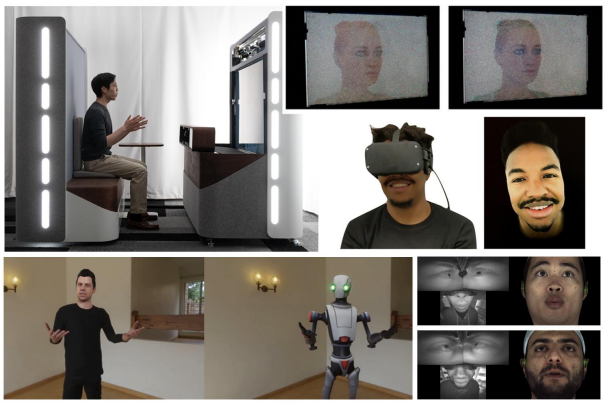
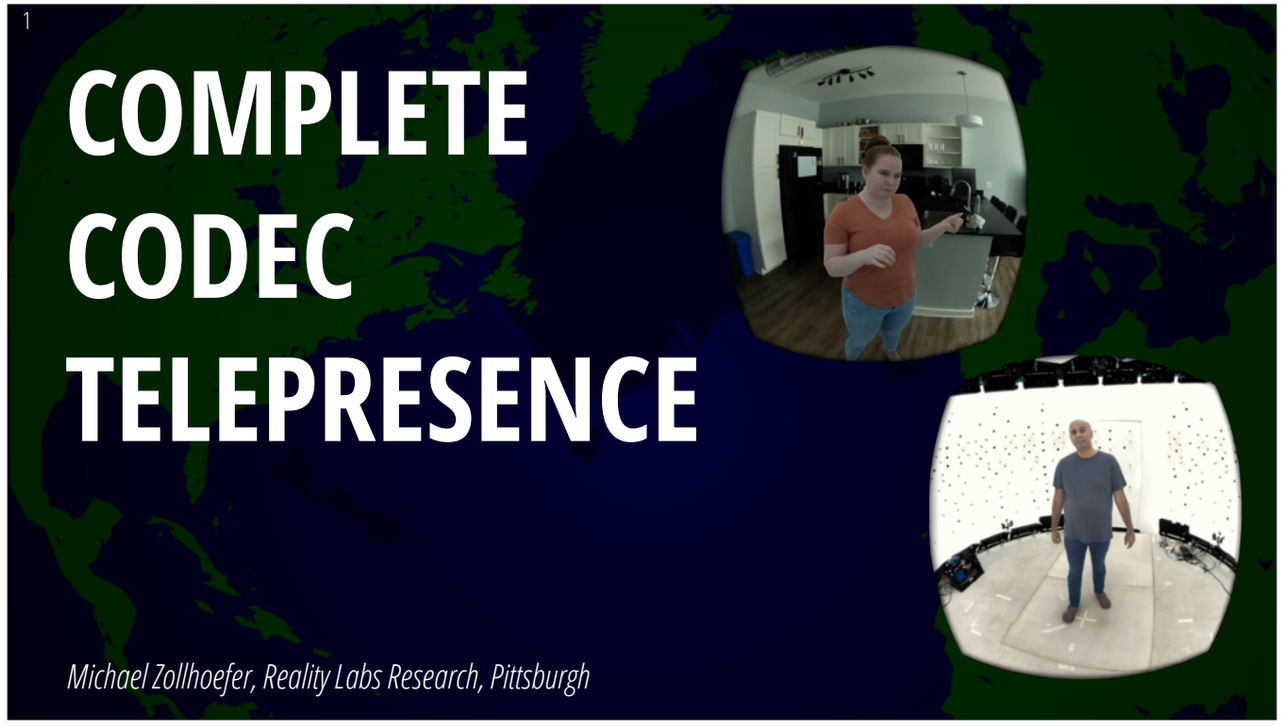
SIGGRAPH 2022
VANCOUVER+ 8-11 AUG

State of the Art in Telepresence

Part 2



SIGGRAPH '22 Courses, August 07-11, 2022, Vancouver, BC, Canada
ACM 978-1-4503-9362-1/22/08.
10.1145/3532720.3535679



Welcome everybody!
I am Michael a Research Scientist from Reality Labs Research in Pittsburgh.
Today, I want to talk about “Complete Codec Telepresence”, how this is related to the concept of the “Trinity of Telepresence”, and provide you with a deep dive into our technology to render photorealistic avatars and spaces.

Globe image from:
https://commons.wikimedia.org/wiki/File:Globe.svg
Public domain from the Creative Commons Corporation
Author: Augiasstallputzer~commonswiki

Other images/videos are Meta internal and not from third party works.

MY JOURNEY

My career in academia and later in industry created many opportunities to travel the world.
I lived on 2 different continents and in 6 different cities so far.
It has been an amazing journey and allowed me to gather many new impressions, interact with different cultures, and make friends all around the globe.

MY JOURNEY

Talking to my colleagues in the field, this live-style seems to be one of the common patterns for many of us working on computer graphics, computer vision, or machine learning, since the best opportunities for career growth normally are not co-located with where we are currently living.
Thus, until I joined RL Pittsburgh to work on telepresence, I was moving roughly once every 2-4 years.

Globe image from:
https://commons.wikimedia.org/wiki/File:Globe.svg
Public domain from the Creative Commons Corporation
Author: Augiasstallputzer~commonswiki

MY JOURNEY

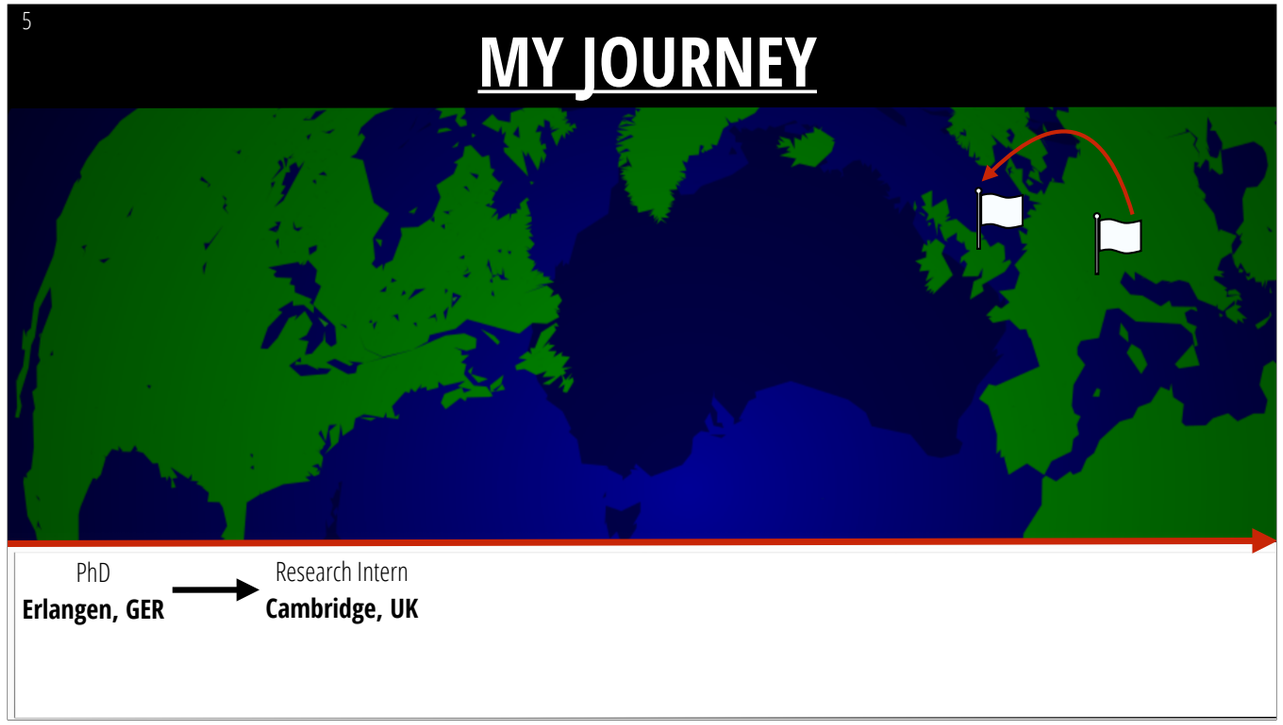
PhD

Erlangen, GER

I started my journey growing up in a small town called “Herzogenaurach” in Germany.
Fun fact, it is also the founding city of both Adidas and Puma.
I decided to go to the nearest university (Erlangen-Nuremberg), since this seemed like a convenient choice that allowed me to stay close to friends and family.
This is were I got exposed to computer graphics and real-time reconstruction research.

Globe image from:
https://commons.wikimedia.org/wiki/File:Globe.svg
Public domain from the Creative Commons Corporation
Author: Augiasstallputzer~commonswiki

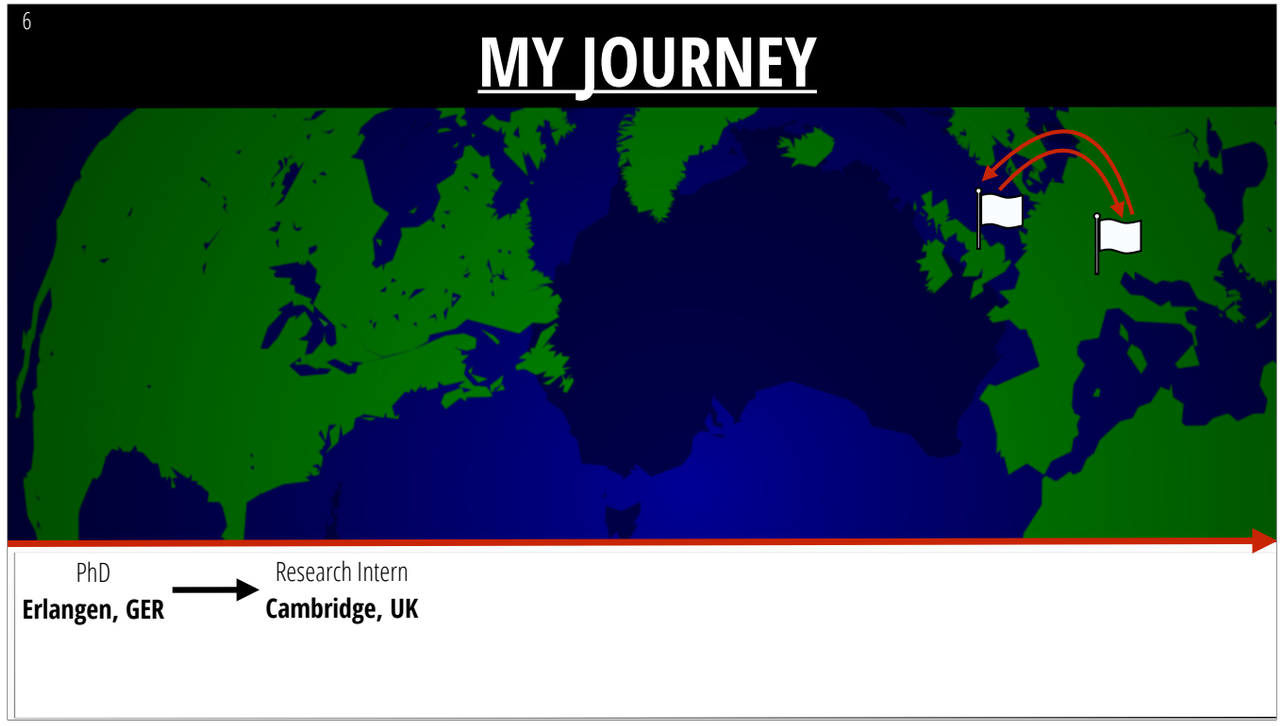
MY JOURNEY



Working on real-time reconstruction created the opportunity for me to do an internship at Microsoft Research in Cambridge, UK to work with Shahram Izardi on real-time deformable surface reconstruction, which we published at Siggraph in 2014. Yes, I’m that old.

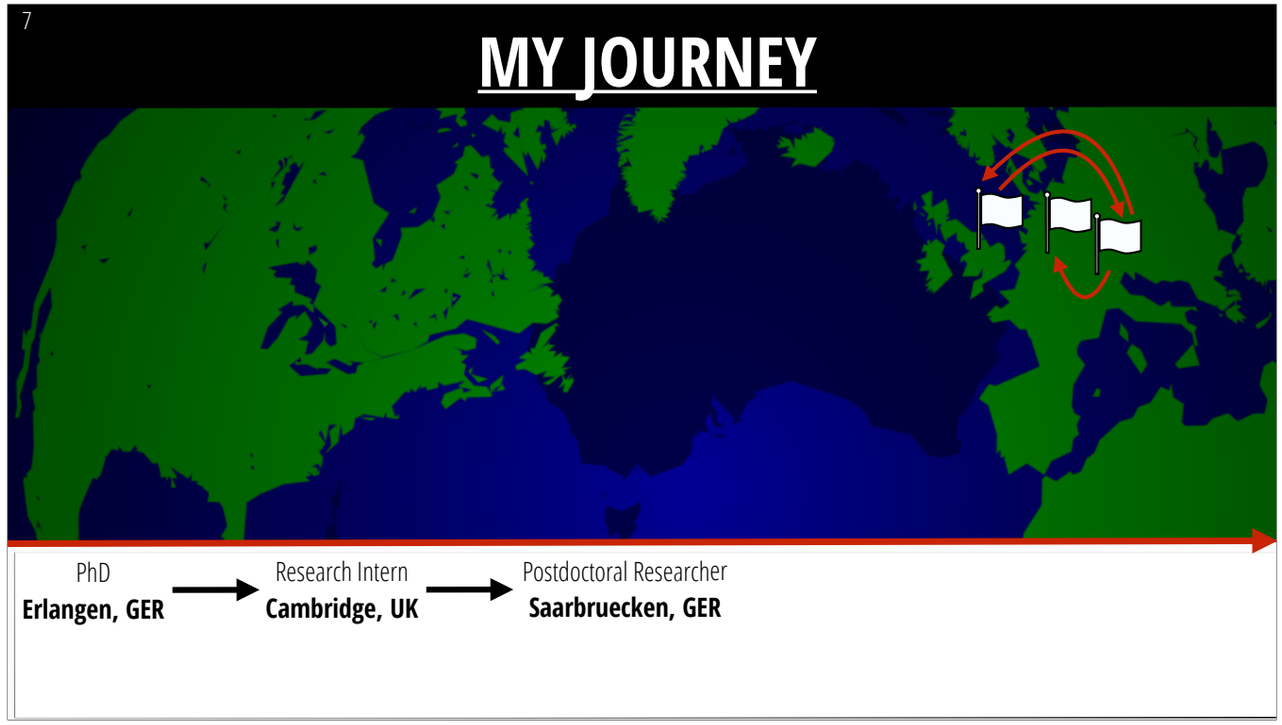
Globe image from:
https://commons.wikimedia.org/wiki/File:Globe.svg
Public domain from the Creative Commons Corporation
Author: Augiasstallputzer~commonswiki

MY JOURNEY



After my internship, I went back to the University of Erlangen-Nuremberg to finished my PhD in the same year.
One of the collaborators during my internship project at Microsoft was Christian Theobalt from MPI Informatics.

Globe image from:
https://commons.wikimedia.org/wiki/File:Globe.svg
Public domain from the Creative Commons Corporation
Author: Augiasstallputzer~commonswiki



He offered me a postdoctoral researcher position and thus the next step in my career was suddenly clear.
I moved to Saarbruecken to join MPI. Saarbruecken was still relatively close to home, a short 4h drive with the car. Given that german highways do not have a speed limit, it is also possible in 3h on a good day.
At MPI I got exposed to computer vision and also had to catch up on deep learning after all PhD students suddenly switched from classical vision to learning-based approaches within only half a year.
As a postdoctoral researcher that had never trained a neural network before, this was a steep learning curve, but it was a fun experience to learn something new.

Globe image from:
https://commons.wikimedia.org/wiki/File:Globe.svg
Public domain from the Creative Commons Corporation
Author: Augiasstallputzer~commonswiki

At the end of my postdoc, an opportunity arose to go to Stanford as a Visiting Professor as part of a research program of the Max Planck Center for Visual Computing and Communication.
This was an opportunity I could not decline, so I moved to Stanford to work on the new emerging field of neural rendering.
Little did I know how hard it would be over the course of the next two year to stay in touch with friends, family, and collaborators in Europe.
Trying to maintain social connections via messaging, phone, VC calls turned out be really hard.

Globe image from:
https://commons.wikimedia.org/wiki/File:Globe.svg
Public domain from the Creative Commons Corporation
Author: Augiasstallputzer~commonswiki

After two years, another opportunity arose. I got an offer to work on VR telepresence at Reality Labs Research in Pittsburgh, where I have been ever since.
This offer resonated with we, since the goal was to create technology for people to better communicate, collaborate, and maintain their social bonds independent of their physical location on our planet.
In short, the goal was to develop a system to “defy” physical distance and bring the world closet together.
In the following, I want to introduce to you the technical components of the system we are working on at Reality Labs Pittsburgh, our research that will enable it, and our vision for the future of telepresence.

Globe image from:
https://commons.wikimedia.org/wiki/File:Globe.svg
Public domain from the Creative Commons Corporation
Author: Augiasstallputzer~commonswiki

Other images/videos are Meta internal and not from third party works.

