An Equal Seat at the Table: Exploring Videoconferencing with Shared Spatial Context combined with 3D Video Representations

Thomas J. Cashman[*]
Tim Hutton
Martin de La Gorce
Tibor Takács
Antonio Criminisi
tcashman@microsoft.com
tihutt@microsoft.com
madelago@microsoft.com
titakac@microsoft.com
acriminisi@microsoft.com
Microsoft
Cambridge, United Kingdom

Milica Đorđević
Goran Dubajić
Đorđe Marjanović
Milena Okošanović
Vukašin Ranković
Ivan Razumenić
mdjordjevic@microsoft.com
gorand@microsoft.com
dmarjanovic@microsoft.com
milenaokosanovic@gmail.com
rvukasin@gmail.com
irazum@microsoft.com
Microsoft Development Center
Belgrade, Serbia

Bojan Roško
Teo Šarkić
Marko Skakun
Miloš Stojanović
Nikola Veličković
Predrag Jovanović
roskobojan@gmail.com
tsark@microsoft.com
markoskakun@microsoft.com
mistojan@microsoft.com
nikolav@microsoft.com
prjova@microsoft.com
Microsoft Development Center
Belgrade, Serbia

Payod Panda
Lev Tankelevitch
Sean Rintel
payod.panda@microsoft.com
t-levt@microsoft.com
serintel@microsoft.com
Microsoft Research
Cambridge, United Kingdom

ABSTRACT

Work video meetings in the traditional grid interface have inclusion, effectiveness, and fatigue problems, due in part to the difficulty of directing or communicating attention. Virtual 3D meeting spaces have value, but representing people in them is a challenge. Avatars face resistance, and 2D video is limited to near-frontal views, constraining the spatial layout. We present a novel experimental system for virtual meeting rooms that predicts 3D video of users in real-time from a standard webcam, positions them in a shared 3D space, and renders a controllable first-person view. We report study results comparing this system to a traditional grid, and to 2D video of people in the same 3D space. While spatial layouts fared better in terms of attention and co-presence, the traditional grid was more comfortable and professional. This is likely due to unsettled 3D design, the need for manual control, and a preference for the familiar.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

social presence, monocular depth estimation, inclusion, spatiality

ACM Reference Format:
Thomas J. Cashman, Tim Hutton, Martin de La Gorce, Tibor Takács, Antonio Criminisi, Milica Đorđević, Goran Dubajić, Đorđe Marjanović, Milena Okošanović, Vukašin Ranković, Ivan Razumenić, Bojan Roško, Teo Šarkić, Marko Skakun, Miloš Stojanović, Nikola Veličković, Predrag Jovanović, Payod Panda, Lev Tankelevitch, and Sean Rintel. 2024. An Equal Seat at the Table: Exploring Videoconferencing with Shared Spatial Context combined with 3D Video Representations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3613905.3650903

1 INTRODUCTION

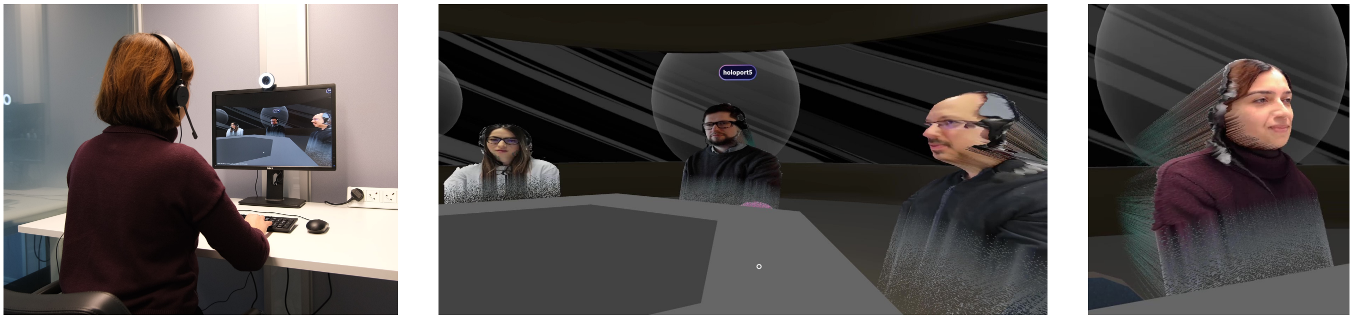
Space is not just where we talk, it's also a part of talk. However, most video meetings are held in traditional 2D grid interfaces. The lack of spatiality in these interfaces challenges users' ability to direct and communicate their attention to others, leading to problems with inclusion, effectiveness, and fatigue. Virtual 3D meeting spaces offer potential solutions, but the representation of participants within these spaces presents its own set of challenges. The use of avatars is sometimes met with resistance in professional settings [19], and 2D video is sub-optimal as users' views can approach or exceed perpendicular orientation to others if the 3D space is to offer tangible benefits (see Figure 3b).

In response to these challenges, this paper explores the potential for 3D virtual meeting spaces to facilitate meetings of people represented with live 3D video, using only commodity hardware. The contributions of this paper are:

- A description of a video meeting system that allows for streaming 3D video using a single standard webcam on a desktop computer. The 3D video is incorporated into a 3D virtual space and users have control of their orientation. Views can be rendered in real-time from a wide range of viewpoints for many-to-many communication.
- An empirical comparison of group meetings using this system with a 3D view in a 3D space, against a 2D view in the same 3D space, and against a traditional 2D interface.

[*] Corresponding author (tcashman@microsoft.com).

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650903



(a) Users of the system sit at a standard PC with a single attached webcam

(b) Rendered first-person perspective view of the 3D virtual space; the view direction is controlled by the user

(c) Each participant is represented using 3D video

Figure 1: Our prototype videoconferencing system predicts 3D geometry corresponding to each user's regular camera feed, and positions all users in a coherent 3D virtual space that is rendered from a first-person perspective. ***Note: We obtained explicit permission from colleagues who were not study participants to include their images in this paper, because showing relative mutual orientation of the 3D representations in the prototypes is crucial to understanding the research.***

We find that while spatial layouts performed better in terms of attention and co-presence, the traditional interface was perceived as more comfortable and professional. The preference for the traditional interface may be attributed to the unsettled nature of 3D design, the need for manual orientation control, and comfort with the familiar. Nevertheless, we conclude that live 3D video in virtual meeting spaces holds great potential for reducing fatigue and improving inclusion and effectiveness in remote meetings.

