清華大學 Tsinghua University
交叉信息研究院 Institute for Interdisciplinary Information Sciences

Lecture 5: Learning-Based Multi-View Stereo

Li Yi

2025.03.20

Announcement

- Assignment 1 due soon
- Assignment 2 is going be released

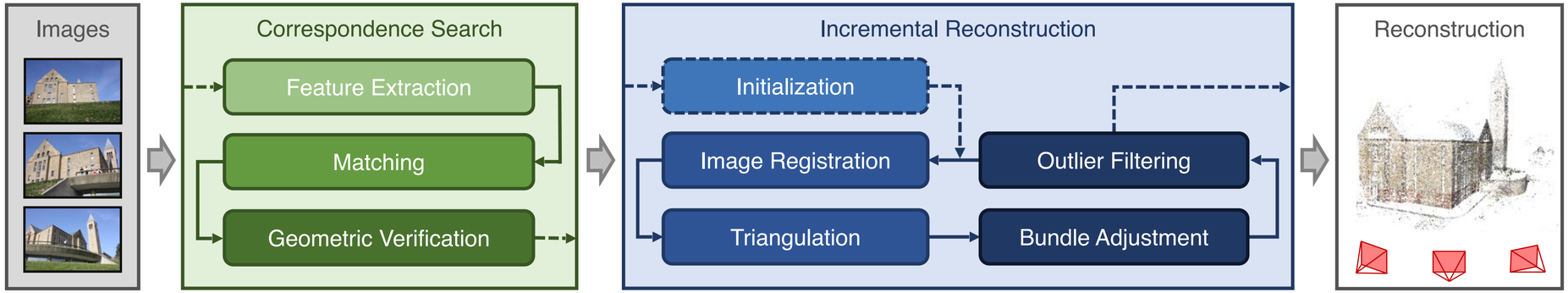
What to Expect in the Rest of the Semester

- Application-based lecture organization
- Go over important 3D learning techniques
- Understand the common thinking style and methodology shared by different topics
- Introduce key technical points but not all the details of a DL pipeline

Recap

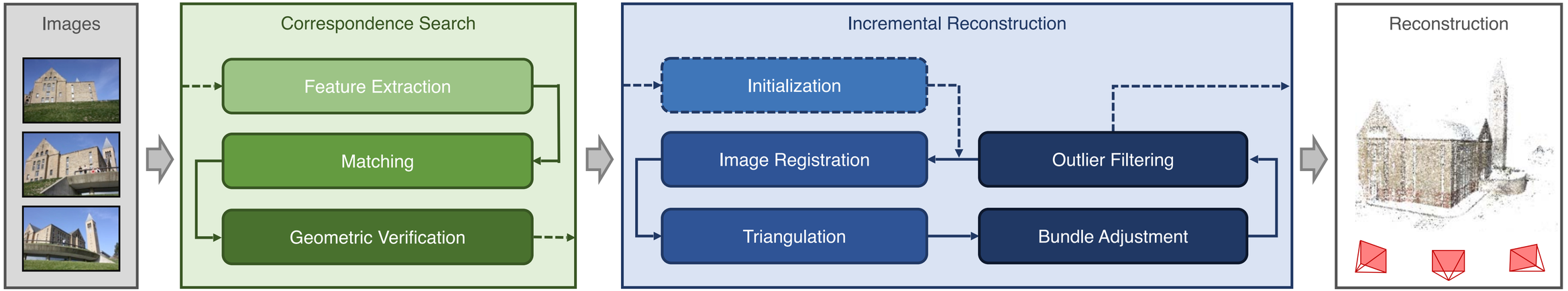
- 2D-3D basics
- Structure from motion
- Learning-based structure from motion

Incremental SfM



Slide credits: Johannes Schönberger

Learning for SfM



Can we improve the robustness/precision via data-driven learning?

Example 1: Improving features and keypoints for matching

Example 2: Improving the matching process via global reasoning

Image credits: Johannes Schönberger

A minimal matching pipeline

SuperGlue: context aggregation + matching + filtering

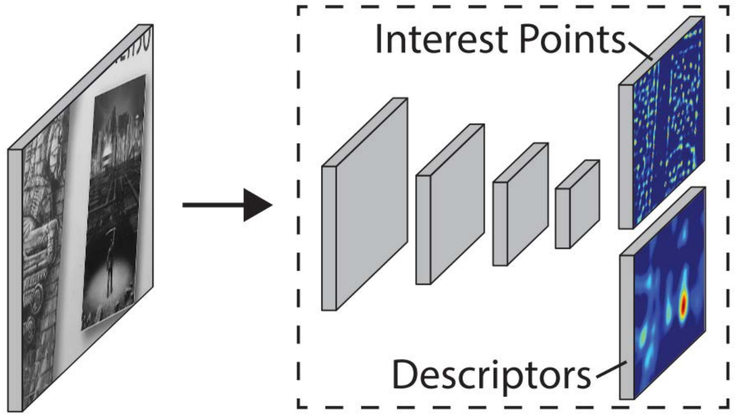
image pair

detection → description → feature matching → outlier filtering → pose estimation

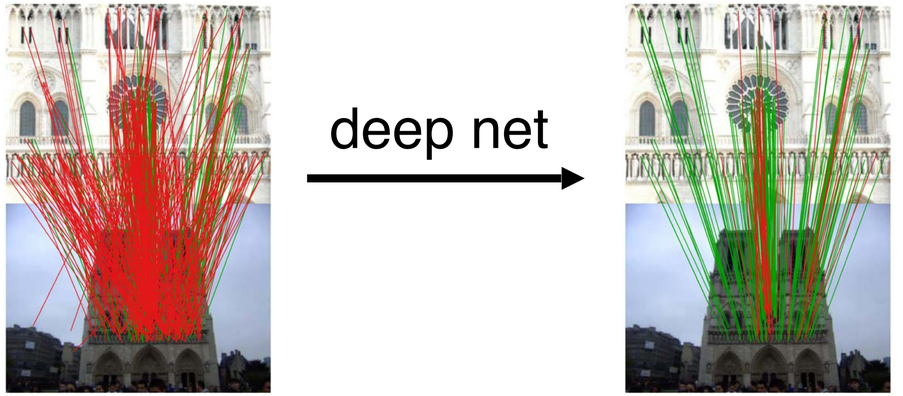
> Classical: SIFT, ORB
> Learned: SuperPoint, D2-Net

Nearest Neighbor Matching

> Heuristics: ratio test, mutual check
> Learned: classifier on set



[DeTone et al, 2018]

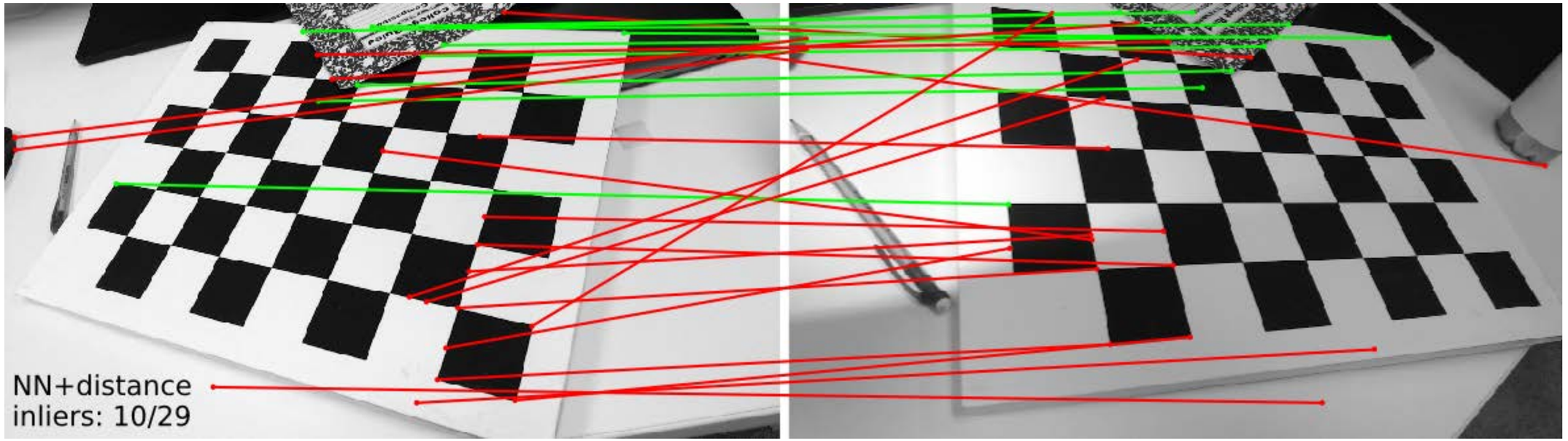


[Yi et al, 2018]

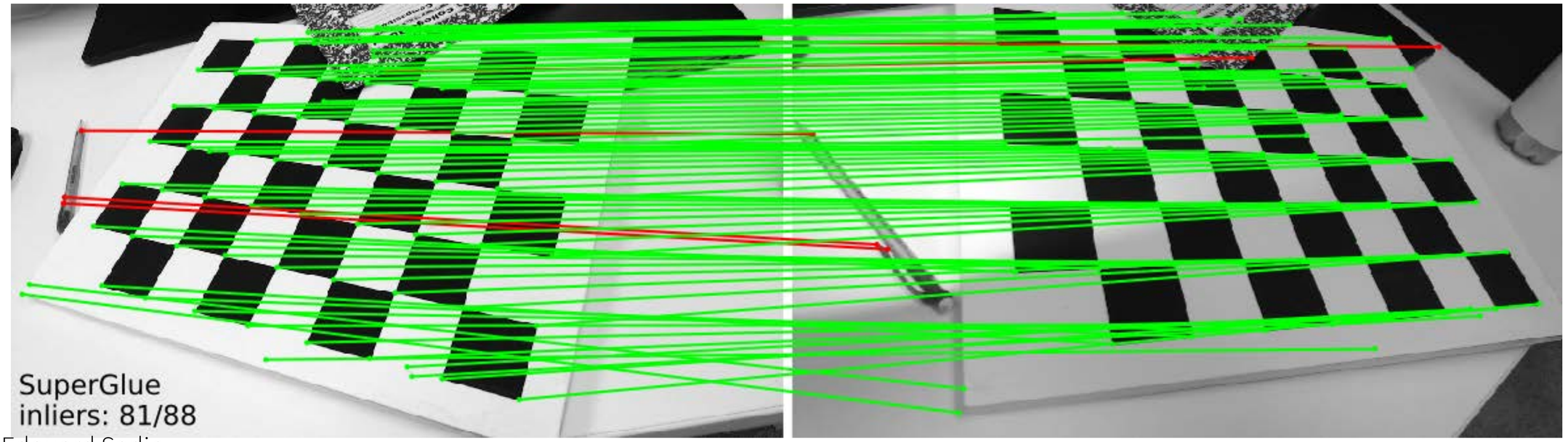
Slide credits: Paul-Edouard Sarlin

The importance of context

no SuperGlue



with SuperGlue



Slide credits: Paul-Edouard Sarlin

Problem formulation

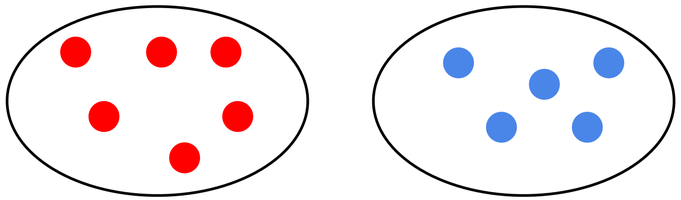
Inputs → 

→ **Outputs**

- Images **A** and **B**
- **2 sets** of **M**, **N** **local features**
  - Keypoints:
    - Coordinates
    - Confidence
  - Visual descriptors:

$$\mathbf{p}_i := (x, y, c)_i$$

$$\mathbf{d}_i$$

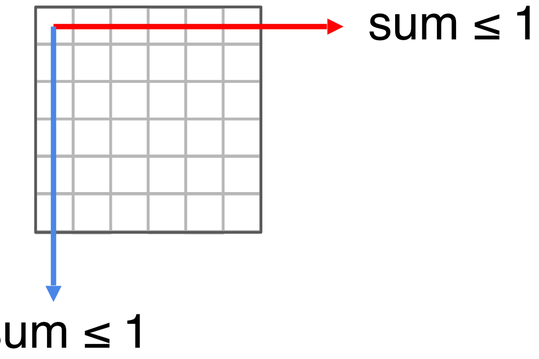


Single a match per keypoint

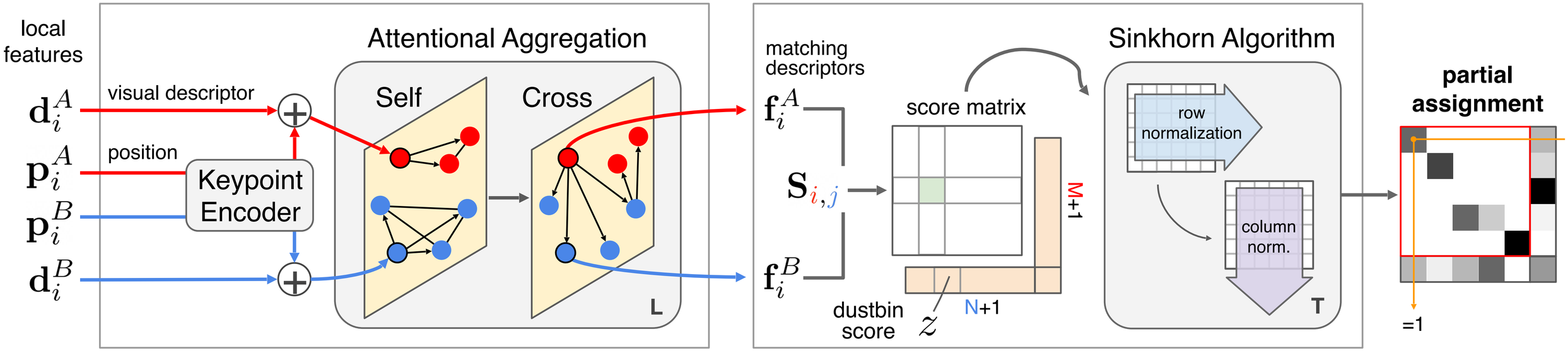
+ occlusion and noise

→ a **soft partial assignment**:

$$\mathbf{P} \in [0, 1]^{M \times N}$$



Slide credits: Paul-Edouard Sarlin



A Graph Neural Network with attention

Encodes **contextual cues** & priors

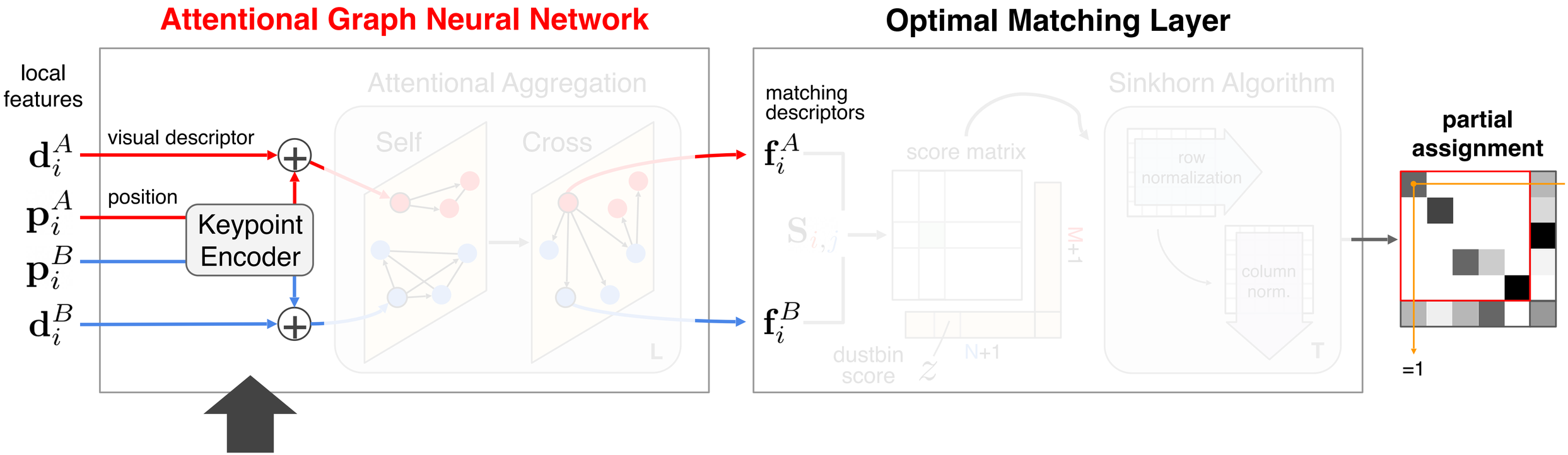
Reasons about the 3D scene

Solving a partial assignment problem

Differentiable **solver**

Enforces the assignment constraints
= **domain knowledge**

Slide credits: Paul-Edouard Sarlin

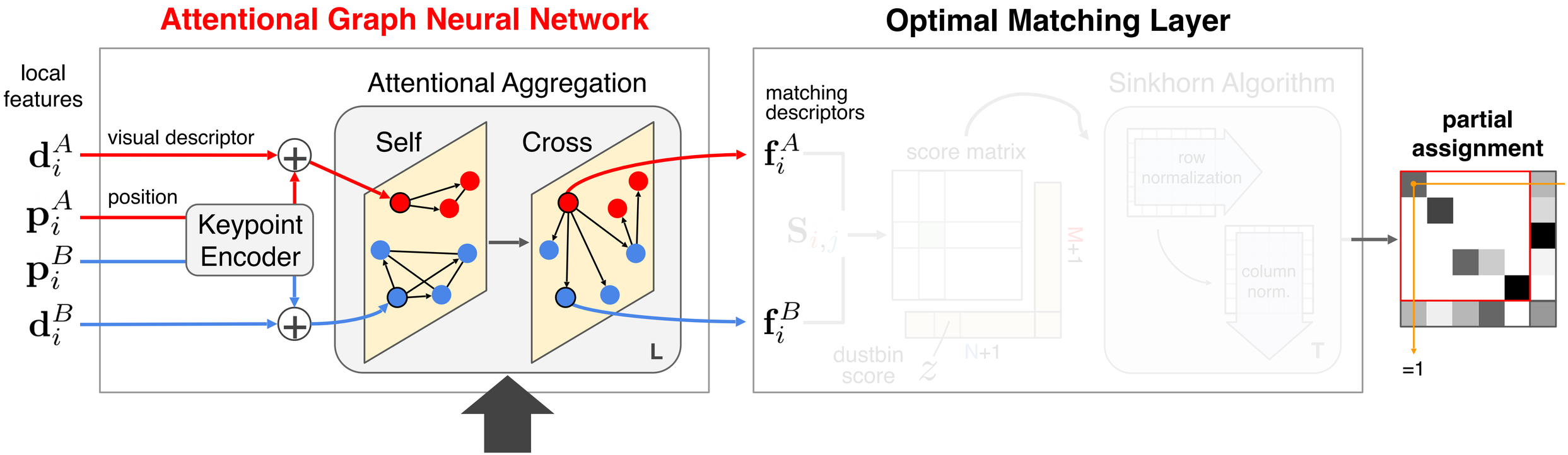


- Initial representation for each keypoints i: ${}^{(0)}\mathbf{x}_i$
- Combines visual appearance and position with an MLP:

$$ {}^{(0)}\mathbf{x}_i = \mathbf{d}_i + \mathrm{MLP}\left(\mathbf{p}_i\right) $$

Multi-Layer Perceptron

Slide credits: Paul-Edouard Sarlin



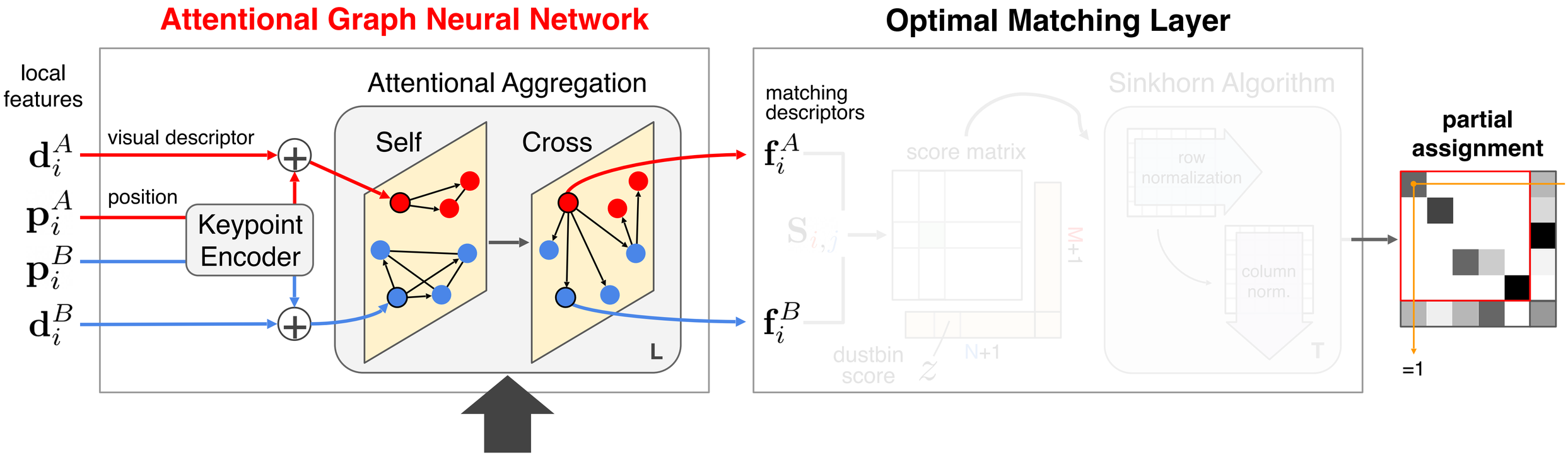
Update the representation based on other keypoints:

- in the same image: "**self**" edges
- in the other image: "**cross**" edges

→ A complete **graph** with two types of edges

$$ {}^{(\ell)}\mathbf{x}_i^A \longrightarrow {}^{(\ell+1)}\mathbf{x}_i^A $$

Slide credits: Paul-Edouard Sarlin



Update the representation using a **Message Passing Neural Network**

$$ {}^{(\ell+1)}\mathbf{x}_i^A = {}^{(\ell)}\mathbf{x}_i^A + \mathrm{MLP}\left(\left[{}^{(\ell)}\mathbf{x}_i^A \,||\, \mathbf{m}_{\mathcal{E}\to i}\right]\right) $$

Similar to a single-head transformer layer

the message

Slide credits: Paul-Edouard Sarlin

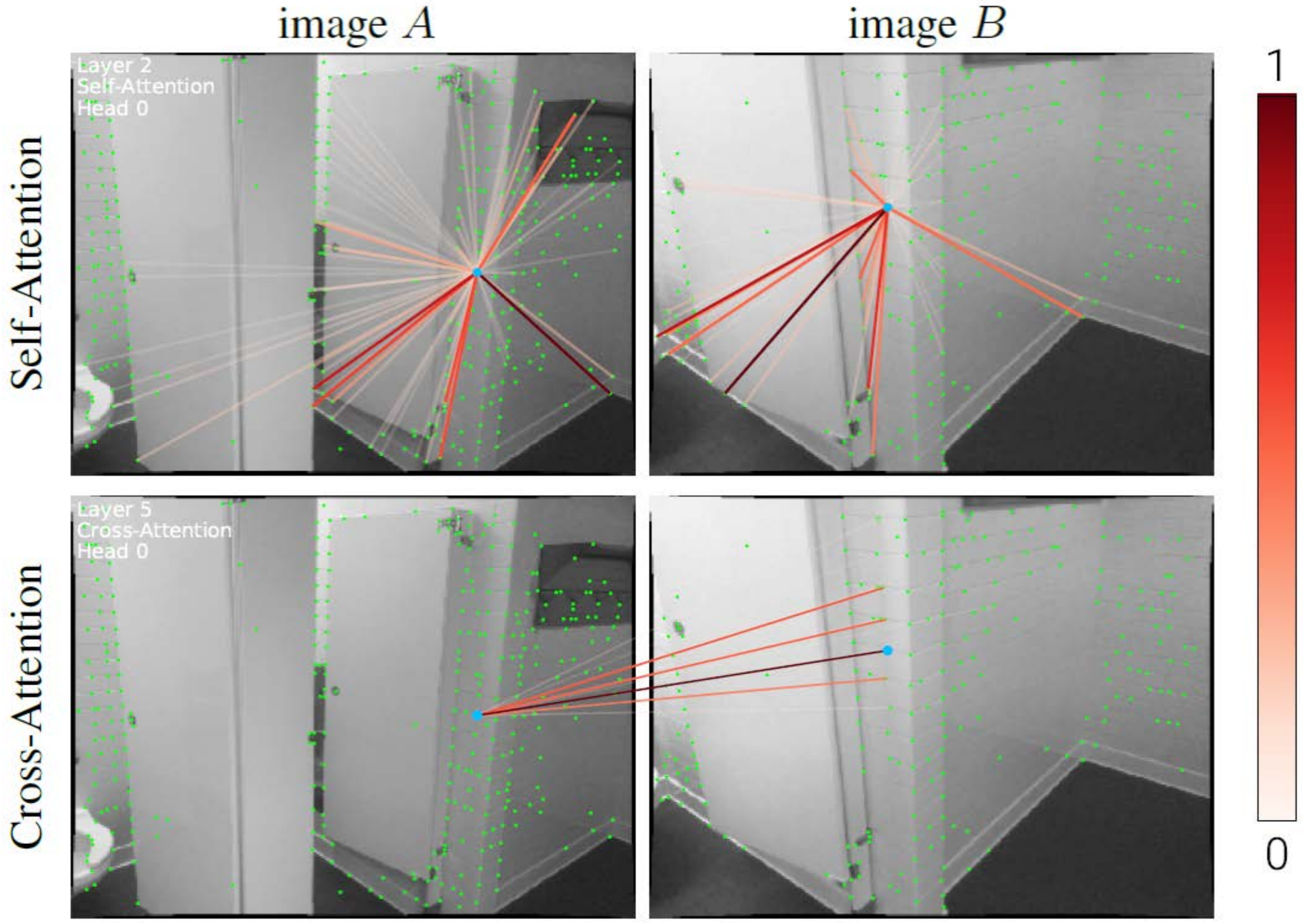
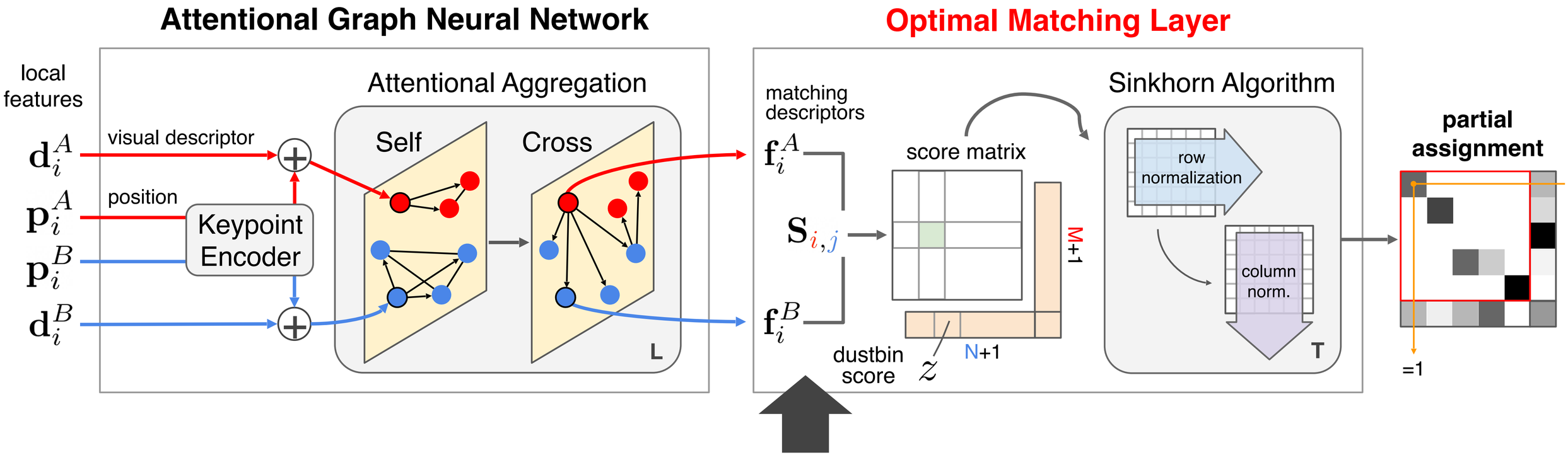


image A
image B
Self-Attention
Cross-Attention
Layer 2
Self-Attention
Head 0
Layer 5
Cross-Attention
Head 0
1
0

