Holoportation: Virtual 3D Teleportation in Real-time

S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev and 17 other Microsoft researchers

In Proceedings of the 29th Annual Symposium on User Interface Software and Technology (pp. 741-754). ACM.

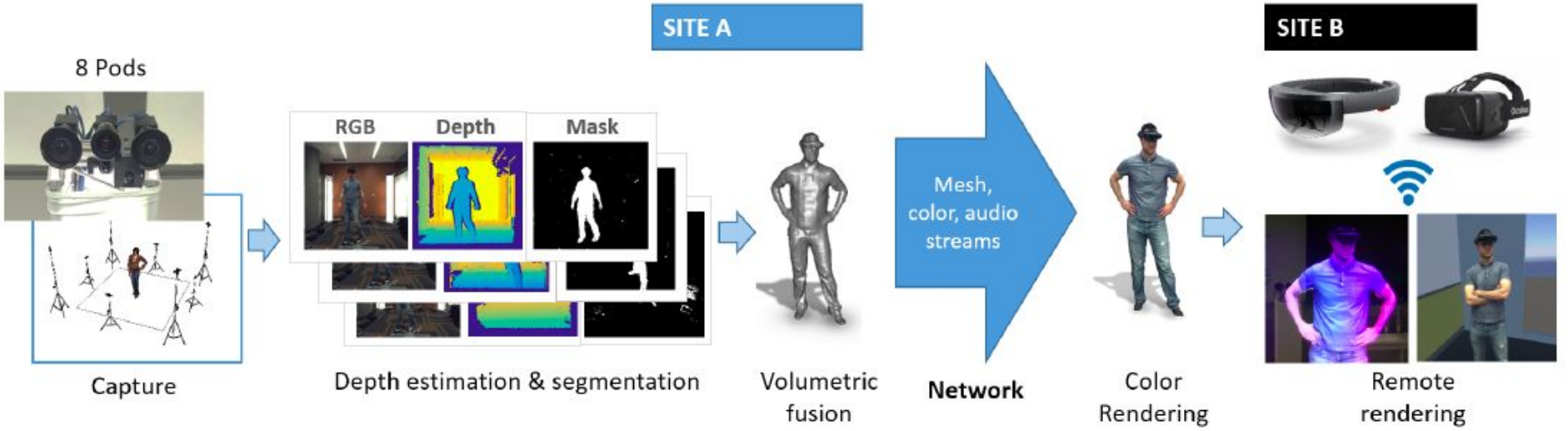

remote
user
++

