AI-Mediated 3D Video Conferencing

Michael Stengel
Koki Nagano
mstengel@nvidia.com
knagano@nvidia.com
NVIDIA
USA

Chao Liu
Matthew Chan
chaoliu@nvidia.com
matchan@nvidia.com
NVIDIA
USA

Alex Trevithick
atrevith@gmail.com
UCSD
USA

Shalini De Mello
Jonghyun Kim
shalinig@nvidia.com
jonghyunk@nvidia.com
NVIDIA
USA

David Luebke
dluebke@nvidia.com
NVIDIA
USA

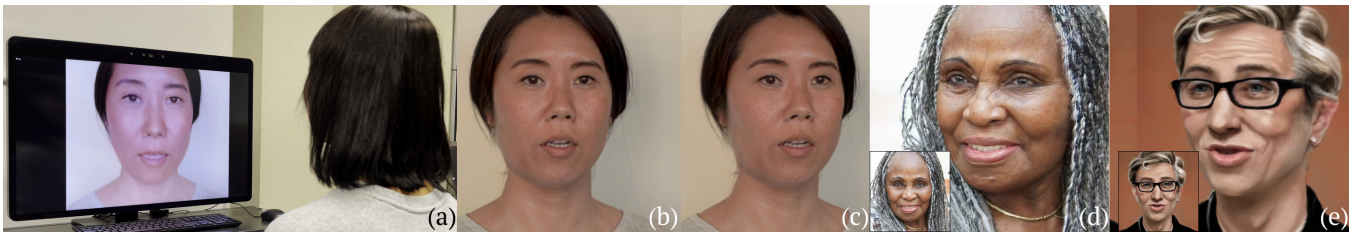

Figure 1: (a) Our system overview using a tracked stereo display. (b, c) A cross-fusible stereo pair of the participant that appears life-size in correct perspective recorded on our display. (d, e) Our method can generate novel views of both photo realistic and stylized telepresence given a single RGB image shown in inset.

ABSTRACT

We present an AI-mediated 3D video conferencing system that can reconstruct and autostereoscopically display a life-sized talking head using consumer-grade compute resources and minimal capture equipment. Our 3D capture uses a novel 3D lifting method that encodes a given 2D input into an efficient triplanar neural representation of the user, which can be rendered from novel viewpoints in real-time. Our AI-based techniques drastically reduce the cost for 3D capture, while providing a high-fidelity 3D representation on the receiver's end at the cost of traditional 2D video streaming. Additional advantages of our AI-based approach include the ability to accommodate both photorealistic and stylized avatars, and the ability to enable mutual eye contact in multi-directional video conferencing. We demonstrate our system using a tracked stereo display for a personal viewing experience as well as a lightfield display for a room-scale multi-viewer experience.

ACM Reference Format:
Michael Stengel, Koki Nagano, Chao Liu, Matthew Chan, Alex Trevithick, Shalini De Mello, Jonghyun Kim, and David Luebke. 2023. AI-Mediated 3D Video Conferencing. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Emerging Technologies (SIGGRAPH '23 Emerging Technologies), August 06-10, 2023.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3588037.3595385

1 INTRODUCTION

Over 50 years ago, companies introduced the first commercially available video conference systems, such as the AT&T Picturephone and Philips Videophone, allowing people to *see* as well as hear each other over long distances. However, decades of advances in computing technology, video encoding, and streaming infrastructure were required before video conferencing was inexpensive and widely adopted. The ultimate goal of video conferencing is to enable immersive communication between remote participants as if they were in the same physical location. While recent advances in 3D video conferencing demonstrate promising capabilities to capture eye contact and other non-verbal cues critical to face-to-face communications, existing systems require expensive 3D acquisition setups that are inaccessible to end users. In this work, we show a set of AI methods that allow users to create compelling 3D telepresence given 2D videos compatible with traditional video conferencing. We argue that our AI-based solution enables democratization of high-fidelity telepresence, while also offering new capabilities that 3D scanning-based methods cannot.

2 RELATED WORK

Previous Emerging Technologies installations showed pre-recorded 3D scans of a life-sized face [Nagano et al. 2013] or pre-recorded full-body digital humans [Jones et al. 2015] using projector arrays. Other previous work [Jones et al. 2009; Lawrence et al. 2021] used custom multi-view 3D acquisition systems for capturing the participant in real-time. However, such 3D scanning systems are expensive and require volumetric video streaming, making the technology less accessible. Instead, our method relies on recent neural implicit representations, encoding the sender's information to a photo-realistic volumetric representation given a single RGB image. As a result, our system allows the sender to stream 2D videos like traditional video conferencing, while decoding high-quality 3D representations on the receiver side that can be rendered in real-time.

3 SYSTEM OVERVIEW

Fig. 2 shows a pipeline of our system, which consists of a *sender* that records and streams 2D videos from a single RGB webcam, and a *receiver* that receives the 2D video and lifts it to 3D for novel view rendering from the viewer's perspective for a corresponding

