(i) = \min(.001/\sigma_i, 1)$ with $\sigma_i = \sqrt{\frac{1}{|N_i|} \sum_{k \in N_i} \min((d_i - d_k)^2, T^2)}$, where d_i is its depth and N_i is its 7×7 pixel neighborhood.

Figure 9 shows how fusion from all 3 pods both improves surface coverage and reduces the noise amplitude. To provide a fair evaluation of image-based fusion, we implement an efficient version of volumetric fusion using an occupancy grid of 32^3 supervoxels to discard unoccupied subregions. On an NVIDIA Titan RTX, this volumetric fusion algorithm takes 2.5 ms to accumulate a TSDF grid of size $1152 \times 896 \times 768$ and 7.5 ms to compute the 6 raycasts (four shadow maps and two eye views), for a total of 10.0 ms. Using the rasterized bounds computation, the total GPU time is reduced to 4.3 ms. Switching to image-based fusion further reduces total time to 1.5 ms — a small fraction of the overall 16.6 ms per-frame budget.

Weighted color blending. For each of the left and right eye views, we obtain the color at each pixel by projectively mapping the color images onto the fused geometry (Figure 10), computing partial visibility using percentage closer filtering on the shadow maps [Reeves et al. 1987], and modulating the blend weight of each color image by the squared cosine of the angle between the surface normal and the camera vector. These computations are performed in an OpenGL fragment shader. Similar to Buehler et al. [2001], we assign greater weight to the contribution of the zoom camera to provide more detail over the face.

Although the use of a backlight helps produce crisp silhouettes in the stereo depth estimation, some temporal flickering remains. We reduce its effect using an edge blending technique [Okun and

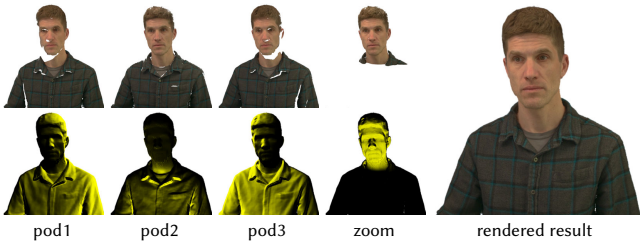


Fig. 10. We project each color image onto the fused surface and combine these using blend weights (yellow) determined from the surface normal.

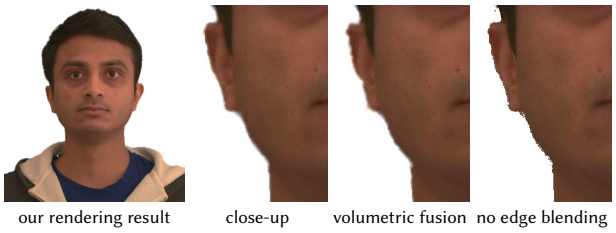


Fig. 11. Our result (image-based fusion with edge blending), modified to instead use volumetric fusion or to omit edge blending.

Zwerman 2010] that adaptively blurs the composite image along depth discontinuities using a Gaussian filter. Figure 11 shows the irregular pixelated silhouette without edge blending. The figure also shows that image-based fusion provides slightly more complete reconstruction than volumetric fusion near the silhouettes.

A few additional features subtly enhance the sense of realism. First, to convey symmetry between local and remote locations, we render a synthetic background that closely resembles the system's backlight unit. Second, to enhance the sense of depth, we cast a soft shadow from the subject onto this virtual background using the shadow maps computed for the two top color cameras. And finally, because our stereo pods cannot see all the way inside an open mouth, we render a dark textured mesh at the back of the tracked mouth to avoid seeing through reconstruction holes to the virtual background.

4.7 Audio

The audio subsystem is designed for high-quality capture of each talker's voice from within their acoustic environment, high-fidelity compression, transmission, and decompression of the extracted voices, and accurate and natural-sounding 3D spatialized rendering of each talker to the opposing listener. We achieve these goals using a novel combination of talker-tracked beamforming, reverberation reduction, WebRTC transmission, talker/listener-tracked virtual audio synthesis, and a split-frequency combination of binaural crosstalk cancellation and amplitude-panning display. Compared to traditional videoconferencing systems, the availability of precise talker and listener tracking is a key enabler for a natural sensation of shared space (Figure 12). To our knowledge, this is the first use of headset-free, head-tracked audio for videoconferencing that spatializes the talkers' voices to emanate from the rendered talkers' mouths.

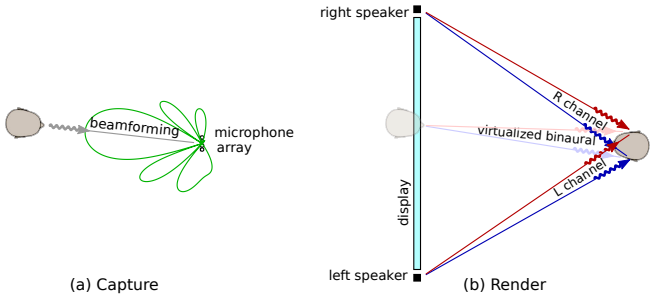


Fig. 12. Audio capture and render. The stereo loudspeakers emit a virtualized binaural signal using a hybrid combination of crosstalk cancellation and amplitude panning, given continuously tracked talker and listener positions.

Capture. Audio is captured at a sampling rate of 44.1 kHz using four cardioid microphones arranged as a linear array in the middle wall, beneath the lower capture pod (Figure 5). Each microphone input is calibrated and equalized to compensate for its frequency response. Capture processing performs the following tasks:

(1) *Ambient noise reduction*: The system transmits only the voices of the two participants from one side to the other. All other acoustic energy (e.g., HVAC, typing, outside talkers) is minimized and, ideally, eliminated.
(2) *Reverberation reduction*: Both sending and receiving sides undergo room reverberation. The system reduces capture-side reverberation so that the remote listener experiences primarily the natural reverberation of their local environment.
(3) *Acoustic echo cancellation (AEC)*: Audio played out of the loudspeakers must be removed from the signal captured by the microphones. This prevents a talker's own voice from echoing back through the remote loudspeaker and microphones.

We implement these three tasks using a combination of techniques:

- The four unidirectional, cardioid microphones [Elko 2004] are oriented in the general direction of the talker to create a base directional-reception pattern providing initial noise and reverberation reduction.
- Tracker-steered, superdirective and noise-constrained optimal-directivity beamforming [Stadler and Rabinowitz 1993] uses the microphone array to sharpen directional reception and further reduce noise and reverberation. The inter-microphone spacing is 0.07 m, for a total array span of 0.21 m.
- Adaptive weighted-prediction-error processing [Caroselli et al. 2017] further reduces reverberation.
- WebRTC VoiceEngine processing [Johnston and Burnett 2012] provides single-channel noise reduction and AEC.