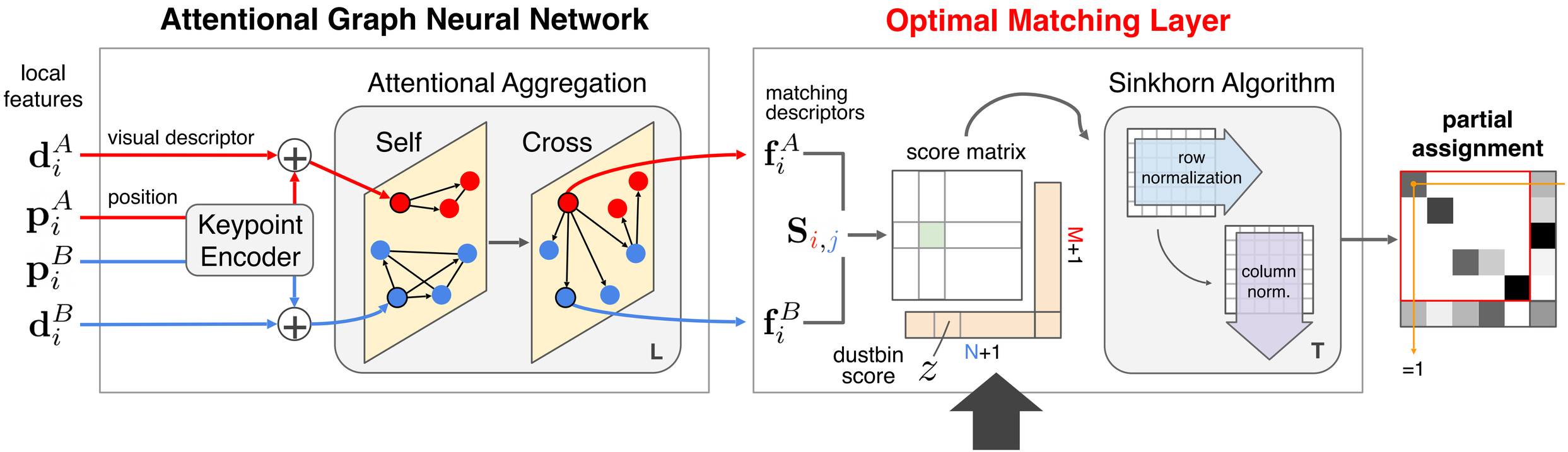


Compute a **score matrix** $\mathbf{S} \in \mathbb{R}^{M \times N}$ for all matches:

$$\mathbf{f}_i^A = \mathbf{W} \cdot {}^{(L)}\mathbf{x}_i^A + \mathbf{b}$$

$$\mathbf{S}_{i,j} = < \mathbf{f}_i^A, \mathbf{f}_j^B >$$

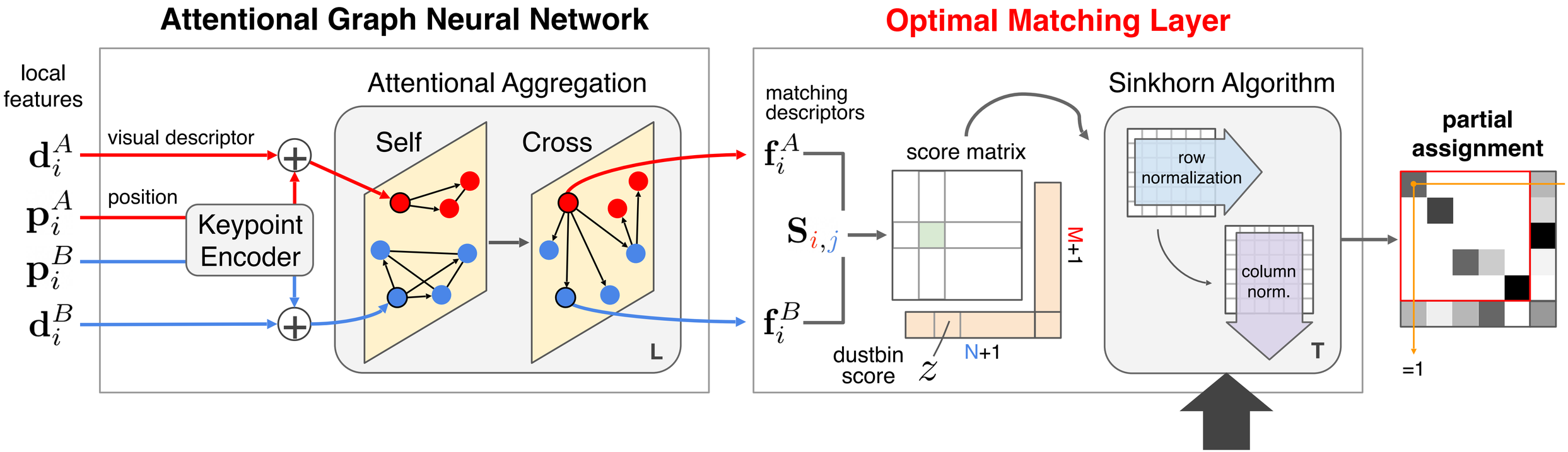
Slide credits: Paul-Edouard Sarlin



- Occlusion and noise: unmatched keypoints are assigned to a **dustbin**
- **Augment** the scores with a learnable dustbin score z

$$\bar{\mathbf{S}}_{i,N+1} = \bar{\mathbf{S}}_{M+1,j} = \bar{\mathbf{S}}_{M+1,N+1} = z \in \mathbb{R}$$

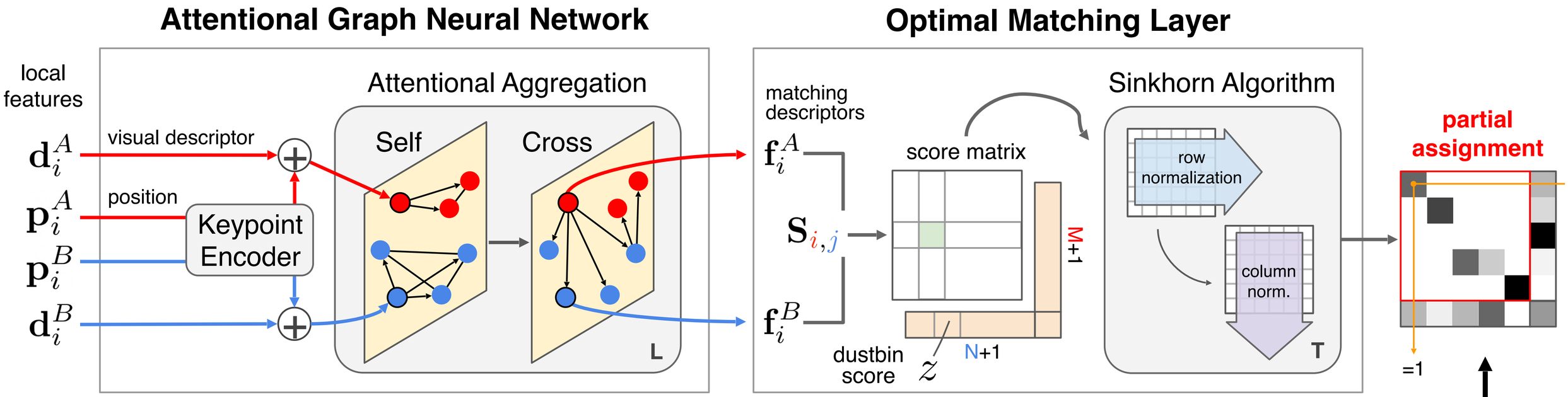
Slide credits: Paul-Edouard Sarlin



- Compute the assignment $\bar{\mathbf{P}}$ that maximizes $\sum_{i,j} \bar{\mathbf{S}}_{i,j} \bar{\mathbf{P}}_{i,j}$
- Solve an **optimal transport** problem
- With the **Sinkhorn algorithm**: differentiable & soft Hungarian algorithm

[Sinkhorn & Knopp, 1967]

Slide credits: Paul-Edouard Sarlin

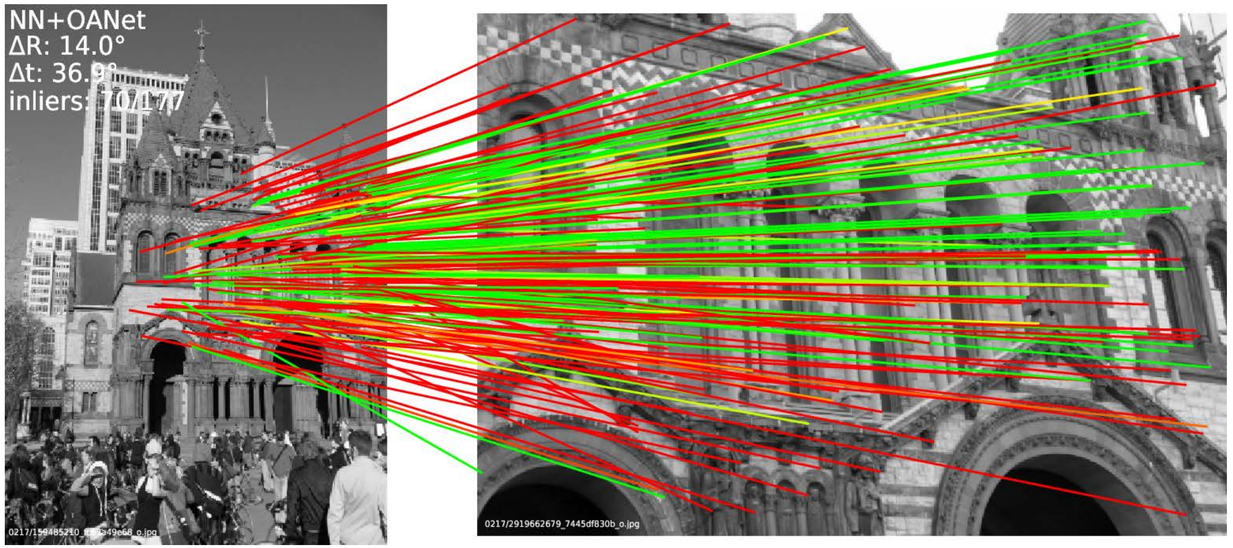


- Compute **ground truth correspondences** from pose and depth
- Find which keypoints should be **unmatched**
- Loss: maximize the log-likelihood $\bar{\mathbf{P}}_{i,j}$ of the GT cells

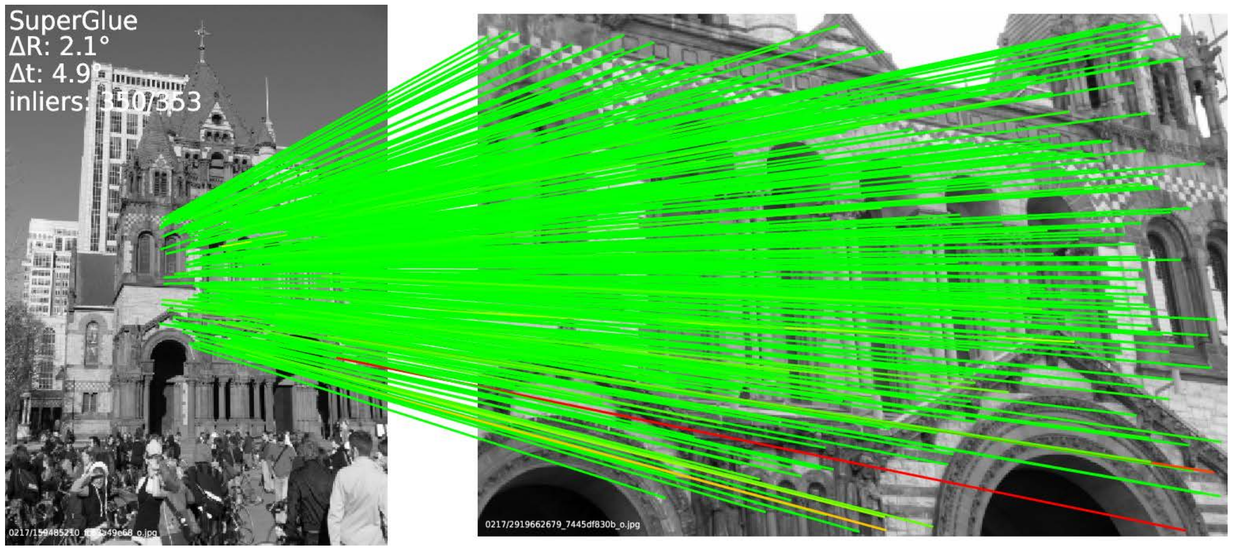
Slide credits: Paul-Edouard Sarlin

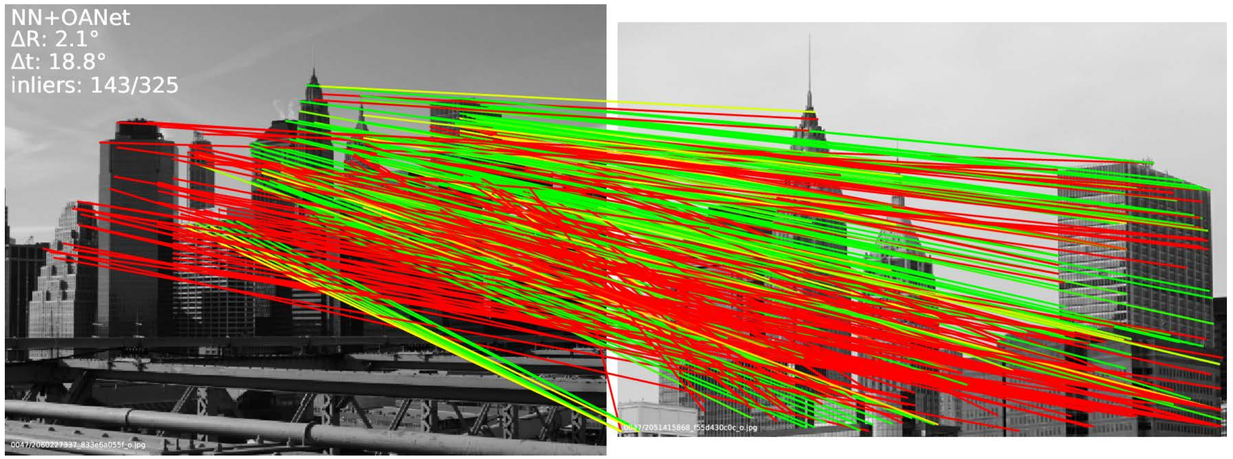
Results: outdoor - SfM

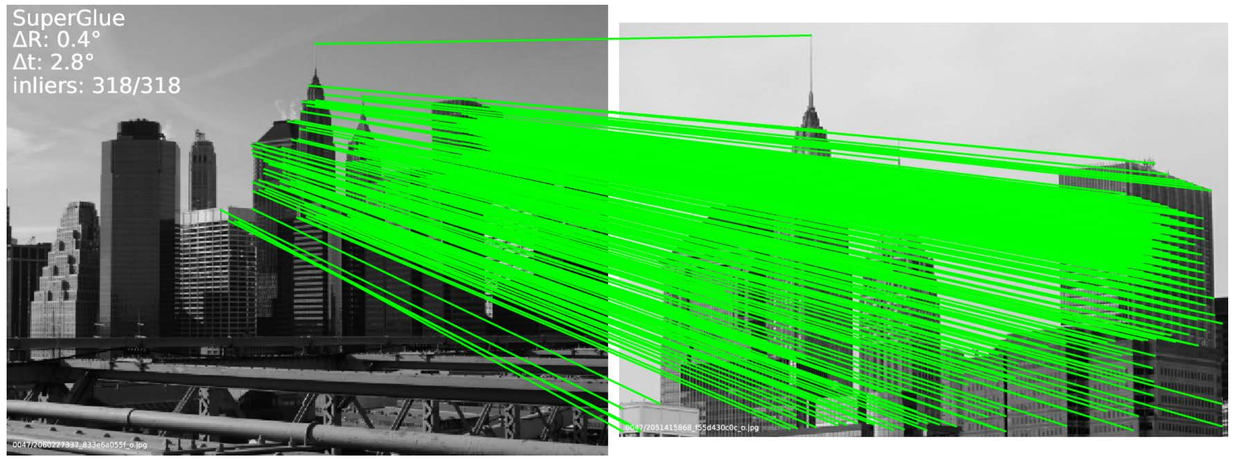
SuperPoint + NN + OA-Net (inlier classifier)



SuperPoint + **SuperGlue**



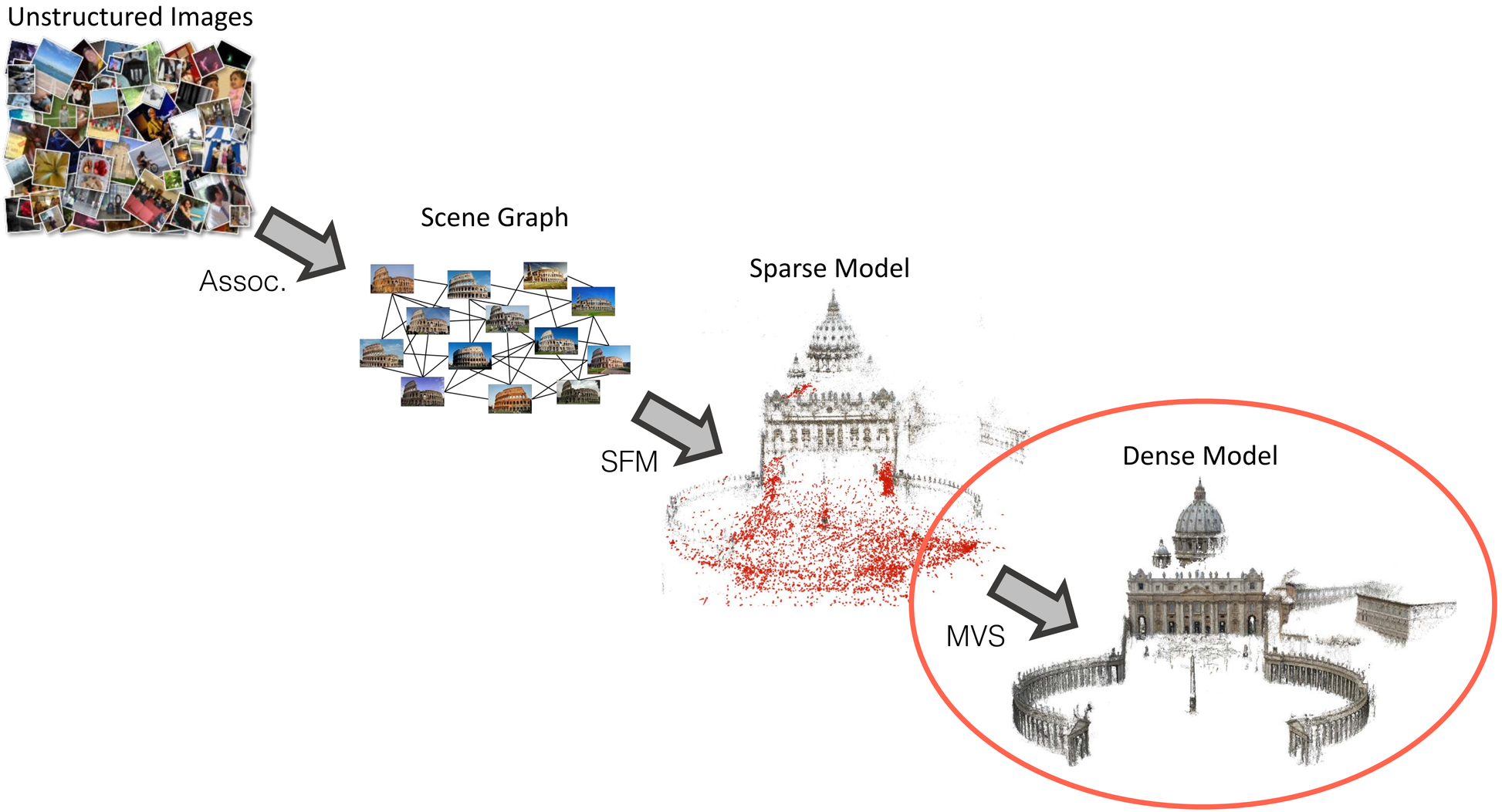




SuperGlue: more **correct matches** and fewer **mismatches**

Slide credits: Paul-Edouard Sarlin

Today's Focus



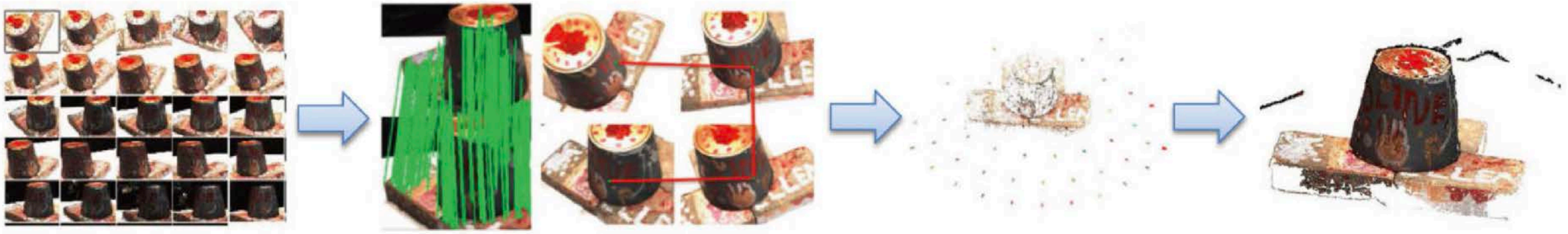
Outline

- Introduction to multi-view stereo (MVS)
- Classic MVS
- Learning-based MVS: a first pipeline
- Learning-based MVS: Improvements
  - Adaptive Space Sampling
  - Depth-Normal Consistency

Outline

- *Introduction to multi-view stereo (MVS)*
- Classic MVS
- Learning-based MVS: a first pipeline
- Learning-based MVS: Improvements
  - Adaptive Space Sampling
  - Depth-Normal Consistency

A Typical Image-Based 3D Reconstruction Pipeline



Input images

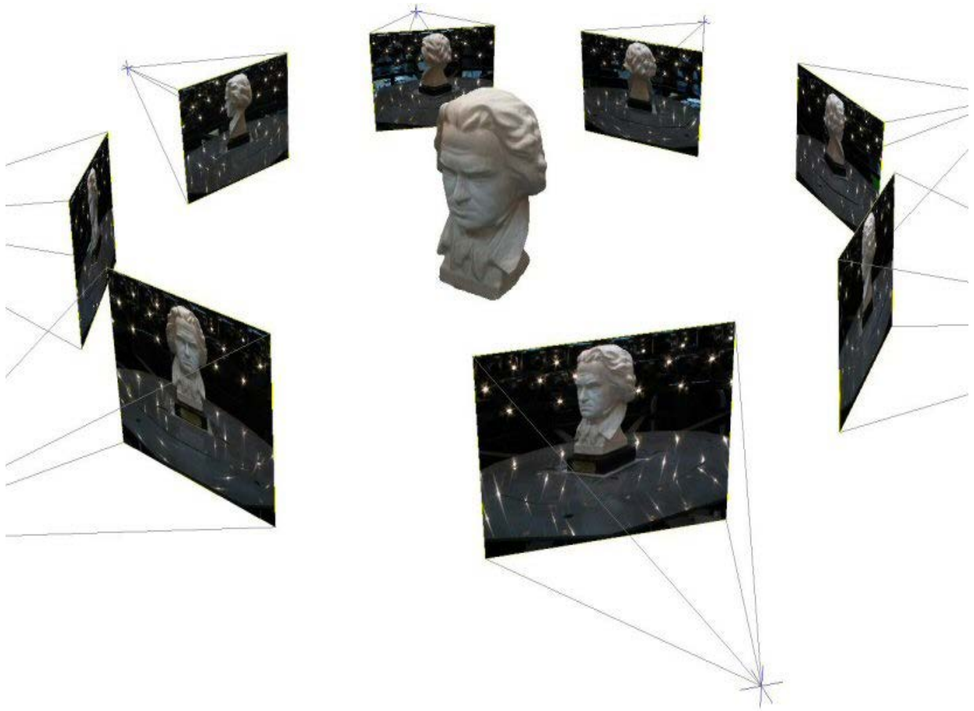
Find matched points across images

Structure from Motion (SFM)

dense reconstruction

Multi-View Stereo

Reconstruct the dense 3D shape from a set of **images** and **camera parameters**