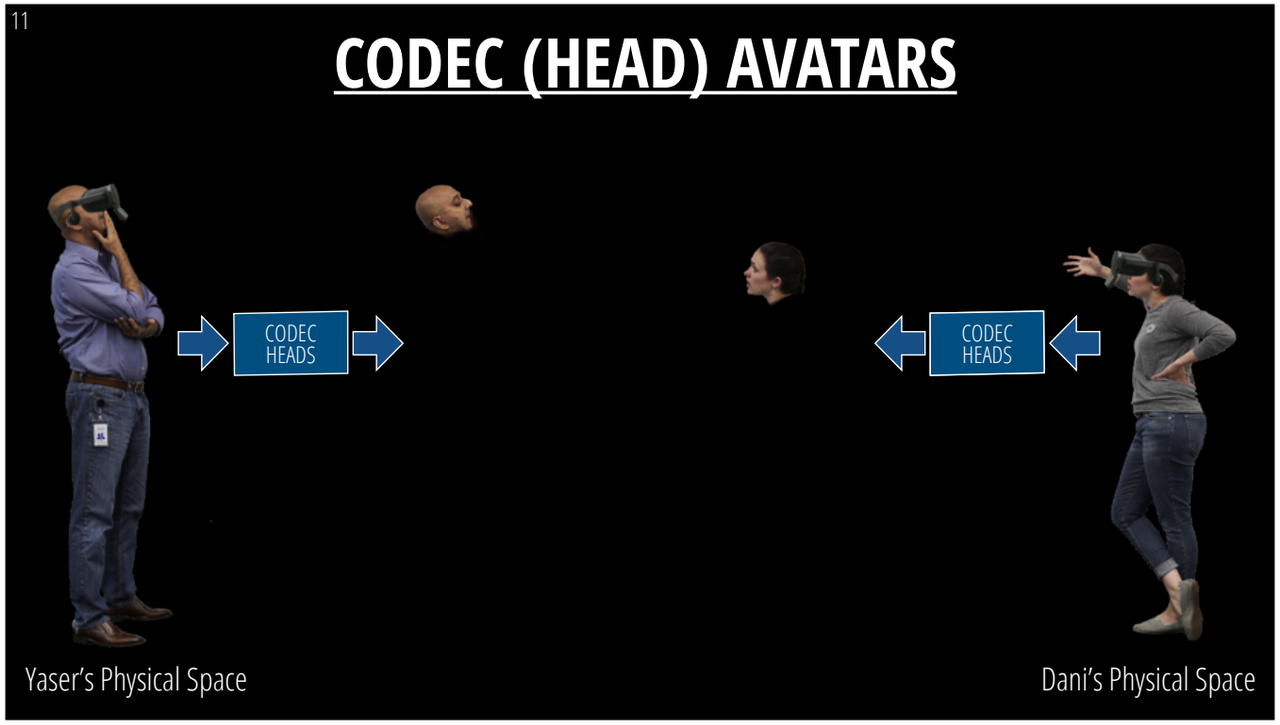
VR TELEPRESENCE

Yaser's Physical Space

Dani's Physical Space

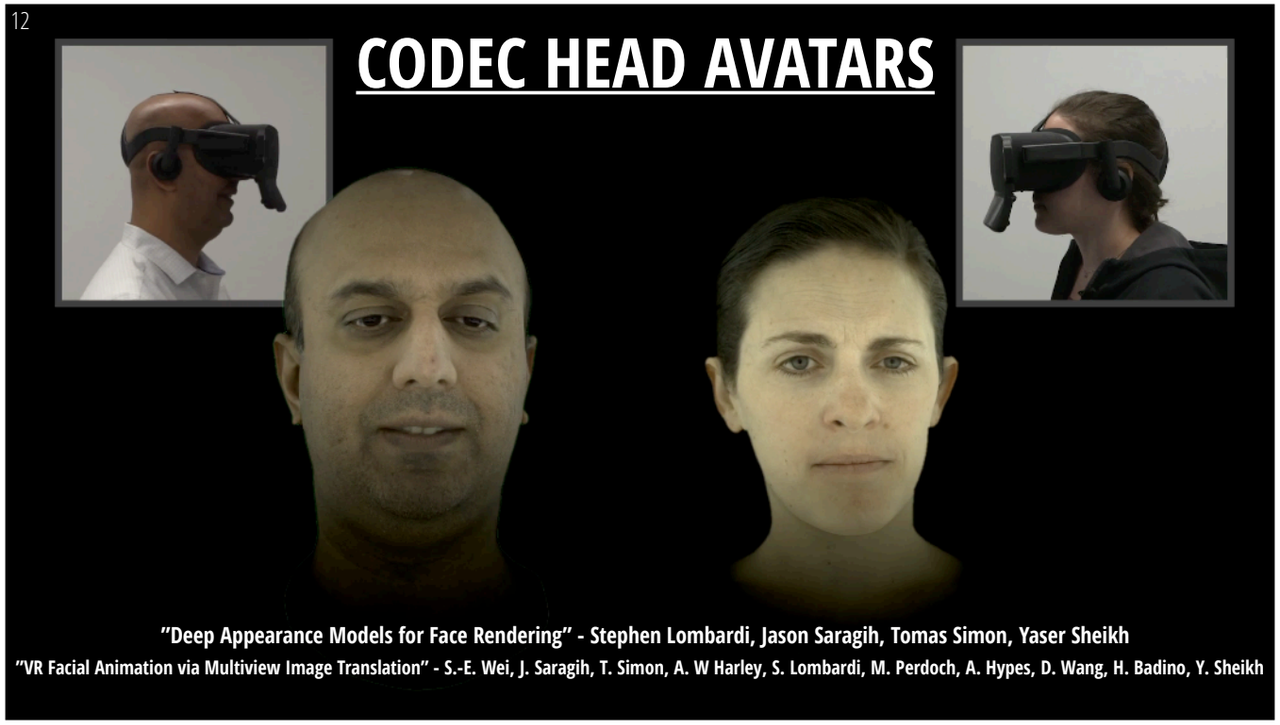
Imagine two people, each of them within the safety of their own home, being able to meet with each other virtually in a photorealistic "Metaverse" that enables them to communicate and interact as if both of them were present in the same shared physical space.
Enabling such an experience is one of the grant challenges for Reality Labs Research in Pittsburgh.
To this end, we want to build a complete codec telepresence system that is indistinguishable from reality.

Images/videos are Meta internal and not from third party works.



Pittsburgh has already taken several steps towards this goal.
The first of which was what we call “Codec (Heads) Avatars”.
Codec Heads are photorealistic digital clones of a human head that can be driven in real-time to enable a symmetric telepresence experience.
The following video gives a great overview of what such an experience looks like.

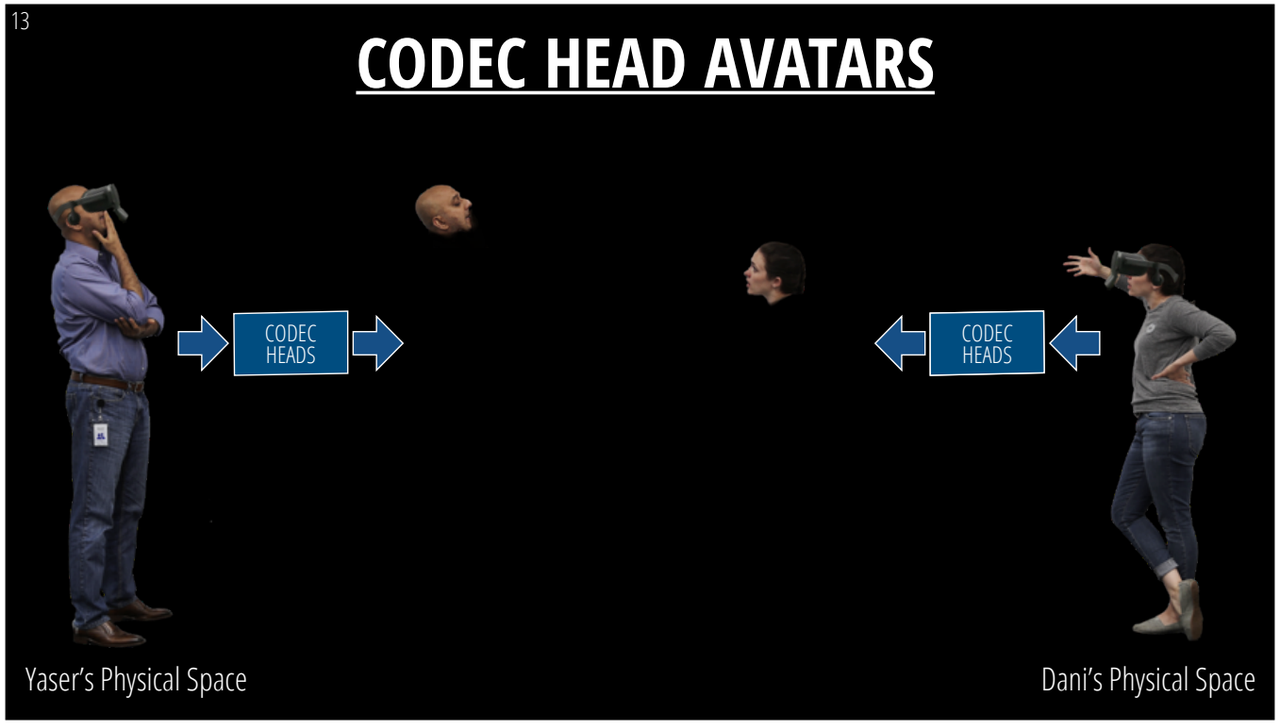
Images/videos are Meta internal and not from third party works.



Play video with sound.

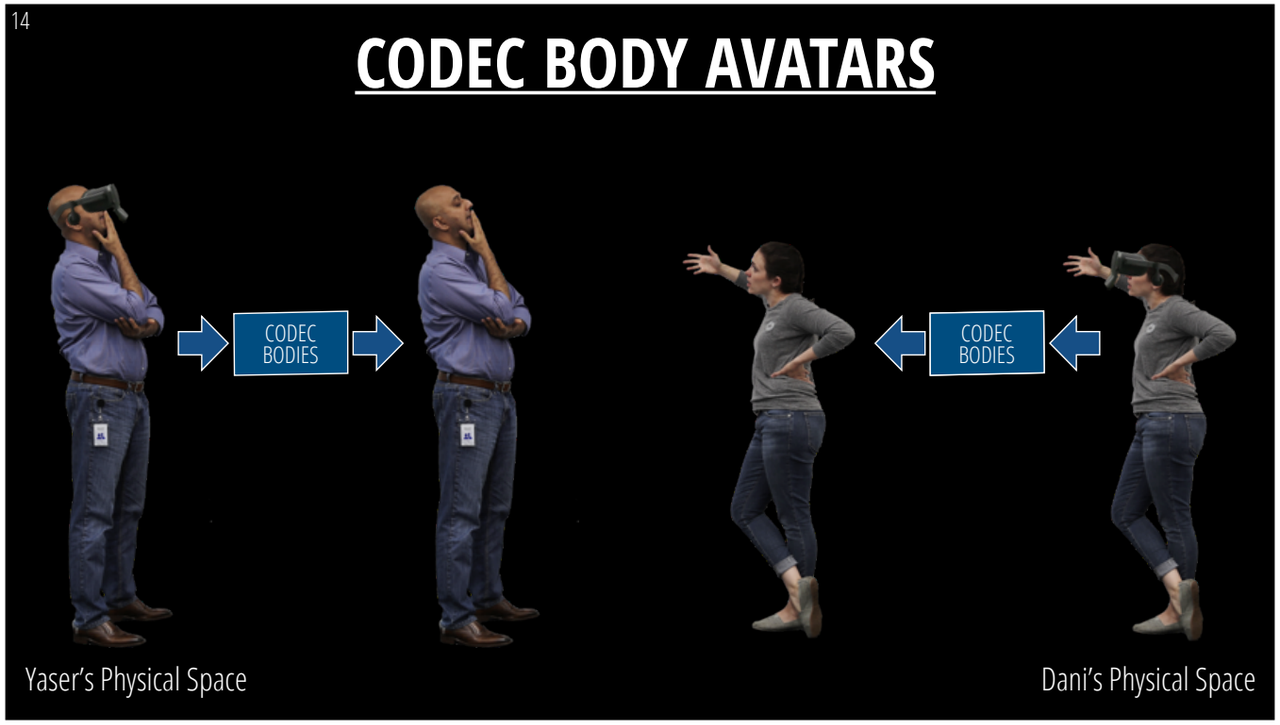
Combination of two approaches:
Deep appearance models for face rendering
Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018.
ACM Trans. Graph. 37, 4, Article 68 (August 2018), 13 pages. https://doi.org/10.1145/3197517.3201401
+
VR facial animation via multiview image translation
Shih-En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. 2019.
ACM Trans. Graph. 38, 4, Article 67 (August 2019), 16 pages. https://doi.org/10.1145/3306346.3323030

Images/videos are Meta internal and not from third party works.



These are really impressive results for real-time photorealistic animation of digital models of humans, but there is clearly something important missing here.
As you probably noticed, only the head is modeled, thus we have a set of floating heads in front of a black void talking to each other, thus clearly these results are not indistinguishable from reality.

Images/videos are Meta internal and not from third party works.



The first step towards “Complete Codec Telepresence” is to push for body complete avatars.
These are avatars that have a realistic head, including hair, hands, arms, and the rest of the body, including clothing.
In the following, I will show you a video that shows you how an asymmetric telepresence experience with full body avatars looks like.

Images/videos are Meta internal and not from third party works.

15

CODEC BODY AVATARS

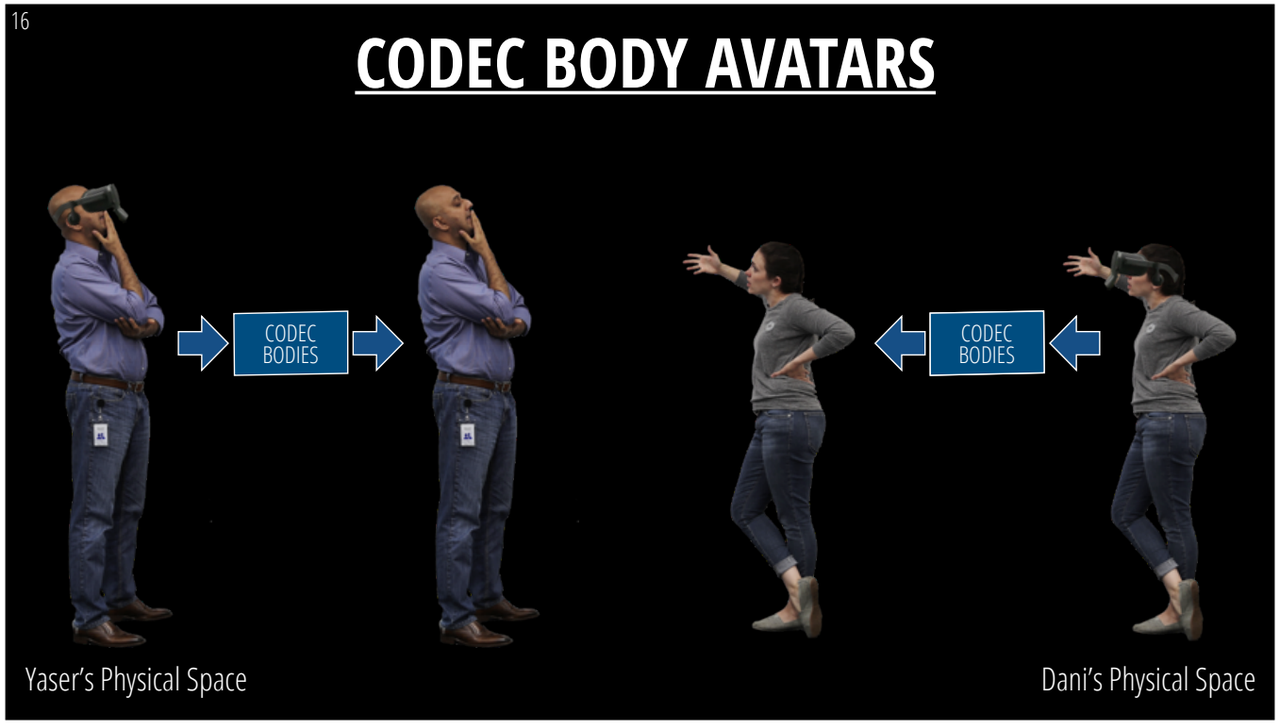
Live VR View

RX Side

[T. Bagautdinov, A. Pahuja, C. Wu, A. Richardt, Y.-S Shih , H. Mehta, E. Fath, M. Zollhoefer, Y. Sheikh]

Play video with sound.

Images/videos are Meta internal and not from third party works.



This is already a big step forward.
Yaser's avatar is "body complete" and Elaine is able to freely walk around him and pick her favorite view-point.
But Yaser's avatar still looks dislocated and floats on top of a black void.
What is missing here?

Images/videos are Meta internal and not from third party works.

