Telepresence

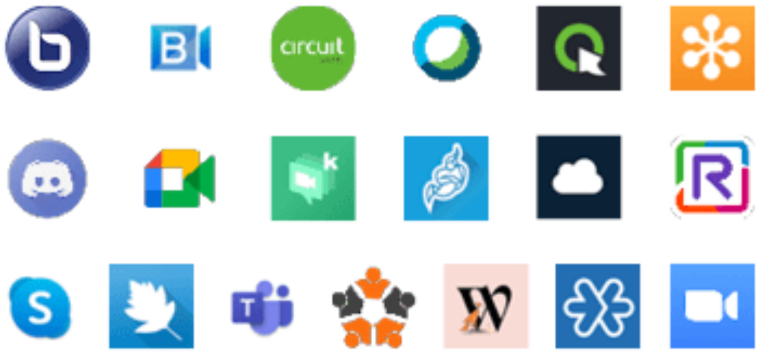
CS 6334 Virtual Reality

Professor Yapeng Tian

The University of Texas at Dallas

Telecommunication

- Improvements in telecommunications have steadily increased both the fidelity and availability of synchronous communication over long-distance networks
- Video-based systems like Skype, Zoom, Meet, and Teams are a recent step forward in bringing people closer together who are far apart



Telepresence

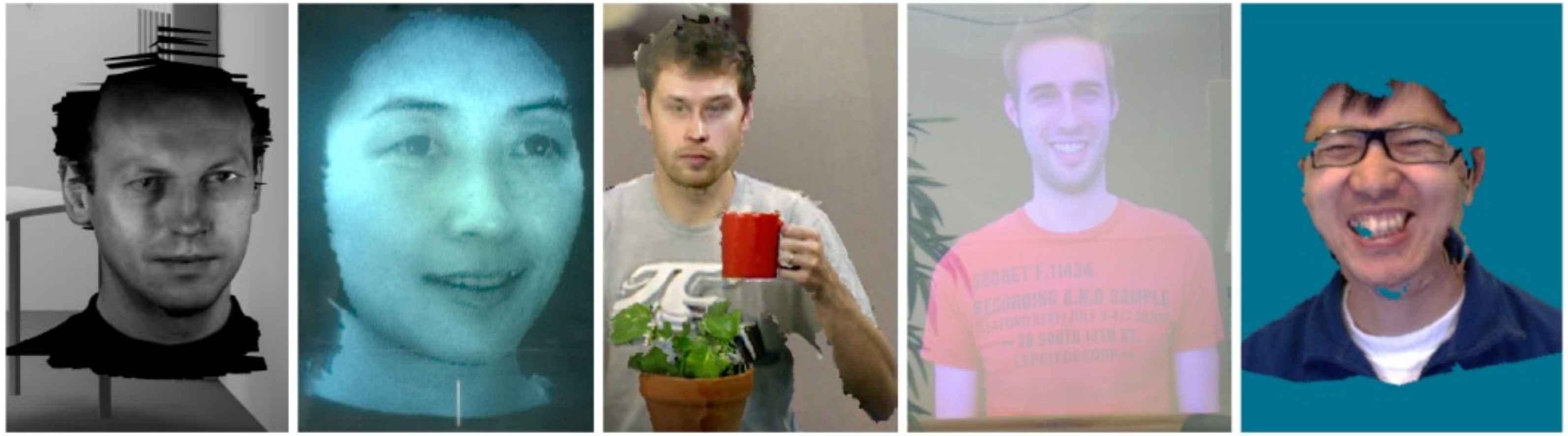
- At the far end of this spectrum is ***telepresence***
  - enabling remote participants to feel copresent, as if they are occupying a shared physical/virtual space

Google

Meta

3D Telepresence Systems

- Past systems have limits in spatial resolution, color fidelity, depth accuracy, audio, and refresh rate (not realistic)

