Folien zur Vorlesung am 26.05.2025

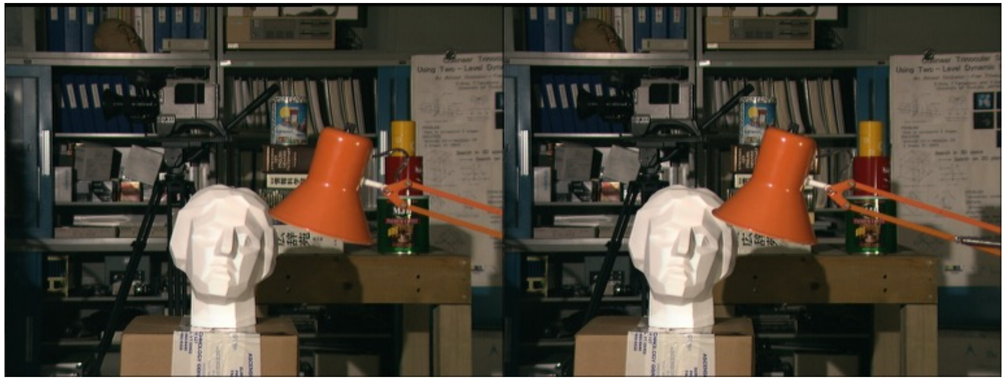
3D Computer Vision

MULTI-VIEW STEREO

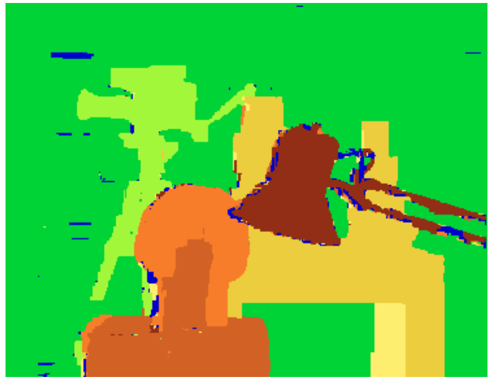
Recommended Reading

- Slides from Noah Snavely (Cornell)
- Szeliski (2nd Edition) Chapter 12.7
- Multi-View Stereo: A Tutorial, Furukawa and Hernandez, 2015
  - http://carlos-hernandez.org/papers/fnt_mvs_2015.pdf

Last time: Binocular (Two-View) Stereo



Left-right (rectified) stereo pair



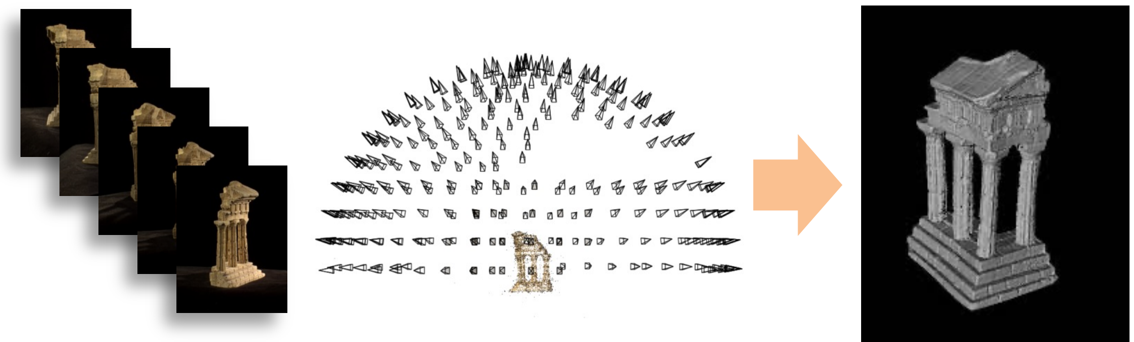
Computed disparity map

Useful for robot perception and navigation, video effects, etc.

Multi-view Stereo

Problem formulation: given several images of the same object or scene, compute a representation of its 3D shape

Binocular Stereo

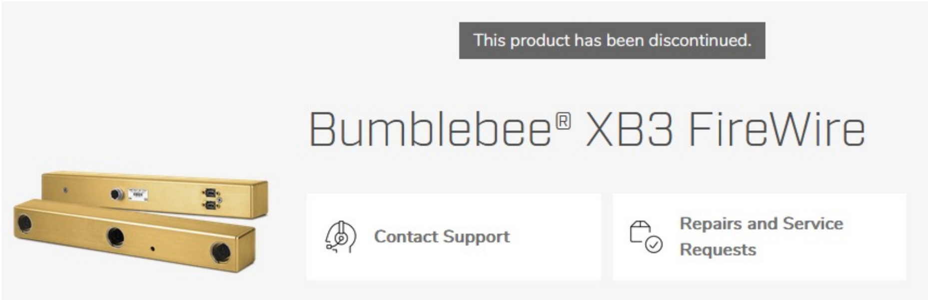


Multi-view stereo

Multi-view Stereo projects

- https://www.youtube.com/watch?v=Bse7YXWdP-c

Multi-view Stereo devices

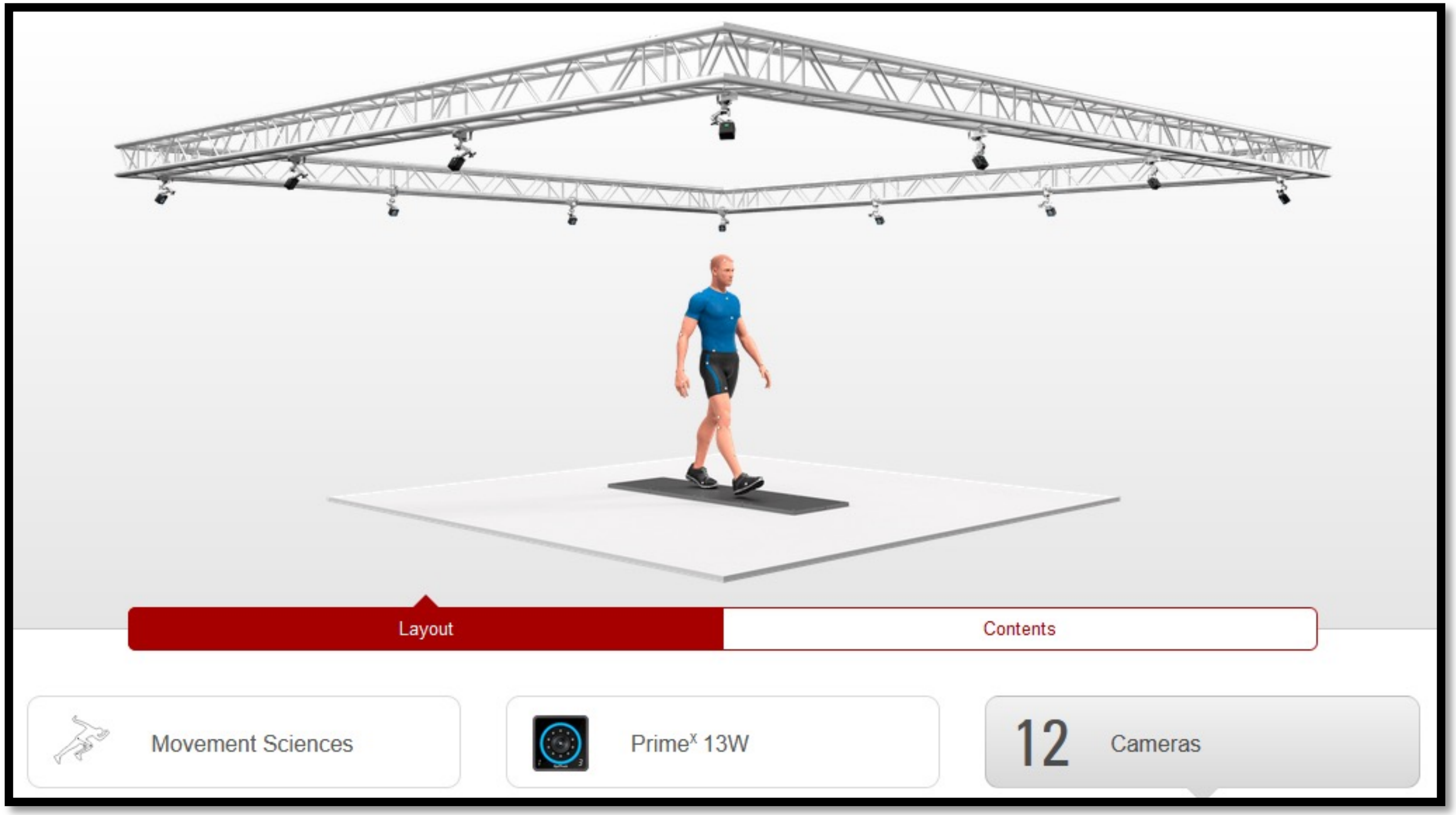


https://www.flir.de/support/products/bumblebee-xb3-firewire/#Overview

- Not available any more…

Point Grey's ProFusion 25

Multi-view Stereo devices



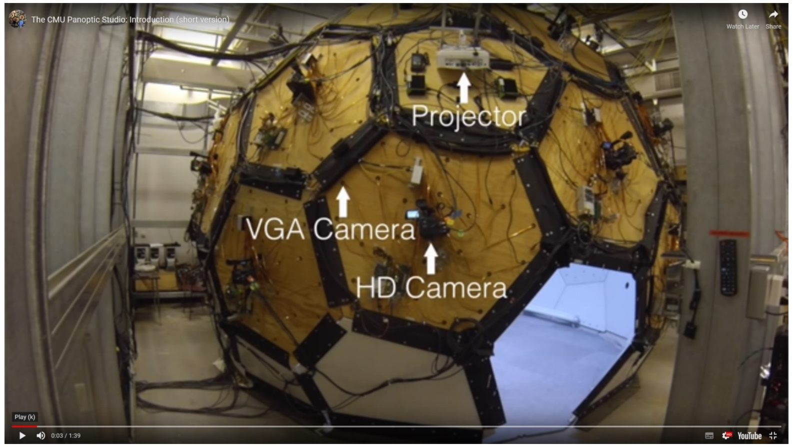
- https://optitrack.com/systems/#movement/primex-13w/12
- Such a system is available at HFU.