2 RELATED WORK

We *can* talk without seeing one another (e.g. on the telephone [16]), but space is a resource for talk when we are in person [27]. Conversational flow involves coordinating talk with mutual bodily orientation [9, 21] and nuanced views of gaze, head, and shoulder poses, arm gestures and facial expressions [5, 6, 10, 20, 26, 28, 29]. We also configure spaces in which we meet as resources. Discussions occur in many-to-many configurations, e.g. around a table [6, 34], while presentations occur in one-to-many configurations, e.g. a presenter facing an audience [28].

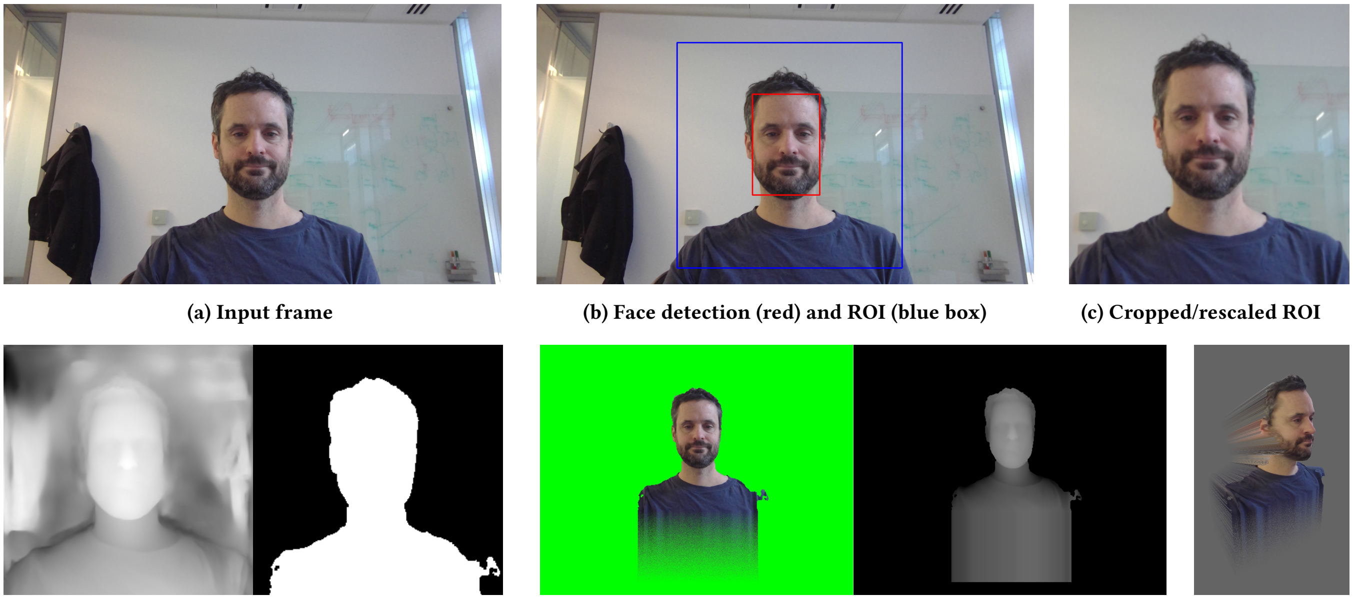
Video meetings fracture the common interactional space [15, 25]. The standard videoconferencing interface for fully remote meetings is a grid of individuals in their respective spaces, and a small self-view in one corner. This creates three problems that are factors in both ineffective meetings and videoconferencing fatigue [32]. *Hypergaze and flattening* [2] are the subjective configuration of *all* meetings as presentations, leaving users feeling stared at and where people look as not representing true focus of attention. Further, there is cognitive load associated with knowing that each attendee is in their unique space, shown against their *differentiated backgrounds*, but having to treat the conversation as happening in one space [17]. The *constant mirror* effect [22] is the cognitive load effect of always seeing oneself.

Video-mediated communication research has long explored spatiality for fully remote meetings [8, 12]. One early system, Hydra [35], provided each user with a small integrated AV unit to be arranged on a desk with others. It demonstrated the value of natural spatial cues, but used specialized hardware. Ensor [18] was an early proponent of software-based spatial room metaphors displayed on standard computer displays for improving the experience of video meetings. More recently, in Perspectives [37], each user occupies a seat at a virtual table in a spatially-consistent virtual room. Every user has a first-person view, and sees all others as 2D live video cutouts. When compared to three Microsoft Teams interfaces (gallery, Together Mode, and Front Row), participants rated Perspectives higher than all other conditions for co-presence, a mental model of where people are, and flow of talk. So spatiality is crucial.

However, the flat and static side-by-side display of video in Perspectives means that conveying attention is very subtle and does not allow for mutual bodily orientation because people are not in 3D. The challenge of spatiality is that while 3D rooms are relatively easy to render, authentic 3D representations of people are difficult. In Virtual Reality (VR), illustrated 3D avatars afford spatial value [7], but if avatars are cartoonish [19] or so hyper-realistic that they create uncanny valley effects [31], they may not be accepted in professional contexts. Hyper-realistic avatars are computationally expensive, have high enrollment complexity, and tend to only be available in dedicated complex hardware setups (e.g. [23, 40]).

Some attempts at enabling 3D representations of people on commodity systems do exist, but are limited. Harrison and Hudson's [11] pseudo-3D view used a webcam and head-tracking to provide motion parallax depth cues, but not 3D views of people. Gaze-2 [39] tracked eye gaze and enabled 2D videos of people on 'billboards' that turned in space towards one another, but the users' faces began to occlude as the angle increased. GazeChat [13] used gaze tracking and neural rendering of users' photographs to show users looking at one another, but did not provide live representations of people, and, because it uses the traditional grid, did not show naturalistic mutual body orientations. Live 3D Portrait [38] showed real-time radiance fields predicted from monocular input, but the view synthesis has limited support for yaw rotation, and the system is currently too expensive for many-to-many settings.

