GazeChat: Enhancing Virtual Conferences with Gaze-aware 3D Photos

ZHENYI HE, New York University, USA
KERU WANG, New York University, USA
BRANDON YUSHAN FENG, University of Maryland, College Park, USA
RUOFEI DU*, Google, USA
KEN PERLIN*, New York University, USA

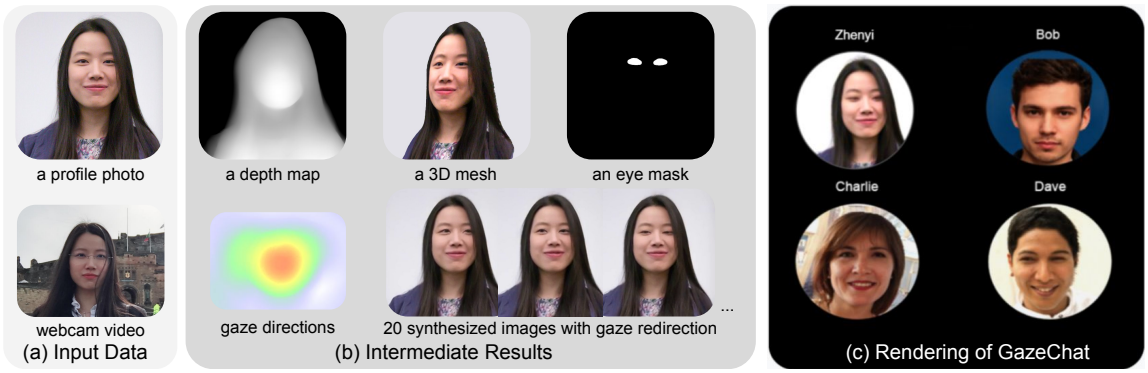

Fig. 1. We present GazeChat, a virtual conference system that leverages eye-tracking technology available with ordinary webcams to render gaze-aware 3D photos with low bandwidth requirement. (a) For each user, we take a profile photo and the webcam video as input, (b) generate a depth image, a 3D mesh reconstructed from the depth map, an eye mask and 20 synthesized gaze images, identify whom the user is looking at through gaze directions, and (c) render gaze-aware 3D photos. In (c), GazeChat shows Zhenyi (user) is looking at Dave, Dave is looking at Charlie, while Bob and Charlie are both looking at the user (Zhenyi).

Communication software such as Clubhouse and Zoom has evolved to be an integral part of many people's daily lives. However, due to network bandwidth constraints and concerns about privacy, cameras in video conferencing are often turned off by participants. This leads to a situation in which people can only see each others' profile images, which is essentially an audio-only experience. Even when switched on, video feeds do not provide accurate cues as to who is talking to whom. This paper introduces GazeChat, a remote communication system that visually represents users as gaze-aware 3D profile photos. This satisfies users' privacy needs while keeping online conversations engaging and efficient. GazeChat uses a single webcam to track whom any participant is looking at, then uses neural rendering to animate all participants' profile images so that participants appear to be looking at each other. We have conducted a remote user study (N=16) to evaluate GazeChat in three conditions: audio conferencing with profile photos, GazeChat, and video conferencing. Based on the results of our user study, we conclude that GazeChat maintains the feeling of presence while preserving more privacy and requiring lower bandwidth than video conferencing, provides a greater level of engagement than to audio conferencing, and helps people to better understand the structure of their conversation.

*Both authors contributed equally to this research.

Manuscript submitted to ACM

CCS Concepts: • **Human-centered computing → Collaborative interaction**.

Additional Key Words and Phrases: eye contact, gaze awareness, video conferencing, video-mediated communication, gaze interaction

ACM Reference Format:
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 (UIST '21), October 10–14, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3472749.3474785

1 INTRODUCTION

Virtual conferences are rapidly becoming the dominant medium for online education, remote collaboration, and casual meetings with families and friends. However, gaze awareness is not conveyed accurately in virtual conferences, so it is difficult to determine who is looking at whom from video feeds. Moreover, users often turn off their cameras in video conferences, due to low network bandwidth, shared environments, or concerns about privacy. This leads to similar challenges as for audio conferences, in which people can only see each others' still profile photos. Motivated by these limitations, we wonder: What if we could augment virtual conferences in commodity hardware by using interactive 3D photos to add proper gaze tracking?

Prior art such as True-view [83], GAZE-2 [79, 80], and MultiView [51, 52] has leveraged multiple cameras or eye trackers to obtain users' gaze directions and synthesize view-dependent images in video-mediated conversations. Additionally, seminal work such as Photoportals [7, 38], MMSpace [56–59], Sirkin *et al.* [71], and TeleHuman [35] has investigated a shared large display, kinetic displays, or a cylinder display to convey gaze awareness in remote group conversations.

Besides, commercial software like Memoji[1] and Loom.ai[2] directly map user's facial keypoints to control points of a 3D avatar for enhancing video conferencing experiences. With the recent advances in eye tracking and neural rendering, we investigate the following questions: What if we do not have access to eye trackers and special displays? How can we design a virtual conferencing system to accommodate users' concerns about privacy and limited bandwidth? Can we enhance virtual conferences with gaze-aware 3D photos for a greater level of engagement?

Towards these goals, we design, deploy, and evaluate GazeChat (Fig. 1), a virtual conferencing system, which conveys gaze awareness with augmented 3D photos. Our focus is on bringing realistic gaze awareness to virtual meetings. Different from previous work like Memoji and Loom.ai, we render the relative gaze rather than absolute gaze: rendering the eyes to reveal gaze awareness information rather than just replicating the actual eye positions. GazeChat consists of four components: a WebRTC[3] configuration to support videoconferencing and logging, a real-time eye-tracking module inspired by `WebGazer.js` [60] to recognize gaze targets, deep-learning modules to infer depth maps and synthesize novel photos by redirecting the gaze, and a rendering module implemented with the `three.js`[4]. In terms of bandwidth, GazeChat only adds a small overhead of gaze data and a one-time packet of 22 images to conventional audio conferences.

To evaluate GazeChat, we conducted four user studies with 16 remote participants (ages 21-38, 7 female and 9 male). In our analyses of video recordings, post-activity questionnaires, and post interviews, we found that GazeChat can effectively engage participants in small-group conversations by visualizing gaze awareness. gaze awareness provides more eye-contact feeling than classic video and audio conferencing, can improve the conversation experience, bring

[1] Memoji: https://support.apple.com/en-us/HT208986
[2] Loom.ai: https://loomai.com/news
[3] WebRTC: Web-based Real-Time Communications, https://www.webrtc.org
[4] three.js: JavaScript 3D library, http://www.threejs.org.

greater social presence and richness, and provide better user engagement than audio conferencing while saving bandwidth and offering privacy protection compared to videoconferencing. Our main contributions are:

(1) Conception, development, and deployment of GazeChat, a virtual conference system that can convey gaze awareness in augmented 3D photos.
(2) A low-cost, low-bandwidth, in-situ pipeline to turn users' profile pictures into animated, gaze-aware 3D photos on ordinary laptops without special hardware.
(3) Reporting evaluation results and reflections about the opportunistic use of GazeChat in virtual conferences (VC) - benefits, limitations, and potential impacts to future VC systems.
(4) Open-sourcing[5]. Our system is web-based and cross-platform compatible, making it easier to adopt for future research. We plan to make our software available to facilitate future development in VC systems with real-time neural rendering of nonverbal cues.

2 RELATED WORK

We review prior art on multi-user experience in distributed collaboration, gaze tracking, gaze redirection technology, and how gaze awareness is integrated into virtual conferences.

2.1 Multi-user Collaboration in Distributed Environments

Distributed multi-user collaboration has been widely researched from the perspective of locomotion, shared proxies, and life-size reconstruction, as well as for different purposes including communication[25], presentation[27, 77], and object manipulation [22, 43]. Your Place and Mine [72] creates experiences that allow everyone to use real walking for locomotion in collaborative VR. Three's Company [78] presents a three-way distributed collaboration system that places remote users either on the same side or around a round table. Besides, Three's Company provides non-verbal cues like body gestures through a shared tabletop interface. Remote users' arm shadows are displayed locally on a tabletop device, which is beneficial for collaborative tasks with shared objects. Tan *et al.* [77] focus on presentation in large-venue scenarios, creating a live video view that seamlessly combines the presenter and the presented material, capturing all graphical, verbal, and nonverbal channels of communication. The concept of Blended Interaction Spaces [54] is proposed to provide the illusion of a single unified space by creating appropriate shared spatial geometries. TwinSpace [66] is a generic framework discussing brainstorming and presentation in cross-reality that combines interactive workspaces and collaborative virtual worlds with large wall screens and projected tabletops. SharedSphere [42] is a wearable MR remote collaboration system that enriches a live captured immersive panorama-based collaboration through MR visualization of non-verbal communication cues.

Immersive collaborative virtual environment (ICVE) and Augmented Reality (AR) are widely used to develop new forms of teleconferencing, which often leverages multiple cameras setup and 3D reconstruction algorithms. EyeCVE [73] uses mobile eye-trackers to drive the gaze of each participant's virtual avatar, thus supporting remote mutual eye-contact and awareness of others' gaze in a perceptually coherent shared virtual workspace. Jones *et al.* [33] design a one-to-many 3D teleconferencing system able to reproduce the effects of gaze, attention, and eye contact. A camera with projected structure-light is set up for reconstructing the remote user. Billinghurst and Kato [8] developed a system that allows virtual avatars and live video of remote collaborators to be superimposed over any real location. Remote participants were mapped to different fiducial markers. The corresponding video images were attached to the marker

[5]The code is available on Github.: https://github.com/snowymo/GazeChat-Enhancing-Virtual-Conferences-with-Gaze-aware-3D-Photos

surface when markers are visible. Room2Room [62] is a telepresence system that leverages projected AR to enable life-size, face-to-face, co-present interaction between two remote participants by performing 3D capture of the local user with RGBD cameras. Holoportation [55] demonstrates real-time 3D reconstructions of an entire space, including people, furniture and objects, using a set of depth cameras. Gestures are preserved via full-body reconstruction and headset removal algorithms are designed to convey eye contact.