SIGGRAPH '23 Emerging Technologies, August 06-10, 2023, Los Angeles, CA, USA

ACM ISBN 979-8-4007-0154-2/23/08.
https://doi.org/10.1145/3588037.3595385

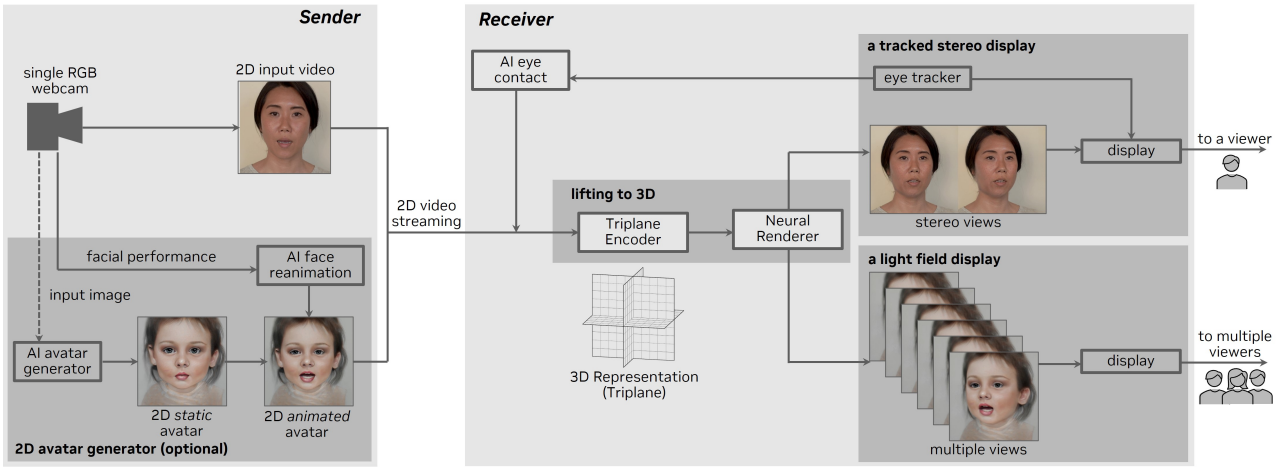

Figure 2: Given a single RGB webcam, our system starts by recording a 2D video of the sender and streaming it to the receiver like traditional video conferencing. The user can optionally choose to generate a 2D avatar or stylize the input picture and then drive the generated avatar using our 2D avatar generator. The streamed 2D video is then encoded to our neural triplanar representation that is rendered from novel views in real-time and eye contact can be achieved by using our AI eye contact module. Our system supports both a tracked stereo display to a single viewer or a light field display to multiple viewers.

3D display. In addition to using a webcam picture, the user can optionally use our 2D avatar generator module that generates and customizes a 2D avatar to be driven by the user [Wang et al. 2021].

Lifting to 3D. We propose a novel Vision Transformer-based encoder which transforms 2D input into an efficient triplanar implicit 3D representation [Chan et al. 2022]. Given a single RGB image of a user, our method automatically creates a frontalized 3D representation of the user, which can be efficiently rendered from novel viewpoints using volumetric rendering. Our 3D lifting module employs generative priors to ensure that generated views are multi-view consistent and photo realistic from novel views and generalizes to anyone in one shot without person-specific training.

Eye contact. We use a state-of-the-art neural method [Zheng et al. 2020] to synthesize a redirected eye gaze given a picture of the user. We then lift the gaze-corrected 2D image to 3D to make eye contact (see Fig. 3).

3D Displays. Our system supports multiple off-the-shelf 3D displays including a tracked stereo display targeted for a single viewer or light field display for multiple viewers. Fig. 1 shows an implementation using a 32 inch 3D display from Dimenco[1], which uses eye tracking and lenticular prisms to show a stereo image pair for the user's eye positions. We also tested our system using a 32 inch Looking Glass display[2] that shows wider range of dense views for multiple participants.

4 DEMO

When visitors approach our show booth, multiple visitors can simultaneously see a life-sized talking head on a light field display that explains the overall technology and our booth layout. This light field display provides an eye-catching at-a-glance demonstration of the technology for casual passers-by and for people waiting in line for the demo. Participants are then invited to try a tracked stereo display located in separated booths, where they can see the other participants on the 3D display and can experience a multi-way AI-mediated 3D video conferencing call.

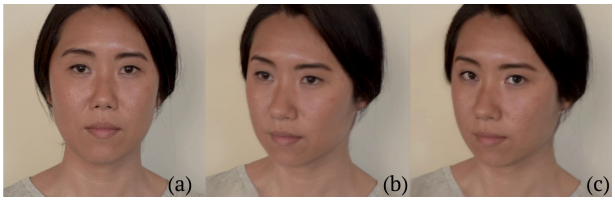

Figure 3: Demonstration of our AI eye contact feature. (a) The subject looks at the camera in the input image. (b) Without the eye contact feature, the subject maintains the original eye gaze and does not make an eye contact when rendered from a novel view. (c) When the eye contact feature is enabled, the subject can make eye contact from arbitrary views.

[1]https://www.dimenco.eu/
[2]https://lookingglassfactory.com/

REFERENCES

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. In *CVPR*.

Andrew Jones, Magnus Lang, Graham Fyffe, Xueming Yu, Jay Busch, Ian McDowall, Mark Bolas, and Paul Debevec. 2009. HeadSPIN: A One-to-Many 3D Video Teleconferencing System. In *ACM SIGGRAPH 2009 Emerging Technologies*.

Andrew Jones, Jonas Unger, Koki Nagano, Jay Busch, Xueming Yu, Hsuan-Yueh Peng, Oleg Alexander, Mark Bolas, and Paul Debevec. 2015. An Automultiscopic Projector Array for Interactive Digital Humans. In *ACM SIGGRAPH 2015 Emerging Technologies*.

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.* (2021).

Koki Nagano, Andrew Jones, Jing Liu, Jay Busch, Xueming Yu, Mark Bolas, and Paul Debevec. 2013. An Autostereoscopic Projector Array Optimized for 3D Facial Display. In *ACM SIGGRAPH 2013 Emerging Technologies*.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *CVPR*.

Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. 2020. Self-learning transformations for improving gaze and head redirection. *NeurIPS* (2020).