In summary, prior work shows that reducing fatigue and improving inclusion and effectiveness in remote meetings involves a complex combination of issues around spatiality and representation,



(a) Input frame

(b) Face detection (red) and ROI (blue box)

(c) Cropped/rescaled ROI

(d) Depth and segmentation prediction. Depth outside of the predicted segmentation is ignored.

(e) Encoded frame, including color signal with segmentation (left) and per-pixel depth estimation (right)

(f) Novel view rendering

Figure 2: Pipeline for 3D estimation from single-view input. Our pipeline runs in real time on each frame received from the input camera, with steps (a) to (e) on the sending client device, and step (f) on the receiving client.

manifested in and brought to the meeting experience. These issues involve spatial presence and co-presence; attentional awareness and control; conversational flow; comfort, distraction, and representational quality; and personal, social, and professional preferences and norms.

On the basis of this review, we developed the following research questions to explore the potential acceptance and value of 3D video representations in virtual meeting spaces:

RQ1: Can we capture 3D views of people using commodity webcams and render them in a 3D virtual environment?

RQ2: What advantages does user-controlled view direction confer for focusing user attention, and conveying direction of attention to other users?

RQ3: Are there advantages to a 3D video representation of users in the 3D virtual environment?

3 TECHNOLOGY

Our experimental system runs a pipeline (Figure 2) that includes real-time 2D video capture and 3D prediction, 3D video streaming, incoming stream processing, and audio and video rendering.

2D video capture and 3D prediction. We used OpenCV [4] to connect to a webcam , and then process each captured frame to predict plausible 3D geometry for the subject in the frame. The main frame processing steps are to:

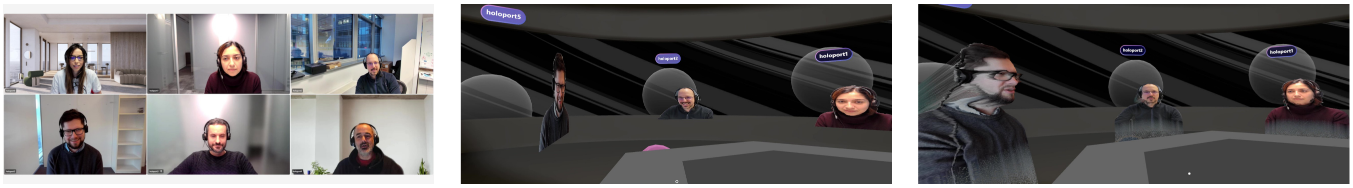
- run an off-the-shelf face detector[1] to locate the user in the frame (Figure 2b);
- crop and rescale an area around the detected face, to form a region of interest (ROI) (Figure 2c);
- pass the ROI to a feedforward convolutional neural network (CNN) that predicts depth for each image pixel, as well as a segmentation of the ROI into foreground and background regions (Figure 2d);
- extend the lower edge of the image, to bridge the gap between the field of view for a typical webcam, and the hexagonal ‘table’ that we placed in the virtual 3D scene (see Figure 3c for an example);
- composite the colour and depth signal, with the extension on the lower edge, into a single stacked image that is suitable for streaming (Figure 2e).

These processing steps are implemented using a combination of the ONNX runtime[2] for CNN evaluation, and CUDA kernels [30] for optimized image processing. We chose a simple solution to extend the prediction below the camera’s field of view, copying the last row of colour and depth pixels downwards. We also added stippling that increases with distance from the captured image to visually indicate that this region is not observed by the camera. This processing pipeline for capture and 3D prediction runs in 35ms per frame on a workstation with an NVIDIA RTX 2080Ti GPU, which allows for real-time prediction and streaming at almost 30 frames per second.

The CNN that performs depth prediction and segmentation is the most technically complex part of this pipeline. We provide details of our training loss in the appendix (see Section C). Once the CNN is trained, it is evaluated in the pipeline on the ROI (Figure 2c), to

[1] https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB

[2] https://onnxruntime.ai/



(a) 2D video, 2D grid

(b) 2D video, 3D space

(c) 3D video, 3D space

Figure 3: Representative views from the three conditions that we tested. For the spatial layouts, (b) and (c), participants saw at most 3 other participants at any one time, and used the keyboard to control their first-person view to look around the scene.

give depth and segmentation predictions (Figure 2d). The end result of the 3D prediction stage is an image that combines the color and depth signals into a single stacked frame in the YUV color space (Figure 2e). The depth signal does not use the UV channels, instead encoding depth purely in the 8-bit Y channel.

Real-time streaming. We transmit these combined frames using WebRTC [3] and the open source Janus server [1] with the videoroom plugin. This gives a media transport solution with a latency typically below 500ms, which is tolerable, although higher than we would expect from a typical videoconferencing system. Each user sends one combined color-and-depth stream alongside the audio signal, and receives *n* streams from the Janus server for a meeting of *n* participants. Our prototype uses the GPU for native decoding of the incoming compressed video streams, alongside the depth prediction processing described above. The end result of this stage is *n* incoming color-and-depth streams.

Incoming stream processing. Each incoming stream is processed to turn the video signal into a 3D mesh, where the texture is derived from the color signal (corresponding to the original captured webcam image) and the vertex positions are determined from the depth signal. We also apply a stylized 'projector' effect, by connecting a subset of vertices in the 3D mesh to a point slightly behind the user's representation (see Figure 2f), to communicate that some portions of the user are not visible to the camera.

Audio and video rendering. The final step for each frame is for the 3D meshes to be rendered in the virtual scene, alongside the geometry for the simple room environment we created for this prototype. The spatialized audio signals are rendered into a single stereo stream that can be consumed using stereo headphones, where audio sources to the left of the participant are more prominent in the left channel of the mixed stream (and correspondingly for the right of the participant). Both audio and video rendering are implemented in Unity.

Participants could control the virtual camera used for the final render using arrow keys on the keyboard; pressing the left or right arrow key rotated their view to the left or right accordingly. This rotation was also reflected for remote participants, where each user was rendered with an orientation which matched their own first-person perspective view (see the accompanying video for details).

4 METHODS

48 participants were recruited by email from a global technology company to take part in one-hour experimental sessions. They comprised 37 males and 11 females, and formed 8 groups of 6 members each (see Table 1 in the appendix for details). All participants completed an informed consent procedure prior to participation.