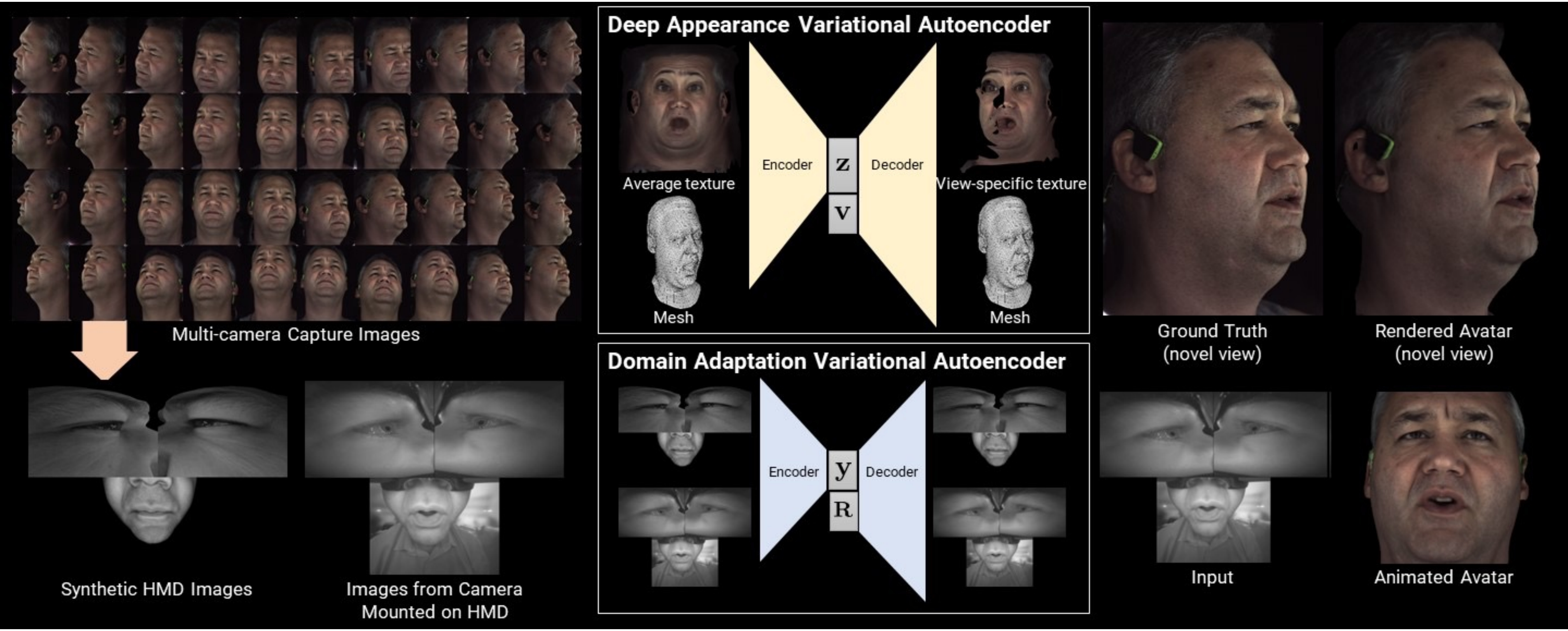


[Gibbs et al. 1999] [Jones et al. 2009] [Maimone et al. 2012] [Kuster et al. 2012] [Zhang et al. 2013]

Meta’s HMD-based Codec Avatar

https://www.youtube.com/watch?v=w52CziLgnAc

Key AI Technique in Meta's Codec Avatar


Multi-camera Capture Images
Synthetic HMD Images
Images from Camera Mounted on HMD
Deep Appearance Variational Autoencoder
Average texture
Mesh
Encoder
z
v
Decoder
View-specific texture
Mesh
Domain Adaptation Variational Autoencoder
Encoder
y
R
Decoder
Ground Truth (novel view)
Rendered Avatar (novel view)
Input
Animated Avatar

Stephen et al. Deep Appearance Models for Face Rendering, TOG, 2018.

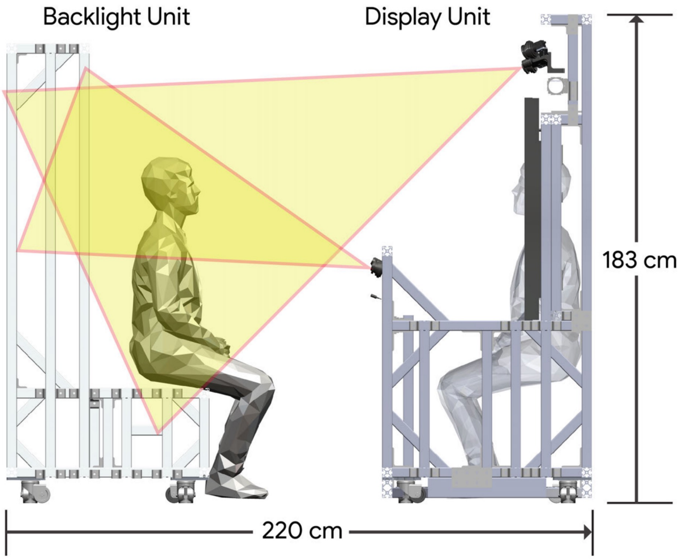
Project Starline:

