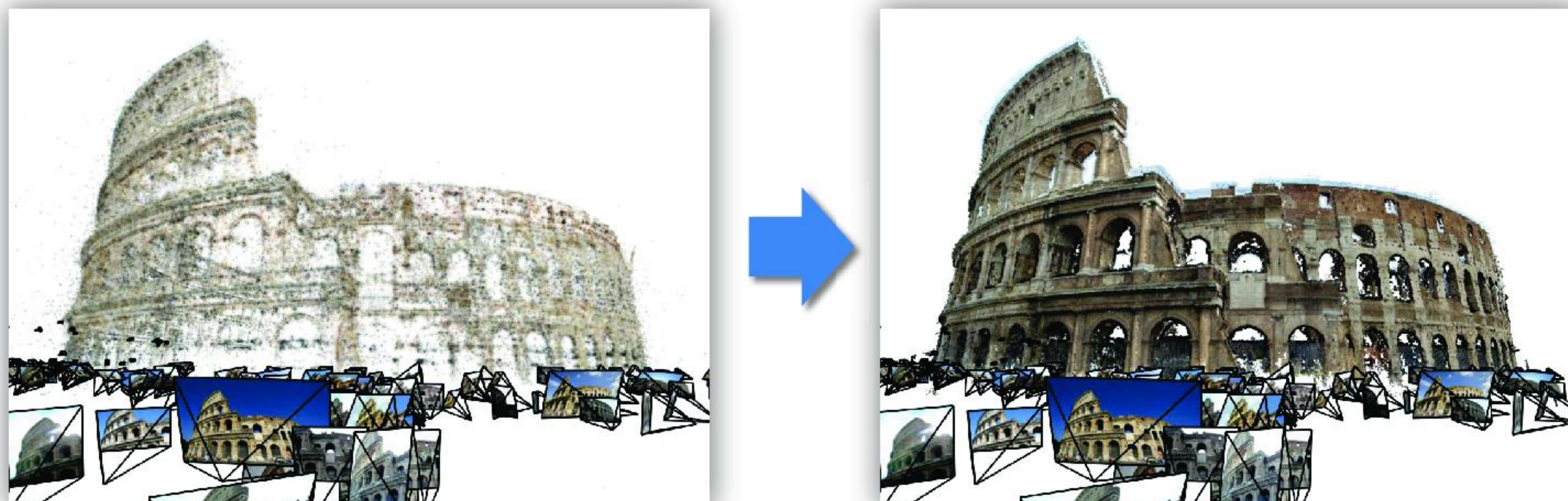
CS5670: Computer Vision

Noah Snavely / Zhengqi Li

Multi-view stereo



Recommended Reading

Szeliski Chapter 11.6

Multi-View Stereo: A Tutorial

Furukawa and Hernandez, 2015

http://www.cse.wustl.edu/~furukawa/papers/fnt_mvs.pdf

Multi-view Stereo

What is stereo vision?

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

Stereo

Multi-view stereo

Multi-view Stereo

Point Grey's Bumblebee XB3

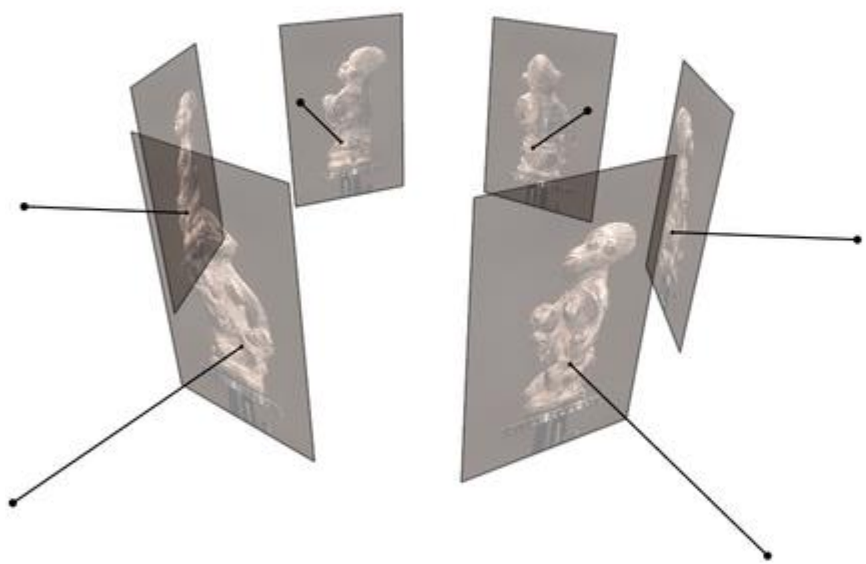
Point Grey's ProFusion 25

CMU's 3D Room

Multi-view Stereo

Input: calibrated images from several viewpoints

Output: 3D object model



Figures by Carlos Hernandez

What is stereo vision?

- **Generic problem formulation: given several images of the same object or scene, compute a representation of its 3D shape.**
- **“Images of the same object or scene”**
  - Arbitrary number of images (from two to thousands)
  - Arbitrary camera positions (camera network or video sequence)
  - Calibration may be initially unknown
- **“Representation of 3D shape Representation of 3D shape ”**
  - Depth maps
  - Meshes
  - Point clouds
  - Patch clouds
  - Volumetric models
  - Layered models

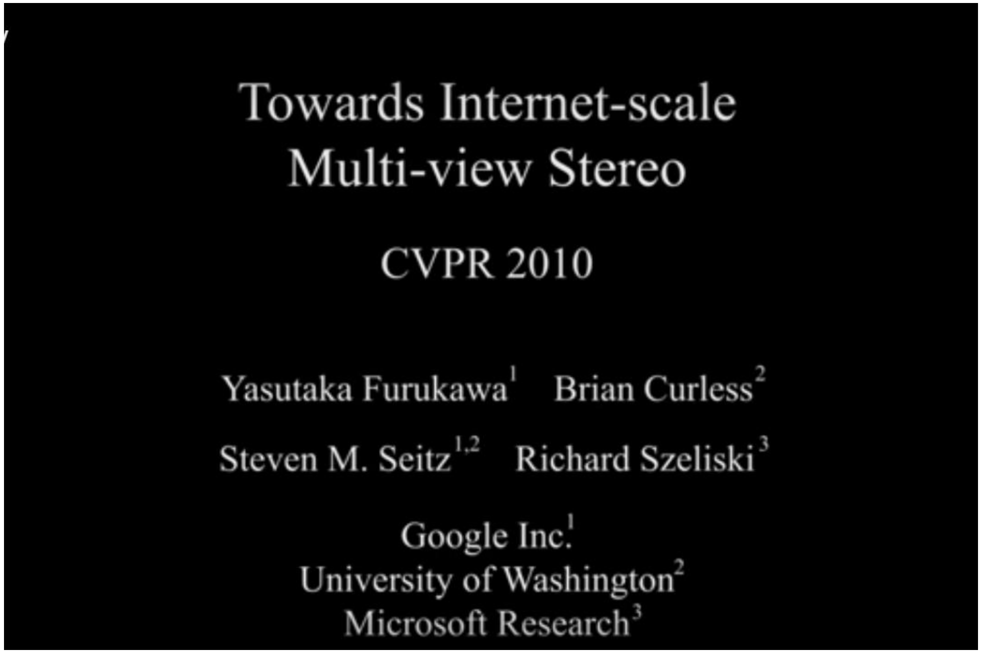
Faugeras, Keriven
1998

Hernandez, Schmitt
2004

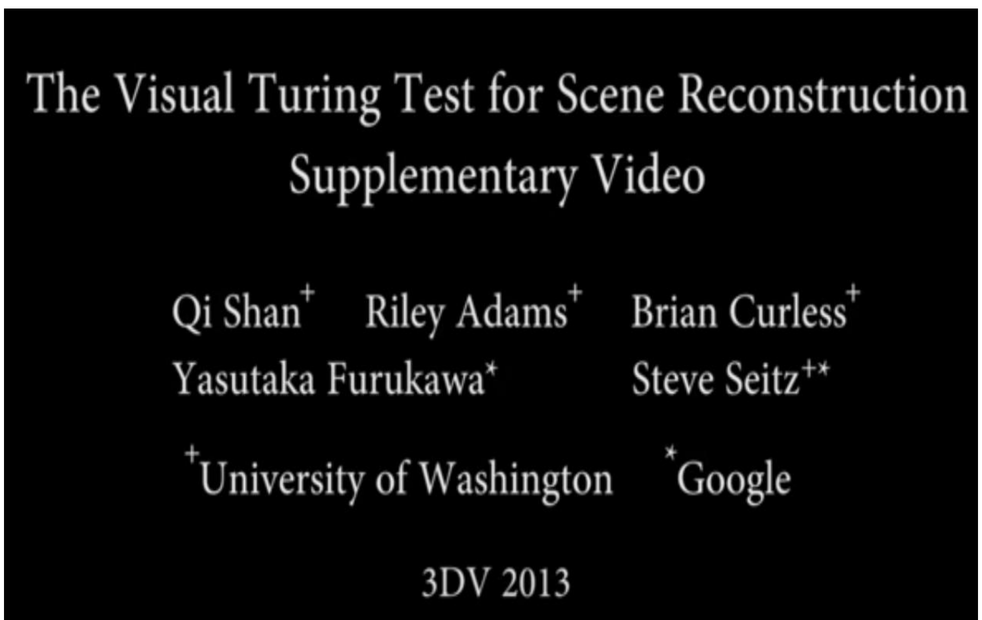
Pons, Keriven, Faugeras
2005

Furukawa, Ponce
2006

Goesele et al.
2007



https://www.youtube.com/watch?v=ofHFOr2nRxU



https://www.youtube.com/watch?v=NdeD4cjLI0c&t=64s

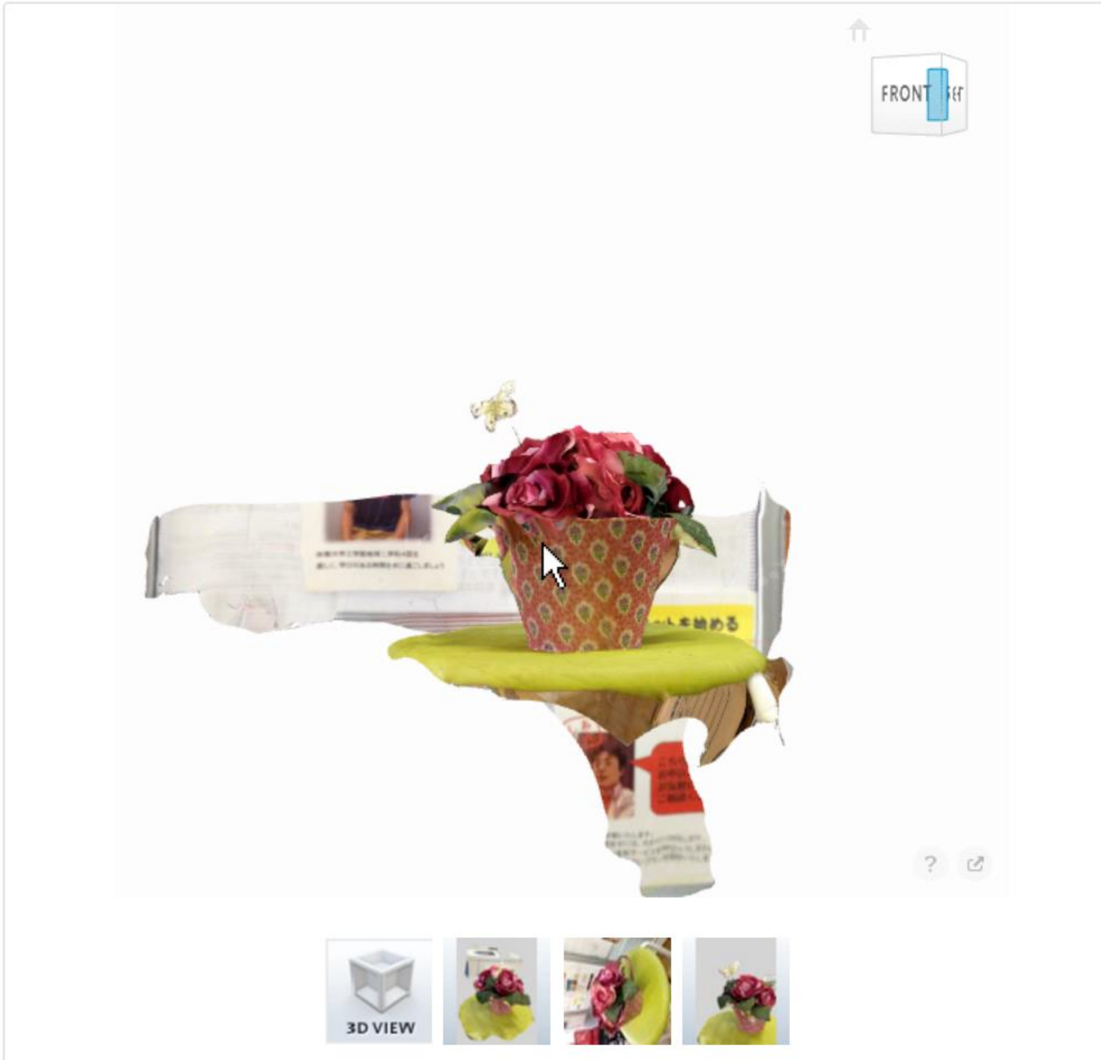
Applications

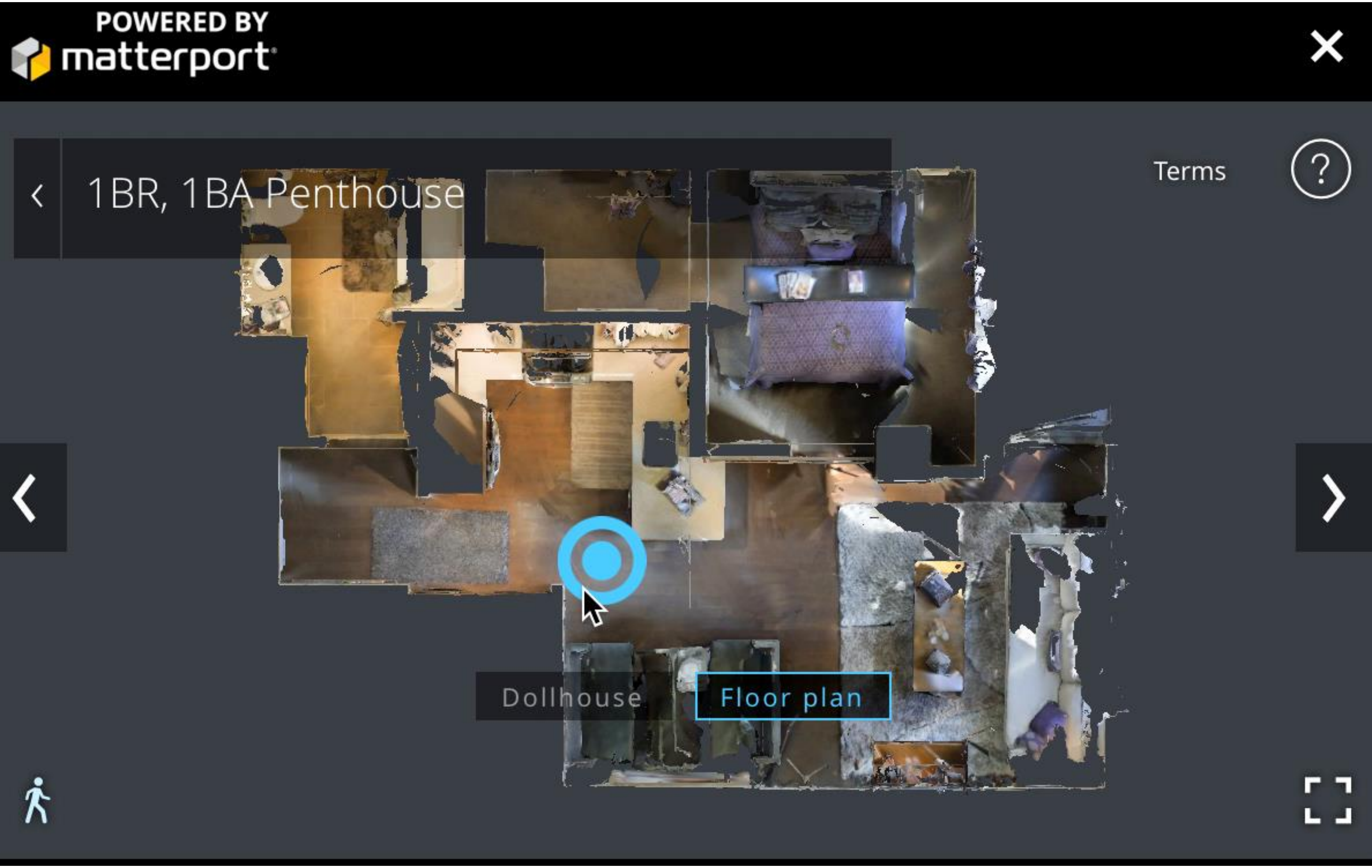

Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Google earth

AUTODESK® 123D® APPS 3D MODELS PROJECTS TOOLS BLOG

でんじろんさん

by じろう でん 266,




POWERED BY
matterport
1BR, 1BA Penthouse
Terms
Dollhouse
Floor plan

Whistle in the Form of Female Figure *600 AD - 900 AD*

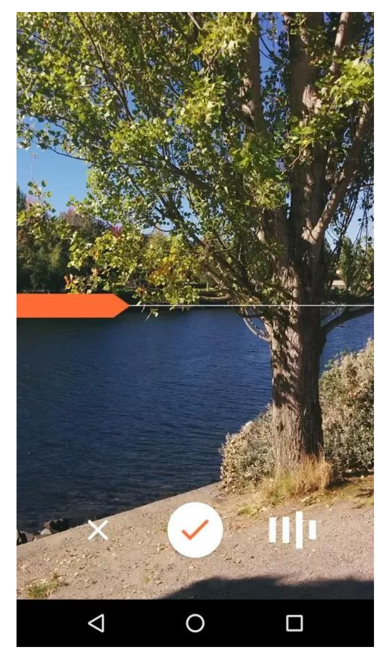
Details
Los Angeles County Museum of Art

Los Angeles County Museum of Art
Sculpture
Mexico

Share
Compare
Saved 0
Discover
Google

Google

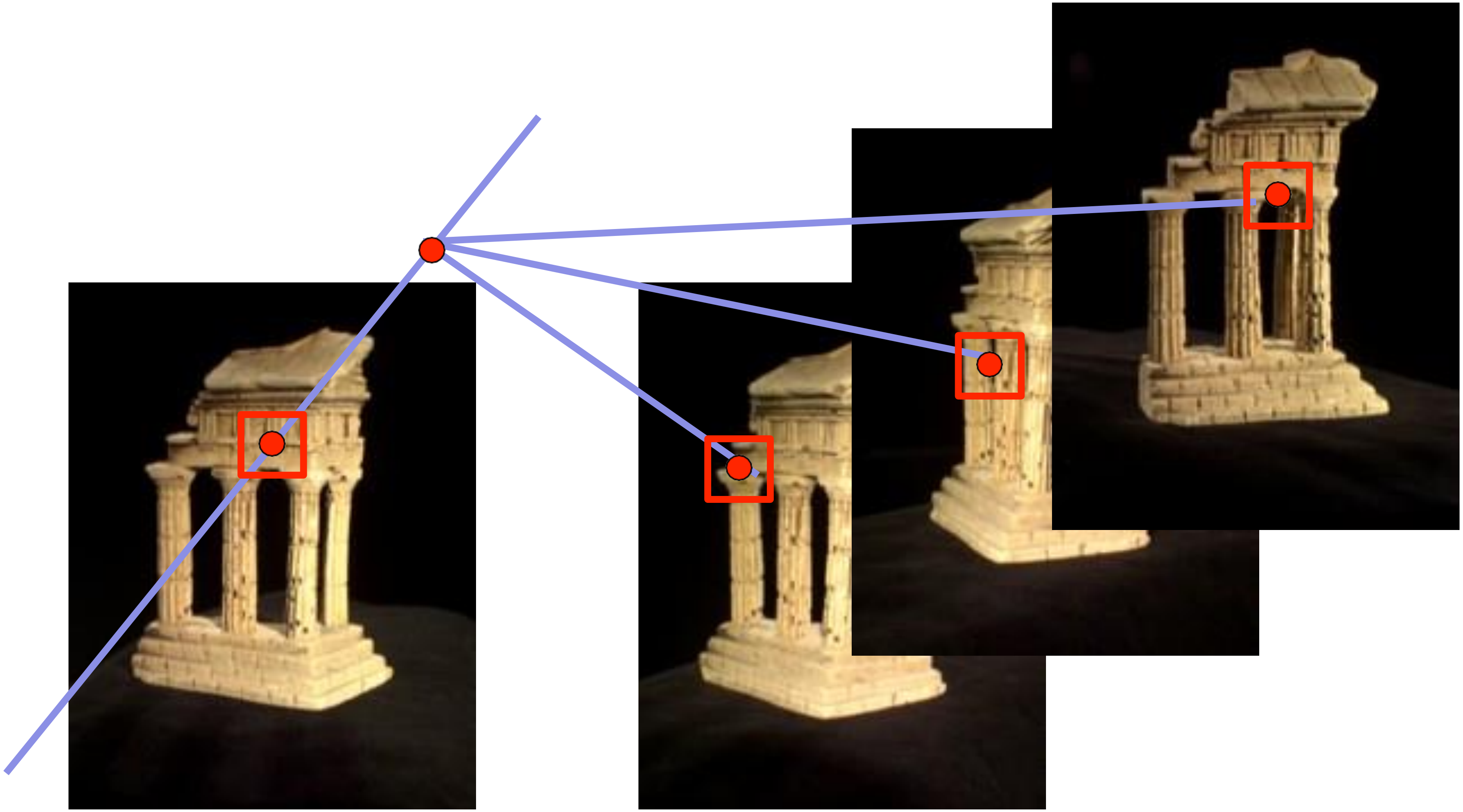
JUMP




cnet

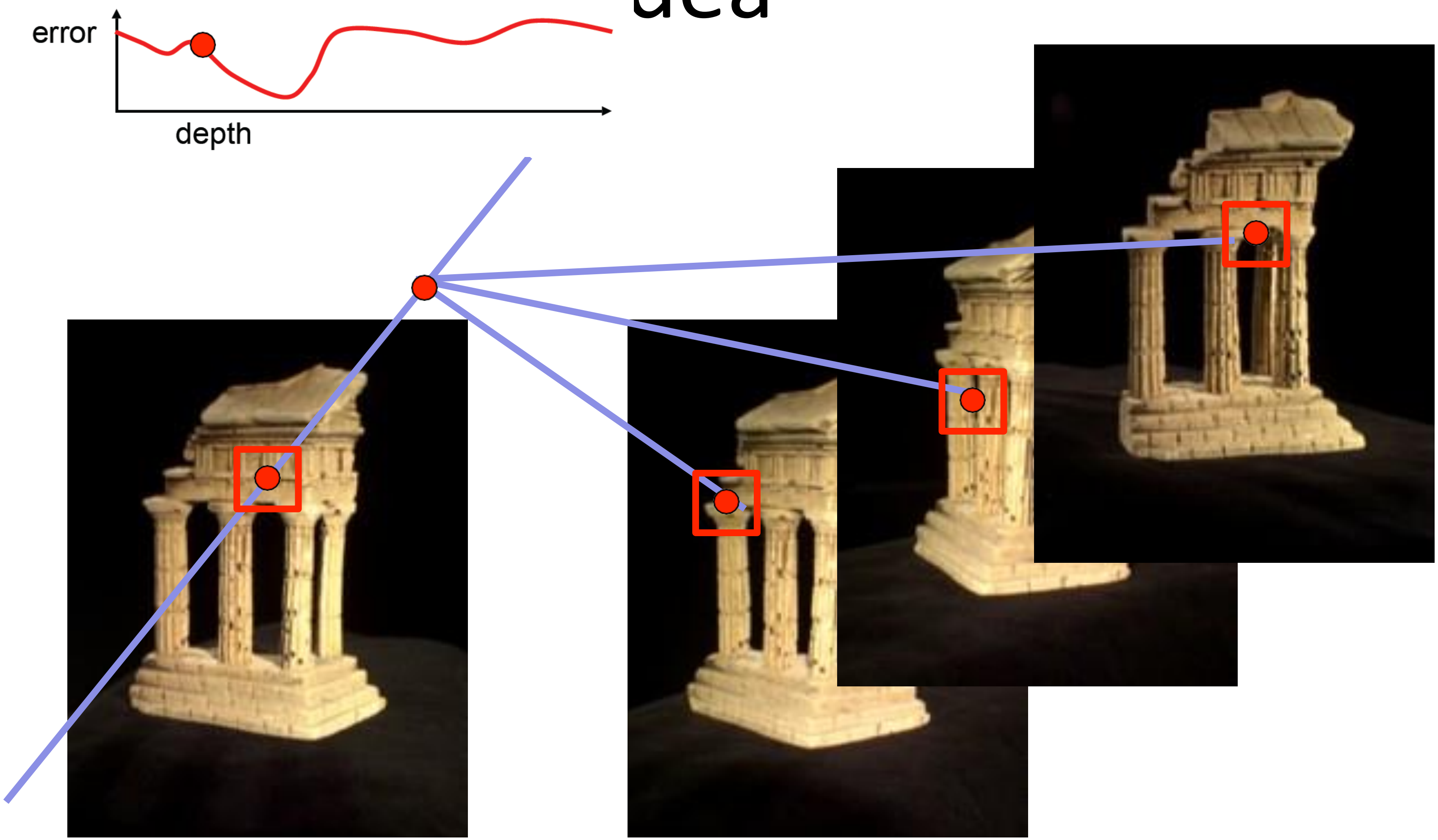
https://code.facebook.com/posts/1755691291326688/introducing-facebook-surround-360-an-open-high-quality-3d-360-video-capture-system?hc_location=ufi