We conducted a within-subjects experiment comparing (A) a traditional 2D videoconferencing system, (B) our system with 3D representations in a 3D space, and (C) a system with 2D representations in the same 3D space. The Supplementary Materials contain the full details of tasks and surveys.

Drawing on [37], each condition began with participants counting up one at a time from 1 to 20. This was to test the hypothesis that counting in order is easier in a spatial environment than in traditional grid, because the spatial environment provides more cues for speaker ordering.

Participants then engaged in a five minute decision-making discussion that encouraged naturalistic conversation with a seed idea and trajectory (e.g. decide on a team-building event). The specific decision made was not relevant to this study, only the stimulus for purposeful talk.

Following their discussion, participants filled out short surveys about their experience. The survey questions were developed to cover the themes we noted above found in prior work, with one or two questions per theme, drawing inspiration from questions reported in [37] and [24], and additional questions to cover the specific capabilities of our prototype conditions. The themes covered were: spatial presence; co-presence; attentional awareness and control; conversational flow; comfort, distraction, and representational quality; and preferences, professionalism, and norms.

Lastly, given the known quality limitations of the 3D representations in our live prototype, we showed participants a recorded video demonstrating a meeting using high-quality 3D representations (see Figure 5 in the appendix).

Limitations. Due to technical constraints and time, condition order was not counterbalanced across participants. All groups except one completed conditions A to C in that order (one group completed condition B last). We also did not have access to identifiers linking participants' survey data across conditions, and therefore report statistical tests assuming independent samples, acknowledging that caution is strongly warranted in interpreting quantitative findings.

5 RESULTS

Our results report responses to the end-of-session surveys. The quantitative results of each question are reported along with representative examples of participants' observations and reasoning drawn from the their input into open text fields at the end of each survey.

Participants' responses for each survey item were compared between conditions using a non-parametric Kruskal-Wallis test, with significant effects followed up with pairwise comparisons using the Wilcoxon rank-sum test (conducted in R). Figure 4 summarises the quantitative survey findings, depicting 95% confidence intervals around the mean for each survey item and experimental condition. To support the interpretation of these findings, we also include select verbatim quotes from participants' responses (identified by the corresponding group number, e.g., G6).

Spatial presence. There was a significant condition effect for items including 'I felt like people were next to each other' (H = 50.87, p < 0.001), 'I felt like people shared the same space' (H = 49.26, p < 0.001), 'I felt like people had a shared understanding of who was next to whom' (H = 89.41, p < 0.001), and 'I felt like people took up physical space' (H = 40.42, p < 0.001). For all items, both spatial conditions scored higher than the standard 2D layout (p < 0.001 for all). Between the spatial conditions, the 3D video representation scored higher only on the last item concerning physical space (p = 0.018). Indeed, participants noted the enhanced *"room-feel"* (G5) afforded by the 3D video.

Accordingly, participants in both spatial conditions commented that the meeting *"felt like a more natural interaction around a table"* (G6), and *"much closer to a personal meeting experience than a normal [remote] meeting"* (G6). People reported a sense of immersion during the meeting, with a G6 participant commenting, *"I forgot I was in a [remote] meeting to some extent, I was no longer looking at my own video but focusing on others instead"*.

Co-presence. For 'I felt present with the other people', there was no significant effect of condition (p = 0.759). However, participants' comments did suggest an enhanced sense of co-presence in the spatial conditions. For example, a G8 participant commented that *"we were sitting in the same place and we could turn around the table"*. Others felt that it was like *"we really sat together"* (G2) and *"were in the same room"* (G5). The sense of co-presence was further enhanced by the spatial audio, which most participants appreciated, e.g., *"the spatial audio made it feel more like I was with people"* (G8).

Attentional awareness. There was a significant effect of condition for the items 'I knew when other people were looking at me' (H = 37.6, p < 0.001) and 'I knew when other people were looking at each other' (H = 34.8, p < 0.001), suggesting that, relative to the standard 2D condition, the spatial layouts provided valuable information about people's attention or *"looking directions"* (G8). Although there was no difference between the 2D and 3D representations (p > 0.1 for both), participants' comments did suggest of the additional value of 3D video: *"it was hard to tell the direction people were facing in [2D video] compared with [3D video]"* (G6).

Attentional control. There was no condition effect for the attentional control items 'I enjoyed looking around at other people' and 'I had control of who I looked at'. Some participants did note the additional control that a 3D spatial layout afforded them, including being able to *"control [. . .] where to look"* (G2) and *"choose who to look at"* (G3).

Conversational flow. There was a significant effect of condition only for the items 'I knew when other people wanted to take a turn at speaking' (H = 6.95, p = 0.031) and 'There were awkward pauses' (H = 14.14, p = 0.001). For the former, the standard 2D condition performed worse than the spatial condition with 2D representation (p = 0.01); for the latter, it performed worse than both spatial conditions (p = 0.001 for 2D representation; p = 0.002 for 3D representation). This aligns with comments about the standard 2D condition about not knowing *"if [they're] about to interrupt someone, who's where or if someone wants to talk or not"* (G3). All other items were not significant (p > 0.1 for all).

Comfort, distraction, and representational quality. There was a significant effect of condition for the item 'It was a comfortable meeting experience' (H = 30.38, p < 0.001), with the standard 2D condition scoring better than both spatial conditions (p < 0.001 for both). There was a similar pattern for distraction ('It was a distracting meeting experience'; H = 24.99, p < 0.001), with the standard 2D condition scoring better than both spatial conditions (p < 0.001 for both).

Participants reported that the limited field of view prevented them from seeing all other participants simultaneously. Both spatial conditions required participants to rotate their view to see others, which a participant from G8 noted *"felt a little non-inclusive as most of the time I could not see them"*. This issue was exacerbated by the keyboard control of the viewing angle, which participants found *"hard"* (G1) and *"a bit strange"* (G1), noting that *"in real life you just turn your head or even just eyes and you can cover all persons sitting in same room"* (G1). This was a key reason for preferring standard 2D videoconferencing, as it enabled people to *"see everyone at once and keep track of everyone's opinions more easily"* (G5). Moreover, several participants felt that the 3D layouts made them *"dizzy"* (G4), though this may have been less strong for the 2D condition.