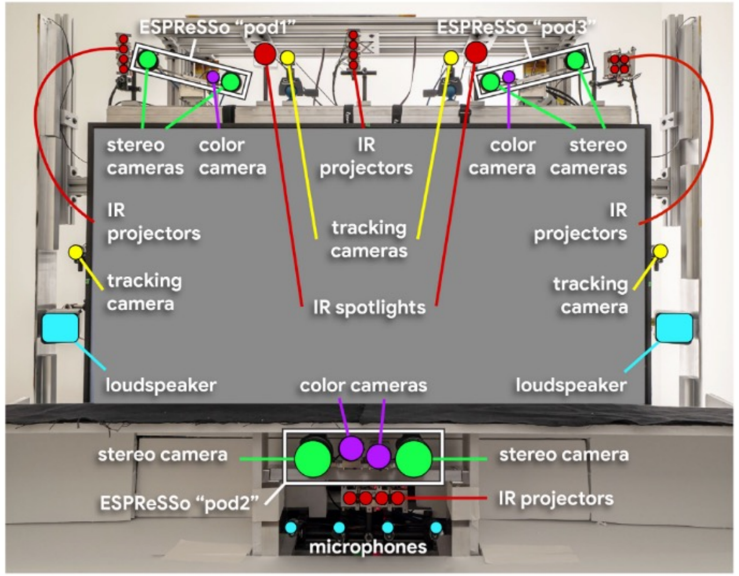
Google's next-generation 3D video chat booth

https://www.youtube.com/watch?v=Q13CishCKXY

System Setup

- The system comprises two main structures
  - Display unit housing a display, cameras, speakers, microphones, illuminators, and computer
  - Backlight unit housing an infrared backlight and also serving as a bench seat





Lighting

- Both units contain white LED strips angled toward the walls and ceiling to produce soft bounce lighting

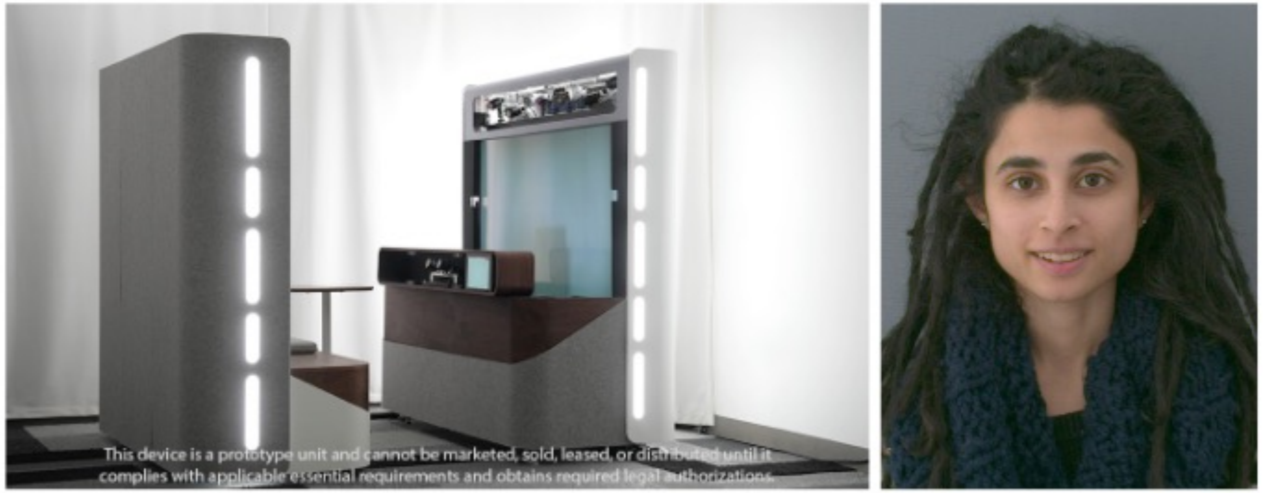
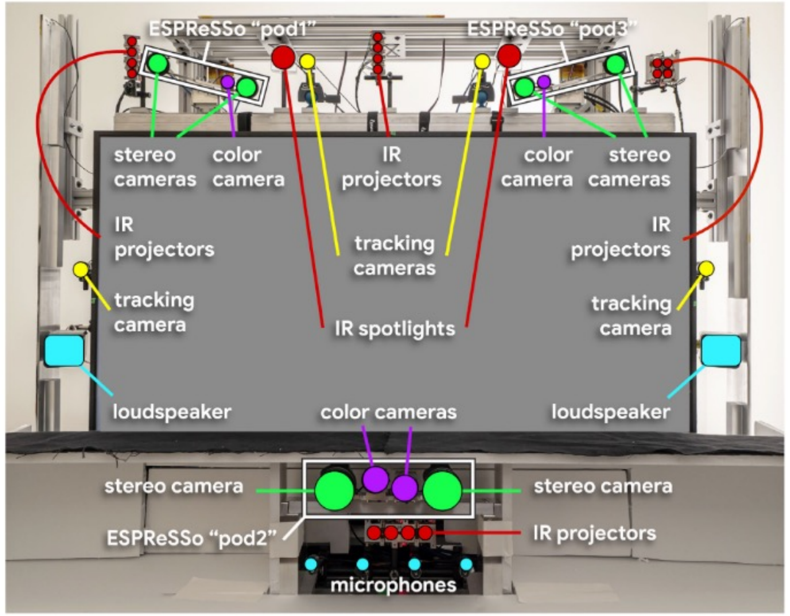