Note that the audio capture system does not perform "blind" acoustic source separation to extract the target talker, but instead uses the 3D mouth tracking system to steer the beamforming toward the talker's mouth during natural conversation (Figure 12a).

Transmission. WebRTC performs compression, transmission, and decompression. Single-channel 44.1 kHz audio is encoded at a target rate of 256 Kbps using the Opus codec (http://opus-codec.org/). The

WebRTC/Opus decoder handles transmission-related factors such as sample-rate mismatch and packet loss concealment.

Render. Stereo loudspeakers (positioned on the sides of the display, 0.1 m below the midline) render tracked and 3D-spatialized audio using a two-step process (Figure 12b): First, the tracked talker and listener positions are combined dynamically with a generic head-related transfer function (HRTF) to yield a real-time-tracked binaural signal. This signal would result in realistic 3D spatialization when presented over headphones. Then, the binaural signal is converted to stereo loudspeaker output using listener-tracked binaural crosstalk cancellation [Gardner 1997; Lentz 2006; Song et al. 2010] with the same HRTF model.

However, we observe that due to short audio wavelengths at high frequencies, inaccuracies in the tracked ear locations can result in audible high-frequency noise. Therefore, we restrict crosstalk cancellation to operate below 1500 Hz, i.e., the low-frequency region in which interaural time difference (ITD) cues dominate perceptual sound localization [Wightman and Kistler 1992]. Above 1500 Hz, we instead weight the loudspeaker outputs using a generalization of vector-based amplitude panning [Pulkki 1997], which considers not just talker tracking but also listener tracking. Assessments indicate that this processing is more robust to tracker error than crosstalk cancellation.

The final loudspeaker signals are calibrated and equalized to compensate for their frequency responses. We introduce an audio delay of about 5 ms using a software FIFO buffer to obtain audio-video synchronization. In addition, a subwoofer boosts the low-frequency energy. The resulting audio subsystem accurately conveys the talker's voice to the opposing listener, reproducing realistic levels, frequency characteristics, and talker/listener 3D spatialization. We measure speech-weighted frequency response errors of 1.5 dB and ITD errors less than 10 μs.

5 ANALYSIS

We first evaluate the effectiveness of our system in two user studies: (1) a small-scale deployment in which participants reported their experiences in remote meetings, and (2) a within-subjects experiment between our system and traditional videoconferencing, evaluating both user sentiments and behaviors.

5.1 Post-meeting surveys in small-scale deployment

To measure the practical effectiveness of our system as a communication tool, we let 117 participants use it for their existing remote work meetings across 3 sites (some separated by >1,000 km). Over a period of nine months, participants held a total of 308 meetings, with an average duration of 35.2 minutes (SD = 16.7). After each meeting, participants were sent a survey to gauge their sentiment of our system relative to the videoconferencing they would ordinarily use. There were 296 survey responses.

Results. Most users (over 87% of survey responses) believed that our system is slightly or much better than traditional videoconferencing across four key communication variables: presence (a sense of "being there" with your meeting partner), attentiveness (being able to pay attention and avoid distractions), personal connection (ability to maintain or establish rapport, trust, and workplace relationships), and reaction-gauging (ability to read the meeting partner's body language and expressions). See Table 5 in Appendix B for complete results.

5.2 Within-subjects experiment

While the strength of the post-meeting survey results in Section 5.1 is external validity (i.e., use for actual remote meetings), its weakness is the lack of causality. Thus, we next conducted a controlled, within-subjects experiment in which each participant had a conversation in both our system and traditional videoconferencing, so we could statistically test for communication variables, as well as move beyond self-report data to include behavioral measures like body language and memory recall.

Experiment setup. All participants ($N = 25$) were recruited within our organization but unconnected to this project. They were offered a small financial incentive (roughly equal in value to 15 USD) for a 30-minute user study session. Each participant had a 5 minute semi-structured conversation with a research confederate using both our system and traditional videoconferencing conditions, in randomized order. The videoconferencing condition had an external setup (e.g., display, seat, viewer distance) similar to our system but used a webcam (Logitech C930e) to stream a 2-D video feed at 720p and 30 Hz. The webcam was placed in front of the confederate at eye-level, so the participant saw a more direct view than typically available in videoconferencing systems, making the viewpoints more similar across conditions. The conversation in each condition was followed by a brief survey of Likert-style self-report measures and a memory recall task. Video recordings (portrait and profile view) of participants were also collected for nonverbal behavioral analysis. To control for content of conversations, we randomly selected questions that have been validated to foster a sense of social intimacy in strangers [Aron et al. 1997].

After each conversation condition, participants rated the extent to which they believed the respective technology facilitated a sense of presence, attentiveness, personal connection, and reaction-gauging, on a scale from 1 ("Not at all") to 5 ("Extremely"). They similarly rated how engaged they were during the conversation and how close they felt toward the meeting partner. They also rated the extent to which the technology facilitated their ability to make eye contact, on a scale from 1 ("Strongly disagree") to 7 ("Strongly agree").

Finally, participants were asked to "write down as many things as you can recall that your conversation partner shared about themself." We counted the number of written words as a proxy of how much they could remember.

For each participant under each condition, we analyzed a clip of the recorded video to count the number of hand gestures, head nods, and eyebrow movements, as detailed in Appendix C.

Results. All statistical analyses use the Wilcoxon-Pratt signed rank test, a non-parametric test to determine whether the differences between pairs are distributed symmetrically around zero. The Wilcoxon-Pratt test also accounts for ties (e.g., when a participant

Table 2. Self-report results. Wilcoxon-Pratt (W-P) statistics indicate that all sentiment improvements are statistically significant ($p < .05$).

	Mean (Standard deviation)			
	Videoconferencing	Our system	W-P	p
Presence, in range [1, 5]	2.88 (.88)	4.52 (.65)	4.43	<.001
Attentiveness [1, 5]	3.52 (.82)	4.36 (.81)	3.18	.001
Personal connection [1, 5]	3.24 (1.01)	4.36 (.70)	3.75	<.001
Reaction-gauging [1, 5]	3.36 (1.08)	4.60 (.50)	3.66	<.001
Engagement [1, 5]	4.12 (.88)	4.84 (.37)	3.07	.002
Closeness [1, 5]	3.40 (.71)	4.44 (.65)	3.75	<.001
Eye contact [1, 7]	5.24 (1.20)	6.20 (1.04)	3.09	.002

Table 3. Measured nonverbal behaviors. Wilcoxon-Pratt (W-P) statistics indicate that all behavior increases are statistically significant ($p < .05$).

	Mean (Standard deviation)			
	Videoconferencing	Our system	W-P	p
Hand gestures	4.68 (2.98)	6.68 (3.55)	2.05	.040
Head nods	7.23 (2.25)	9.09 (3.07)	2.70	.007
Eyebrow movements	3.32 (2.08)	4.95 (3.98)	2.44	.015

has the same score for both measures). We performed these analyses using the "coin" package in R [Hothorn et al. 2008].