Clearly, a shared virtual space.
Note, that adding a virtual space requires us to also model the interactions of the avatars with the environment in terms of light and sound transport.
For example, the shadows the avatar casts on the wall and the ones the avatar casts on the floor.
In addition, we have to relight the avatars correctly based on the incoming radiance from the surrounding space.
In the following, I will show you a video that shows you how this looks like.

Images/videos are Meta internal and not from third party works.

COMPLETE CODEC TELEPRESENCE

Live VR View

RX Side

[T. Bagautdinov, A. Pahuja, C. Wu, A. Richardt, Y.-S Shih , H. Mehta, E. Fath, M. Zollhoefer, Y. Sheikh]

Play video with sound.

Images/videos are Meta internal and not from third party works.

A two-sided, symmetric version of this experience, that is able to teleport both people to an arbitrary environment, is what I am referring two as “Complete Codec Telepresence”.
It will allow you in the future to “teleport” to a shared virtual space with your family, friends, and colleagues, such that you can communicate and interact with them as if all of you would be physical present in the same space.

Images/videos are Meta internal and not from third party works.

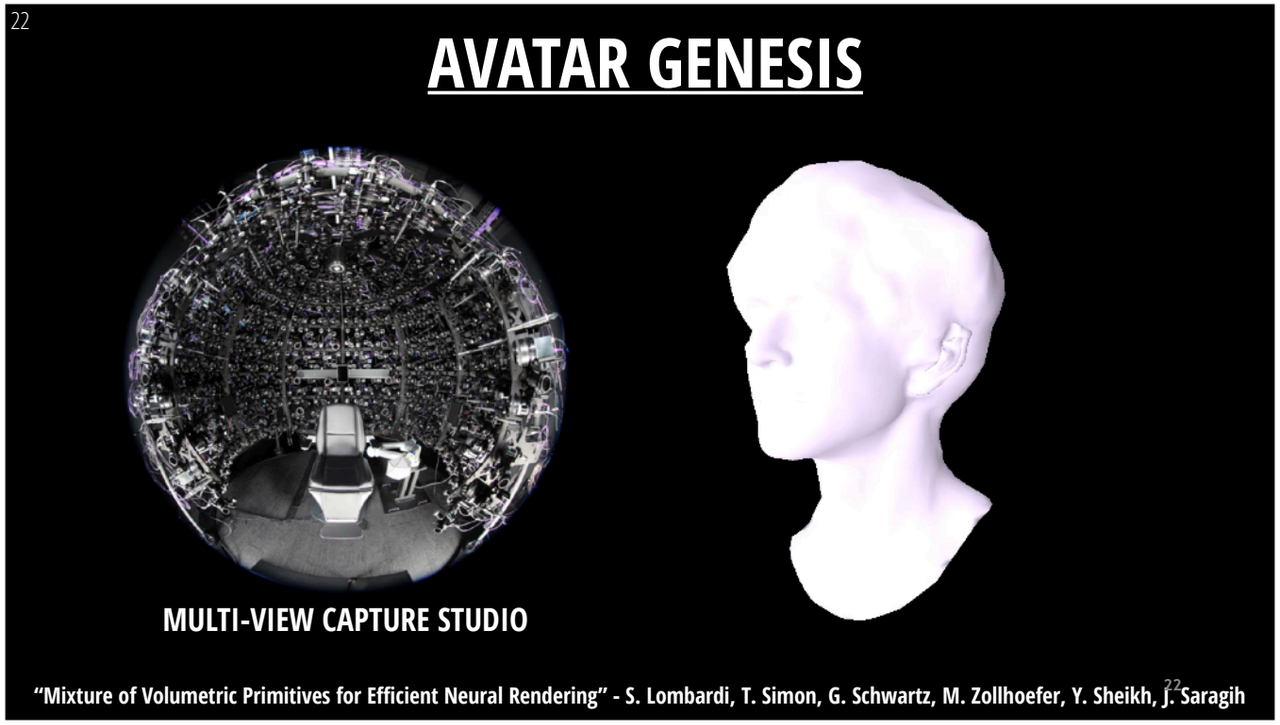
TRINITY OF TELEPRESENCE

To achieve “Complete Codec Telepresence” we need to work on the “Trinity of Telepresence”.

TRINITY OF TELEPRESENCE

The first of the three components of the trinity are photorealistic and drivable representations of humans, which we call “avatars”.

Images/videos are Meta internal and not from third party works.



Our avatars are created based on data-driven machine learning approaches that are fueled by large amounts of captured data from multi-camera capture systems.
Here you can see our volumetric upper body avatars in action.

Video results from:
Mixture of volumetric primitives for efficient neural rendering
Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih.
ACM Trans. Graph. 40, 4, Article 59 (August 2021), 13 pages. https://doi.org/10.1145/3450626.3459863
Fair Use.
This is a course/talk with the purpose of education/teaching.
Video result from related work for illustrative purpose.

VOLUMETRIC AVATARS

"Mixture of Volumetric Primitives for Efficient Neural Rendering" - S. Lombardi, T. Simon, G. Schwartz, M. Zollhoefer, Y. Sheikh, J. Saragih

The newest version of these avatars is based on a mixture of volumetric primitives.
This enables the avatars to be more complete and have realistic hair.
A full body version of such volumetric avatars will also be presented in the technical paper session of this Siggraph.
These avatars feature the full head, upper body, lower body, hair, clothing, and more.

Video results from:
Mixture of volumetric primitives for efficient neural rendering
Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih.
ACM Trans. Graph. 40, 4, Article 59 (August 2021), 13 pages. https://doi.org/10.1145/3450626.3459863
Fair Use.
This is a course/talk with the purpose of education/teaching.
Video result from related work for illustrative purpose.

“Mixture of Volumetric Primitives for Efficient Neural Rendering” - S. Lombardi, T. Simon, G. Schwartz, M. Zollhoefer, Y. Sheikh, J. Saragih

To be useful for telepresence, we have to be able to render our avatars in real-time.
More precisely, we have to be able to render two high resolution images, e.g., for Quest2 2x1832x1920pixels at >72Hz.
This leaves only ~14ms to render the binocular imagery.

Video results from:
Mixture of volumetric primitives for efficient neural rendering
Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih.
ACM Trans. Graph. 40, 4, Article 59 (August 2021), 13 pages. https://doi.org/10.1145/3450626.3459863
Fair Use.
This is a course/talk with the purpose of education/teaching.
Video result from related work for illustrative purpose.

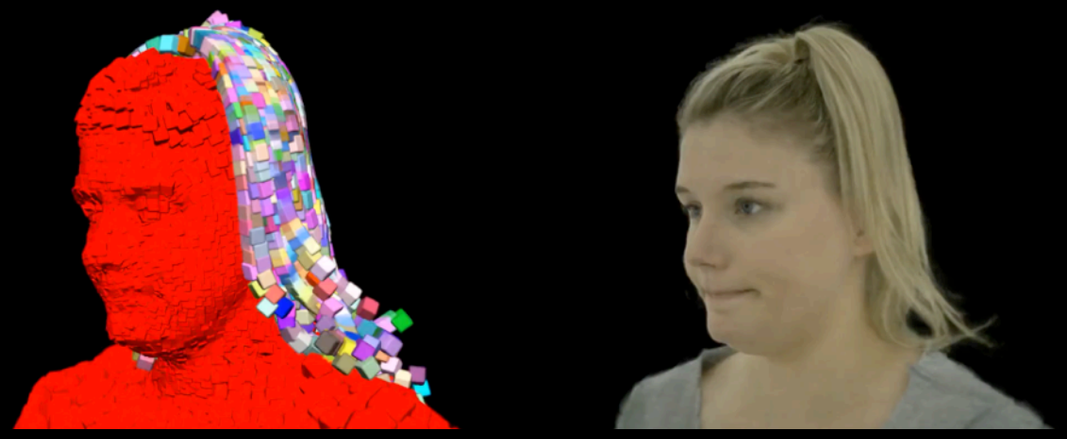
HAIR TRACKING

"HVH: Learning a Hybrid Neural Volumetric Representation for Dynamic Hair Performance Capture" - Z. Wang, G. Nam, T. Stuyck, S. Lombardi, M. Zollhoefer, J. Hodgins, C. Lassner

One important component of human heads is hair.
Tracking hair is highly challenging due to the large number of hair (100k+), challenging non-rigid motion, and strong occlusions.
Here, you can see the output of a hair tracking approach we have developed.

Video results from:
HVH: Learning a Hybrid Neural Volumetric Representation for Dynamic Hair Performance Capture.
Wang, Ziyan & Nam, Giljoo & Stuyck, Tuur & Lombardi, Stephen & Zollhöfer, Michael & Hodgins, Jessica & Lassner, Christoph
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
Fair Use.
This is a course/talk with the purpose of education/teaching.
Video result from related work for illustrative purpose.

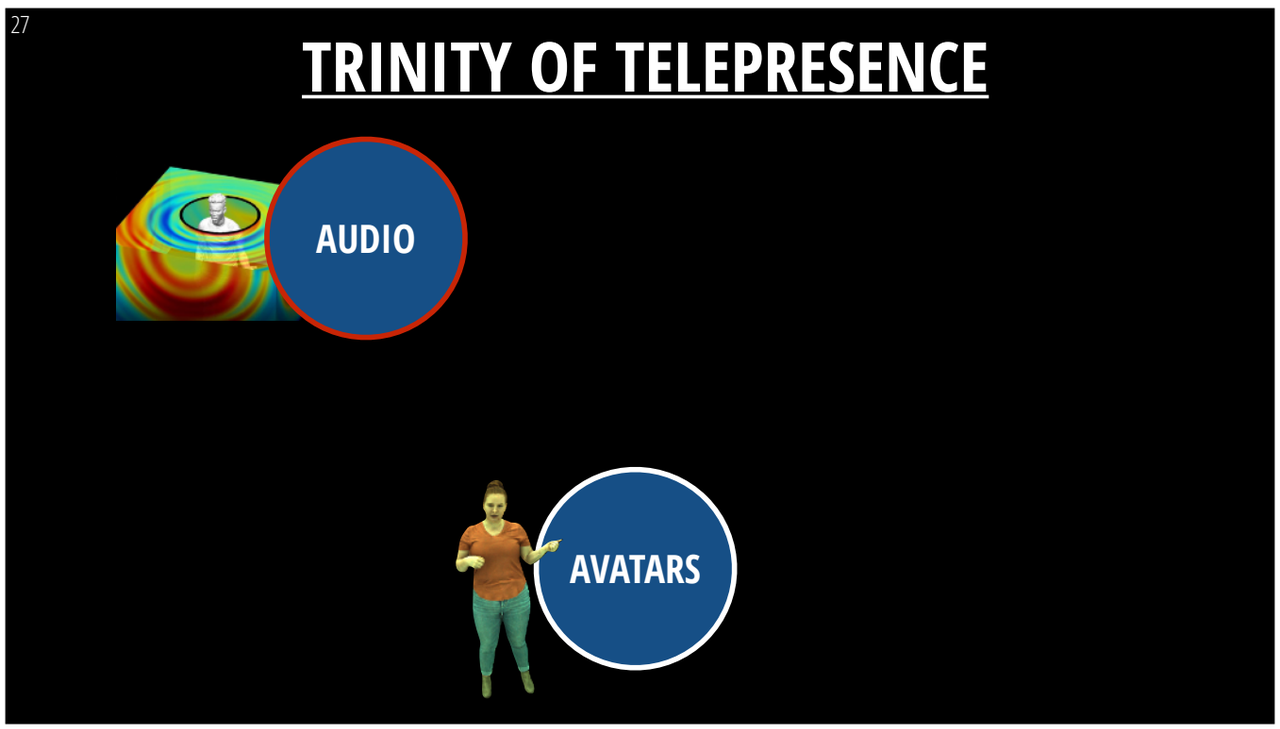
NEURAL HAIR RENDERING



"HVH: Learning a Hybrid Neural Volumetric Representation for Dynamic Hair Performance Capture" - Z. Wang, G. Nam, T. Stuyck, S. Lombardi, M. Zollhoefer, J. Hodgins, C. Lassner

This strand based tracking enables us to place a set of volumetric primitives along the guide strands.
This more physically based representation of hair allows for high quality hair rendering and animation.

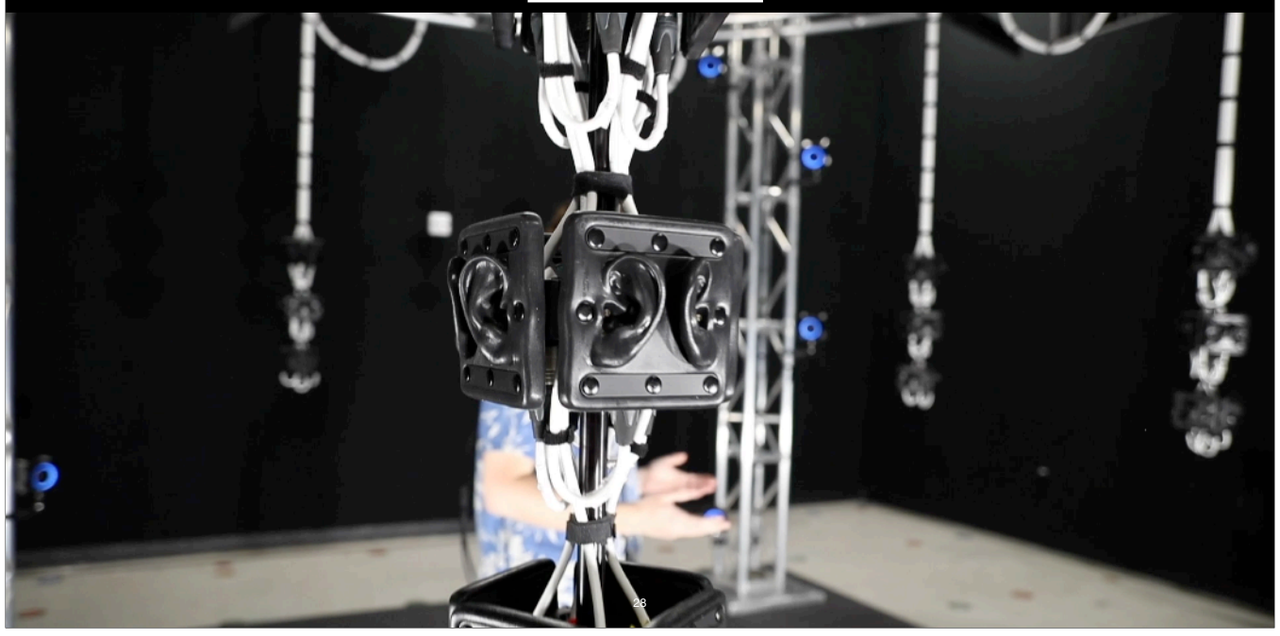
Video results from:
HVH: Learning a Hybrid Neural Volumetric Representation for Dynamic Hair Performance Capture.
Wang, Ziyan & Nam, Giljoo & Stuyck, Tuur & Lombardi, Stephen & Zollhöfer, Michael & Hodgins, Jessica & Lassner, Christoph
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
Fair Use.
This is a course/talk with the purpose of education/teaching.
Video result from related work for illustrative purpose.



Besides a photorealistic renderable representation of yourself that looks and moves like you, we also require avatars that sounds like you.
By this, I do not only mean the tone of your voice, but we also have to be able to generate binaural audio, i.e., creating a stereo signal that takes the relative position and orientation between the receiver and the transmitter into account.

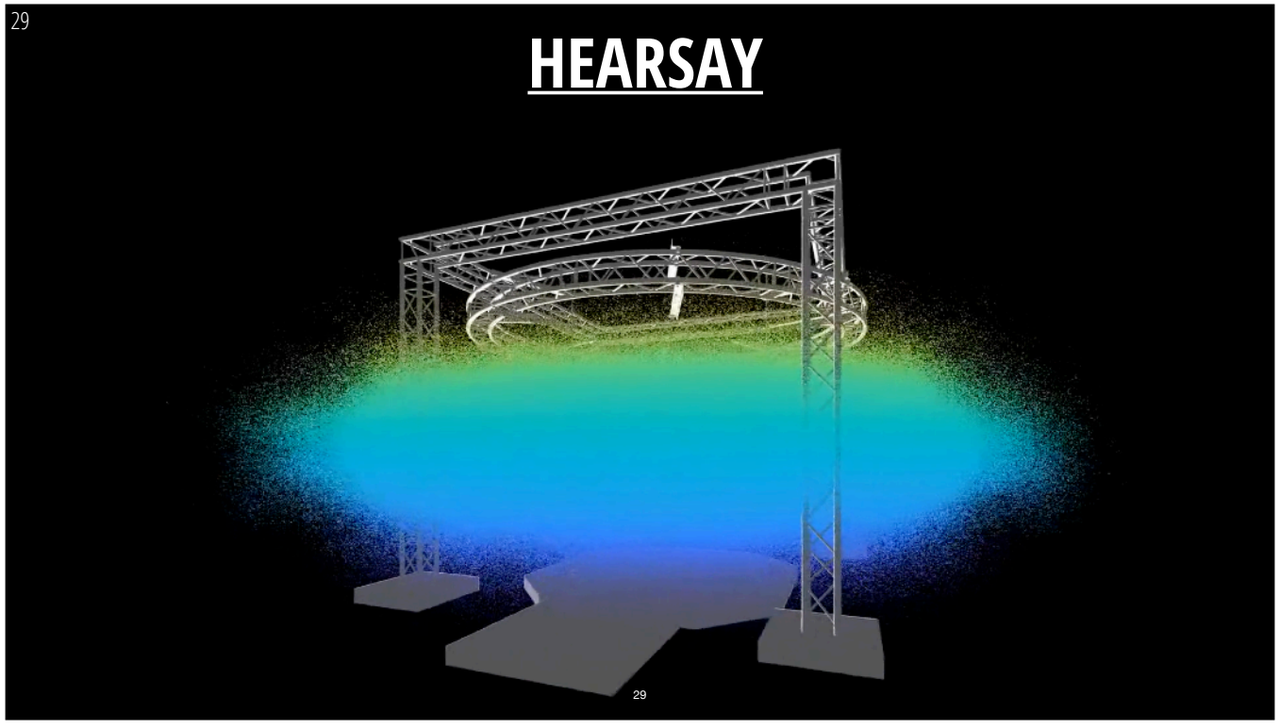
Images/videos are Meta internal and not from third party works.