Multi-view stereo: Basic idea



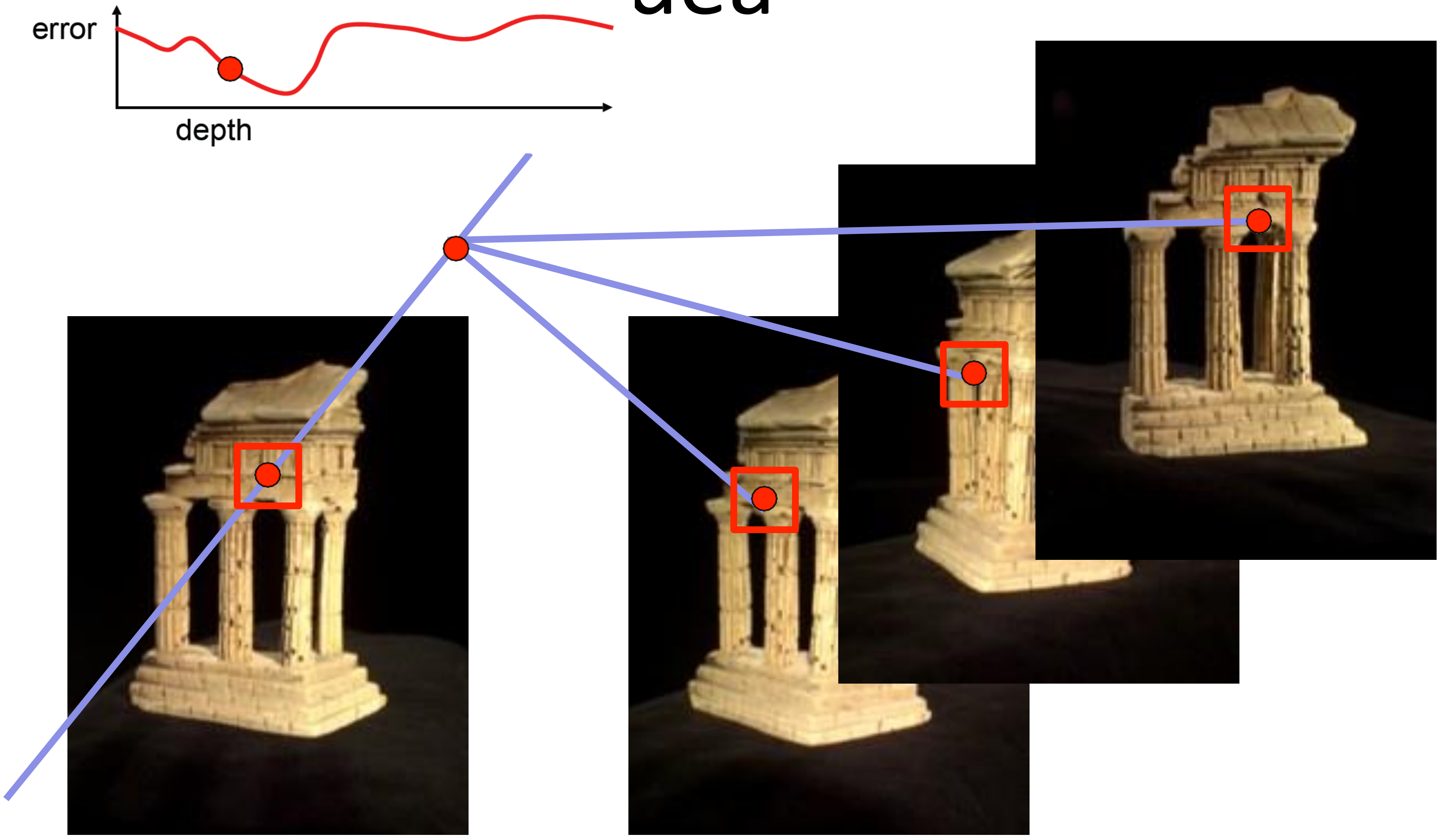
Source: Y. Furukawa

Multi-view stereo: Basic idea



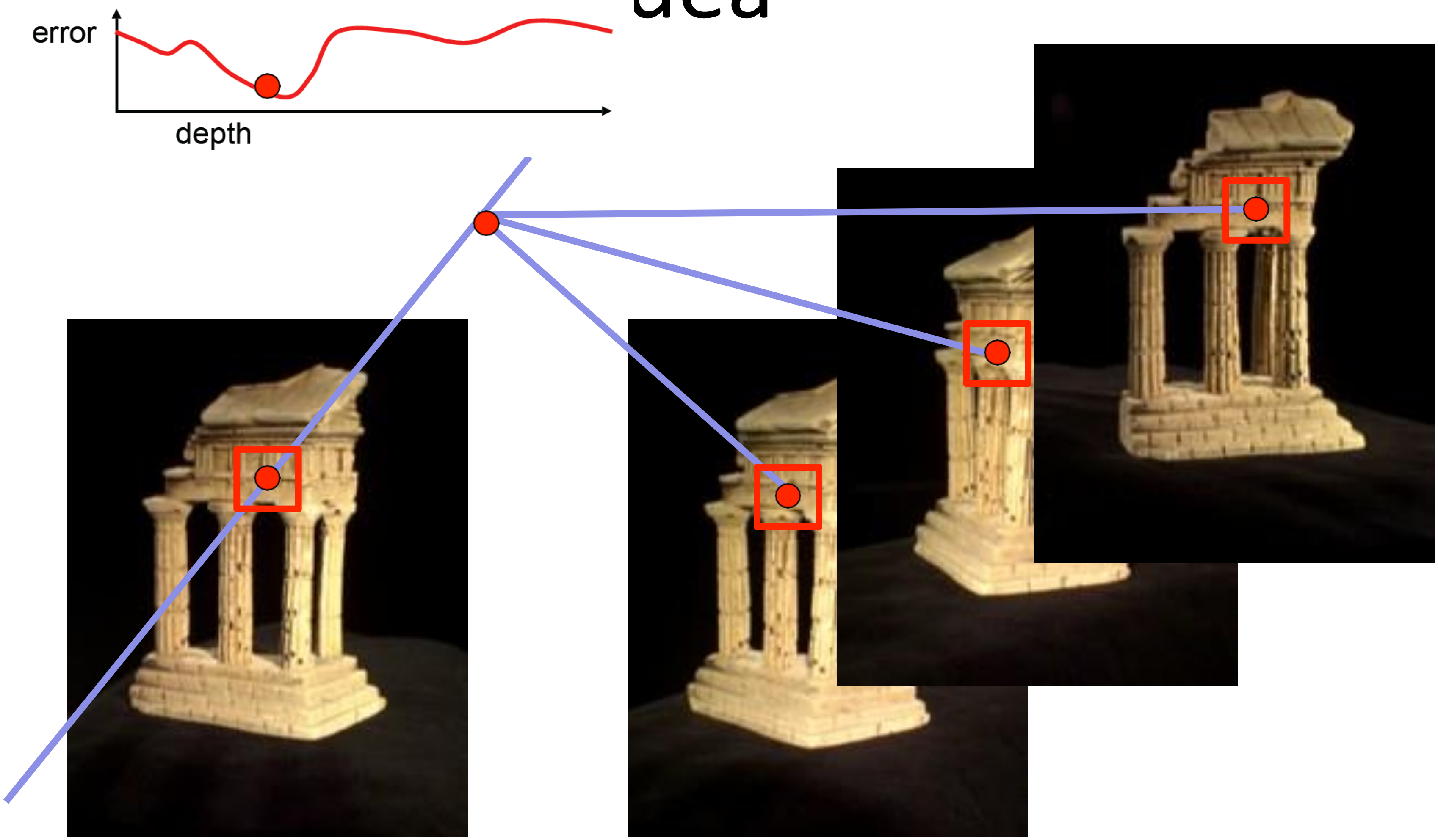
Source: Y. Furukawa

Multi-view stereo: Basic idea

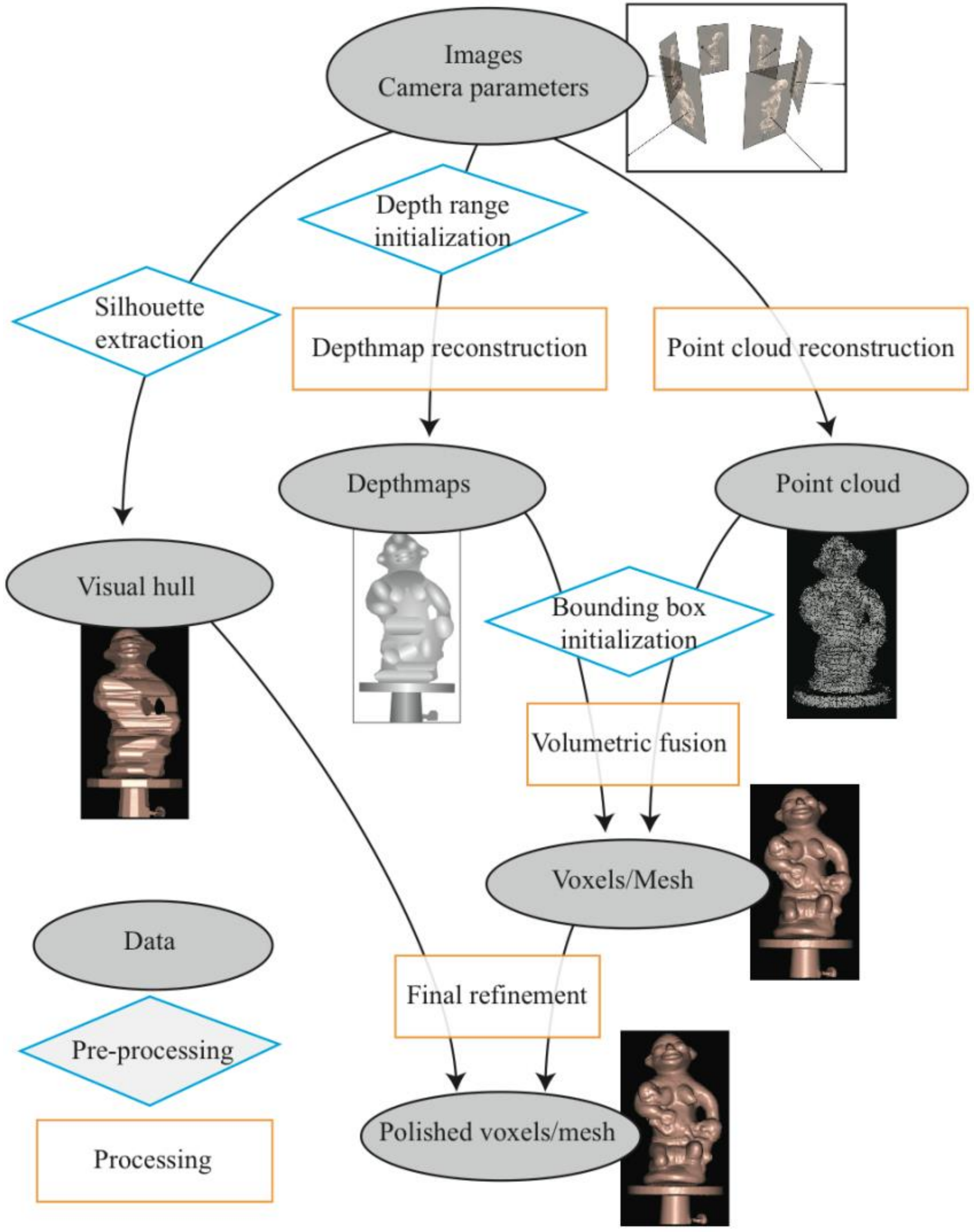


Source: Y. Furukawa

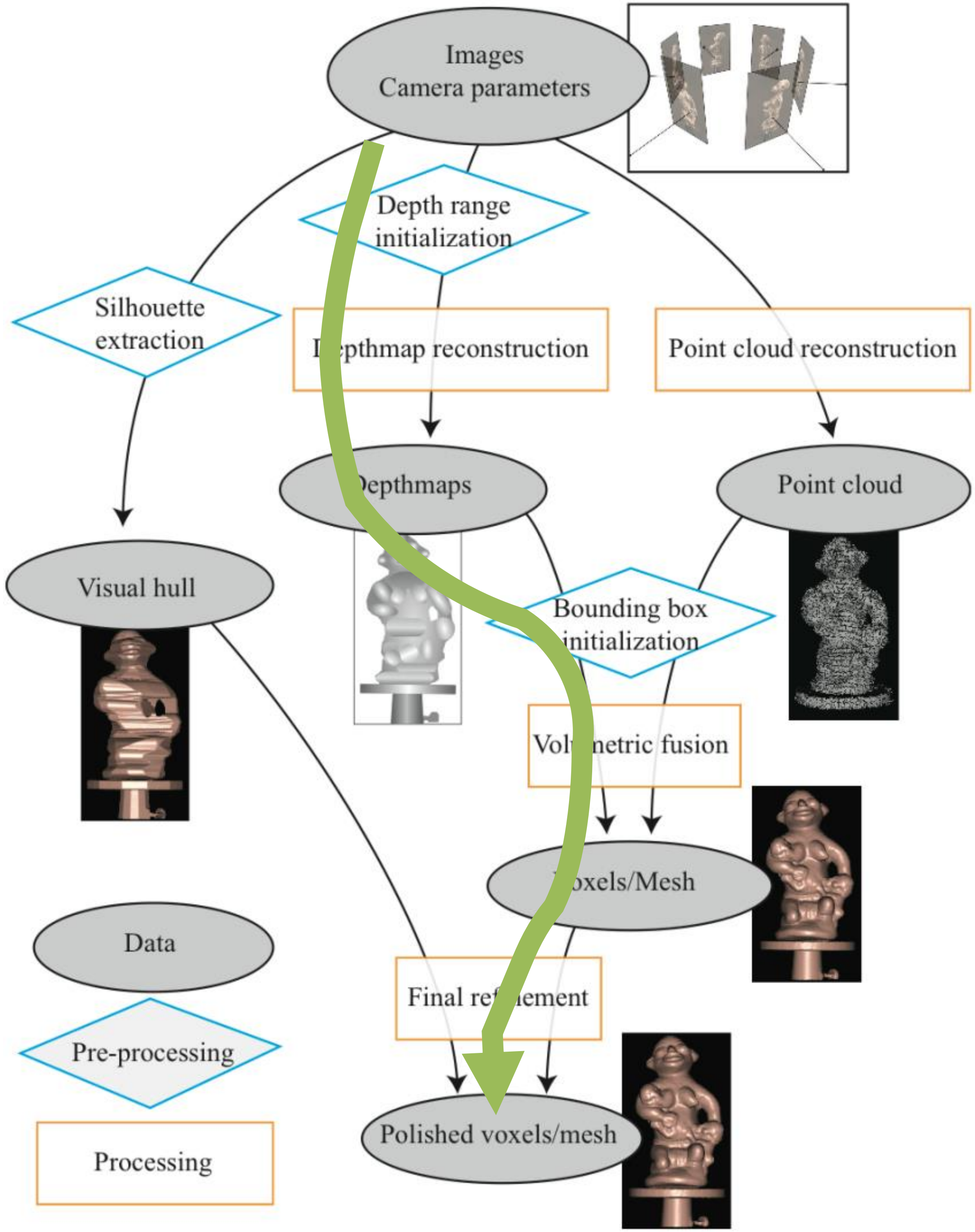
Multi-view stereo: Basic idea

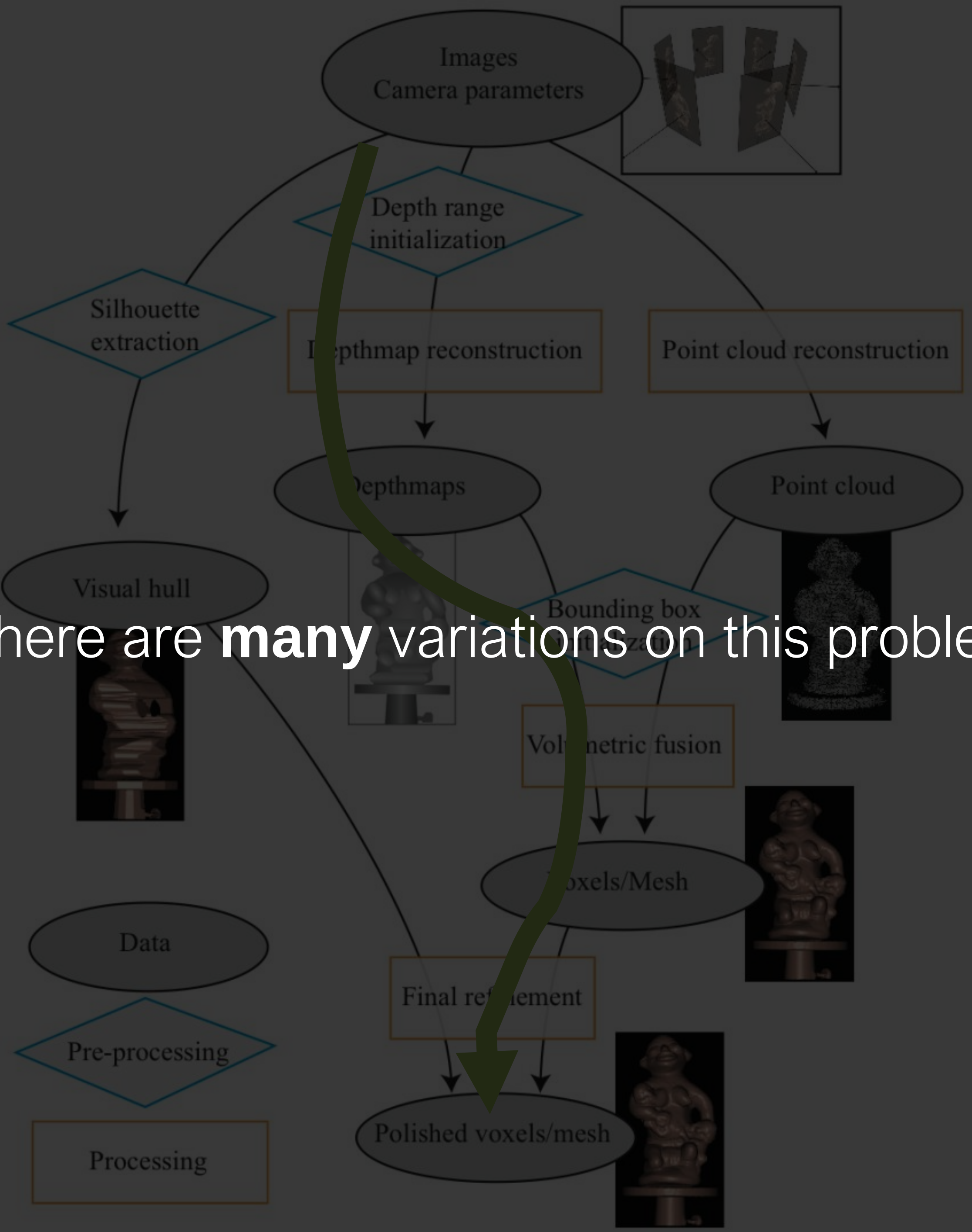


Source: Y. Furukawa



Images
Camera parameters
Depth range
initialization
Silhouette
extraction
Depthmap reconstruction
Point cloud reconstruction
Depthmaps
Point cloud
Visual hull
Bounding box
initialization
Volumetric fusion
Voxels/Mesh
Data
Final refinement
Pre-processing
Polished voxels/mesh
Processing

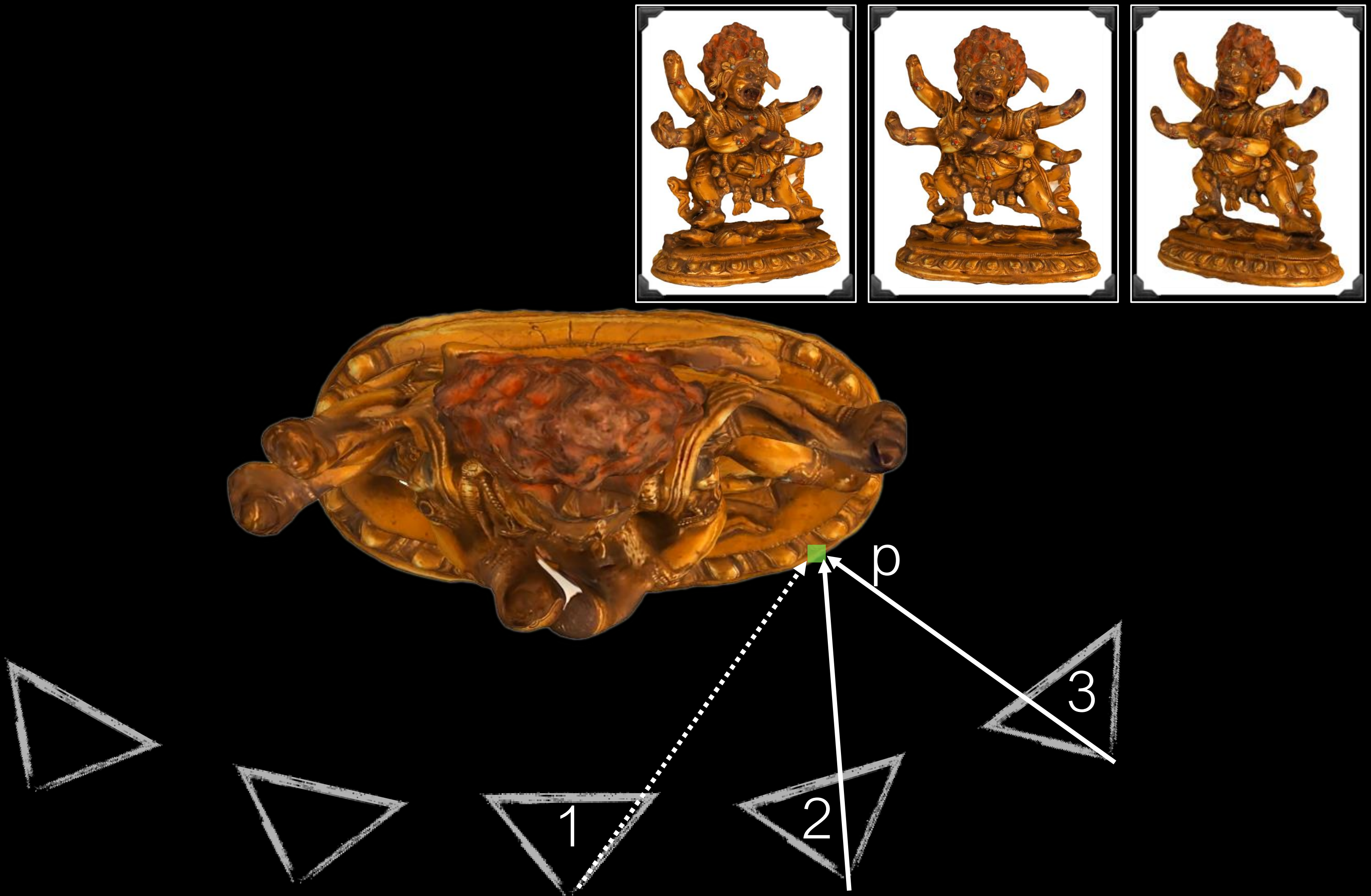

Images
Camera parameters
Depth range initialization
Silhouette extraction
Depthmap reconstruction
Point cloud reconstruction
Depthmaps
Point cloud
Visual hull
Bounding box initialization
Volumetric fusion
Voxels/Mesh
Data
Final refinement
Pre-processing
Polished voxels/mesh
Processing



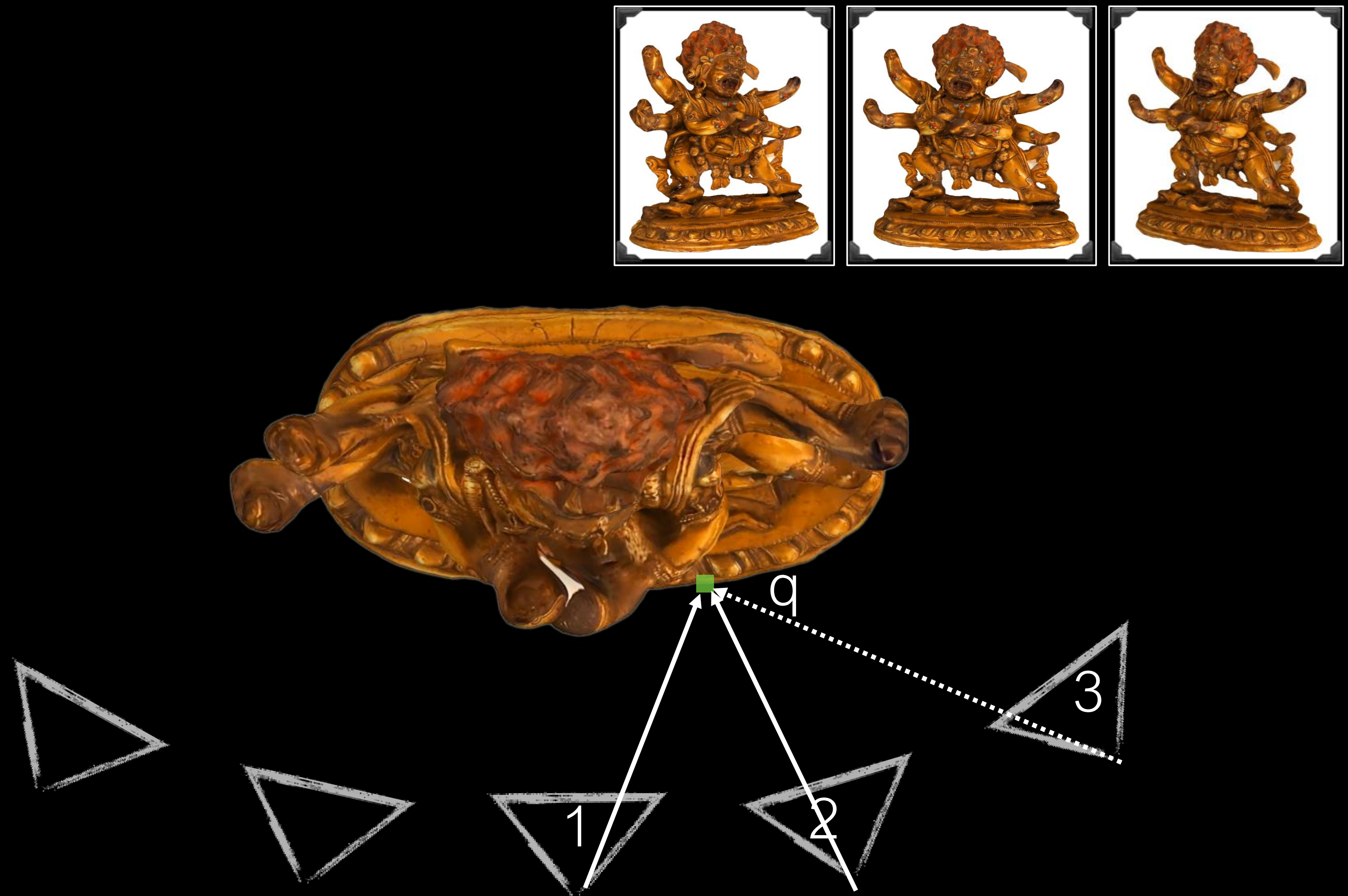
There are **many** variations on this problem.