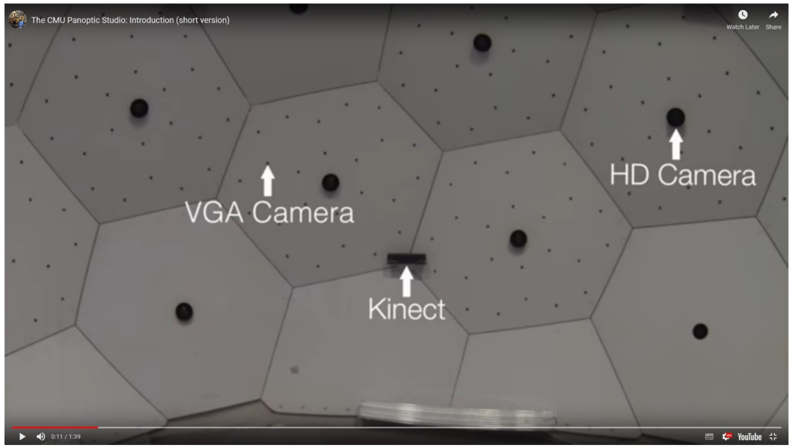
Multi-view Stereo as a service

https://renderpeople.com/about-us/

Multi-view Stereo world scale

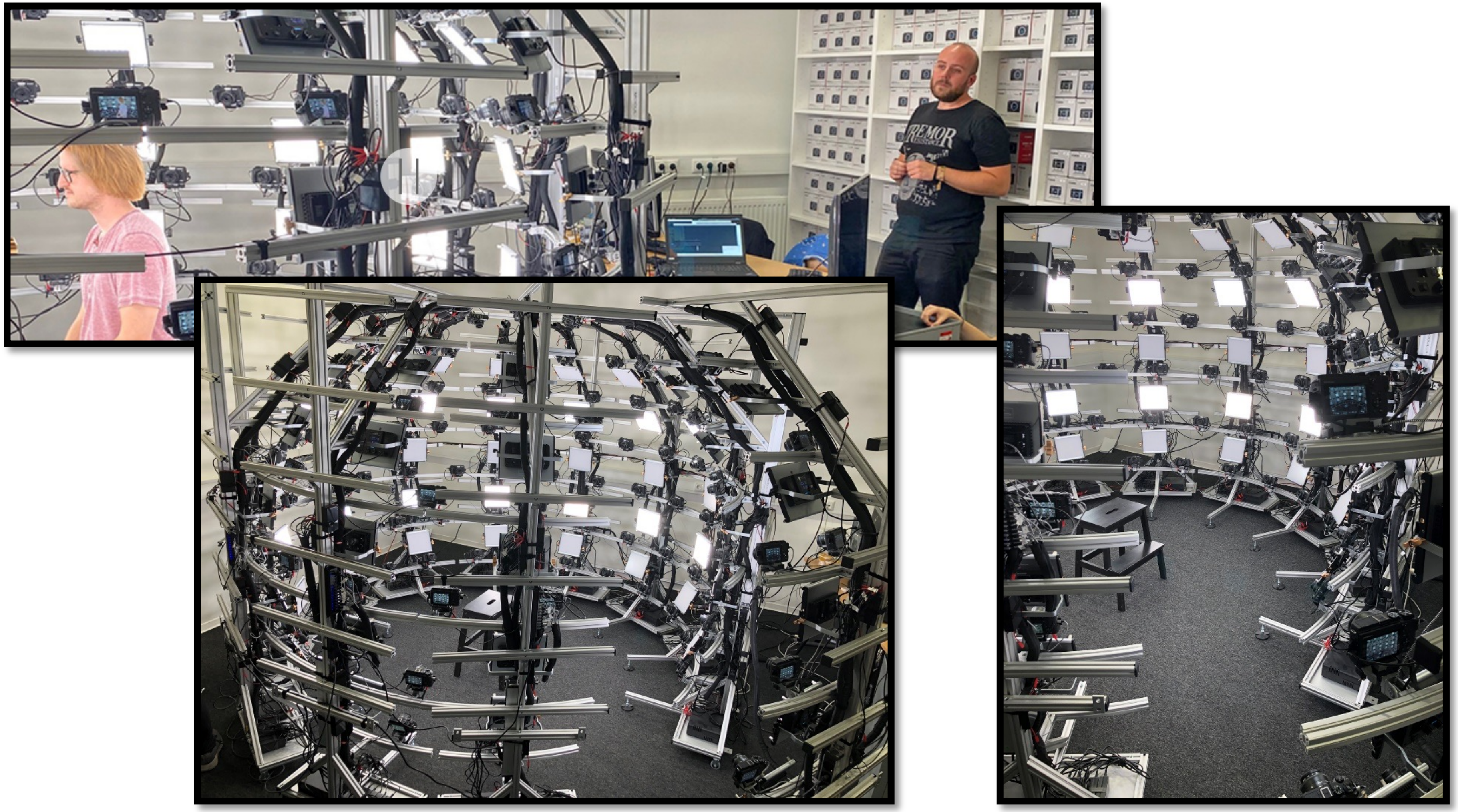
Multi-view Stereo research
(Carnegie Mellon)





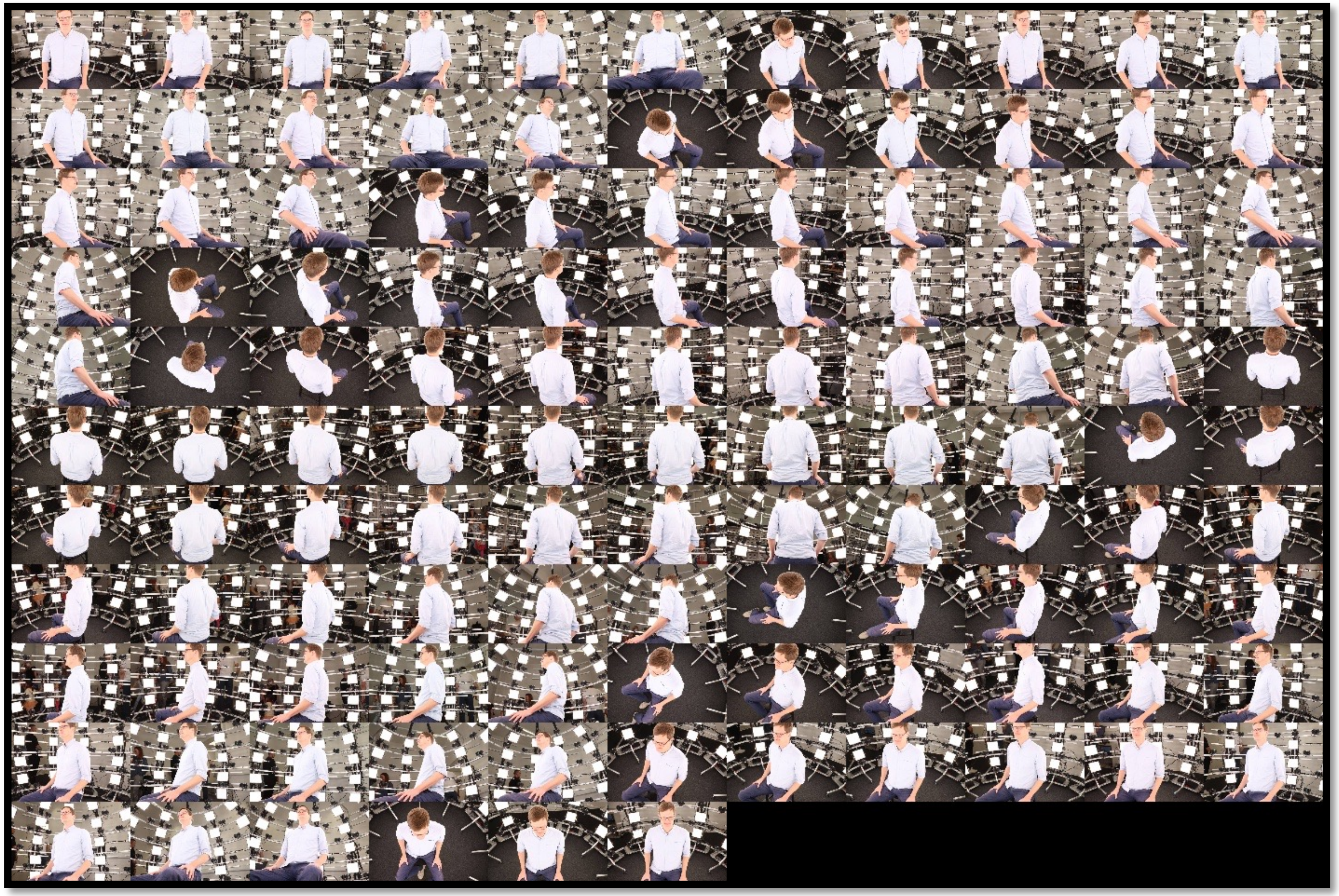
http://domedb.perception.cs.cmu.edu/

Multi-view Stereo research
(Bauhaus Uni Weimar)

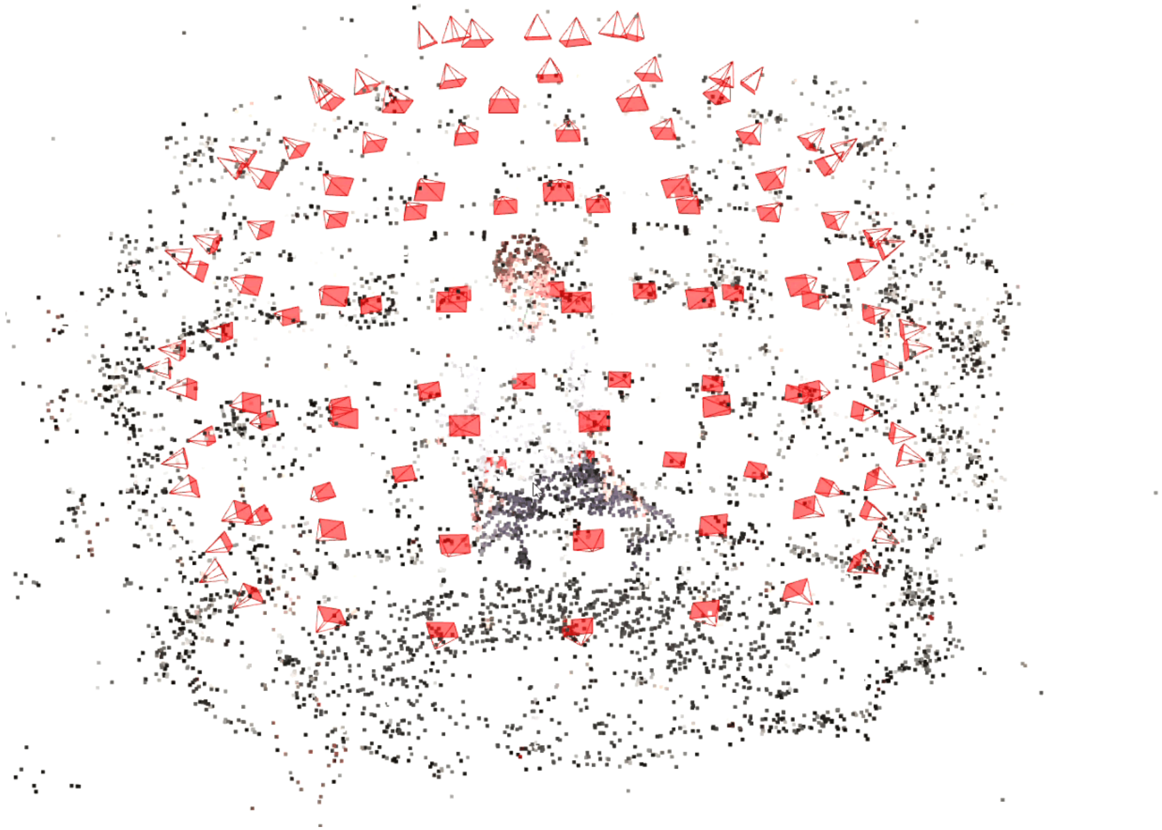


https://www.uni-weimar.de/de/medien/.../computer-vision/.../3d-realitycapture-scanlab/