HEARSAY



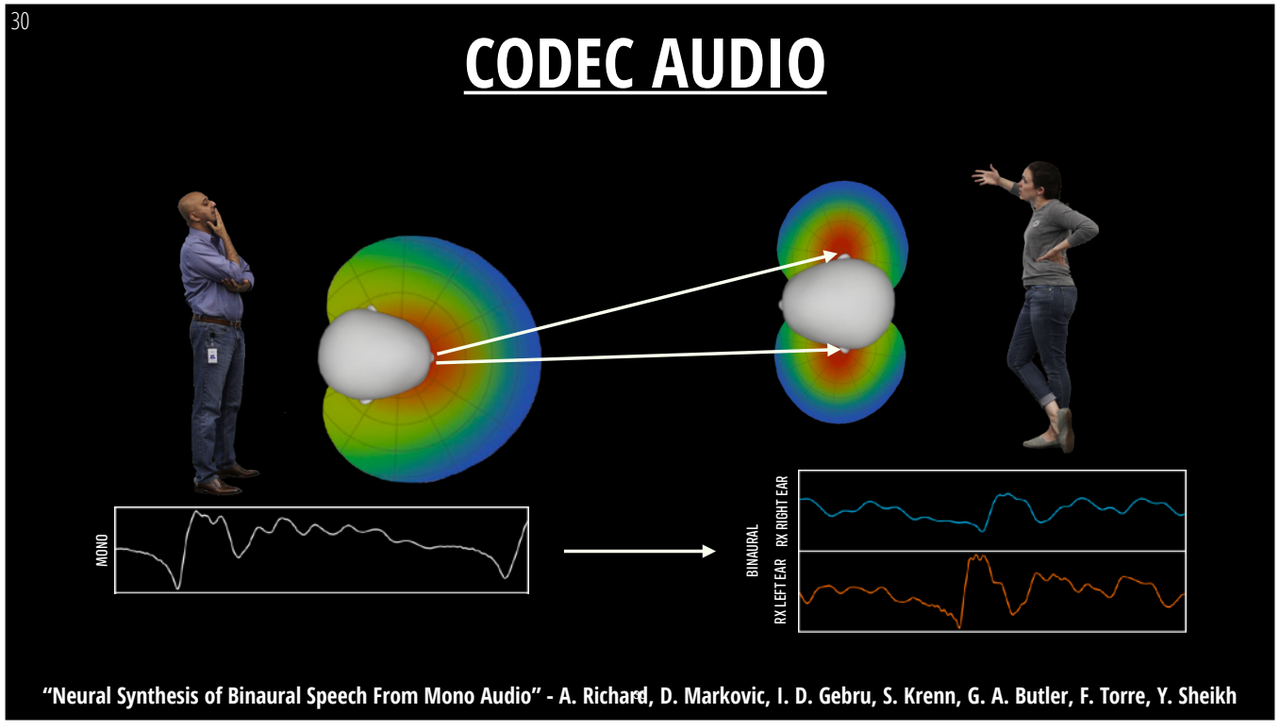
Similar to how we create the visual representation of our avatars, we learn sound binauralization using a lot of data.
This is where Hearsay comes into play.
Hearsay is one of our audio capture stages and has hundreds of ears (to bake a standard HRTF into the captured signal) with microphones.
This is in direct analogy to our multi-camera capture systems, which have hundreds of cameras.

Images/videos are Meta internal and not from third party works.



This enables us to obtain dense spatial sampling in the entire space by exploiting symmetry and a free-field sound propagation assumption.
Each point here represents one second of recorded audio.

Images/videos are Meta internal and not from third party works.



This data enables us to train models that can predict how the transmitter sounds from the perspective of the two ears of the receiver.

Approach as in:
Neural Synthesis of Binaural Speech From Mono Audio
Alexander Richard and Dejan Markovic and Israel D. Gebru and Steven Krenn and Gladstone Alexander Butler and Fernando Torre and Yaser Sheikh
International Conference on Learning Representations, 2021

Images/videos are Meta internal and not from third party works.

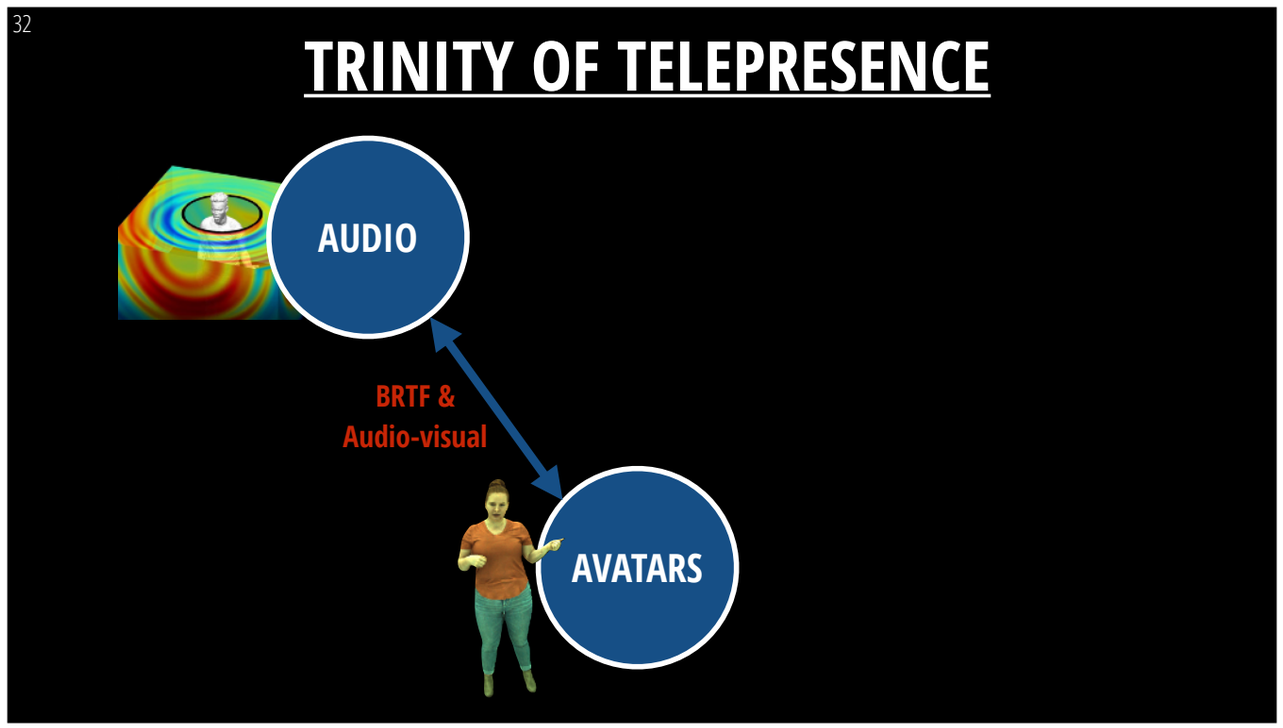
CODEC AUDIO

"Neural Synthesis of Binaural Speech From Mono Audio" - A. Richard, D. Markovic, I. D. Gebru, S. Krenn, G. A. Butler, F. Torre, Y. Sheikh

Here is one example.

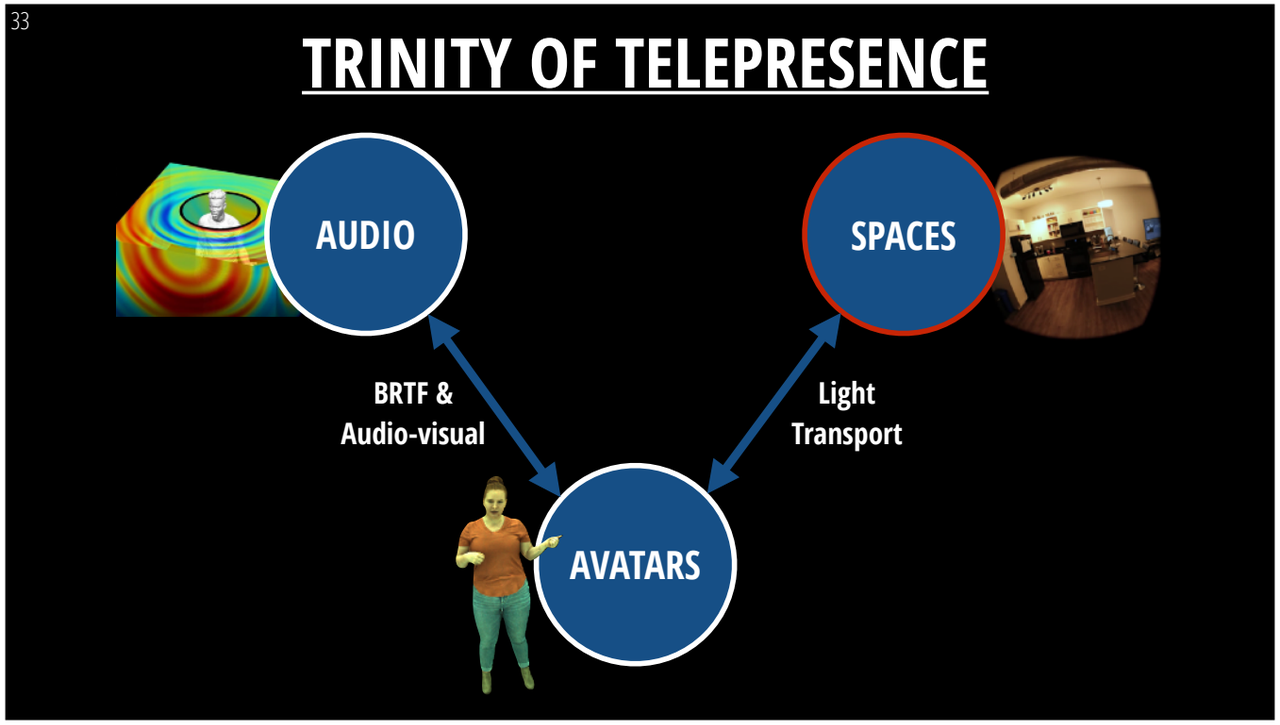
Approach as in:
Neural Synthesis of Binaural Speech From Mono Audio
Alexander Richard and Dejan Markovic and Israel D. Gebru and Steven Krenn and Gladstone Alexander Butler and Fernando Torre and Yaser Sheikh
International Conference on Learning Representations, 2021

Images/videos are Meta internal and not from third party works.



The representations for audio and avatars are tightly connected.
Audio can be used to improve avatar animation, i.e., enforce lip closure, while the shape of the avatar (and their ears) can informs us about the BRTF.

Images/videos are Meta internal and not from third party works.



The final and third component of the Trinity of Telepresence is the surrounding space.

Images/videos are Meta internal and not from third party works.

CODEC SPACES

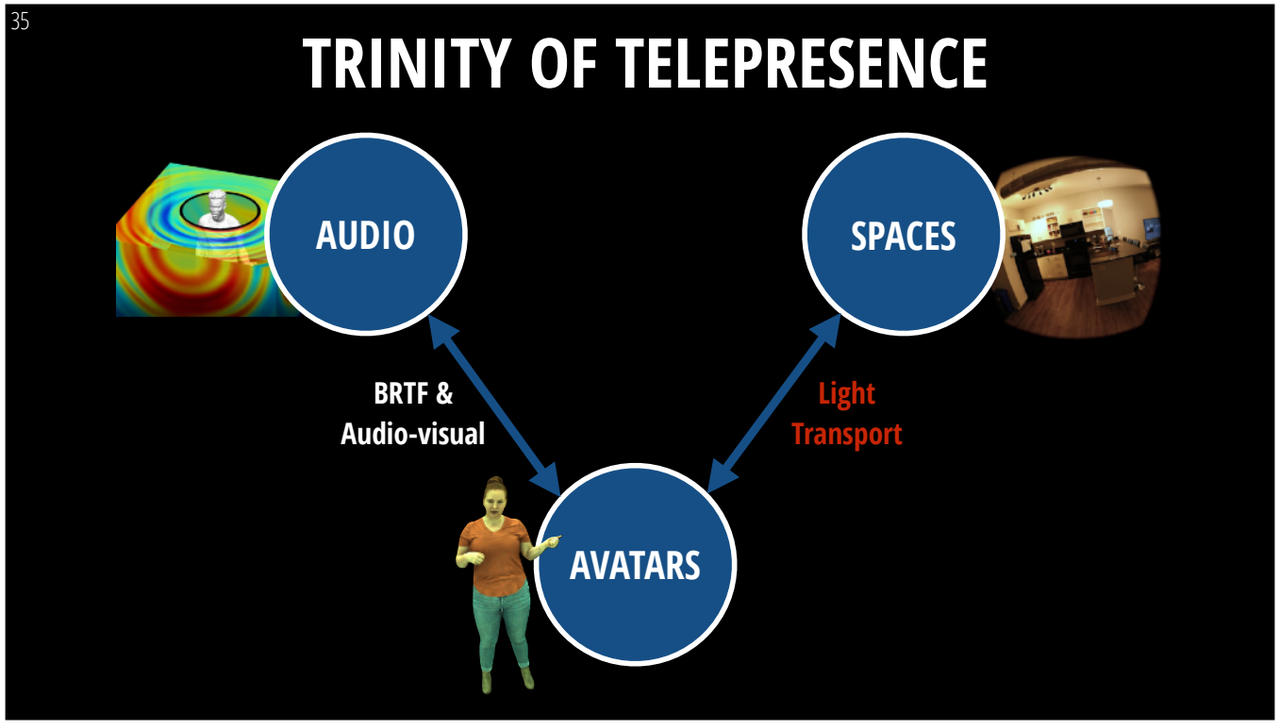
Live VR View

RX Side

[Y.-S. Shih, V. Agrawal, M. Zollhoefer]

To this end, we require real-time 360-degree novel view synthesis techniques that are fast enough to run at VR frame rates and resolutions (~12.5ms to render two high resolution images).

Images/videos are Meta internal and not from third party works.



The representations for spaces and avatars are tightly connected via the light transport in the scene.
The avatars influence the appearance of the space, e.g., by casting shadows on the floor.
In addition, the space influences the appearance of the avatar, i.e., to seamlessly blend in the avatar it has to be relit based on its location and orientation within the space.

Images/videos are Meta internal and not from third party works.

RELIGHTABLE AVATARS

Relightable avatar capture works similar to fully-lit avatar capture, but we also have to observe the human under various different illumination condition. To this end, we employ time multi-flexed lighting in our multi-camera capture system.

Combination of two approaches:
Deep relightable appearance models for animatable faces
Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn McPhail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih.
ACM Trans. Graph. 40, 4, Article 89 (August 2021), 15 pages. https://doi.org/10.1145/3450626.3459829
+
Mixture of volumetric primitives for efficient neural rendering
Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih.
ACM Trans. Graph. 40, 4, Article 59 (August 2021), 13 pages. https://doi.org/10.1145/3450626.3459863

Images/videos are Meta internal and not from third party works.