A key issue with the 3D representations in the live setting was their limited quality. Participants disliked the *"flicker"* (G2) and *"lots of artifacts"* (G3), and commented that the 3D video looked *"so weird and creepy"* (G2). Accordingly, one G5 participant concluded that *"it would take a lot of improvement"* to use these 3D videos in a professional context. These issues were particularly noticeable in 3D videos positioned close to participants, who saw them from the side, and in turn found them *"distracting"* (G1). These quality issues were a key reason why some participants ultimately preferred the 2D representations, as exemplified by a comment in G6: *"I wished the 2D was 3D but without the distracting shaders/effects"*.

For both 2D and 3D representations, people also noted the absence of hand movement and other body language. In G8, a participant commented that *"Some people usually use their hands when talking and seeing those kind of body movements is good which is not happening in these [. . .] experiences"*. Similarly, in G8 a participant said that the 2D representations *"[don't] really provide a presence of a human. Body language is still mostly hidden"*.

Preferences, professionalism, and norms. Ultimately, standard 2D videoconferencing was still overall preferred by a majority of participants (26 people [54%], compared to 13 people [27%] who preferred 2D video in a 3D space, and 9 people [18%] who preferred the 3D video in a 3D space.) A key reason was the standard interface's appropriateness *"especially for business"* (G2), and *"for bigger and*

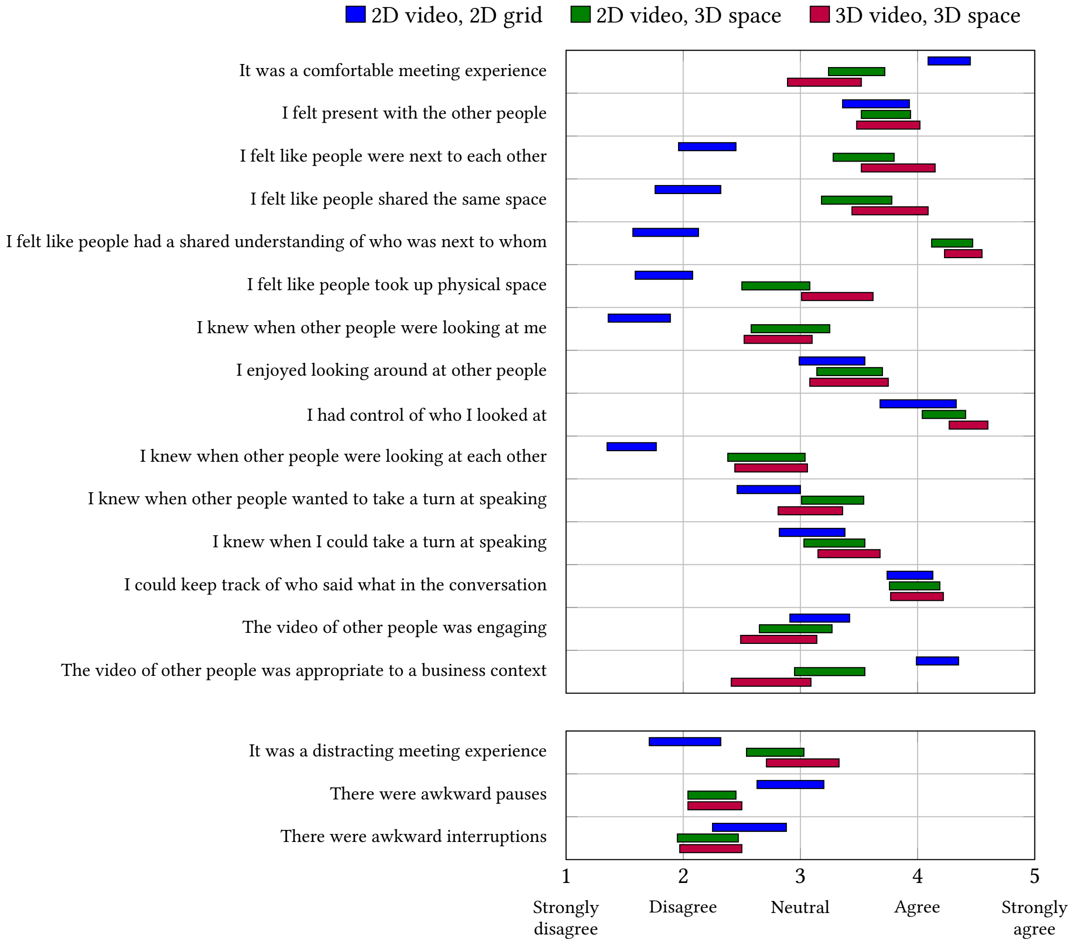
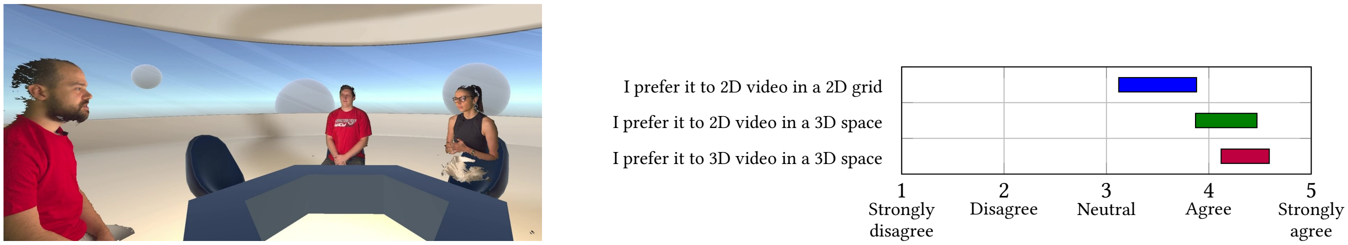


Figure 4: 95% confidence intervals for the mean score of participant agreement with the statements listed. Each of the 48 participants completed a survey immediately after completing a task in each of the three conditions, and reported their agreement with the statements using a 5-point Likert scale. The three last statements are shown separately to aid interpretation, as agreement to these statements is associated with a *negative* meeting experience.



(a) A frame of the recording that participants viewed at the end of each study group session.