1. Goldlucke et al. “A Super-resolution Framework for High-Accuracy Multiview Reconstruction”

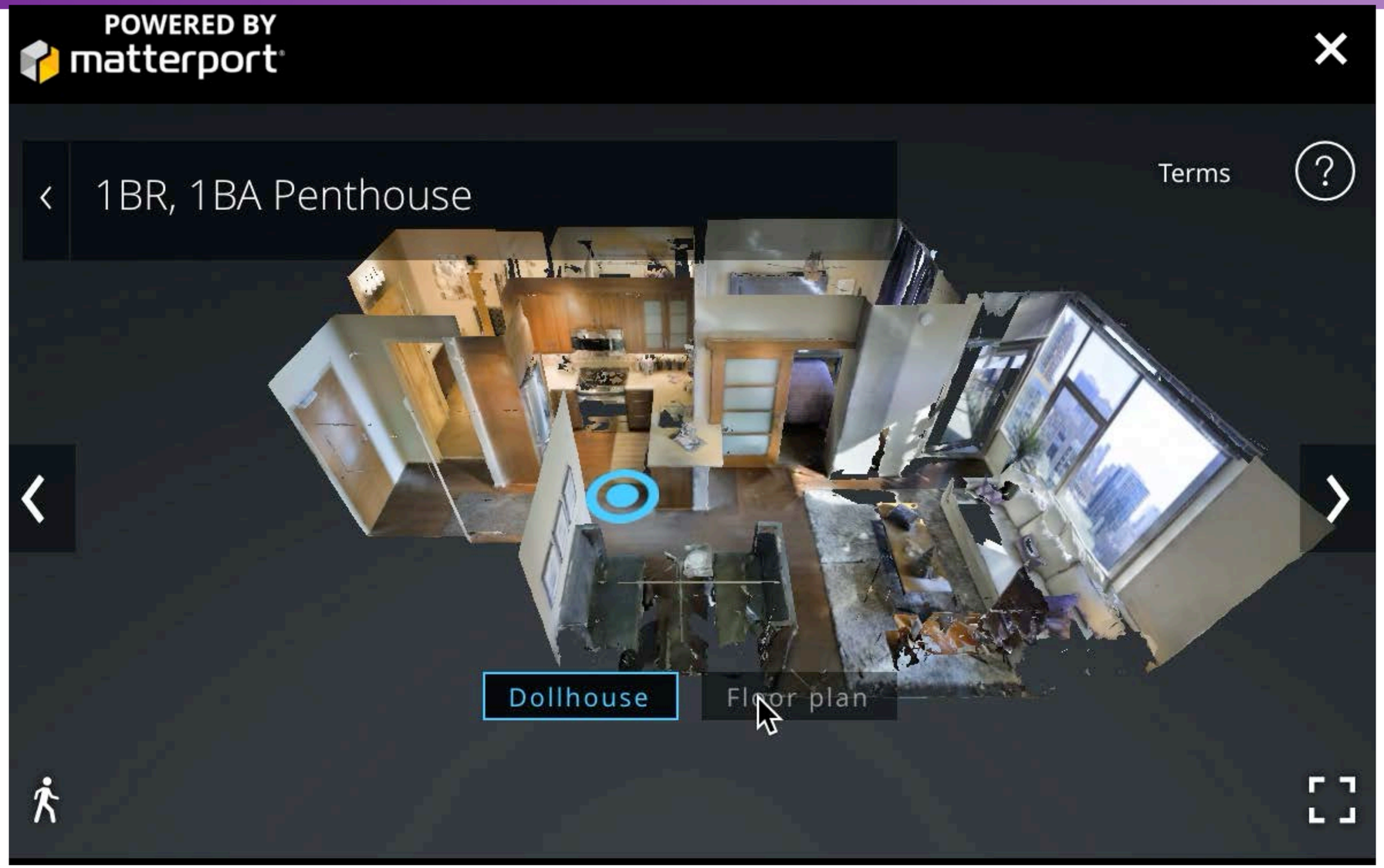
Application

Source: N. Snavely

Application

Source: N. Snavely

Application



Source: N. Snavely

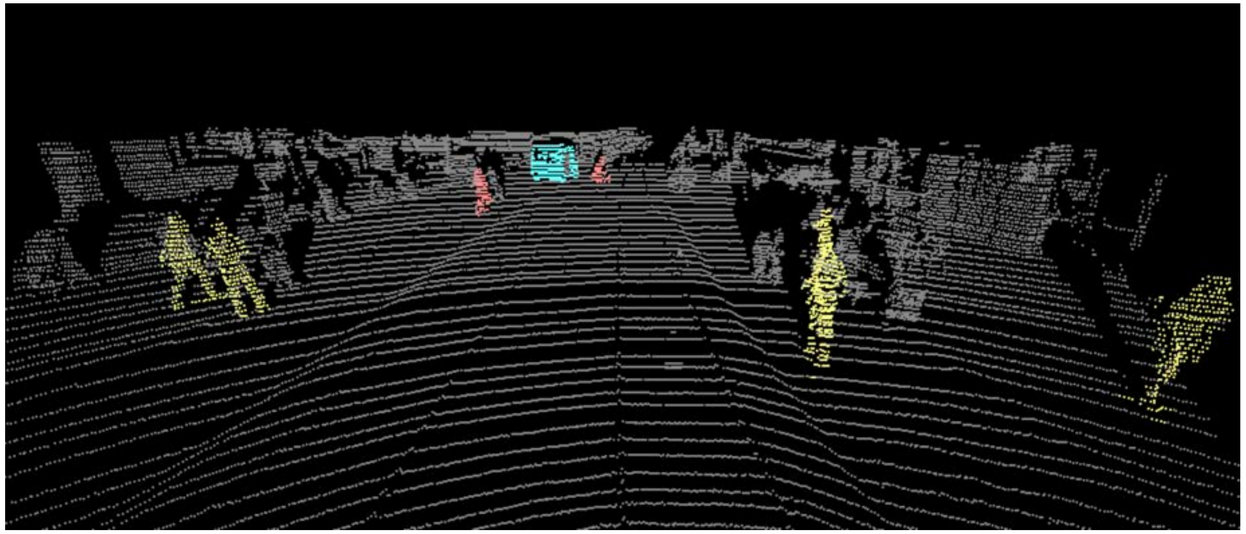
Application

- Enable inspection in hard to reach areas with drone photos and 3D reconstruction
- Create 3D model from images
- Provide tools to inspect on images and map interactions to 3D

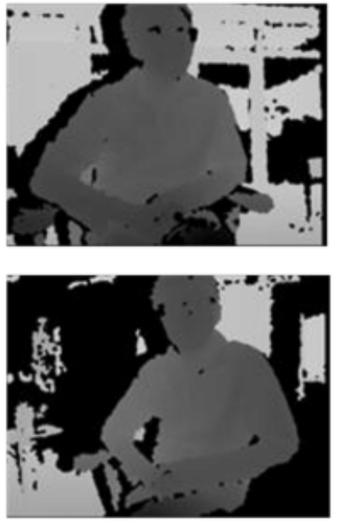
Source: D. Hoiem

Why MVS Given the Development of Depth Sensors?

- Measurement of depth sensors is either sparse or with limited range
- Texture and lighting information is missing
- Nice association between image and 3D



Sparse Lidar point cloud



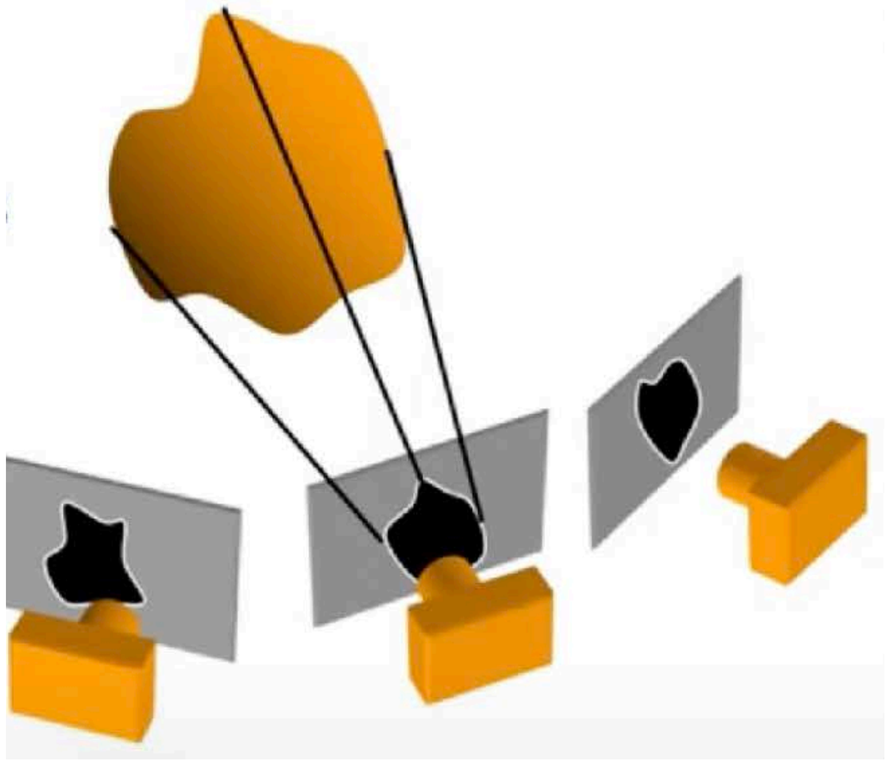
Kinect point cloud with limited range

Outline

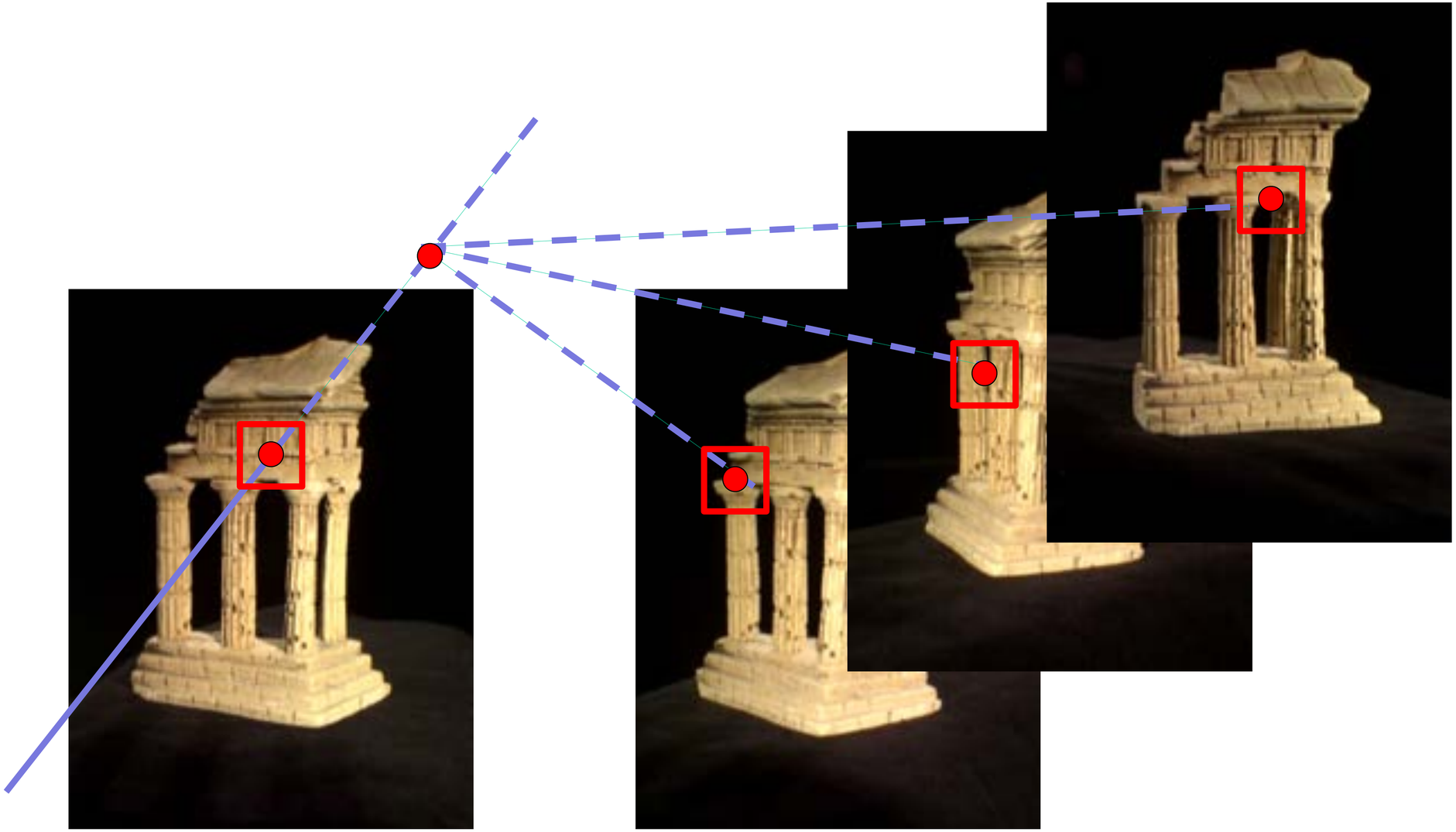
- Introduction to multi-view stereo (MVS)
- ***Classic MVS***
- Learning-based MVS: a first pipeline
- Learning-based MVS: improvements
  - Adaptive Space Sampling
  - Depth-Normal Consistency

Reconstruction from Silhouettes

- Approach
  - Back-project each silhouette
  - Intersect back-projected volumes
- Pros
  - Easy to implement, fast
  - Accelerated with Octrees
- Cons
  - Requires identification of silhouettes
  - Not photo-consistent
  - No concavities