Prior art values the importance of immersion and non-verbal cues especially for eye contact information. Various devices are included such as cameras, markers, big screens, and headsets. We design GazeChat to investigate how we can use minimal hardware to provide essential information for distributed communication.

2.2 Eye Tracking

There have been extensive research focusing on developing eye-tracking devices such as the Tobii eye tracker [63] and on using webcams for eye tracking. Such methods typically involve an explicit calibration phase and are less accurate than infrared eye trackers [23]. One of the early-stage appearance-based methods employed video images for neural networks [5]. Recent work like Lu *et al.* introduced an adaptive linear regression model that requires sparse calibration however is sensitive to head movement [46]. Later Lu *et al.* overcame this by using synthetic images for head poses, however, need extensive calibration [47]. Another trend of research takes advantage of image salience to estimate gaze for calibration purposes [74], and salience is only a rough estimate of where a user is looking. Alnajar *et al.* designed a webcam eye tracker that supports self-calibration, though still requires users to look at the "ground truth" gaze patterns [1]. Similarly, PACE [30] and TurkerGaze [84] also predicts gaze information through webcam. Differently, GazeChat focuses on providing region-level gaze awareness instead of identifying the pixel-level gaze information. We support adapting webcam-based eye tracking technology for GazeChat use as well as eye-tracking device such as the Tobii eye tracker.

2.3 Conveying Gaze Awareness in Collaborative Tasks

Buxton started a series work researching shared space in remote collaboration since decades ago [11], especially discussed how eye contacts behave in such kind of videoconferencing [68]. Sellen *et al.* [68] present Hydra, a prototype for supporting four-way videoconferencing. Three picture-in-picture devices were used to represent three remote participants. Separate devices also native support different view points. Hydra raises the common motivation for conventional videoconferencing and has strong impacts for following work like MMSpace [56]. Additionally, deep discussion on personal space and social space was made and different principles were proposed for mediaspace, meaningspace, and meetignspace [10].

To better describe gaze information in the context, we use the term "gaze awareness" to represent the information we want to convey during videoconferencing. Gaze awareness is related to gaze direction information. It is an ability to perceive an accurate spatial relationship between an observing person and the object, that is being observed [50]. In this work, we focus on gaze awareness related to person. Many prior work visualizes gaze awareness for various purposes. "An eye for design" [15] breaks down attributes of eye movements and is inspiring. Eye-write [40] and the follow-up work "Effects of Shared Gaze" [41] emphasize the effects of gaze awareness on a shared screen. In the meantime, "Look together" [89] enhances collaborative search via gaze. Gaze is a complementary modality to be included in videoconferencing and is effective for multimodal communication [9]. EyeCVE [73] uses mobile eye-trackers to drive the gaze of each participant's virtual avatar, thus supporting remote mutual eye contact and awareness of others' gaze in a perceptually coherent shared virtual workspace. LookAtChat [26] used symbols to show gaze awareness without

changing original video feed. Jones *et al.* [33] design a one-to-many 3D teleconferencing system able to reproduce the effects of gaze, attention, and eye contact. A camera with projected structure-light is set up for reconstructing the remote user.

Inspired by previous work, GazeChat integrates gaze awareness into virtual conferences as well as separating personal space and social space. We will elaborate on the design and validate the work.

2.4 Gaze Correction and Redirection

To enable gaze redirection, the traditional approaches are typically based on 3D modeling [6, 81] by fitting eye texture and shape against 3D morphable models, but they are not ideal for handling images with eyeglasses and the high variance of facial details. Some others [13, 39, 85, 90] render a scene containing the face of a subject from a given viewpoint to mimic gazing at the camera.

Various hardware setups have been explored for gaze correction including hole in screen, long distance, and half-silver mirror. The hole in screen concept is about drilling a hole in the screen and placing a camera. Long distance uses a screen at a far distance while placing the camera as close as possible [75]. Half-silver mirror allows a user to see through a half-transparent mirror while being observed by a well-positioned camera at the same time. This idea was adapted in ClearBoard [24, 32] and Li *et al.*'s transparent display [44]. Despite their advantages in terms of system complexity and costs, such solutions are rarely used outside of laboratory due to the availability of hardware. In the meantime, quite a few 2D video-based (or image-based) approaches are proposed for eye contact including eye correction with a single camera [2, 3] and multiple cameras [14] while applying image-based approaches like texture remapping and image warp [20]. However, the technology was not sufficiently accurate to avoid visual artifacts . 3D video-based solutions including 3D reconstruction is another trend for maintaining eye contact while the head is reconstructed. RGB camera [83], depth camera [90], Kinect [39], or motion capture system [49] are used for 3D reconstruction. Eng *et al.* [18] propose a gaze correction solution for a 3D teleconferencing system with a single color/depth camera. A virtual view is generated in the virtual camera location with hole filling algorithms. Nourbakhsh also used one webcam to apply gaze redirection for one-to-one video conferencing[53]. Compared to single camera setup, multiple cameras are popularly used for providing gaze [4] in videoconferencing.

The deep learning era gave rise to various learning-based methods [19, 28, 36, 37, 61, 86]. These methods mostly train neural networks that predict the flow field for warping the eye pixels in the original image, and additional techniques such as inpainting [88] and latent space interpolation [82] have been proposed to improve improve the visual quality and redirection precision. However, these methods still fall short of reliably generating to data in the wild, with large variations in head poses, gaze angles, and lighting. A crucial reason for the lack of success for many learning-based method is the limited training data, since it is very difficult to collect a large amount of high-quality eye gaze images with correctly labeled gaze angles.

The First Order Motion model [70] estimates unsupervised keypoints from the input images and predicted a dense motion field to warp the source features to the target pose. Thanks to its unsupervised nature, this method benefits from a much larger training data set and produces results compelling visual quality. Although it is not specifically designed for the task of gaze redirection, we manage to repurpose the FOM model to generate the redirected eye gazes at desired angles.

2.5 Eye Contacts and Gaze Visualization Applied in Video-mediated Conversation

True-view [83] was implemented with two cameras (one on the left and the other on the right). The synthesised virtual camera view image at the middle viewpoint is generated to provide correct views of each other and the illusion of close proximity. GAZE-2 [79, 80] utilizes an eye tracker with three cameras. The eye tracker is used for selecting a proper camera closest to where the user is looking. GAZE-2 prototypes an attentive virtual meeting room to experiment with camera selection. In each meeting room, each user's video image is automatically rotated in 3D toward the participant he is looking at. All the video images are placed horizontally so the video image turns left or right when the corresponding camera is chosen. Likewise, MultiView [51, 52] is a videoconferencing system that supports collaboration between remote groups of people with three cameras. Additionally, MultiView allows multiple users to be co-located in one site by generating a personal view for each user even though they look upon the same projection surface, which they achieve by using a retro-reflective material. Photoportals [7, 38] groups local users and remote users together through a large display. All users are tracked and roughly reconstructed through multiple cameras and then rendered within a virtual environment. MMSpace [56–59] provided realistic social telepresence in symmetric small group-to-group conversations through "kinetic display avatars". Kinetic display avatars can change pose and position by automatically mirroring the remote user's head motions. One camera is associated with one transparent display. Both camera and display can be turned to provide corresponding video input image and output angle. Sirkin *et al.* [71] developed a kinetic video conferencing proxy with a swiveling display screen to indicate which direction in which the satellite participant was looking for maintaining gaze and gestures to mediate interaction. Instead of rendering a video image on a rectangular display, a cylinder display is proposed in TeleHuman [35] with 6 Kinects and a 3D projector.

GazeChat is designed to be used with *a minimum requirement of a laptop/PC and a single webcam*. While multi-view cameras and external hardware may yield better eye tracking and 3D rendering solutions, such systems typically require very high computational power and exclusive hardware setups. Since it is possible for users with low-cost video conferencing setup to learn to interpret gaze direction to a very high degree of accuracy [21], we decided not to apply extensive image-based manipulation on video streams but rather to focus on the design of a widely accessible online system to empower video conferencing users with real-time visualization of eye contacts.

3 GAZECHAT

In this section, we elaborate on how GazeChat is designed and implemented including eye tracking, image synthesis, network infrastructure, and rendering. We chose to start by using gaze-only information because facial expressions and head gestures are not directly helpful for augmenting gaze awareness. By open sourcing GazeChat, we envision that future work can easily leverage our framework and add new features for better remote meetings. To use the GazeChat system, users first register GazeChat by uploading their profile photos. The server then generates corresponding depth maps and synthesizes 20 images to represent each user as seen from different angles (subsection 3.2). At the beginning of the chat, the server sends each client all the synthesized profile photos and depth maps only once for real-time rendering (subsection 3.3). During the chat, each client runs an eye-tracking algorithm and classifies whom the user is looking at on screen (subsection 3.1). Meanwhile, each client sends its gaze target and audio levels to the server for the server to broadcast (subsection 3.4). The clients stream audio in a peer-to-peer manner and render gaze-aware 3D photos based on the local layouts of the profile photos.