This device is a prototype unit and cannot be marketed, sold, leased, or distributed until it complies with applicable essential requirements and obtains required legal authorizations.

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

Visual Capture

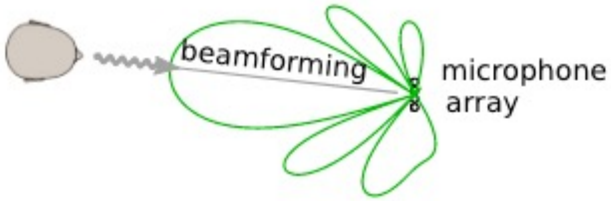
- Three synchronized stereo RGBD capture pods
  - Two above the display, and one in the "middle wall" below the display
  - The lower pod includes an extra color camera, zoomed into the subject's face
- Four tracking cameras
  - Two above the display and one on each side, capture high-speed wide-angle images for real-time 3D localization of the eyes, ears, and mouth
- Four color and three depth streams from the RGBD capture pods can be captured

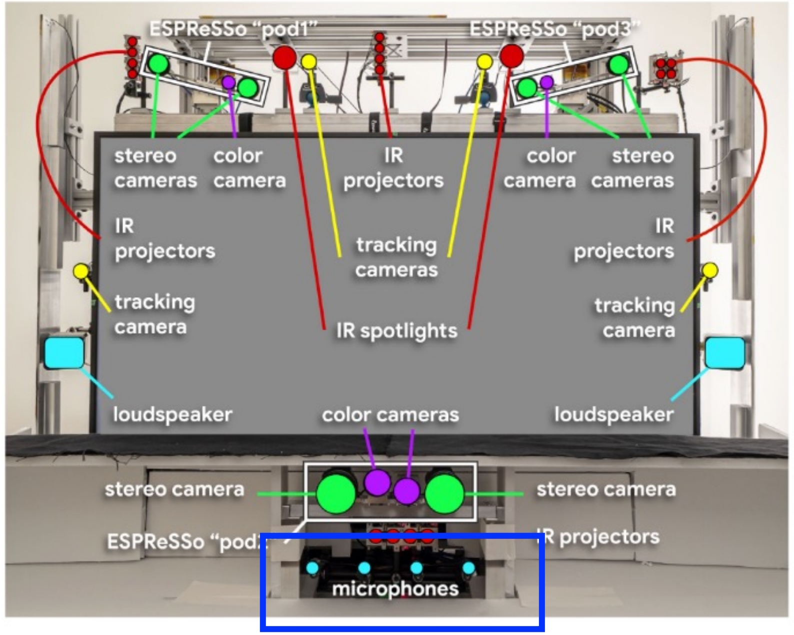


The visual capture will be compressed on the GPU and transmitted alongside tracked 3D face points using WebRTC