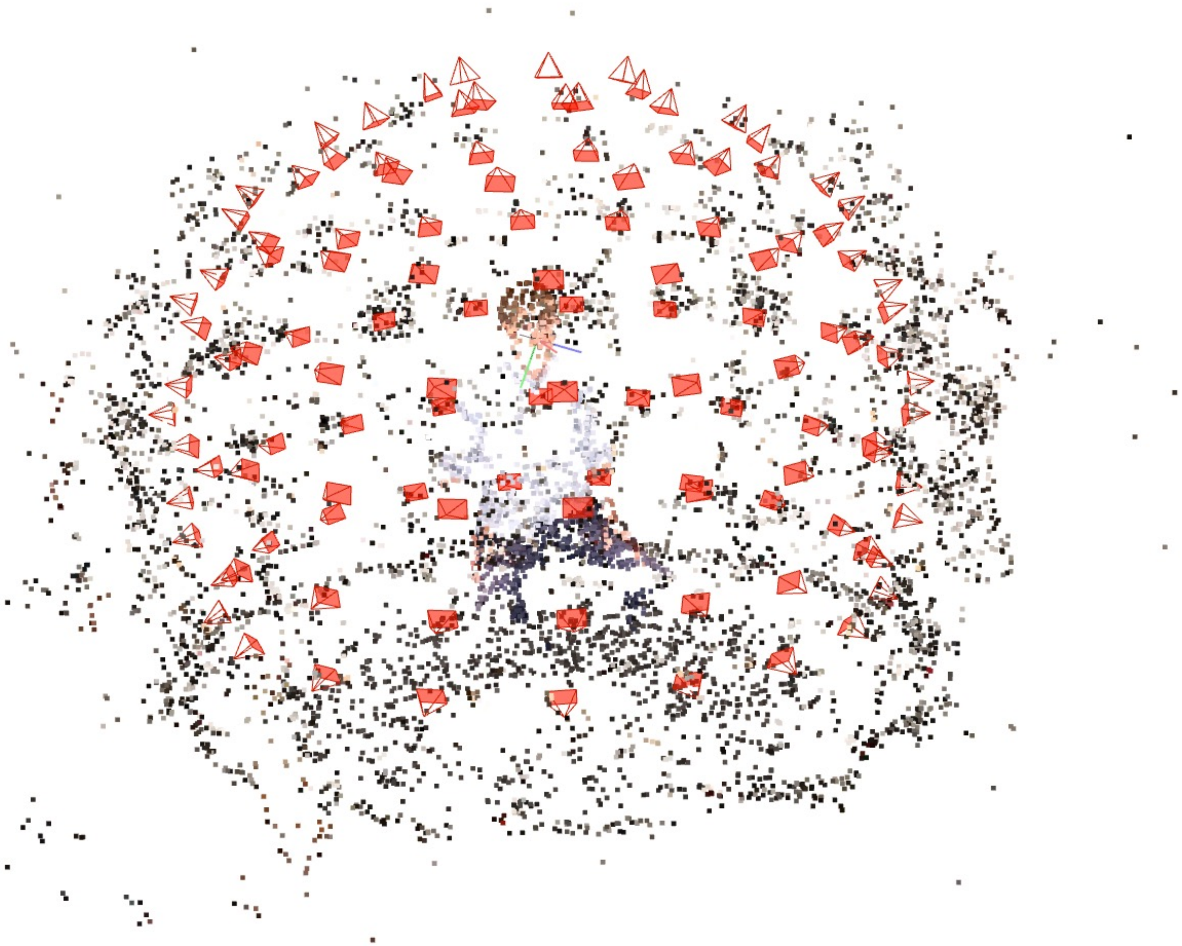
Multi-view Stereo research
(Bauhaus Uni Weimar)



Multi-view Stereo research
(Bauhaus Uni Weimar)



Multi-view Stereo research
(Bauhaus Uni Weimar)

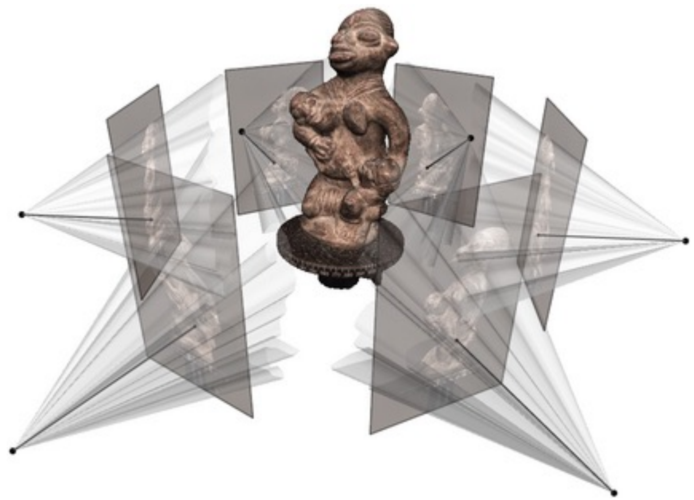


ENOUGH EXAMPLES → THEORY

Multi-view Stereo principle

Input: calibrated images from several viewpoints (known intrinsics and extrinsics / projection matrices)

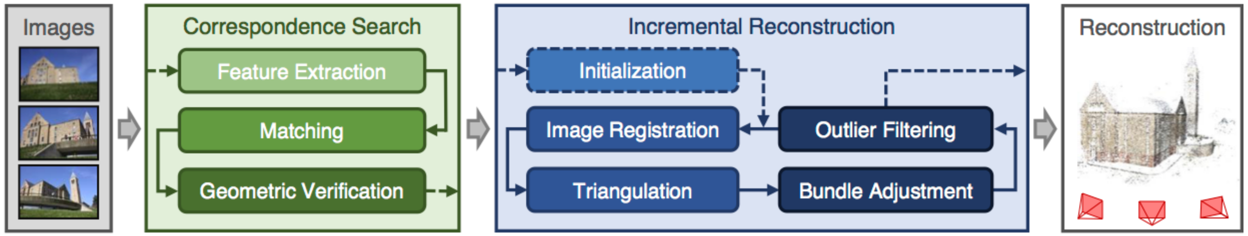
Output: 3D object model



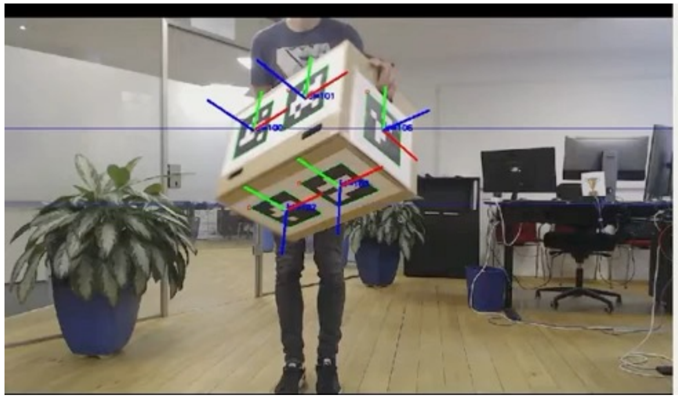
Figures by Carlos Hernandez

How to get the camera parameters?

- COLMAP → explanation later in Structure-From-Motion chapter



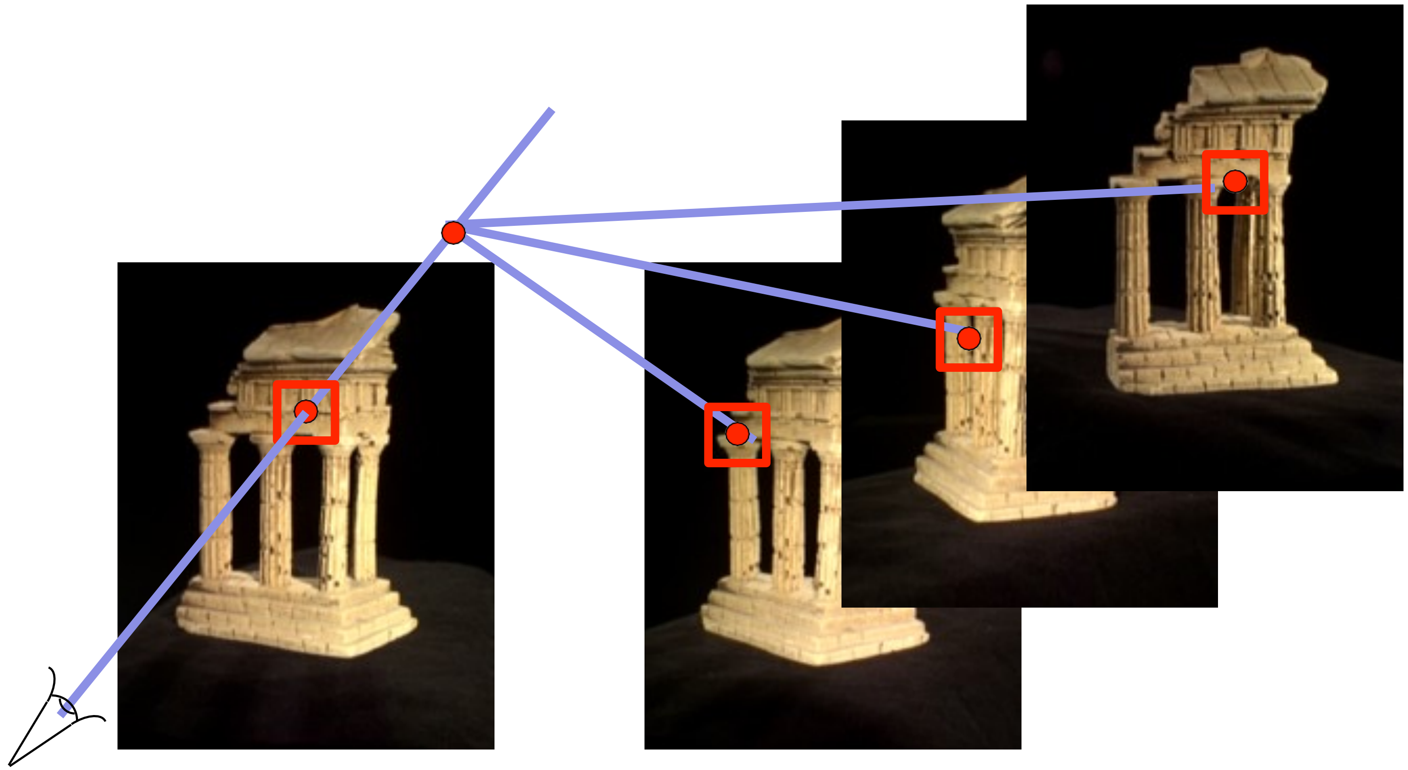
- Manual calibration with known targets (e.g. chessboards, ARuCo, …)