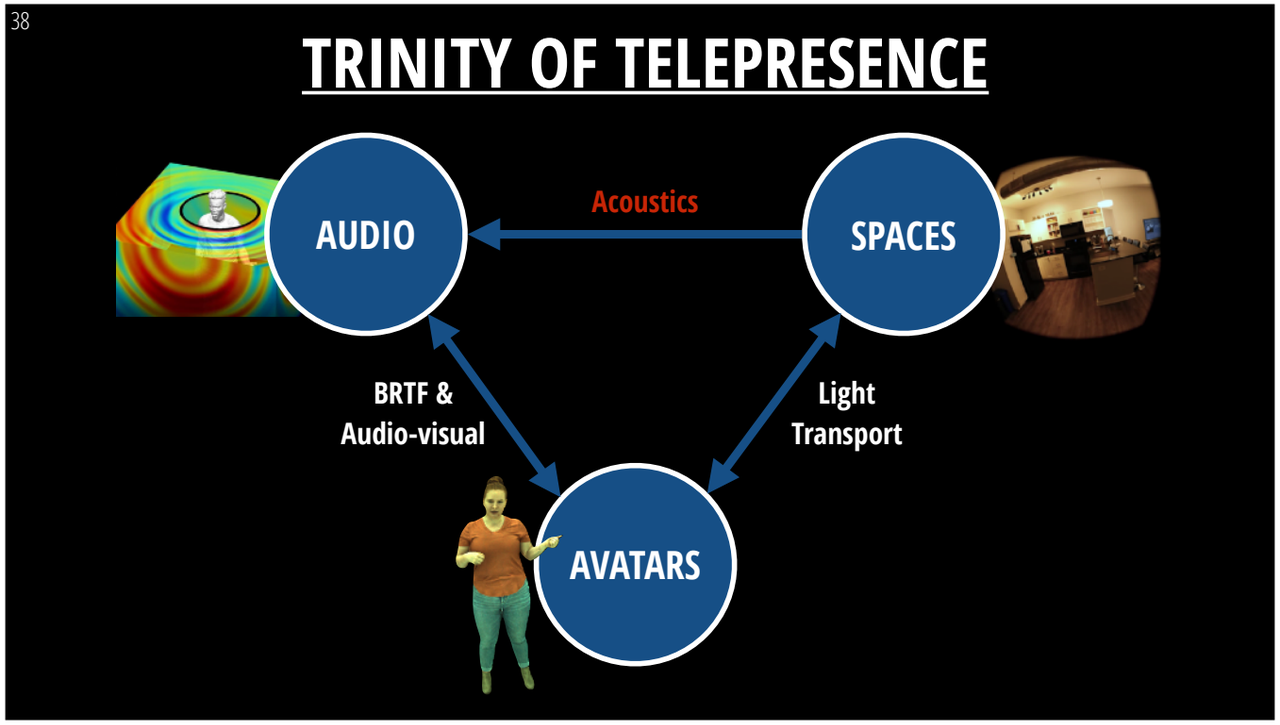
“Deep Relightable Appearance Models for Animatable Faces” - S. Bi, S. Lombardi, S. Saito, T. Simon, S.-E. Wei, K. McPhail, R. Ramamoorthi, Y. Sheikh, J. Saragih

“Mixture of Volumetric Primitives for Efficient Neural Rendering” - S. Lombardi, T. Simon, G. Schwartz, M. Zollhoefer, Y. Sheikh, J. Saragih

Given the collected data, we can train neural networks that enables us to predict how a person would look like under various illumination conditions, e.g., lit by a point light source or an environment map.

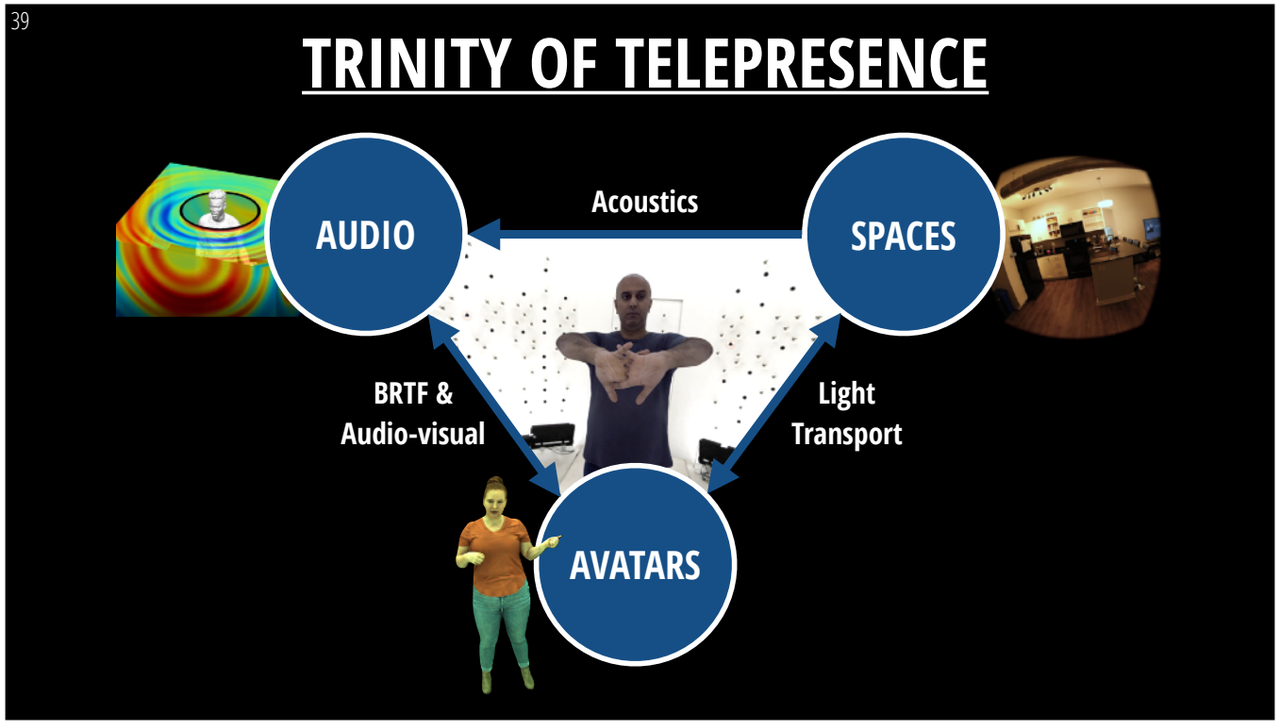
Combination of two approaches:
Deep relightable appearance models for animatable faces
Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn McPhail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih.
ACM Trans. Graph. 40, 4, Article 89 (August 2021), 15 pages. https://doi.org/10.1145/3450626.3459829
+
Mixture of volumetric primitives for efficient neural rendering
Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih.
ACM Trans. Graph. 40, 4, Article 59 (August 2021), 13 pages. https://doi.org/10.1145/3450626.3459863

Images/videos are Meta internal and not from third party works.



The final connection of the trinity of telepresence is between audio and spaces.
Audio and spaces are connected via the sounds field in the room, i.e., acoustics.
For example, the voice of a person sounds different in a fully furnished apartment and in a concert hall.

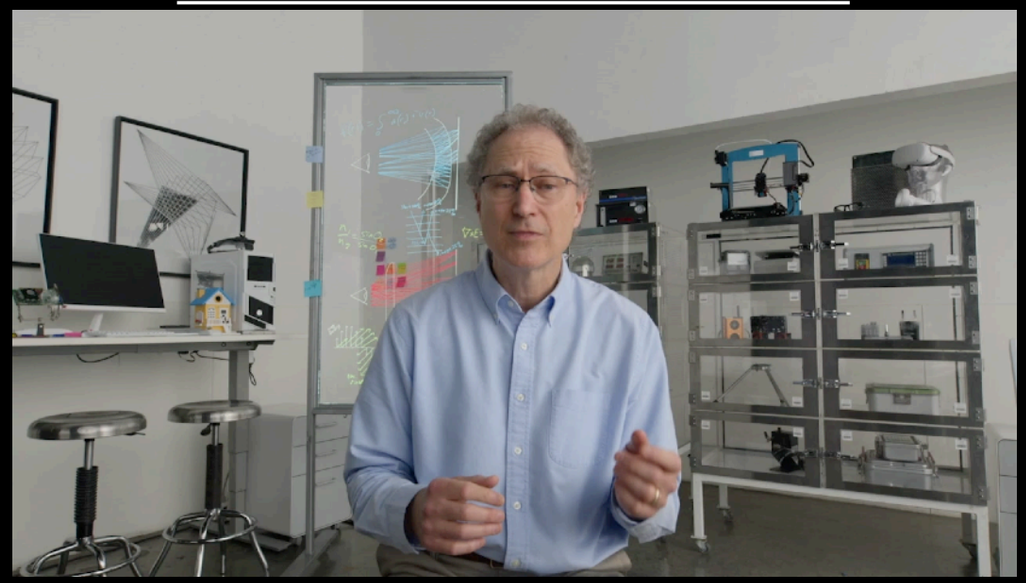
Images/videos are Meta internal and not from third party works.



All of these components together, including their interactions, are required to achieve “Complete Codec Telepresence”.

Images/videos are Meta internal and not from third party works.

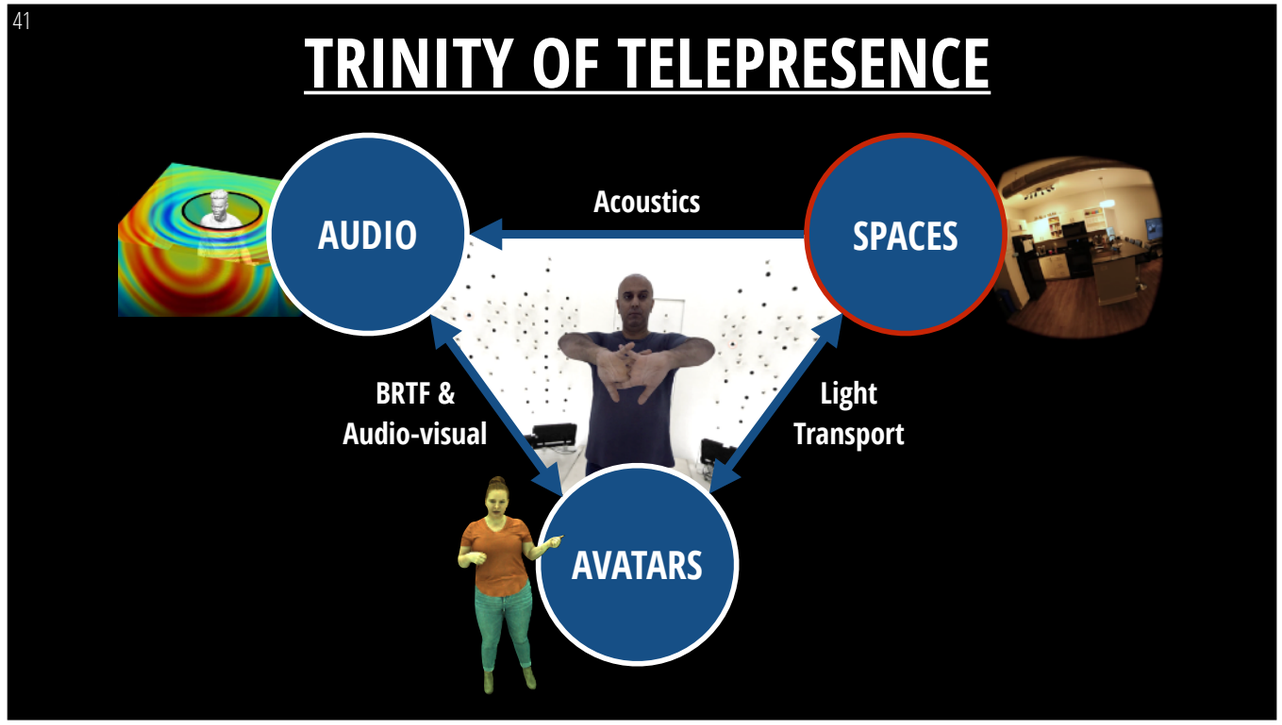
TRINITY PLAYBACK DEMO



[T. Bagautdinov, A. Pahuja, C. Wu, A. Richardt, Y.-S Shih , H. Mehta, E. Fath, M. Zollhoefer, Y. Sheikh]

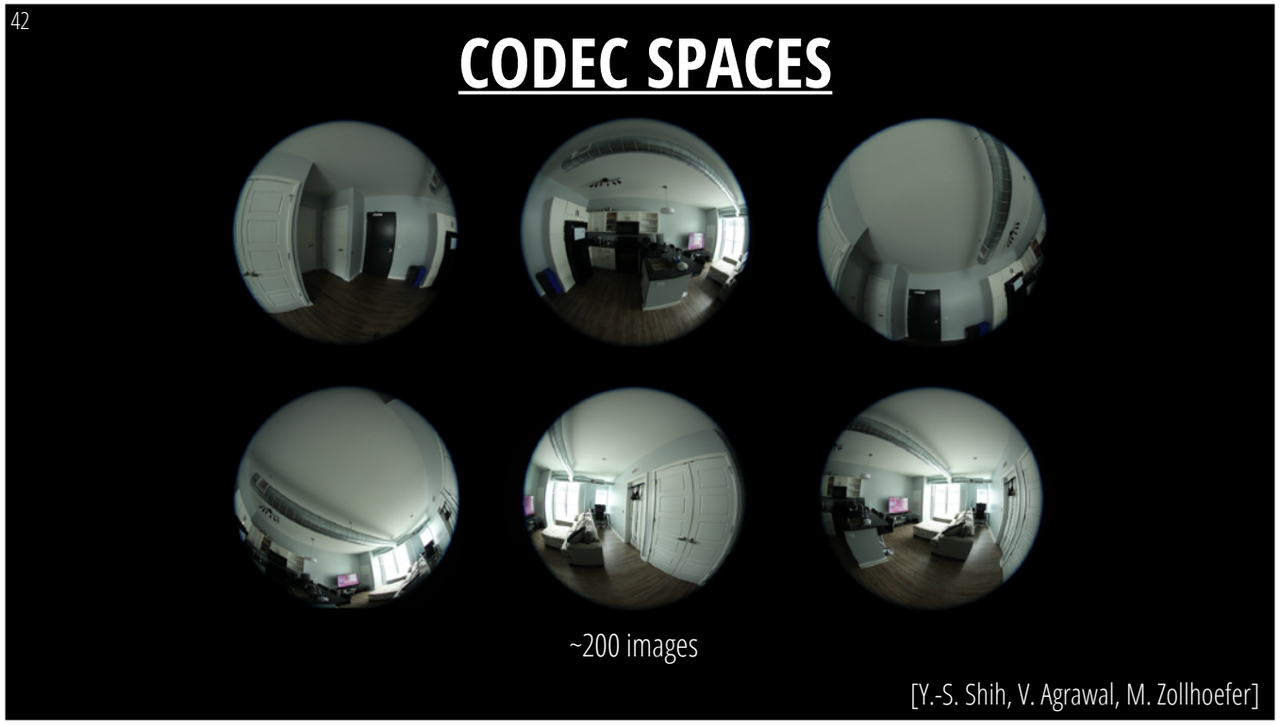
We have shown a demo that combines all of these components to produce the following video that was featured at FB Connect.

Images/videos are Meta internal and not from third party works.



In the following, we will dive deeper into how we create the photorealistic renderings of the space for this demo.
Jason will talk more about the avatars.

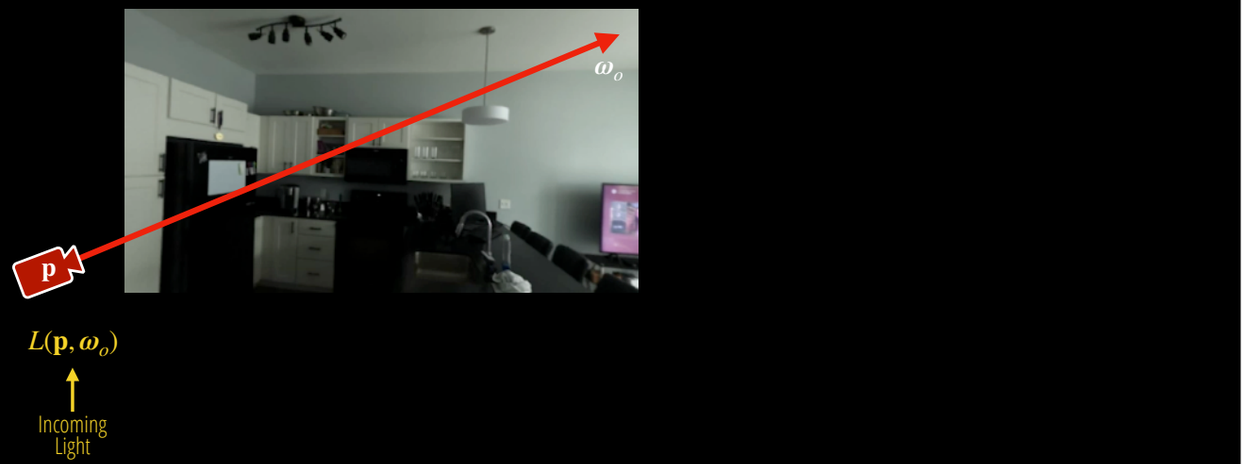
Images/videos are Meta internal and not from third party works.



We learn our photorealistic environments from a few hundred fisheye images that have been captured by a DSLR camera.
We employ a 180-degree fisheye lens to observe large parts of the space with only a few images.

Images/videos are Meta internal and not from third party works.

VOLUMETRIC RENDERING



Our approach is based on volumetric rendering.
To compute the incoming light for a pixel, we shoot a ray (with starting position p and direction w_o) into the scene.

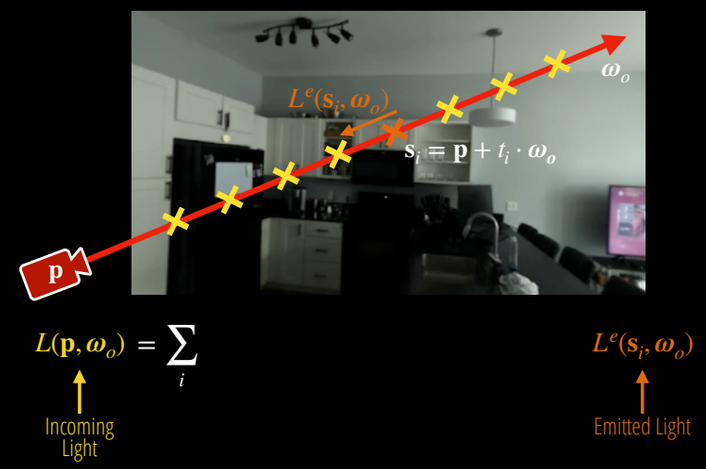
Images/videos are Meta internal and not from third party works.