Audio: Capture

- Audio is captured using four microphones arranged as a linear array in the middle wall, beneath the lower capture pod
- Capture processing performs the following tasks
  - Ambient noise reduction
  - Reverberation reduction
  - Acoustic echo cancellation

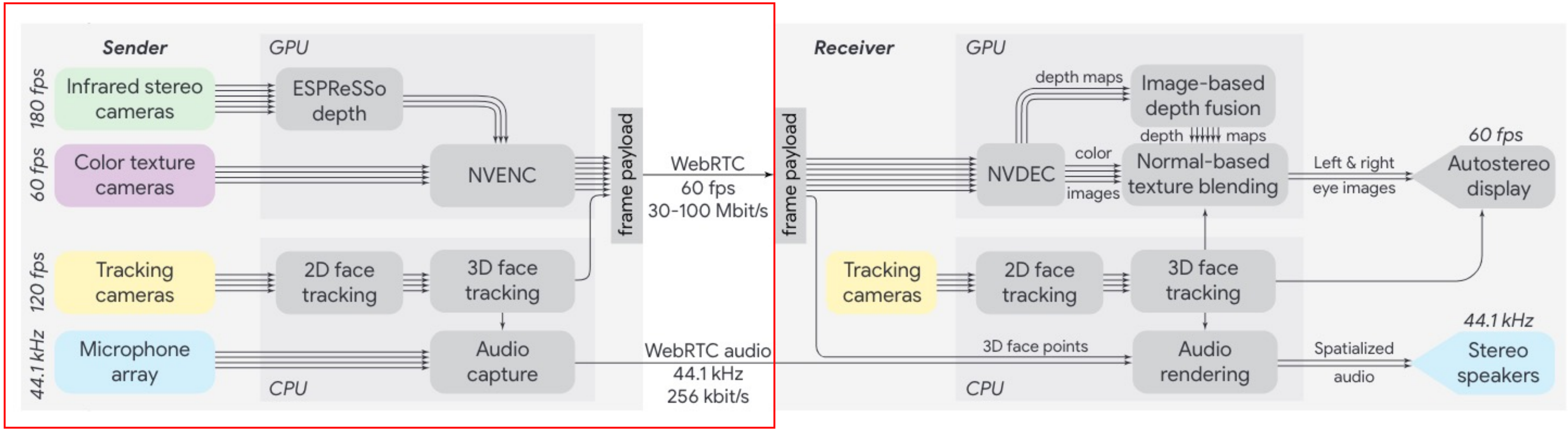




3D Face Tracking

- Precise 3D tracking of user facial features is crucial
    - The eye locations determine stereo viewpoints for rendering
    - The mouth position enables beamforming in audio capture
    - Both the mouth and ear locations contribute to spatialized audio rendering
- Tracking approach
    - For each captured image, we detect the face and locate 34 facial landmarks
    - We determine the 2D locations of five features (eyes, mouth, and ears) as weighted combinations of nearby landmarks
    - For each feature found in at least two of the four tracking cameras, we use triangulation to obtain its 3D position

Data Flow in the System



Four color images, three depth maps, and one audio

Left and right eye images (3D reconstruction) and spatial audio

Visual Rendering: 3D Reconstruction

- On the receiving client, given the 3 depth maps and 4 color images, the system renders a fused 3D surface of the local user