3.1 Eye Tracking

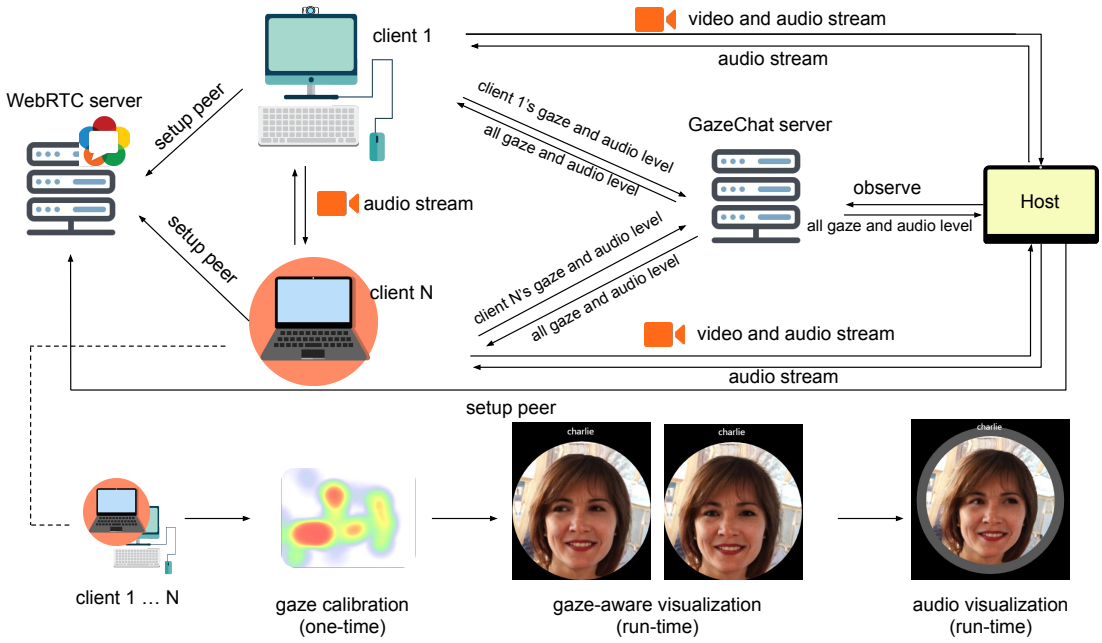

Fig. 2. GazeChat framework. GazeChat is built in spirit of peer-to-peer network. From the perspective of multi-user setup, A WebRTC server is designed for each client to talk to each other directly. A GazeChat server is provided for light-weight message transmission. A host node is implemented for observation from remote. The host can see each client's video image as well as their rendering result. From the perspective of each client, gaze-aware visualization is rendered after a profile photo is provided. Audio level is visualized via a gray circle with radius proportional to the volume.

We have considered different profile placement solutions including circles, lines, auditorium, etc, to provide spatial information that is missing in virtual meetings. From what prior art like GAZE-2, Hydra, and MMSpace concluded and given our specific user scenarios, we chose to place profile images in a 2x2 arrangement for small-group conversation. We leverage WebGazer [60] to calibrate and obtain raw gaze positions in each client. Constrained by the webcam setup within an ordinary laptop or PC, WebGazer is a state-of-the-art, off-the-shelf software solution for eye tracking. Different from WebGazer, which reports gaze coordinates, GazeChat focuses on "who is looking at whom" for video conferencing. We smooth the coordinate outputs from WebGazer with 1€ Filter [12] and then classify the data to understand which client is being looked at. When a user looks at no one or at another screen, our algorithm refrains from visualizing their gaze. GazeChat also supports external eye-tracking devices such as the Tobii tracker bar, which provides similar gaze coordinate information.

Our system expects users to reach an accuracy of 80% during the calibration session and ensure that the size of the face is larger than 25% of the video frame. To improve accuracy, we fine-tune the parameters of 1€ Filter and video stream placement with a Tobii eye tracker.

First, we tune parameter *mincutoff* in 1€ Filter to ensure that gaze coordinates are not jittering and *beta* to ensure that the result is not introducing too much latency. A straight line is used as the gaze path for the fine-tuning step. An animated dot moves from the start to the end of the line. The animated dot is a circle with a 10-pixel radius (from experimental data). We move our cursor to follow the dot several times and record the cluster of cursor positions.

(a) GazeChat with area ID annotated (b) Result of identifying gaze area

Fig. 3. Eye tracking accuracy. (a) shows how we annotate each area. Area 1 — 4 are representing gaze-aware 3D photos for all participants including self. Area 0 is the area classified as the viewer is looking at theirself video image. (b) shows the error rate for each area.

The sum of the average distance between cursor and animated dot is set as a reference value. We next move our gaze to follow the animated dot and apply the same calculation. *Mincutoff* is tuned to ensure that the distance of gaze is comparable to the reference value. *Beta* is tuned to ensure that average latency is less than 5 ms between the raw gaze position and the smoothed gaze position. We set *mincutoff* to be 0.3 and *beta* to be 0.3 for smoothing the raw gaze positions.

Second, we adjust the placement of the video stream. We use a Tobii tracker bar for ground truth. Our goal is to reach over 90% accuracy in identifying which area is being gazed at. All participants are rendered as gaze-aware 3D photos, including viewer's self. In the meantime, each viewer's video image is rendered on the top left, only visible to that viewer. The viewer's video image will be annotated with a red frame when the viewer's face is not in a good position for eye-tracking, so the viewer can adjust their pose accordingly. We chose to detect whether the smoothed gaze coordinates is in the area of any given gaze-aware 3D photo or other areas like the viewer's self image. We tune the size of each gaze-aware 3D photo and the distance between neighbour representations (horizontally and vertically) to maximize accuracy. Thus, the size of the face is neither too large nor too small and good for gaze detection. Distance between 3D photos is tuned according to the human's fovea and peripheral vision and the precision of gaze detection as well so users can notice the changes on neighbour profiles while gazing. We calculate the accuracy of our algorithms based on the coordinates reported by the Tobii eye tracker. The error rate is calculated as the ratio of incorrect area identification to total area identification. As Fig.3 shows, our algorithms report 92.5% accuracy on average. Next, placement data is recorded with the size of the screen so the algorithms behaves the same on different screens.

3.2 Image Synthesis

Our goal is to automatically synthesize images of our participants with their eye gaze fixating at different angles. We ask the participants to provide a profile picture of themselves beforehand, which is commonly done by users of online conferencing platforms. In order to synthesize realistic images where the participant's eyes are looking at various gaze angles, we utilize the pretrained FOM model [70]. A crucial requirement of the FOM model is that the synthesized movements of the static input image are modeled after the image frames from a "driving video". We capture a selfie video using a smartphone and use it as the driving video for the profile image of all participants. Specifically, we generate the

driving video such that the eyes appear to rotate along a full circle, and we extract 20 frames where the gaze is looking at the following angles: $\{\phi_i = \frac{\pi}{10} i\}, i = 1, 2, ..., 20$. Example synthesized gaze are shown in Fig. 4.

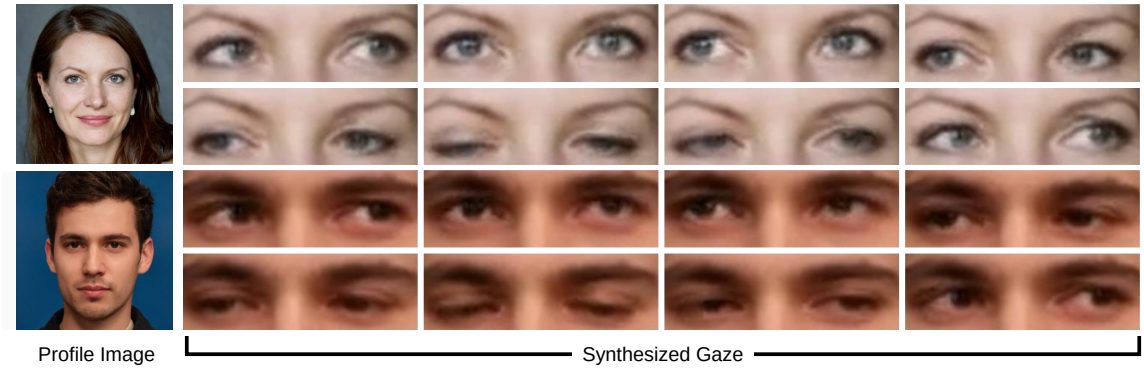

Fig. 4. On the left we show example profile images, which are photo-realistic "deep-fake" portraits produced by StyleGAN [34]. We synthesize their gaze images as described in subsection 3.2, and on the right we present the zoomed-in results around the eye region.

3.3 Rendering

As shown in Fig. 5, our renderer uses intermediate results of a depth image, synthesized gaze images, and an eye mask to obtain a gaze-redirected 3D photo. We infer the depth (relative distance from the camera) of each image pixel from the provided profile picture using the MiDaS model [64]. To create the 3D mesh, we use the screen-space depth meshing technique described in [17, 29]. This technique relies on a densely tessellated quad, in which each vertex is displaced based on the re-projected depth value. We rotate the mesh based on users' gaze direction. We also create an eye mask image based on the facial landmark points [65] to constrain the texture sampling to be within our region of interest.