As shown in Table 2, participants reported that our system better facilitated presence, attentiveness, personal connection, ability to read partner's nonverbal behavior, and ability to make eye contact than traditional videoconferencing. Also, participants reported feeling closer to the conversation partner and more engaged.

Participants wrote more words after their conversation within our system ($M = 57.3, SD = 30.8$) compared to the videoconferencing condition ($M = 44.8, SD = 24.4$), suggesting that they recalled roughly 28% more meeting content (W-P statistic = 1.93, $p = .053$).

As shown in Table 3, participants also exhibited significantly more nonverbal behaviors (hand gestures, head nods, and eyebrow movements) in our system. Nonverbal behaviors like these are critical to interpersonal communication by conveying information (e.g., a conversation partner's internal states like emotion) [Hall et al. 2019] and facilitating rapport via mirroring (unconsciously replicating someone's nonverbal behavior promotes interpersonal connection) [Chartrand and Bargh 1999].

These convergent findings across multiple measures suggest that our system offers a more engaged communication experience that may be more similar to face-to-face interactions than traditional videoconferencing, even with the visual shortcomings of our system's 3D reconstruction.

5.3 Audio realism study

We also conducted a user study based on signal detection theory [Green and Swets 1966] to measure the realism of our system's audio capture-and-render pipeline, as detailed in Appendix D. This study evaluates how well users are able to distinguish between an audio signal presented from a centrally located loudspeaker in front of the display and a virtual source simulated at the same location using our system. Discriminability is quantified using a sensitivity index [Stanislaw and Todorov 1999], which ranges from −3.72 (perfect negative discrimination ability in this study), through 0.0 (no discrimination ability), to 3.72 (perfect discrimination ability). We obtain an average sensitivity index value of 0.32, which is consistent with low discrimination ability between auditory stimuli from the real-world and from our system. When our system's captured audio stimuli are replaced with "ideal" stimuli captured with a body-worn microphone, the sensitivity index decreases to −0.10, which is consistent with reduced discrimination ability. These results validate our hybrid split-frequency rendering method (combining crosstalk cancellation and amplitude panning).

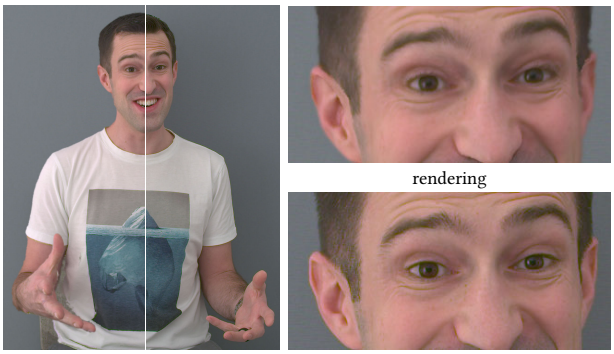


Fig. 13. Comparison of a rendering to a real photo. In the rendering, the synthetic background normally shown in our system is replaced by an image of the empty endstation to enable meaningful qualitative comparisons.

5.4 Reconstruction fidelity

To assess the visual fidelity of our system we add a centrally located and calibrated "witness" camera that records ground-truth photos during a session within our system. We compare these photos to generated renderings for the same camera parameters (Figure 13). These comparisons indicate the accuracy of our capture-and-render pipeline across all stages: calibration, stereo capture, geometry fusion, color blending, and rendering. Note that such a frontal view is challenging to reconstruct due to the oblique angles of the color and depth source cameras. Appendix E includes more such comparisons, together with quantitative error measures.

5.5 System latency

To determine our system's end-to-end latency, we measured the amount of time required for an off-to-on transition of an LED to appear on the display with the system running in loopback mode, using the full WebRTC stack but with two network endpoints on the same PC. We observe an average latency of 105.8 ms (standard deviation 9.1 ms), which is within the 250 ms upper bound required for human participants to perceive a synchronous conversation [Chen et al. 2004]. Note that this is different from the local motion-to-render latency discussed in Section 4.4, which must be much lower to deliver a comfortable and compelling 3D viewing experience.

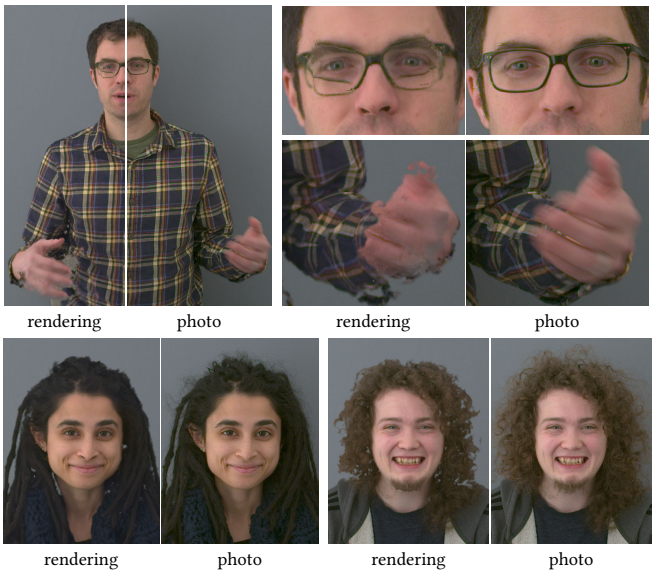


Fig. 14. Limitations. Top: eyeglasses cause texture reprojection artifacts, and fast hand motions yield incomplete reconstructions. Bottom: dark and frizzy hair leads to holes and missing hair strands.

6 CONCLUSION AND FUTURE WORK

We have presented a fully bidirectional and encumbrance-free communication system that reproduces the experience of being physically copresent with another person at a distance. Our communication system is the first that is demonstrably better than traditional 2D videoconferencing, as measured through multiple user studies. Participants using our system reported improvements in presence, attentiveness, reaction-gauging, and engagement, along with greater meeting content recall. These user studies also reveal higher rates of nonverbal behaviors (e.g., head nods, eyebrow movements) that are known to be key indicators of positive meeting dynamics.

We have tested our system across a wide variety of individuals and scenes. Although it achieves a level of audiovisual fidelity not demonstrated in previous telepresence systems, some effects are still not well captured in our system. Thin and semitransparent geometry (e.g., hair and eyeglasses), deep concavities, and fast motion may lead to errors or holes in the reconstructed depth maps, resulting in incorrect geometry and texturing errors (Figure 14). Further work is needed to overcome such artifacts, perhaps by incorporating learned priors or temporal fusion [Dou et al. 2016] into the rendering pipeline.