Why MVS?

- Different points on the object’s surface will be more clearly visible in some subset of cameras
  - Could have high-res closeups of some regions
  - Some surfaces are foreshortened from certain views



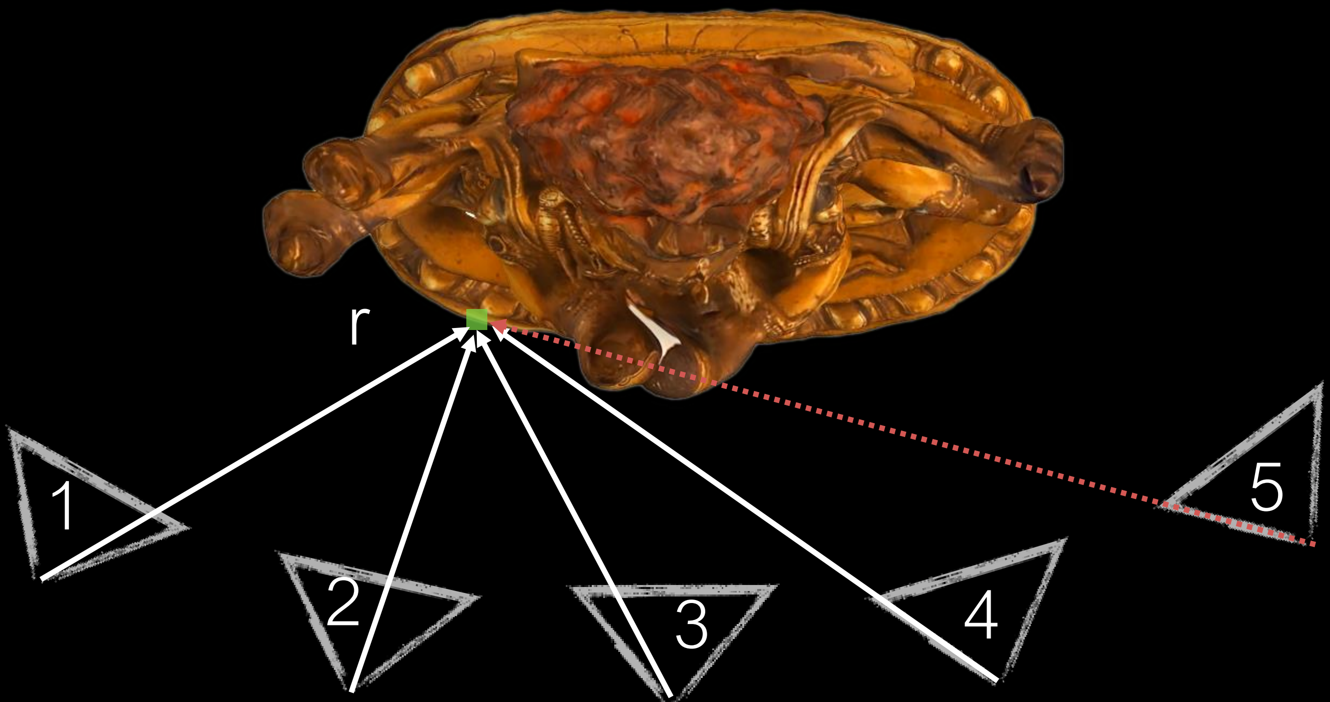
Cameras 2 and 3 can more clearly see point p.



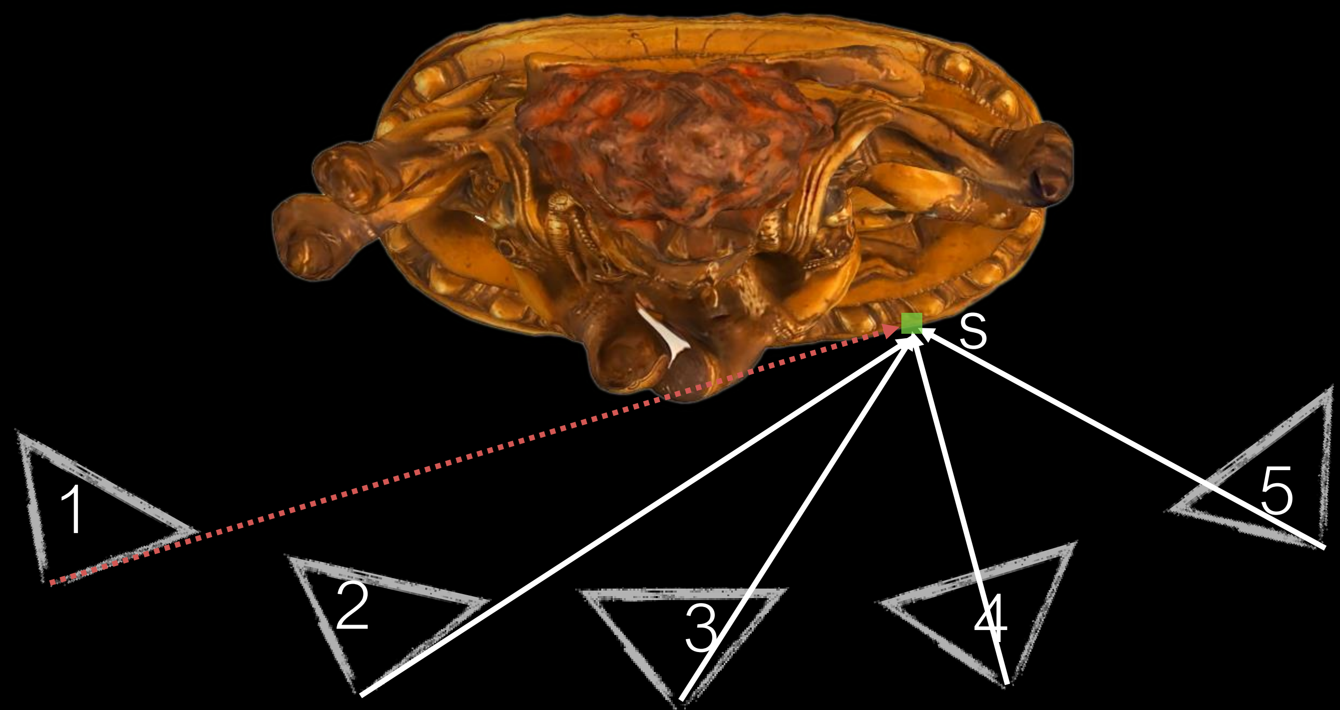
Cameras 1 and 2 can more clearly see point q.

Why MVS?

- Different points on the object’s surface will be more clearly visible in some subset of cameras
  - Could have high res close-ups of some regions
  - Some surfaces are foreshortened from certain views
- Some points may be occluded entirely in certain views



Camera 5 can’t see point r.

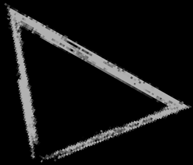


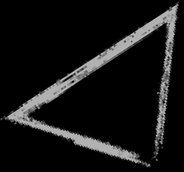
Camera 1 can't see point s.

Why MVS?

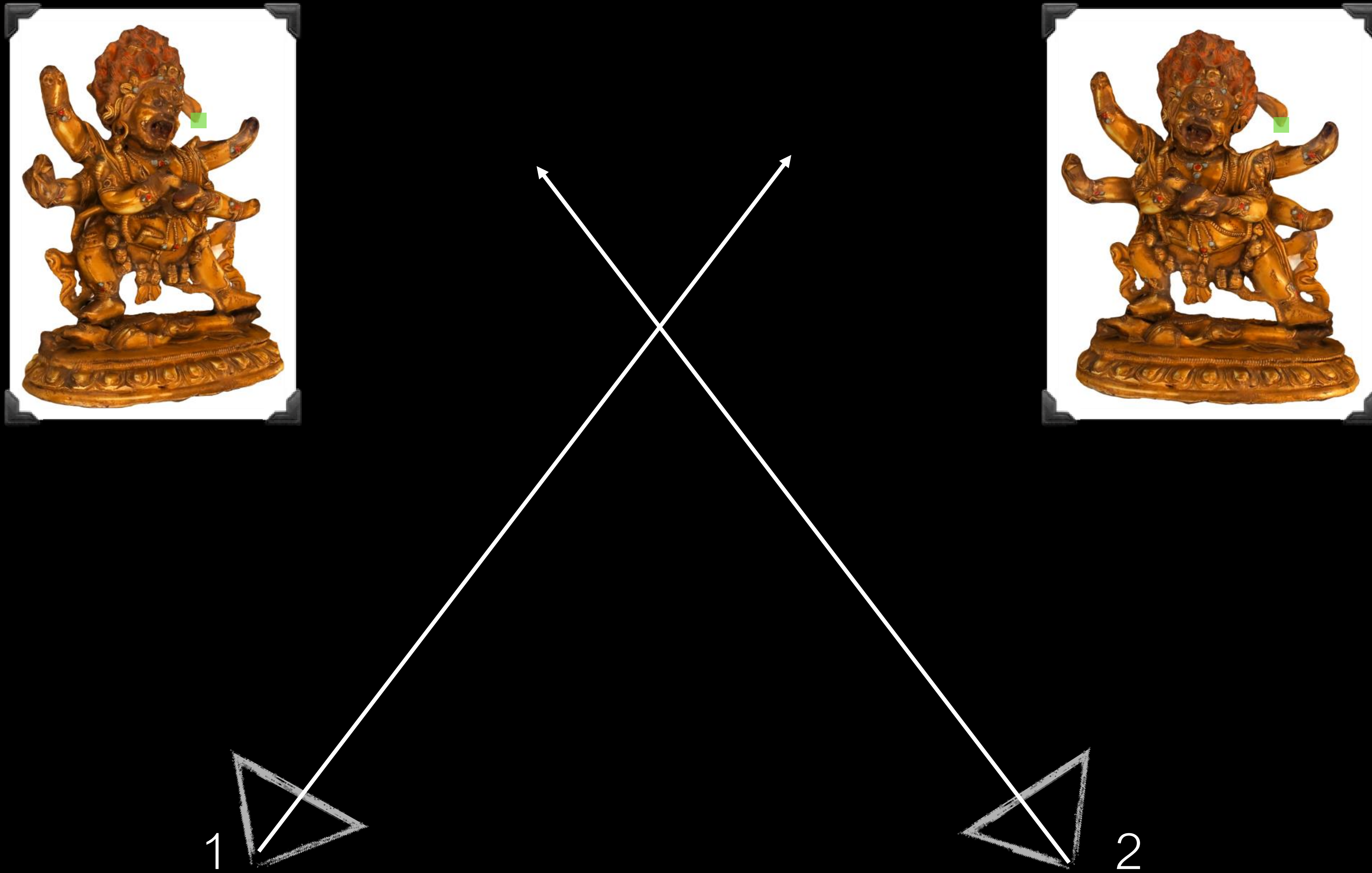
- Different points on the object’s surface will be more clearly visible in some subset of cameras
  - Could have high res closeups of some regions
  - Some surfaces are foreshortened from certain views
- Some points may be occluded entirely in certain views
- More measurements per point can reduce error

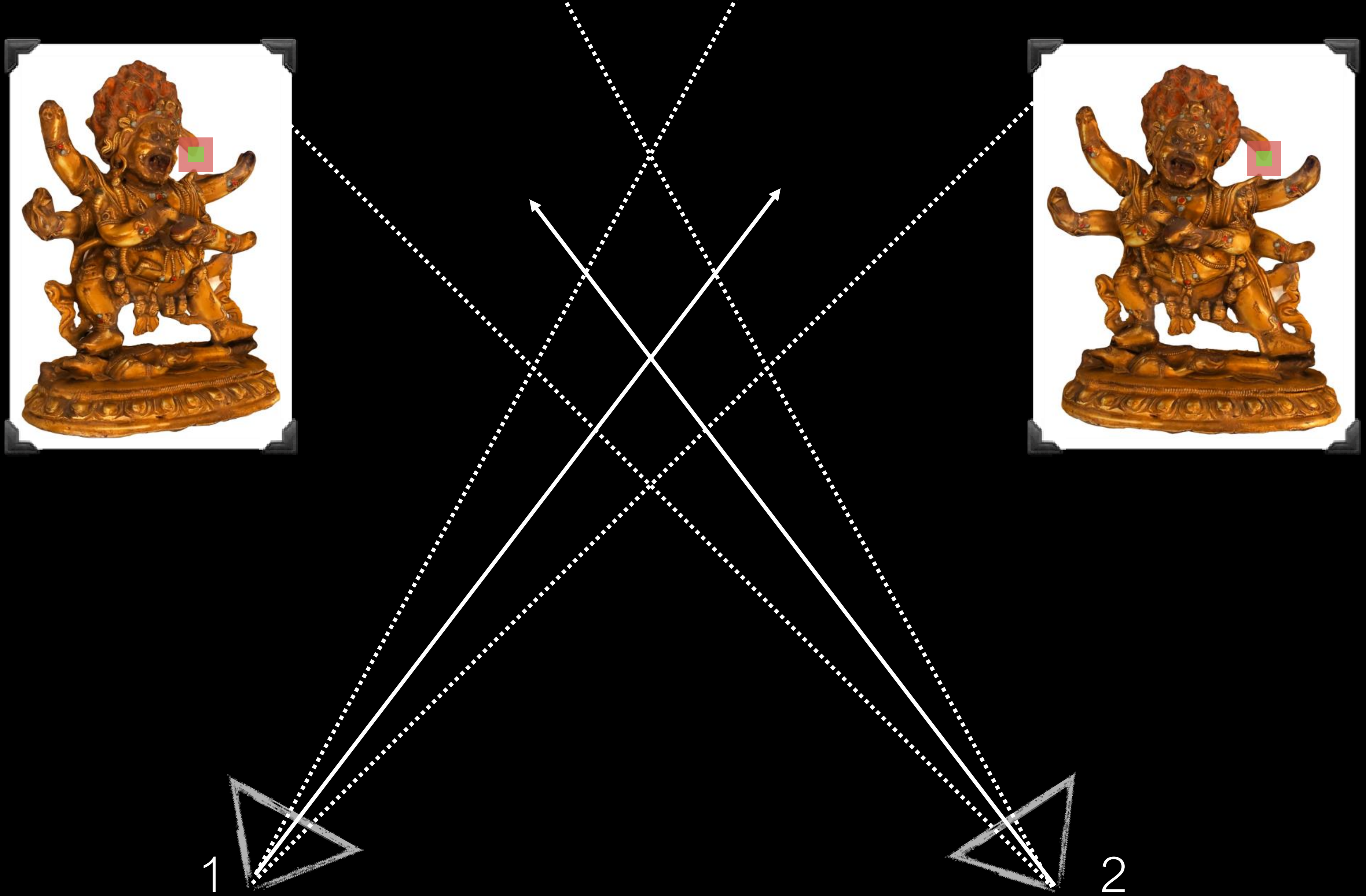
1



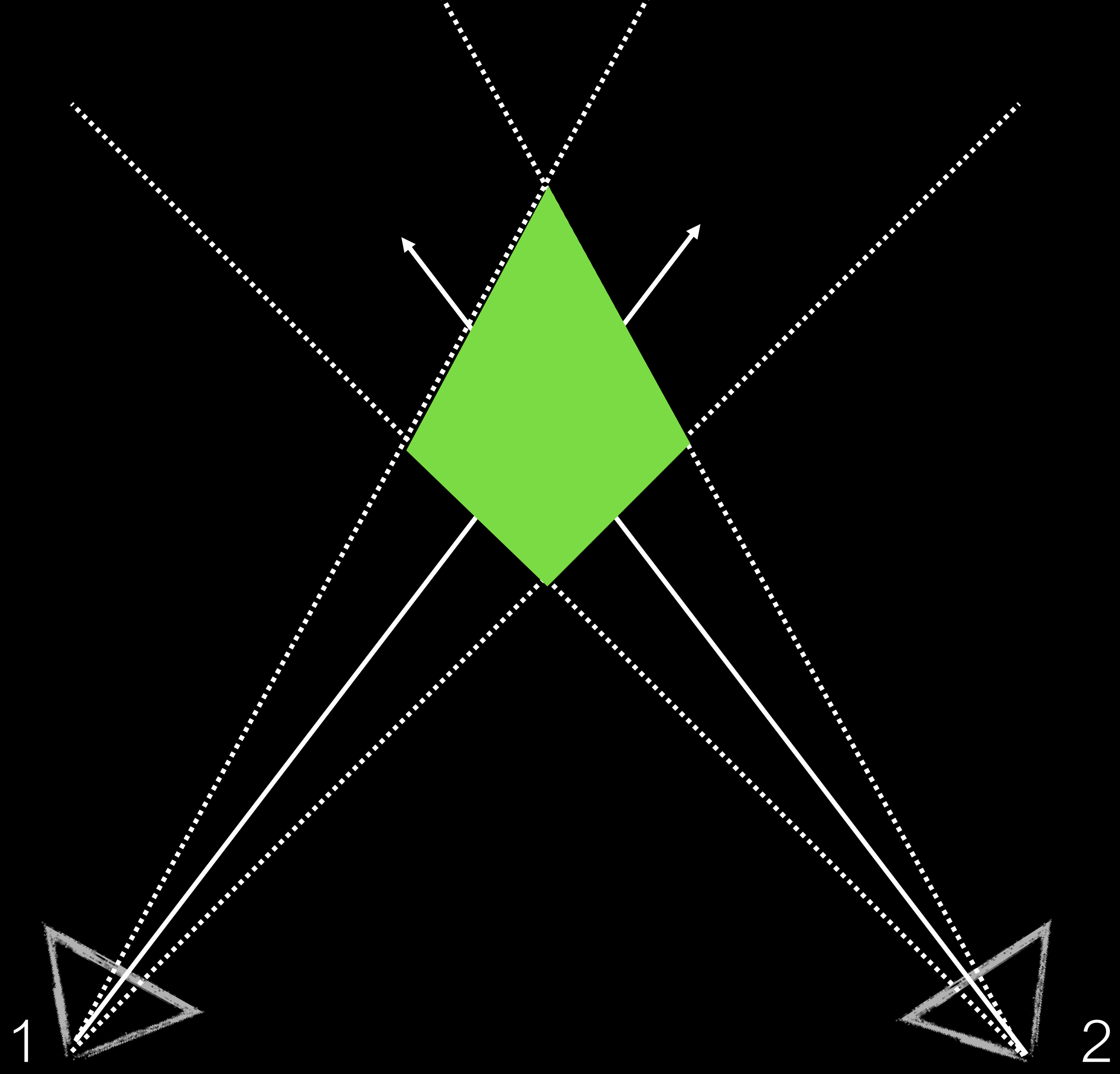


2

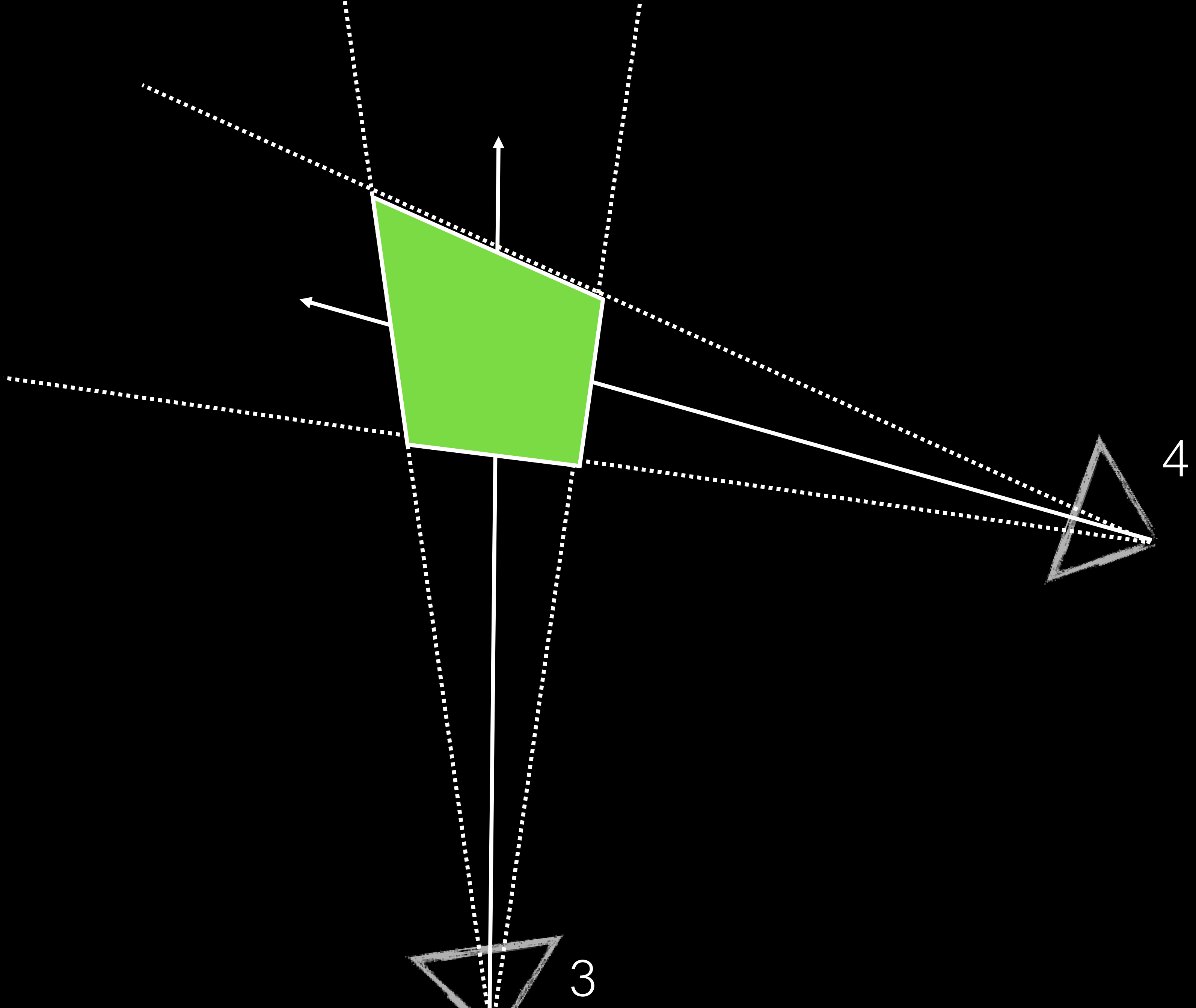

1
2



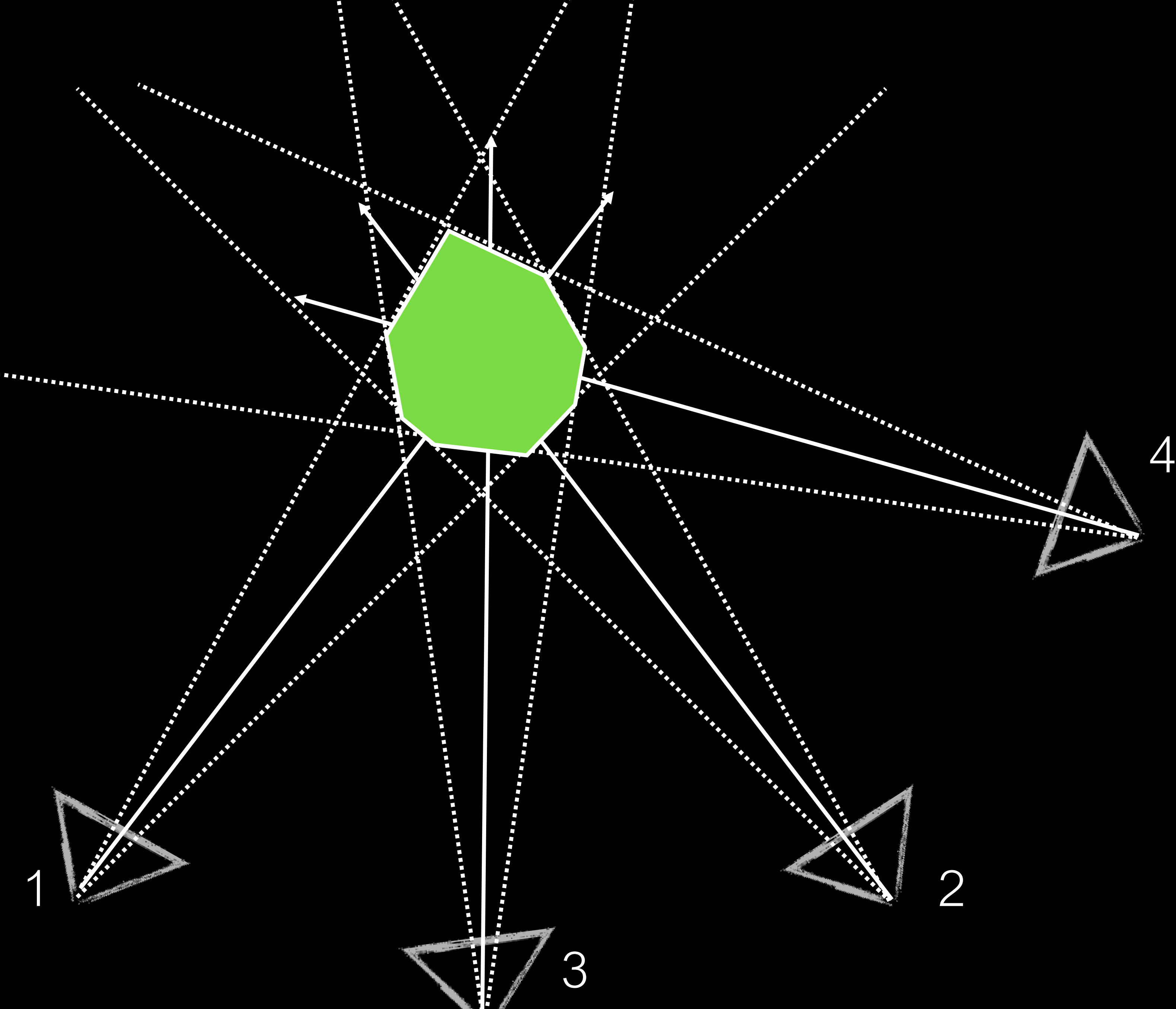
Estimated points contain some error.