During rendering, we need to actively query the real-time gaze point, compute the corresponding gaze angle, and then display the nearest synthesized gaze. Let the current gaze point to be at (x, y) and the screen center to be coordinate system origin at $(0, 0)$. We compute the rotation angle as $\phi_{cur} = \arctan(\frac{y}{x})$ and its corresponding index $i_{cur} = \frac{10 \cdot \phi_{cur}}{\pi}$. If the gaze position is close to the profile's center, we use the original image's eyes in the rendering to achieve the effect of looking straight forward. If the gaze is far from the center, from the set of synthesized images $\{\text{Img}_1, \text{Img}_2, ..., \text{Img}_{20}\}$ and their corresponding gaze angles $\{\phi_i\}$ we pick the i-th and $(i+1)$-th image such that $i = \lfloor i_{cur} \rfloor$ and $i + 1 = \lceil i_{cur} \rceil$. As demonstrated in Fig. 6, we use alpha blending to obtain the final image $\text{Img}_{\text{final}} = (1 - \alpha)\text{Img}_i + \alpha\text{Img}_{(i+1)}$, where $\alpha = i_{cur} - i$. The rotation of the profile mesh follows the gaze by rotating $0.1 \cos \phi_{cur}$ and $0.1 \sin \phi_{cur}$ alone the x and y axes respectively.

3.4 GazeChat Infrastructure and Eye-contact Variance

As Fig. 2 demonstrates, GazeChat employs a WebRTC server as well as peer-to-peer networking. For each newly-joined client, it talks to the WebRTC server (including Internet Connectivity Establishment server and Signaling) first to establish peer-to-peer connection with existing clients. Hence, the clients can send and receive video (optionally) and audio streams with each other. Next, the GazeChat server maintains the identifier of each client after the WebRTC connection is established. For each client, gaze awareness data and audio level information are processed locally and sent to the GazeChat server for further use. Afterwards, the server broadcasts the information to all of the clients and

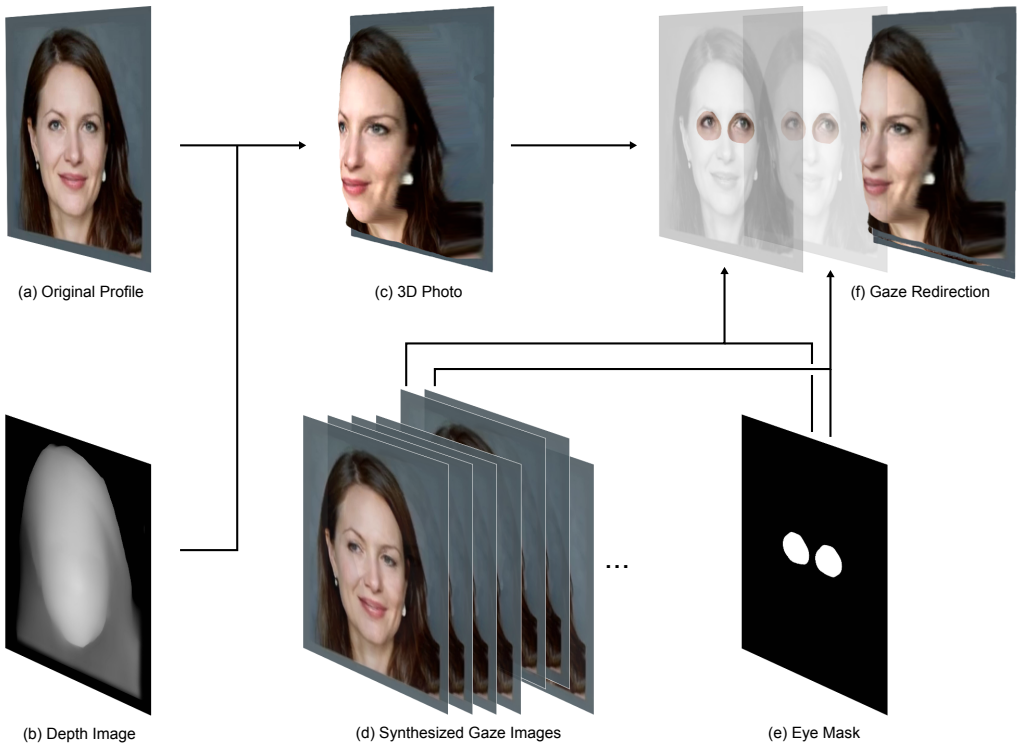

Fig. 5. The construction of the gaze-aware 3D photo. We first generate the 3D photo in (c) using the original profile as texture and the depth image as geometry reference. We then select proper synthesized gaze images for 3D photo's eye texture blending. The chosen photos' eye regions are constrained using the eye mask and blended to replace the original 3D photo's eyes for gaze redirection.

the each client renders the result accordingly. Audio level is visualized as a gray circle behind each gaze-aware 3D photo, with radius proportional to volume.

Regarding rendering gaze awareness data, it is constructed as a pair of source viewer ID and destination viewer ID (which could be empty because the source viewer may not be looking at anyone). GazeChat will locally calculate the angle for each viewer according to gaze awareness data. We further design two variances in terms of eye contact rendering. We later note two variances as GazeChat(Eye) and GazeChat(Third) in the figures. The first one means GazeChat with EyeContact. When a user is gazing at the viewer, GazeChat will render this user's synthesized image as looking directly to the viewer. Thus the viewer feels as though this user is looking into her/his eyes. The other one is GazeChat with ThirdPerson perspective. GazeChat treats all gaze information the same, so that the gaze angle is calculated according to the source viewer placement and destination viewer placement. That is to say, when a user is gazing at the viewer, GazeChat will render the synthesized image as looking toward the viewer's synthesized image. When a user is looking at no one, the user will be rendered as looking at her/his corresponding corner.

Host mode (Fig. 2) is designed for remote management and evaluation use. The host can see each participant's video image (given consent) and their rendering results as well as talk to all participants. Also, all gaze awareness and audio level data can be recorded by the host for further analysis. Without host mode, GazeChat does not require participants to stream video images.

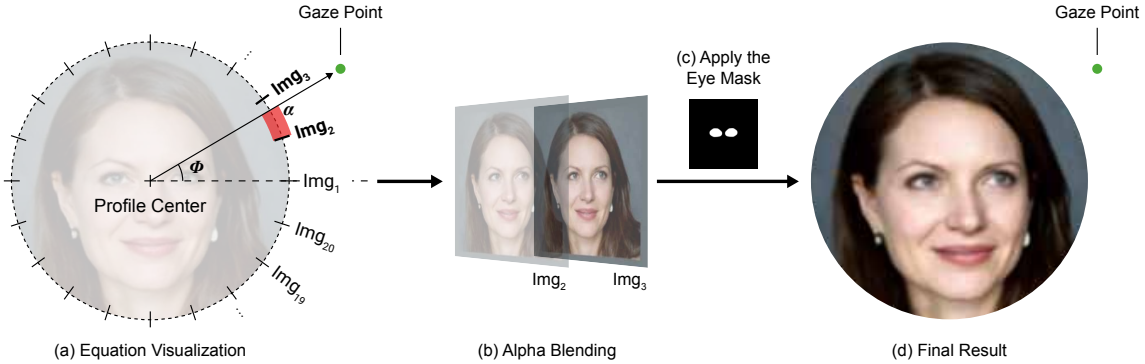

Fig. 6. Render the eyes to follow the gaze point. (a) shows how the equations in subsection 3.3 determine the images used for eye texture blending and their blending ratio. The gaze point in this example results in Img_2 and Img_3 with a blending ratio of $1-\alpha:\alpha$. The eye regions of the blended images in (b) are chosen using an eye mask in (c) to replace the original 3D photo's eyes for gaze redirection. (d) demonstrates the final rendering result in which the person looks at the gaze point.

4 EVALUATION: SMALL-GROUP DEBATE

We conducted a remote user study to examine how our system works in the wild and how the different eye-contact design perform in terms of conversation, subjective feedback, and user preferences, compared with a traditional video-conferencing setup where no gaze information is visualized, and a Clubhouse-like audio conferencing condition where only static profile is provided. The user study follows a within-subject design in three conditions: videoconferencing(VIDEO), audio conferencing(AUDIO), and our system: GazeChat(Eye) and GazeChat(Third). The conditions were counterbalanced to avoid bias in the following combinations: VIDEO–AUDIO–GazeChat, AUDIO–GazeChat–VIDEO, GazeChat–VIDEO–AUDIO. The DVs were conversation experience defined in Sellen's work [69], user experience defined in Schrepp *et al.*'s work [67] and Hung *et al.*'s work [31], and selected Temple Presence Inventory (TPI) questions [45]. We processed the data through an analysis of variance (ANOVA). All tests for significance were made at the $\alpha = 0.05$ level. The error bars in the graphs show standard error.

4.1 Participants and Apparatus

We recruited a total of 16 participants at least 18 years old with normal or corrected-to-normal vision (7 females and 9 males; age range: $21-38$, $M = 24.5$, $SD = 4.7$) via social media and email lists. The participants have a diverse background from both academia and industry. None of the participants had been involved with GazeChat before. We assign participants into four 4-person groups for the user study. Following COVID-19 regulations, the study was conducted remotely in personal homes. Participants used their personal computers with a webcam, visited the website we provided through Google Chrome browser, and experienced different conditions as instructed by the host. We instructed participants to take the user study in a quiet and brightly lit room where faces in the webcam are clearly visible from the background. For the duration of the study, participants' behavior, including their conversations and video streams, were observed and recorded with their consent.