The major computational steps in our system are the depth-from-stereo calculation, 3D face tracking, compression, geometry fusion via raycasting, and color blending. All of these operations scale roughly linearly with respect to the size of the inputs (RGB and IR images) and outputs (display resolution). We believe the depth-from-stereo and compression steps provide the biggest opportunities for improving the overall efficiency of the system. Regarding improving compression, the color and depth views in the transmitted video streams have much redundancy. Video compression standards include extensions to exploit this redundancy to reduce overall bandwidth [e.g., Vetro et al. 2011]. Unfortunately these extensions are primarily aimed at camera arrays and presently lack real-time encoding implementations. It should be possible to adapt these extensions to make use of the contextual knowledge about the location and movement of both participants.

Finally, we anticipate that future increases in display pixel density and new display architectures will make it possible to extend systems like ours to support multiple concurrent viewers. This will create opportunities for bringing groups of people together more fully than current communication systems allow.

ACKNOWLEDGMENTS

We wish to thank our incredible teammates and colleagues at Google, without whom this work would not have been possible. In particular we are grateful for the guidance and support of Clay Bavor and Andrew Nartker. We also thank Matthew DuVall, TJ Hayes, and Jeff Prouty for their help with the video, and the SIGGRAPH reviewers for their feedback.

REFERENCES

Michael Argyle and Mark Cook. 1976. *Gaze and mutual gaze.* Cambridge U Press.

Arthur Aron, Edward Melinat, Elaine. N. Aron, Robert D. Vallone, and Renee J. Bator. 1997. The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin* 23, 4 (1997), 363–377.

H. Harlyn Baker, Donald Tanguay, Irwin Sobel, Dan Gelb, Michael E. Goss, W. Bruce Culbertson, and Thomas Malzbender. 2002. The Coliseum immersive teleconferencing system. In *Proceedings of the International Workshop on Immersive Telepresence*, Vol. 6.

Danielle Blanch-Hartigan, Mollie A. Ruben, Judith A. Hall, and Marianne S. Mast. 2018. Measuring nonverbal behavior in clinical interaction: A pragmatic guide. *Patient Education and Counseling* 101, 12 (2018), 2209–2218.

Atanas Boev, Kalle Raunio, Mihail Georgiev, Atanas Gotchev, and Karen Egiazarian. 2008. OpenGL-based control of semi-active 3D display. In *Proceedings of the 3DTV Conference*. 125–128.

Reinhard Borner, Bernd Duckstein, Oliver Machui, Hans Roder, Thomas Sinnig, and Thomas Sikora. 2000. A family of single-user autostereoscopic displays with head-tracking capabilities. *IEEE Transactions on Circuits and Systems for Video Technology* 10, 2 (2000), 234–243.

Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. 2001. Unstructured lumigraph rendering. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*. ACM, 425–432.

Calibu Contributors. 2014. Calibu Calibration Library. https://github.com/arpg/calibu. [Online; accessed 17-December-2019].

Joe Caroselli, Izhak Shafran, Arun Narayanan, and Richard Rose. 2017. Adaptive multichannel dereverberation for automatic speech recognition. In *Interspeech 2017*. 3877–3881.

Tanya L. Chartrand and John A. Bargh. 1999. The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology* 76, 6 (1999), 893–910.

Jiawen Chen, Dennis Bautembach, and Shahram Izadi. 2013. Scalable real-time volumetric surface reconstruction. *ACM Trans. Graph.* 32, 4 (July 2013).

Renjie Chen, Andrew Maimone, Henry Fuchs, Ramesh Raskar, and Gordon Wetzstein. 2014. Wide field of view compressive light field display using a multilayer architecture and tracked viewers. *Journal of the Society for Information Display* 22, 10 (2014), 525–534.

Yan Chen, Toni Farley, and Nong Ye. 2004. QoS requirements of network applications on the Internet. *Information Knowledge Systems Management* 4, 1 (2004), 55–76.

Hang Chu, Shugao Ma, Fernando de la Torre, Sanja Fidler, and Yaser Sheikh. 2020. Expressive telepresence via modular codec avatars. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Cisco Systems, Inc. 2011. Cisco TelePresence System T3 System Assembly Guide. (December 2011). https://www.cisco.com/c/dam/en/us/td/docs/telepresence/endpoint/t3/guides/t3_system_assembly_guide.pdf

Antonio Criminisi, Jamie Shotton, Andrew Blake, and Philip HS Torr. 2003. Gaze manipulation for one-to-one teleconferencing. In *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 3. 13–16.

Brian Curless and Marc Levoy. 1996. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '96)*. ACM, 303–312.

Liyanage C De Silva, Mitsuho Tahara, Kiyoharu Aizawa, and Mitsutoshi Hatori. 1995. A teleconferencing system capable of multiple person eye contact (MPEC) using half mirrors and cameras placed at common points of extended lines of gaze. *IEEE Transactions on Circuits and Systems for Video Technology* 5, 4 (1995), 268–277.

Neil A Dodgson. 2005. Autostereoscopic 3D displays. *Computer* 38, 8 (2005), 31–36.

Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. 2016. Fusion4D: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–13.

Mingsong Dou, Ying Shi, Jan-Michael Frahm, Henry Fuchs, Bill Mauchly, and Mod Marathe. 2012. Room-sized informal telepresence system. In *IEEE Virtual Reality Workshops (VRW)*. 15–18.

John V. Draper, David B. Kaber, and John M. Usher. 1998. Telepresence. *Human Factors* 40, 3 (1998), 354–375.

DVE. 2014. DVE Unveils First-of-Its-Kind Holographic Presentation Room. (Jan 2014). https://www.prnewswire.com/news-releases/dve-unveils-first-of-its-kind-holographic-presentation-room-powered-by-microsoft-240082321.html

Gary Elko. 2004. Differential microphone arrays. In *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benetsy (Eds.). Springer, Boston, 11–65.

FaceDetector. 2019. https://developers.google.com/android/reference/com/google/android/gms/vision/face/FaceDetector.

Christian Frueh, Avneesh Sud, and Vivek Kwatra. 2017. Headset removal for virtual and mixed reality. In *ACM SIGGRAPH 2017 Talks*.

Henry Fuchs, Andrei State, and Jean-Charles Bazin. 2014. Immersive 3d telepresence. *Computer* 47, 7 (2014), 46–52.

Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. 2016. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *Proceedings of the European Conference on Computer Vision*. 311–326.

William Gardner. 1997. Head tracked 3-D audio using loudspeakers. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 898–901.

Simon J Gibbs, Constantin Arapis, and Christian J Breiteneder. 1999. TELEPORT –Towards immersive copresence. *Multimedia Systems* 7, 3 (1999), 214–221.

David M. Green and John A. Swets. 1966. *Signal Detection Theory and Psychophysics*. Wiley & Sons, New York, USA.