Estimated points contain some error.



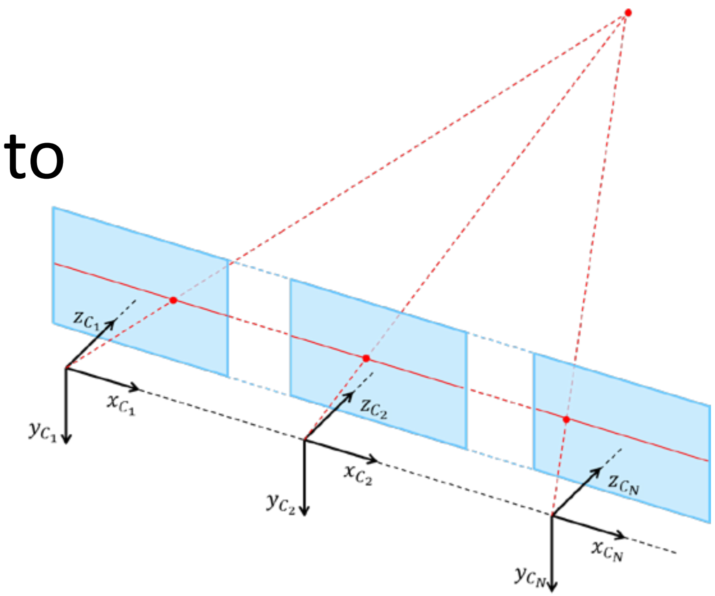
Estimated points contain some error.



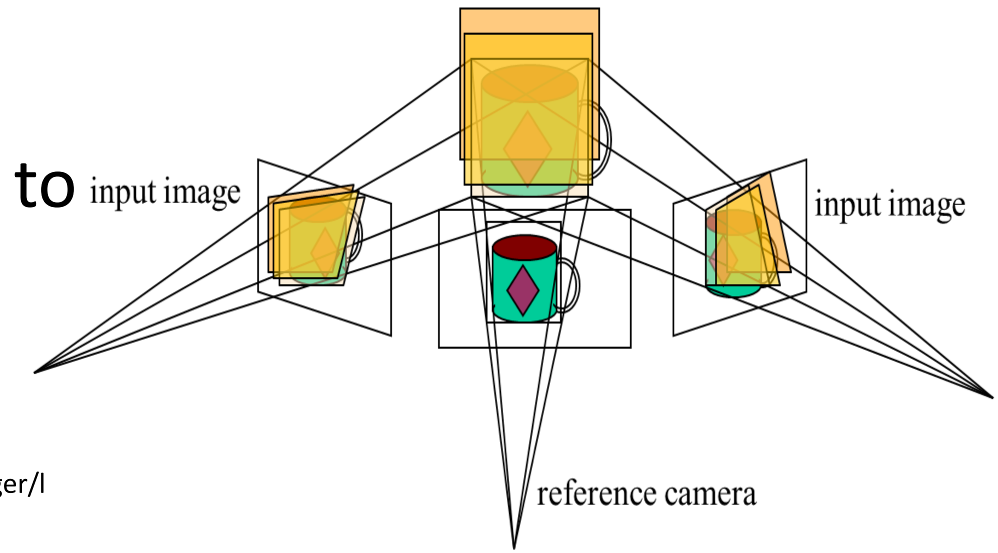
Additional views reduce error.

Depth maps reconstruction

- **Multiple-baseline stereo**
  – Rectification of several cameras onto common plane
  – Problems with wide baselines and distortions after rectification

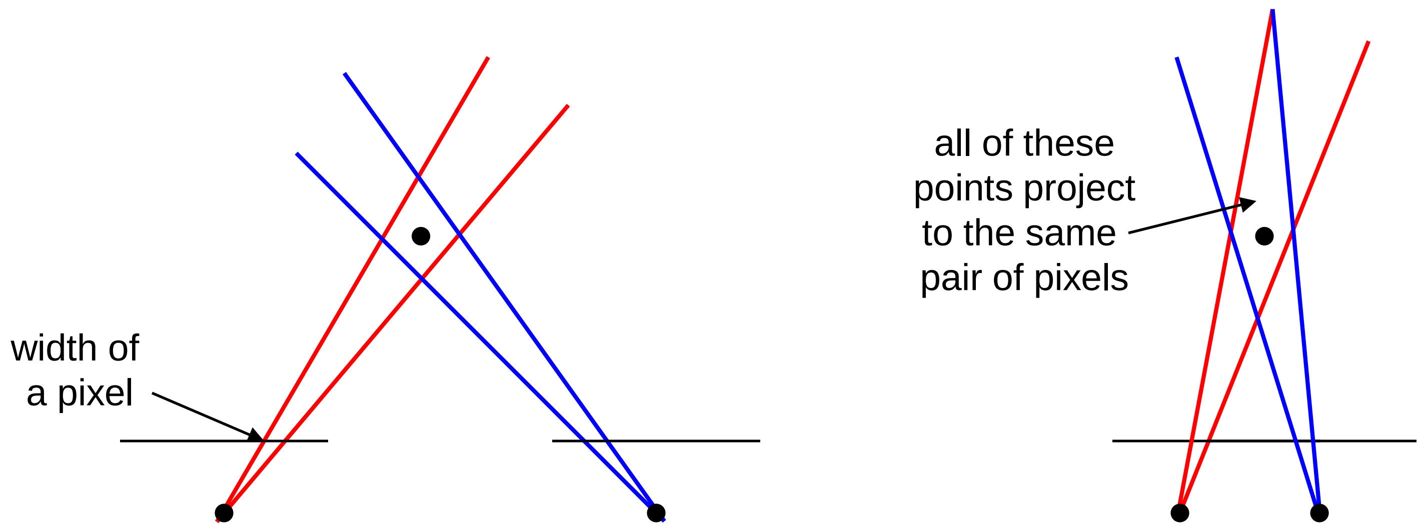


- **Plane sweep stereo**
  – Choose a reference view
  – Sweep family of planes at different depths with respect to the reference camera



http://www.uio.no/studier/emner/matnat/its/UNIK4690/v16/forelesninger/lecture_8_3_multiple_view_stereo.pdf

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

- Pick a reference image, and slide the corresponding window along the corresponding epipolar lines of all other images, using inverse depth relative to the first image as the search parameter

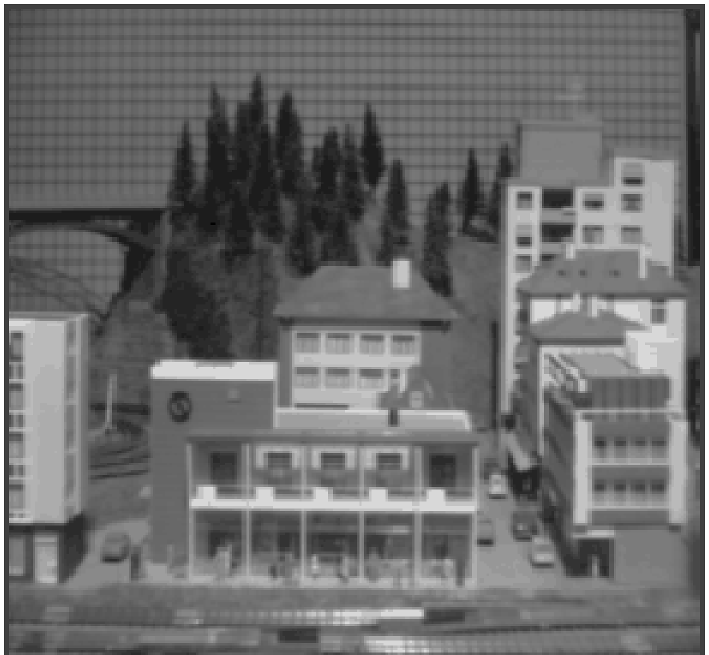
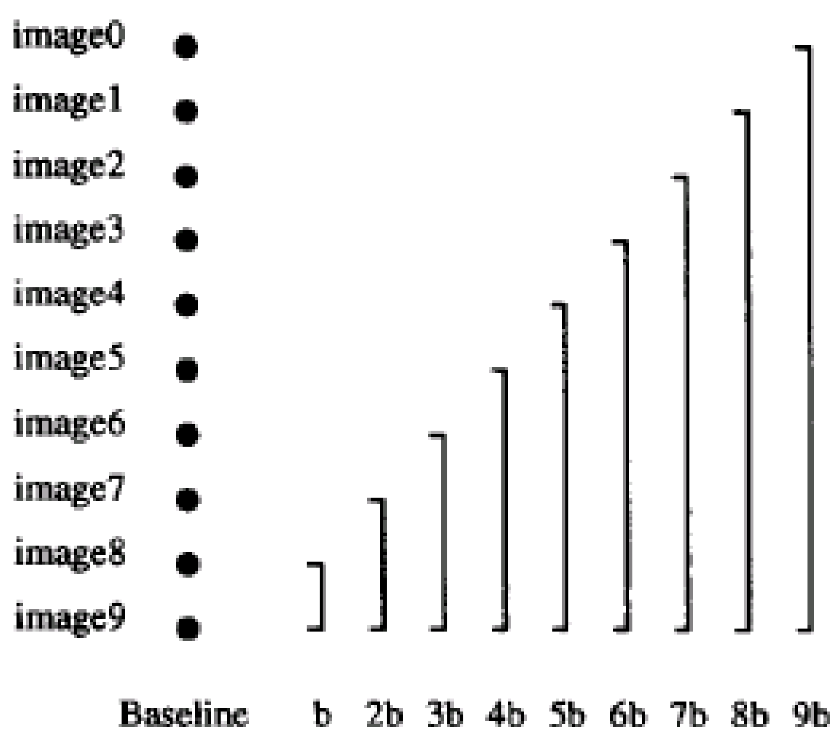
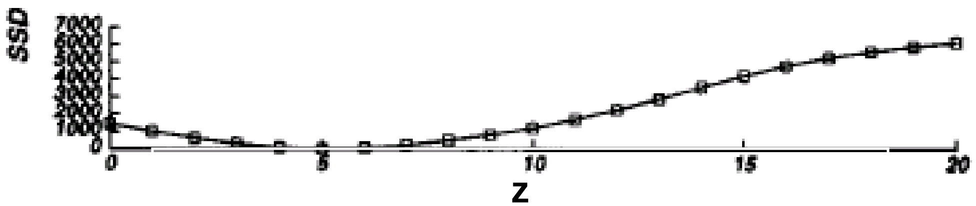


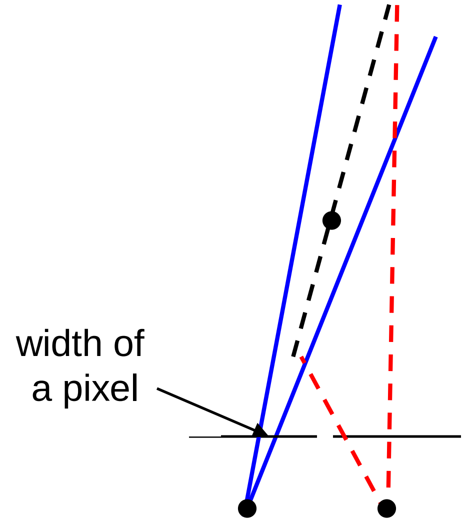
Figure 2: An example scene. The grid pattern in the background has ambiguity of matching.



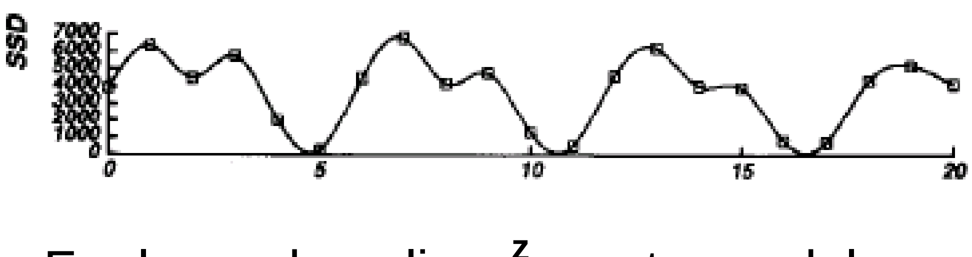
Multiple-baseline stereo

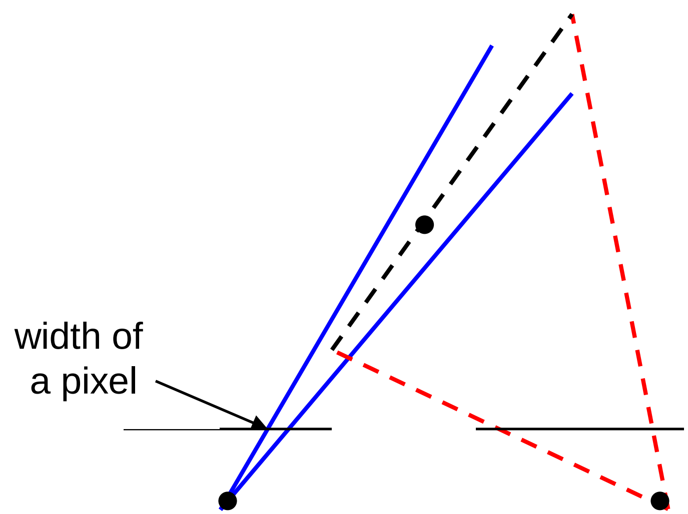


pixel matching score



- 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

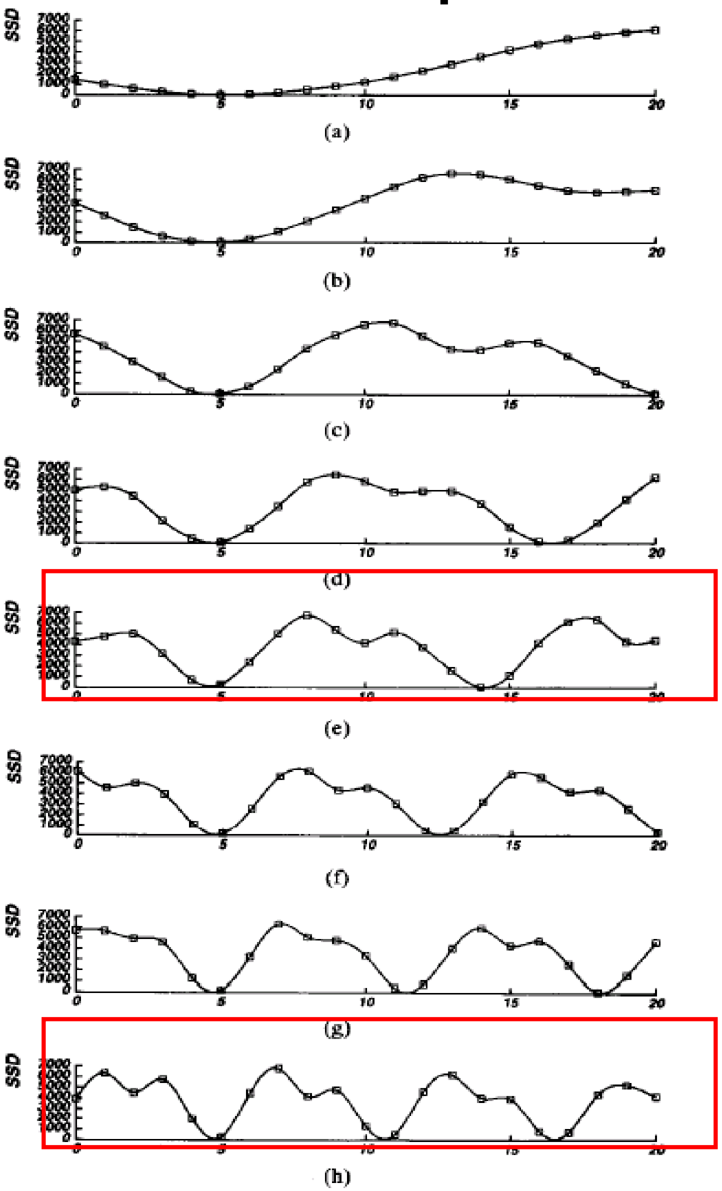


Fig. 5. SSD values versus inverse distance: (a) $B = b$; (b) $B = 2b$; (c) $B = 3b$; (d) $B = 4b$; (e) $B = 5b$; (f) $B = 6b$; (g) $B = 7b$; (h) $B = 8b$. The horizontal axis is normalized such that $8bF = 1$.

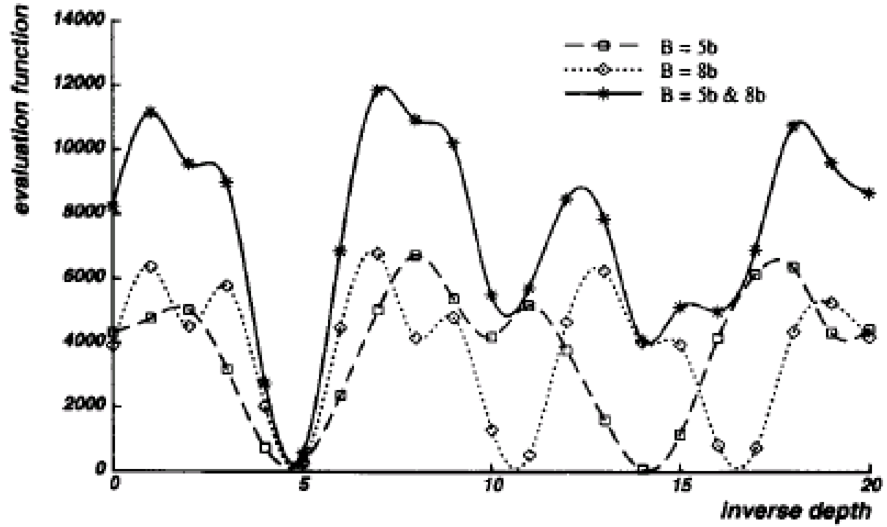


Fig. 6. Combining two stereo pairs with different baselines.

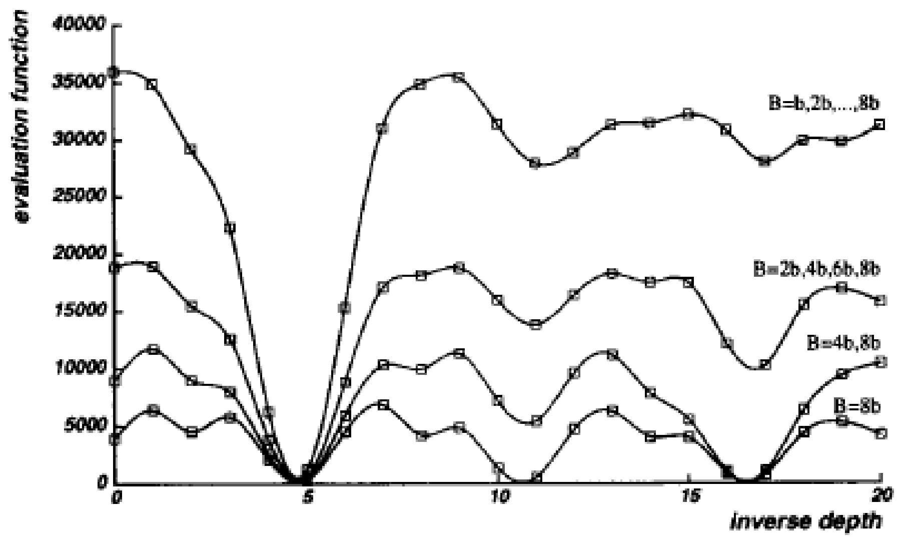
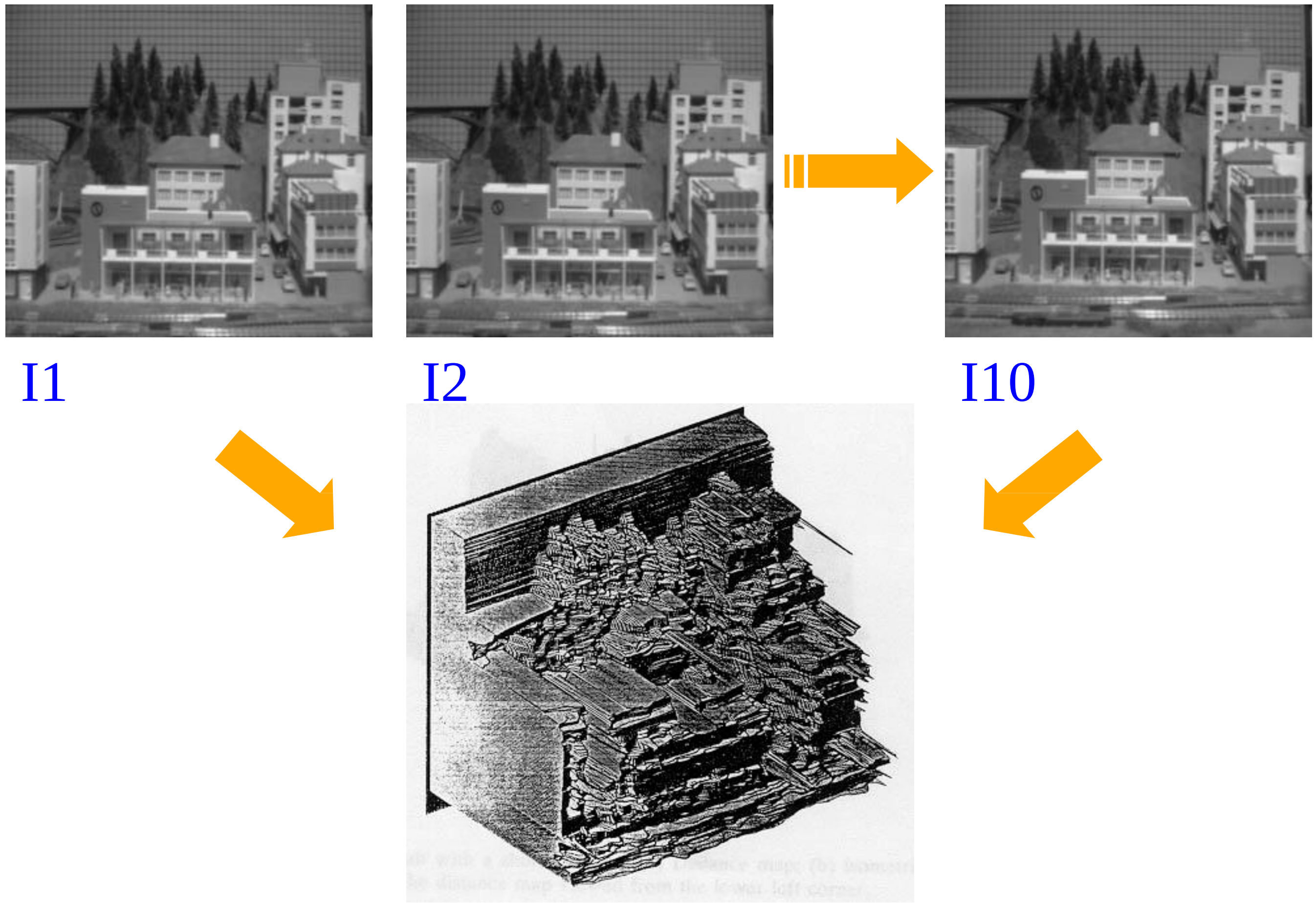


Fig. 7. Combining multiple baseline stereo pairs.