4.2 Procedure

Our remote user study is scheduled using conventional calendar and videoconferencing tools (Zoom). Once all participants were online, the host briefly introduces GazeChat system with a tutorial video and asks all participants to fill in consent forms. After the tutorial, the host instructs all participants to enter a condition in the user study website (https://chat.3dvar.com). Participants are instructed to mute their video and audio streams in Zoom to prevent echoing and save networking bandwidth. Meanwhile, the participants can still follow the host's instructions from Zoom and the host can monitor the experiments with the **host mode** in GazeChat. The user study session of each condition consists of three parts: gaze calibration (~5 min), warm-up conversation (~3 min), and a debate (~10 min). We next describe the three parts in more detail:

Gaze calibration. Participants are required to first calibrate their gaze individually. Our system adapts the calibration procedure of WebGazer[60]: A box rendered around participant's face mesh turns green when the participant is at the center of the camera view and close enough. Next, the participant calibrates 9 points on the screen and the accuracy of gaze point is reported. We suggest that the participant proceed after reaching over 80% pixel-level accuracy.

Warm-up conversation. To reduce the novelty effect caused by our system, we had a "warm-up conversation" to familiarize users with the system and current condition. In the beginning of each condition, researchers briefly describe how gaze information is visualized for the current condition. Later on, participants pick a topic and one by one give a short speech for around 30 seconds.

Debate. After the warm-up conversation, the group is instructed to debate. We chose a debate as the user study task since it has key features: individual speech and group discussion.

At the end of each study session, we ask the participants to fill an online questionnaire about conversation experience, user experience, and TPI [45] with a 7-point Likert scale for the condition they just completed (~5 min). Hence, the study session of each condition lasts for around 30 to 45 minutes. At the end of all the three conditions, we ask the participants to fill demographic information, scale of usability in general, and rank the conditions. Lastly, participants are interviewed about LookAtChat, reasons for their ranking, and gave suggestions. On average, the experiment takes about 100 to 120 minutes in total.

4.3 Results

We validated that the data is normally distributed through the Shapiro-Wilk test and satisfies the assumptions of an analysis of variance (ANOVA). All tests for significance were made at the $\alpha = 0.05$ level. The error bars in the graphs show the standard error. Symbol * means $p <= .05$, ** means $p <= .01$, and *** means $p <= .001$.

TPI - Social Presence and Richness. The results for the ratings of social presence and richness questions that have significant results over all conditions are illustrated in Fig. 7. For the question *"How often did you want to or did you make eye-contact with someone you saw/heard?"* ($M_{GazeChat(Eye)} = 5.1, M_{GazeChat(Third)} = 5.3, M_{VIDEO} = 4.3$, and $M_{AUDIO} = 3.1$), a one-way within-subjects ANOVA was conducted to test the influence of conditions on the ratings. The main effects for condition ($F(3, 45) = 8.9, p < .001^{***}$) was significant. Post hoc t-tests with Holm correction showed a significant difference between GazeChat(Eye) and AUDIO as well as GazeChat(Third) and AUDIO ($t(15) > 1.04, p < .001^{***}$) with a 'large' effect size (Cohen's d > 1). GazeChat also has larger ratings than VIDEO while

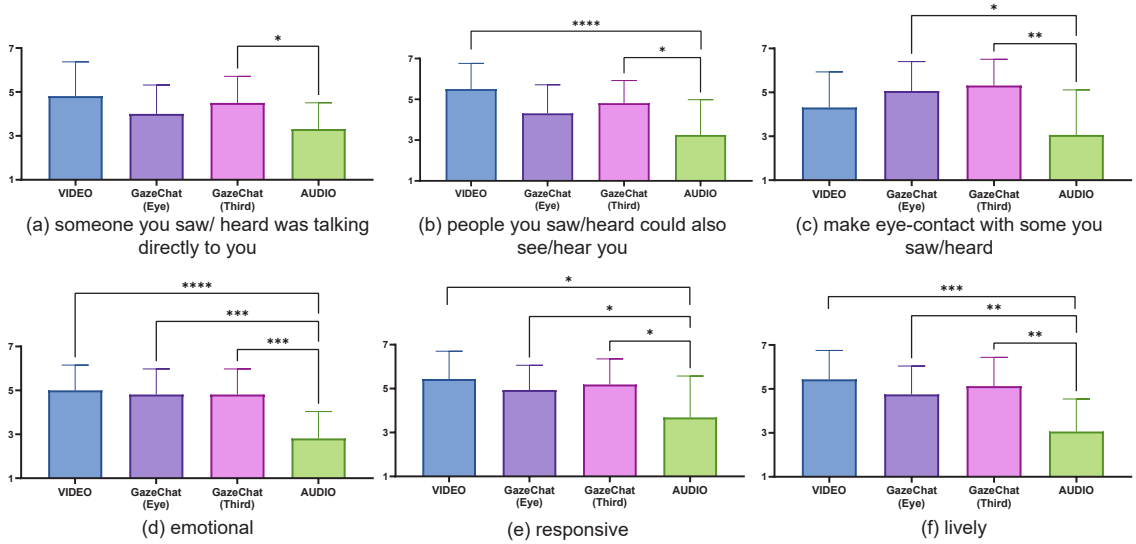

Fig. 7. Summary of results regarding TPI between AUDIO, VIDEO, and GazeChat in variance of GazeChat(Eye) and GazeChat(Third). The error bars in the graphs show the standard error. We reported 6 of them that have significant effects through ANOVA and all of them have significant impacts in post hoc tests, $p \leq .05$, **: $p \leq .01$, ***: $p \leq .001$. GazeChat(Third) condition has notably better scores than AUDIO for all TPI questions. VIDEO has better scores than GazeChat except for the feeling of eye-contact in (c).

not significant. The results indicate that GazeChat with EyeContact design and ThirdPerson design provided notably more eye-contact experience than AUDIO.

The results for the questions *"How often did you have the sensation that people you saw/heard was talking directly to you" and "... could also see/hear you?"* are similar (Fig.7(a) and (b)). VIDEO has highest scores among all. The main effects for condition ($F(3, 45) = 4.22, p < .01^{**}$) was significant while only GazeChat(Third) is notably better than AUDIO from post hoc t-tests ($t(15) > 3.335, p < .027^{***}$). The results indicate that GazeChat has better sensation of talking and feeling of being seen/heard than AUDIO though lower score than VIDEO. GazeChat with Third Person design provided notably more sentation than AUDIO.

The results for social richness questions are similar (Fig.7(d — f). The main effects for condition were significant. VIDEO has highest ratings, GazeChat has slightly lower ratings than VIDEO, while both VIDEO and GazeChat are better than AUDIO. Post hoc t-tests demonstrate that GazeChat with two variance were both notably better than AUDIO. The results indicate that participants experience similar social richness from GazeChat compared to VIDEO.

User Engagement and User Experience. The results for the ratings of user engagement experience questions that have significant results over all layouts are illustrated in Fig.8. Results show that GazeChat(Eye) and GazeChat(Third) are interesting ($M_{GazeChat(Eye)} = 5.4$, $M_{GazeChat(Eye)} = 5.3$), as well as novel ($M_{GazeChat(Eye)} = 5.7$, $M_{GazeChat(Eye)} = 5.5$). A one-way within-subjects ANOVA was conducted to test the influence of conditions on the ratings. The main effects for condition ($F(3, 45)_{GazeChat(Eye)} = 18.38, F(3, 45)_{GazeChat(Third)} = 9.6, p < .001^{***}$) were both significant. Post hoc t-tests showed a significant difference between GazeChat and AUDIO ($t(19) > 4.4, p < .001^{***}$) with a 'large' effect size (Cohen's d > 1.37). The results indicate that GazeChat provided notably more user engagement than AUDIO.

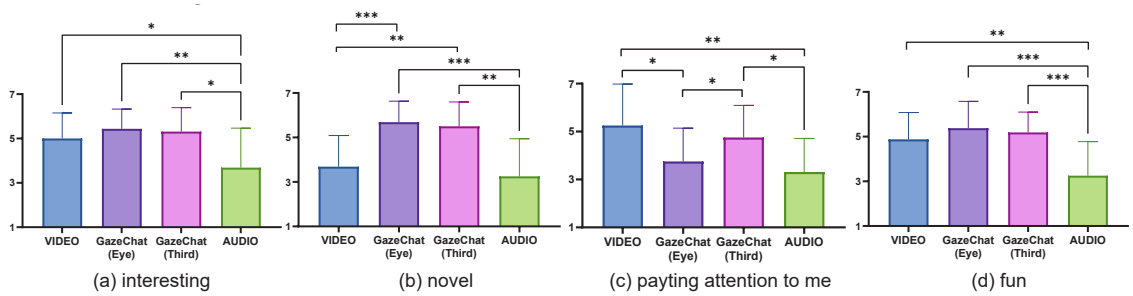

Fig. 8. Summary of significant results regarding User Engagement, Conversation Experience, and User Experience between AUDIO, VIDEO, and GazeChat. We reported 4 of them that have significant effects through ANOVA and significant impacts in post hoc tests, p <= .05, **: p <= .01, ***: p <= .001. GazeChat with EyeContact and ThirdPerson condition has notably better scores than AUDIO.

Similar to questions regarding user experience, GazeChat has significantly better scores than AUDIO and not significant but better ratings than VIDEO (Fig. 8(d)).