VOLUMETRIC RENDERING

We take multiple discrete sample points s_i along the ray into account.

Images/videos are Meta internal and not from third party works.

VOLUMETRIC RENDERING



$$L(\mathbf{p}, \omega_o) = \sum_i$$

Incoming Light

$L^e(\mathbf{s}_i, \omega_o)$

Emitted Light

For each of the sample points s_i we have to determine the light that is emitted towards the camera.

Images/videos are Meta internal and not from third party works.

VOLUMETRIC RENDERING

$$L(\mathbf{p}, \omega_o) = \sum_i \Big(1 - \exp\big(-\delta_i \cdot \sigma(\mathbf{s}_i)\big)\Big) \cdot L^e(\mathbf{s}_i, \omega_o)$$

Incoming Light

Step Size

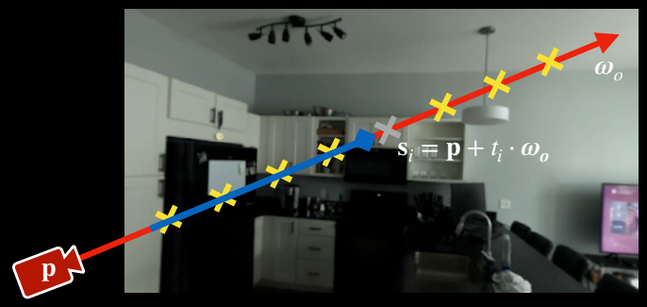
Volume Density

Emitted Light

This emitted light has to be multiplied by the volume density that is associated with the ray segment
This quantity is dependent on the step size delta_i of the ray marcher and the infinitesimal volume density sigma_i.

Images/videos are Meta internal and not from third party works.

VOLUMETRIC RENDERING



$$L(\mathbf{p}, \omega_o) = \sum_i T(\mathbf{s}_i, \omega_o) \cdot \left(1 - \exp\left(-\delta_i \cdot \sigma(\mathbf{s}_i)\right)\right) \cdot L^e(\mathbf{s}_i, \omega_o)$$

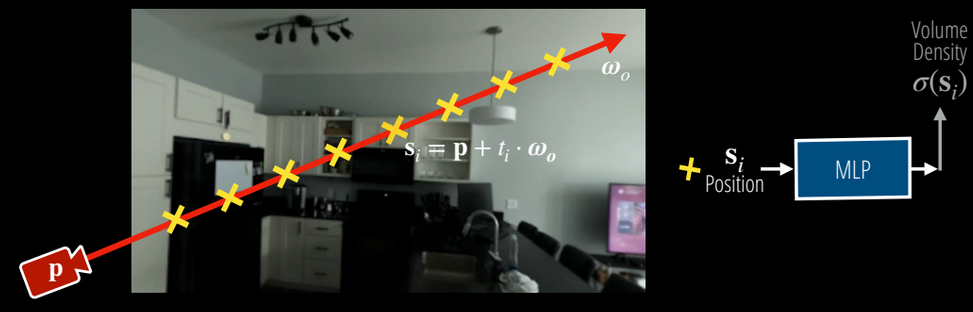
Incoming Light — Transmittance — Step Size — Volume Density — Emitted Light

$$T(\mathbf{s}_i, \omega_o) = \exp\left(-\sum_{j=0}^{i-1} \delta_j \cdot \sigma(\mathbf{s}_j)\right) \quad \delta_i = t_i - t_{i-1}$$

Finally, we have to also take into account how much of the light is absorbed by earlier sample points on the way to the camera. To this end, we multiply by the transmittance.
If we know the emitted light as well as the volume density per sample point, we can compute all other quantities from that and can obtain the incoming light via a sum.
One big question is how to parameterize or store the per-sample-point quantities.

Images/videos are Meta internal and not from third party works.

NEURAL RADIANCE FIELDS



$$L(\mathbf{p}, \omega_o) = \sum_i T(\mathbf{s}_i, \omega_o) \cdot \Big(1 - \exp\big(-\delta_i \cdot \sigma(\mathbf{s}_i)\big)\Big) \cdot L^e(\mathbf{s}_i, \omega_o)$$

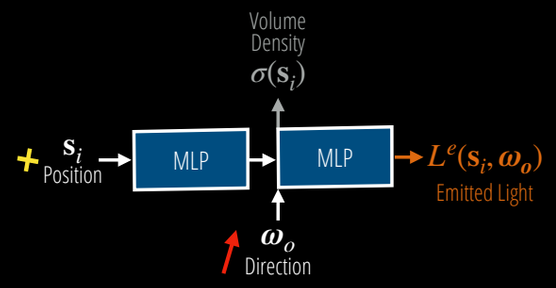
Incoming Light · Transmittance · Step Size · Volume Density · Emitted Light

$$T(\mathbf{s}_i, \omega_o) = \exp\Big(-\sum_{j=0}^{i-1} \delta_j \cdot \sigma(\mathbf{s}_j)\Big) \qquad \delta_i = t_i - t_{i-1}$$

One of the insights of recent coordinate-based neural scene representations such as NeRF is that these quantities can be well parameterized via a multi-layer perceptron. Volume density is only depend on the 3D position in space and is thus independent of the viewing direction.

Images/videos are Meta internal and not from third party works.

NEURAL RADIANCE FIELDS



$$L(\mathbf{p}, \omega_o) = \sum_i T(\mathbf{s}_i, \omega_o) \cdot \Big(1 - \exp\big(-\delta_i \cdot \sigma(\mathbf{s}_i)\big)\Big) \cdot L^e(\mathbf{s}_i, \omega_o)$$

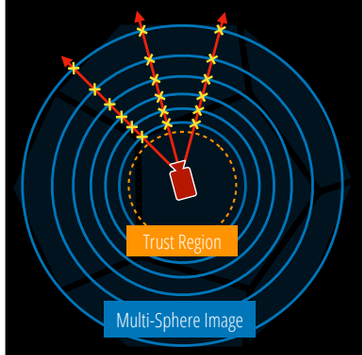
Incoming Light — Transmittance — Step Size — Volume Density — Emitted Light

$$T(\mathbf{s}_i, \omega_o) = \exp\Big(-\sum_{j=0}^{i-1} \delta_j \cdot \sigma(\mathbf{s}_j)\Big) \quad \delta_i = t_i - t_{i-1}$$

In contrast, the emitted light is view dependent, thus we additionally feed the ray direction to the MLP.
This helps in modeling view-dependent materials in the scene.
MLP-based models, while producing high fidelity results, are slow to evaluate.
Rendering a single image can take around 30s.
Thus, these approaches are not applicable to real-time VR telepresence applications.

Images/videos are Meta internal and not from third party works.

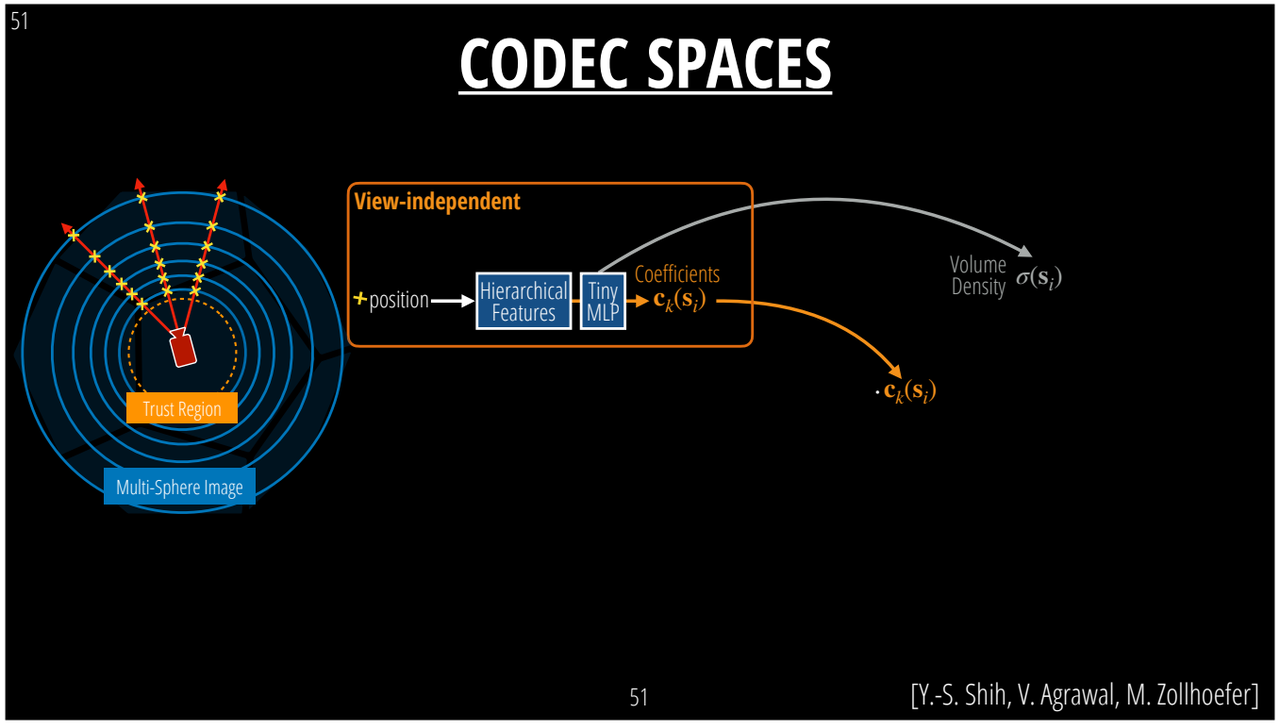
CODEC SPACES



[Y.-S. Shih, V. Agrawal, M. Zollhoefer]

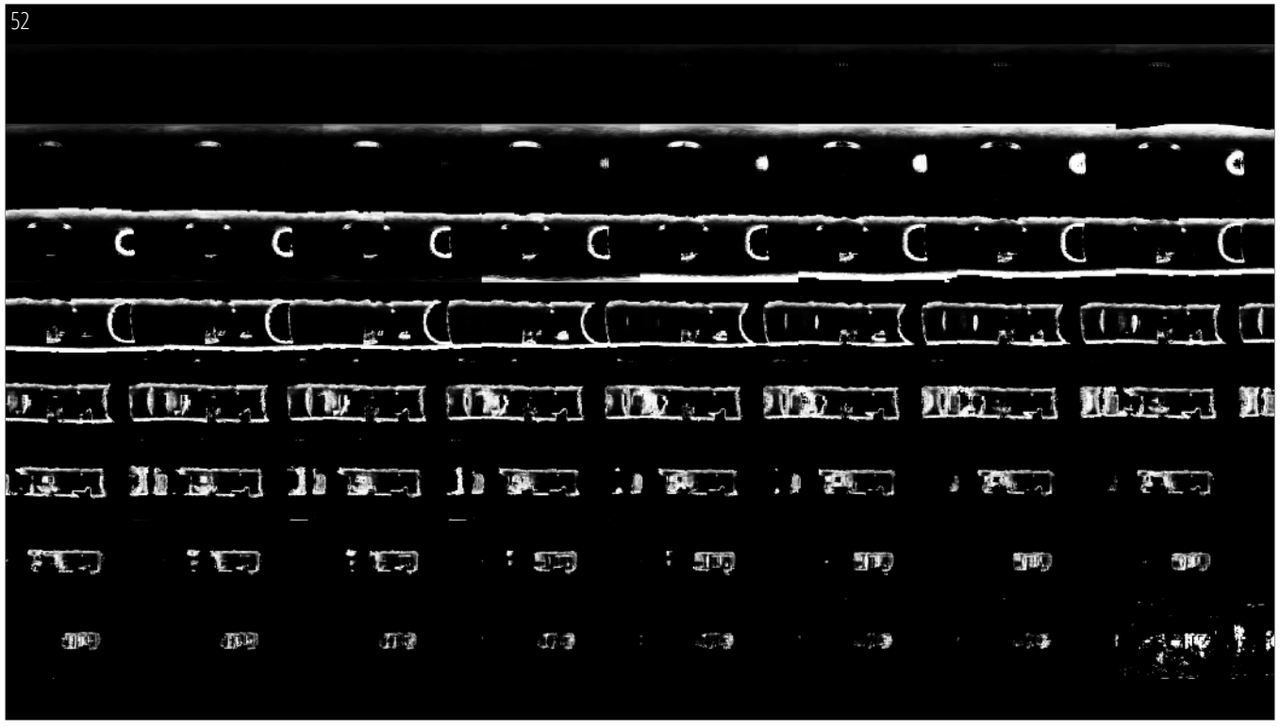
In contrast, our model is based on a set of proxy spheres and we restrict the sample points of the volumetric renderer to this proxy.

Images/videos are Meta internal and not from third party works.



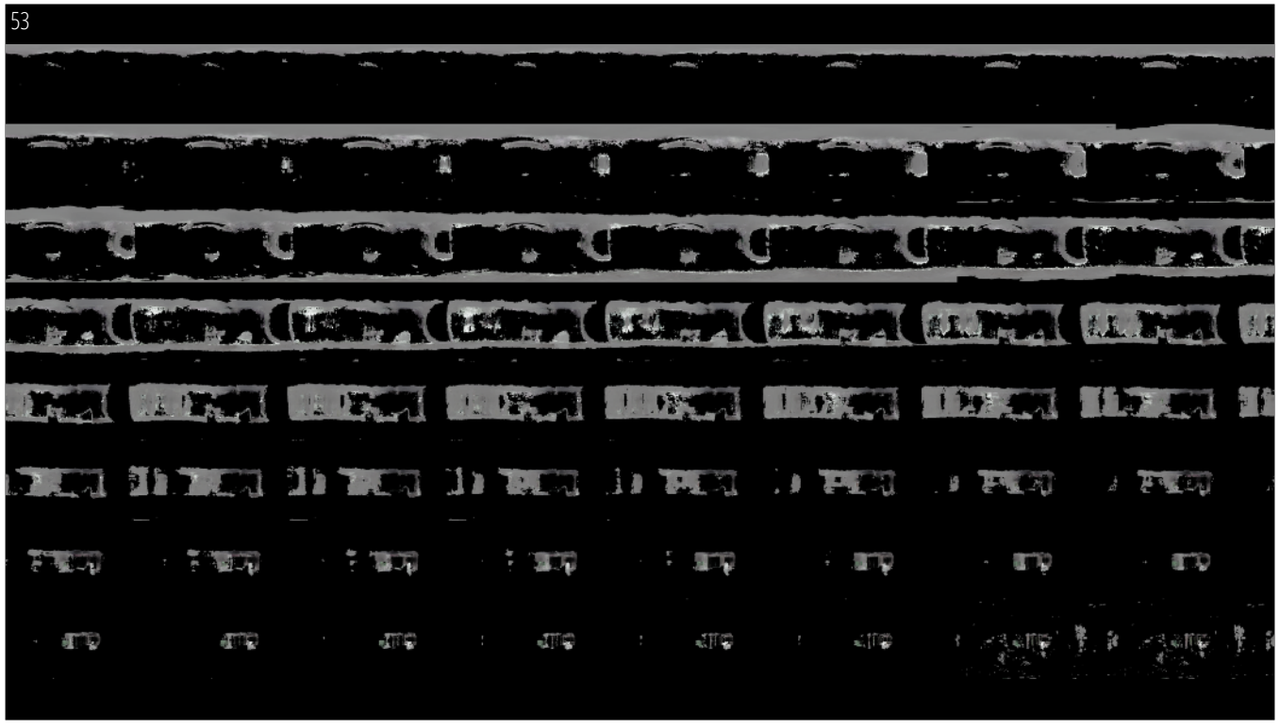
In addition, our model is split into view-dependent and view-independent networks.
This enables us to later “cache” the results, which enables real-time rendering of Codec Spaces in VR.
The view-independent network takes the sample position as input, accesses a hierarchical feature embedding, and runs a tiny MLP-based neural network to retrieve the volume density and a set of coefficients.
Since the input to the network is 3D, after training, we can cache the output of the MLP into a layered texture map.
Thus, at test time, we only have to sample textures for each of the sample points along the camera rays instead of having to evaluate neural networks.

Images/videos are Meta internal and not from third party works.