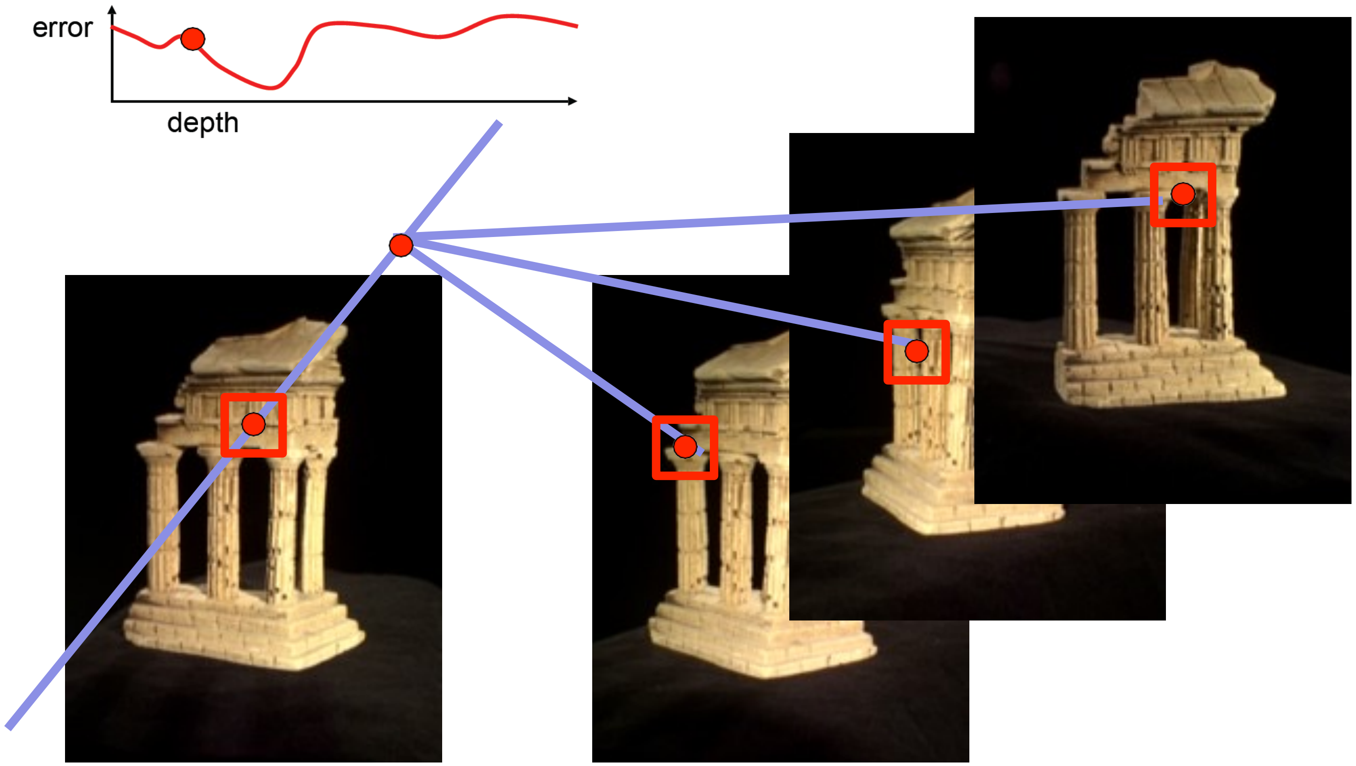
- This is still a hot research topic.

[video source: https://www.youtube.com/watch?v=HGDxFJALNsY]

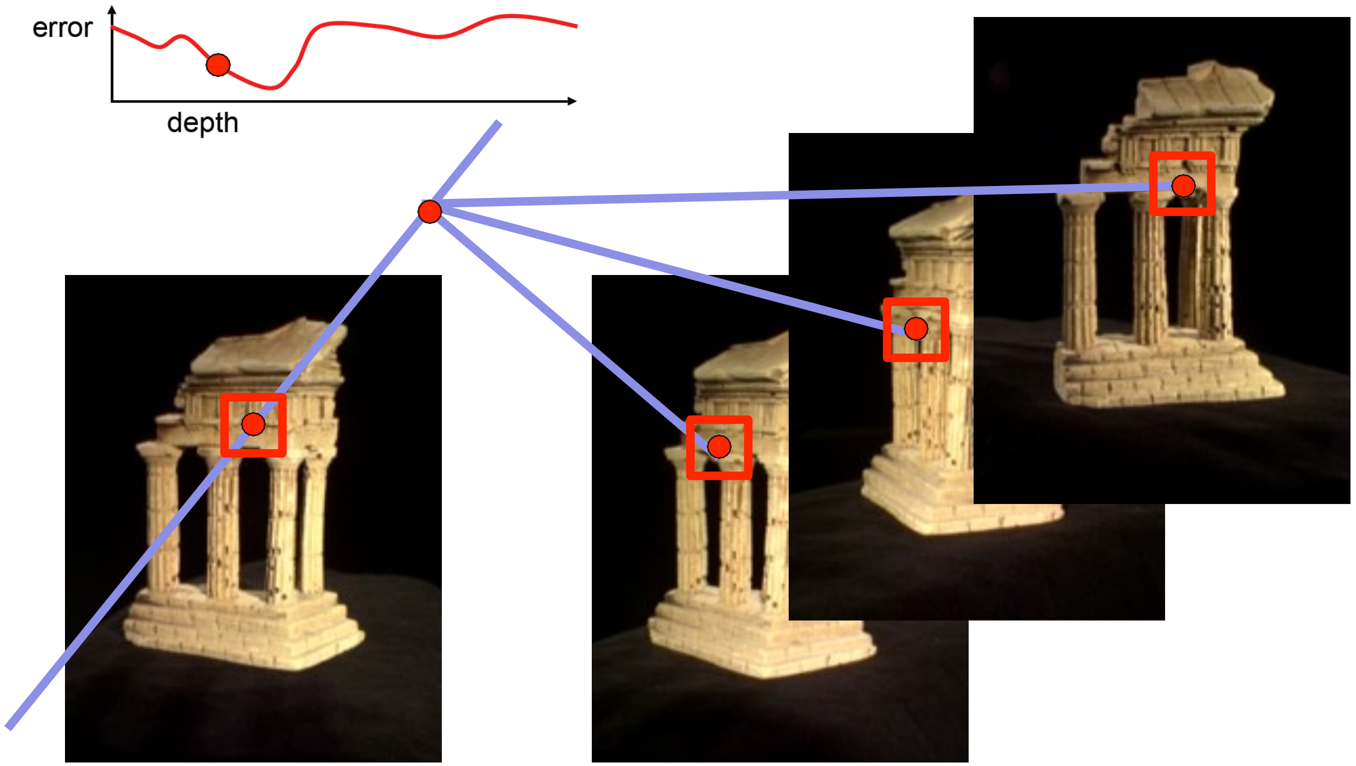
Multi-view Stereo: Basic idea



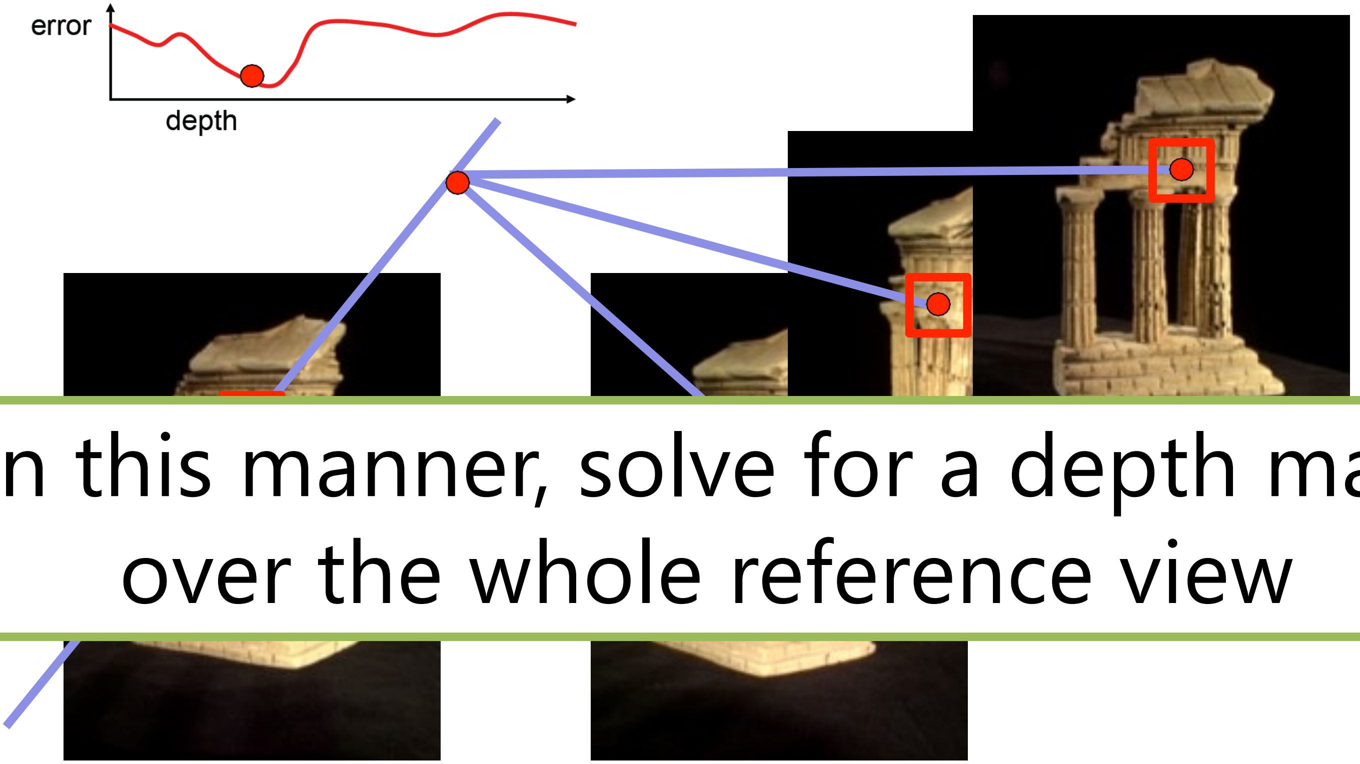
Multi-view Stereo: Basic idea



Multi-view Stereo: Basic idea



Multi-view Stereo: Basic idea



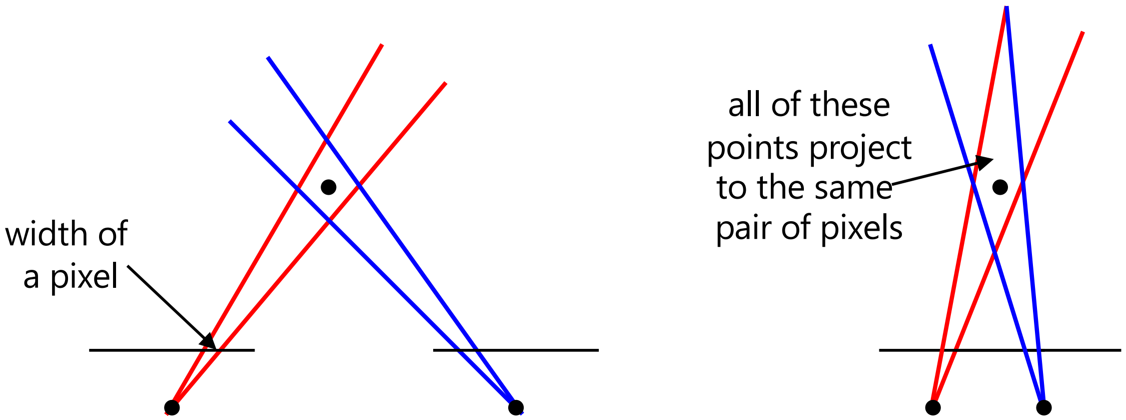
In this manner, solve for a depth map over the whole reference view