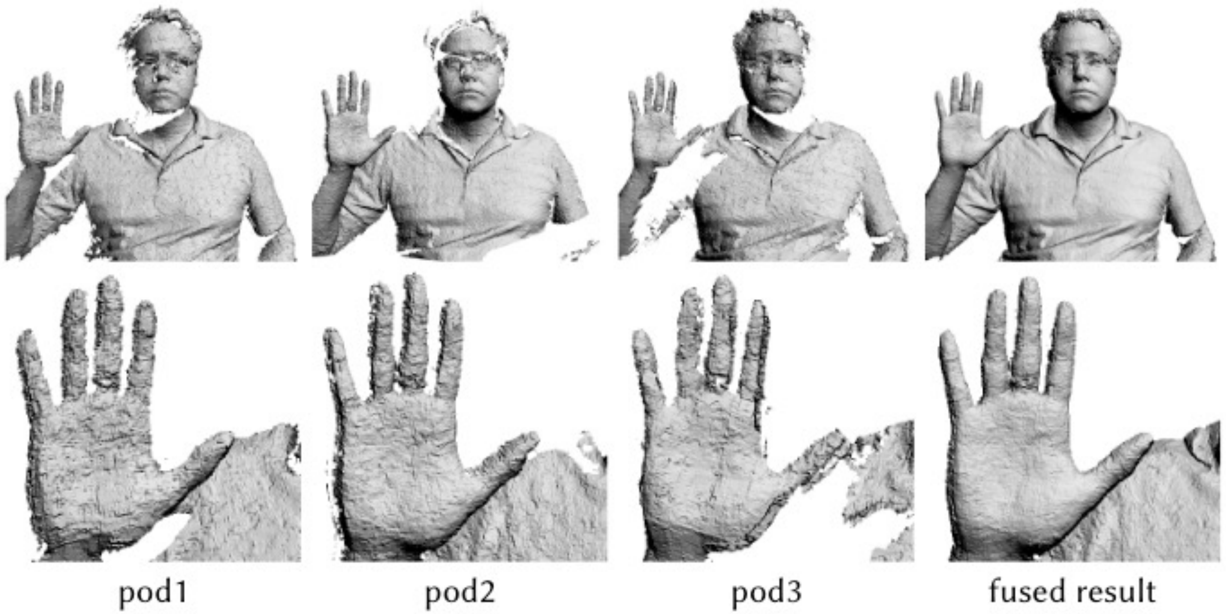
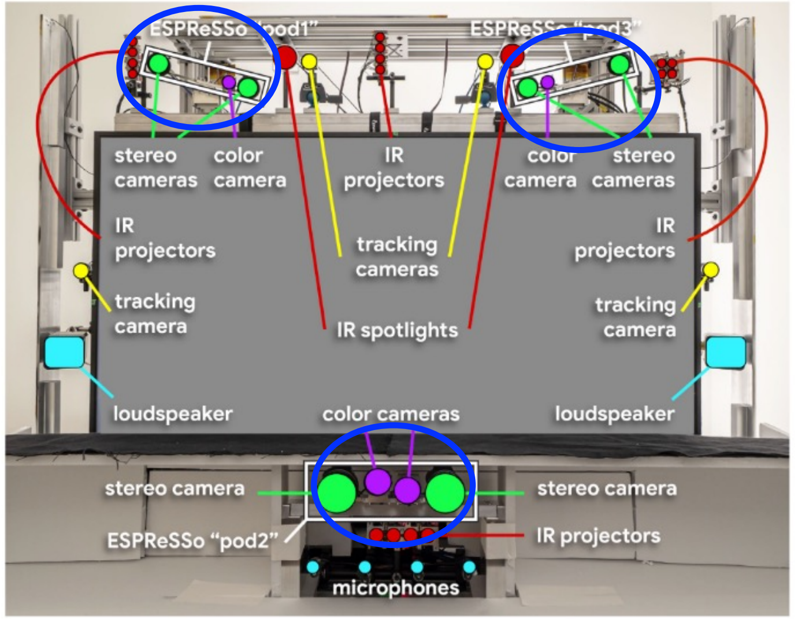


Fig. 9. Contribution of each stereo depth image and resulting fused surface.