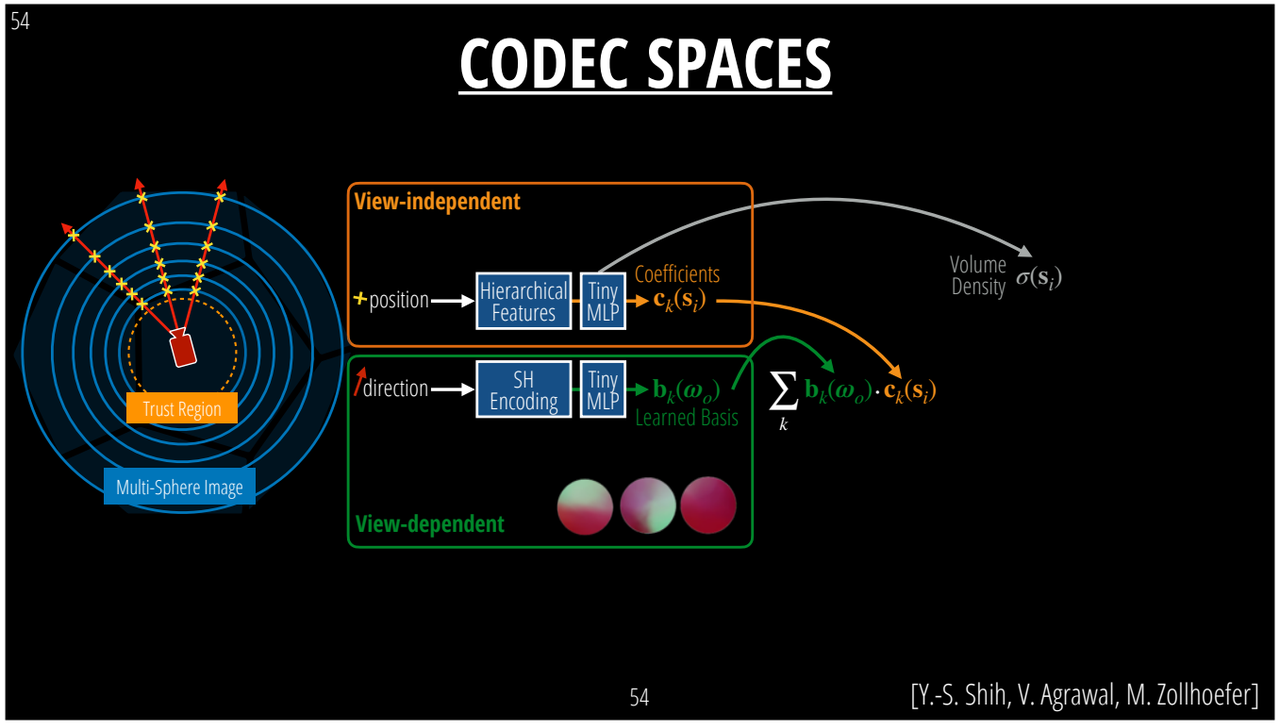
The textures look like this for the opacity and …

Images/videos are Meta internal and not from third party works.



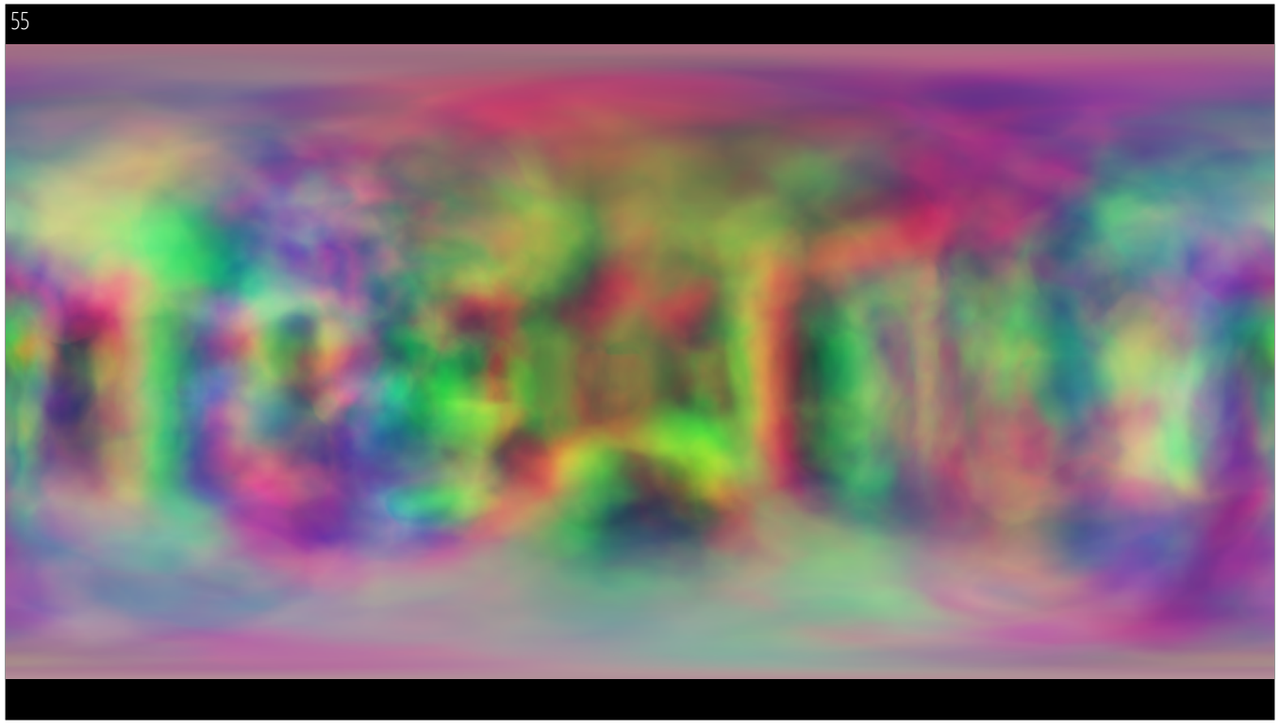
… like this for the coefficients.

Images/videos are Meta internal and not from third party works.



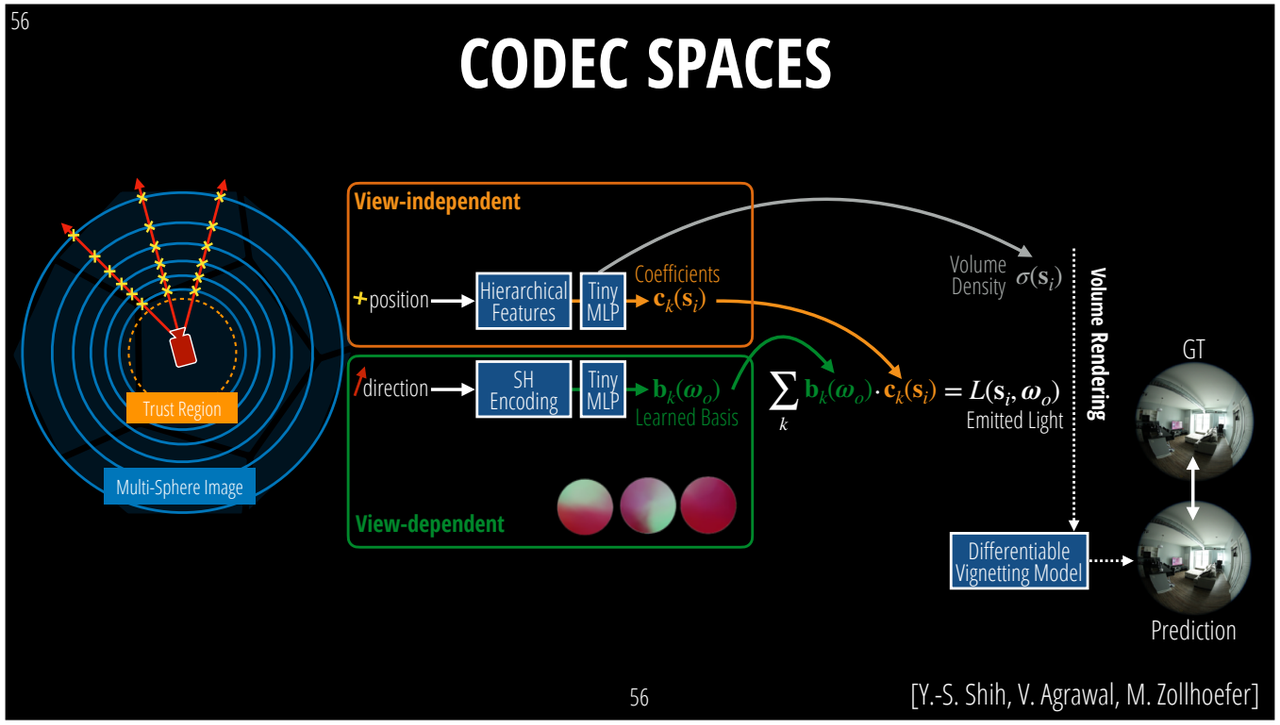
In addition to the view-independent part, our network also has a view-dependent part that consumes the viewing direction and outputs a set of evaluated basis functions. Since the basis functions are only view dependent (two degrees of freedom), they can be cached into a single texture map after training, which enables us to access them via simple texture reads during test time.

Images/videos are Meta internal and not from third party works.



For the basis functions, the textures are in lat-long coordinates and look like this.

Images/videos are Meta internal and not from third party works.



By combining the coefficients and the basis functions via a dot product, we obtain the view dependent color.
The color is aggregated based on the volume density to obtain the final image.
We also employ a differentiable vignetting module to compensate for lens vignetting of the used fisheye lens.
We can train our model in only a few hours by comparing the prediction results to GT images of resolution 4k vs 2k.

Images/videos are Meta internal and not from third party works.

CODEC SPACES

Live VR View

RX Side

[Y.-S. Shih, V. Agrawal, M. Zollhoefer]

Now, after I have explained the technical detail, let’s have again a look at the video that shows real-time VR rendering of the learned scene representation.

Images/videos are Meta internal and not from third party works.

CODEC SPACES - 360-DEGREE RENDERING

[Y.-S. Shih, V. Agrawal, M. Zollhoefer]

This video demonstrates the full 360-degree (panoramic) novel view synthesis capabilities of our approach.

Images/videos are Meta internal and not from third party works.

CODEC SPACES - SPARSITY

[Y.-S. Shih, V. Agrawal, M. Zollhoefer]

We train our model based on a novel occlusion-aware depth variance loss.
This enables us to obtain a much sparser scene representation than a standard radiance field would learn.
The video shows a swipe through the learned volume.
In the beginning, the result of aggregating all layers is shown.
Afterwards, the layers are disabled one-by-one in front-to-back order.
As can be seen, the model learns objects at their correct depth and the learned representation is sparse.
The sparsity enables us to compress the model by only storing occupied space using a texture atlas.

Images/videos are Meta internal and not from third party works.

CODEC SPACES - COMPARISON

Here, we show a comparison to a 3D reconstruction by Agisoft Metashape, a state-of-the-art multi-view stereo reconstruction approach.
Codec Spaces obtain highly realistic results with view-dependent effects, while the 3D reconstructions suffer from reconstruction failures and incomplete geometry.

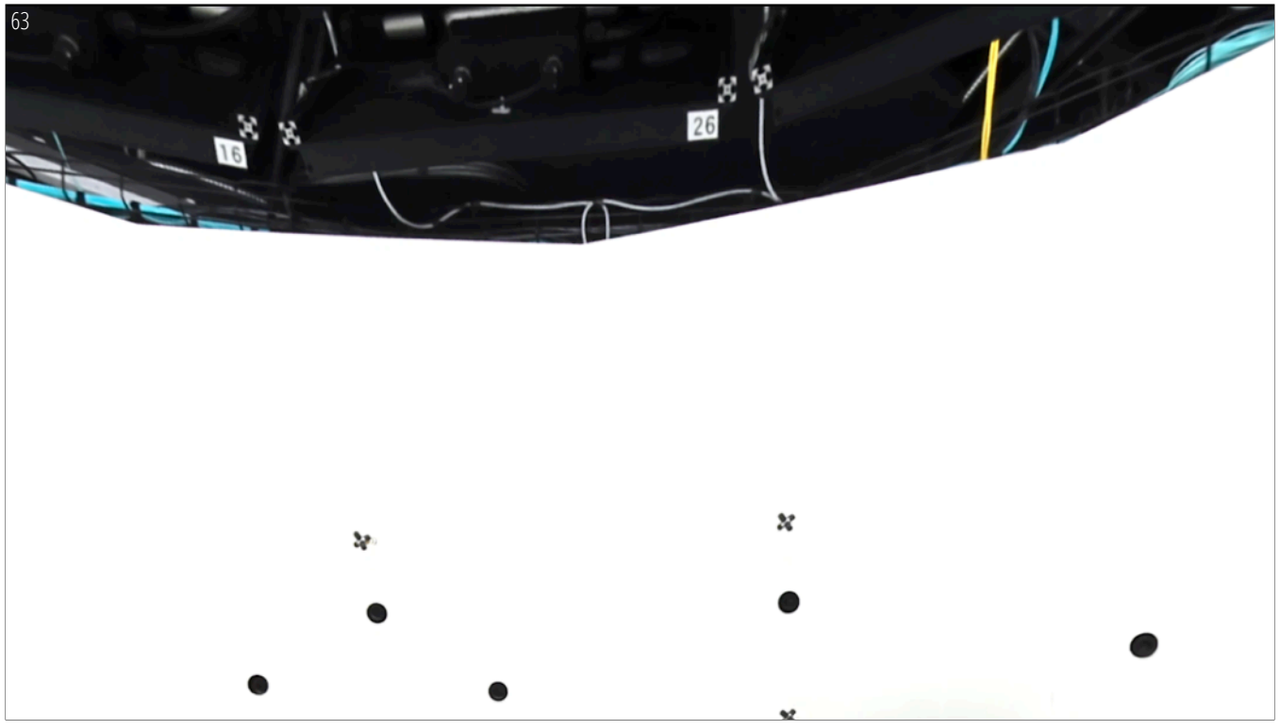
Images/videos are Meta internal and not from third party works.

High-res view synthesis results for an apartment scene.
The specular reflections on the TV are well modeled and the user can look into any direction.

Images/videos are Meta internal and not from third party works.

Our model also enables us to create virtual representations of the capture setups in our lab.
Here, you can see me walking around in a virtual version of Sociopticon, our full-body capture studio.
We used a similar way to synthesize the background for the Trinity playback demo that I showed earlier.

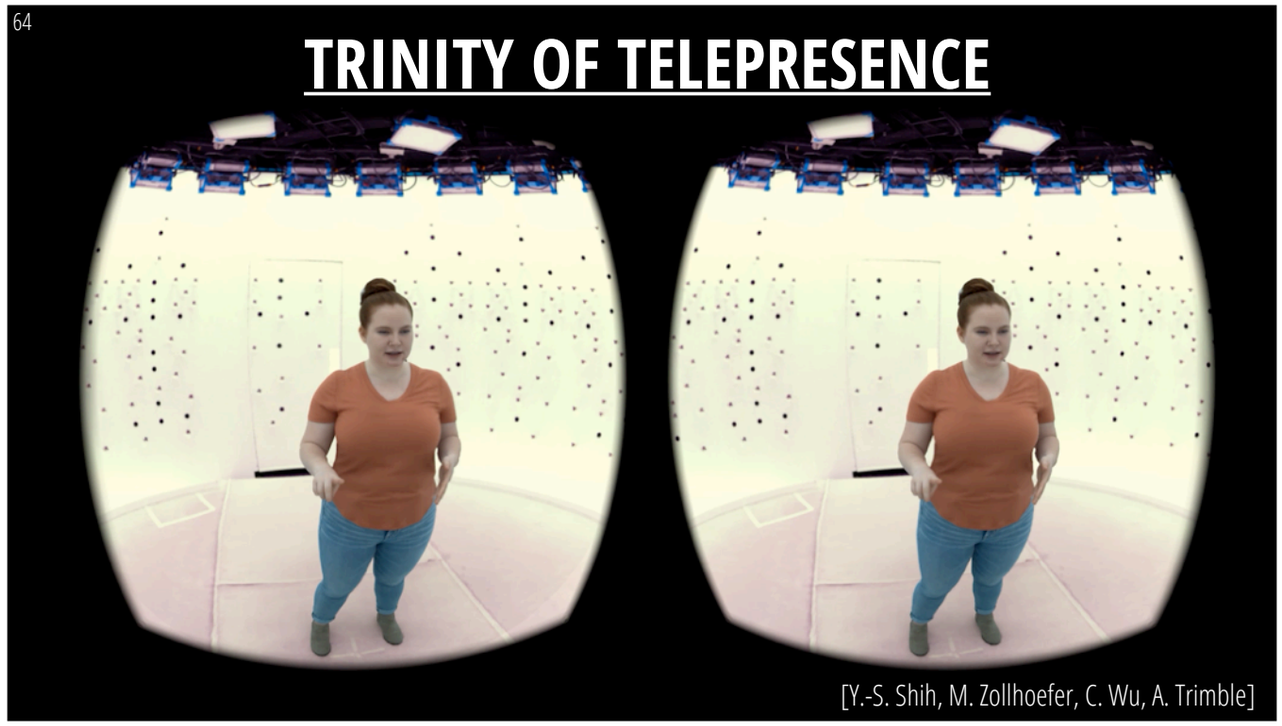
Images/videos are Meta internal and not from third party works.



Here is a flip test of our rendering with respect to the ground truth images.
Our synthesized results are almost indistinguishable from the ground truth.