Conversation Experience. The result for the ratings of question *"I knew when people were listening or paying attention to me."* is shown in Fig. 8(c) with $M_{VIDEO} = 5.3, M_{GazeChat(Eye)} = 3.8, M_{GazeChat(Thrid)} = 4.8$. A one-way within-subjects ANOVA was conducted to test the influence of condition on the ratings. The main effects for layout ($F(3, 45) = 9.25, p < .001^{***}$) were significant. Post hoc t-tests with Holm correction showed a significant difference between GazeChat (Third) and AUDIO ($t(19) = 3.47, p = .007^{**}$) with a 'large' effect size (Cohen's d = .87). The results indicate that GazeChat with Third Person design provided notably more feeling of attention then AUDIO.

Preference and Subjective Feedback. 10 participants preferred the VIDEO and 6 participants preferred GazeChat (4 preferred the GazeChat with EyeContact, and 2 preferred GazeChat with ThirdPerson). Participants who preferred VIDEO because "*It makes me feel closer to the other participants.*" (P11, F). However, participants stated that it is common that it is "*ambiguous that who is talking to whom*" so that "*finding the right time to speak is hard.*" (P15, M). No participants rate AUDIO since "*people don't become immediately aware that I am talking to them, and they would interrupt me to confirm with me, or I need to repeat myself since they were not listening*" (P2, F).

Regarding GazeChat, there is no significant difference between GazeChat with EyeContact and GazeChat with ThridPerson from participants' feedback. Participants acknowledged that GazeChat protects privacy by displaying only the augmented photo while "*maximizing the feeling of presence with the vividly rendered dynamic photos.*"(P9, M). The gaze-aware 3D photo also helps to clarify who is talking to whom, thus makes the conversation more straightforward. Participants can understand "*the logical structure of the conversation*" (P2, F) better since they manage to "*know who is talking to whom, so the conversation is simpler to follow*" (P2, F). GazeChat allows the participants to speak "*in turns*", "*make immediate comments*", and "*ask questions*" (P16, M) easier by introducing the eye contact into the conferencing. Moreover, P(10) commented that under some content, things like facial expression/body movement might be trivial but distracting, and by only displaying the gaze-aware 3D photos2, GazeChat improves "*how we are allocating our attention to the critical stuff regards the conversation.*"

5 DISCUSSION

The results from the evaluation show that GazeChat provides a greater sense of presence, better privacy protection, and enhanced social engagement, but also suggest limitations and opportunities for improvement.

5.1 Insights

We conducted four interviews with each group of participants to learn their thoughts on the online conferencing systems they used. Combined with the survey results, we summarize that stability, privacy protection, efficiency, and social engagement are four essential things they care about when using an online conferencing system.

Balance stability and social engagement. There is a tradeoff between stability and social engagement among existing online conferencing systems. For example, disabling video saves bandwidth but results in a more limited "*feeling of intimacy*" (P11, F) with other participants, while full video conferencing is highly engaging but could cease to be functional under poor internet conditions. Finding an intermediate step between full video and pure audio conferencing, such as hiding the actual video feed while visualizing only meaningful non-verbal cues, is a way to balance the system's stability and social engagement. In GazeChat, we require low bandwidth to pass around only the audio and gaze position, while "*maximizing the feeling of presence with the vividly rendered dynamic profiles.*" (P9, M). Future online conferencing systems should allow participants to determine how detailed is the information they want to pass to and receive from others based on their internet speed.

Replace full video feed with other visual cues for privacy protection. Turning on the camera is often required in small online group conferencing for better communication and engagement. However, privacy concerns such as an unwillingness to expose personal information make participants feel "*stressful and awkward*" (P3, F) to enable the webcam, especially when they are "*in bad shape*" (P3, F) or "*in a messy room*" (P1, M). Future online conferencing systems should transmit enough information for others to understand whether/to whom the participant is paying attention and how he/she is reacting without requiring the full use of video. Displaying only an augmented photo while embedded eye contact as cues for who is talking/listening to whom, GazeChat can satisfy users' privacy concerns and the need for attention information at the same time.

Clarifying who is talking/listening to whom for efficiency. Clarifying the relationship of who is talking or listening to whom can greatly improve online conferencing efficiency. It can make the conversation simpler to follow and can better deliver the right message to the right person. The relationship between speaker and listener can be inferred from users' gaze information because people tend to look at the 3D photo of the person they are paying attention to. Conferencing systems can improve communication efficiency by include gaze awareness via enhancing"*intuitive conversation*". This allows the user to"*pick up a direct conversation with a person smoothly*" since he/she "*knows I am staring at him*". "*Things should work the same way as if we were having offline meetings.*" (P8, F). With gaze awareness, GazeChat helps participants better understand the structure of the conversation, "*reduce interruptions*" (P15, M), and "*encourage smooth interactions*" (P14, M). Future conferencing systems could enable different kinds of visual cues for gaze awareness to help participants process gaze information, such as highlighting the user profile to show eye contact, so as to better understand who is talking to whom.

Filter out irrelevant information to improve efficiency. The amount of received attention is correlated with the amount of information provided. Sometimes showing too much trivial content, such as "*displaying the non-speakers in a lecture*

setting might distract the listener from what truly matters" (P13, F). Filtering out irrelevant information may improve conversation efficiency. Which information is irrelevant depends on the purpose of the meeting. Some participants mentioned that GazeChat is suitable for scenarios like debates because participants focus more on "*the speaker's opinion and to whom he is talking*" (P9, M) rather than facial expressions or body movements. It would be interesting to explore how the value of any given detail changes depending on the context of the conversation. Future online conferencing systems could give users the freedom to select which information to emphasize and which to ignore.

5.2 Limitations and Future Direction

The results of the user study show that GazeChat is novel and interesting while providing more eye-contact information, but also reveal limitations and potential for improvement.

Limited visual cues. Since GazeChat is designed to study how gaze information influences online conferencing experience, only the eyes and the head of the profile are animated while the other parts are static. This rendered result might look "unnatural" for some users. The profile image might become more natural were it to include some other subtle movements and expressions. Moreover, participants' real facial expressions and the body movements, such as nodding, shaking heads, and raising hands, are embedded within the video feed. Such information can greatly facilitate interactions and help participants to better understand other persons' immediate responses. While the current version of GazeChat only focuses on gaze awareness, future research may build upon the system to keep track of the actual head position [48], body movements [76], and hand gestures [87], then incorporate these features into the system.

Unstable gaze tracking. Currently, the gaze tracking algorithm only supports one user in front of the webcam. Accuracy is limited by the individual calibration procedure, the user's position in front of the camera, and environment lighting. It would be interesting in future work to explore the addition of other devices with cameras, such as cell phones, to provide assistance in gaze tracking.

Limitations of user study. In the user study, as the participants' ages spanned $21 - 38$, the results may not generalize to other populations, such as young students or older adults, who may prefer more or less eye contact in video conferences. A long-term study in the future might be needed to better assess the longitudinal impact of the proposed system on the user's teleconferencing experience and behavior.

6 CONCLUSION

In this paper, we introduced GazeChat, a web-based group chat system which enhances virtual conferences with gaze-aware 3D photos. Motivated by users turning off cameras in conventional video conferences due to low bandwidth or concerns about privacy, we propose to track users' gaze directions with a webcam, transmitting only a small overhead of gaze awareness information in addition to audio streams, rendering gaze-aware 3D photos to convey "who is looking at whom". We conducted a remote user study of 16 participants to examine the benefits and limitations of the interface, as well as the potential impacts of varied user engagement and experience. The quantitative results indicate that GazeChat provides a better feeling of eye-contact than VIDEO and AUDIO conditions and a significantly better experience in terms of user engagement than AUDIO. Among all conditions, participants found GazeChat to be novel, interesting, and fun.

As an initial effort to use gaze-aware 3D photos for virtual conferences, we believe our open-sourced work may inspire more designs that convey nonverbal cues for remote conversations. Such features may eventually be integrated

with video conferencing software or mixed reality social platforms [16] to increase social engagement and to improve conversational experience.

REFERENCES

[1] Fares Alnajar, Theo Gevers, Roberto Valenti, and Sennay Ghebreab. 2013. Calibration-Free Gaze Estimation Using Human Gaze Patterns. In *Proceedings of the IEEE International Conference on Computer Vision*. 137–144. https://doi.org/10.1109/ICCV.2013.24

[2] Russell Lennart Andersson and Homer H Chen. 1997. Method for Achieving Eye-to-Eye Contact in a Video-Conferencing System. US Patent 5,675,376.

[3] Russell L Andersson, Tsuhan Chen, and Barin G Haskell. 1996. Video Conference System and Method of Providing Parallax Correction and a Sense of Presence. US Patent 5,500,671.

[4] Mark Ashdown, Kenji Oka, and Yoichi Sato. 2005. Combining Head Tracking and Mouse Input for a GUI on Multiple Monitors. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. 1188–1191. https://doi.org/10.1145/1056808.1056873

[5] Shumeet Baluja and Dean Pomerleau. 1994. Non-Intrusive Gaze Tracking Using Artificial Neural Networks. In *Advances in Neural Information Processing Systems*, J. Cowan, G. Tesauro, and J. Alspector (Eds.), Vol. 6. Morgan-Kaufmann. https://proceedings.neurips.cc/paper/1993/file/19b650660b253761af189682e03501dd-Paper.pdf

[6] Michael Banf and Volker Blanz. 2009. Example-Based Rendering of Eye Movements. *Computer Graphics Forum* 28, 2 (2009), 659–666. https://doi.org/10.1111/j.1467-8659.2009.01406.x

[7] Stephan Beck, Andre Kunert, Alexander Kulik, and Bernd Froehlich. 2013. Immersive Group-to-Group Telepresence. *IEEE Transactions on Visualization and Computer Graphics* 19, 4 (2013), 616–625. https://doi.org/10.1109/TVCG.2013.33

[8] Mark Billinghurst and Hirokazu Kato. 2000. Out and About—real World Teleconferencing. *BT Technology Journal* 18, 1 (2000), 80–82. https://doi.org/10.1023/A:1026582022824

[9] Richard A Bolt. 1980. "Put-That-There" Voice and Gesture at the Graphics Interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*. 262–270. https://doi.org/10.1145/800250.807503

[10] Bill Buxton. 2009. Mediaspace-Meaningspace-Meetingspace. In *Media Space 20+ Years of Mediated Life*. Springer, 217–231. https://doi.org/10.1145/3170427.3173033

[11] William Buxton. [n.d.]. Telepresence: Integrating Shared Task and Person Spaces. In *Proceedings of Graphics Interface*, Vol. 92.