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).

Multibaseline Stereo

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.

Problem: *visibility*

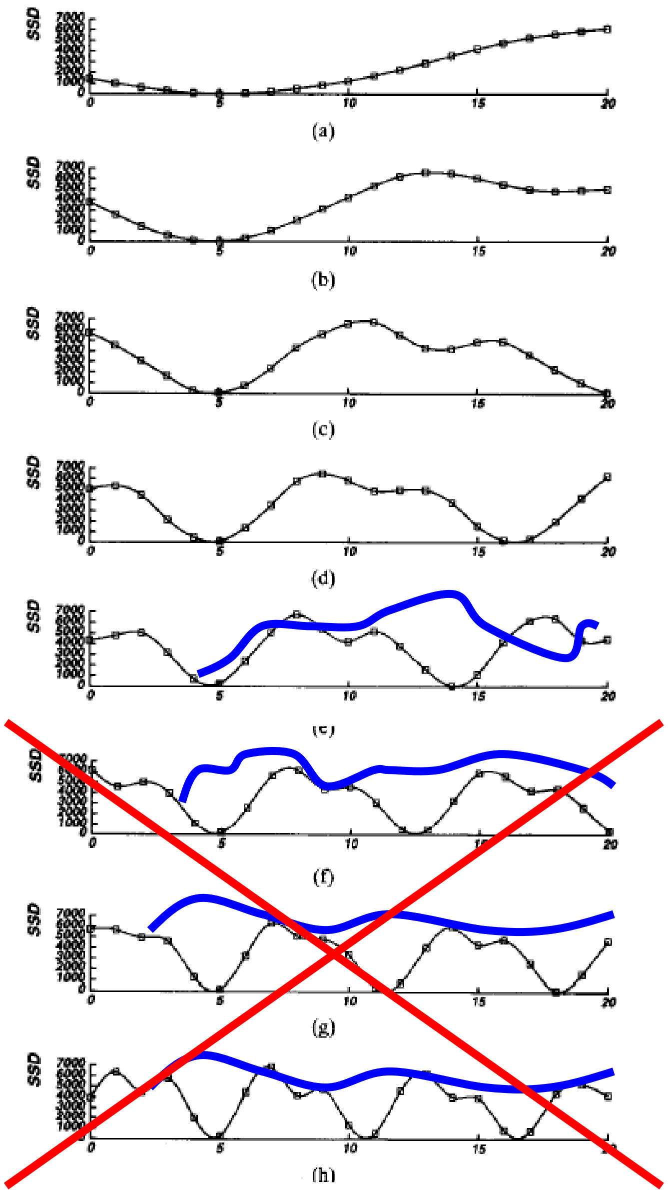


Fig. 5. SSD values versus inverse distance: (a) $B = b$; (b) $B = 2b$; (c) $B = 3b$; (d) $B = 4b$; (e) $B = 5b$; (f) $B = 6b$; (g) $B = 7b$; (h) $B = 8b$. The horizontal axis is normalized such that $8bF = 1$.

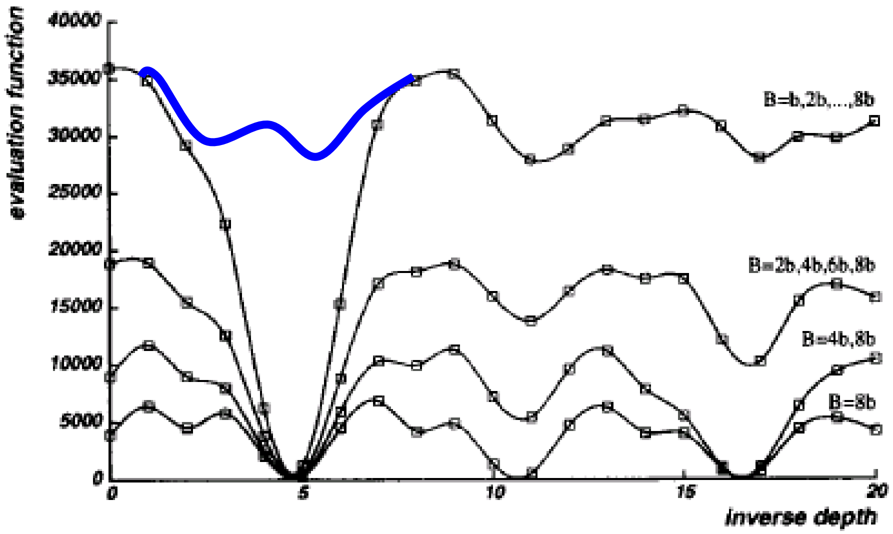


Fig. 7. Combining multiple baseline stereo pairs.

Some Solutions

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

Photo-consistency measures (matching score)

Given a set of N input images and a 3D point p seen by all the images, one can define the photo-consistency of p w.r.t. each pair of images I_i and I_j as:

$$\mathcal{C}_{ij}(p) = \rho(I_i(\Omega(\pi_i(p))), I_j(\Omega(\pi_j(p)))), \tag{2.1}$$

where $\rho(f, g)$ is a similarity measure that compares two vectors, $\pi_i(p)$ denotes the projection of p into image i, $\Omega(x)$ defines a support domain around point x, and $I_i(x)$ denotes the image intensities sampled within the domain. Every photo-consistency measure can be described as a particular choice of ρ and Ω.

http://carlos-hernandez.org/papers/fnt_mvs_2015.pdf

Popular matching scores

- SSD (Sum Squared Distance)

$$\sum_{x,y} |W_1(x,y) - W_2(x,y)|^2$$

- SAD (Sum of Absolute Difference) $\sum_{x,y} |W_1(x,y) - W_2(x,y)|$
- 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 $\overline{W_i} = \frac{1}{n}\sum_{x,y} W_i \quad \sigma_{W_i} = \sqrt{\frac{1}{n}\sum_{x,y}(W_i - \overline{W_i})^2}$

  – what advantages might NCC have?

Summary

Table 2.1: Summary table of different similarity measures used to compute photo-consistency.

Measure	required Ω	invariance
Sum of Squared Differences (SSD)	no	none
Sum of Absolute Differences (SAD)	no	none
Normalized Cross Correlation (NCC)	yes	bias/gain
Census	yes	bias/gain
Rank	yes	bias/gain/rotation
Mutual Information (MI)	yes	any bijection

http://carlos-hernandez.org/papers/fnt_mvs_2015.pdf

Plane sweep stereo

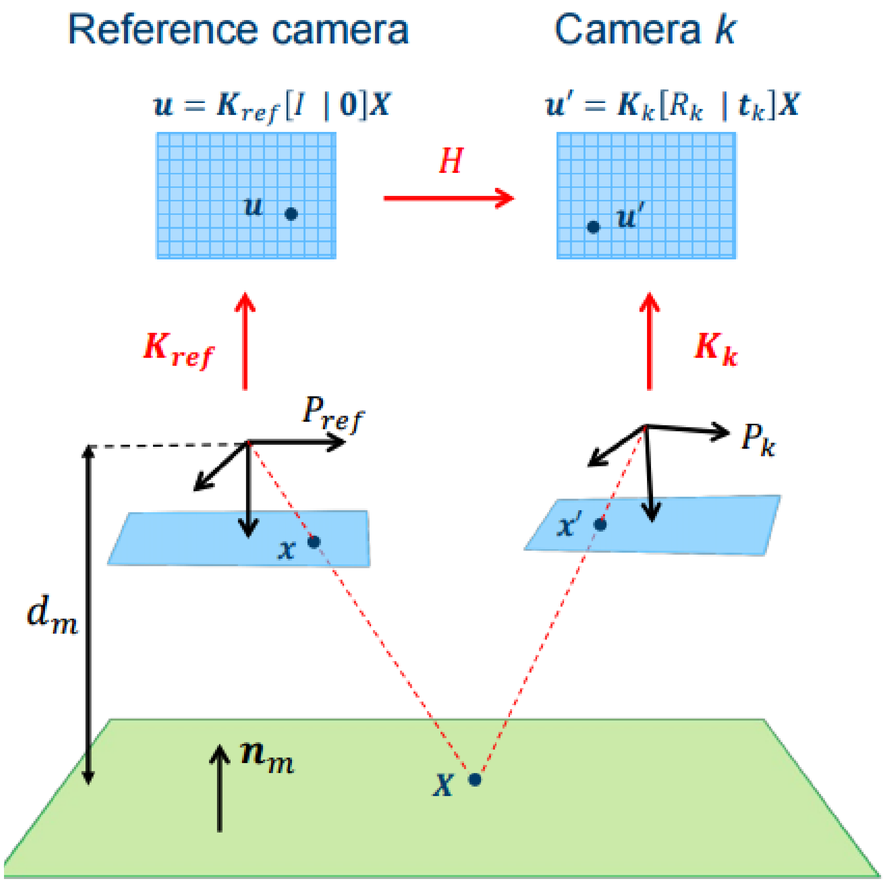
- The family of depth planes
in the coordinate frame of the reference view

$$\Pi_m = \begin{bmatrix} \boldsymbol{n}_m^T & -d_m \end{bmatrix}$$

- The mapping from the reference camera P_{ref} onto the plane Π_m and back to camera P_k is described by the homography induced by the plane Π_m

$$H_{\Pi_m, P_k} = K_k \left(R_k - \boldsymbol{t}_k \boldsymbol{n}_m^T / d_m \right) K_{ref}^{-1}$$

- The mapping from P_k to P_{ref} induced by Π_m is the inverse homography H_{Π_m, P_k}^{-1}



D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang and M. Pollefeys, Real-Time Plane-Sweeping Stereo with Multiple Sweeping Directions, CVPR 2007

Plane sweep stereo

1. Map each target image I_k to the reference image I_{ref} for each depth plane Π_m with the homography $H^{-1}_{\Pi_m, P_k}$ giving the warped images $\breve{I}_{k,m}$

2. Compute the similarity between I_{ref} and each $\breve{I}_{k,m}$
   – Zero Mean Normalized Cross Correlation (ZNCC) between small patches *W* around each pixel

3. Compute the figure-of-merit for each depth plane by combining the similarity measurements for each image *k*

$$M(u,v,\Pi_m) = \sum_k \underset{W}{ZNCC}(I_{ref}, \breve{I}_{k,m})$$

4. For each pixel, select the depth plane with best fit

$$\tilde{\Pi}(u,v) = \arg\max_m M(u,v,\Pi_m)$$

Plane sweep through oriented planes

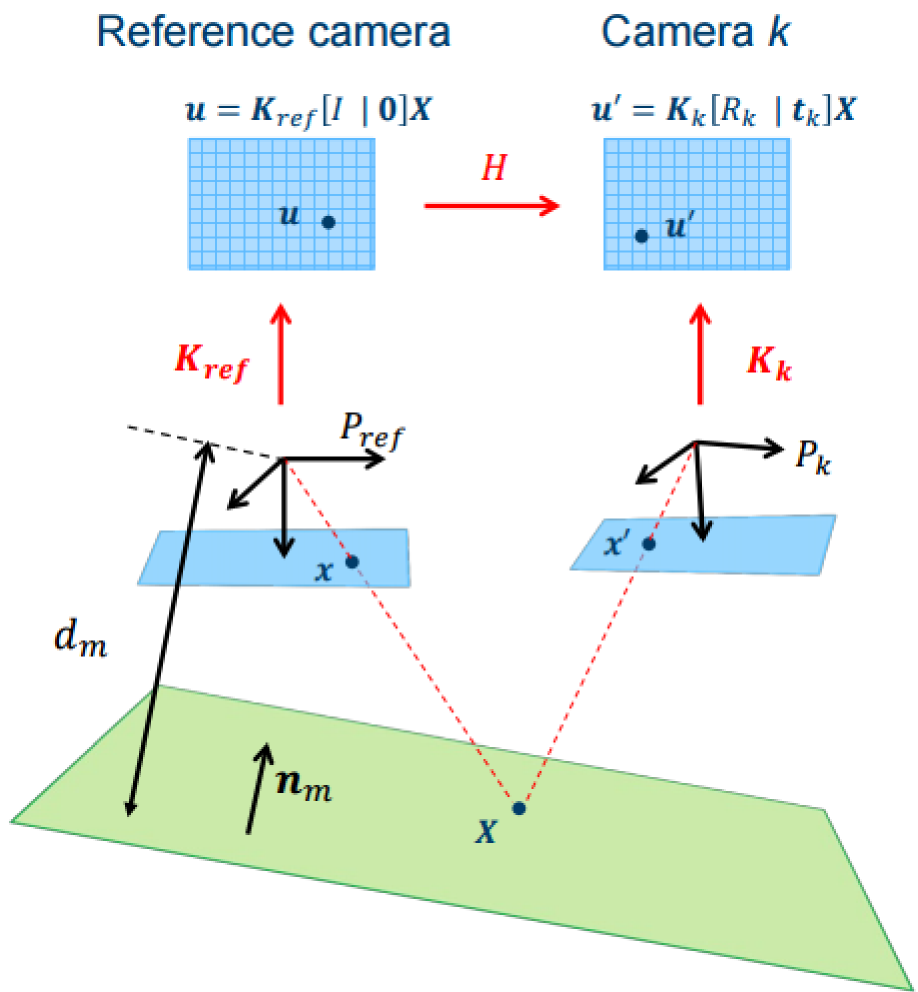
- Fronto-parallel

$$\boldsymbol{n}_m = \begin{bmatrix} 0 & 0 & -1 \end{bmatrix}^T$$

$$Z_m(u,v) = d_m$$

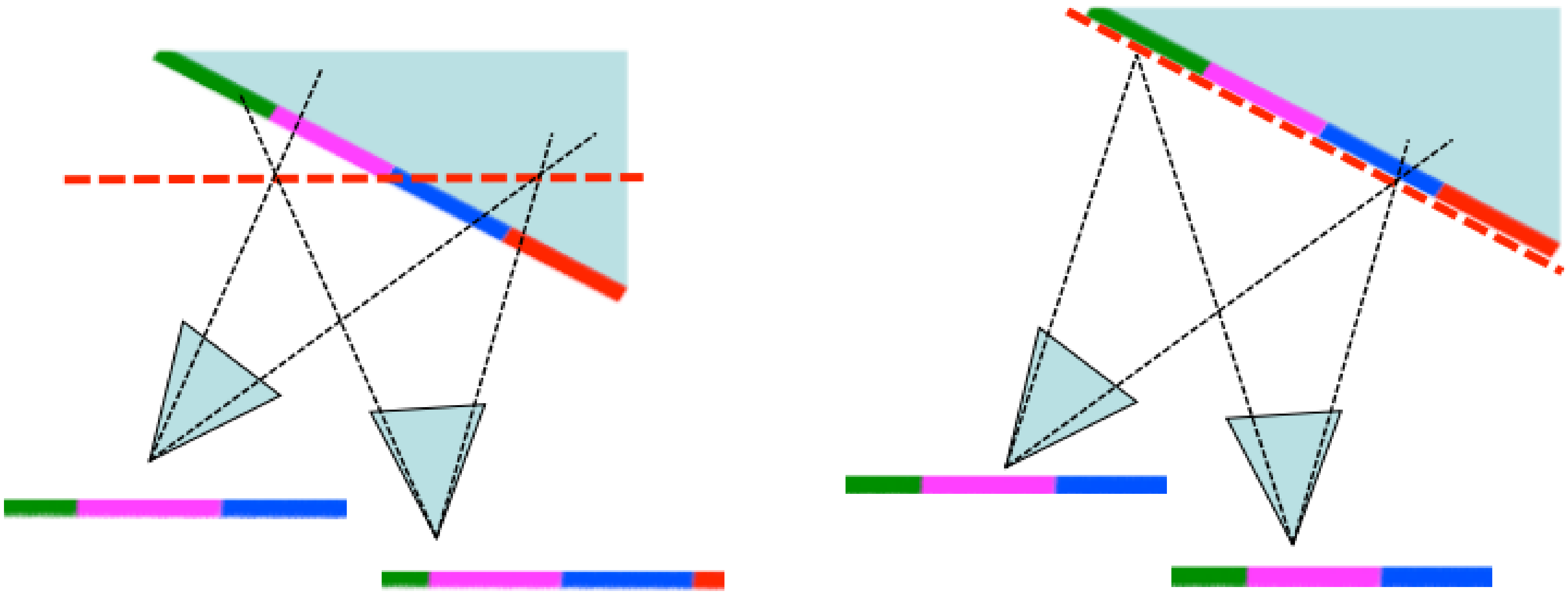
- Other plane orientations

$$Z_m(u,v) = \frac{-d_m}{\begin{bmatrix} u & v & 1 \end{bmatrix} K_{ref}^{-T} \boldsymbol{n}_m}$$



Plane Sweep: Enhanced Robustness through oriented planes

- Aligning sweeping direction to surface orientation reduces photometric inconsistencies at correct depth

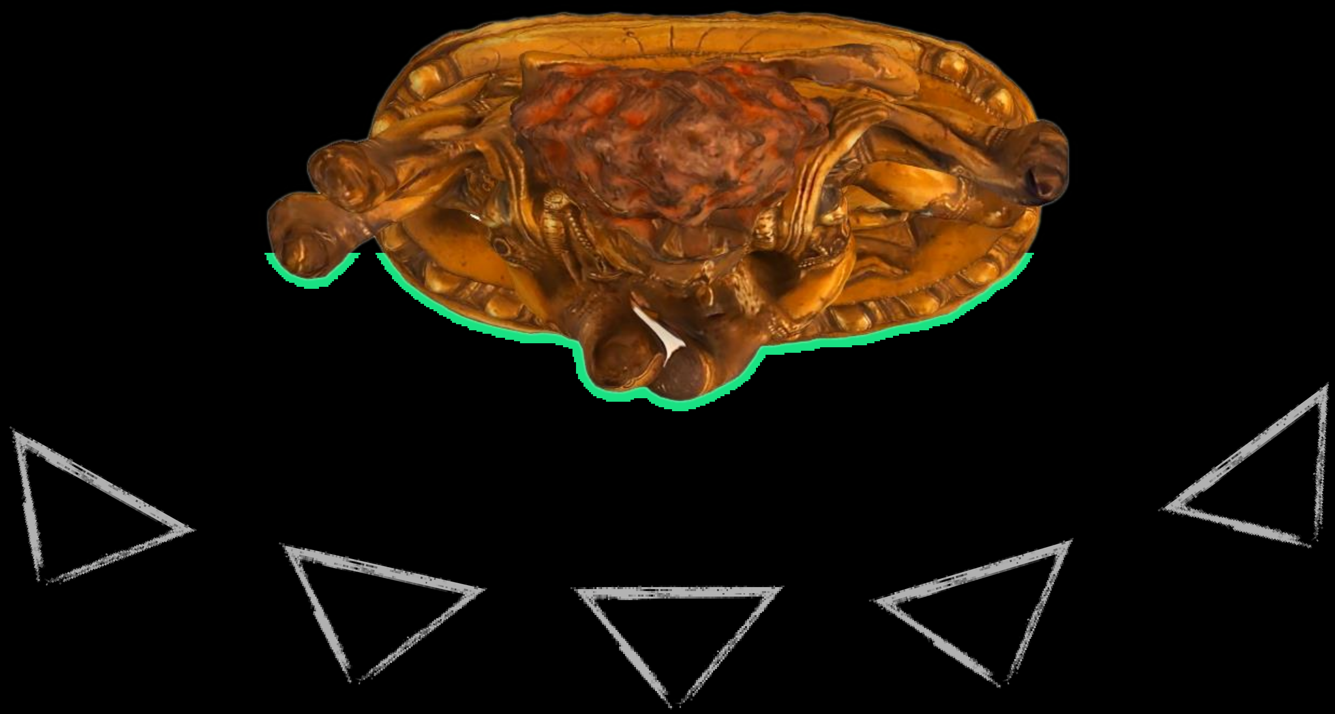


https://demuc.de/tutorials/cvpr2017/dense-modeling.pdf

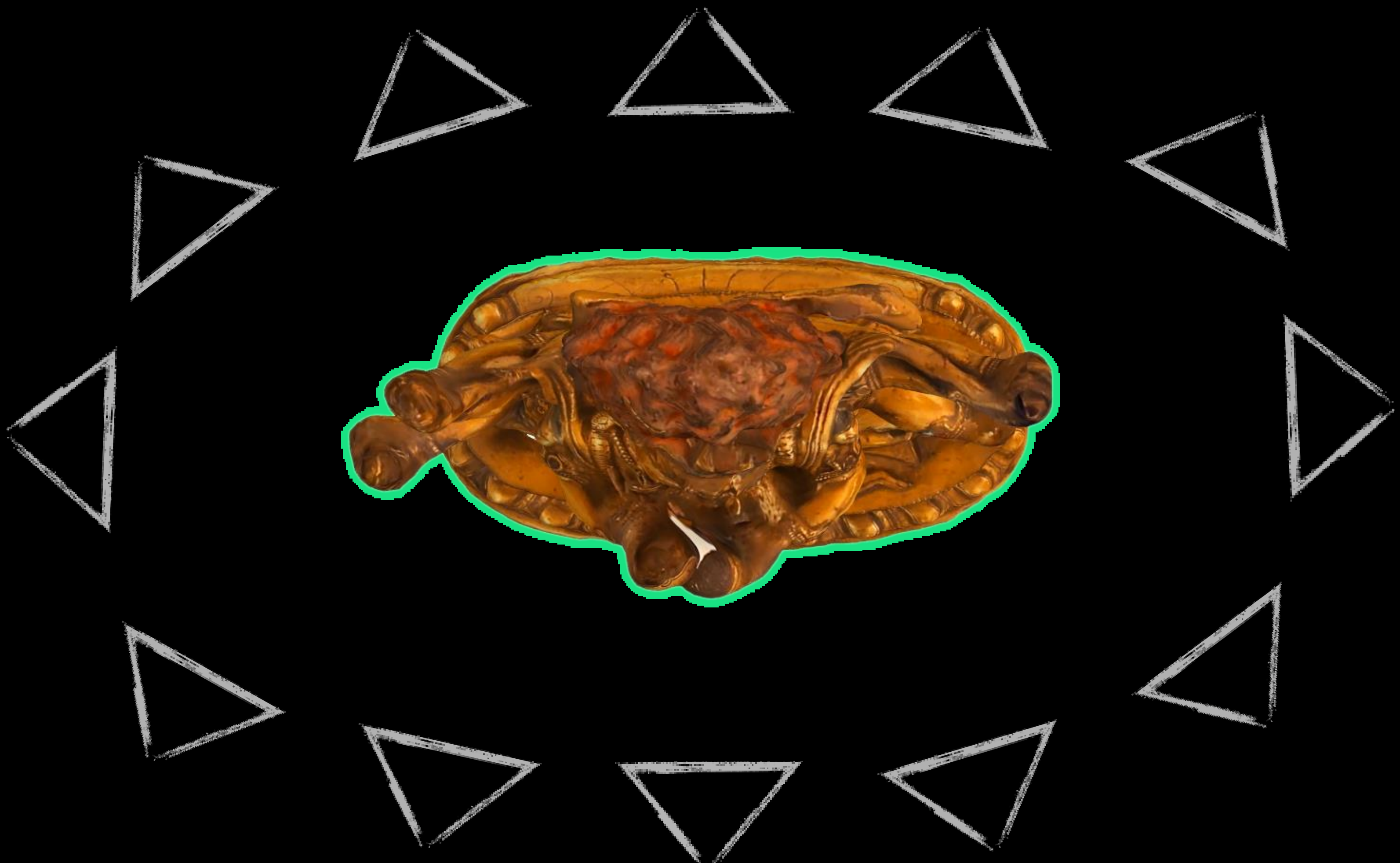
Streetside reconstructions by plane sweeping stereo