Markus Hadwiger, Christian Sigg, Henning Scharsach, Khatja Bühler, and Markus Gross. 2005. Real-time ray-casting and advanced shading of discrete isosurfaces. In *Computer graphics forum*, Vol. 24. Wiley Online Library, 303–312.

Edward T. Hall. 1963. A system for the notation of proxemic behavior. *American Anthropologist* 65, 5 (1963), 1003–1026. http://www.jstor.org/stable/668580

Judith A. Hall, Terrence G. Horgan, and Nora A. Murphy. 2019. Nonverbal communication. *Annual Review of Psychology* 70 (2019), 271–294.

Zhe He, Adrian Spurr, Xucong Zhang, and Otmar Hilliges. 2019. Photo-realistic monocular gaze redirection using generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*.

Joel A. Hesch, Anastasios I. Mourikis, and Stergios I. Roumeliotis. 2008. Mirror-based extrinsic camera calibration. In *Algorithmic Foundation of Robotics VIII, Selected Contributions of the Eight International Workshop on the Algorithmic Foundations of Robotics, WAFR 2008, Guanajuato, Mexico, December 7-9, 2008*. 285–299. https://doi.org/10.1007/978-3-642-00312-7_18

Hewlett-Packard. 2005. HP unveils Halo Collaboration Studios: Life-like communication leaps across geographic boundaries. (Dec 2005). https://www8.hp.com/us/en/hp-news/press-release.html?id=170674

Torsten Hothorn, Kurt Hornik, Mark A. Van de Wiel, and Achim Zeileis. 2008. Implementing a class of permutation tests: The coin package. *Journal of Statistical Software* 28, 8 (2008), 1–23.

IEEE. 1969. Recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics* 17 (1969), 225–246. Issue 3.

Alan B. Johnston and Daniel C. Burnett. 2012. *WebRTC: APIs and RTCWEB Protocols of the HTML5 Real-Time Web*. Digital Codex LLC, USA.

Andrew Jones, Magnus Lang, Graham Fyffe, Xueming Yu, Jay Busch, Ian McDowall, Mark Bolas, and Paul Debevec. 2009. Achieving eye contact in a one-to-many 3D video teleconferencing system. *ACM Transactions on Graphics* 28, 3 (2009).

Silvio Jurk and René de la Barré. 2014. A new tracking approach for XYZ-user-adaptation of stereoscopic content. In *Proceedings of the Electronic Displays Conference*, Vol. 3.

Peter Kauff and Oliver Schreer. 2002. An immersive 3D video-conferencing system using shared virtual team user environments. In *Proceedings of the International Conference on Collaborative Virtual Environments (CVE)*. 105–112.

Daniil Kononenko and Victor Lempitsky. 2015. Learning to look up: Realtime monocular gaze correction using machine learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4667–4675.

Claudia Kuster, Nicola Ranieri, Henning Zimmer, Jean-Charles Bazin, Chengzheng Sun, Tiberiu Popa, and Markus Gross. 2012. Towards next generation 3D teleconferencing systems. In *3DTV-CON: The True Vision-Capture, Transmission and Display of 3D Video*. IEEE, 1–4.

Jaron Lanier. 2001. Virtually there. *Scientific American* 284, 4 (2001), 52–61.

Tobias Lentz. 2006. Dynamic crosstalk cancellation for binaural synthesis in virtual reality environments. *J. Audio Eng. Soc* 54, 4 (2006), 283–294.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics* 37, 4 (2018).

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural volumes: Learning dynamic renderable volumes from images. *ACM Transactions on Graphics* 38, 4 (2019).

C. Neil Macrae, Bruce M. Hood, Alan B. Milne, Angela C. Rowe, and Malia F. Mason. 2002. Are you looking at me? Eye gaze and person perception. *Psychological Science* 13, 5 (2002), 460–464.

Andrew Maimone, Jonathan Bidwell, Kun Peng, and Henry Fuchs. 2012. Enhanced personal autostereoscopic telepresence system using commodity depth cameras. *Computers & Graphics* 36, 7 (2012), 791 – 807.

Andrew Maimone and Henry Fuchs. 2011. Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras. In *IEEE International Symposium on Mixed and Augmented Reality*. 137–146.

Andrew Maimone, Xubo Yang, Nate Dierk, Andrei State, Mingsong Dou, and Henry Fuchs. 2013. General-purpose telepresence with head-worn optical see-through displays and projector-based lighting. In *IEEE Virtual Reality (VR)*. 23–26.

Aditi Majumder, W. Brent Seales, M. Gopi, and Henry Fuchs. 1999. Immersive teleconferencing: A new algorithm to generate seamless panoramic video imagery. In *Proceedings of the ACM International Conference on Multimedia*. ACM, New York, NY, USA, 169–178. https://doi.org/10.1145/319463.319485

Wojciech Matusik and Hanspeter Pfister. 2004. 3D TV: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. *ACM Transactions on Graphics* 23, 3 (2004).

Calvin S McCamy, Harold Marcus, James G Davidson, et al. 1976. A color-rendition chart. *J. App. Photog. Eng* 2, 3 (1976), 95–99.

Lothar Muhlbach, Martin Bocker, and Angela Prussog. 1995. Telepresence in videocommunications: A study on stereoscopy and individual eye contact. *Human Factors* 37, 2 (1995), 290–305.

Matthias Niessner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. 2013. Real-time 3D reconstruction at scale using voxel hashing. *ACM Trans. Graph.* 32, 6, Article 169 (Nov. 2013), 11 pages. https://doi.org/10.1145/2508363.2508374

Harris Nover, Supreeth Achar, and Dan B Goldman. 2018. ESPReSSo: Efficient slanted PatchMatch for real-time spacetime stereo. In *2018 International Conference on 3D Vision (3DV)*. IEEE, 578–586.

Jeffrey A. Okun and Susan Zwerman (Eds.). 2010. *The VES Handbook of Visual Effects: Industry Standard VFX Practices and Procedures*. Focal Press, 569.

Manuel M Oliveira, Gary Bishop, and David McAllister. 2000. Relief texture mapping. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM, 359–368.

Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D teleportation in real-time. In *Proceedings of the Symposium on User Interface Software and Technology (UIST)*.

Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek, and Andrew Wilson. 2016. Room2Room: Enabling life-size telepresence in a projected augmented reality environment. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*.

Ken Perlin, Salvatore Paxia, and Joel S Kollin. 2000. An autostereoscopic display. In *Proceedings of SIGGRAPH*. 319–326.

Plantronics Inc. 2019. Polycom RealPresence Immersive Studio. https://www.polycom.com/hd-video-conferencing/realpresence-immersive-video-telepresence.html.

Ville Pulkki. 1997. Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc* 45, 6 (1997), 456–466.

Ramesh Raskar, Greg Welch, Matt Cutts, Adam Lake, Lev Stesin, and Henry Fuchs. 1998. The Office of the Future: A unified approach to image-based modeling and spatially immersive displays. In *Proceedings of SIGGRAPH*.