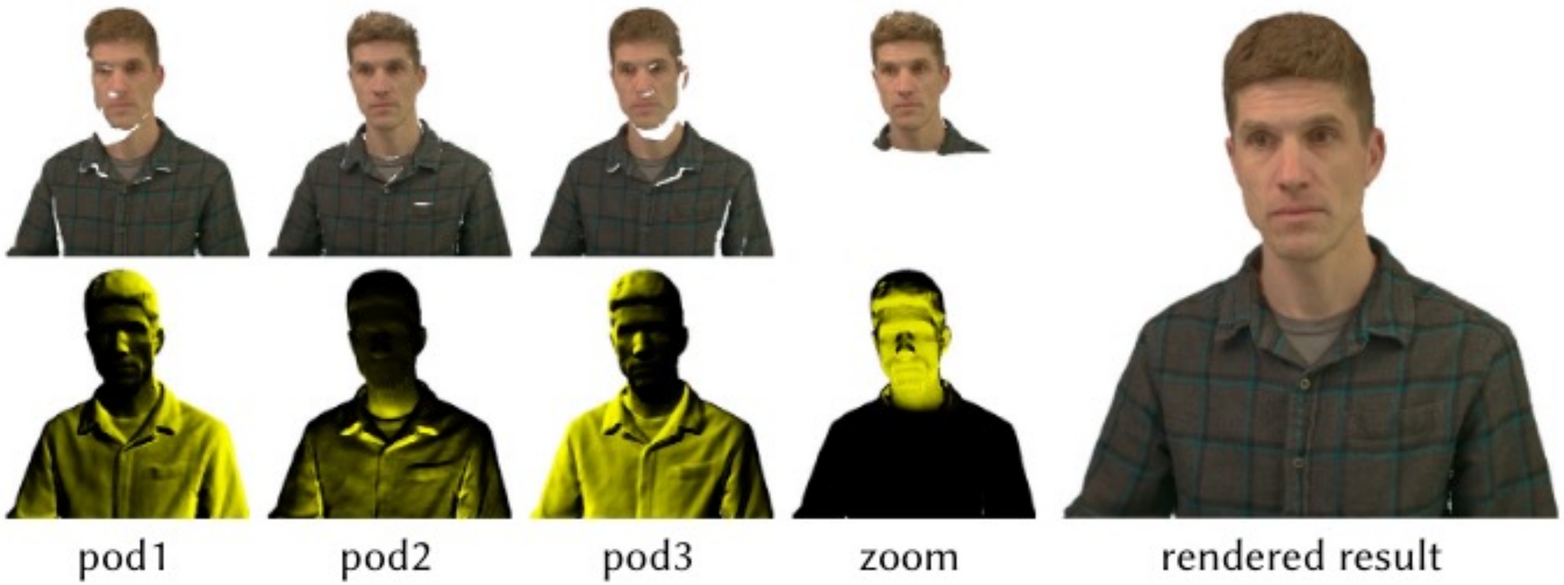


head-tracked autostereoscopic display

Brian Curless and Marc Levoy. A Volumetric Method for Building Complex Models from Range Images. 1996.

Visual Rendering: Weighted Color Blending

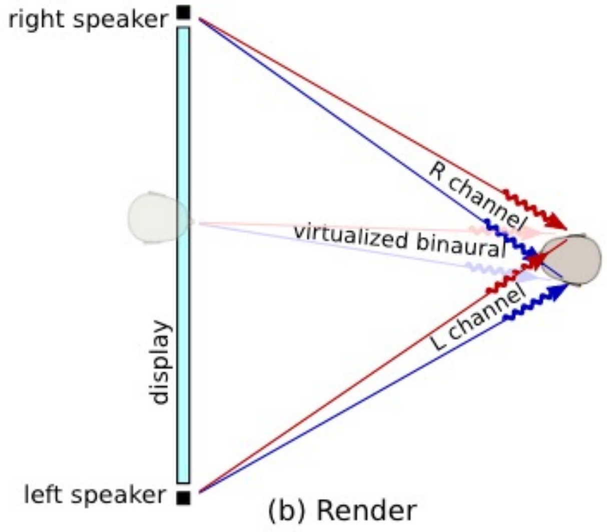
- Project each color image onto the fused surface and combine these using blend weights (yellow) determined from the surface normal
  - The blend weight of each color image is modulated by the squared cosine of the angle between the surface normal and the camera vector



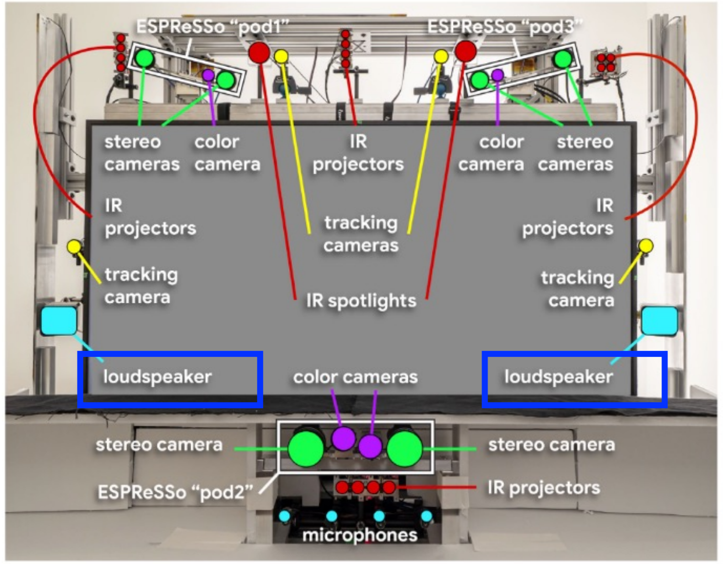
Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. SIGGRAPH, 2001.