http://carlos-hernandez.org/papers/fnt_mvs_2015.pdf

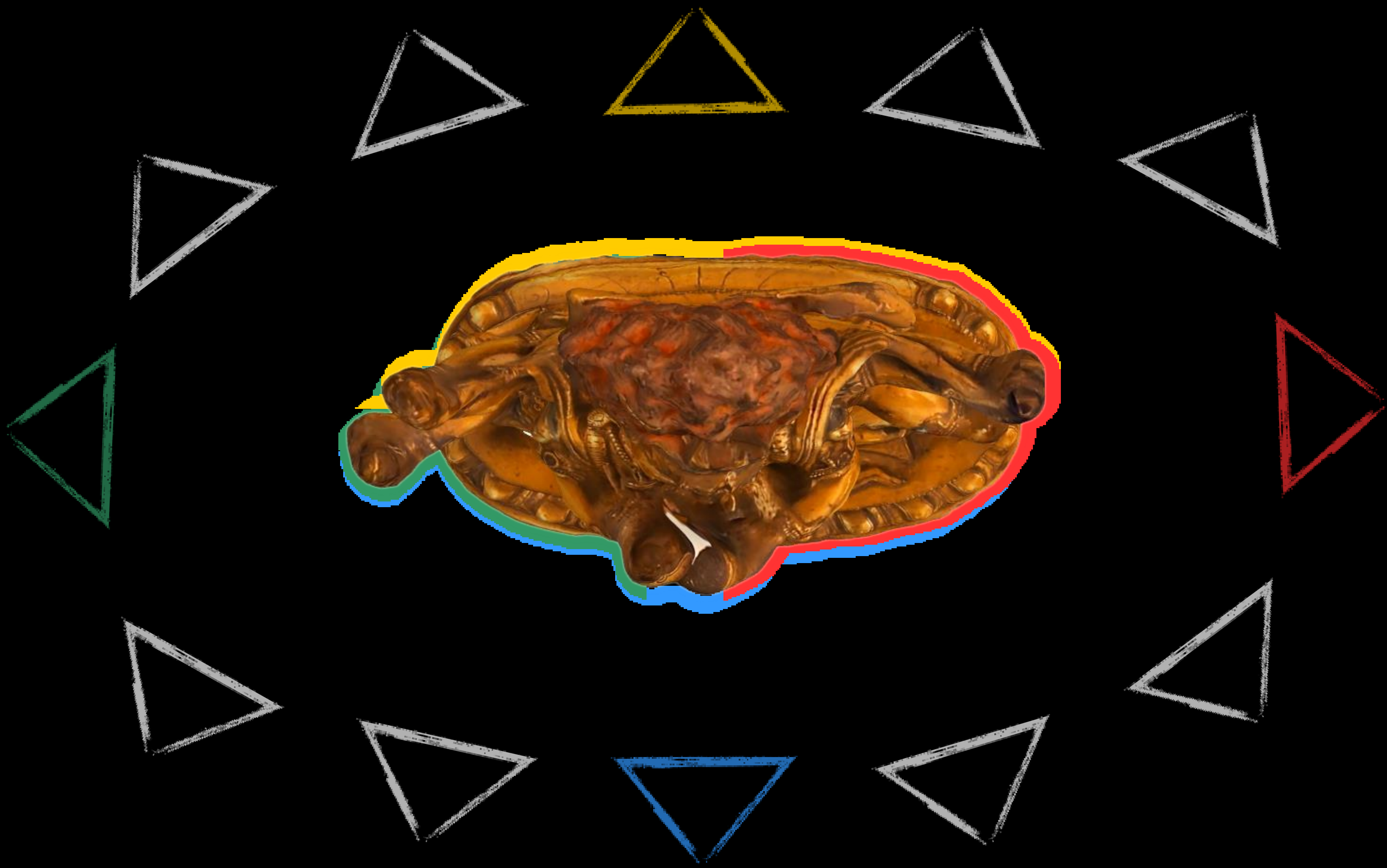
Single depth map often isn't enough



Really want full coverage



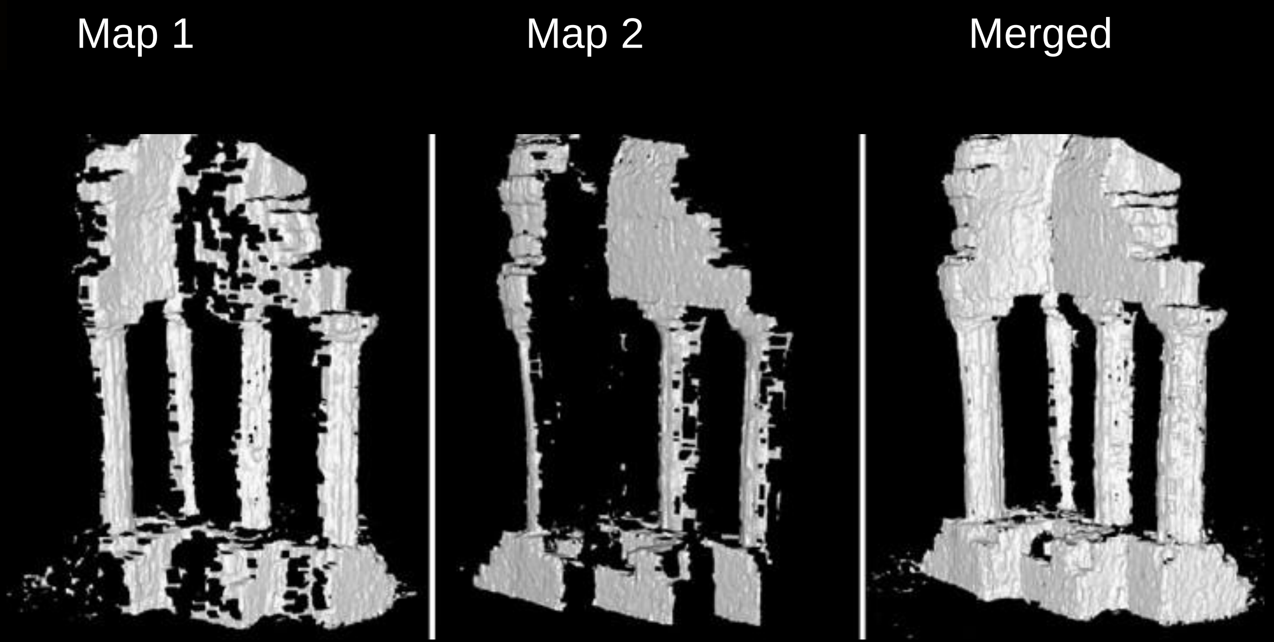
Idea: Combine many depth maps



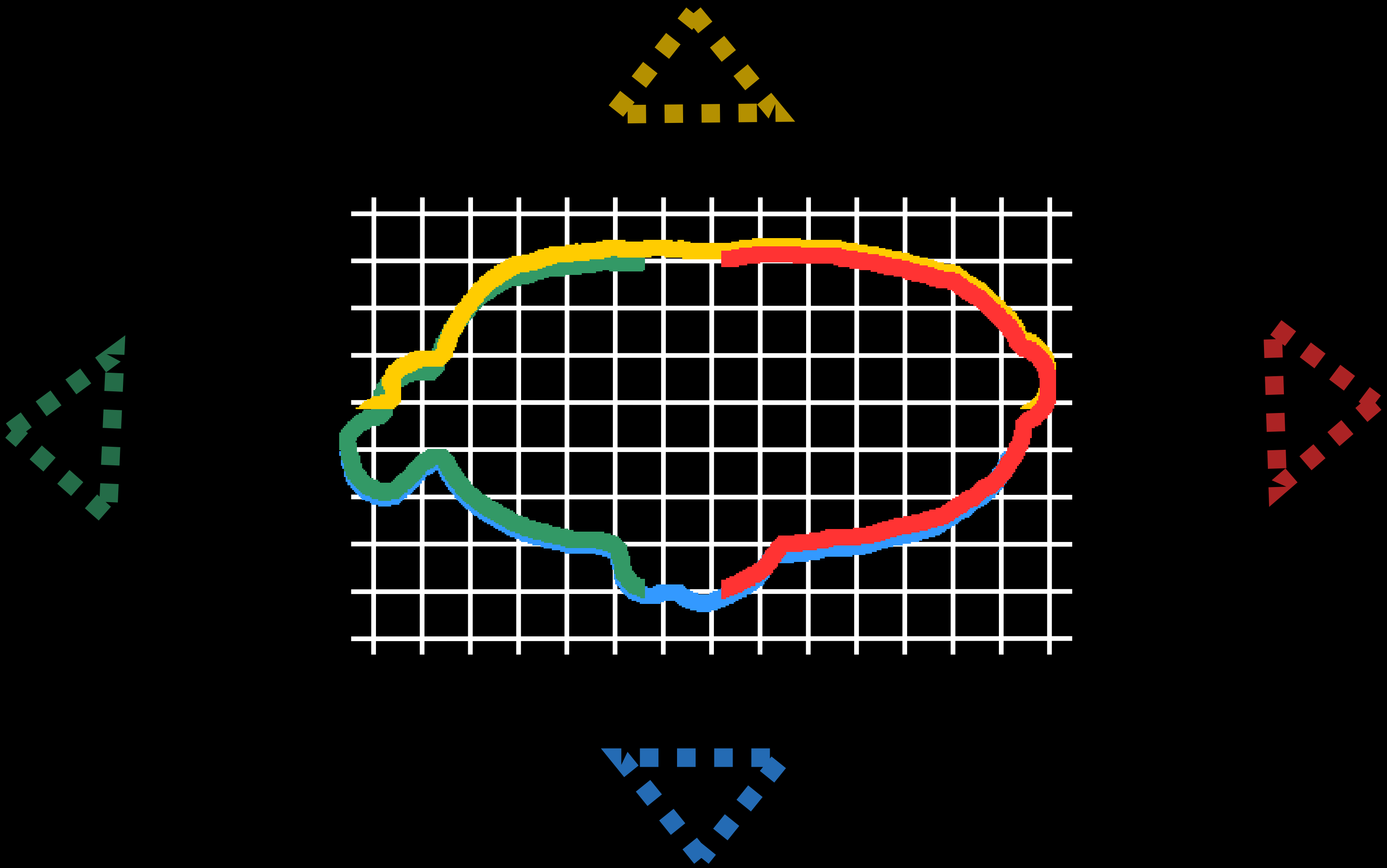
Many depth maps, each with error. How can we fuse these?

Merging depth maps

- Given a group of images, choose each one as reference and compute a depth map w.r.t. that view using a multi-baseline approach
- Merge multiple depth maps to a volume or a mesh (see, e.g., Curless and Levoy 96)

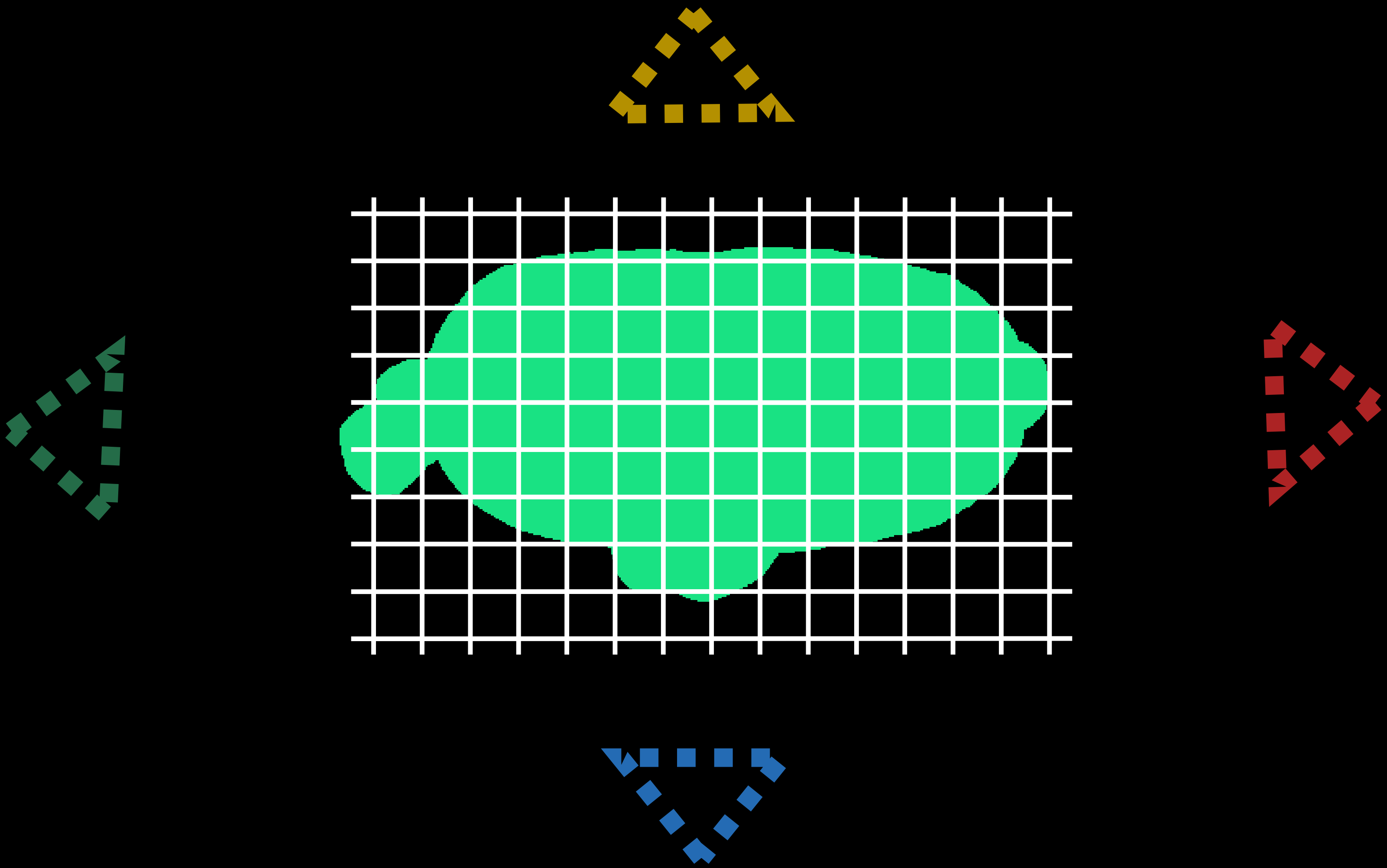


Volumetric fusion



A common world-space coordinate system.

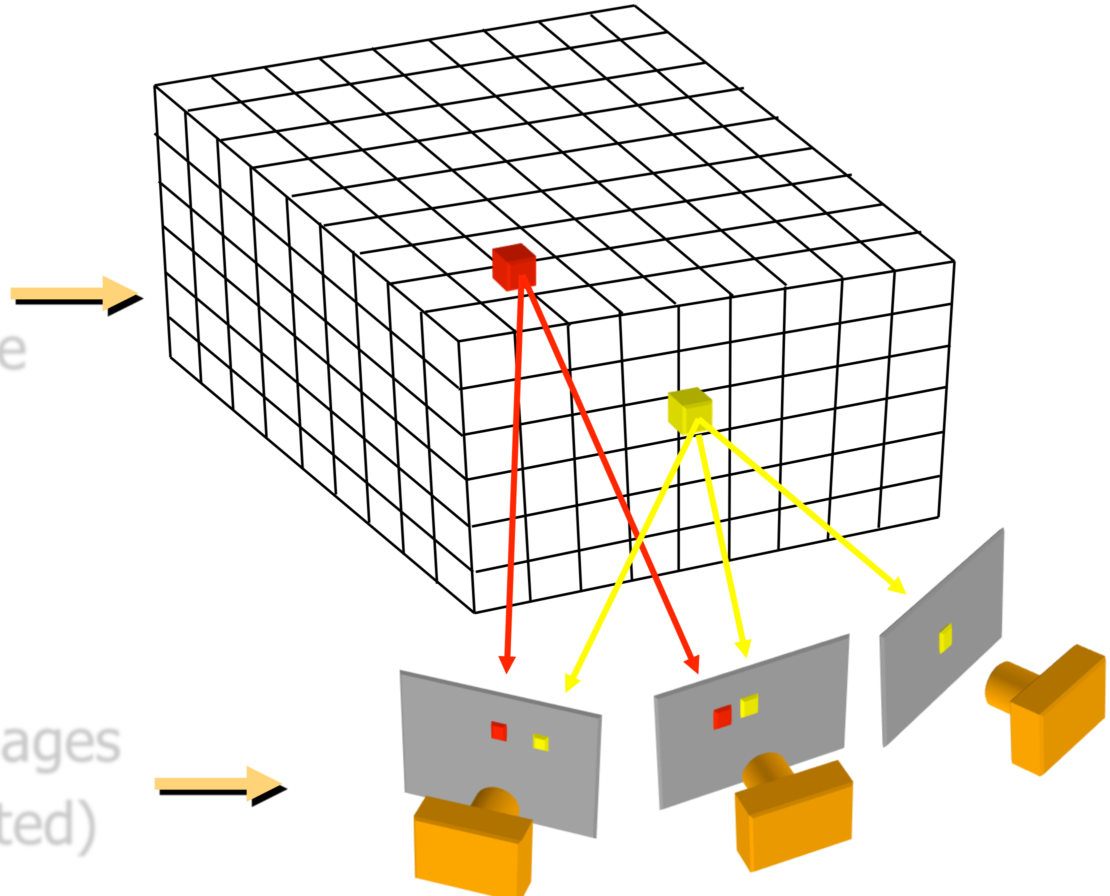
Volumetric fusion



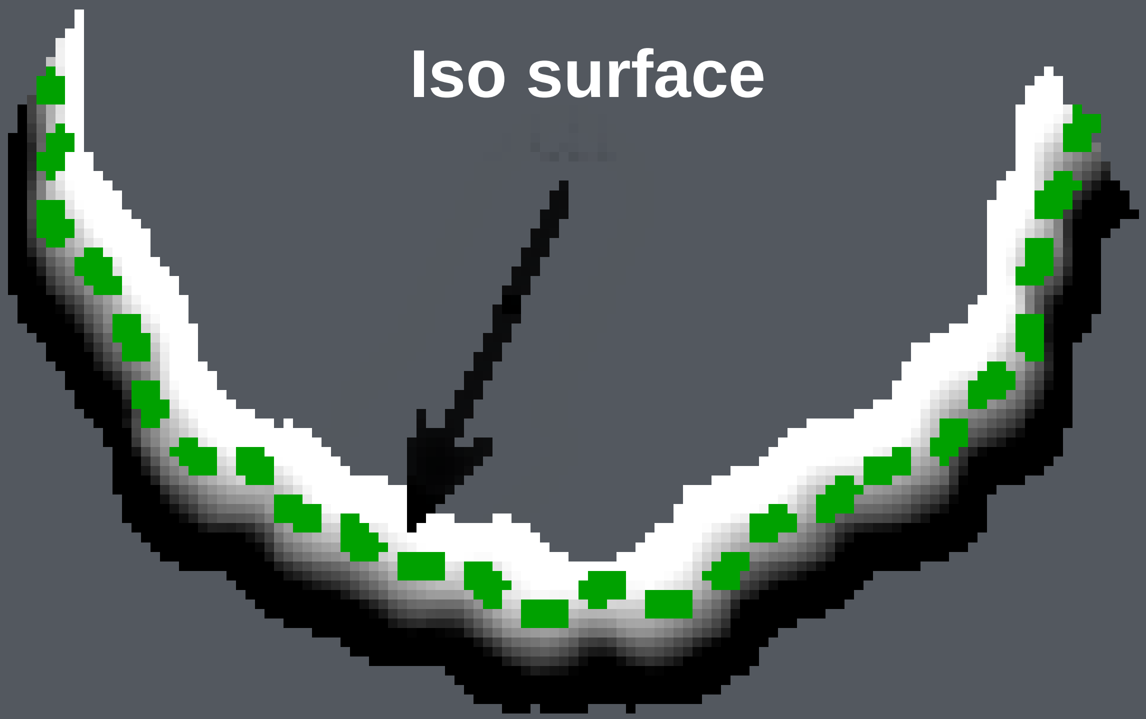
A common world-space coordinate system.

Volumetric stereo

- In plane sweep stereo, the sampling of the scene depends on the reference view
- We can use a voxel volume to get a view independent representation



GOAL: Assign RGB values to voxels in V
***photo-consistent* with images**


Iso surface

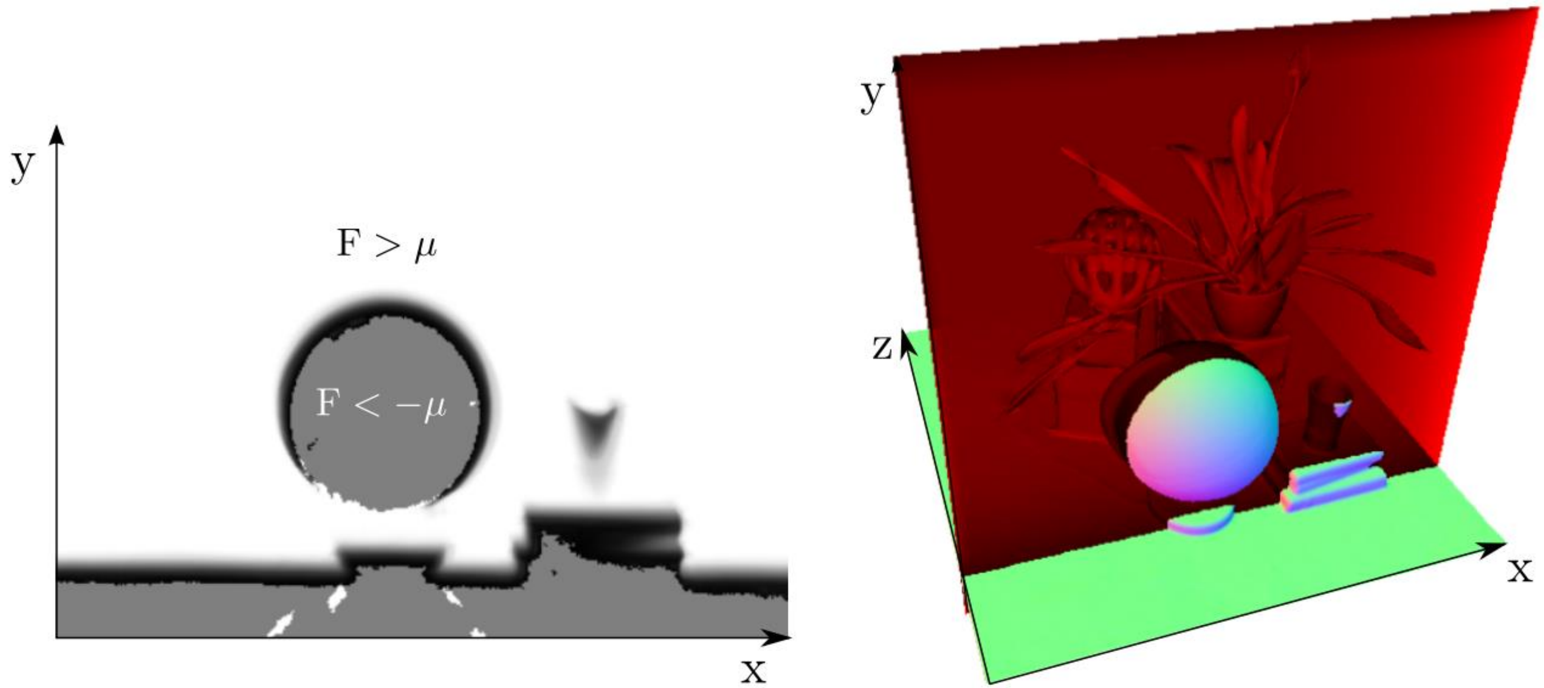


Figure 4: A slice through the truncated signed distance volume showing the truncated function $F > \mu$ (white), the smooth distance field around the surface interface $F = 0$ and voxels that have not yet had a valid measurement(grey) as detailed in eqn. 9.

Newcombe, et. al. ISMAR 2011.

KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera*

Shahram Izadi[1], *David Kim*[1,3], *Otmar Hilliges*[1], *David Molyneaux*[1,4], *Richard Newcombe*[2],
Pushmeet Kohli[1], *Jamie Shotton*[1], *Steve Hodges*[1], *Dustin Freeman*[1,5],
Andrew Davison[2], *Andrew Fitzgibbon*[1]

[1]Microsoft Research Cambridge, UK [2]Imperial College London, UK
[3]Newcastle University, UK [4]Lancaster University, UK [5]University of Toronto, Canada

Figure 1: KinectFusion enables real-time detailed 3D reconstructions of indoor scenes using only the depth data from a standard Kinect camera. A) user points Kinect at coffee table scene. B) Phong shaded reconstructed 3D model (the wireframe frustum shows current tracked 3D pose of Kinect). C) 3D model texture mapped using Kinect RGB data with real-time particles simulated on the 3D model as reconstruction occurs. D) Multi-touch interactions performed on any reconstructed surface. E) Real-time segmentation and 3D tracking of a physical object.