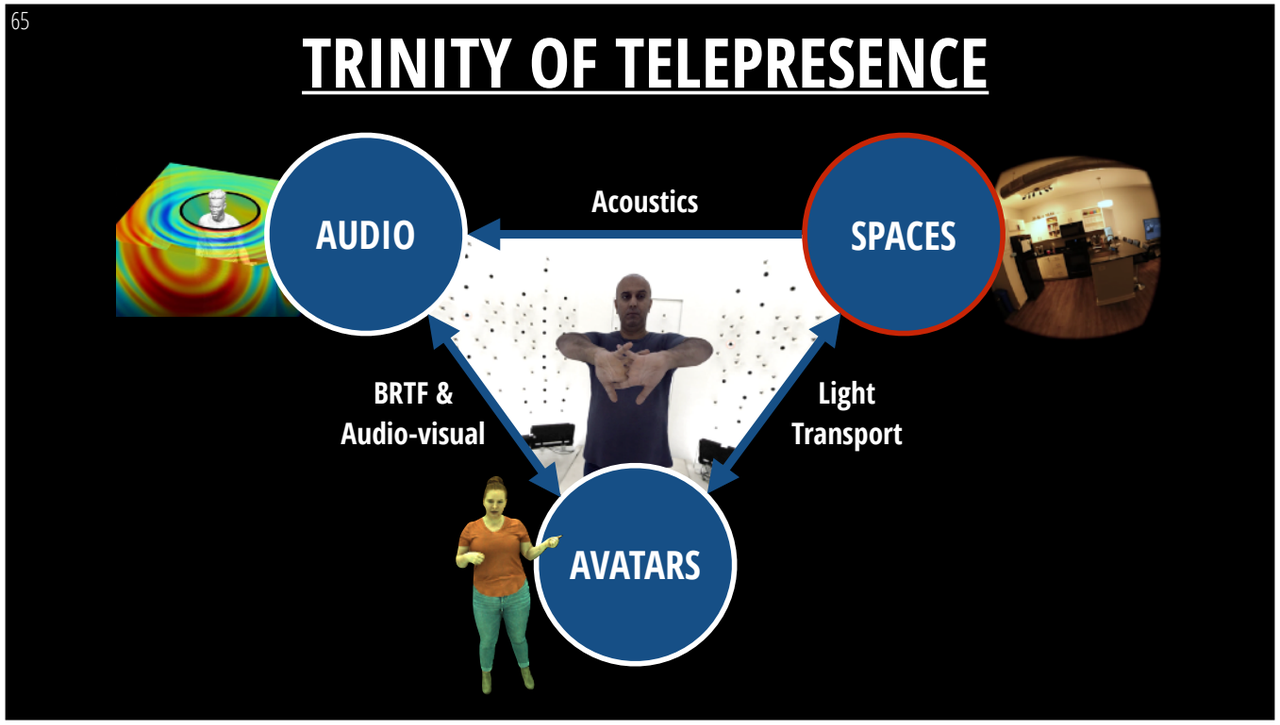
Images/videos are Meta internal and not from third party works.

TRINITY OF TELEPRESENCE



Here is another results that combines avatars, audio, and spaces.

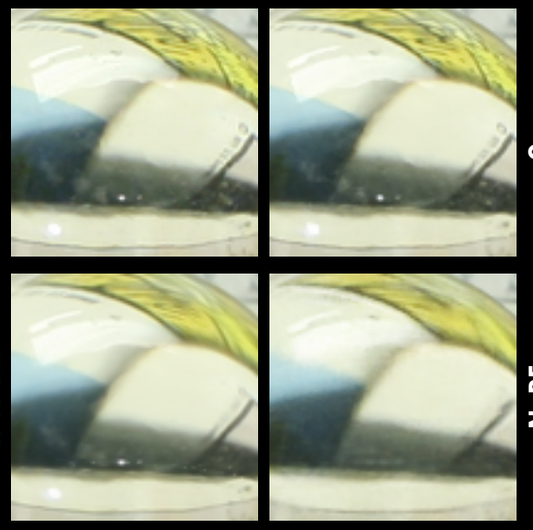
Images/videos are Meta internal and not from third party works.



We have talked about the Trinity of Telepresence and seen how research at the intersection of avatars, audio, and spaces is required to achieve complete codec telepresence.
There are still many exciting research problems that have to be solved to fully realize the trinity of telepresence and deliver an experience that is indistinguishable from reality.
For spaces, we will require the ability to …

Images/videos are Meta internal and not from third party works.

CURVED REFLECTIONS



"Learning Neural Light Fields with Ray-Space Embedding Networks" - B. Attal, J.-B. Huang, M. Zollhoefer, J. Kopf, C. Kim

Model curved reflections at real time frame rates, which is not possible with current radiance field based approaches.
One promising direction that be have explored so far are neural light field networks that allow to synthesize a pixel's color with as few as 1 MLP evaluation, i.e., directly mapping a ray to its color.

Video results from:
Learning Neural Light Fields with Ray-Space Embedding Networks
Benjamin Attal and Jia-Bin Huang and Michael Zollhoefer and Johannes Kopf and Changil Kim
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
Dataset: http://lightfield.stanford.edu/acq.html
Fair Use.
This is a course/talk with the purpose of education/teaching.
Video result from related work for illustrative purpose.

DYNAMIC SCENES

“Neural 3D Video Synthesis from Multi-view Video”- T. Li, M. Slavcheva, M. Zollhoefer, S. Green, C. Lassner, C. Kim, T. Schmidt, S. Lovegrove, M. Goesele, R. Newcombe, Z. Lv

Ideally, in the future, we will also need a scene representation to model dynamic spaces in real-time.
We have explored a first version of this in a CVPR paper.

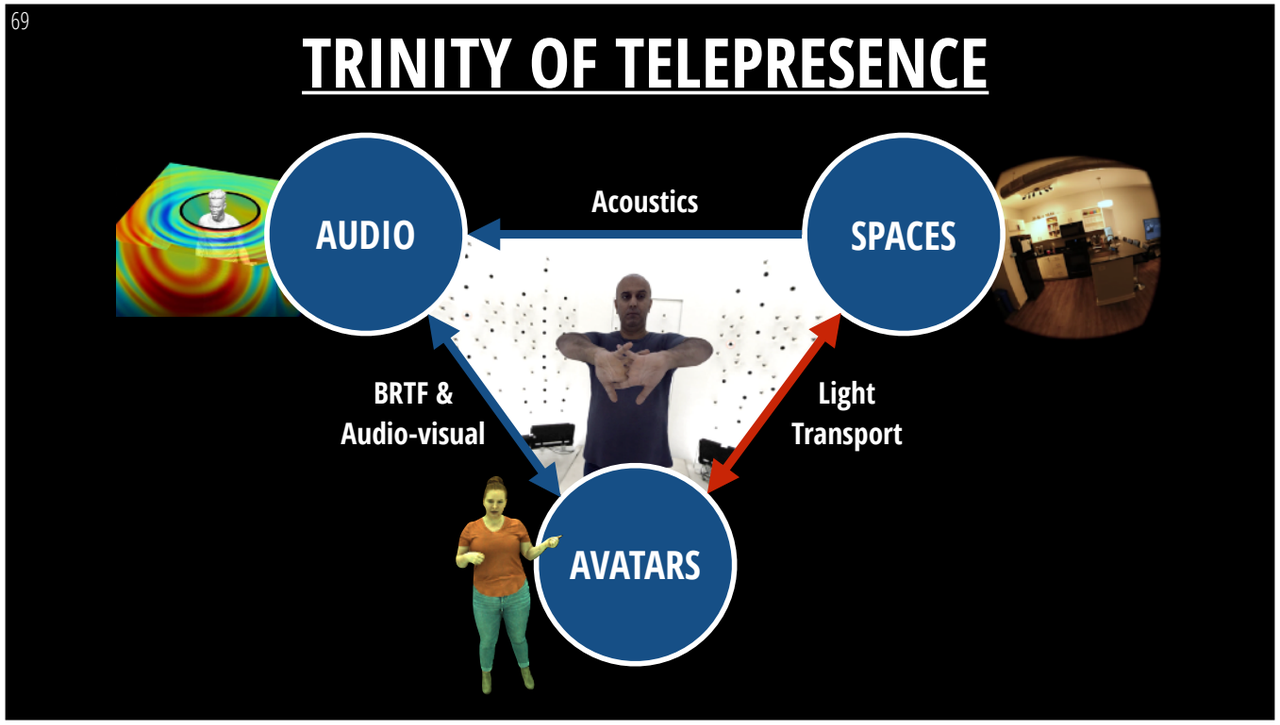
Video results from:
Neural 3D Video Synthesis from Multi-view Video
Li, Tianye and Slavcheva, Mira and Zollhoefer, Michael and Green, Simon and Lassner, Christoph and Kim, Changil and Schmidt, Tanner and Lovegrove, Steven and Goesele, Michael and Newcombe, Richard and Lv, Zhaoyang, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
Fair Use.
This is a course/talk with the purpose of education/teaching.
Video result from related work for illustrative purpose.
Dataset: https://github.com/facebookresearch/Neural_3D_Video
We/Meta are the owners of this dataset.

LIVE PLAYBACK

“Neural 3D Video Synthesis from Multi-view Video”- T. Li, M. Slavcheva, M. Zollhoefer, S. Green, C. Lassner, C. Kim, T. Schmidt, S. Lovegrove, M. Goesele, R. Newcombe, Z. Lv

These dynamic radiance fields can be played back in VR on a mobile HMC, such as Quest2.
Currently, only playback of the training data is possible, no generalization to new motions or interactions with objects.
Please note, the person in this scene is not an avatar, i.e., not drivable, but is represented jointly with the radiance field of the scene.

Video results from:
Neural 3D Video Synthesis from Multi-view Video
Li, Tianye and Slavcheva, Mira and Zollhoefer, Michael and Green, Simon and Lassner, Christoph and Kim, Changil and Schmidt, Tanner and Lovegrove, Steven and Goesele, Michael and Newcombe, Richard and Lv, Zhaoyang, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
Fair Use.
This is a course/talk with the purpose of education/teaching.
Video result from related work for illustrative purpose.
Dataset: https://github.com/facebookresearch/Neural_3D_Video
We/Meta are the owners of this dataset.



While we already have relightable heads and upper bodies, in the future, more research will be required into relightable full bodies to enable us to “teleport” people into arbitrary spaces.

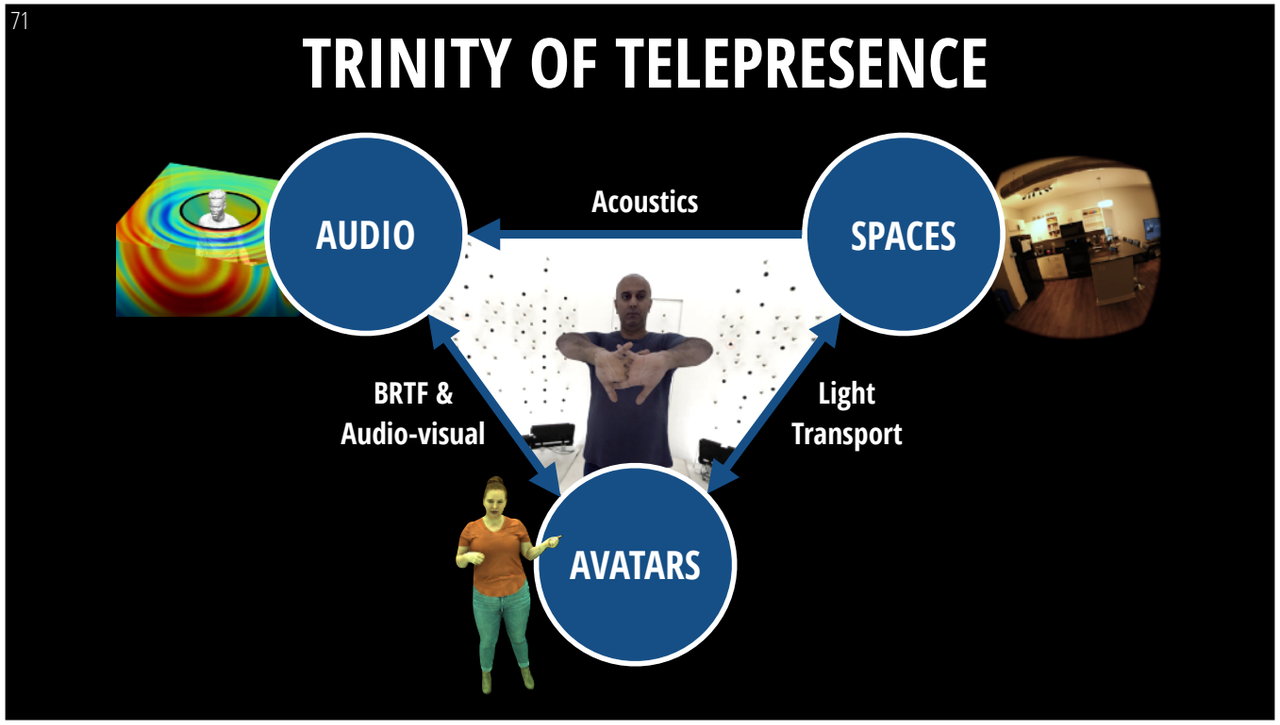
Images/videos are Meta internal and not from third party works.

THE NEED FOR RELIGHTABLE BODIES

[T. Bagautdinov, A. Pahuja, C. Wu, Y.-S Shih, M. Zollhoefer, Y. Sheikh]

As you can see here, currently, our full-body avatars do not perfectly blend into a new space, due to the lack of relightability.
This is especially noticeable here for the feet and legs of Yaser, since the light should be blocked by the kitchen isle and thus the feet and legs should be much darker than the upper body.

Images/videos are Meta internal and not from third party works.



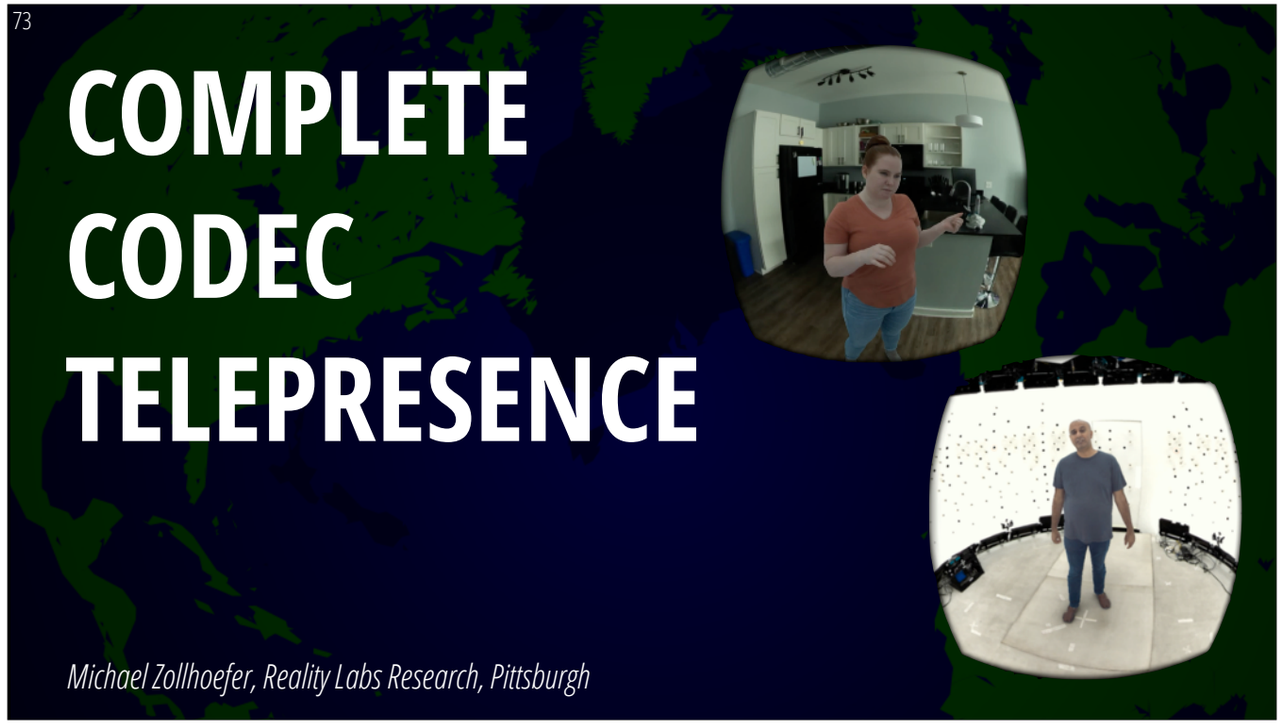
As you can see, there are still many exciting research problems ahead of us that have to be solved to deliver a complete telepresence experience that is indistinguishable from reality.

Images/videos are Meta internal and not from third party works.

Looking back at the history of communication technologies, each innovation, such as mail, the phone, the internet, or video calls, has made communication between people across the world easier and more efficient.
All these technologies have brought the world closer together and thus virtually “shrunk” the size of our planet.
VR telepresence is the next evolution of communication and has the promise to “shrink” our planet further.
In the future, this will bring the world even closer together by enabling anybody to communicate and interact with anyone, anywhere, at any time, as if everyone would be sharing the same physical space.

Globe image from:
https://commons.wikimedia.org/wiki/File:Globe.svg
Public domain from the Creative Commons Corporation
Author: Augiasstallputzer~commonswiki

Other images/videos are Meta internal and not from third party works.



https://commons.wikimedia.org/wiki/File:Globe.svg
Public domain from the Creative Commons Corporation
Author: Augiasstallputzer~commonswiki

Other images/videos are Meta internal and not from third party works.