Multi-view Stereo: Advantages

- Can match windows using more than 1 neighbor, giving a **stronger match signal**
- If you have lots of potential neighbors, can **choose the best subset** of neighbors to match per reference image
- Can reconstruct a depth map for each reference frame, and then merge into a **complete 3D model**

PREREQUISITES AND PROCESS DETAILS

Choosing the Stereo baseline



Large Baseline **Small Baseline**

- What's the optimal baseline?
  - Too small: large depth error
  - Too large: difficult search problem

The Effect of Baseline on Depth Estimation

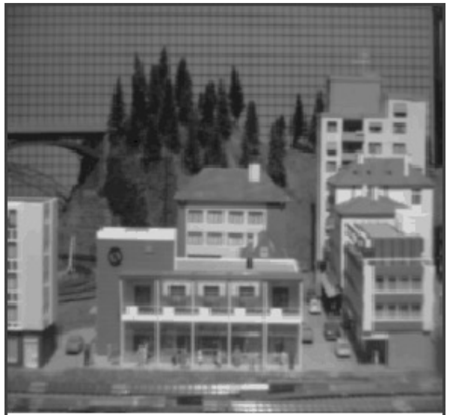
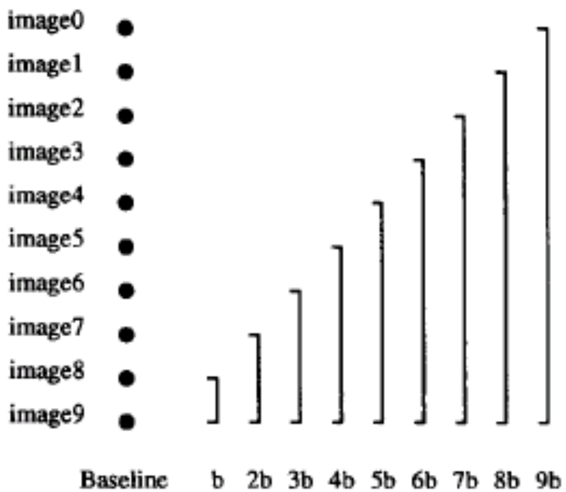
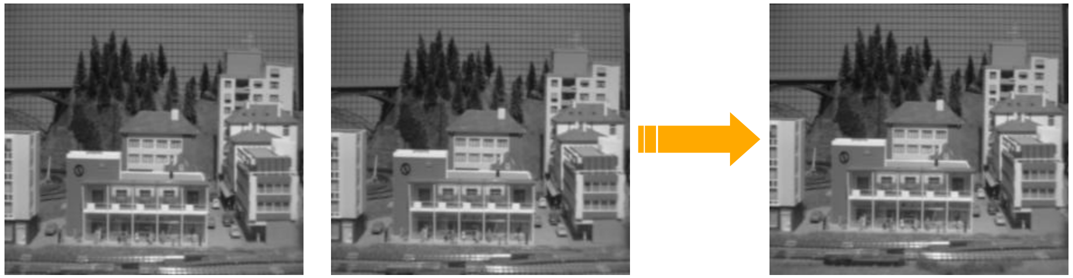


Figure 2: An example scene. The grid pattern in the background has ambiguity of matching.

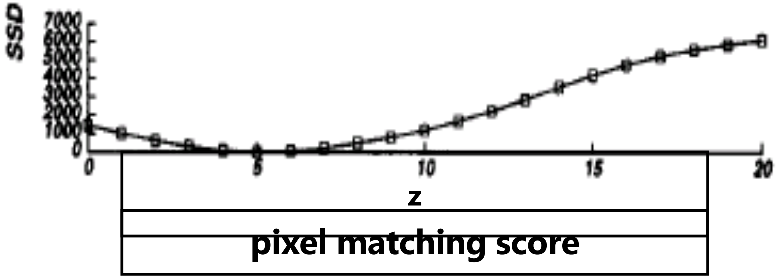


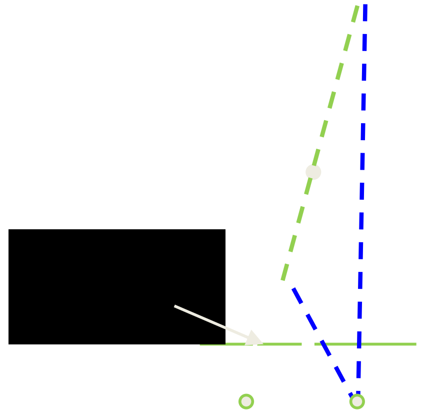


I_1 I_2 I_{10}

M. Okutomi and T. Kanade, "A Multiple-Baseline Stereo System," IEEE Trans. on Pattern Analysis and Machine Intelligence, 15(4):353-363 (1993).

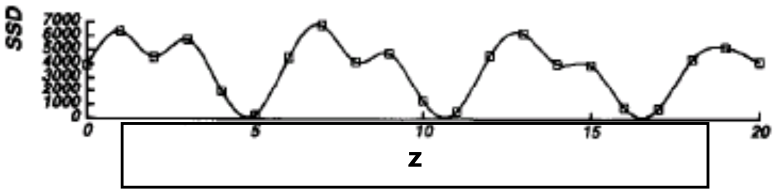
Multiple-baseline Stereo

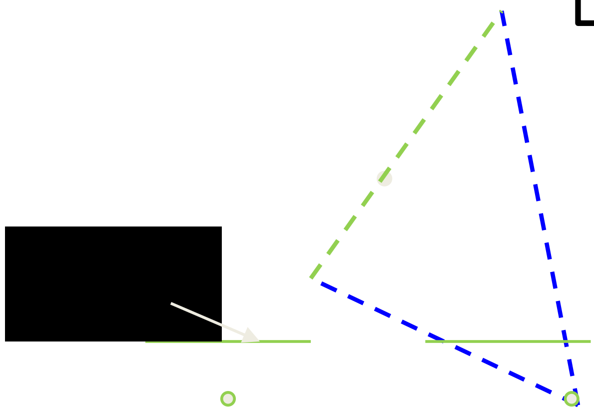




$$z = \frac{b\, f_x}{d}$$

- For short baselines, estimated depth will be less precise due to narrow triangulation





- For larger baselines, must search larger area in second image

M. Okutomi and T. Kanade, "A Multiple-Baseline Stereo System," IEEE Trans. on Pattern Analysis and Machine Intelligence, 15(4):353-363 (1993).

Multiple-baseline Stereo

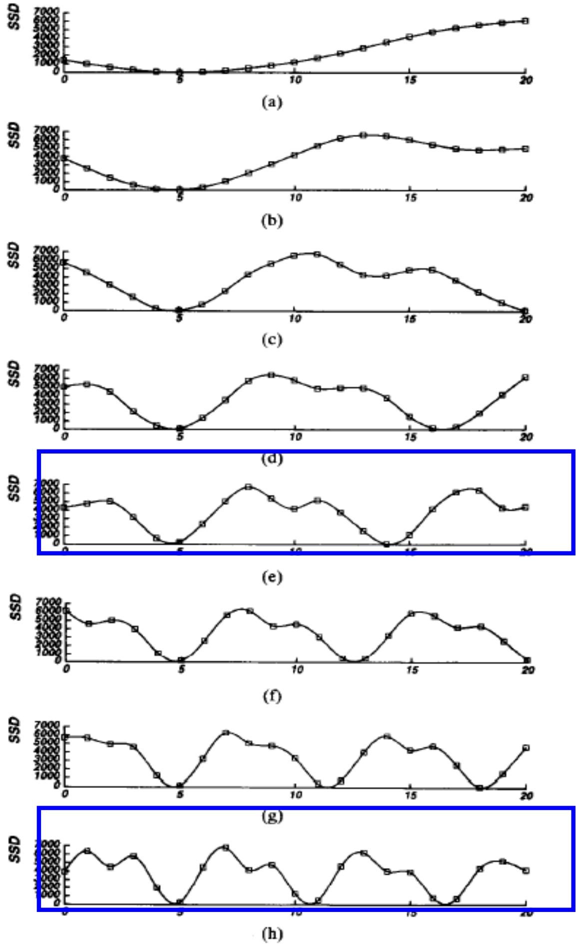


Fig. 5. SSD values versus inverse distance: (a) $B = b$; (b) $B = 2b$; (c) $B = 3b$; (d) $B = 4b$; (e) $B = 5b$; (f) $B = 6b$; (g) $B = 7b$; (h) $B = 8b$. The horizontal axis is normalized such that $8bF = 1$.

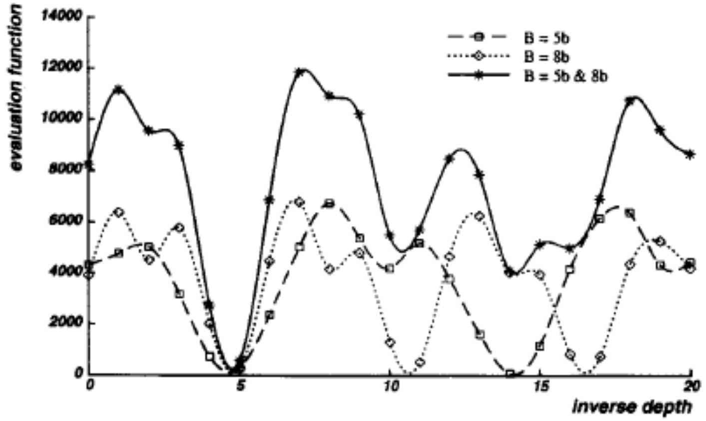


Fig. 6. Combining two stereo pairs with different baselines.