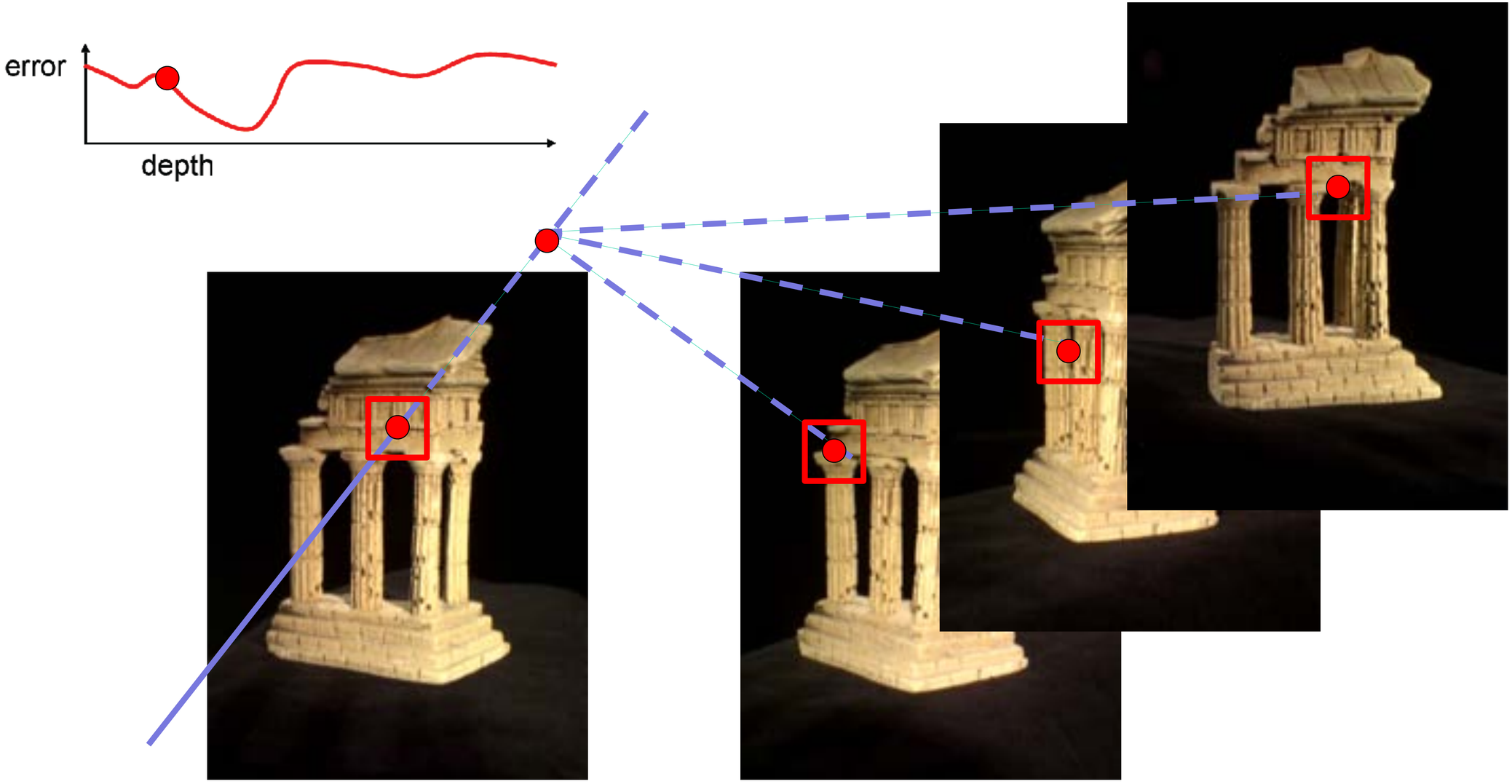


Reconstruction from Photometric Consistency



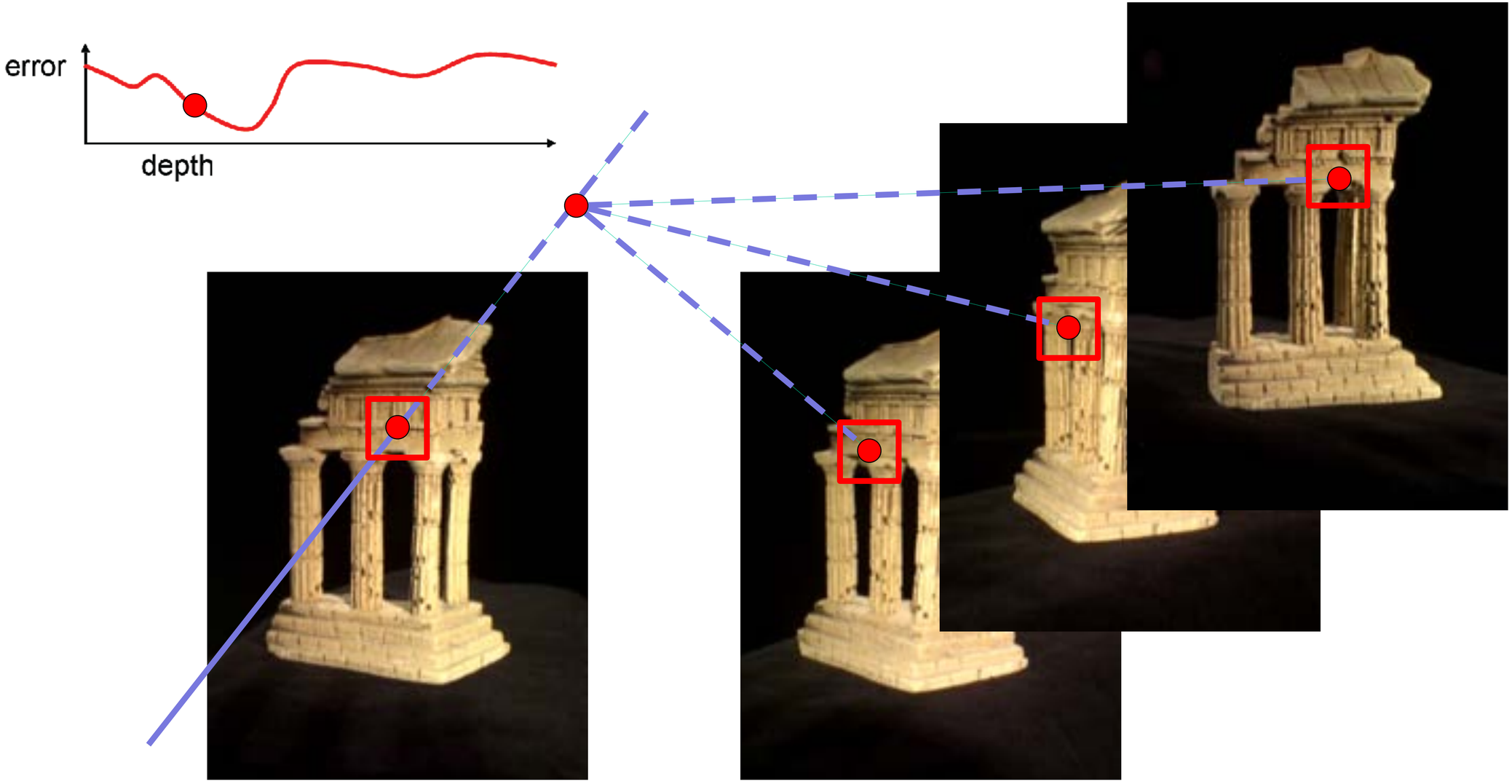
Source: Y. Furukawa

Reconstruction from Photometric Consistency



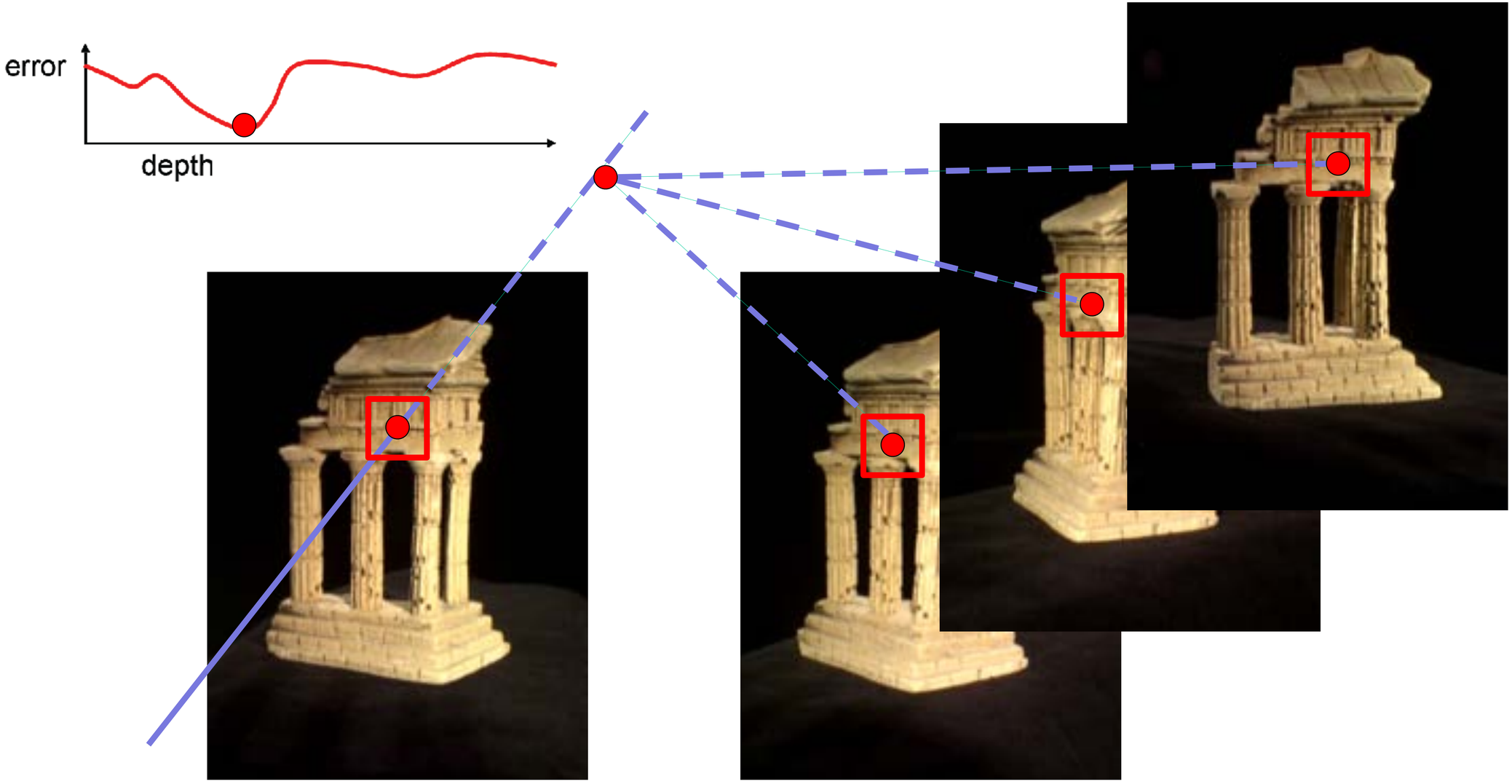
Source: Y. Furukawa

Reconstruction from Photometric Consistency



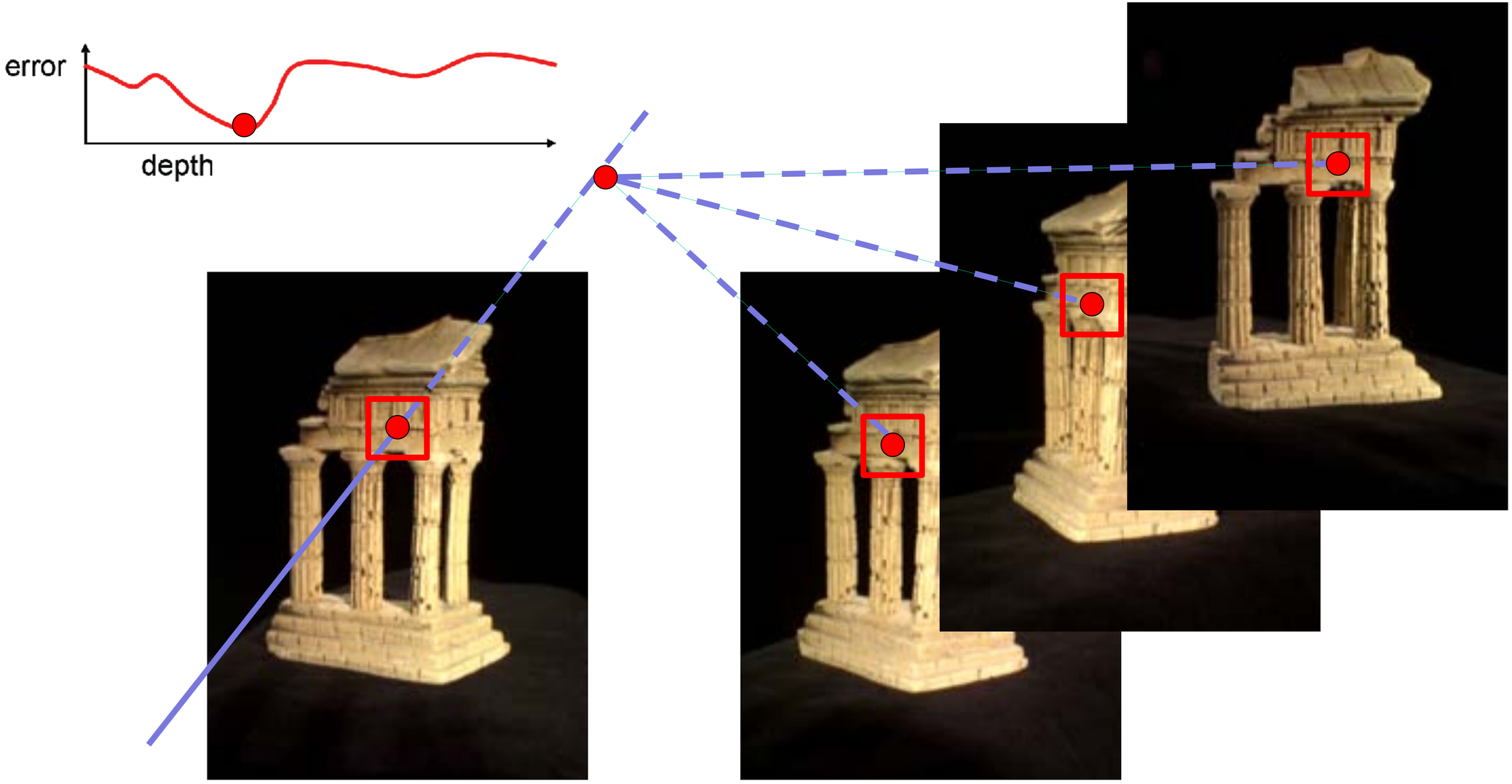
Source: Y. Furukawa

Reconstruction from Photometric Consistency



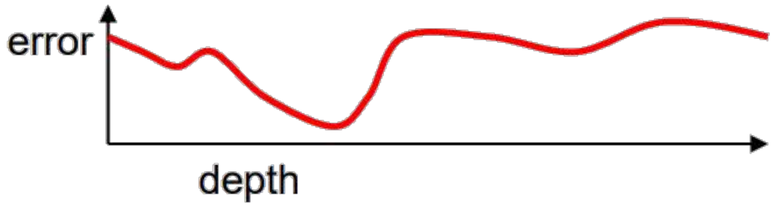
Source: Y. Furukawa

Reconstruction from Photometric Consistency



Source: Y. Furukawa

Reconstruction from Photometric Consistency



Correspondence issue

Triangulation error

Source: Y. Furukawa

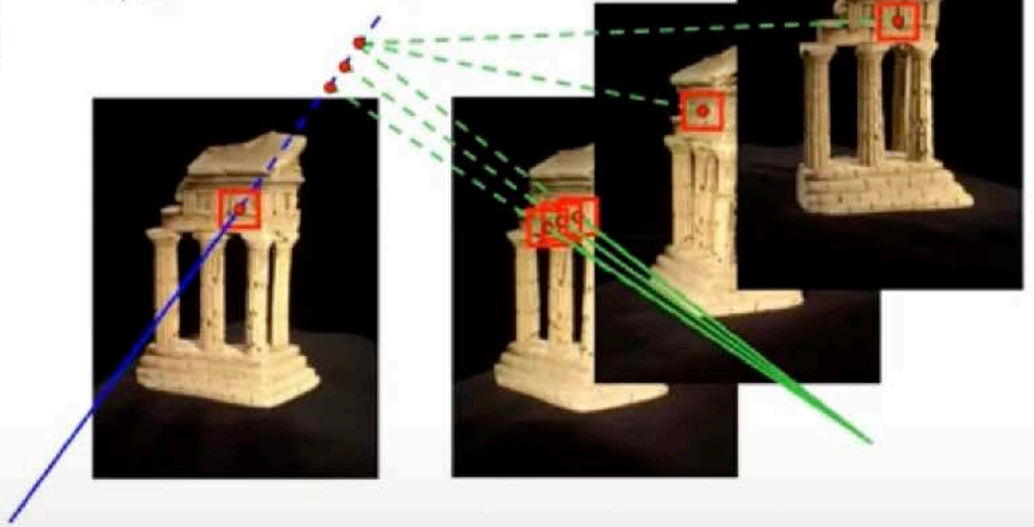
Reconstruction from Photometric Consistency

NCC (Normalized Cross Correlation)

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

SSD (Sum Squared Distance)

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



Pros

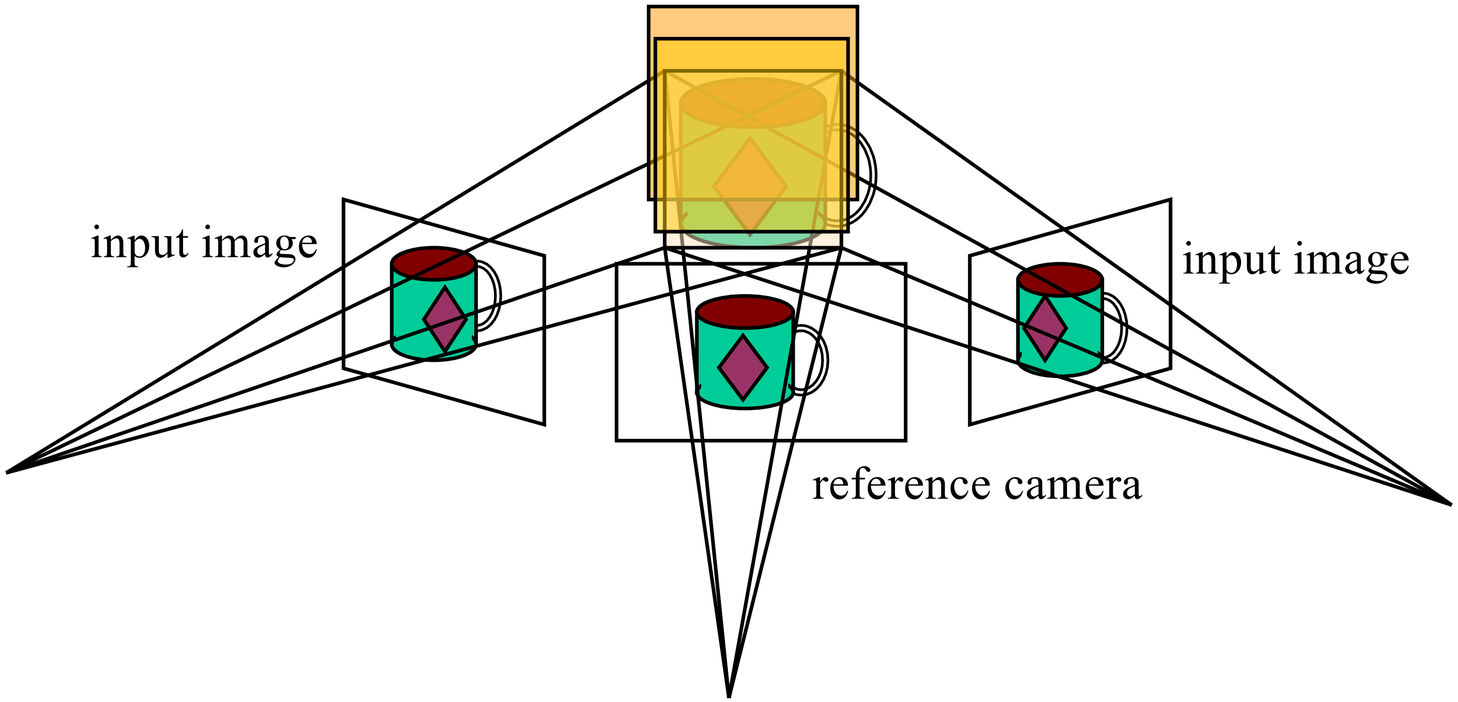
- Accurate
- Concave shape

Cons

- Requires texture
- Sensitive to Non-lambertian area

Source: Y. Furukawa

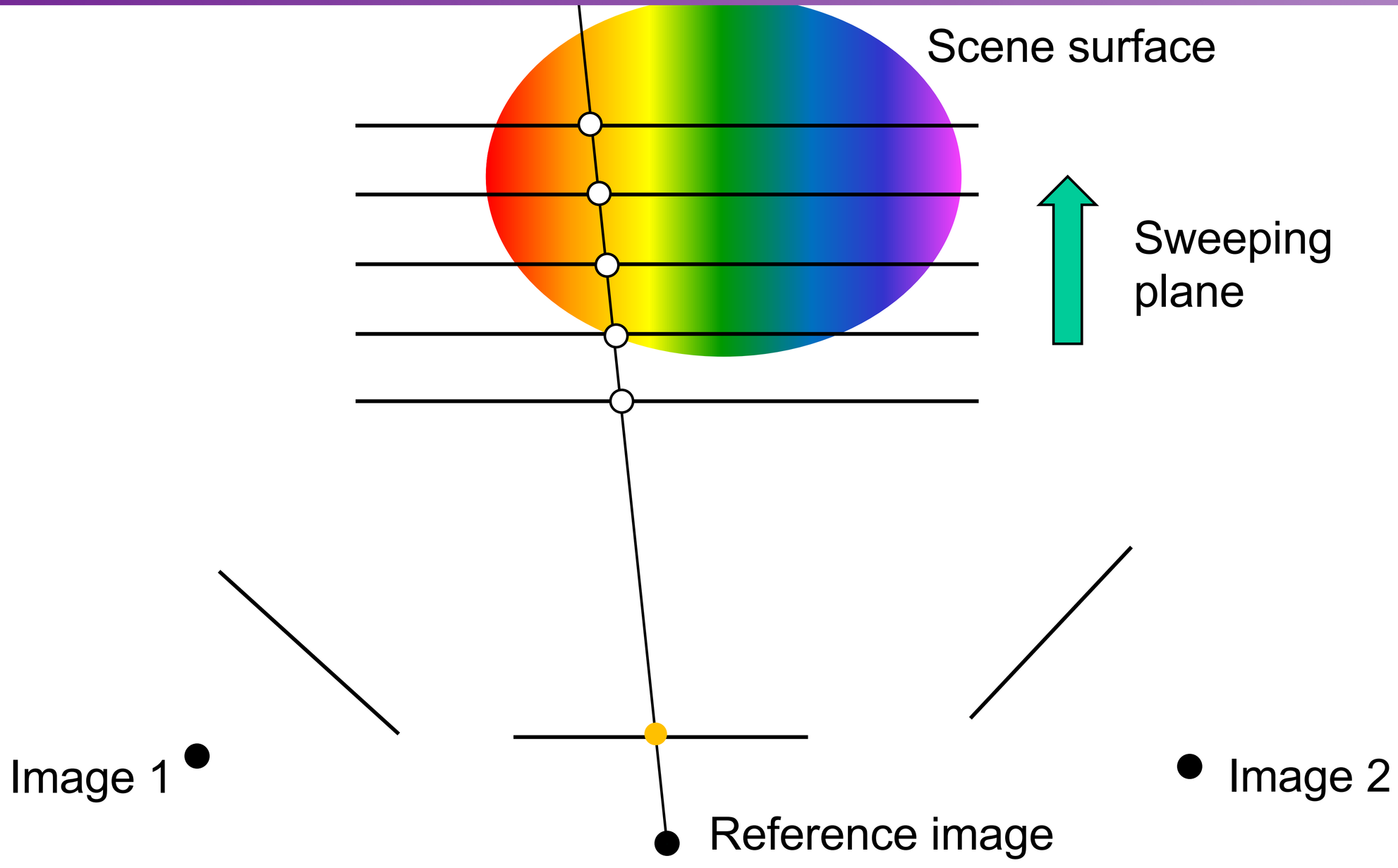
Plane Sweep Stereo



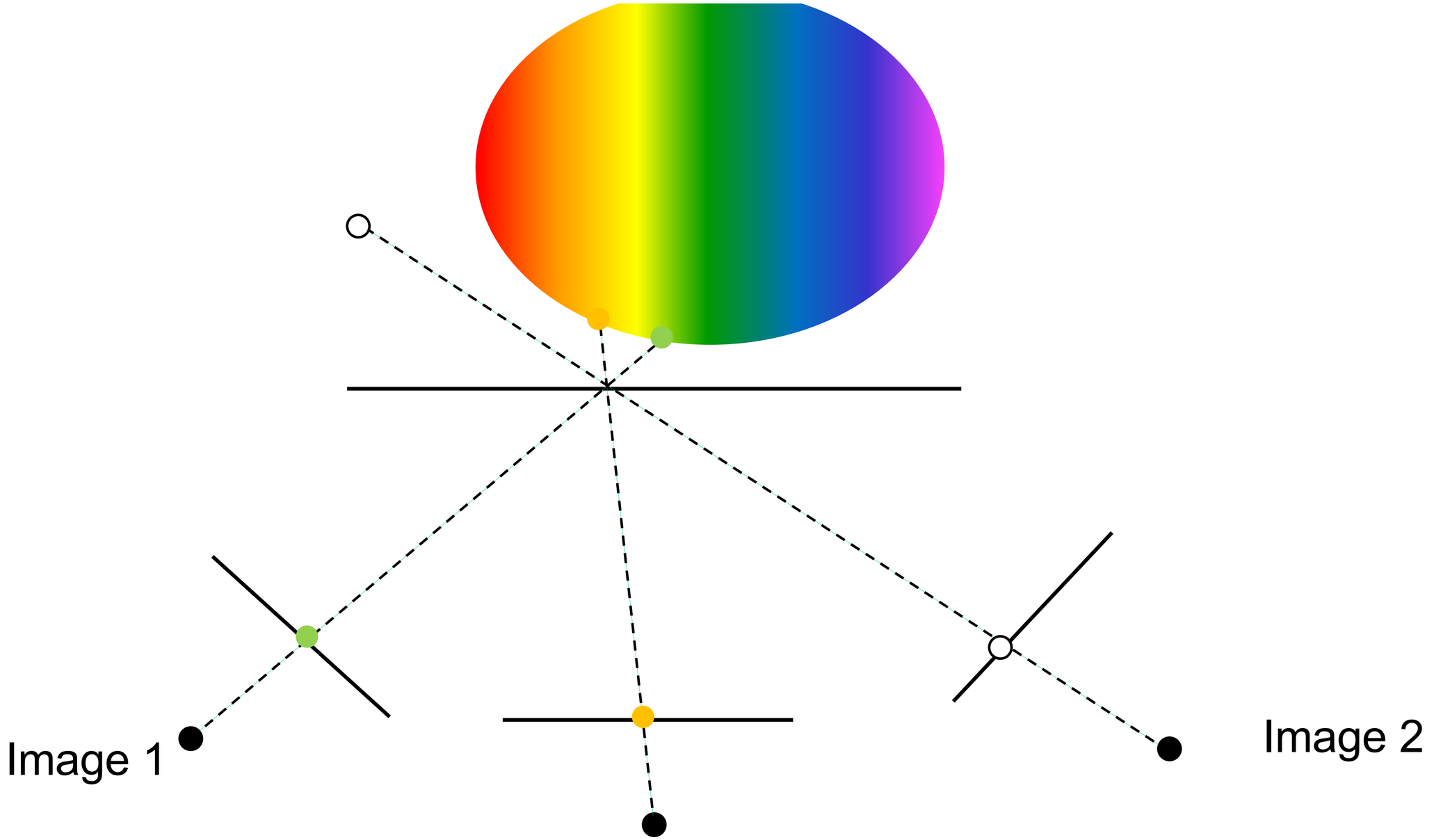
- Sweep plane across a range of depths w.r.t. a reference camera
- For each depth, project each input image onto that plane (homography) and compare the resulting stack of images