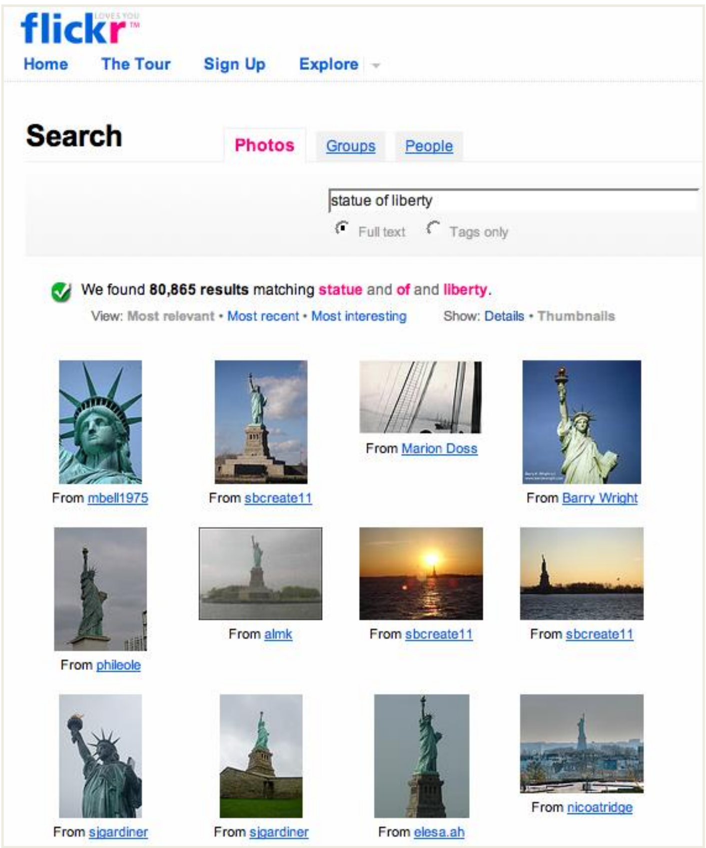
https://www.youtube.com/watch?v=quGhaggn3cQ

Questions?

Questions?

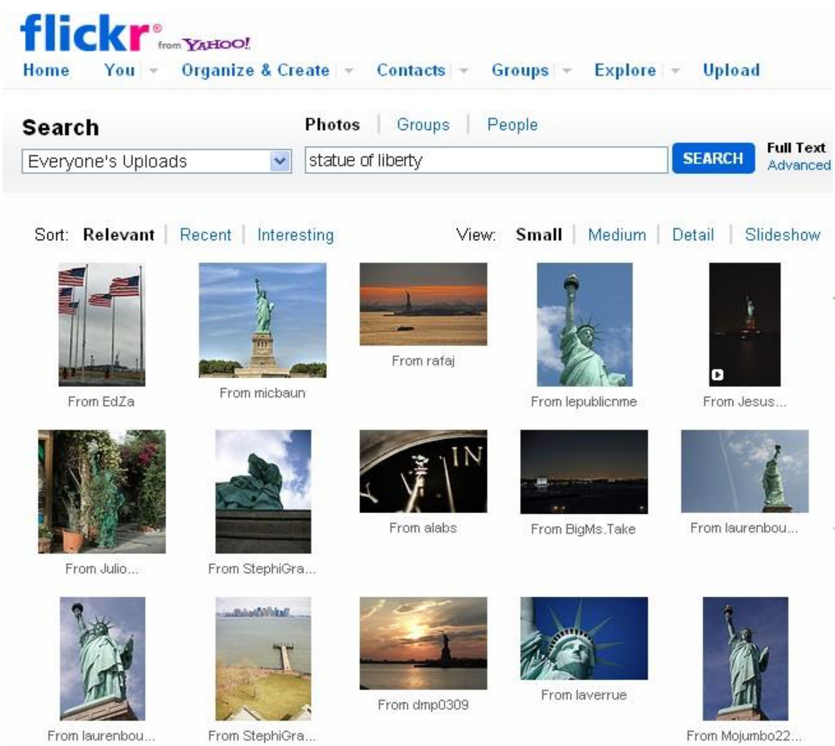
Multi-view stereo from Internet Collections

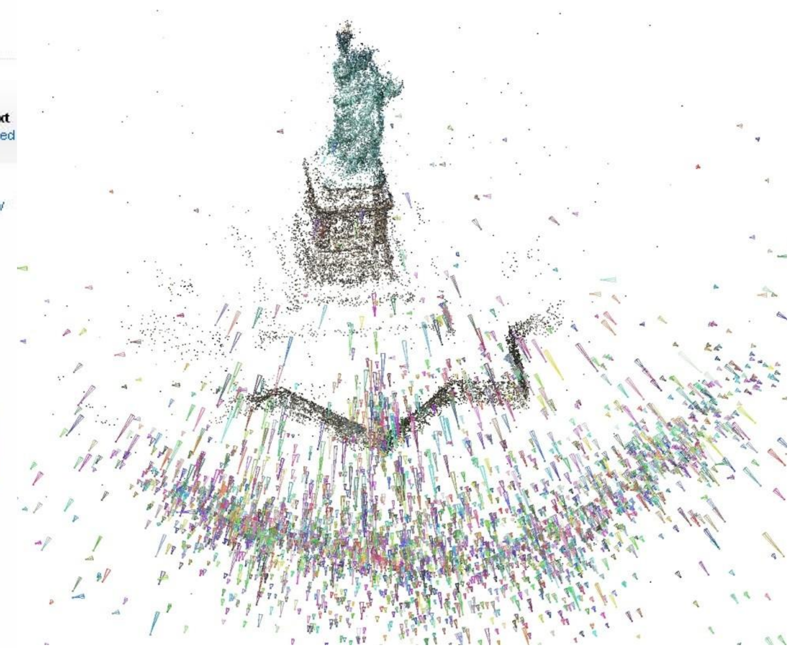
[Goesele, Snavely, Curless, Hoppe, Seitz, ICCV 2007]



Stereo from community photo collections

- Need structure from motion to recover unknown camera parameters
- Need view selection to find good groups of images on which to run dense stereo





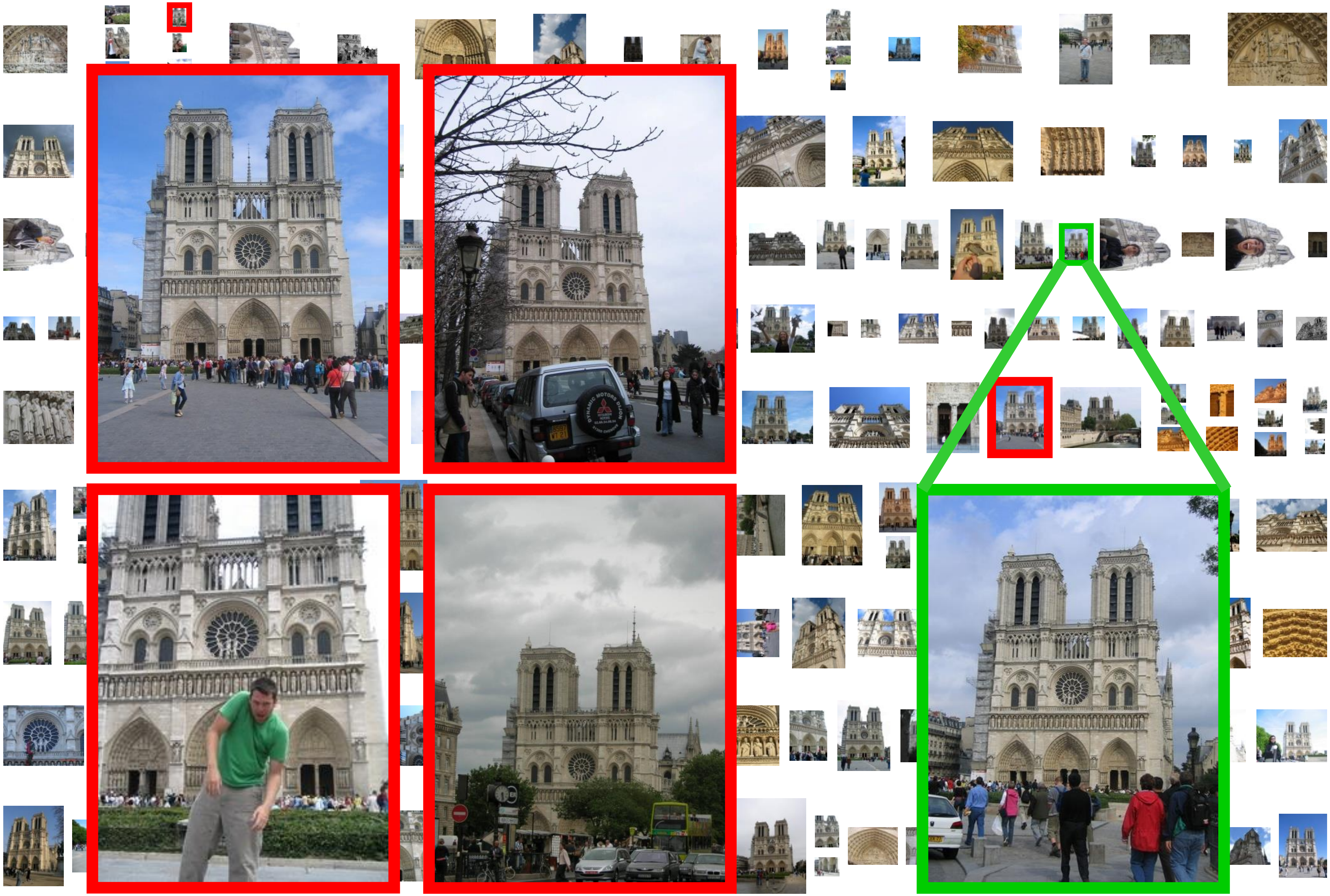
Challenges

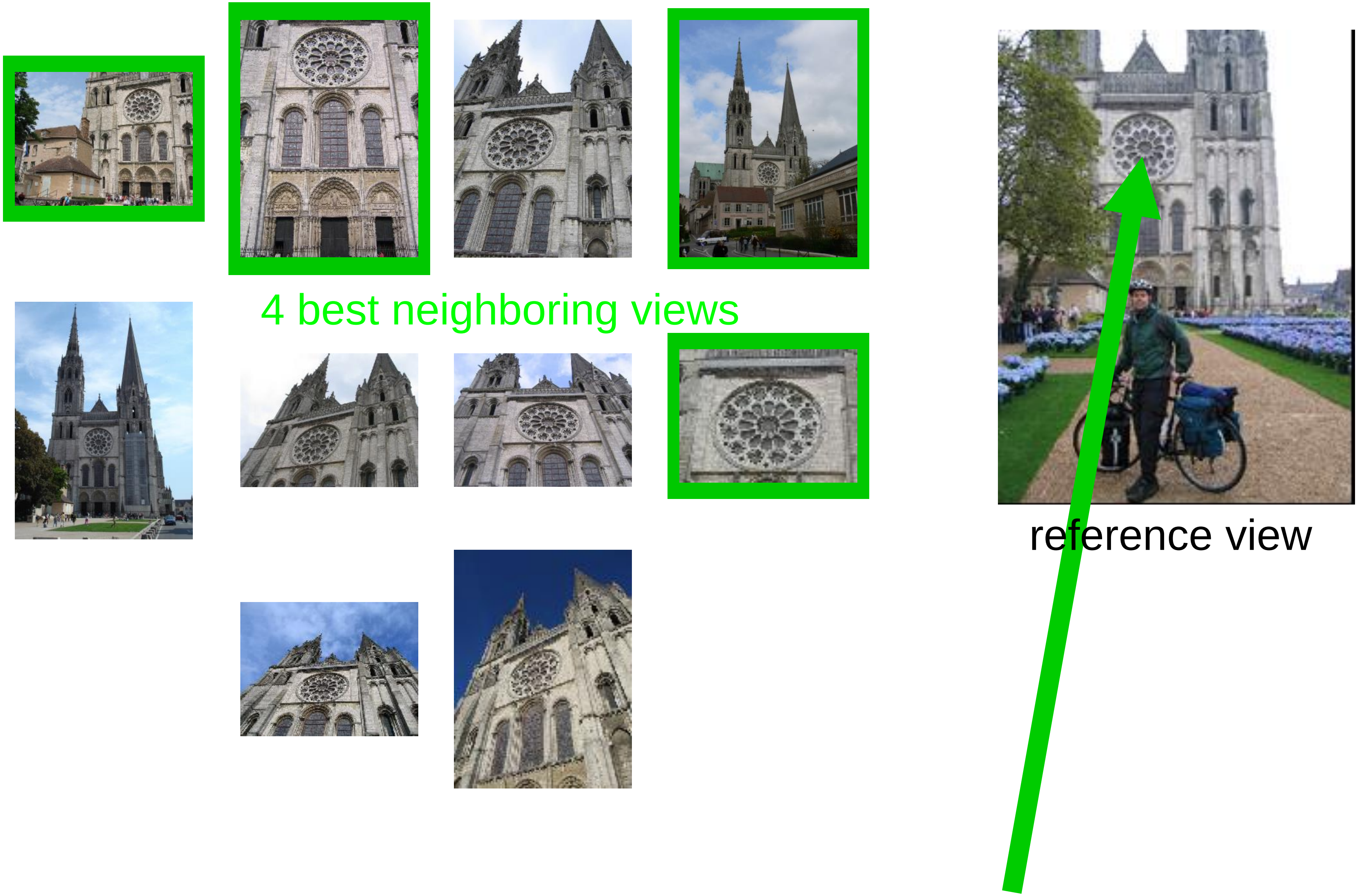
- appearance variation

- resolution

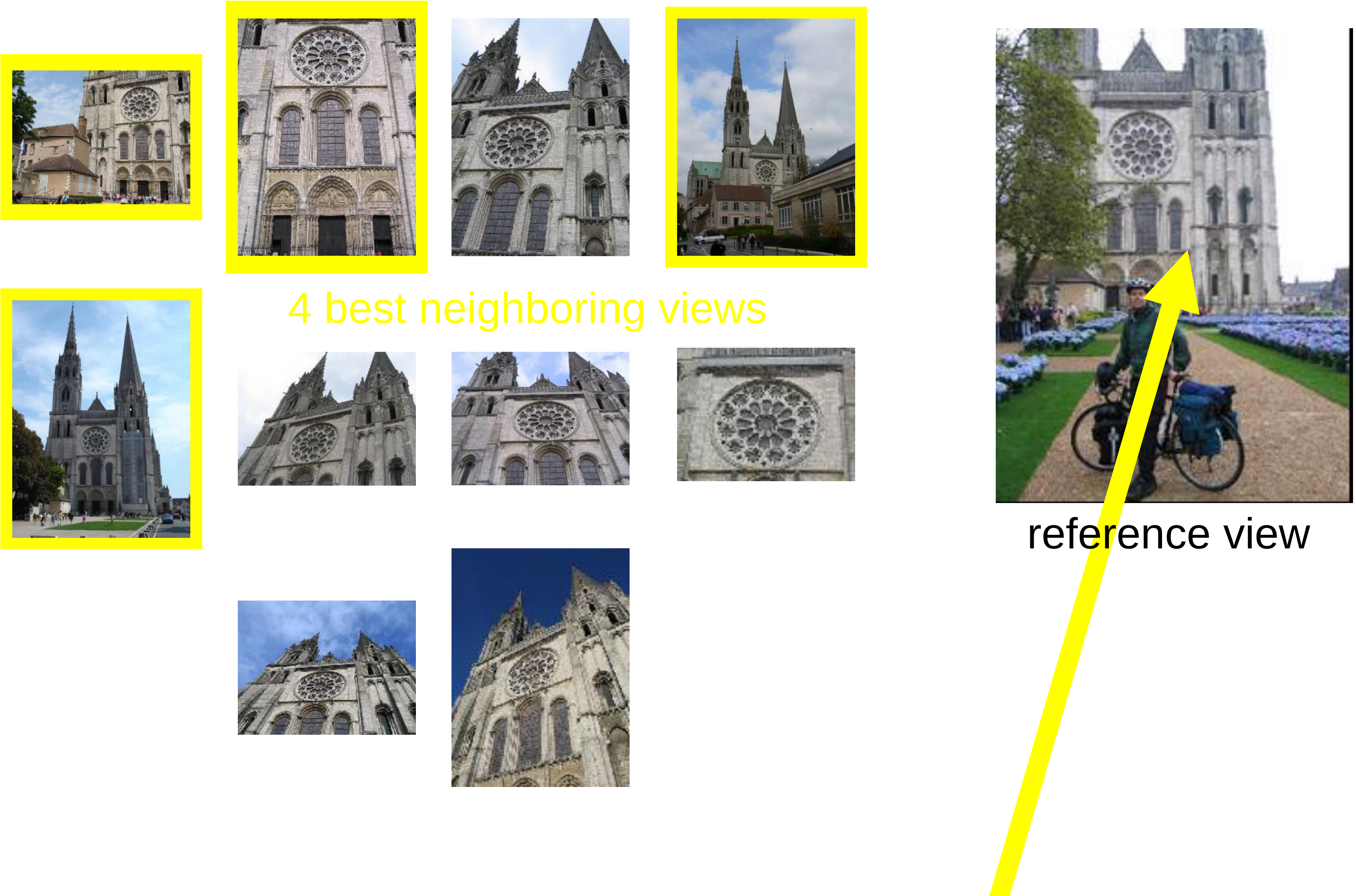
- massive collections

82,754 results for photos matching **notre** and **dame** and **paris**.

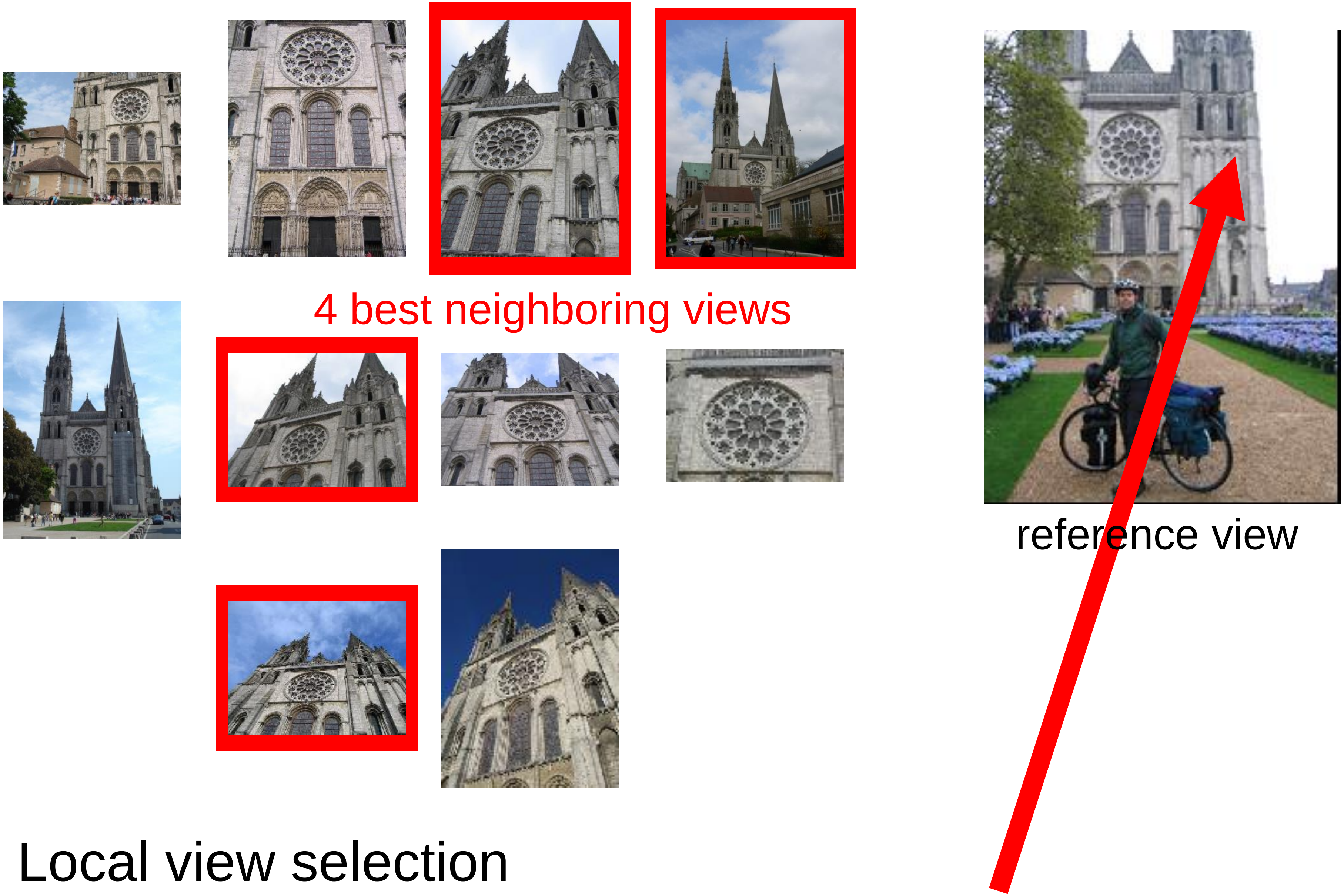




- Automatically select neighboring views for each point in the image
- Desiderata: good matches AND good baselines



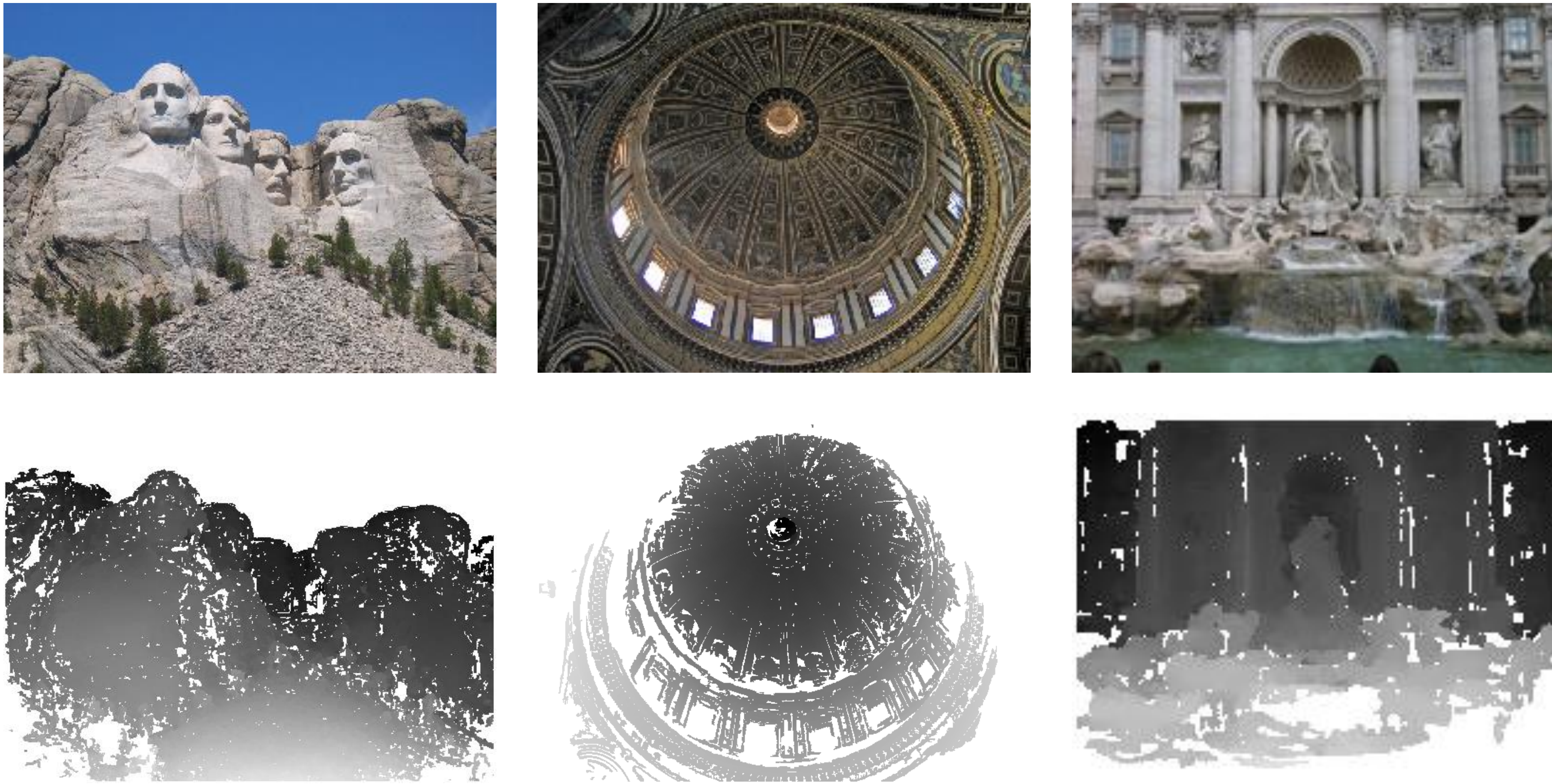
- Automatically select neighboring views for each point in the image
- Desiderata: good matches AND good baselines