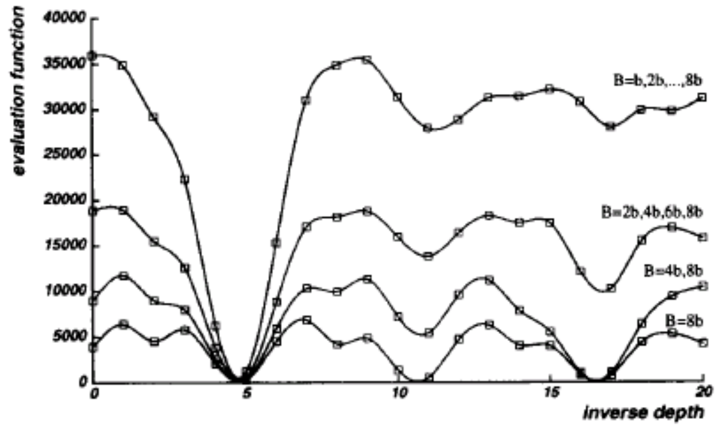
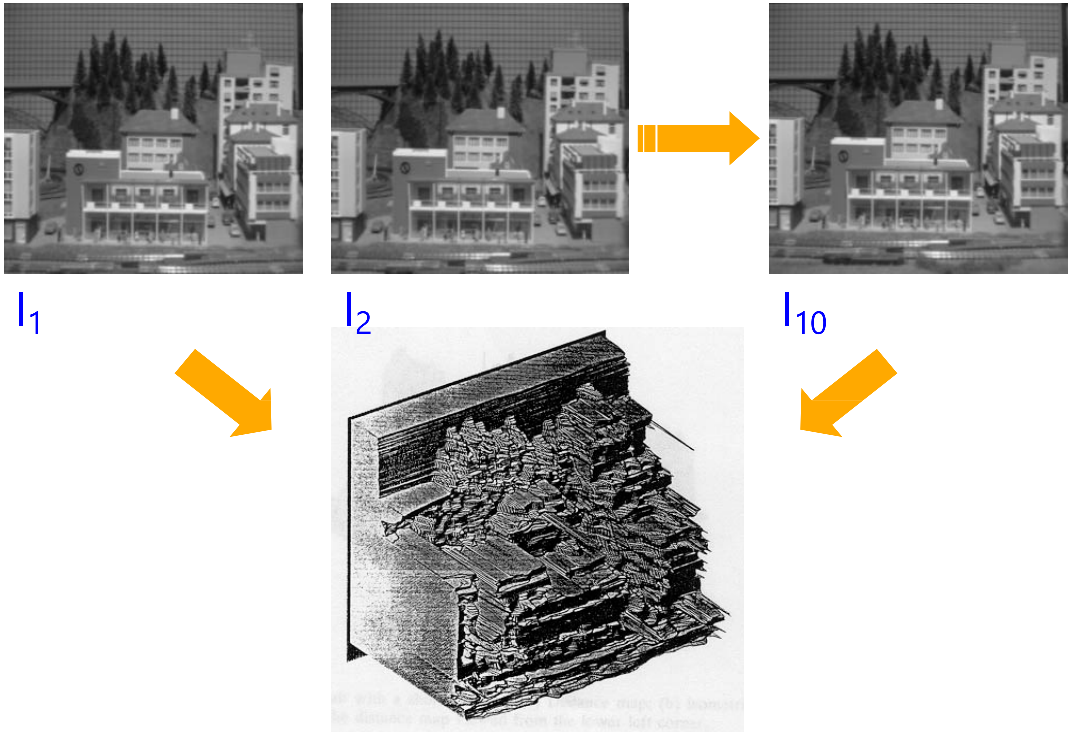


Fig. 7. Combining multiple baseline stereo pairs.

Multiple-baseline Stereo: Results



M. Okutomi and T. Kanade, *A Multiple-Baseline Stereo System,* IEEE Trans. on Pattern Analysis and Machine Intelligence, 15(4):353-363 (1993).

Multiple-baseline Stereo: Summary

Basic Approach

- Choose a reference view
- Use your favorite stereo algorithm BUT
  - replace two-view SSD with **SSSD** over all baselines
  - **SSSD**: the SSD values are computed first for each pair of stereo images, and then add all together from multiple stereo pairs.

Limitations

- Only gives a depth map (not an "object model")
- Won't work for widely distributed views.

Multiple-baseline Stereo: Problem

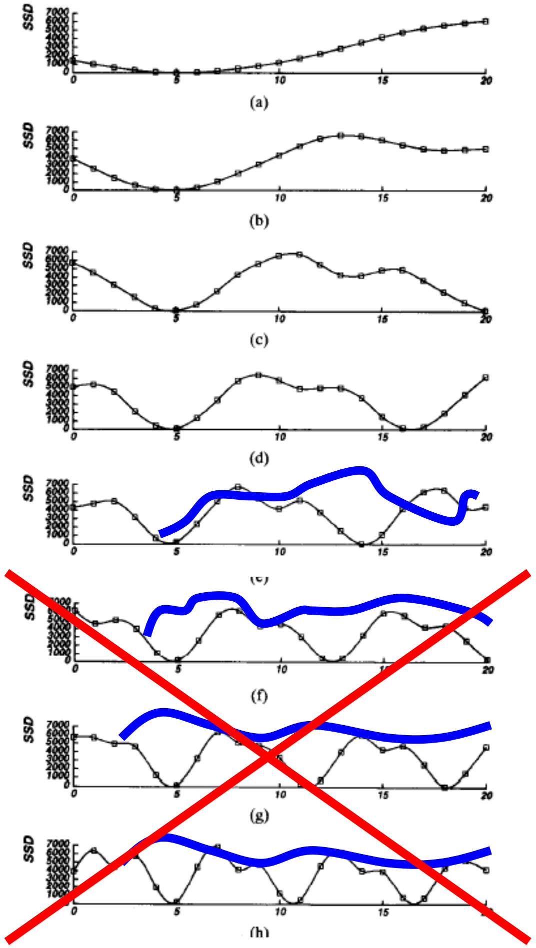


Fig. 5. SSD values versus inverse distance: (a) $B = b$; (b) $B = 2b$; (c) $B = 3b$; (d) $B = 4b$; (e) $B = 5b$; (f) $B = 6b$; (g) $B = 7b$; (h) $B = 8b$. The horizontal axis is normalized such that $8bF = 1$.

Visibility

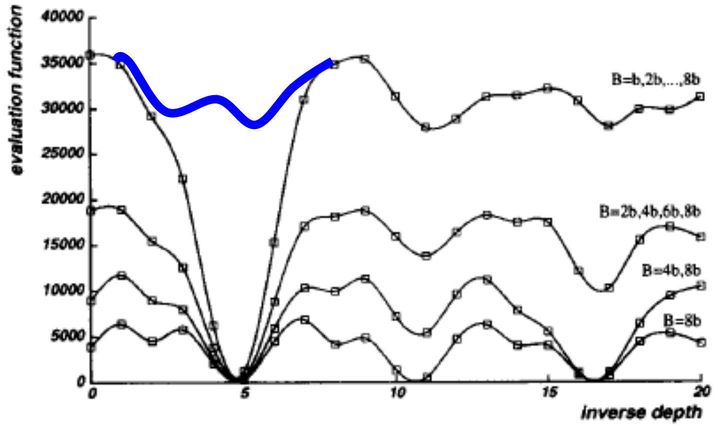


Fig. 7. Combining multiple baseline stereo pairs.

Some solutions

- Match only nearby photos [Narayanan 98]
- Use NCC instead of SSD, ignore NCC values > threshold [Hernandez & Schmitt 03]

Popular matching scores*

- SSD (Sum of Squared Differences) $\sum_{x,y} |W_1(x,y) - W_2(x,y)|^2$
- SAD (Sum of Absolute Differences) $\sum_{x,y} |W_1(x,y) - W_2(x,y)|$
- ZNCC (Zero-mean Normalized Cross Correlation)

$$\frac{\sum_{x,y}(W_1(x,y) - \overline{W_1})(W_2(x,y) - \overline{W_2})}{\sigma_{W_1}\sigma_{W_2}}$$

  – where $\overline{W_i} = \frac{1}{n}\sum_{x,y} W_i \qquad \sigma_{W_i} = \sqrt{\frac{1}{n}\sum_{x,y}(W_i - \overline{W_i})^2}$

  – what advantages might NCC have?

* also known as photo-consistency measures

Popular matching scores

* Original paper: Ramin Zabih and John Woodfill. 1994. Non-parametric local transforms for computing visual correspondence. In Proceedings of the third European conference on Computer Vision (Vol. II) (ECCV '94). Springer-Verlag, Berlin, Heidelberg, 151–158.

- Census*
  - Given a comparison operator

    $$\xi(a, b) = 1 \text{ if } a < b, 0 \text{ otherwise}$$

  - and a support domain Ω (neighbourhood) centered at p, census computes a bit string that describes whether a pixel in the support domain is brighter or darker than p

    $$census(f) = \bigotimes_{q \in \Omega} \xi(f(p), f(q)),$$

  - where ⊗ is the concatenation operator. The census score is computed as the Hamming distance of the two bit strings, which can be computed as the L1 norm of their difference:

    $$\rho_{census}(f, g) = |census(f) - census(g)|_1,$$

  - with values in [0,N], where N is the size of the domain Ω.

Is there more than just depth maps?

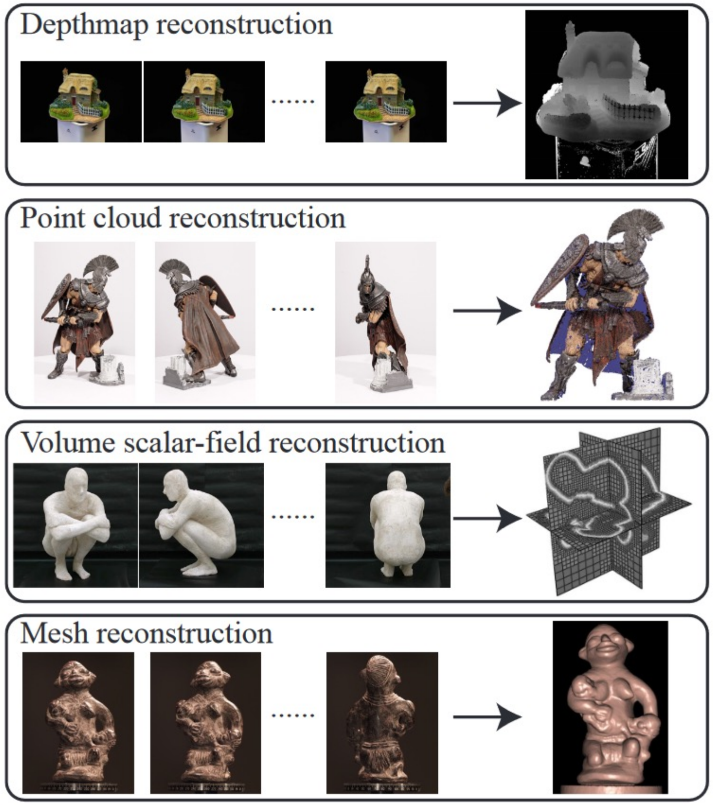


Figure 3.1: MVS algorithms can be classified based on the output scene representation. The four popular representations are a depthmap(s), a point cloud, a volume scalar-field, and a mesh. Note that a point cloud is very dense and may look like a textured mesh model, but is simply a collection of 3D points. Reconstruction examples are from state-of-the-art MVS algorithms presented in [48], [74], [94], and [93] respectively, from top to bottom.