William T. Reeves, David H. Salesin, and Robert L. Cook. 1987. Rendering antialiased shadows with depth maps. *SIGGRAPH Comput. Graph.* 21, 4 (Aug. 1987), 283–291. https://doi.org/10.1145/37402.37435

Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando de la Torre, and Yaser Sheikh. 2021. Audio- and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 41–50.

Drew Schmidt and Christian Heckendorf. 2017. ngram: Fast n-gram tokenization. R-package.

Oliver Schreer, Nicole Brandenburg, Serap Askar, and Emanuele Trucco. 2001. A virtual 3d video-conferencing system providing semi-immersive telepresence: A real-time

solution in hardware and software. In *Proceedings of the International Conference on eWork and eBusiness*. 184–190.

Myung-Suk Song, Cha Zhang, Dinei Florencio, and Hong-Goo Kang. 2010. Personal 3D audio system with loudspeakers. In *IEEE International Conference on Multimedia and Expo*. 1600–1605.

Sony. 2008. Sony 3D Telepresence. (2008). https://www.tzmc.us/sony/3d_telepresence/index.htm

Robert W. Stadler and William M. Rabinowitz. 1993. On the potential of fixed arrays for hearing aids. *J. Acoust. Soc. Am.* 94 (1993), 1332–1342.

Harold Stanislaw and Natasha Todorov. 1999. Calculation of signal detection theory measures. *Behaviors Research Methods, Instruments, & Computers* 31 (1999), 137–149.

Christopher H. Sterling and George Shiers. 2000. *History of Telecommuniations Technology*. Scarecrow Press.

Jian Sun, Yin Li, Sing Bing Kang, and Heung-Yeung Shum. 2006. Flash matting. *ACM Trans. Graph.* 25, 3 (July 2006), 772–778.

Tim Szigeti, Kevin McMenamy, Roland Saville, and Alan Glowacki. 2009. *Cisco TelePresence Fundamentals* (1st ed.).

Anthony Vetro, Thomas Wiegand, and Gary J Sullivan. 2011. Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard. *Proc. IEEE* 99, 4 (2011), 626–642.

Bruce Warren. 2003. *Photography: The Concise Guide*. Thomson Learning Delmar. https://books.google.com/books?id=w0XJQFxD_S4C

Shih-En Wei, Jason Saragih, Tomas Simon, Adam W Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. 2019. VR facial animation via multiview image translation. *ACM Transactions on Graphics* 38, 4 (2019), 67.

Gordon Wetzstein, Douglas Lanman, Matthew Hirsch, and Ramesh Raskar. 2012. Tensor displays: Compressive light field synthesis using multilayer displays with directional backlighting. In *ACM Transactions on Graphics (Proc. SIGGRAPH*.

Frederic L Wightman and Doris J Kistler. 1992. The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America* 91, 3 (1992), 1648–1661.

Wikipedia. 2021. Exponential smoothing — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Exponential%20smoothing&oldid=1039465483#Double_exponential_smoothing. [Online; accessed 27-August-2021].

Lior Wolf, Ziv Freund, and Shai Avidan. 2010. An eye for an eye: A single camera gaze-replacement method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 817–824.

Ruigang Yang, Celso Kurashima, Andrew Nashel, Herman Towles, Anselmo Lastra, and Henry Fuchs. 2002. Creating adaptive views for group video teleconferencing–An image-based approach. *Screen* 100, P2 (2002).

Ruigang Yang and Zhengyou Zhang. 2002. Eye gaze correction with stereovision for video-teleconferencing. In *European Conference on Computer Vision*. 479–494.

Cha Zhang, Qin Cai, Philip A. Chou, Zhengyou Zhang, and Ricardo Martin-Brualla. 2013. Viewport: A distributed, immersive teleconferencing system with infrared dot pattern. *IEEE MultiMedia* 20, 1 (2013), 17–27.

Zhengyou Zhang. 2000. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence* 22, 11 (2000), 1330–1334.

A ADDITIONAL SYSTEM DETAILS

Table 4. Specifications of system hardware components.

Camera sensors:	
RGB	Basler acA1920-155uc 1600×1200 @ 60 Hz
IR	Basler acA1300-200um 1280×1024 @ 180 Hz
Tracker	Basler acA1300-200um 1280×1024 @ 120 Hz
Lenses:	
RGB pod1,3	Thorlabs MVL12M23 (12 mm EFL, f/1.4, 2/3")
RGB pod2	Thorlabs MVL8M23 (8 mm EFL, f/1.4, 2/3")
RGB pod2 zoom	Thorlabs MVL16M23 (16 mm EFL, f/1.4, 2/3")
IR	Theia ML410M (4–10 mm EFL, f/1.4, 1/1.7")
Tracker	Thorlabs MVL8M23 (8 mm EFL, f/1.4, 2/3")
Autostereoscopic display:	
Panel resolution	7680×4320 @ 60 Hz LCD
Peak luminance	210 cd/m^2
Contrast	3000:1
Lenticular lens	63 mm IPD at 1.25 m distance
Operating region (capture, tracking, and display):	
Volume	1.3 × 1.0 × 1.0 m (W×H×D), centered at a point that is 1.25 m in front of the display center

B SMALL-SCALE DEPLOYMENT SURVEY RESULTS

Table 5. Responses in post-meeting surveys of small-scale deployment (Section 5.1), indicating <1> "Much worse than traditional videoconferencing (TVC)"; <2> "Slightly worse than TVC"; <3> "Same as TVC"; <4> "Slightly better than TVC"; <5> "Much better than TVC". The distributions show that a large majority of participants using our system in their work meetings believe it is slightly or much better than their ordinary videoconferencing experience across these four communication variables.

	Response counts				
	<1>	<2>	<3>	<4>	<5>
Presence	1	2	3	88	193
Attentiveness	1	7	24	99	157
Body language	0	2	34	87	165
Personal connection	0	5	33	100	149

C DETAILS FOR WITHIN-SUBJECT EXPERIMENT

Here are additional details for the experiment in Section 5.2.

Memory recall test. We used an R-package "ngram" [Schmidt and Heckendorf 2017] to count the number of written words. Two participants were excluded from these analyses ($N = 23$) because they did not follow instructions and wrote immaterial information.

Analysis of nonverbal behaviors. For each participant, we recorded four 5-minute videos (profile and portrait view, for each condition), though video data was lost for three participants because of a technical malfunction so $N = 22$ for all nonverbal behavior analyses. Following standard protocols in the field of nonverbal behavior [Blanch-Hartigan et al. 2018; Hall et al. 2019], we analyzed a "thin slice" (1 minute[2]) which has been found to accurately represent nonverbal behavior trends among people. We chose the last minute of each conversation to capture when participants were more likely to be at-ease and speaking naturally (e.g., the beginning of conversations were sometimes influenced by participants' shyness or reactions to new technology).