(b) 95% confidence intervals for mean participant agreement with statements that compare our high-quality video with the three live conditions.

Figure 5: We concluded each group session by showing participants a recording which demonstrated offline 3D reconstruction with higher-quality 3D video than the live 3D condition. This allowed us to explore the extent to which the current 3D reconstruction quality is a limiting factor for acceptance of 3D video representations.

more serious meetings" (G5). This was due to its *"familiarity"* (G6), *"simplicity"* (G1), and absence of *"disturbing and/or distracting"* (G2) aspects. This is supported by a significant effect of condition for the item 'The video of other people was appropriate to a business context' (H = 38.12, p < 0.001), with the standard 2D condition scoring higher than both spatial conditions (p < 0.001). Thus, alongside the representational quality issues, there are potential social norms that might slow broader adoption in professional contexts.

On the above item, the spatial condition with 2D representations scored higher than the 3D representations (p = 0.035), and was overall preferred by more people (13 people or 27%). Some participants found the 2D representations were 'good enough' to support a spatial layout, despite their limited representational quality. People noted that they seemed like the *"best of both compromise with current capabilities"* (G6), or that they are *"less realistic but very easy to see"* (G2). In contrast, others still found them to be *"odd"* (G1), *"weird"* (G5), or *"confusing and distracting"* (G3). See Figure 4 for quantitative results on the effect of 2D and 3D representations on meeting dynamics and user perception.

6 IMPLICATIONS AND CONCLUSIONS

We presented a novel system that captures 3D video of people from a webcam, and renders them in a 3D virtual environment for many-to-many communication. In answering our first research question, RQ1, in the affirmative, we found that the CNN that performs depth prediction and segmentation was the most technically complex aspect. We did not directly compare to avatar representations in this study, but the 3D spatial conditions scored no lower than the 2D video grid for agreement with the statement 'I knew when other people wanted to take a turn at speaking', indicating that the 3D videoconferencing conditions were able to preserve conversational cues that are typically absent with avatar-based systems.

Our results support earlier findings [37] on the value of shared spatial context, with participants reporting a greater understanding of others' attention and better flow of talk. However, with reference to RQ2, the ability to control view direction was received with mixed results, with participants finding it helpful to understand when other people were looking at each other, but also reporting friction from the fact that it wasn't possible to see the whole scene at once. This is a key difference to most videoconferencing studies which use an interface in which one can see all participants (e.g. [13, 37]). We hypothesize that this friction is the reason why we don't see a majority preference for our first-person views over a 2D video grid. The tension is not necessarily easy to resolve, as widening the field of view for participants would remove the reason for participants to change their first-person perspective view to attend to different parts of the scene, potentially depriving other attendees of the valuable attention signal.

Turning to RQ3, we did not see a strong preference for 3D video over 2D video representations, with both judged similarly by participants for the conversational factors we measured, and more participants preferring 2D over 3D video representations when presented in a shared 3D space. This suggests that 3D video needs to improve to be accepted and to be effective. We chose to fill in areas not captured by the webcam with rays that produced a 'projector' effect akin to depictions of holograms in science fiction media. However, this choice was not treated as an acceptable workaround when participants compared it to our pre-recorded video with higher-quality reconstructions (see Figure 5 in the appendix), further suggesting that the quality of 3D video is key for acceptability and utility in shared 3D spaces.

ACKNOWLEDGMENTS

We are grateful for the valuable advice and assistance from colleagues that contributed to this work. We extend our thanks to Đorđe Nijemčević, Marko Bežulj, Vladimir Čarapić, and the Belgrade team that made our hardware prototypes possible. We thank the authors of the Perspectives study: John Tang, Kori Inkpen and colleagues, for inspiring this research. We are grateful for the work of many colleagues in Microsoft Mesh, without whom our prototypes would not have been possible: Anton Lipov, Dalibor Kukoleca, Ed Hanway, Josh Wittner, Michelle Felizardo, Shafiq Rahman, Tadas Baltrušaitis and the Human Synthetics team. We are grateful for help from Nathan Jones in running our user study, and to colleagues who gave permission for us to use their images for this paper. Finally, we acknowledge the contribution from our 48 study participants in giving their time and sharing their observations.

REFERENCES

[1] A. Amirante, T. Castaldi, L. Miniero, and S. P. Romano. 2014. Janus: A General Purpose WebRTC Gateway. In *Proceedings of the Conference on Principles, Systems and Applications of IP Telecommunications* (Chicago, Illinois) *(IPTComm '14)*. ACM, New York, NY, USA, Article 7, 8 pages. https://doi.org/10.1145/2670386.2670389

[2] Jeremy N Bailenson. 2021. Nonverbal overload: A theoretical argument for the causes of Zoom fatigue. *Technology, Mind, and Behavior* 2, 1 (Feb 2021). https://doi.org/10.1037/tmb0000030

[3] Niklas Blum, Serge Lachapelle, and Harald Alvestrand. 2021. WebRTC: real-time communication for the open web platform. *Commun. ACM* 64, 8 (jul 2021), 50–54. https://doi.org/10.1145/3453182

[4] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* 25, 11 (2000), 120–123.

[5] Judee K Burgoon, Valerie Manusov, and Laura K Guerrero. 2021. *Nonverbal communication*. Routledge, New York.

[6] Ceclia E. Ford. 2008. *Women Speaking Up: Getting and Using Turns in Workplace Meetings*. Palgrave Macmillan, New York.

[7] Cintia Ramalho Caetano da Silva, Ana Cristina Bicharra Garcia, and Joel Maurício Corrêa da Rosa. 2011. SLMeetingRoom: A model of environment to remote support meetings, oriented tasks with small groups for Second Life. In *Proceedings of the 2011 15th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, New York, 524–531. https://doi.org/10.1109/CSCWD.2011.5960122

[8] Kathleen E. Finn, Abigail J. Sellen, and Sylvia B. Wilbur (Eds.). 1997. *Video-Mediated Communication*. L. Erlbaum Associates Inc., Mahwah NJ, USA.

[9] Erving Goffman. 1963. *Behavior in public places: Notes on the social organization of gatherings*. The Free Press of Glenco, New York.