Figures by Carlos Hernandez

Process

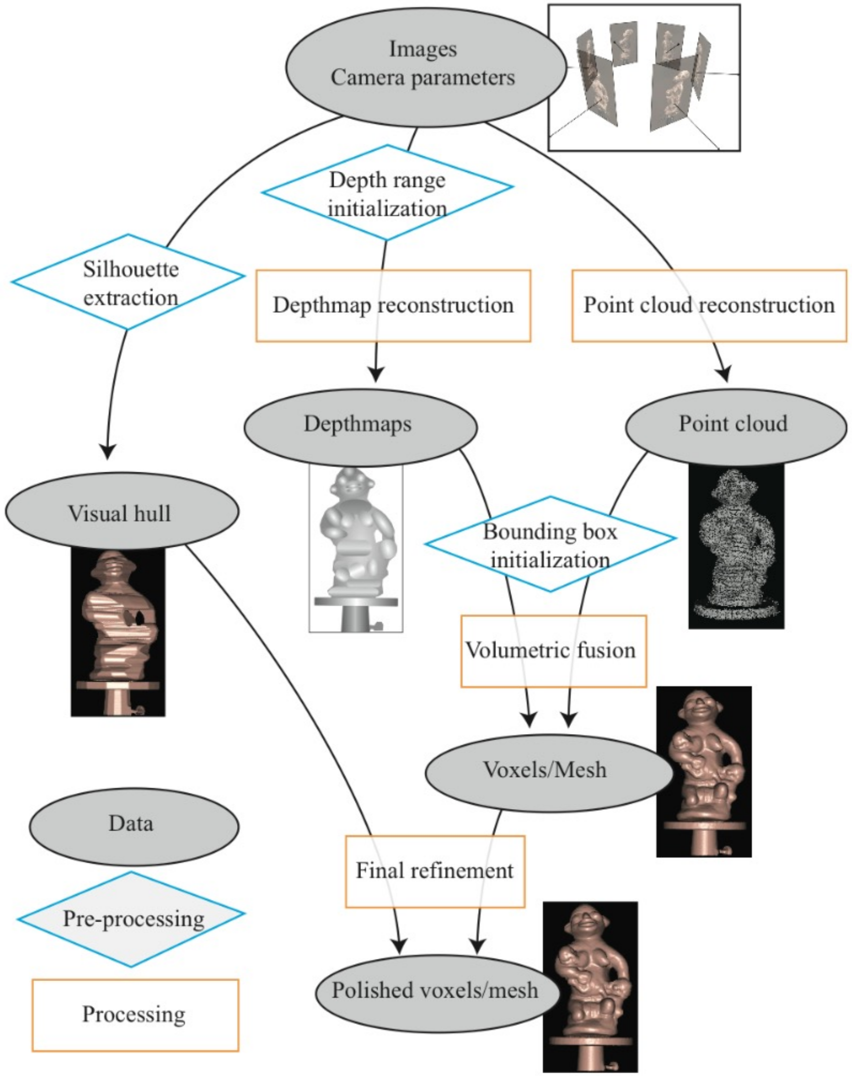
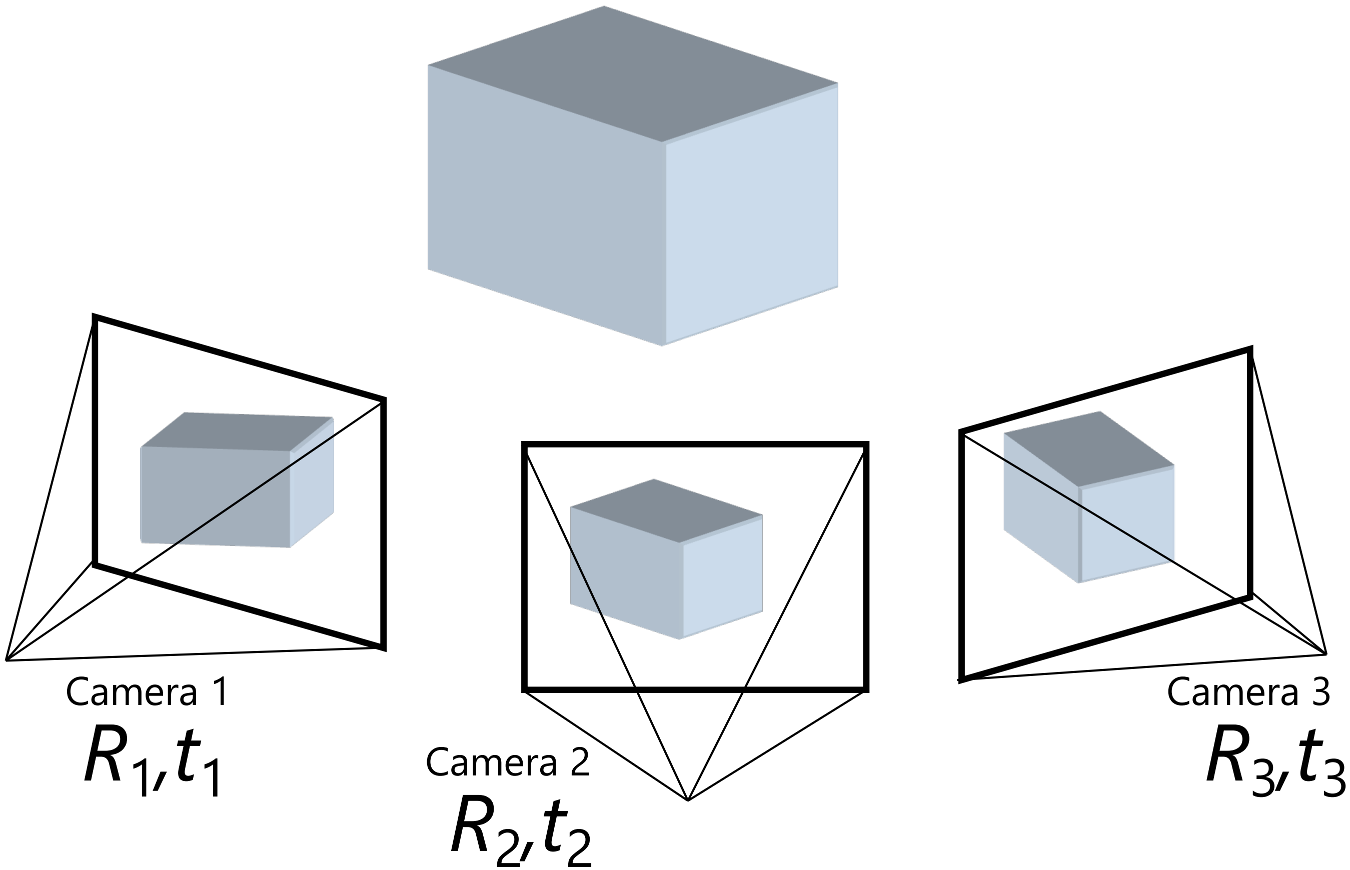


Figure 3.2: MVS processing often consists of multiple stages (algorithms), and the figure illustrates typical processing flow. 3D data representation determines how different algorithms can be put together, and gray ovals correspond to data representations, which are connected by either an MVS algorithm or a pre-processing step. Note that "Polished voxels/mesh" is not necessarily the goal of every MVS system. Depending on the applications, the final step of the MVS system would be different.

Figures by Carlos Hernandez

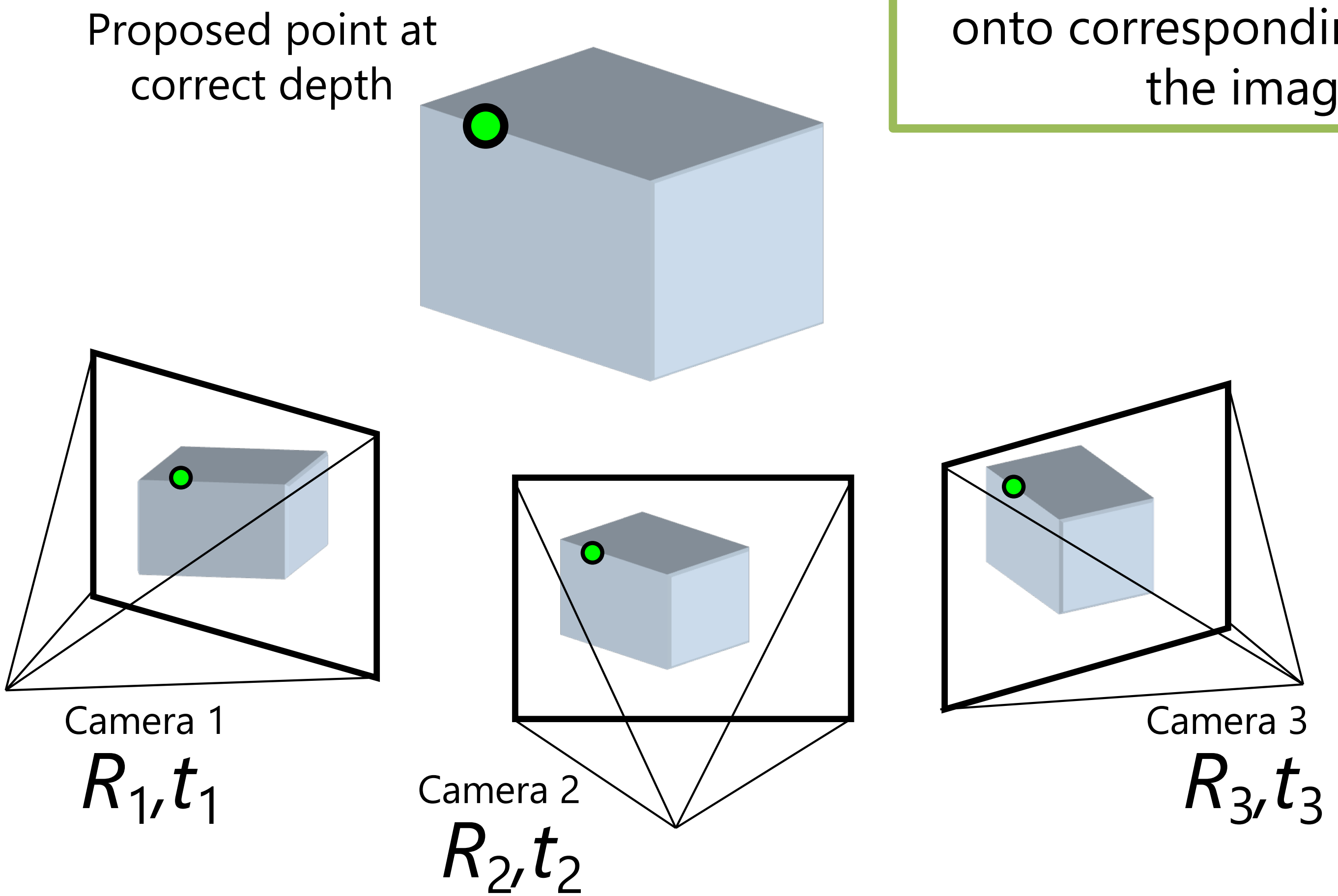
Plane-Sweep Stereo: Assumption

We know the camera poses.