We developed a coding scheme for a select group of nonverbal behaviors: hand gestures, head nods, and eyebrow movements. Other popularly analyzed nonverbal behaviors like body posture were avoided because the form-factor of our system (e.g., a booth-style sitting area with a straight wall against the user's back) may restrict movement. Two researchers trained in the behavioral sciences analyzed the videos for frequency of each behaviors. Hand gestures were defined as any arm or hand movement (not including fidgeting or self-touching like scratching a nose), and 'one' hand gesture was counted as the moment the movement began to the moment the hands/arms returned to a neutral position. Head nods were defined as any head movement (shaking up and down or side to side) and did not include head tilts (e.g., moving head in one direction, often to communicate confusion or thinking, without quickly moving back to original position). 'One' head nod was counted as the moment the head began moving to the moment it stopped (i.e., if someone nodded their head continually for 10 seconds, that counted as one head nod). Eyebrow movements were defined as any movement of the eyebrows (raised or furrowed), and 'one' eyebrow movement was counted as the moment the movement began to when the eyebrow(s) returned to a neutral position.

Discussion. Using a within-subjects design helps avoid interpersonal variability in idiosyncratic behaviors like gesticulation tendencies by comparing an individual's behavior in one condition to that same individual's behavior in a different condition. By having both communication experiences in the same environment, we control for extraneous variables like screen size, distance from screen, seat ergonomics, and lighting — maximizing internal validity. Future work should include an in-person condition to validate our hypothesis that our system's facilitation of more nonverbal behavior mirrors how individuals behave in face-to-face interaction. Additionally, the experiment detailed herein examines our system as an aggregate experience, in that we are unable to determine the relative influence of each technical offering (e.g., correct eye-gaze, stereoscopic display, motion parallax). Follow up research might isolate each of these variables to examine their individual influence on communication outcomes.

Across self-report measures and nonverbal behavior data, we found that our system fosters a communication experience significantly different from traditional videoconferencing. Participants rated our system as significantly better at fostering various elements of communication like presence, attentiveness, and personal connection, as well as physically demonstrating more nonverbal behaviors compared to the traditional videoconferencing condition. Our data also suggest that participants may remember more of their conversations (e.g., wrote 27.87% more when recalling what they talked about) in our system compared to traditional videoconferencing.

[2]In an effort to standardize the length of analyzed recording time, we examined a 2-minute thin slice in the case of hand-gestures. We employed this strategy because people typically use hand gestures when speaking, but not listening, and we ultimately wanted 60-second analysis periods, as in the case of our other nonverbal measures. We assumed an even split of speaking and listening over the 2-minute slice.

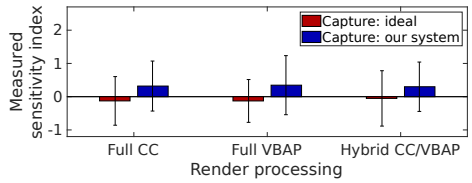


Fig. 15. Average sensitivity index across listeners for the 6 tested conditions. Note that he "Hybrid CC/VBAP" and "Capture: our system" pair of conditions correspond to how our system's audio pipeline operates.

D AUDIO REALISM STUDY

To assess the ability of listeners to discriminate between a source emitting from an actual loudspeaker located at the center of the display and our system's virtual loudspeaker spatialized to the same location, we used a signal-detection-theory task [Green and Swets 1966]. In this assessment, the central loudspeaker represented a repeatable and consistent 'talker' and permitted a large set of data to be collected.

Methods:

- 24 listeners with self-reported normal-hearing were assessed.
- 2 capture and 3 render combinations (yielding 6 conditions for our system in total) were considered.
  - Capture: 'ideal' (from a close-talking microphone) or 'our system' (from the actual system).
  - Render: head-tracked binaural crosstalk cancellation (CC), vector-based amplitude panning (VBAP), or hybrid lowpass/highpass CC/VBAP set by adjusting f_{cut} to 22.05, 0, or 1.5 kHz, respectively. Head-tracking was active and available: listener had full freedom of movement. Note that the hybrid CC/VBAP condition corresponds to our system's audio pipeline.
- Discrimination ability between the actual and virtual sources was measured.
  - Single-interval, two-alternative forced-choice runs were used to assess each condition for our system.
  - A run consisted of N Virtual and N actual loudspeaker (Real) randomly-ordered presentations ($2N$ total).
  - Stimuli were randomly-selected from a set of IEEE sentences [IEEE 1969] spoken by 8 different talkers.
  - Two training (N = 10) and two test (N = 18) runs were presented per system condition. All training runs were conducted prior to the test runs. Condition order was randomized within the training and test blocks.
  - Listeners were requested and encouraged to move their head during presentation.
  - Listeners classified presentations as either 'Virtual' or 'Real'.
  - The final 32 test-run presentations were used to estimate the listener probabilities (i) correctly classifying Real as Real and (ii) incorrectly classifying Virtual as Real. These were in turn used to estimate the sensitivity index and response bias (which may also be referred to as d-prime or d' and c, respectively, [Stanislaw and Todorov 1999]) for the run.

Results:

- Figure 15 presents measured sensitivity index averaged results across listener for the 6 system conditions (2-capture x 3-render).
  - Sensitivity-index = 0 ⇔ inability to discriminate Virtual from Real.
  - Sensitivity-index > 0 ⇔ increasing ability to discriminate Virtual from Real (> 0 for correct classification and < 0 for incorrect classification)
  - For this 32-presentation experiment, the sensitivity-index range spans [-3.72, 3.72].
- Averaged sensitivity indices ranged from -0.13 to -0.05 for ideal (body-worn-microphone) capture and from 0.3 to 0.35 for capture with our system. These values are both near to zero and are well below the perfect-discrimination limits of -/+ 3.72 and are consistent with low discrimination ability.
- Average bias (not plotted) ranged from 0.42 to 0.60, indicating a tendency to classify presentations as Real.
- In terms of the underlying probabilities,
  - Across all six test conditions the listeners correctly classified Real as Real 74.4% of the time.
  - For ideal and system capture, listeners incorrectly classified Virtual as Real 78.3% and 66.1% of the time, respectively, averaged across the three render conditions.

Discussion: Both the sensitivity-index/bias results and the underlying probabilities of classification-as-real reveal

- listener tendency toward identifying both Real and Virtual stimuli as Real;
- low discrimination ability between Virtual and Real stimuli overall; and
- greater contribution of our system's capture (as opposed to render) to any observed discrimination ability.

This indicates that, while our system's audio is generally perceived as real audio, there remain areas for future exploration – in particular in the capture dimension.