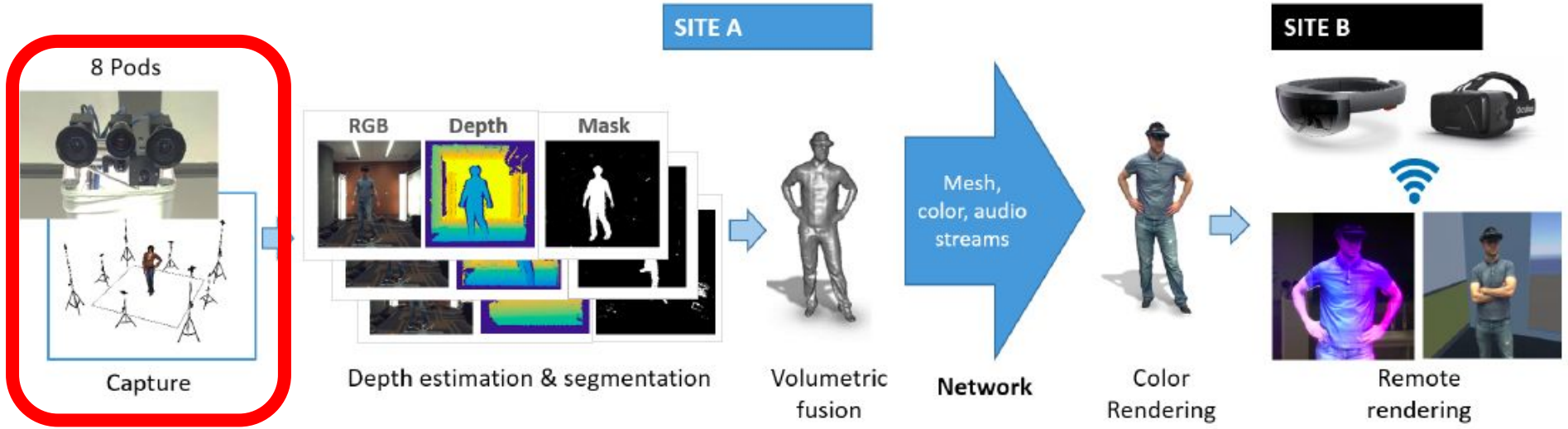

AR first-person view


VR first-person view

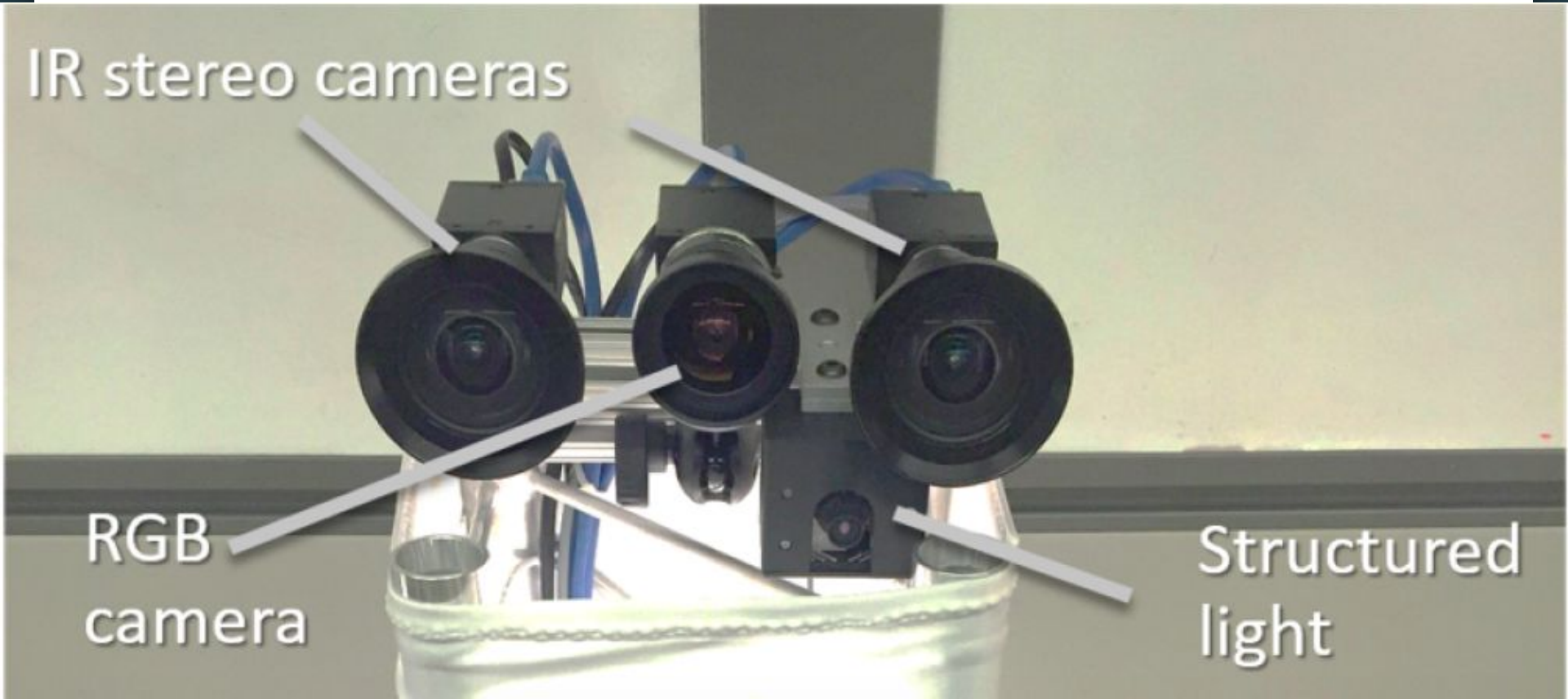
Setup of each Station

- 16 Near Infra-Red cameras (NIR) -> Depth
- 8 RGB cameras
- 8 Structure lights
- 1 HMD
- 10 Gbps link
- 5 PCs with
  - an Intel Core i7 3.4 Ghz CPU, 16 GB of RAM and 2 NVIDIA Titan X GPUs