[12] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 2012. 1€ Filter: a Simple Speed-Based Low-Pass Filter for Noisy Input in Interactive Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2527–2530. https://doi.org/10.1145/2207676.2208639

[13] A. Criminisi, J. Shotton, A. Blake, and P. Torr. 2003. Gaze Manipulation for One-to-One Teleconferencing. *Proceedings Ninth IEEE International Conference on Computer Vision* (2003), 191–198 vol.1. https://doi.org/10.1109/ICCV.2003.1238340

[14] Antonio Criminisi, Jamie Shotton, Andrew Blake, and Philip HS Torr. 2003. Gaze Manipulation for One-to-One Teleconferencing. In *ICCV*, Vol. 3. 13–16. https://doi.org/10.5555/946247.946637

[15] Sarah D'Angelo and Darren Gergle. 2018. An Eye for Design: Gaze Visualizations for Remote Collaborative Work. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3173574.3173923

[16] Ruofei Du, David Li, and Amitabh Varshney. 2019. Geollery: a Mixed Reality Social Media Platform. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 685. https://doi.org/10.1145/3290605.3300915

[17] Ruofei Du, Eric Turner, Maksym Dzitsiuk, Luca Prasso, Ivo Duarte, Jason Dourgarian, Joao Afonso, Jose Pascoal, Josh Gladstone, Nuno Cruces, Shahram Izadi, Adarsh Kowdle, Konstantine Tsotsos, and David Kim. 2020. DepthLab: Real-Time 3D Interaction With Depth Maps for Mobile Augmented Reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 829–843. https://doi.org/10.1145/3379337.3415881

[18] Wei Yong Eng, Dongbo Min, Viet-Anh Nguyen, Jiangbo Lu, and Minh N Do. 2013. Gaze Correction for 3D Tele-Immersive Communication System. In *IVMSP 2013*. IEEE, IEEE, 1–4. https://doi.org/10.1109/IVMSPW.2013.6611942

[19] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. 2016. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European conference on computer vision*. Springer, 311–326. https://doi.org/10.1007/978-3-319-46475-_20

[20] Jim Gemmell, Kentaro Toyama, C Lawrence Zitnick, Thomas Kang, and Steven Seitz. 2000. Gaze Awareness for Video-Conferencing: a Software Approach. *IEEE MultiMedia* 7, 4 (2000), 26–35. https://doi.org/10.1109/93.895152

[21] David M Grayson and Andrew F Monk. 2003. Are You Looking at Me? Eye Contact and Desktop Video Conferencing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 10, 3 (2003), 221–243. https://doi.org/10.1145/937549.937552

[22] Raja Gumienny, Lutz Gericke, Matthias Quasthoff, Christian Willems, and Christoph Meinel. 2011. Tele-Board: Enabling Efficient Collaboration in Digital Design Spaces. In *Proceedings of the 2011 15th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 47–54. https://doi.org/10.1109/CSCWD.2011.5960054

[23] Dan Witzner Hansen and Arthur EC Pece. 2005. Eye Tracking in the Wild. *Computer Vision and Image Understanding* 98, 1 (2005), 155–181. https://doi.org/0.1016/j.cviu.2004.07.013

[24] Beverly L Harrison, Hiroshi Ishii, Kim J Vicente, and William AS Buxton. 1995. Transparent Layered User Interfaces: an Evaluation of a Display Design to Enhance Focused and Divided Attention. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 317–324. https://doi.org/10.1145/223904.223945

[25] Zhenyi He, Ruofei Du, and Ken Perlin. 2020. CollaboVR: A Reconfigurable Framework for Creative Collaboration in Virtual Reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 542–554. https://doi.org/10.1109/ISMAR50242.2020.00082

[26] Zhenyi He, Ruofei Du, and Ken Perlin. 2021. Who Is Looking at Whom? Visualizing Gaze Awareness for Remote Small-Group Conversations. *ArXiv Preprint ArXiv:2107.06265* (2021). https://arxiv.org/pdf/2107.06265

[27] Zhenyi He, Karl Toby Rosenberg, and Ken Perlin. 2019. Exploring configuration of mixed reality spaces for communication. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6. https://doi.org/10.1145/3290607.3312761

[28] Zhe He, A. Spurr, Xucong Zhang, and Otmar Hilliges. 2019. Photo-Realistic Monocular Gaze Redirection Using Generative Adversarial Networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 6931–6940.

[29] Otmar Hilliges, David Kim, Shahram Izadi, Malte Weiss, and Andrew Wilson. 2012. HoloDesk: Direct 3D Interactions With a Situated See-Through Display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2421–2430. https://doi.org/10.1145/2207676.2208405

[30] Michael Xuelin Huang, Tiffany CK Kwok, Grace Ngai, Stephen CF Chan, and Hong Va Leong. 2016. Building a Personalized, Auto-Calibrating Eye Tracker From User Interactions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5169–5179. https://doi.org/10.1145/2858036.2858404

[31] Ya-Hsin Hung and Paul Parsons. 2017. Assessing User Engagement in Information Visualization. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1708–1717. https://doi.org/10.1145/3027063.3053113

[32] Hiroshi Ishii and Minoru Kobayashi. 1992. ClearBoard: a Seamless Medium for Shared Drawing and Conversation With Eye Contact. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 525–532. https://doi.org/10.1145/142750.142977

[33] 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 (TOG)* 28, 3 (2009), 1–8. https://doi.org/10.1145/1531326.1531370

[34] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8110–8119. https://doi.org/10.1109/CVPR42600.2020.00813

[35] Kibum Kim, John Bolton, Audrey Girouard, Jeremy Cooperstock, and Roel Vertegaal. 2012. TeleHuman: Effects of 3d Perspective on Gaze and Pose Estimation With a Life-Size Cylindrical Telepresence Pod. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2531–2540. https://doi.org/10.1145/2207676.2208640

[36] Daniil Kononenko, Yaroslav Ganin, Diana Sungatullina, and Victor Lempitsky. 2018. Photorealistic Monocular Gaze Redirection Using Machine Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2018), 2696–2710. https://doi.org/10.1109/TPAMI.2017.2737423

[37] Daniil Kononenko and Victor Lempitsky. 2015. Learning to Look Up: Realtime Monocular Gaze Correction Using Machine Learning. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 4667–4675. https://doi.org/10.1109/CVPR.2015.7299098

[38] André Kunert, Alexander Kulik, Stephan Beck, and Bernd Froehlich. 2014. Photoportals: Shared References in Space and Time. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1388–1399. https://doi.org/10.1145/2531602.2531727

[39] Claudia Kuster, Tiberiu Popa, Jean-Charles Bazin, Craig Gotsman, and Markus Gross. 2012. Gaze Correction for Home Video Conferencing. *ACM Transactions on Graphics (TOG)* 31, 6 (2012), 1–6. https://doi.org/10.1145/2366145.2366193

[40] Grete Helena Kütt, Kevin Lee, Ethan Hardacre, and Alexandra Papoutsaki. 2019. Eye-Write: Gaze Sharing for Collaborative Writing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3290605.3300727

[41] Grete Helena Kütt, Teerapaun Tanprasert, Jay Rodolitz, Bernardo Moyza, Samuel So, Georgia Kenderova, and Alexandra Papoutsaki. 2020. Effects of Shared Gaze on Audio-Versus Text-Based Remote Collaborations. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25. https://doi.org/10.1145/3415207

[42] Gun A Lee, Theophilus Teo, Seungwon Kim, and Mark Billinghurst. 2018. A User Study on Mr Remote Collaboration Using Live 360 Video. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 153–164. https://doi.org/10.1109/ISMAR.2018.00051

[43] Daniel Leithinger, Sean Follmer, Alex Olwal, and Hiroshi Ishii. 2014. Physical Telepresence: Shape Capture and Display for Embodied, Computer-Mediated Remote Collaboration. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. ACM, 461–470. https://doi.org/10.1145/2642918.2647377

[44] Jiannan Li, Saul Greenberg, Ehud Sharlin, and Joaquim Jorge. 2014. Interactive Two-Sided Transparent Displays: Designing for Collaboration. In *Proceedings of the 2014 Conference on Designing Interactive Systems*. ACM, 395–404. https://doi.org/10.1145/2598510.2598518

[45] Matthew Lombard, Theresa B Ditton, and Lisa Weinstein. 2009. Measuring Presence: the Temple Presence Inventory. In *Proceedings of the 12th Annual International Workshop on Presence*. 1–15. https://doi.org/10.1177/2158244020922878

[46] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. 2011. Inferring Human Gaze From Appearance Via Adaptive Linear Regression. In *2011 International Conference on Computer Vision*. IEEE, 153–160.