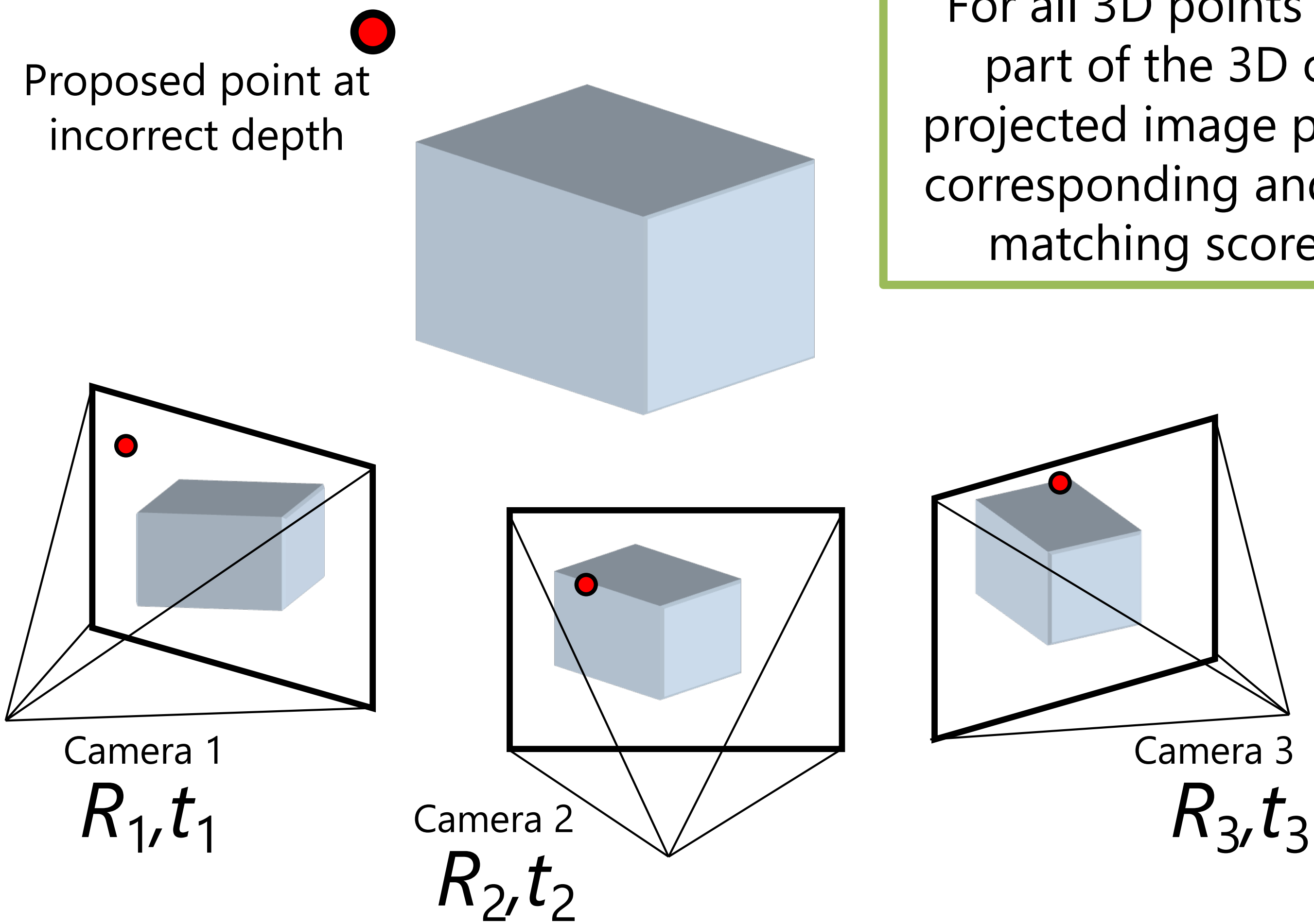


Plane-Sweep Stereo: Assumption

Object points are projected onto corresponding points in the images.



Plane-Sweep Stereo: Assumption

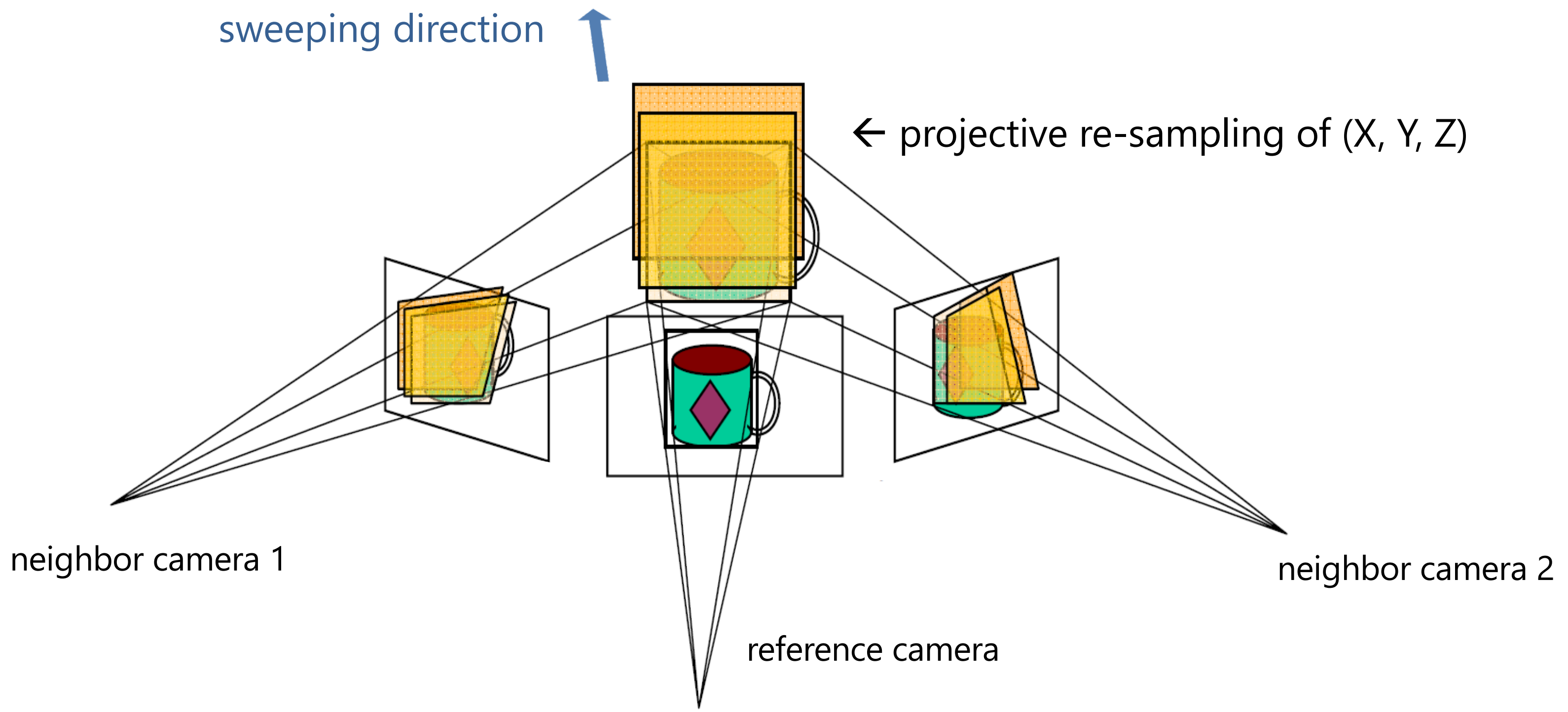


For all 3D points that are not part of the 3D object, the projected image points are not corresponding and hence their matching scores are bad.

Plane-Sweep Stereo

Sweep family of planes parallel to the reference camera image plane

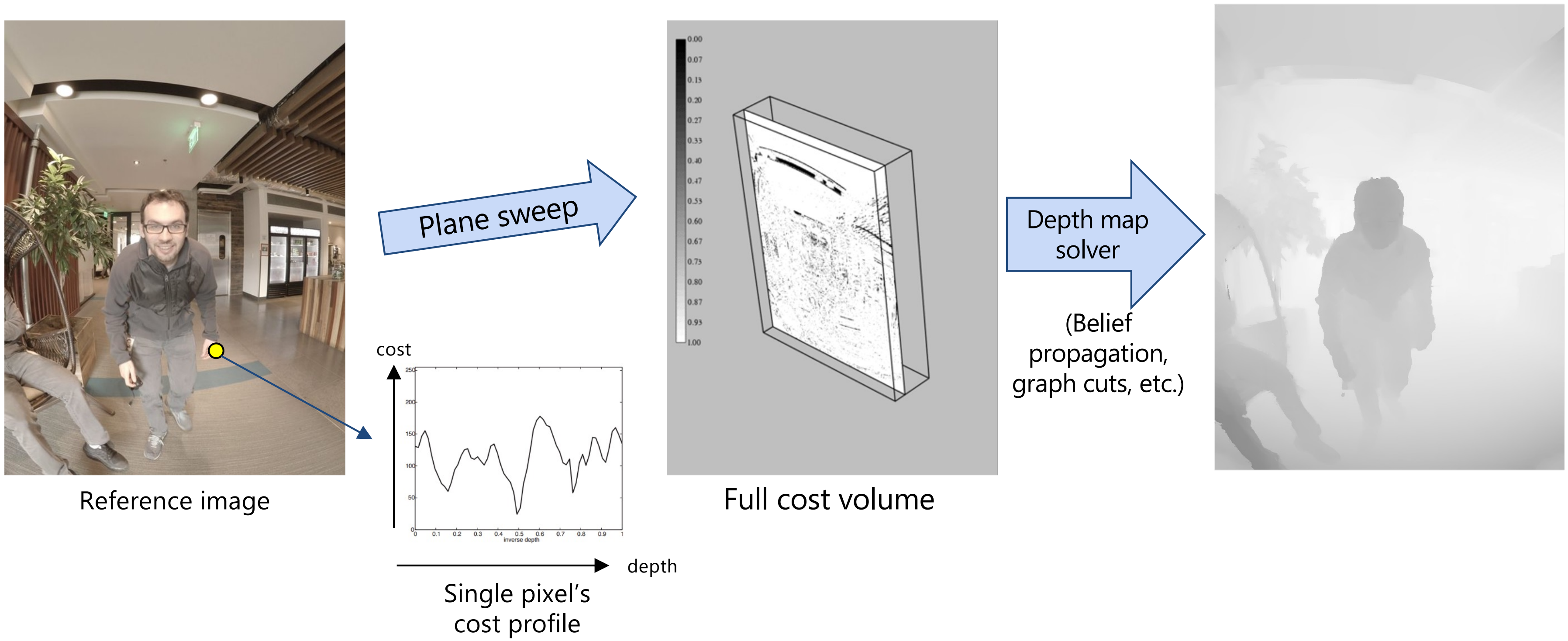
Reproject neighbors onto each plane (via homography) and compare reprojections



Plane-Sweep Stereo

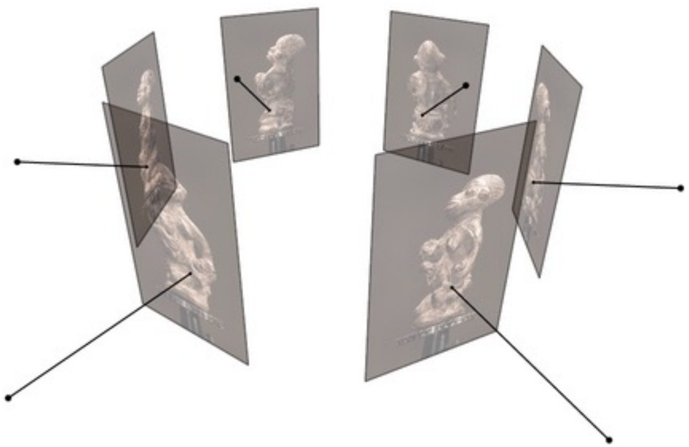
Another example

Cost Volumes -> Depth Maps



Fusing multiple depth maps

- Compute depth map per image
- Fuse the depth maps into a 3D model



Figures by Carlos Hernandez

Another approach: NeRF

- Represent scenes as functions from (x, y, z) to RGB and alpha (transparency), use volume rendering to render images

NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020

https://www.matthewtancik.com/nerf