SITE A
SITE B
8 Pods
RGB
Depth
Mask
Mesh, color, audio streams
Capture
Depth estimation & segmentation
Volumetric fusion
Network
Color Rendering
Remote rendering


8 Pods
Capture
SITE A
RGB
Depth
Mask
Depth estimation & segmentation
Volumetric fusion
Mesh, color, audio streams
Network
Color Rendering
SITE B
Remote rendering

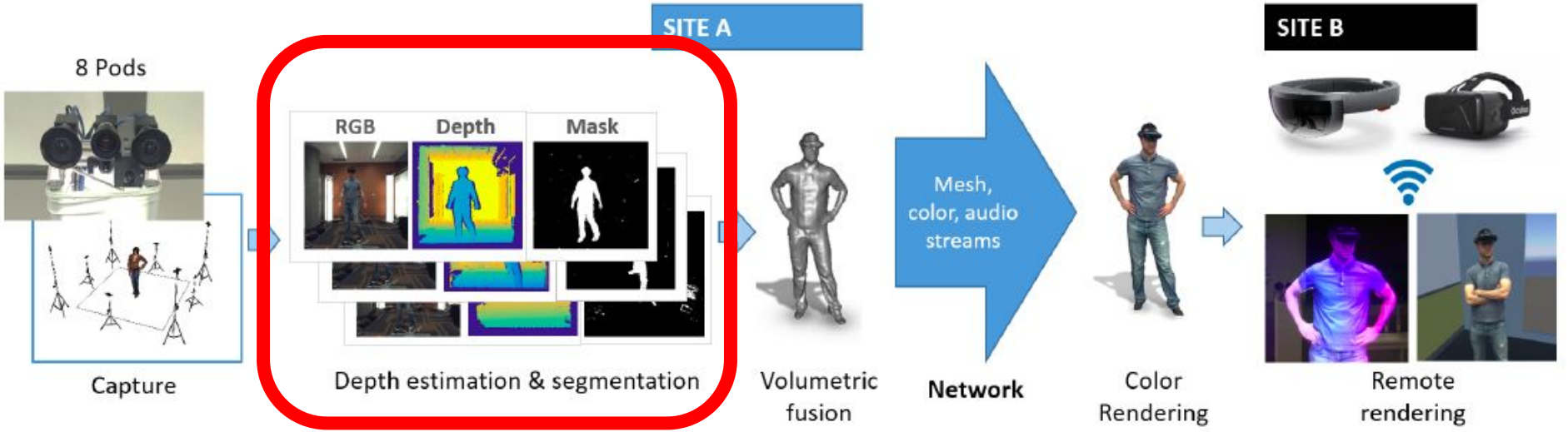
Capture Pods



Capture Pods

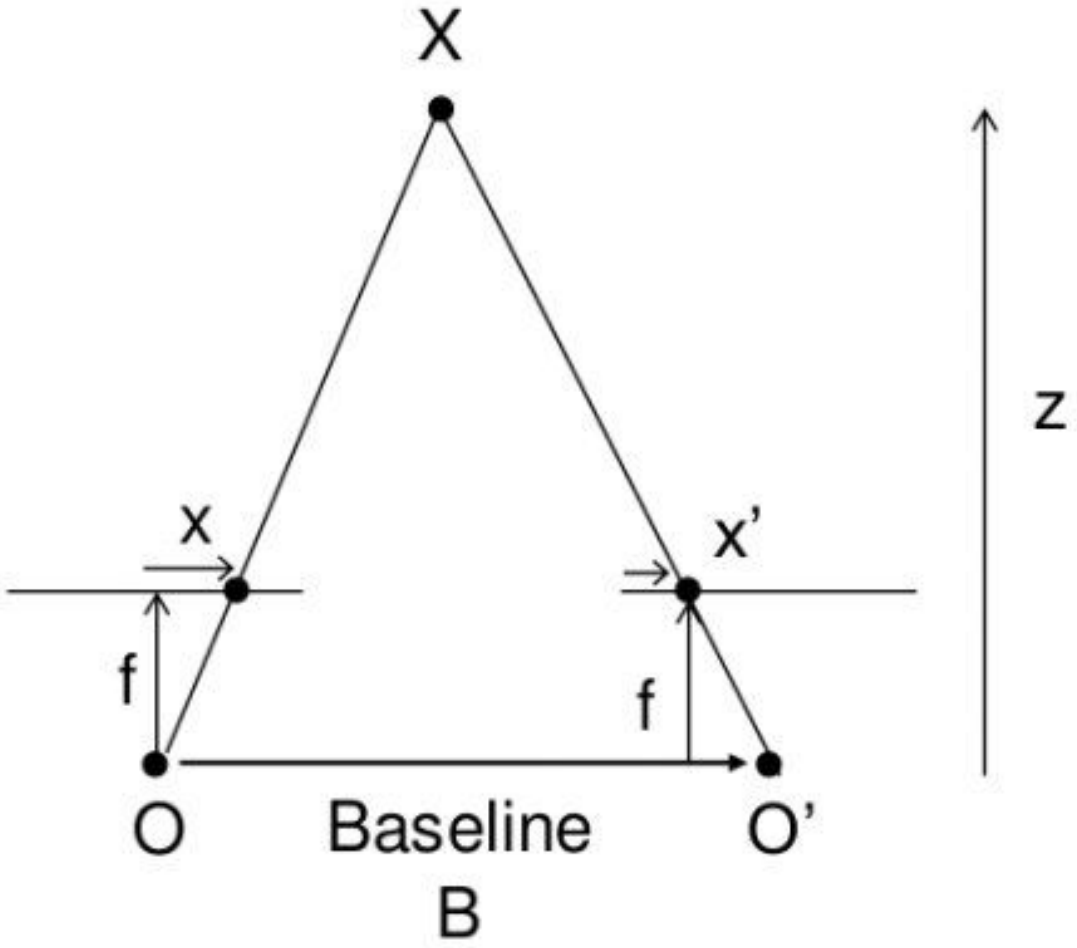
- All the pods are synchronized using an external trigger running at 30fps.
- Depth streams
  - use [1] for computing the camera parameters
- RGB streams
  - individual white balancing
  - makes the signal consistent across all the RGB cameras by using linear mapping.

[1] Zhang, Z. A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. 22, 11 (Nov. 2000), 1330–1334.


SITE A
SITE B
8 Pods
RGB
Depth
Mask
Mesh, color, audio streams
Capture
Depth estimation & segmentation
Volumetric fusion
Network
Color Rendering
Remote rendering