[47] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. 2012. Head Pose-Free Appearance-Based Gaze Sensing Via Eye Image Synthesis. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR 2012)*. IEEE, 1008–1011.

[48] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[49] 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 *2013 IEEE Virtual Reality (VR)*. IEEE, 23–26. https://doi.org/10.1109/VR.2013.6549352

[50] Andrew F Monk and Caroline Gale. 2002. A Look Is Worth a Thousand Words: Full Gaze Awareness in Video-Mediated Conversation. *Discourse Processes* 33, 3 (2002), 257–278. https://doi.org/10.1207/S15326950DP330_4

[51] David Nguyen and John Canny. 2005. MultiView: Spatially Faithful Group Video Conferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 799–808. https://doi.org/10.1145/1240624.1240846

[52] David T Nguyen and John Canny. 2007. Multiview: Improving Trust in Group Video Conferencing Through Spatial Faithfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1465–1474. https://doi.org/abs/10.1145/1240624.1240846

[53] Farhad Nourbakhsh. 2015. Gaze Direction Adjustment for Video Calls and Meetings. US Patent 8,957,943.

[54] Kenton O'hara, Jesper Kjeldskov, and Jeni Paay. 2011. Blended Interaction Spaces for Distributed Team Collaboration. *ACM Transactions on Computer-Human Interaction (TOCHI)* 18, 1 (2011), 3. https://doi.org/10.1145/1959022.1959025

[55] 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, 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 29th Annual Symposium on User Interface Software and Technology (UIST)*. ACM, 741–754. https://doi.org/10.1145/2984511.2984517

[56] Kazuhiro Otsuka. 2016. MMSpace: Kinetically-Augmented Telepresence for Small Group-to-Group Conversations. In *Virtual Reality (VR), 2016 IEEE*. IEEE, 19–28. https://doi.org/10.1109/VR.2016.7504684

[57] Kazuhiro Otsuka. 2017. Behavioral Analysis of Kinetic Telepresence for Small Symmetric Group-to-Group Meetings. *IEEE Transactions on Multimedia* 20, 6 (2017), 1432–1447. https://doi.org/10.1109/TMM.2017.2771396

[58] Kazuhiro Otsuka, Shiro Kumano, Ryo Ishii, Maja Zbogar, and Junji Yamato. 2013. Mm+ Space: Nx 4 Degree-of-Freedom Kinetic Display for Recreating Multiparty Conversation Spaces. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*. 389–396. https://doi.org/10.1145/2522848.2522854

[59] Kazuhiro Otsuka, Shiro Kumano, Dan Mikami, Masafumi Matsuda, and Junji Yamato. 2012. Reconstructing Multiparty Conversation Field by Augmenting Human Head Motions Via Dynamic Displays. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. 2243–2248. https://doi.org/10.1145/2212776.2223783

[60] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI, AAAI, 3839–3845. https://doi.org/pdf/10.1145/3020165.3020170

[61] Seonwook Park, Shalini De Mello, P. Molchanov, U. Iqbal, Otmar Hilliges, and J. Kautz. 2019. Few-Shot Adaptive Gaze Estimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 9367–9376.

[62] 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 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1716–1725. https://doi.org/10.1145/2818048.2819965

[63] Yogi Tri Prasetyo, Retno Widyaningrum, and Chiuhsiang Joe Lin. 2019. Eye Gaze Accuracy in the Projection-Based Stereoscopic Display As a Function of Number of Fixation, Eye Movement Time, and Parallax. In *2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. IEEE, IEEE, 54–58.

[64] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2020).

[65] Ravi Ranjan. 2017. Detect Facial Features. https://github.com/raviranjan0309/Detect-Facial-Features.

[66] Derek F Reilly, Hafez Rouzati, Andy Wu, Jee Yeon Hwang, Jeremy Brudvik, and W Keith Edwards. 2010. TwinSpace: an Infrastructure for Cross-Reality Team Spaces. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*. ACM, 119–128. https://doi.org/10.1145/1866029.1866050

[67] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Construction of a Benchmark for the User Experience Questionnaire (UEQ). *IJIMAI* 4, 4 (2017), 40–44. https://doi.org/10.9781/ijimai.2017.4457

[68] 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*. 651–652. https://doi.org/10.1145/142750.143070

[69] Abigail J Sellen. 1995. Remote Conversations: the Effects of Mediating Talk With Technology. *Human-Computer Interaction* 10, 4 (1995), 401–444. https://doi.org/10.1207/s15327051hci100_2

[70] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. *Advances in Neural Information Processing Systems* 32 (2019), 7137–7147. https://arxiv.org/abs/2003.00196

[71] David Sirkin, Gina Venolia, John Tang, George Robertson, Taemie Kim, Kori Inkpen, Mara Sedlins, Bongshin Lee, and Mike Sinclair. 2011. Motion and Attention in a Kinetic Videoconferencing Proxy. In *IFIP Conference on Human-Computer Interaction*. Springer, Springer, 162–180. https://doi.org/10.1007/978-3-642-23774-_16

[72] Misha Sra, Aske Mottelson, and Pattie Maes. 2018. Your Place and Mine: Designing a Shared VR Experience for Remotely Located Users. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 85–97. https://doi.org/10.1145/3196709.3196788

[73] William Steptoe, Robin Wolff, Alessio Murgia, Estefania Guimaraes, John Rae, Paul Sharkey, David Roberts, and Anthony Steed. 2008. Eye-Tracking for Avatar Eye-Gaze and Interactional Analysis in Immersive Collaborative Virtual Environments. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*. 197–200. https://doi.org/10.1145/1460563.1460593

[74] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2010. Calibration-Free Gaze Sensing Using Saliency Maps. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, IEEE, 2667–2674. https://doi.org/10.1109/CVPR.2010.5539984

[75] Tony Tam, Joseph A Cafazzo, Emily Seto, Mary Ellen Salenieks, and Peter G Rossos. 2007. Perception of Eye Contact in Video Teleconsultation. *Journal of Telemedicine and Telecare* 13, 1 (2007), 35–39. https://doi.org/10.1258/135763307779701239

[76] Feitong Tan, Danhang Tang, Mingsong Dou, Kaiwen Guo, Rohit Pandey, Cem Keskin, Ruofei Du, Deqing Sun, Sofien Bouaziz, Sean Fanello, Ping Tan, and Yinda Zhang. 2021. HumanGPS: Geodesic PreServing Feature for Dense Human Correspondence. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://arxiv.org/abs/2103.15573

[77] Kar-Han Tan, Dan Gelb, Ramin Samadani, Ian Robinson, Bruce Culbertson, and John Apostolopoulos. 2010. Gaze Awareness and Interaction Support in Presentations. In *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. ACM, New York, NY, USA, 643–646. https://doi.org/10.1145/1873951.1874041

[78] Anthony Tang, Michel Pahud, Kori Inkpen, Hrvoje Benko, John C Tang, and Bill Buxton. 2010. Three's Company: Understanding Communication Channels in Three-Way Distributed Collaboration. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*. ACM, 271–280. https://doi.org/10.1145/1718918.1718969

[79] Roel Vertegaal. 1999. The GAZE Groupware System: Mediating Joint Attention in Multiparty Communication and Collaboration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 294–301. https://doi.org/10.1145/302979.303065

[80] 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*. 521–528. https://doi.org/10.1145/642611.642702

[81] E. Wood, T. Baltrusaitis, Louis-Philippe Morency, P. Robinson, and Andreas Bulling. 2018. GazeDirector: Fully Articulated Eye Gaze Redirection in Video. *Computer Graphics Forum* 37 (2018).

[82] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and WenSen Feng. 2020. Controllable Continuous Gaze Redirection. *Proceedings of the 28th ACM International Conference on Multimedia* (2020). https://doi.org/abs/10.1145/3394171.3413868

[83] L-Q Xu, A Loffler, PJ Sheppard, and D Machin. 1999. True-View Videoconferencing System Through 3-D Impression of Telepresence. *BT Technology Journal* 17, 1 (1999), 59–68.

[84] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. 2015. Turkergaze: Crowdsourcing Saliency With Webcam Based Eye Tracking. *ArXiv Preprint ArXiv:1504.06755* (2015).

[85] Ruigang Yang and Z. Zhang. 2004. Eye Gaze Correction With Stereovision for Video-Teleconferencing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26 (2004), 956–960. https://doi.org/10.1109/TPAMI.2004.27

[86] Yu Yu, Gang Liu, and Jean-Marc Odobez. 2019. Improving Few-Shot User-Specific Gaze Adaptation Via Gaze Redirection Synthesis. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 11929–11938.

[87] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214* (2020).

[88] Jichao Zhang, Jingjing Chen, Hao Tang, Wei Wang, Yan Yan, E. Sangineto, and N. Sebe. 2020. Dual In-Painting Model for Unsupervised Gaze Correction and Animation in the Wild. *Proceedings of the 28th ACM International Conference on Multimedia* (2020). https://doi.org/10.1145/3394171.3413981

[89] Yanxia Zhang, Ken Pfeuffer, Ming Ki Chong, Jason Alexander, Andreas Bulling, and Hans Gellersen. 2017. Look Together: Using Gaze for Assisting Co-Located Collaborative Search. *Personal and Ubiquitous Computing* 21, 1 (2017), 173–186. https://doi.org/10.1007/s00779-016-0969-x

[90] Jiejie Zhu, Ruigang Yang, and Xueqing Xiang. 2011. Eye Contact in Video Conference Via Fusion of Time-of-Flight Depth Sensor and Stereo. *3D Research* 2, 3 (2011), 5.