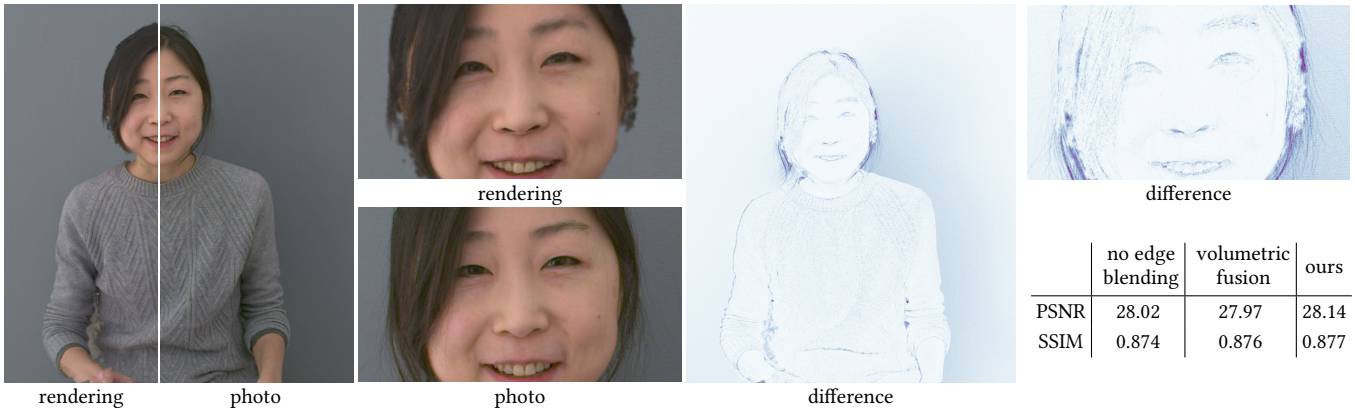
E ADDITIONAL RENDERING COMPARISONS

The rendering comparison in Figure 16 shows that compression of the depth images using the standard HEVC video codec does not significantly affect the quality of the final renderings.

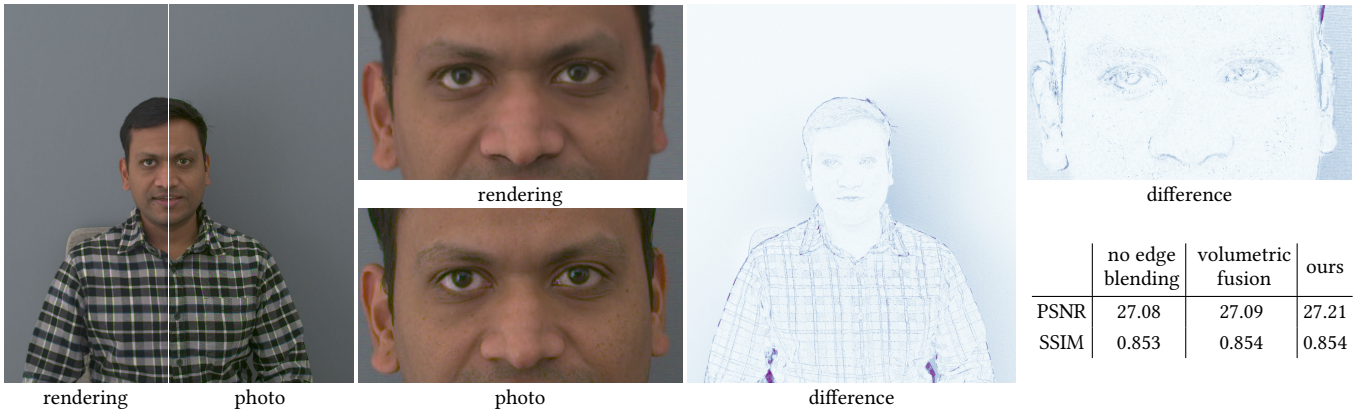
using uncompressed depths using video-compressed depths

Fig. 16. Comparison of rendering using the stereo-reconstructed depth images and those same depth images after the video compression of Section 4.5 (using 10-bit quantization and default quality QP=14). The boundaries of the forearms are slightly worse, but there is otherwise little difference.

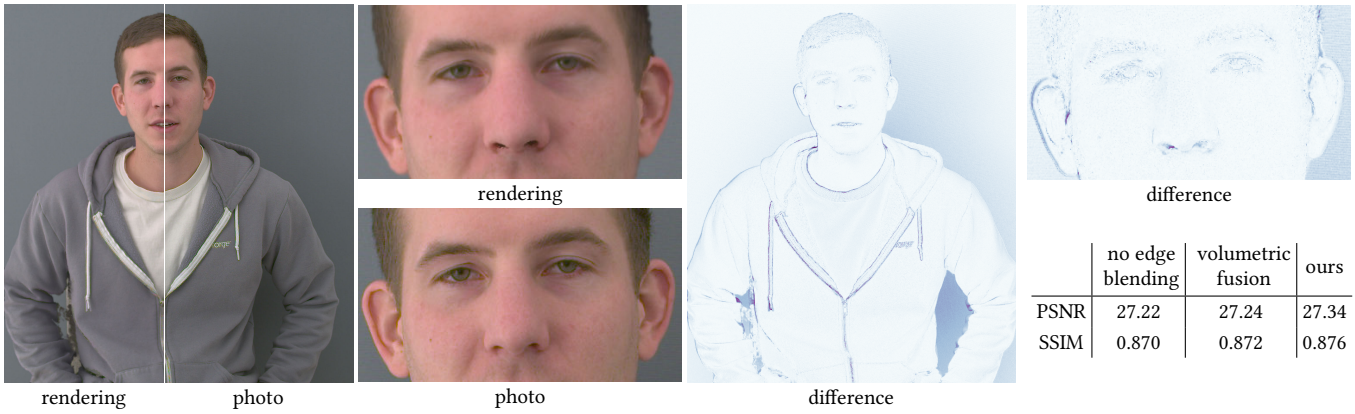
Figure 17 shows additional comparisons between rendering results and groundtruth photos (as in Figure 13), together with quantitative error metrics. We compute the PSNR and SSIM metrics over the whole image including the composited background pixels. The results of our system are compared with volumetric fusion and with omission of edge blending.



	no edge blending	volumetric fusion	ours
PSNR	28.02	27.97	28.14
SSIM	0.874	0.876	0.877



	no edge blending	volumetric fusion	ours
PSNR	27.08	27.09	27.21
SSIM	0.853	0.854	0.854



	no edge blending	volumetric fusion	ours
PSNR	27.22	27.24	27.34
SSIM	0.870	0.872	0.876

Fig. 17. Three additional results (like Figure 13), here including visualizations of the difference between the renderings and real photos. Quantitative metrics show that our approach using image-based fusion and edge blending performs better than the baselines. Note how the error accumulates around silhouette edges and on the missing shadows on the background (which our real-time system emulates but are not used here).