[10] Charles Goodwin. 2002. Time in Action. *Current Anthropology* 43, S4 (Aug. 2002), S19–S35. https://doi.org/10.1086/339566

[11] Chris Harrison and Scott E. Hudson. 2008. Pseudo-3D Video Conferencing with a Generic Webcam. In *2008 Tenth IEEE International Symposium on Multimedia*. IEEE, New York, 236–241. https://doi.org/10.1109/ISM.2008.12

[12] Steve Harrison (Ed.). 2009. *Media Space 20+ Years of Mediated Life*. Springer-Verlag, London. https://doi.org/10.1007/978-1-84882-483-6

[13] Zhenyi He, Keru Wang, Brandon Yushan Feng, Ruofei Du, and Ken Perlin. 2021. GazeChat: Enhancing Virtual Conferences with Gaze-aware 3D Photos. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 769–782. https://doi.org/10.1145/3472749.3474785

[14] Charlie Hewitt, Tadas Baltrusaitis, Erroll Wood, Lohit Petikam, Louis Florentin, and Hanz Cuevas Velasquez. 2023. *Procedural Humans for Computer Vision*. Technical Report MSR-TR-2023-3. Microsoft. https://www.microsoft.com/en-us/research/publication/procedural-humans-for-computer-vision/

[15] Jon Hindmarsh, Mike Fraser, Christian Heath, Steve Benford, and Chris Greenhalgh. 1998. Fragmented Interaction: Establishing Mutual Orientation in Virtual

Environments. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '98)*. Association for Computing Machinery, New York, NY, USA, 217–226. https://doi.org/10.1145/289444.289496

[16] Robert Hopper. 1992. *Telephone conversation.* Vol. 724. Indiana University Press, Bloomington, IN.

[17] Angel Hsing-Chi Hwang, Cheng Yao Wang, Yao-Yuan Yang, and Andrea Stevenson Won. 2021. Hide and Seek: Choices of Virtual Backgrounds in Video Chats and Their Effects on Perception. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 303 (oct 2021), 30 pages. https://doi.org/10.1145/3476044

[18] J. Robert Ensor. 1997. Virtual Meeting Rooms. In *Video-Mediated Communication*, Kathleen E. Finn, Abigail J. Sellen, and Sylvia B. Wilbur (Eds.). L. Erlbaum Associates Inc., Mahwah NJ, USA, 415–434.

[19] Sasa Junuzovic, Kori Inkpen, John Tang, Mara Sedlins, and Kristie Fisher. 2012. To see or not to see: a study comparing four-way avatar, video, and audio conferencing for work. In *Proceedings of the 2012 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '12)*. Association for Computing Machinery, New York, NY, USA, 31–34. https://doi.org/10.1145/2389176.2389181

[20] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26 (Jan. 1967), 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

[21] Adam Kendon. 2010. Spacing and Orientation in Co-present Interaction. In *Development of Multimodal Interfaces: Active Listening and Synchrony: Second COST 2102 International Training School, Dublin, Ireland, March 23-27, 2009, Revised Selected Papers*, Anna Esposito, Nick Campbell, Carl Vogel, Amir Hussain, and Anton Nijholt (Eds.). Springer, Berlin, Heidelberg, 1–15. https://doi.org/10.1007/978-3-642-12397-9_1

[22] Kristine M. Kuhn. 2022. The constant mirror: Self-view and attitudes to virtual meetings. *Computers in Human Behavior* 128 (March 2022), 107110. https://doi.org/10.1016/j.chb.2021.107110

[23] Jason Lawrence, Danb 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 (Dec. 2021), 16 pages. https://doi.org/10.1145/3478513.3480490

[24] Matthew Lombard, Theresa B. Ditton, and Lisa Weinstein. 2009. Measuring Telepresence: The Temple Presence Inventory. In *Proceedings of the 12th Annual International Workshop on Presence* (Los Angeles, California, USA). International Society for Presence Research, https://ispr.info/, 1–15.

[25] Paul Luff, Christian Heath, Hideaki Kuzuoka, Jon Hindmarsh, Keiichi Yamazaki, and Shinya Oyama. 2003. Fractured Ecologies: Creating Environments for Collaboration. *Human–Computer Interaction* 18, 1-2 (2003), 51–84. https://doi.org/10.1207/S15327051HCI1812_3

[26] Vassiliki Markaki and Lorenza Mondada. 2012. Embodied orientations towards co-participants in multinational meetings. *Discourse Studies* 14, 1 (Feb. 2012), 31–52. https://doi.org/10.1177/1461445611427210

[27] Lorenza Mondada. 2009. Emergent focused interactions in public places: A systematic analysis of the multimodal achievement of a common interactional space. *Journal of pragmatics* 41, 10 (2009), 1977–1997.

[28] Lorenza Mondada. 2011. The interactional production of multiple spatialities within a participatory democracy meeting. *Social Semiotics* 21, 2 (April 2011), 289–316. https://doi.org/10.1080/10350330.2011.548650

[29] Lorenza Mondada. 2013. Embodied and spatial resources for turn-taking in institutional multi-party interactions: Participatory democracy debates. *Journal of Pragmatics* 46, 1 (Jan. 2013), 39–68. https://doi.org/10.1016/j.pragma.2012.03.010

[30] NVIDIA Corporation. 2007. *NVIDIA CUDA Compute Unified Device Architecture Programming Guide.* NVIDIA Corporation, Santa Clara, CA.

[31] Vrushank Phadnis, Kristin Moore, and Mar Gonzalez Franco. 2023. Avatars in Work Meetings: Correlation Between Photorealism and Appeal. arXiv:2304.01405 [cs.HC]

[32] Alexander Raake, Markus Fiedler, Katrin Schoenenberg, Katrien De Moor, and Nicola Döring. 2022. Technological Factors Influencing Videoconferencing and Zoom Fatigue. http://arxiv.org/abs/2202.01740

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18.* Springer, Cham, 234–241.

[34] Helen B Schwartzman. 1989. *The meeting: Gatherings in organizations and communities.* Springer, New York.

[35] Abigail Sellen, Bill Buxton, and John Arnott. 1992. Using Spatial Cues to Improve Videoconferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, USA) *(CHI '92)*. Association for Computing Machinery, New York, NY, USA, 651–652. https://doi.org/10.1145/142750.143070

[36] Irwin Sobel. 2014. An Isotropic 3x3 Image Gradient Operator. Presentation at Stanford A.I. Project 1968.