R. Collins, A space-sweep approach to true multi-image matching, CVPR 1996

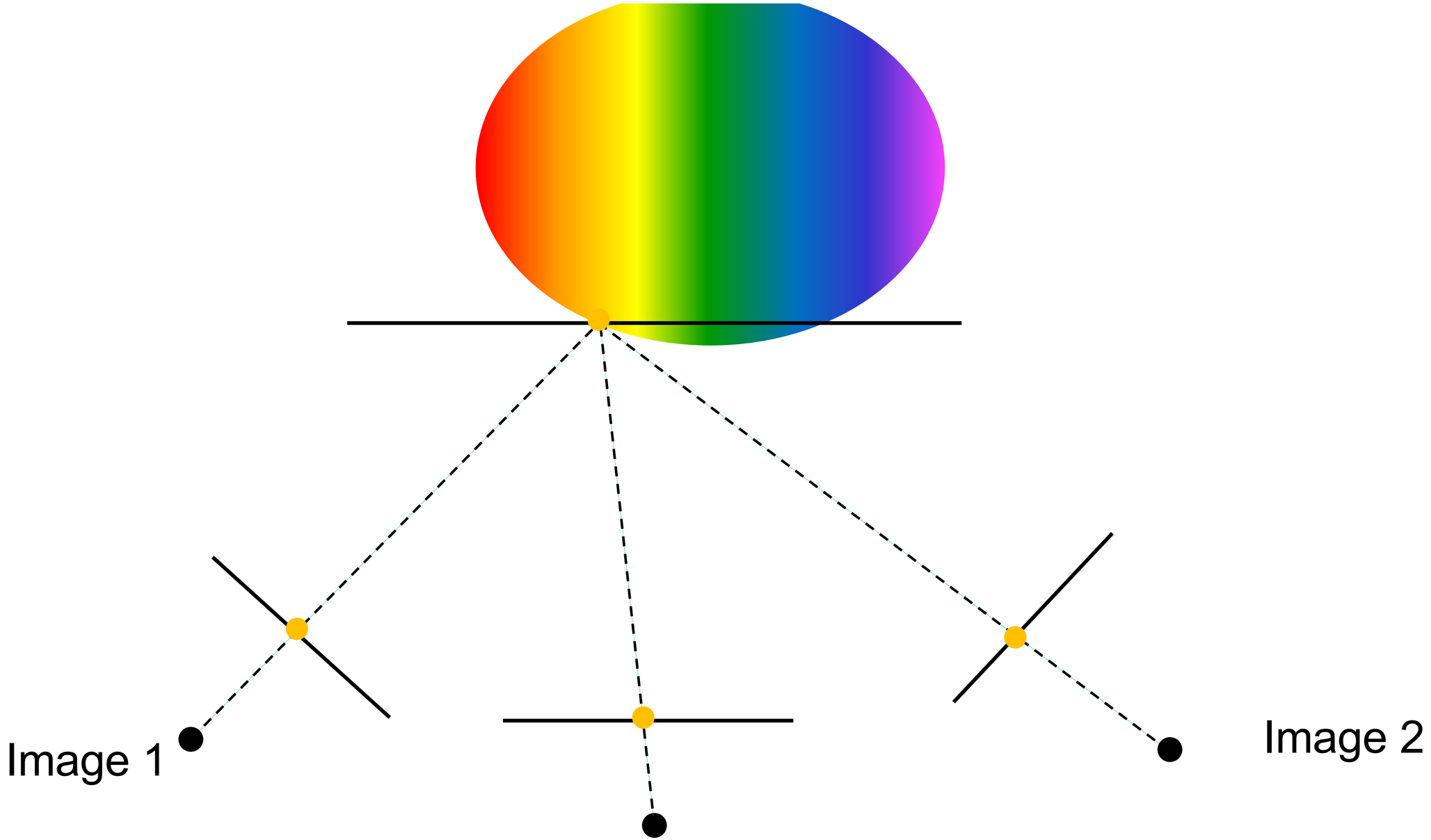
Plane Sweep Stereo



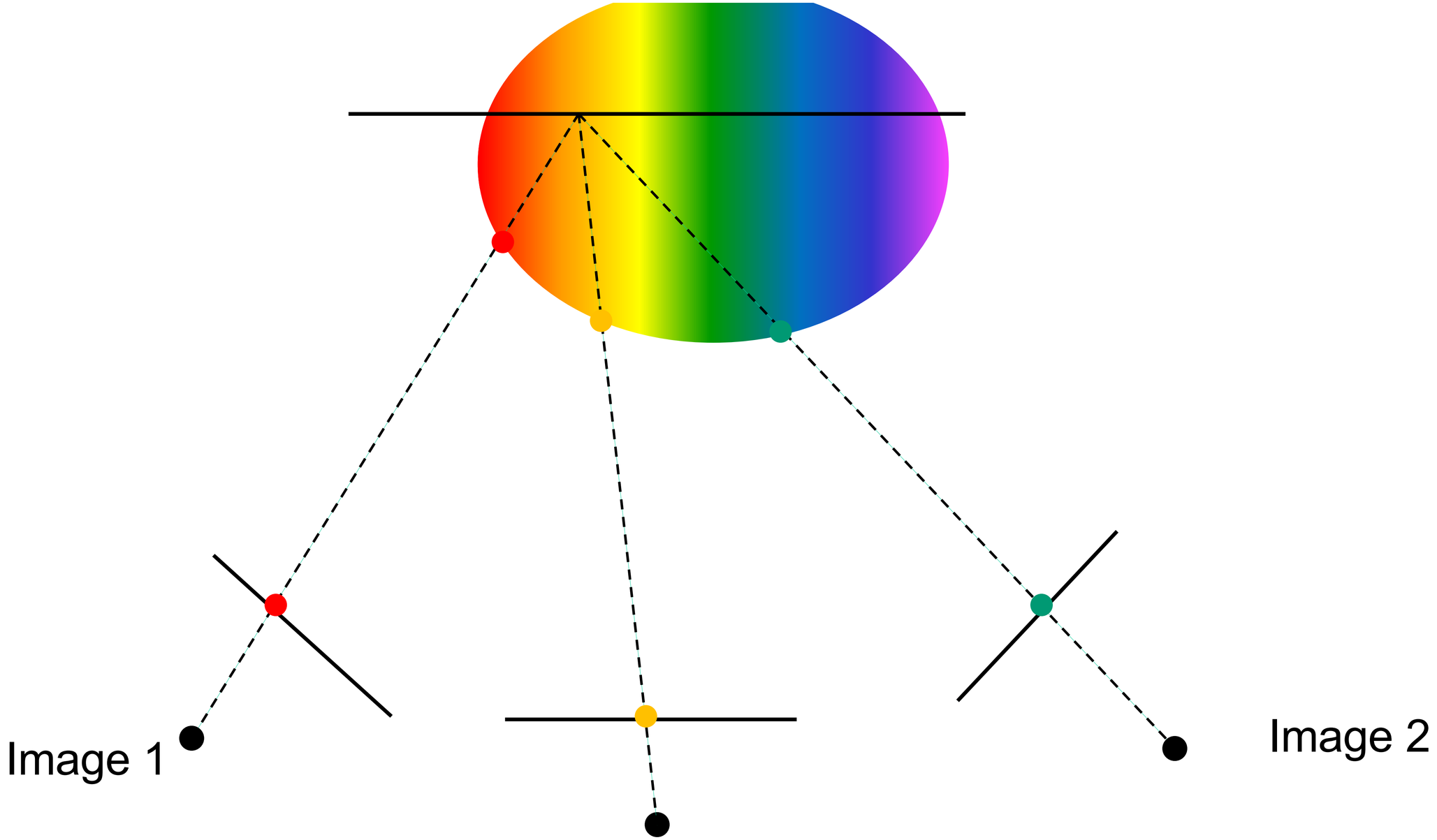
Plane Sweep Stereo



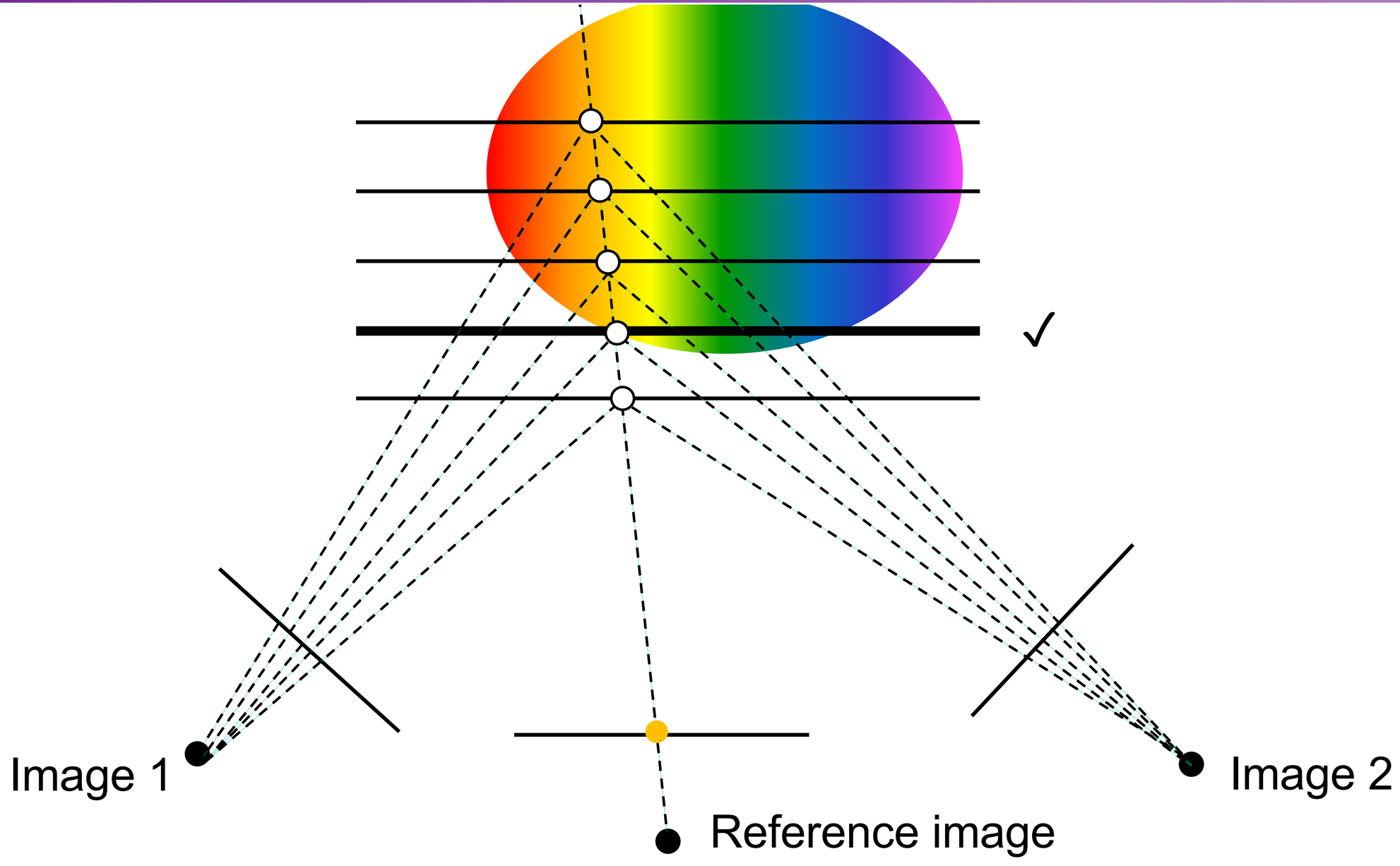
Plane Sweep Stereo



Plane Sweep Stereo



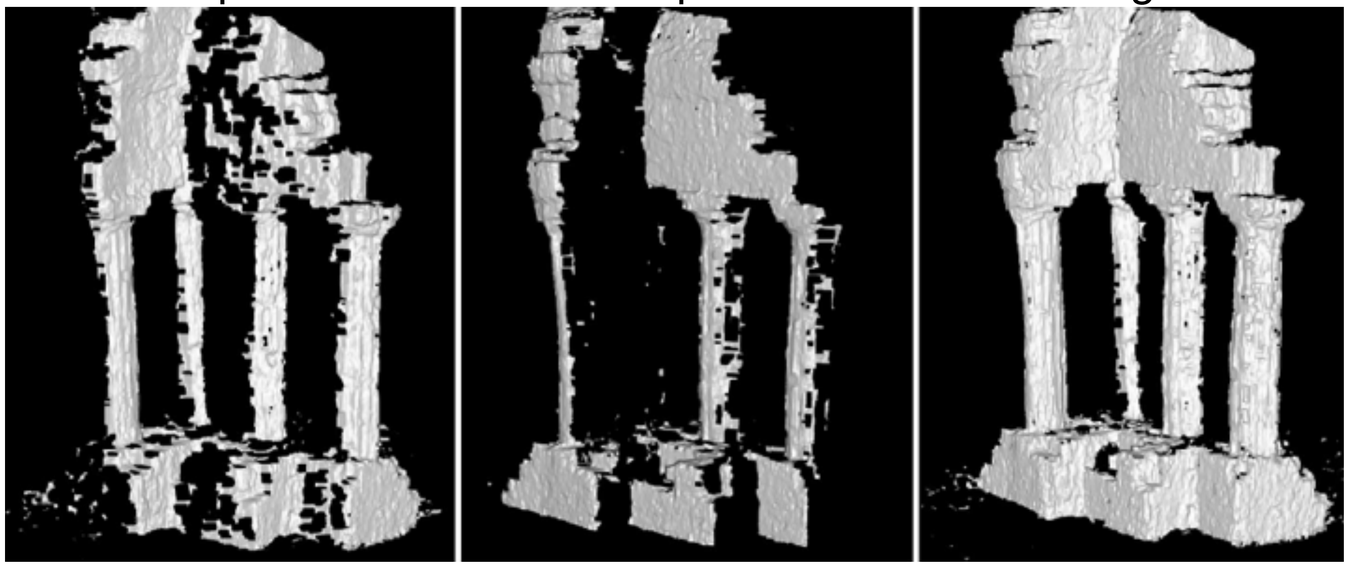
Plane Sweep Stereo



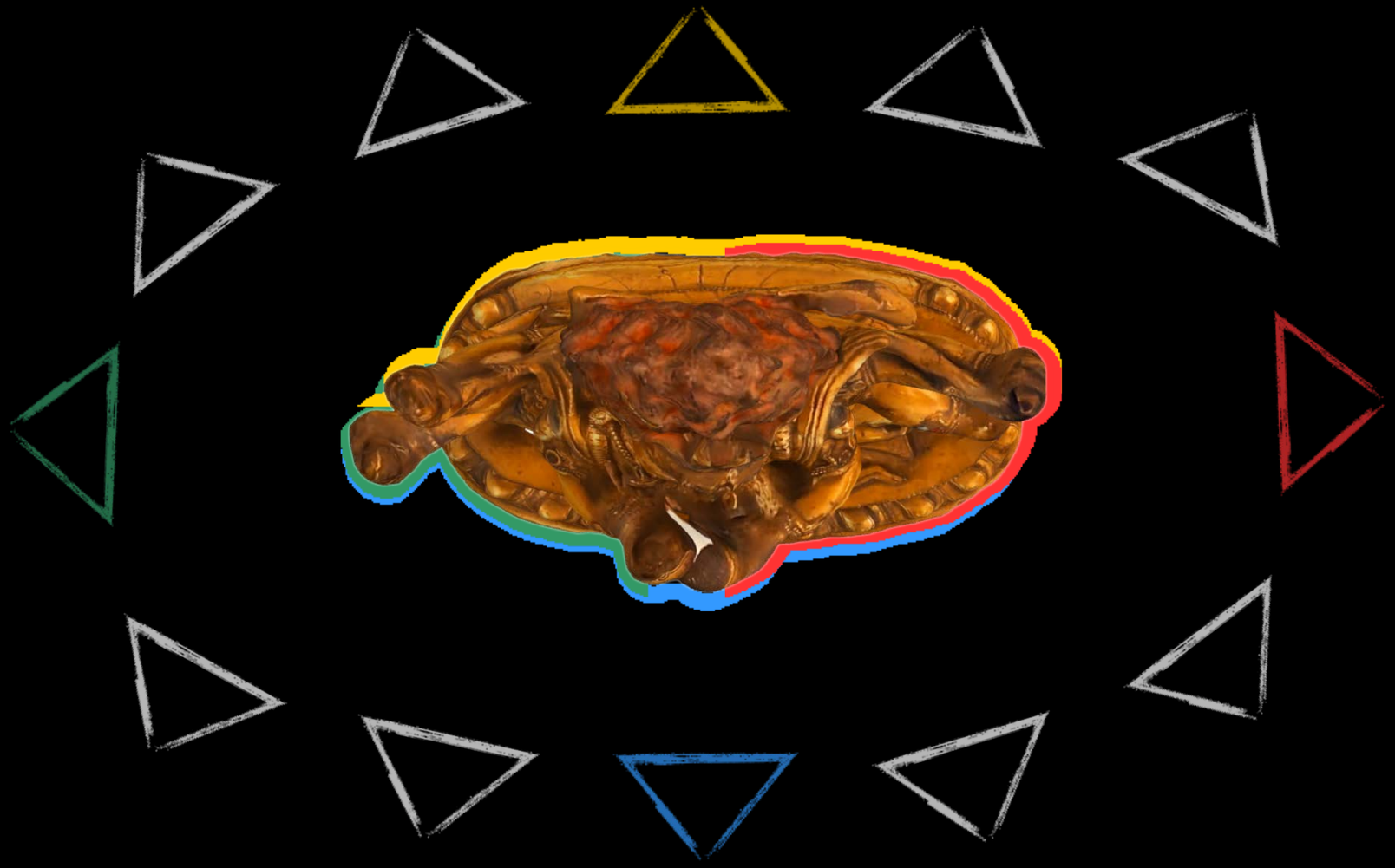
Merge Depth Map

- Given a group of images, compute a depth map using each view as a reference
- Merge multiple depth maps into a volume or a mesh (see, e.g., Curless and Levoy, 1996)

Map 1 Map 2 Merged

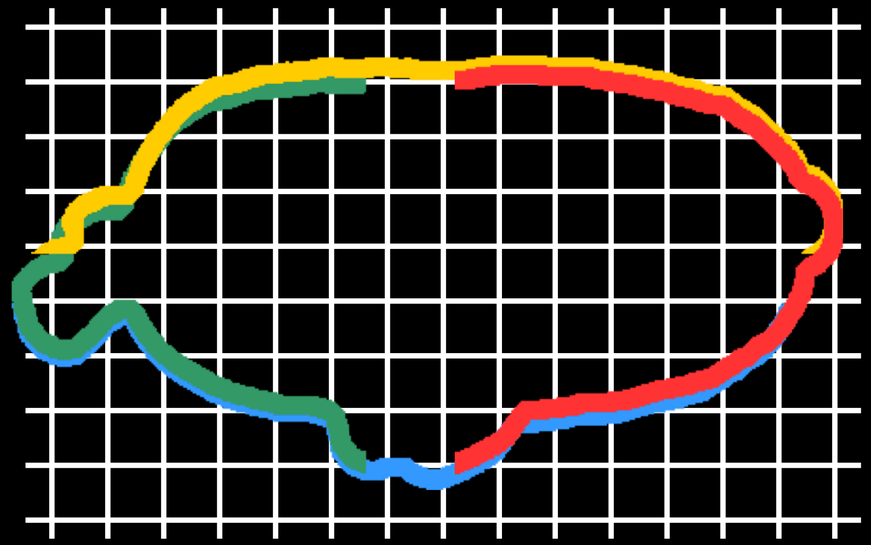


Volumetric Fusion



Source: N. Snavely

Volumetric Fusion

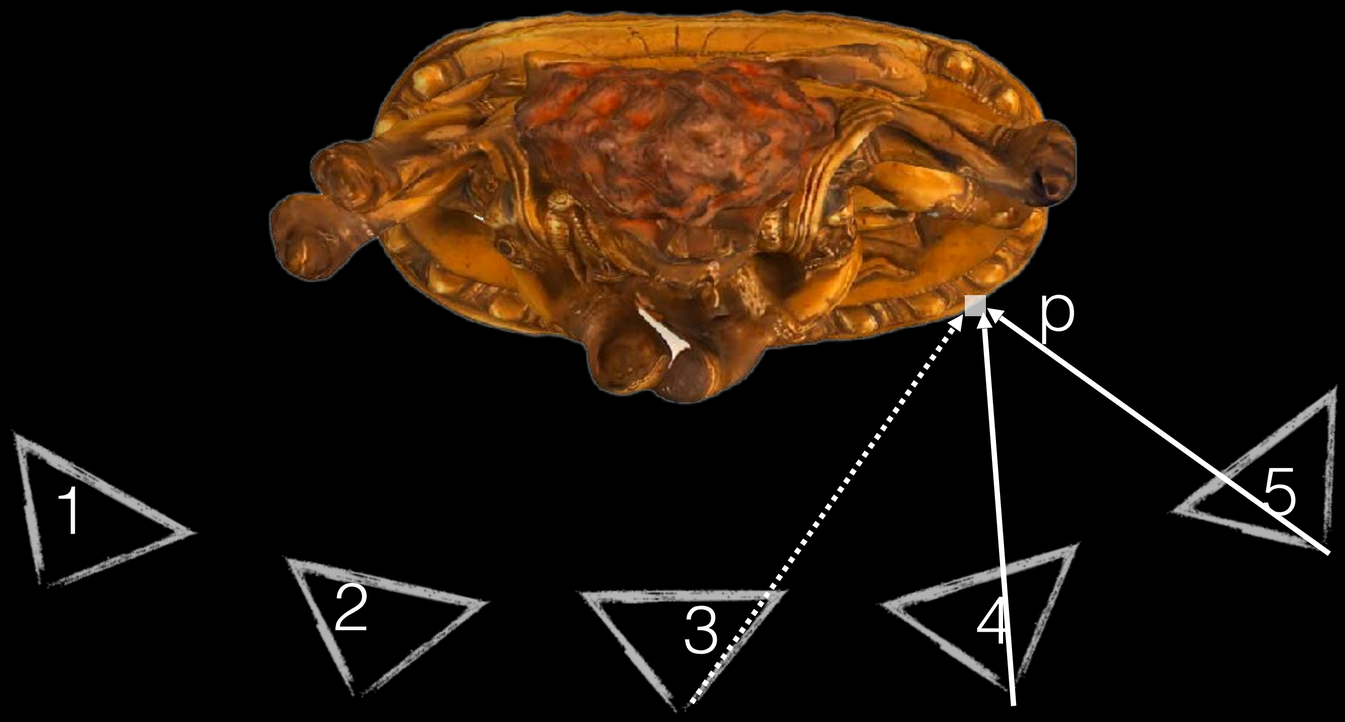


Source: N. Snavely

Multi-View Stereo v.s. Two-View Stereo

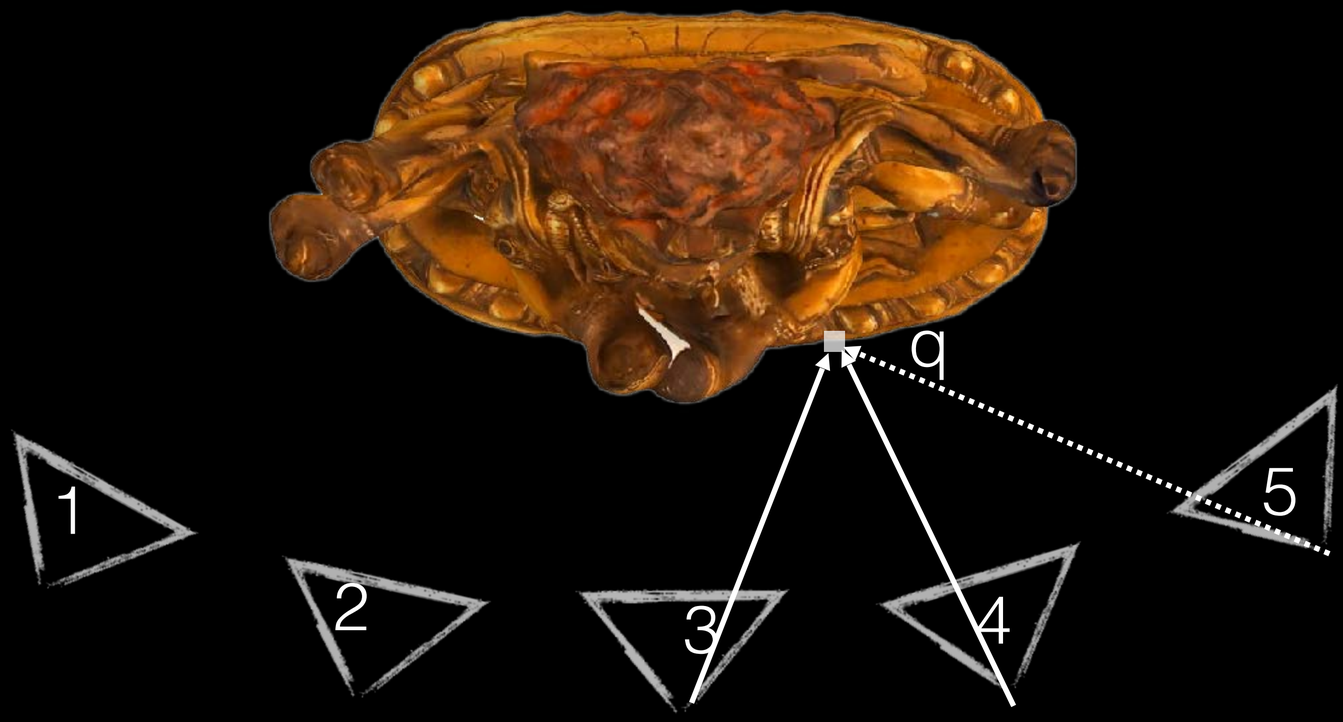
- Different points on the object's surface will be more clearly visible in some subset of cameras
  - Could have high-res closeups of some regions
  - Some surfaces are foreshortened from certain views
  - Some points may be occluded entirely in certain views

Source: N. Snavely



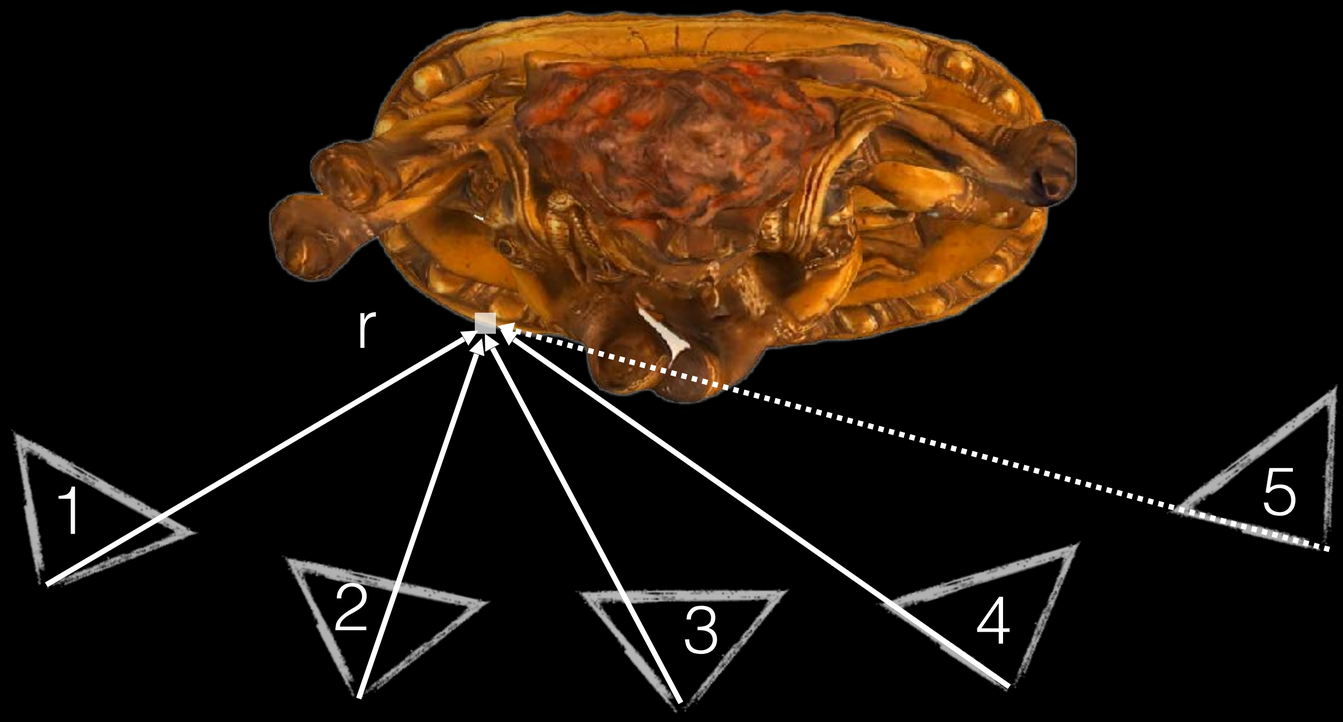
Cameras 4 and 5 can more clearly see point p.

Source: N. Snavely



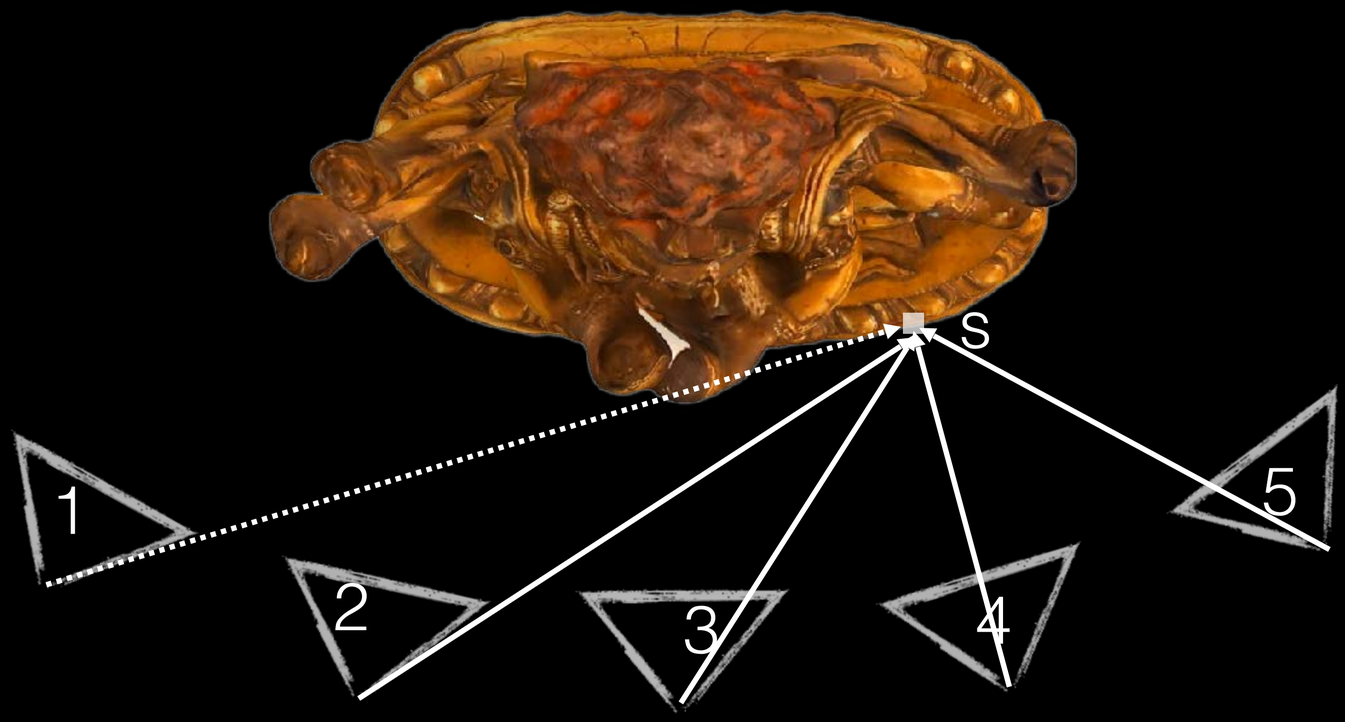
Cameras 3 and 4 can more clearly see point q.

Source: N. Snavely



Camera 5 can’t see point r.

Source: N. Snavely



Camera 1 can’t see point s.