Audio Rendering

- Stereo loudspeakers render tracked and 3D-spatialized audio
  - The tracked talker and listener positions are combined dynamically with a generic head-related transfer function to yield a real-time-tracked binaural signal
  - The binaural signal is converted to stereo loudspeaker output using listener-tracked binaural crosstalk cancellation with the same HRTF model



(b) Render



Crosstalk cancellation (XTC) yields high-spatial-fidelity reproduction of binaural audio through loudspeakers allowing a listener to perceive an accurate 3-D image of a recorded soundfield.

Further Reading

- Lawrence et al. Project starline: a high-fidelity telepresence system, SIGGRAPH, 2021
- https://blog.google/technology/research/project-starline/
- Stephen et al. Deep Appearance Models for Face Rendering, TOG, 2018.

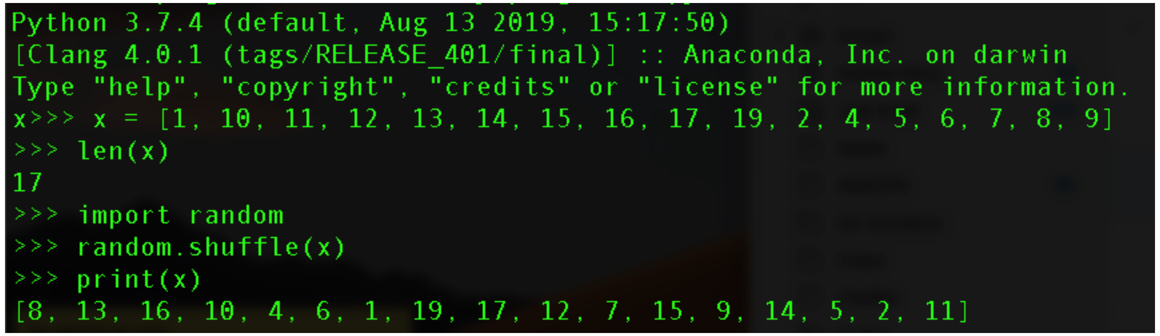
Guest Lecture (Wednesday 11/30)

- Prof. Jeff Price
- School of Arts, Humanities, and Technology
- Research areas
  - VR/AR
  - Game Design
  - 3D Modeling
  - Animation
  - Design

Quiz 5 (attendance)

Presentation Order

- **The presentation order was randomly generated**



```
Python 3.7.4 (default, Aug 13 2019, 15:17:50)
[Clang 4.0.1 (tags/RELEASE_401/final)] :: Anaconda, Inc. on darwin
Type "help", "copyright", "credits" or "license" for more information.
x>>> x = [1, 10, 11, 12, 13, 14, 15, 16, 17, 19, 2, 4, 5, 6, 7, 8, 9]
>>> len(x)
17
>>> import random
>>> random.shuffle(x)
>>> print(x)
[8, 13, 16, 10, 4, 6, 1, 19, 17, 12, 7, 15, 9, 14, 5, 2, 11]
```

- Set 1 (Monday)
  - 8, 13, 16, 10, 4, 6, 1, 19

- Set 2 (Wednesday)
  - 17, 12, 7, 15, 9, 14, 5, 2, 11

Project Presentation

- Presentation
  - Introduction: Project title, group members, system overview (2min)
  - Method: System implementation (2min)
  - Demo: a video demo to showcase your system (3min)
  - QA (1min)
- Each group has 8 minutes for the presentation and questions
  - Please use slides to describe your VR/AR system
  - All group members should show up
  - Show a demo of the system
- Evaluation criteria
  - The grading will be based on the overall quality of the presentation in terms of content, clarity, demo and question answering