Depth Estimation

- Passive stereo for depth estimation



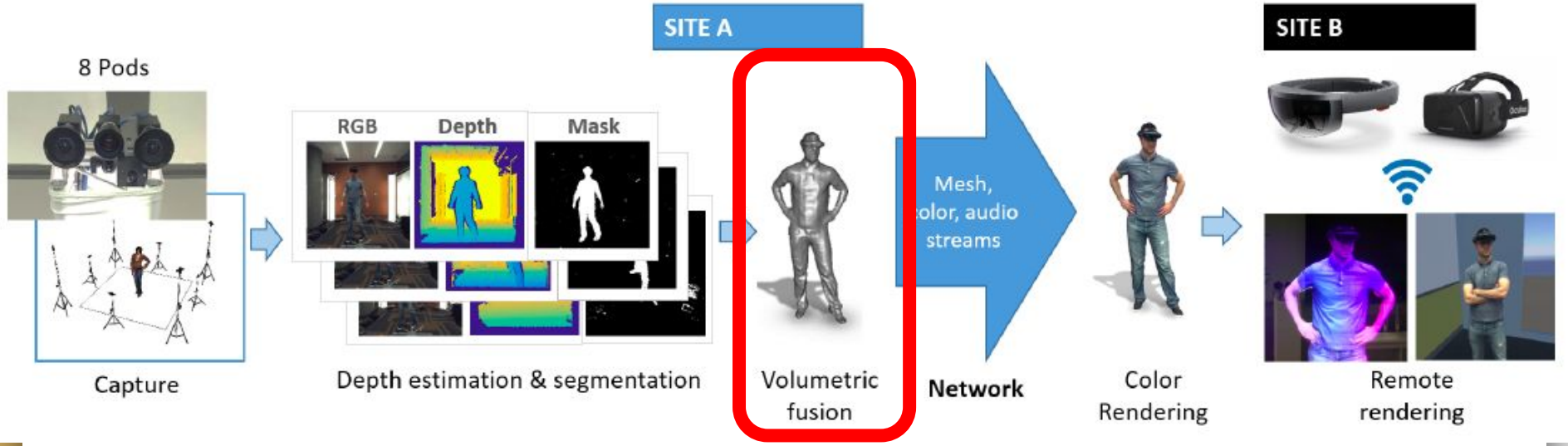
Depth Estimation

- Active stereo for depth estimation.
  - 2 NIR cameras
  - one or more random IR dot pattern projector
- Each IR dot serves as a texture in the scene to help estimate depth even in case of texture-less surfaces.

Foreground Segmentation

- Provides 2D silhouettes [1]
  - achieving temporally consistent 3D reconstructions
  - compressing the data sent over the network.

[1] Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S. R., Kowdle, A., Escolano, S. O., Rhemann, C., Kim, D., Taylor, J., Kohli, P., Tankovich, V., and Izadi, S. Fusion4d: Real-time performance capture of challenging scenes. ACM Trans. Graph. 35, 4


SITE A
SITE B
8 Pods
RGB
Depth
Mask
Mesh,
color, audio
streams
Capture
Depth estimation & segmentation
Volumetric
fusion
Network
Color
Rendering
Remote
rendering

Temporally Consistent 3D Reconstruction

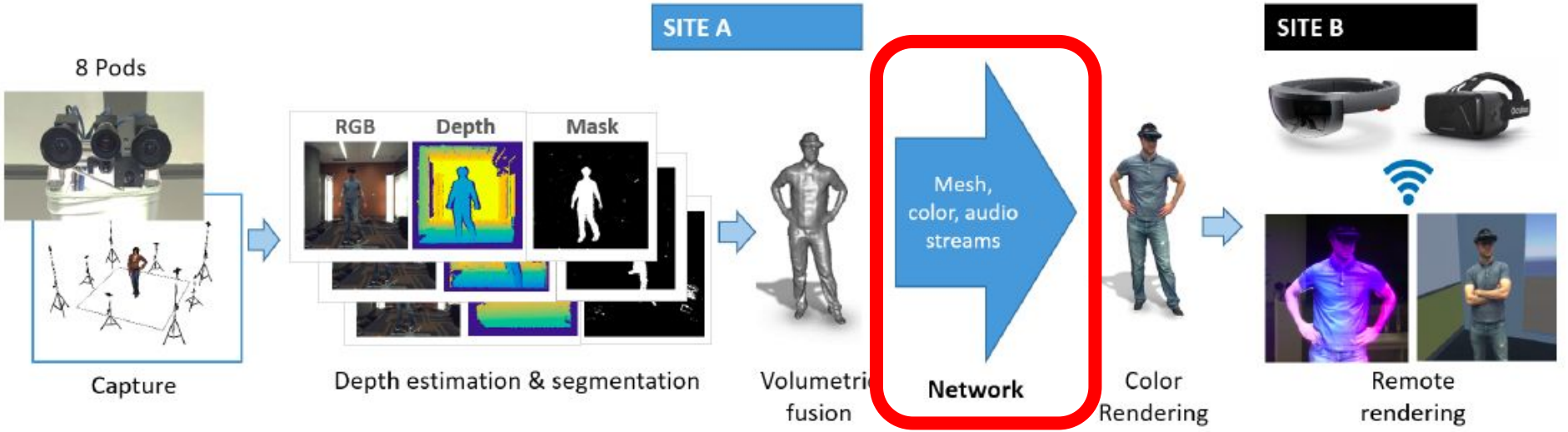
- Tracks the mesh and fuses the data across cameras and frames.[1]
- Marching cubes polygonalization of the volumetric data
- Color Texturing
  - regular visibility tests
  - majority voting scheme for colors
    - to classify each view as trusted, the color candidates for this view must agree with a number of colors from other views that can see this point
    - the number of agreeing views should be maximum.