Source: N. Snavely

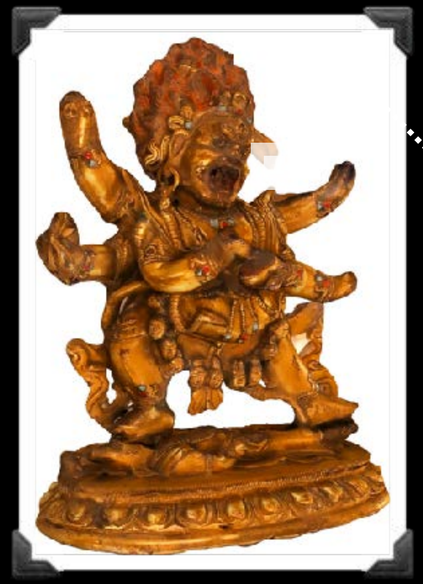
Multi-View Stereo v.s. Two-View Stereo

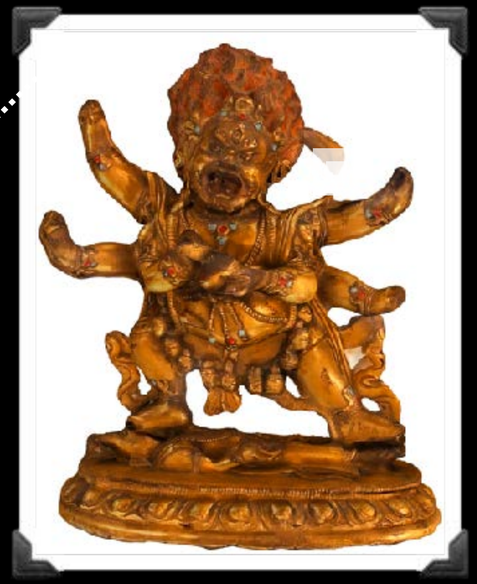
- Different points on the object's surface will be more clearly visible in some subset of cameras
  - Could have high-res closeups of some regions
  - Some surfaces are foreshortened from certain views
  - Some points may be occluded entirely in certain views
- More measurements per point can reduce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

Source: N. Snavely

Source: N. Snavely

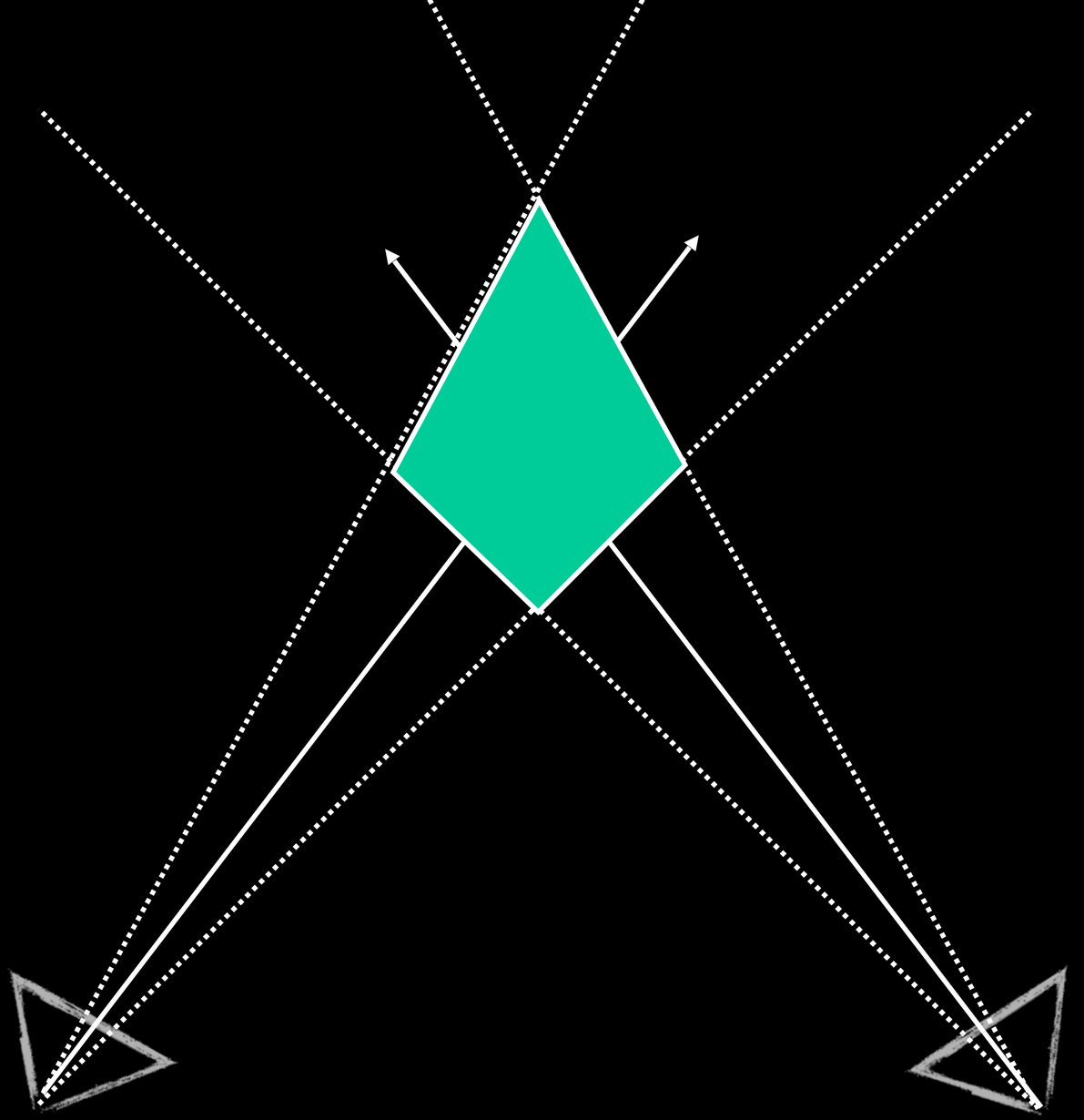
Source: N. Snavely





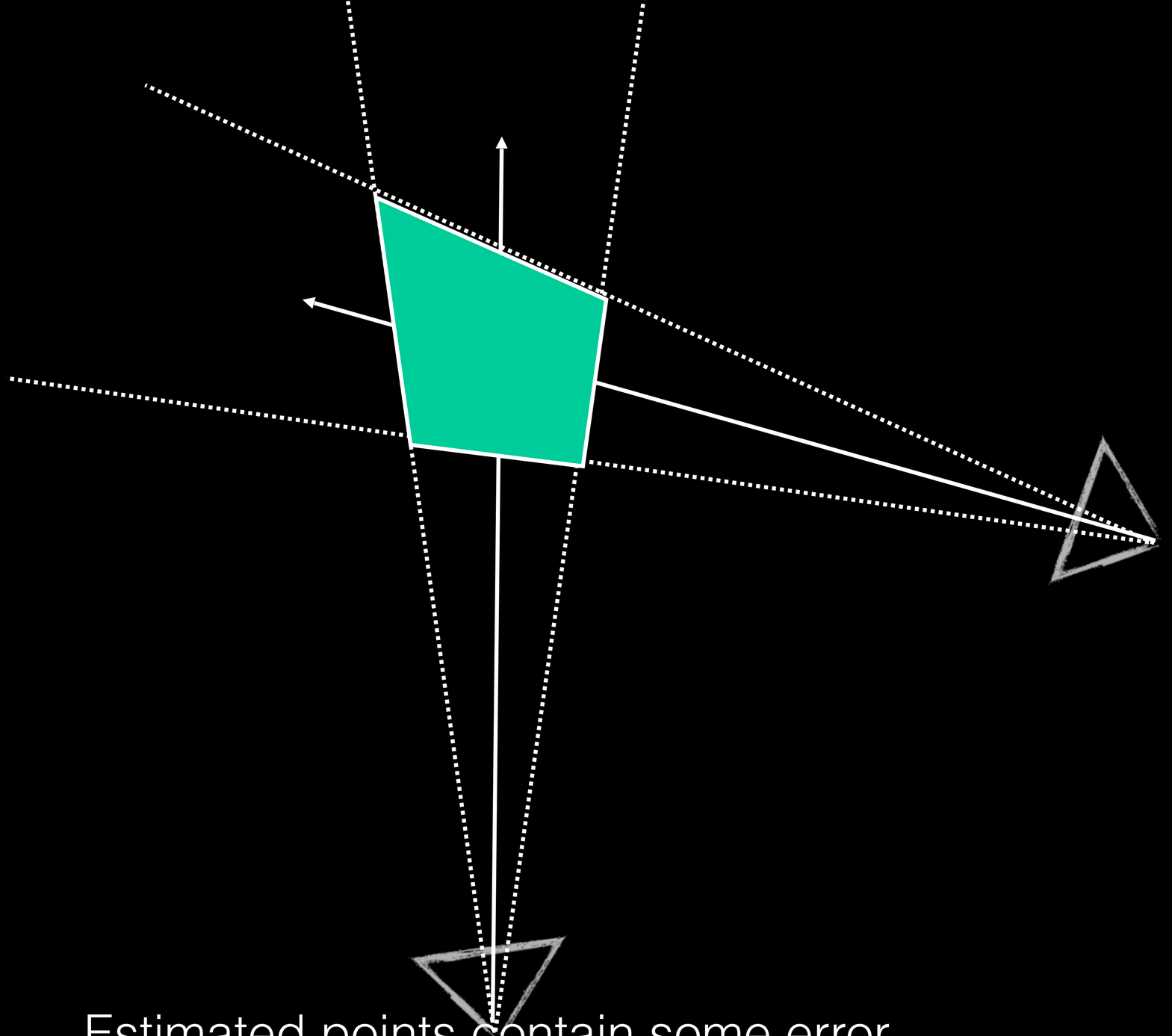
Estimated points contain some error.

Source: N. Snavely



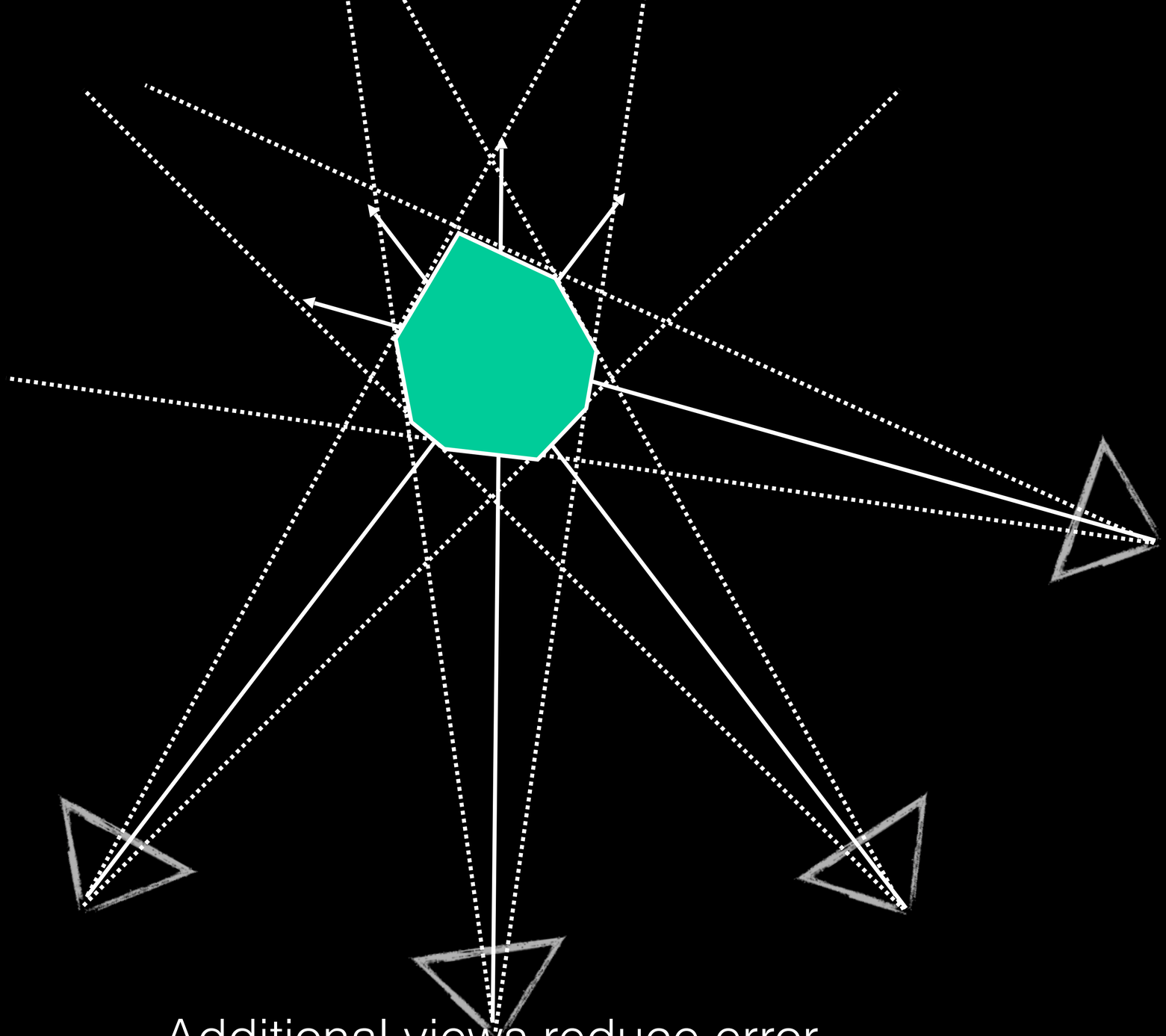
Estimated points contain some error.

Source: N. Snavely



Estimated points contain some error.

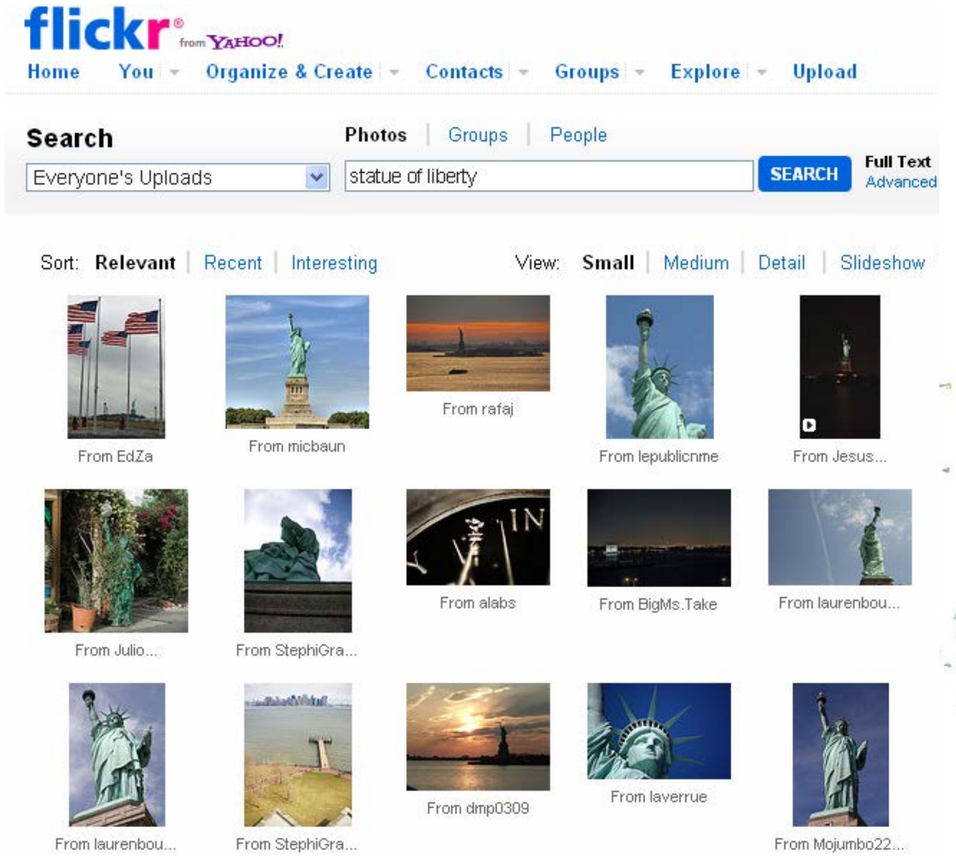
Source: N. Snavely

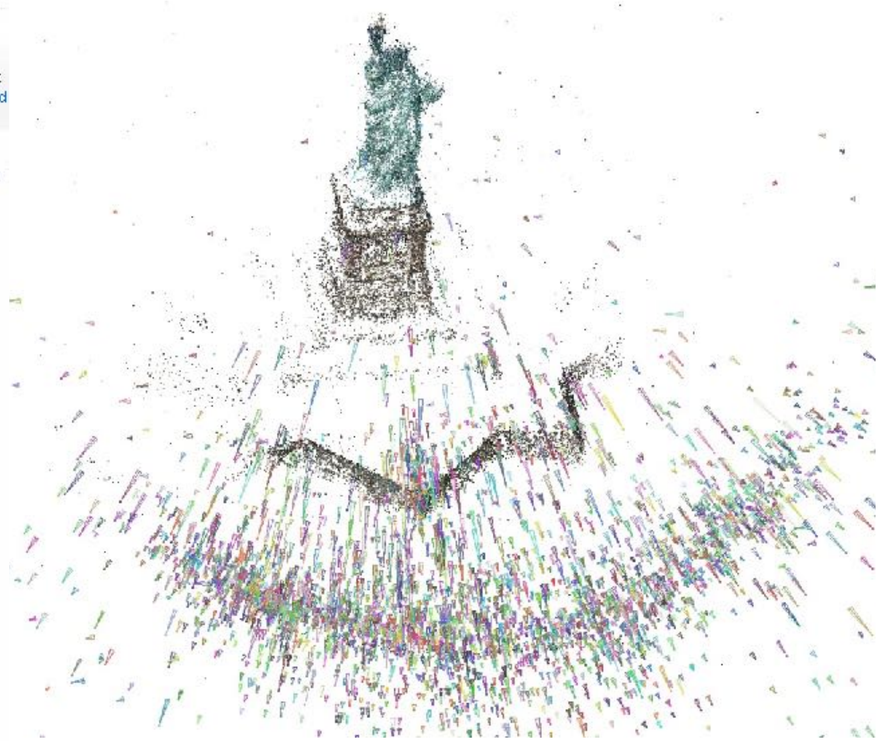


Additional views reduce error.

Source: N. Snavely

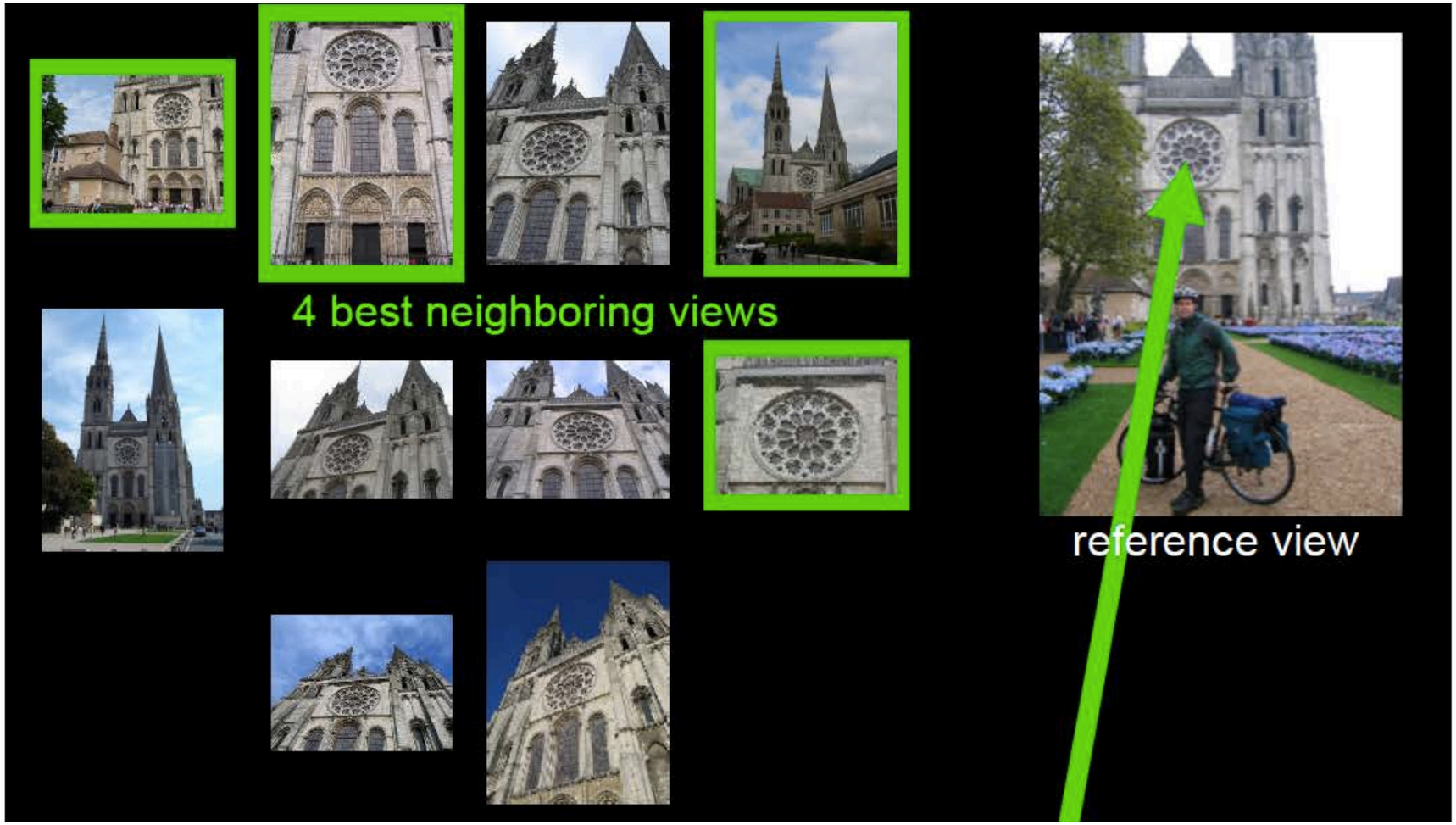
Stereo from Community Photo Collections





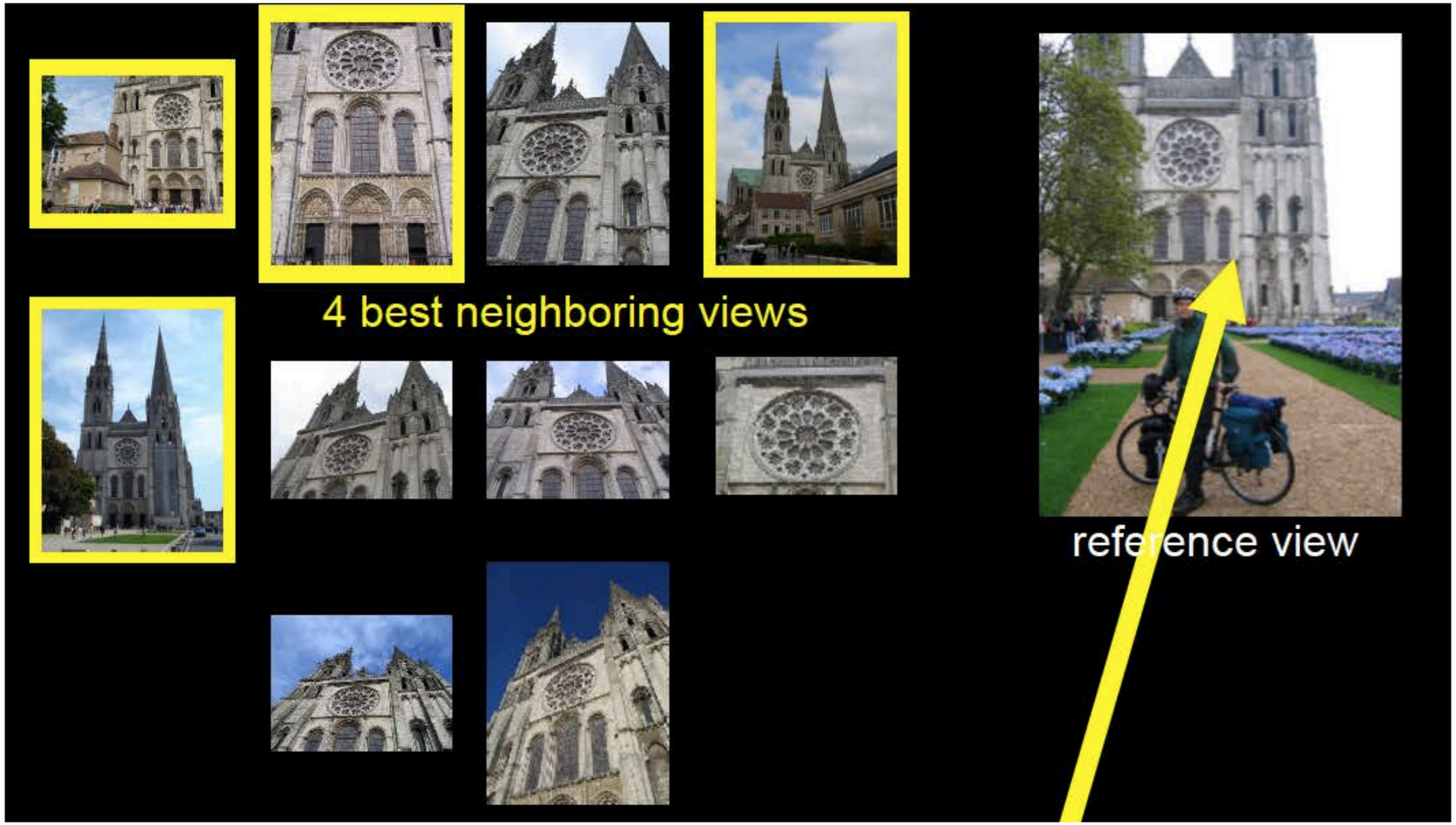
- Need *structure from motion* to recover unknown camera parameters
- Need *view selection* to find good groups of images on which to run dense stereo

Local View Selection



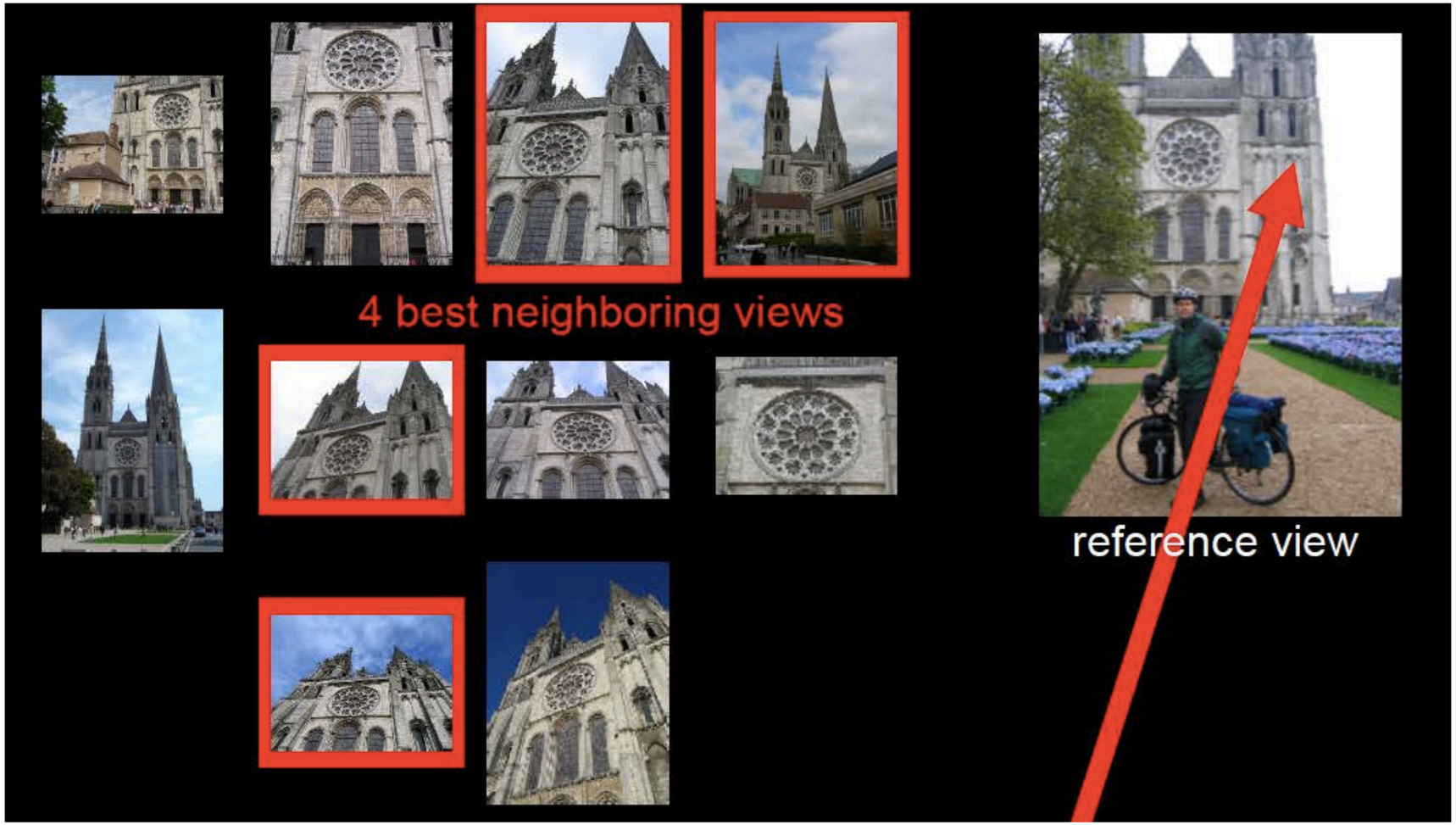
M. Goesele et al., Multi-View Stereo for Community Photo Collections, ICCV 2007