Local view selection

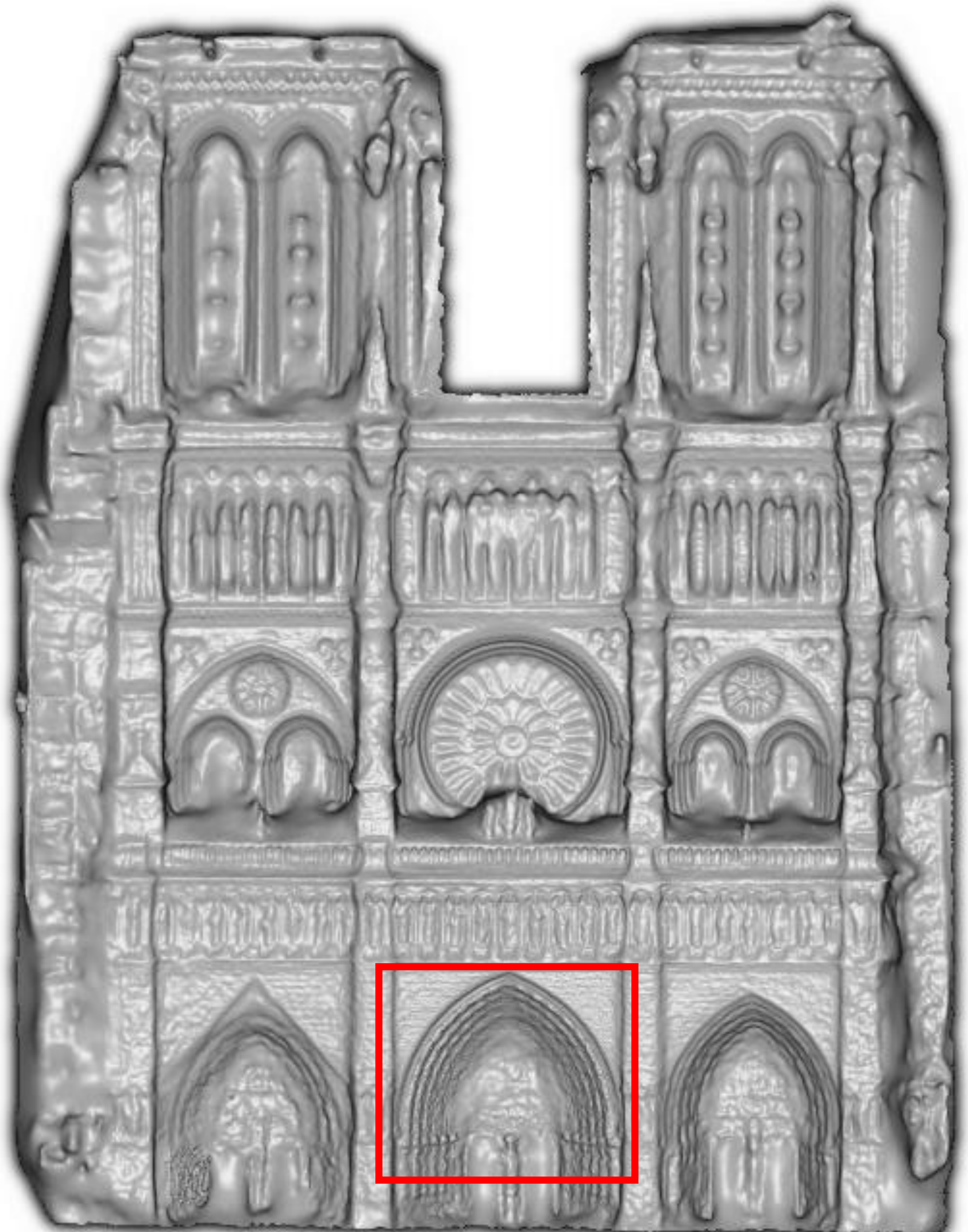
- Automatically select neighboring views for each point in the image
- Desiderata: good matches AND good baselines

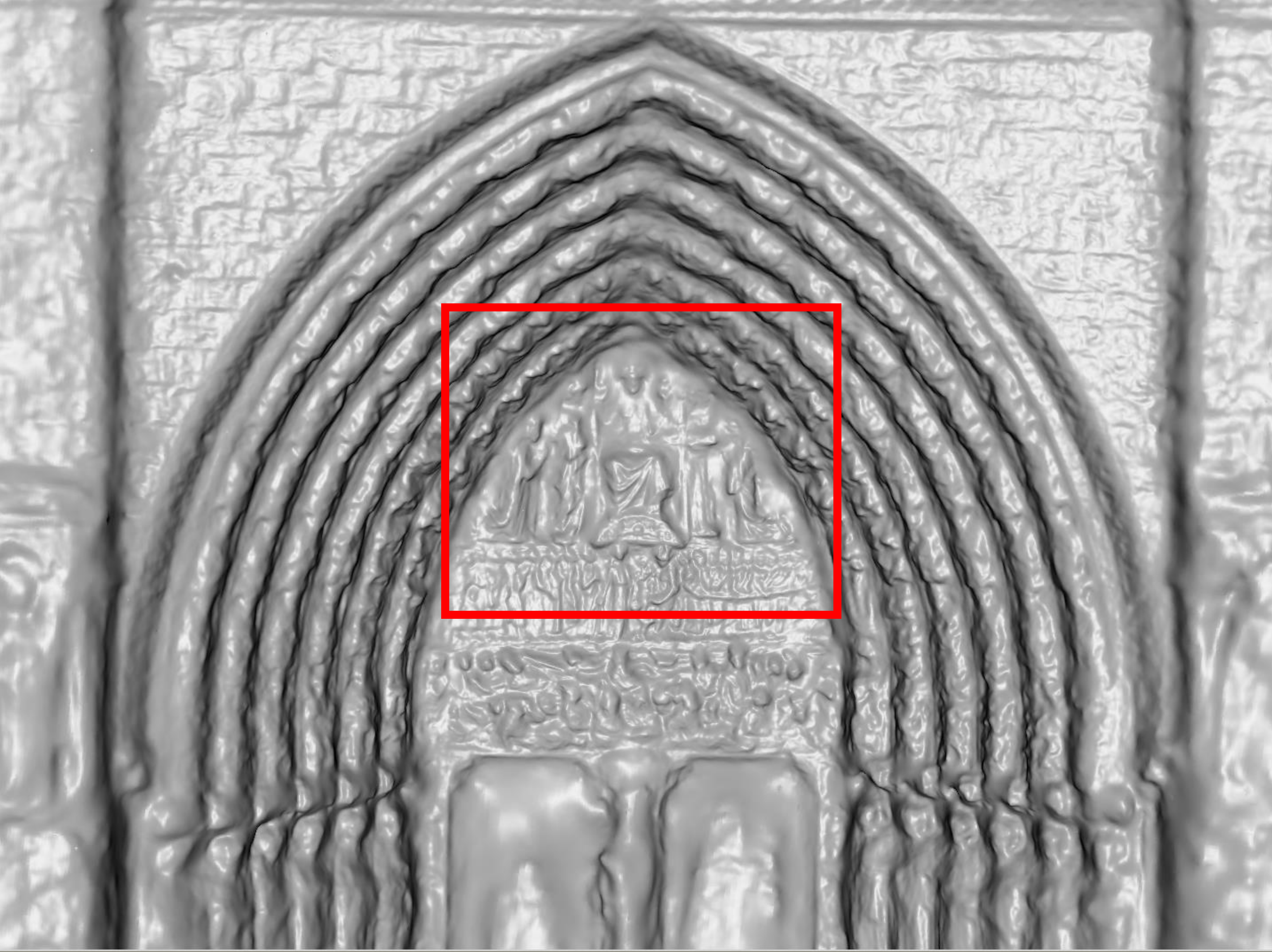
Results

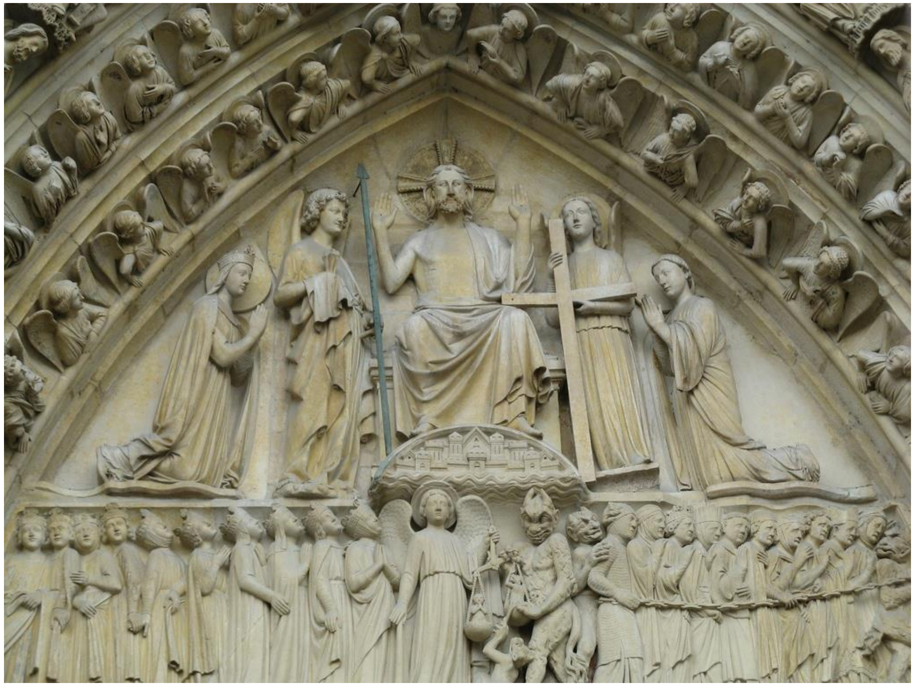


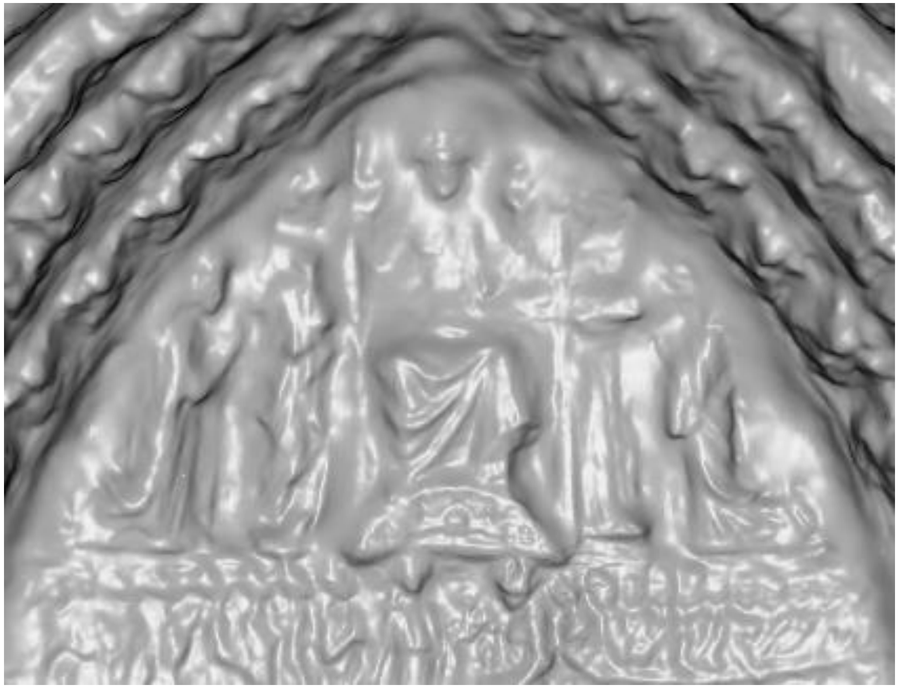
Notre Dame de Paris

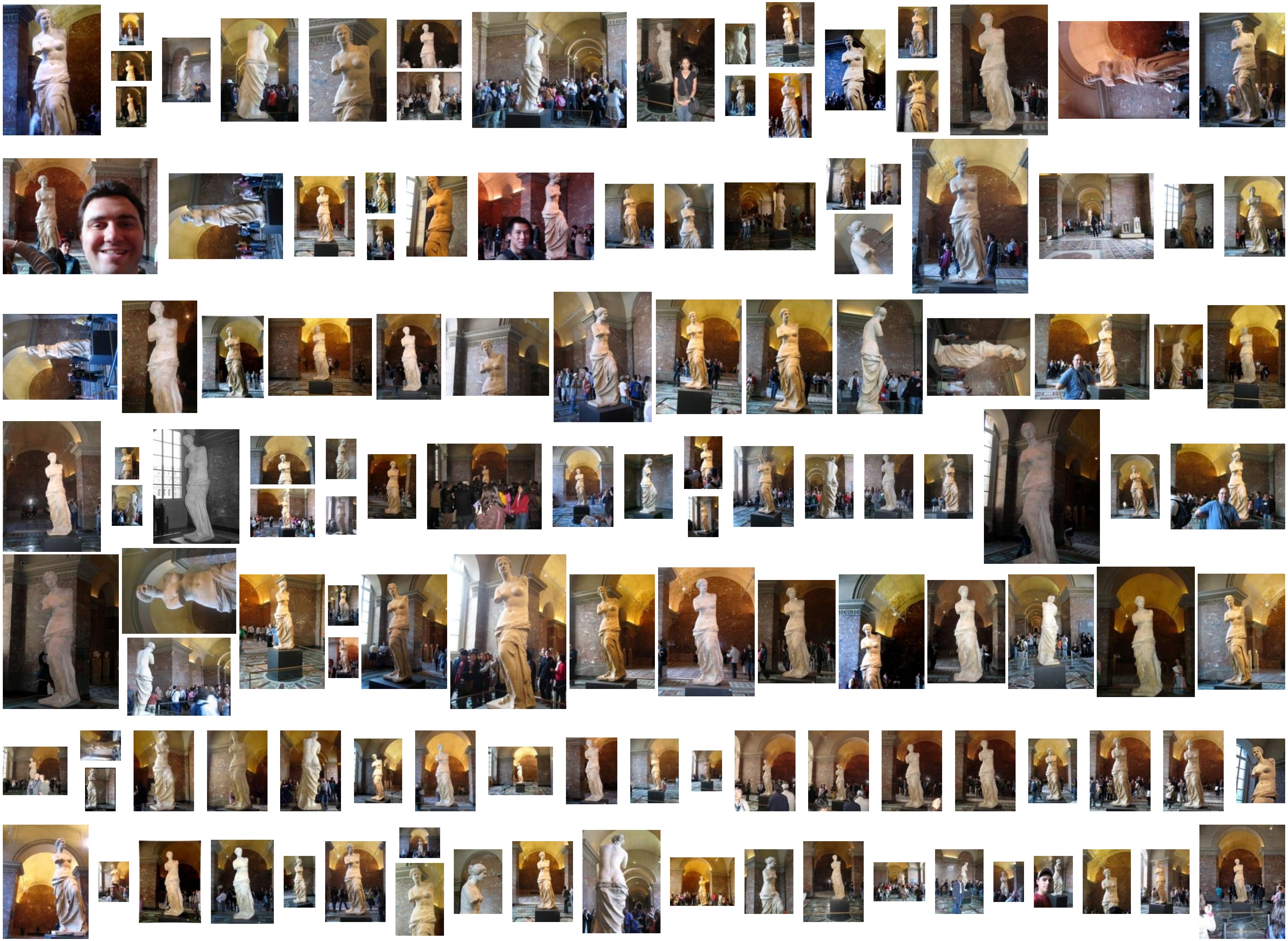
653 images
313 photographers



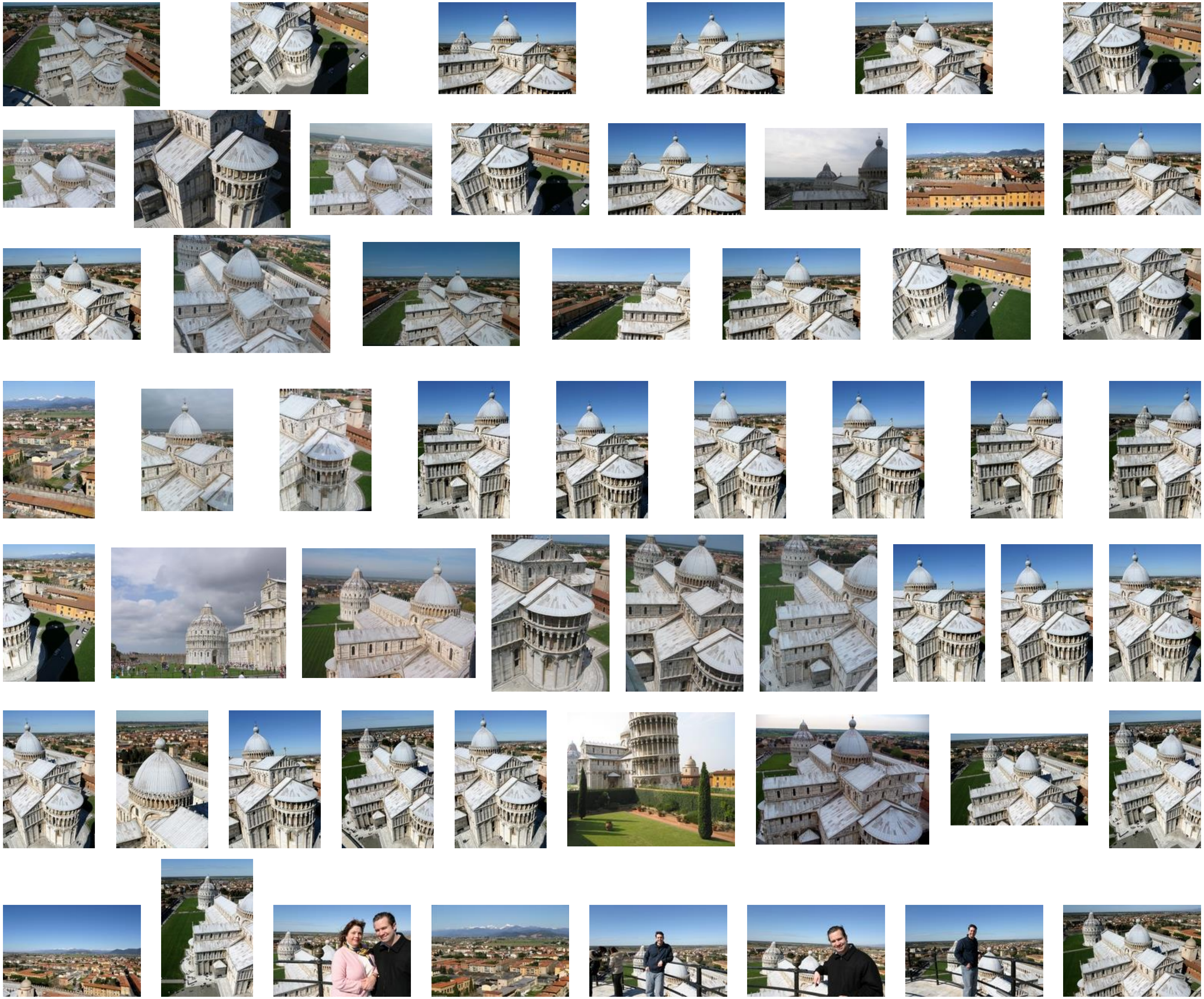


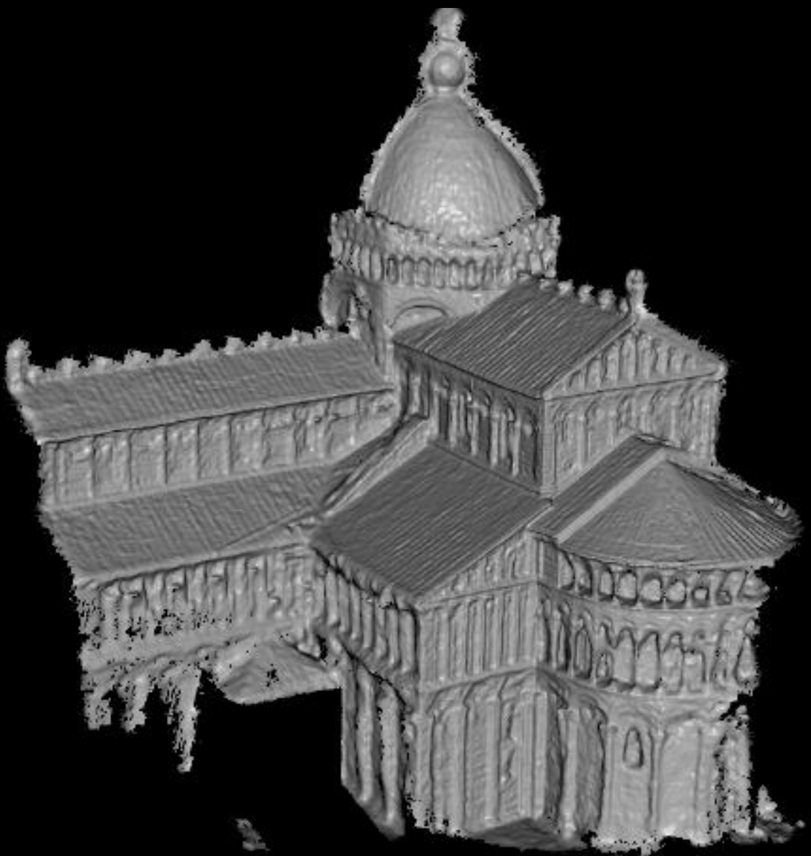




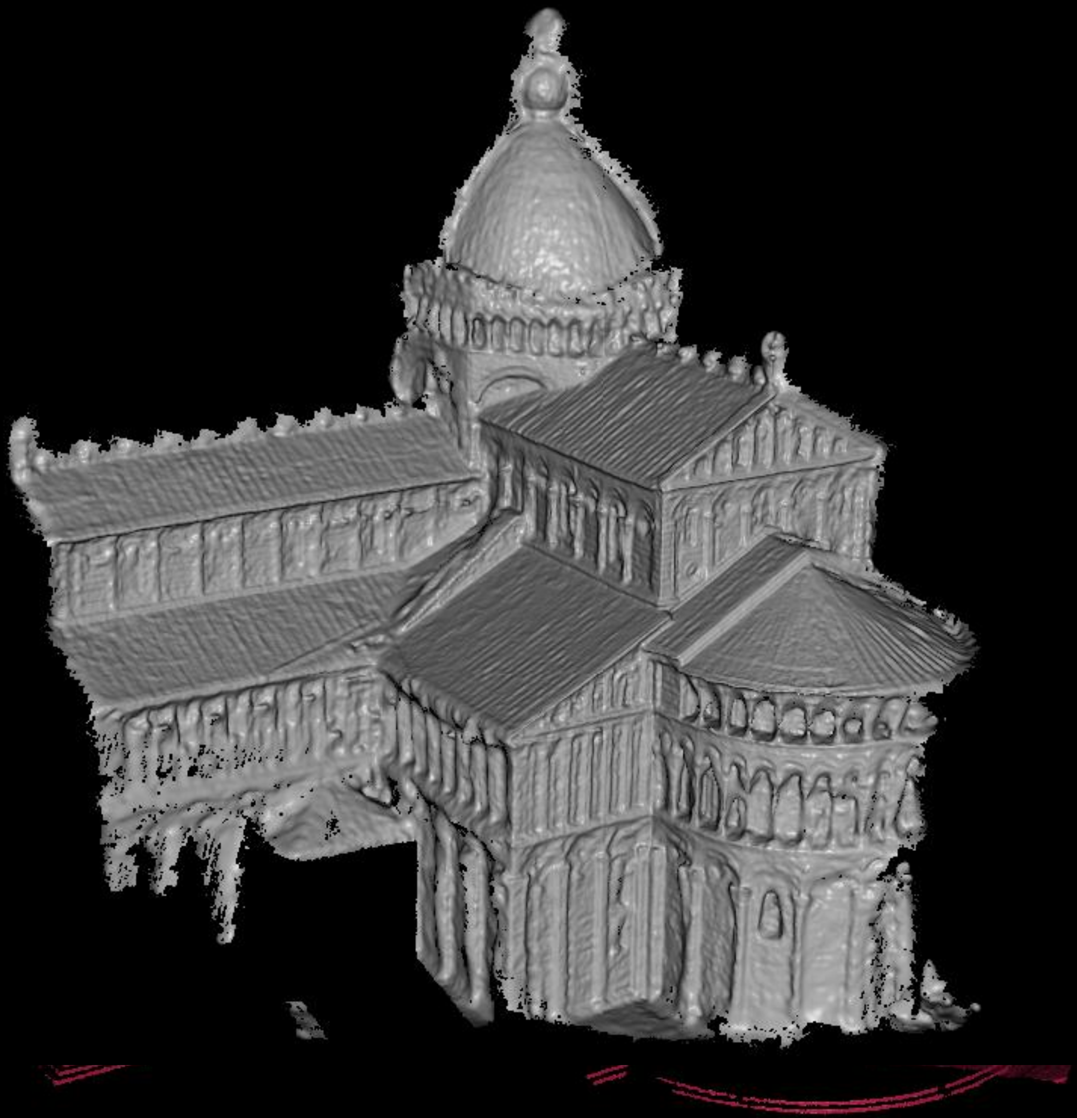


merged model of Venus de Milo



merged model of Pisa Cathedral



Accuracy compared to laser scanned model:
90% of points within *0.25%* of ground truth