[1] Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S. R., Kowdle, A., Escolano, S. O., Rhemann, C., Kim, D., Taylor, J., Kohli, P., Tankovich, V., and Izadi, S. Fusion4d: Real-time performance capture of challenging scenes. ACM Trans. Graph. 35, 4 (July 2016), 114:1–114:13.

Spatial Audio

- Synthesize each remote audio source
- Capture from the position and orientation
- Transforms the head pose information from the remote user’s room coordinate system to the local user’s room coordinate system
- Spatializes the audio source at the proper location and orientation.
- Head related transfer function (HRTF) [1]

[1] Gilkey, R. H., and Anderson, T. R., Eds. Binaural and Spatial Hearing in Real and Virtual Environments. Psychology Press, 2009.

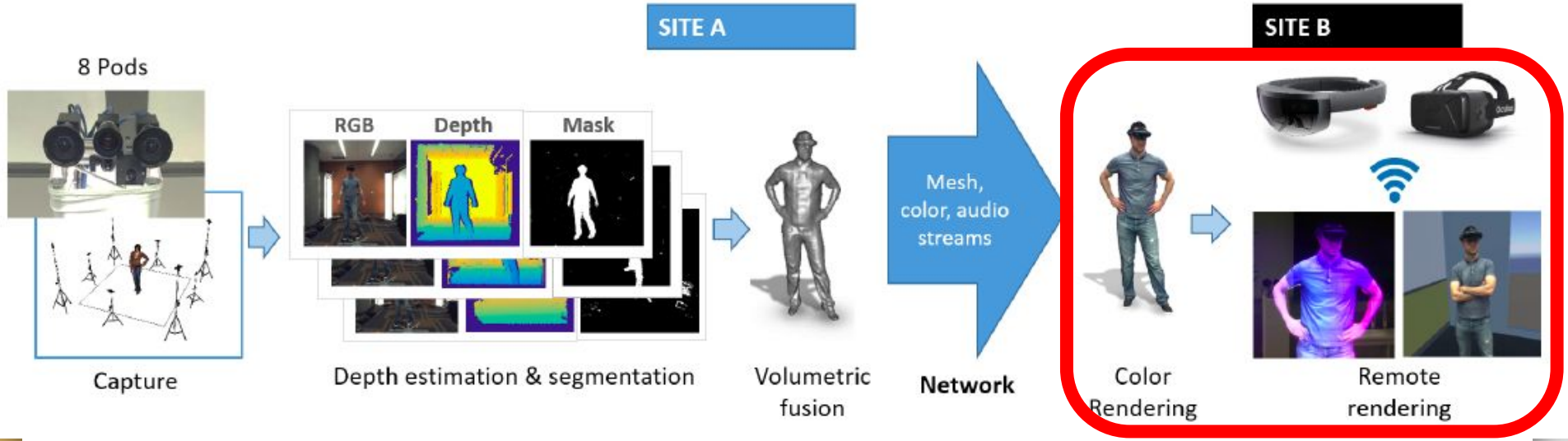

SITE A
SITE B
8 Pods
RGB
Depth
Mask
Mesh, color, audio streams
Network
Capture
Depth estimation & segmentation
Volumetri fusion
Color Rendering
Remote rendering

Compression

- To be real time and the highest quality, perform only a very lightweight real time compression
  - LZ4 compression
  - from 32MB to 3MB per frame

Transmission

- 1-2 Gbps transfer rate over TCP
- 10 Gbps link between stations
- Support 5-6 viewing clients
- The audio+pose data is transmitted independently, bi-directionally.
- Do not provide AV sync.