Local View Selection



M. Goesele et al., Multi-View Stereo for Community Photo Collections, ICCV 2007

Local View Selection

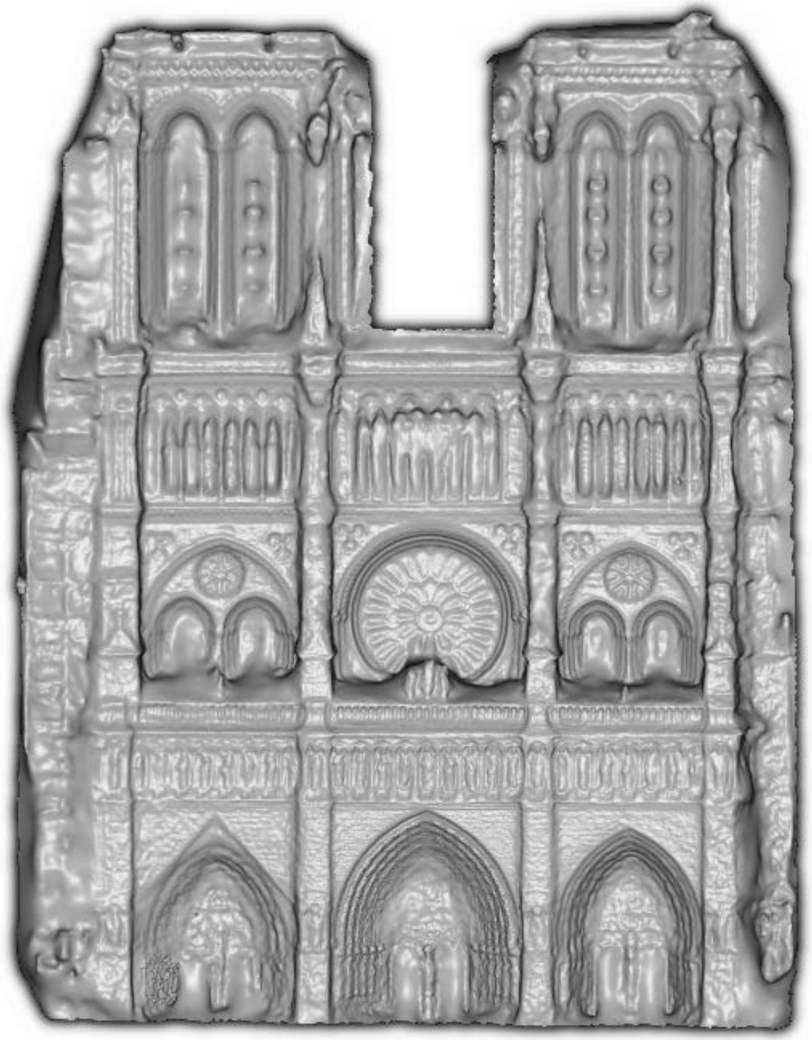


M. Goesele et al., Multi-View Stereo for Community Photo Collections, ICCV 2007

Local View Selection

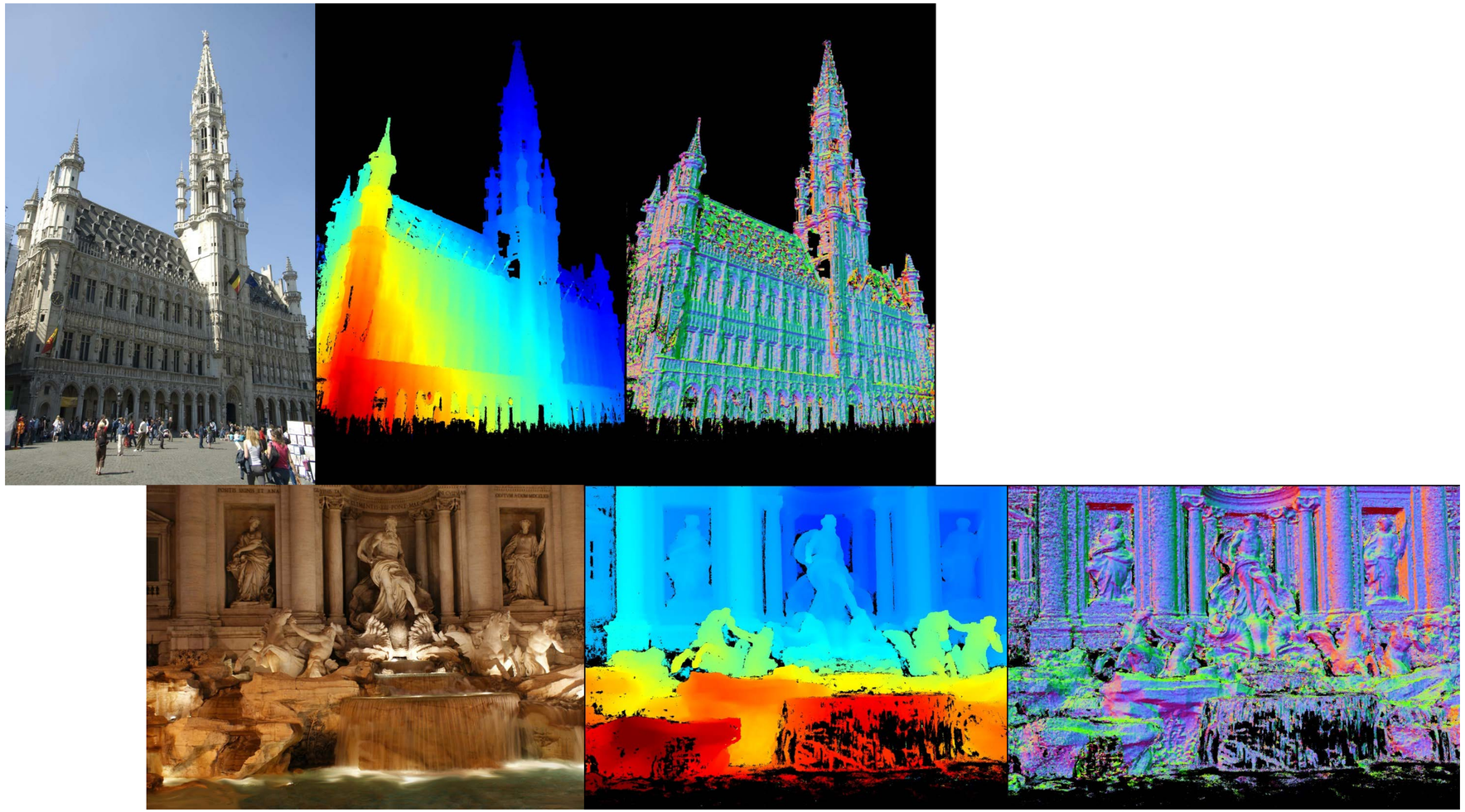
Notre Dame de Paris

653 images
313 photographers



M. Goesele et al., Multi-View Stereo for Community Photo Collections, ICCV 2007

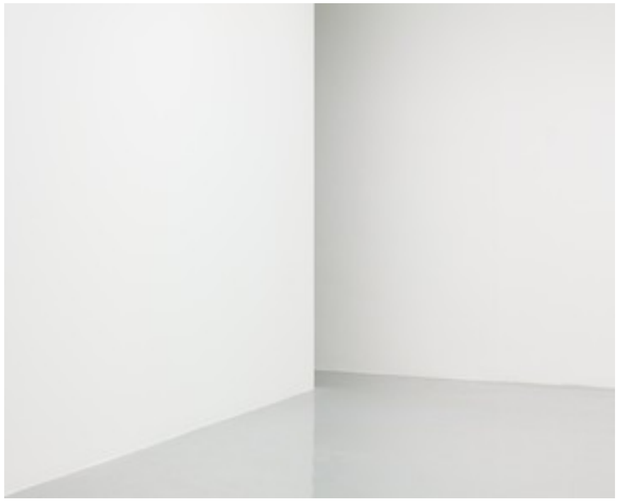
Stereo from Community Photo Collections



https://colmap.github.io/

Schönberger, Johannes L., Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. "Pixelwise view selection for unstructured multi-view stereo." In European Conference on Computer Vision, pp. 501-518. Springer, Cham, 2016.

Limitations of Classical MVS



Textureless Area

Reflection /Transparency

Repetitive patterns

Yi Ma, An Introduction to Holistic 3D Reconstruction, ICCV 2019

Learning-Based MVS

Learned feature → more robust matching

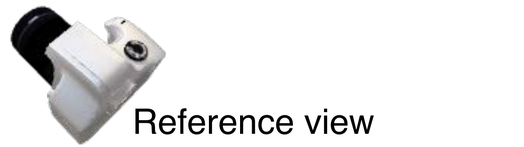
Shape prior → more complete reconstruction

Slides credit: Hao Su

Outline

- Introduction to multi-view stereo (MVS)
- Classic MVS
- ***Learning-based MVS: a first pipeline***
- Learning-based MVS: Improvements
  - Adaptive Space Sampling
  - Depth-Normal Consistency

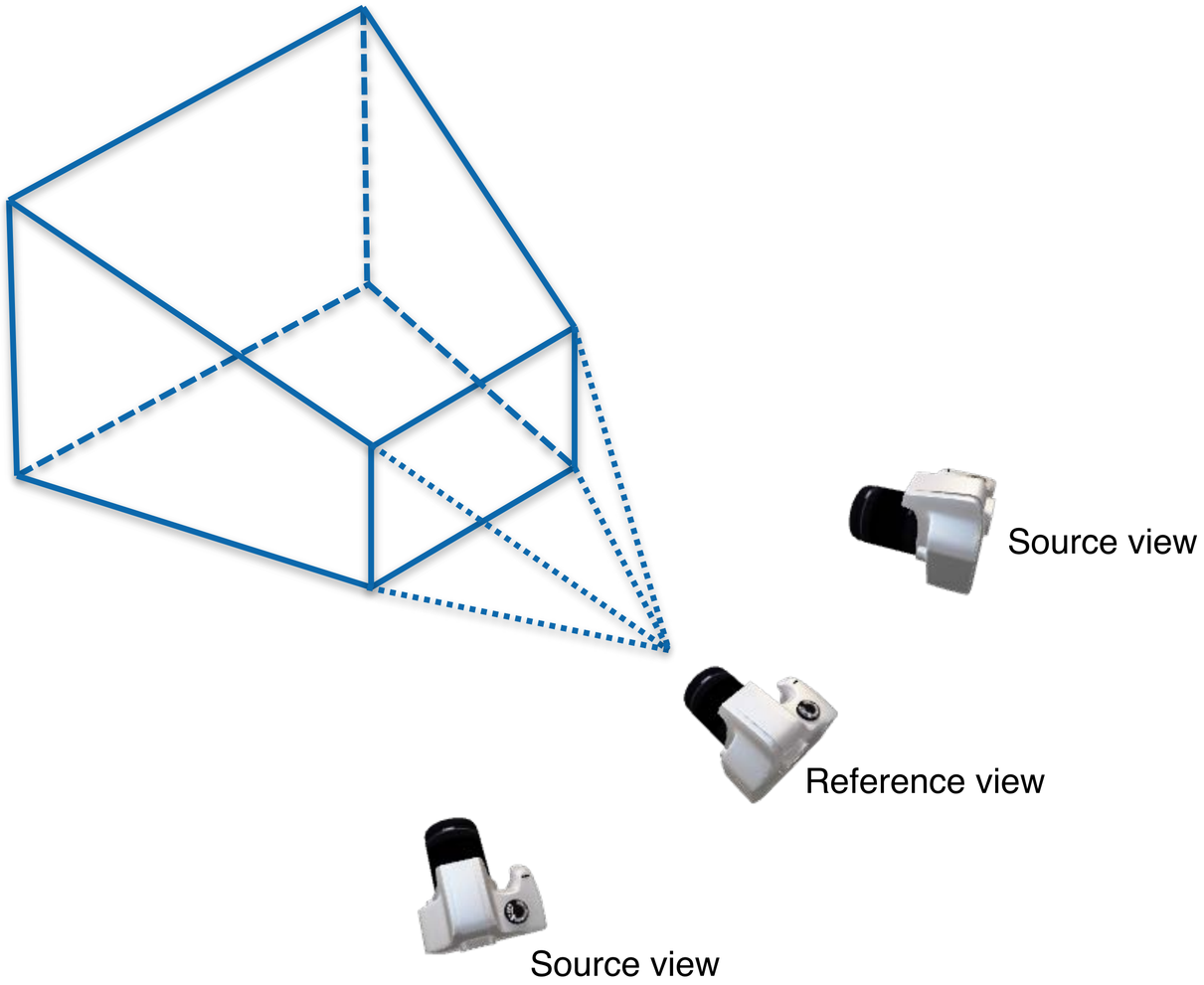
Volumetric Stereo



Slides credit: Hao Su

Volumetric Stereo

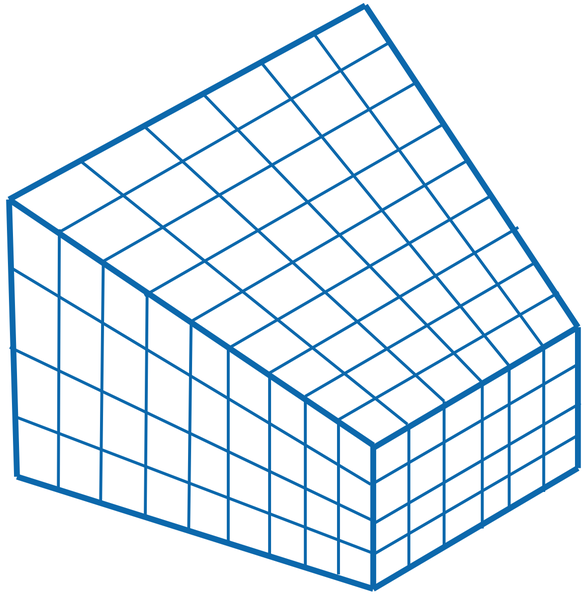
Reference view
frustum



Slides credit: Hao Su

Volumetric Stereo

Reference view
frustum voxelization



Source view

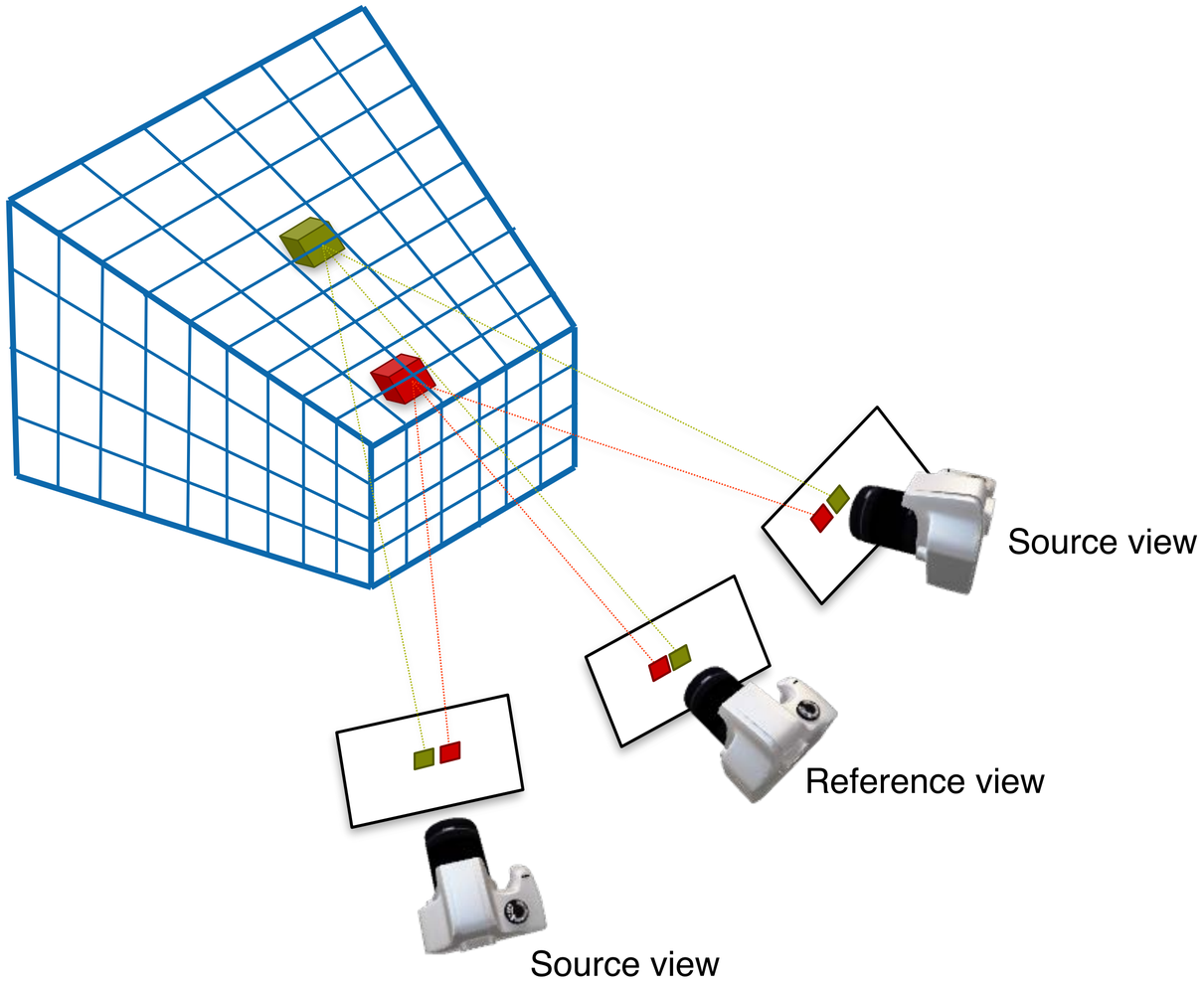
Reference view

Source view

Slides credit: Hao Su

Volumetric Stereo

Image feature warping

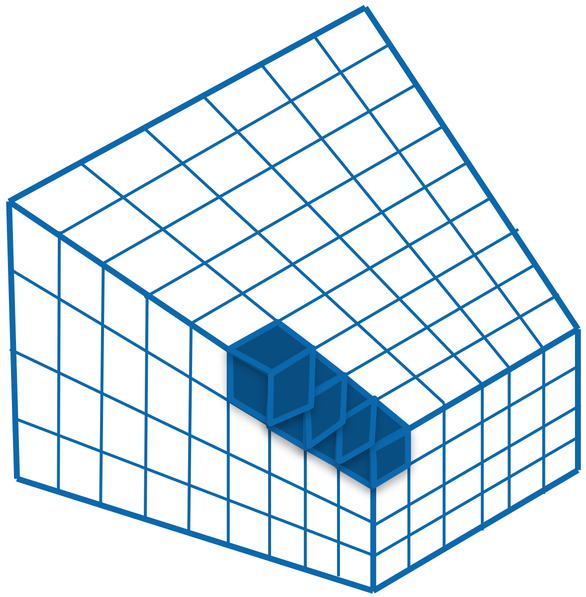


How to handle frustum difference?

Slides credit: Hao Su

Volumetric Stereo

3D CNNs



Source view

Reference view

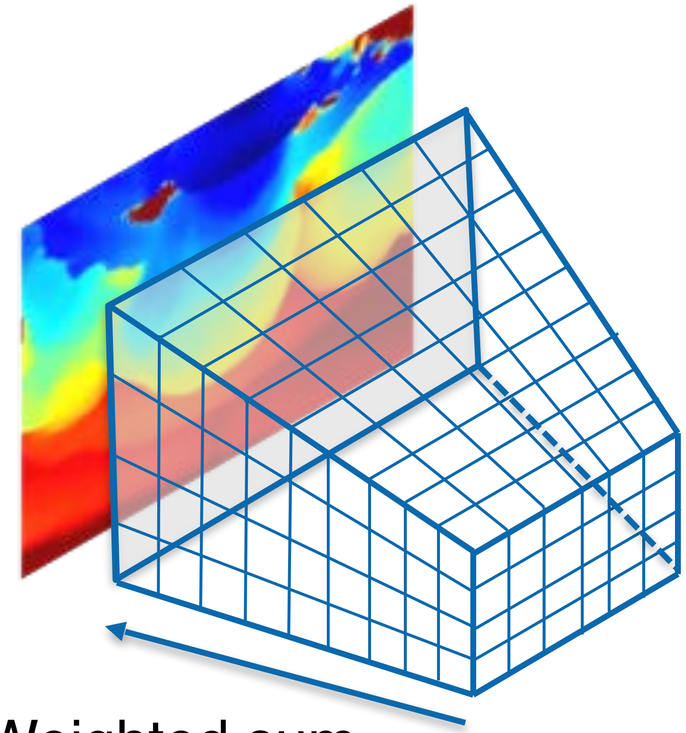
Source view

To suppress noise due to factors violating the consistency assumption
(e.g., non-Lambertian surfaces or object occlusions)

Slides credit: Hao Su

Volumetric Stereo

Reference view
depth prediction



Winner-take-all is non-differentiable

Weighted sum
along view light

Source view

Reference view

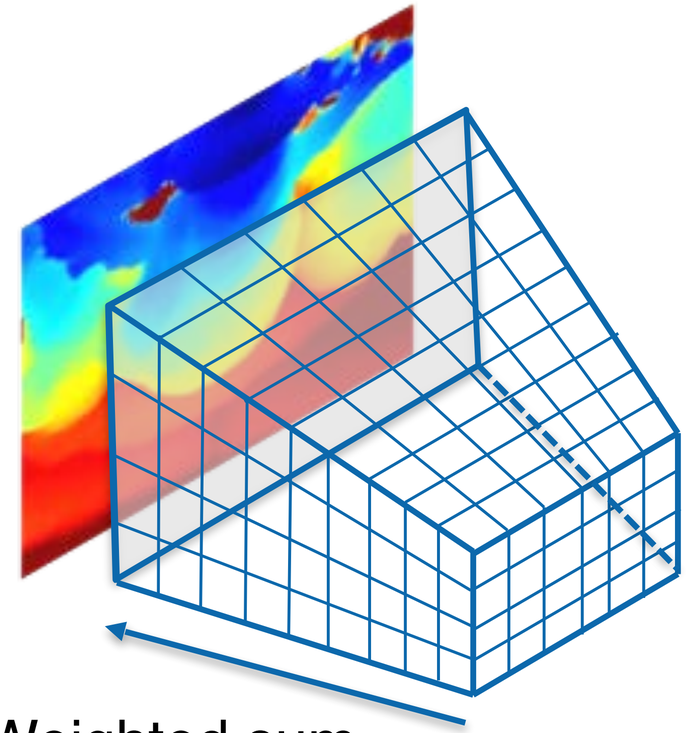
$$\mathbf{D} = \sum_{d=d_{min}}^{d_{max}} d \times \mathbf{P}(d)$$

Source view

Slides credit: Hao Su

Volumetric Stereo

Reference view
depth prediction



$$Loss = \sum_{p \in \mathbf{P}_{\text{valid}}} \left\| d(p) - \hat{d}(p) \right\|_1$$

Valid pixels — GT depth — Depth prediction

Need GT depth from either
reference view or other views

Weighted sum
along view light

$$\mathbf{D} = \sum_{d=d_{min}}^{d_{max}} d \times \mathbf{P}(d)$$

Source view

Reference view

Source view

Slides credit: Hao Su

Pipeline Summary

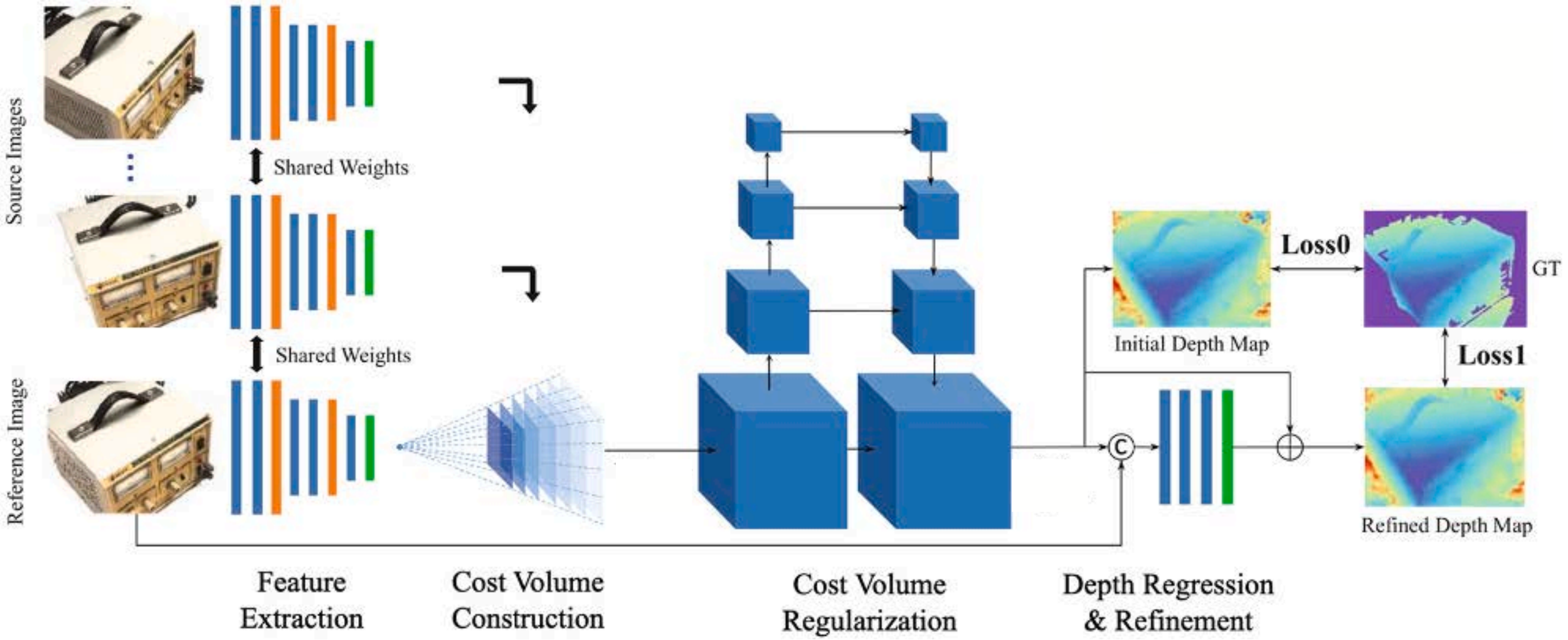


Image source: MVSNet

Issues

- Issues
  - Quality and speed tradeoff
  - Flying points when there is abrupt depth change