[37] John C. Tang, Kori Inkpen, Sasa Junuzovic, Keri Mallari, Andrew D. Wilson, Sean Rintel, Shiraz Cupala, Tony Carbary, Abigail Sellen, and William A.S. Buxton. 2023. Perspectives: Creating Inclusive and Equitable Hybrid Meeting Experiences. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 351 (Oct. 2023), 25 pages. https://doi.org/10.1145/3610200

[38] Alex Trevithick, Matthew Chan, Michael Stengel, Eric Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. 2023. Real-Time Radiance Fields for Single-Image Portrait View Synthesis. *ACM Trans. Graph.* 42, 4, Article 135 (July 2023), 15 pages. https://doi.org/10.1145/3592460

[39] Roel Vertegaal, Ivo Weevers, Changuk Sohn, and Chris Cheung. 2003. GAZE-2: Conveying Eye Contact in Group Video Conferencing Using Eye-Controlled Camera Direction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 521–528. https://doi.org/10.1145/642611.642702

[40] Yizhong Zhang, Jiaolong Yang, Zhen Liu, Ruicheng Wang, Guojun Chen, Xin Tong, and Baining Guo. 2022. VirtualCube: An Immersive 3D Video Communication System. *IEEE Transactions on Visualization and Computer Graphics* 28, 5 (2022), 2146–2156. https://doi.org/10.1109/TVCG.2022.3150512

A PARTICIPANT DEMOGRAPHICS

Participant demographics are reported in Table 1.

Table 1: Demographics for each of the eight groups that participated in our study.

Group	Female	Male	18-29	30-44	45-59	Total
1	0	6	2	4	0	6
2	2	4	2	4	0	6
3	2	4	3	3	0	6
4	1	5	3	3	0	6
5	2	4	6	0	0	6
6	0	6	3	3	0	6
7	1	5	1	4	1	6
8	3	3	0	5	1	6
Total	11	37	20	26	2	48

B HIGH-QUALITY RECORDING AND PREFERENCES

After all study conditions, we showed participants a recorded video demonstrating a meeting using high-quality 3D representations which benefited from a multi-camera capture setup and more processing time than is available in a live meeting. We asked participants about their preference of this high-quality system relative to the three conditions that they experienced in the study session. The importance of representational quality is supported by participants' preferences for the higher-quality 3D video they were shown over all three conditions in the user study. An image of the video and results are shown in Figure 5; see the Supplementary Material video for further detail.

C TRAINING OUR DEPTH AND SEGMENTATION PREDICTION NETWORK

For the i^{th} training image with N_i pixels we define ground-truth segmentation variables $s_i := \{s_{ij}\}_{j=1}^{N_i} \subset \mathbb{R}$, with $s_{ij} := 1$ for foreground pixels j and $s_{ij} := 0$ for background pixels. For each pixel, we also define ground-truth depth values $d_i := \{d_{ij}\}_{j=1}^{N_i} \subset \mathbb{R}$ and corresponding normals $n_i := \{n_{ij}\}_{j=1}^{N_i} \subset \mathbb{R}^3$. (Where d_{ij} and n_{ij}

are not used in the loss for background pixels j.) We denote as $F_i := \sum_j s_{ij}$ the number of pixels in the ground-truth foreground.

We then train a standard U-net [33] to minimize a multitask loss $\mathcal{L}(\theta) = \sum_i \mathcal{L}(\theta; \mathcal{I}_i) + \mathcal{L}_{\text{consistency}}(\theta)$ over CNN weights θ, where $\mathcal{L}(\theta; \mathcal{I}_i)$ measures the error for the predicted segmentation, depth and normals in image $\mathcal{I}_i := \{s_i, d_i, n_i\}$:

$$\begin{aligned}\mathcal{L}(\theta; \mathcal{I}_i) = \frac{\alpha_s}{N_i} \sum_{j=1}^{N_i} \beta(\hat{s}_{ij}(\theta), s_{ij}) + \frac{\alpha_d}{F_i} \sum_{j=1}^{N_i} s_{ij} (\hat{d}_{ij}(\theta) - d_{ij})^2 \\ + \frac{\alpha_n}{F_i} \sum_{j=1}^{N_i} s_{ij} (\hat{n}_{ij}(\hat{d}_i(\theta)) - n_{ij})^2.\end{aligned} \tag{1}$$

Here $\beta(x, y)$ is the binary cross-entropy loss $-[x \log(y) + (1 - x) \log(1 - y)]$, and $\hat{d}_{ij}(\theta) \in \mathbb{R}$, $\hat{s}_{ij}(\theta) \in \mathbb{R}$ are the predictions of the neural network for depth and segmentation, respectively, for training image i given the weights θ. Our computation for the predicted normals $\hat{n}_{ij}(\hat{d}_i(\theta)) \in \mathbb{R}^3$ uses a Sobel filter [36] convolved with the depth predictions $\hat{d}_i(\theta)$ to derive estimated normals from the predicted depth. We found that including normal accuracy in the loss improved convergence for the depth and segmentation tasks, and improved smoothness of the depth predictions.

The loss $\mathcal{L}_{\text{consistency}}$ is defined over pairs of training images which are configured to have identical ground-truth depth $d_i = d_{i+1}$ and segmentations $s_i = s_{i+1}$ (for i even), but which use differing augmentations to produce two different simulated variations on the color input frame. We can then minimize the difference between the predicted depths:

$$\mathcal{L}_{\text{consistency}}(\theta) = \frac{\alpha_c}{F_i} \sum_{i \text{ even}} \sum_{j=1}^{N_i} s_{ij} (\hat{d}_{ij}(\theta) - \hat{d}_{i+1,j}(\theta))^2. \tag{2}$$

This consistency loss encourages the network to be less sensitive to noise and contrast, and thus more stable in its predictions, which helps to reduce jitter in depth estimates when the network is applied to a video sequence.

Our CNN is trained with the combined loss $\mathcal{L}(\theta)$ on a set of synthetic images of humans [14] which provide the ground-truth data for each image $\mathcal{I}_i$.