SITE A
SITE B
8 Pods
RGB
Depth
Mask
Mesh, color, audio streams
Capture
Depth estimation & segmentation
Volumetric fusion
Network
Color Rendering
Remote rendering

Render Offloading

- HMD transfer 6Dof pose to PC

  -> PC predicts a headset pose at render time

  performs scene rendering with that pose for each eye

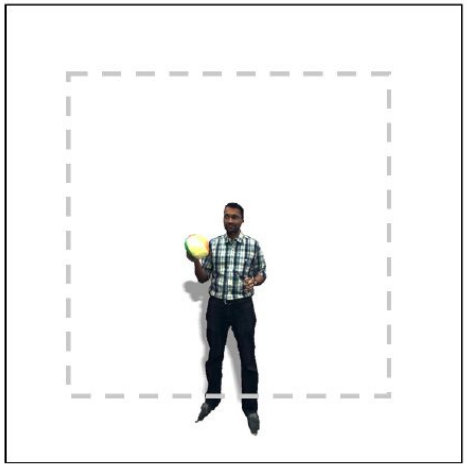
  encodes them and transmit it to the HMD

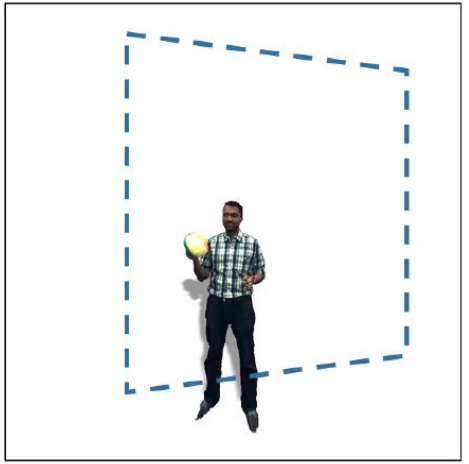
  -> HMD decode the stream and reprojecte to the latest user pose.

Latency Compensation

- The orientation misprediction
  - be compensated by rendering into a larger FoV (field of view) centered around the predicted user
- The small misprediction in rotation
  - renders with actual

    display FoV



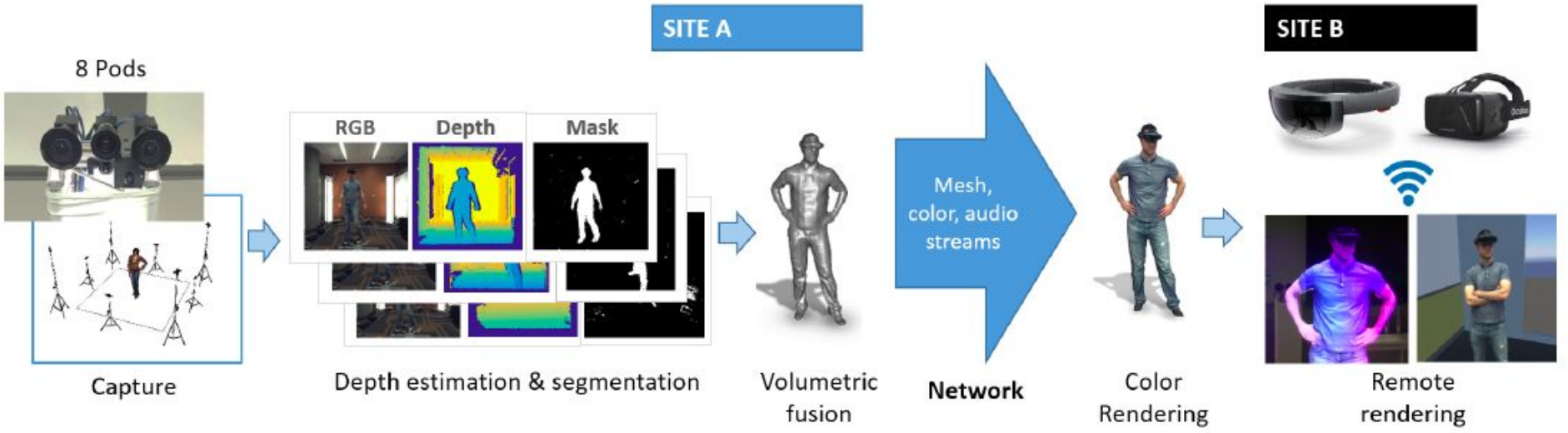


Latency Compensation

- The positional misprediction
  - perform view interpolation techniques as in [1]

[1] Lee, K., Chu, D., Cuervo, E., Kopf, J., Degtyarev, Y., Grizan, S., Wolman, A., and Flinn, J. Outatime: Using speculation to enable low-latency continuous interaction for mobile cloud gaming. In Proceedings of the 13th Annual International Conference on Mobile Systems Applications, and Services, ACM (2015), 151–165.

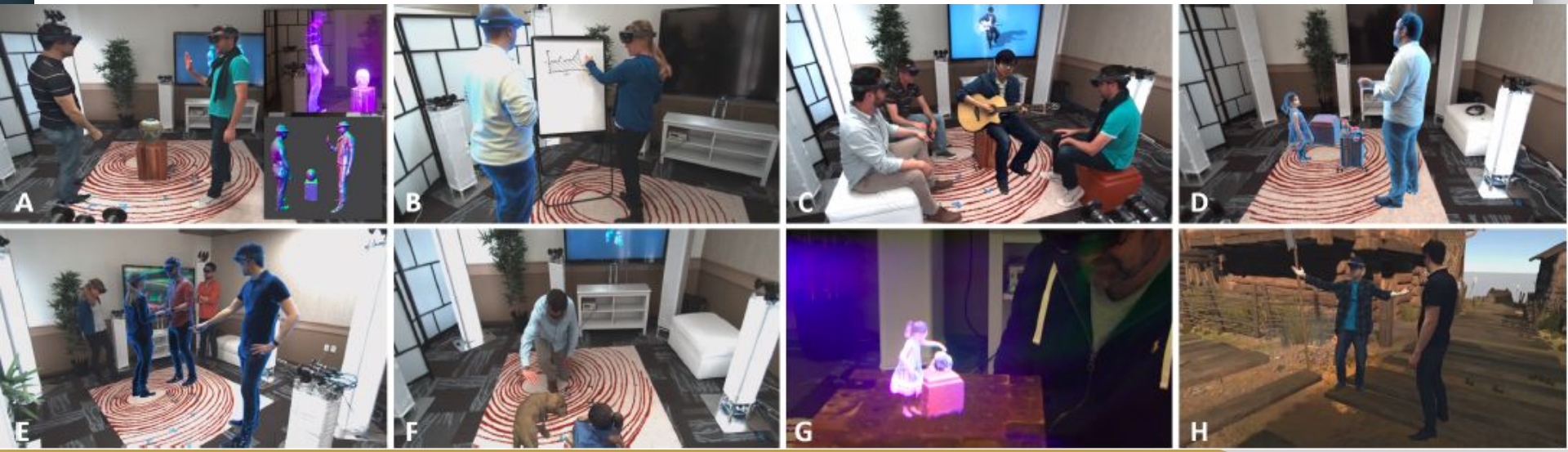
Pipeline



29 ms	29 ms	6 ms

Applications

- One-to-one
- One-to-many
- A Body in VR



Experiment: Setup

- 24 4MP resolution Grasshopper PointGrey cameras.
- Stereo cameras is 15 centimeters
  - giving an average error of 3 millimeters at 1 meter distance
  - 6 millimeters at 1.5 meter distance.
- 10 participants

Experiment: Social Interaction Task: Tell-a-lie

- State 3 pieces of information about themselves with one of the statements being false.
- The partner need to identify the false fact by asking any five questions.
- Accurately interpret verbal and non-verbal communication.

Experiment: Physical Interaction Task: Building Blocks

Results and Discussion

- 70% positive, 30% not so positive
- Interpersonal space awareness
- In AR
  - felt more realistic/natural
- In VR
  - users failed to determine if the block they were about to touch was real or not
  - suffer from latency

Limitations

- The amount of high-end hardware required to run the system is very high
- Currently, a 10 Gigabit Ethernet connection is used to communicate between rooms
- Need good compressing algorithm
- The 3D reconstruction of smaller geometry such as fingers produced artifacts
- Direct eye contact through headset removal is challenging,

Conclusion

- Design and implement an end-to-end system for high-quality and real-time capture, transmission and rendering of people, spaces, and objects in full 3D
- Demonstrated many different interactive scenarios
  - one-to-one communication
  - one to-many broadcast scenarios
  - live/real-time interaction
  - the ability to record and playback
- Need too many high-end hardwares
- Direct eye contact is challenging