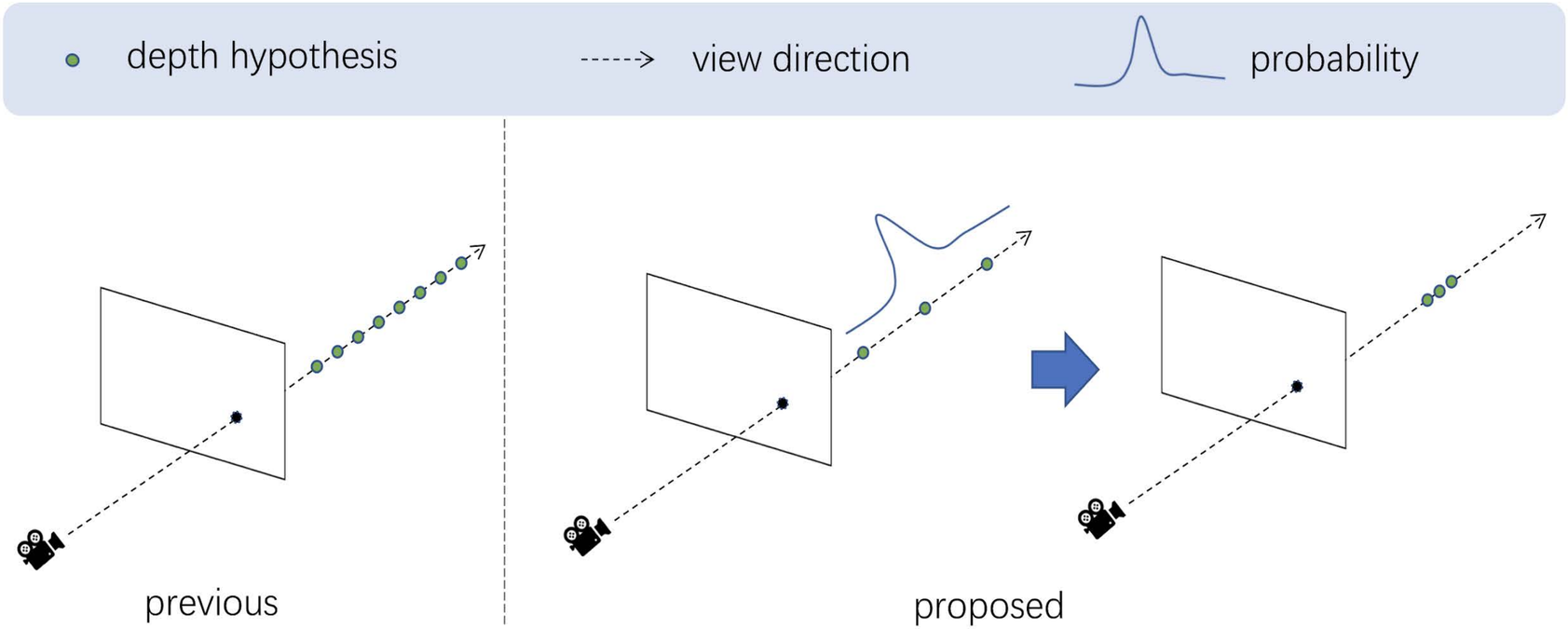
Issues

- Issues
  - Quality and speed tradeoff
  - Flying points when there is abrupt depth change
- Possible solutions
  - Depth estimation following a coarse to fine strategy
  - Stronger loss function regularizing flying points

Outline

- Introduction to multi-view stereo (MVS)
- Classic MVS
- Learning-based MVS: a first pipeline
- Learning-based MVS: Improvements
  - ***Adaptive Space Sampling***
  - Depth-Normal Consistency

Coarse-to-fine Sampling

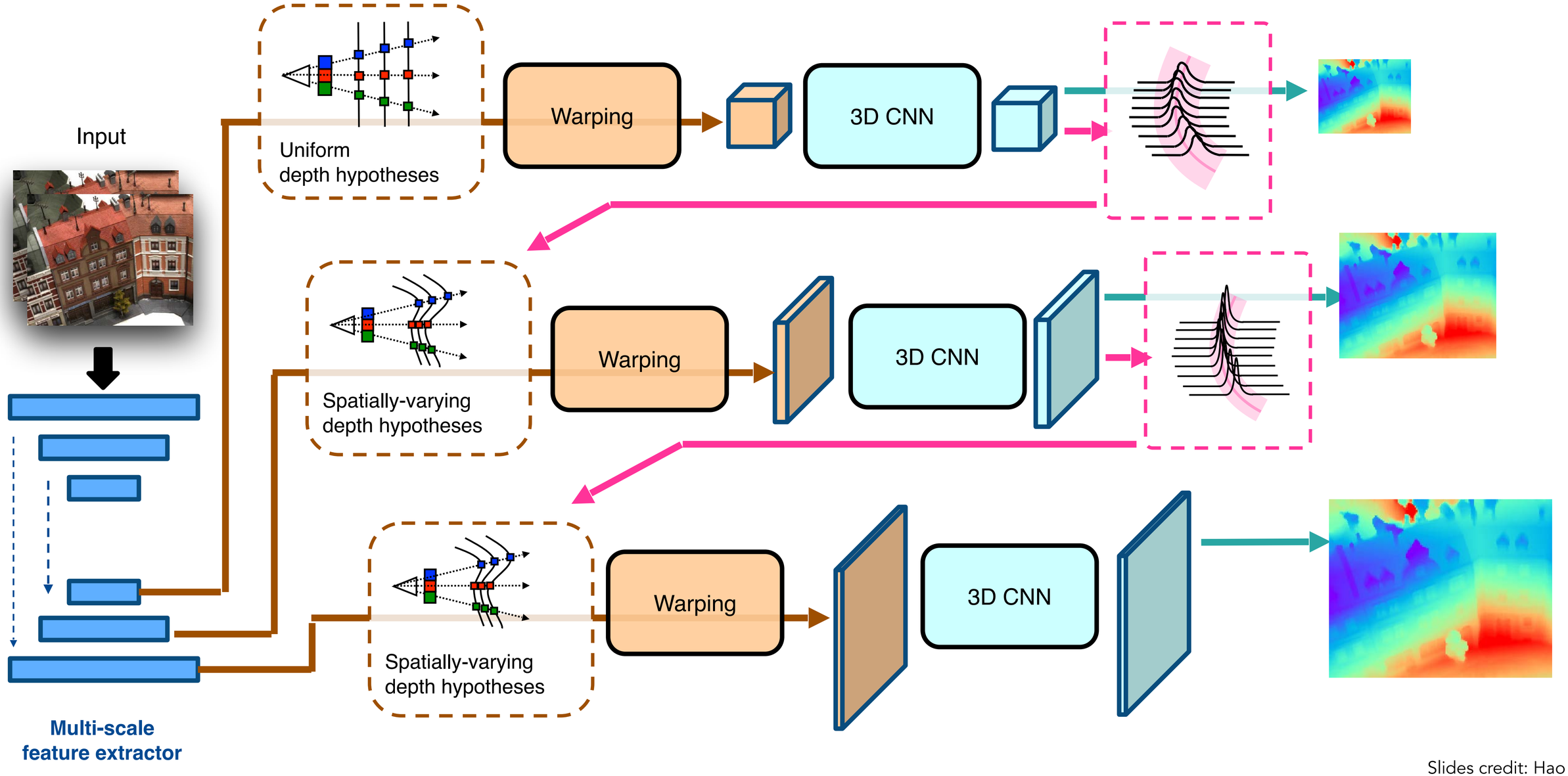


- Analyze per-pixel confidence intervals
- Narrow down the sampling range based on uncertainty

Cheng, Shuo, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. "Deep stereo using adaptive thin volume representation with uncertainty awareness." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2524-2534. 2020.

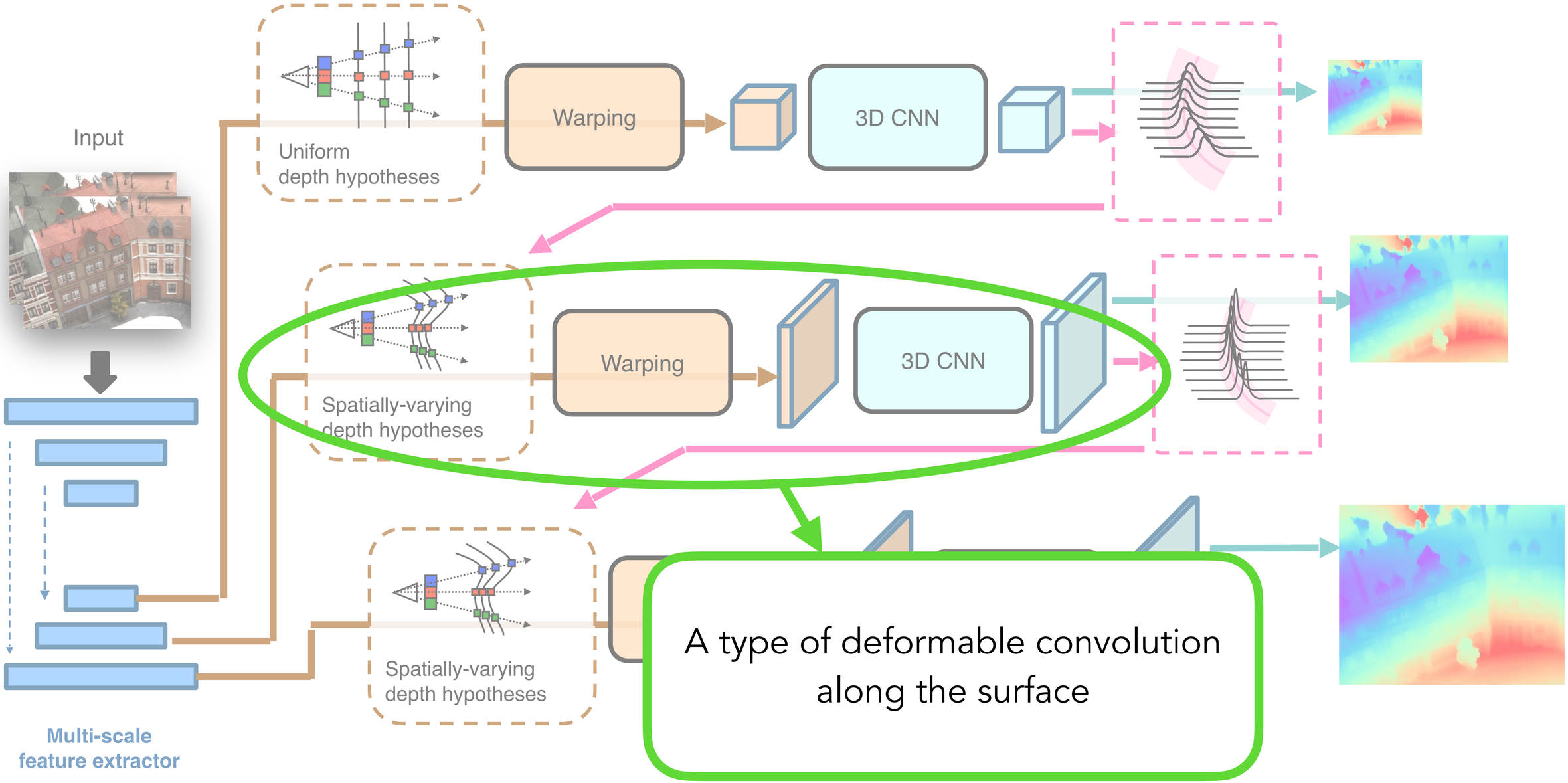
Slides credit: Hao Su

Cascaded Depth Prediction



Slides credit: Hao Su

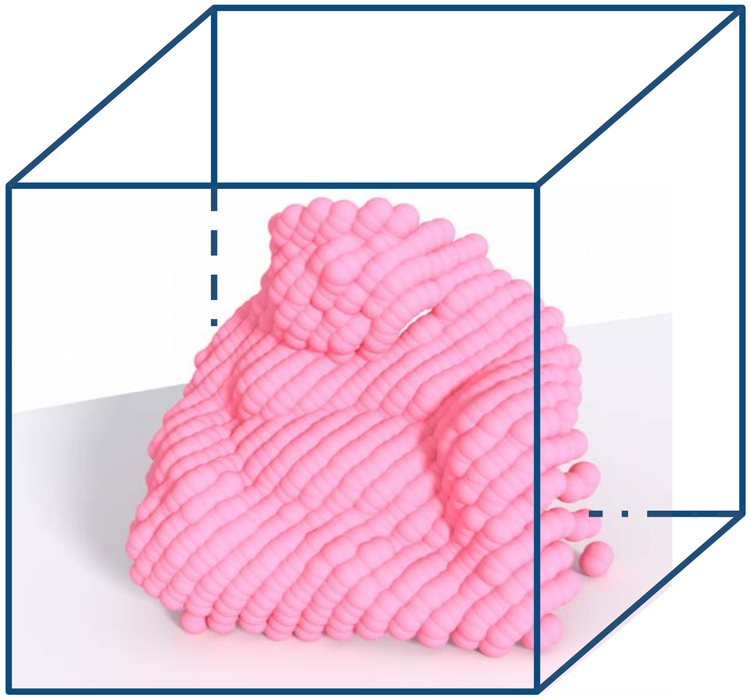
Cascaded Depth Prediction



Point-based Multi-View Stereo Network

Point cloud representation

- Suitable for sparse occupancy
- Memory-efficient

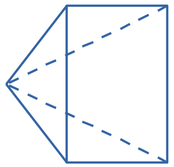


Chen, Rui, Songfang Han, Jing Xu, and Hao Su. "Point-based multi-view stereo network." In Proceedings of the IEEE International Conference on Computer Vision, pp. 1538-1547. 2019.

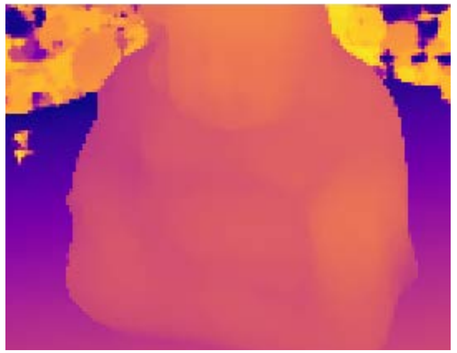
Slides credit: Hao Su

Initial Point Cloud

Estimate **low-resolution** depth map with existing methods



Reference camera



Coarse Depth map

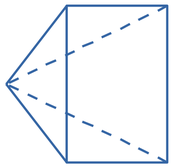
Unprojection

Initial point cloud

Slides credit: Hao Su

Initial Point Cloud

Estimate **low-resolution** depth map with existing methods



Reference camer

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

2D Image Coordinates

Intrinsic properties (Optical Centre, scaling)

Extrinsic properties (Camera Rotation and translation)

3D World Coordinates

Coarse Depth map

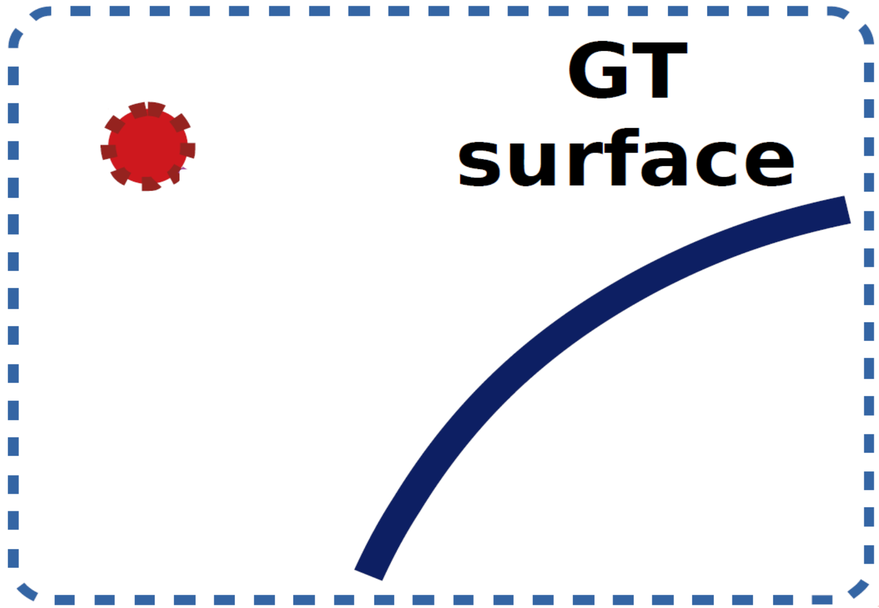
Initial point cloud

Slides credit: Hao Su

Point Flow

Goal:

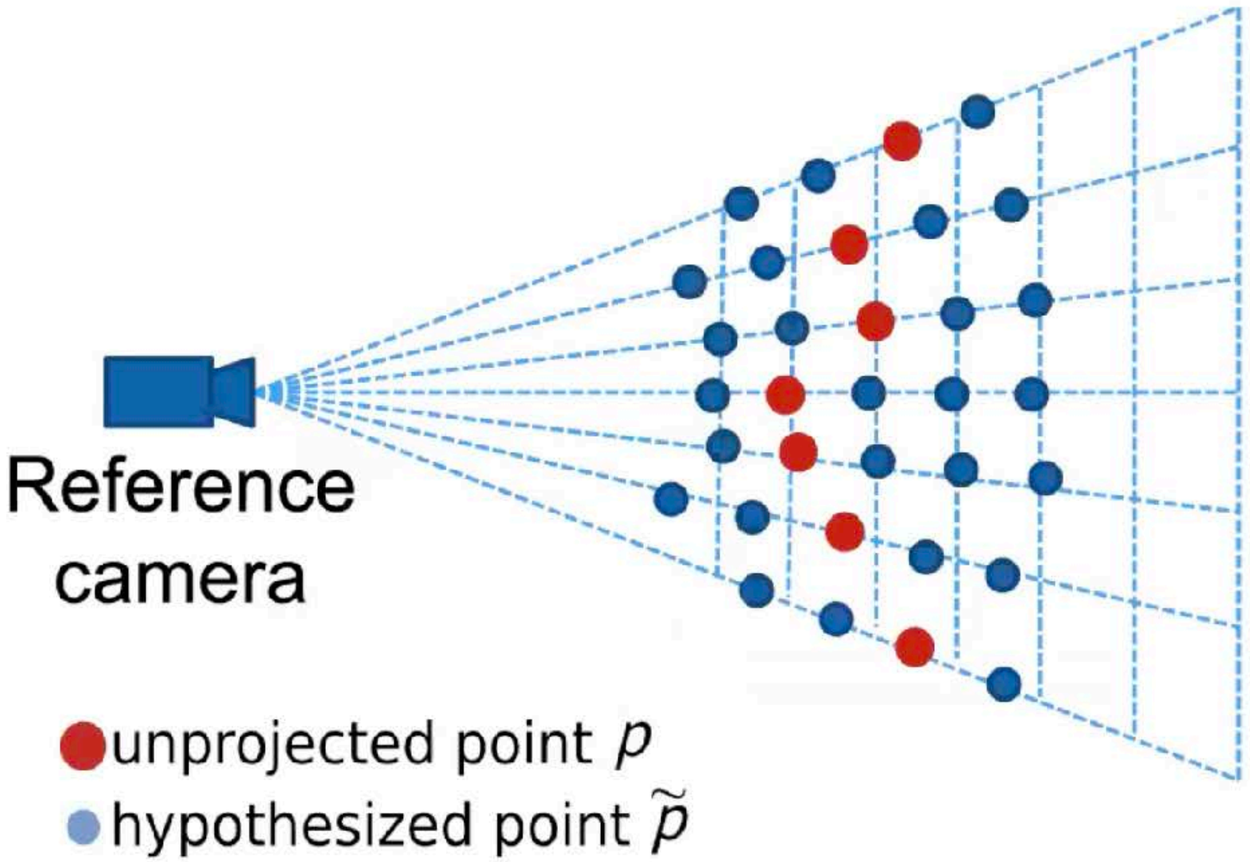
Refine the input depth map by moving the unprojected points along camera direction



Slides credit: Hao Su

Point Hypothesis

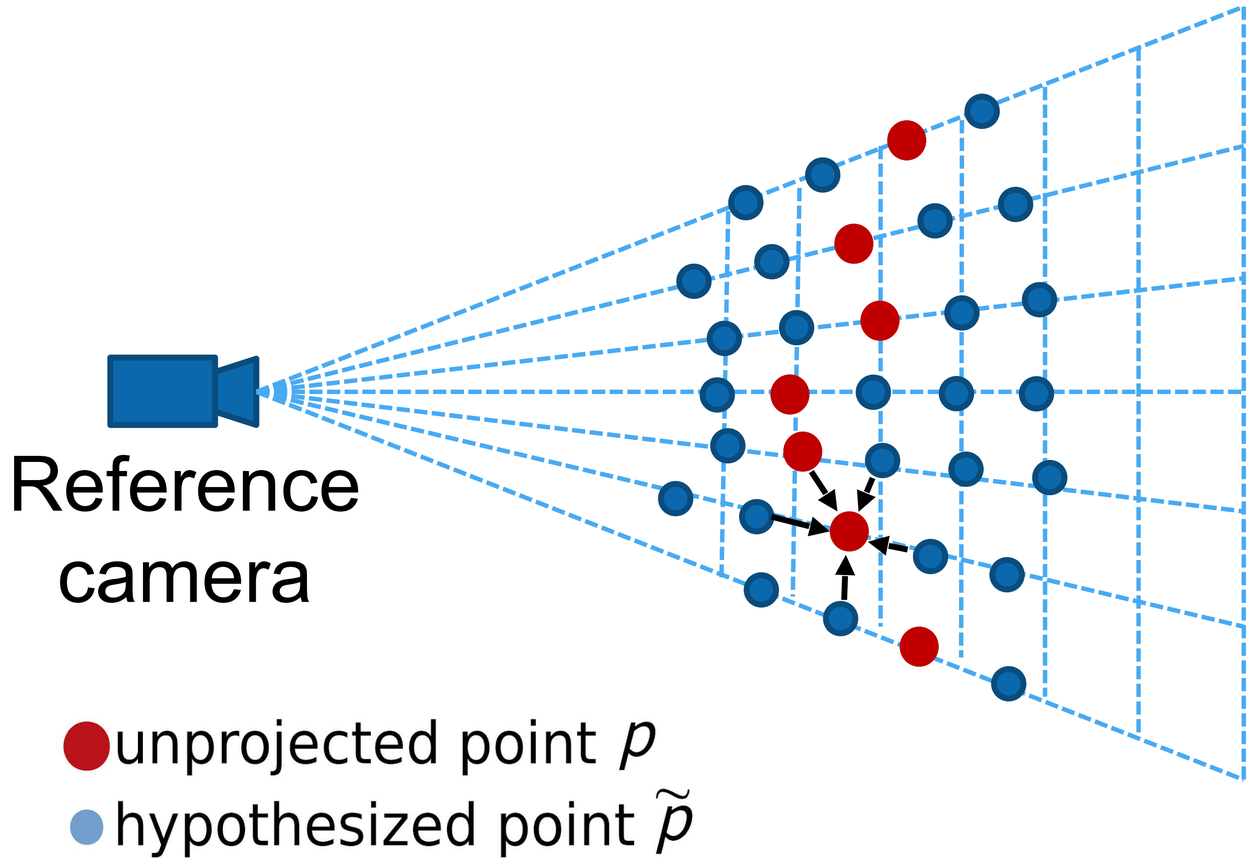
Point hypothesis: A sequence of points with different offsets



Slides credit: Hao Su

Flow Prediction

Flow prediction as expected offset



expected offset

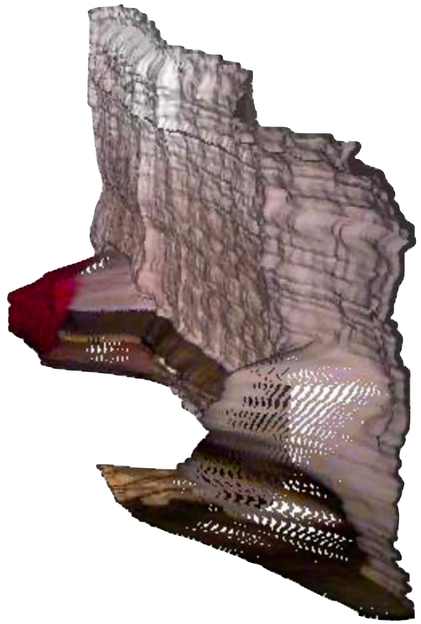
$$\Delta d_p = \mathbf{E}(ks) = \sum_{k=-m}^{m} ks \times \text{Prob}(\tilde{\mathbf{p}}_k)$$

Slides credit: Hao Su

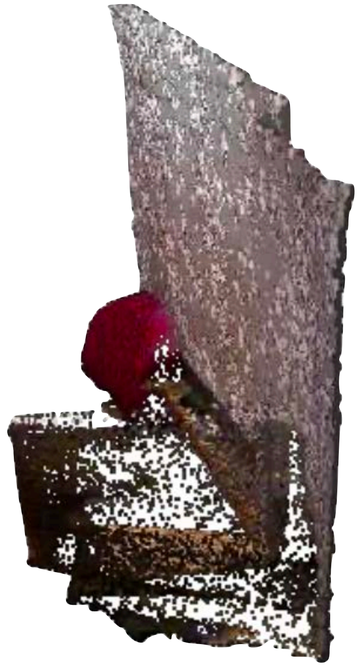
Outline

- Introduction to multi-view stereo (MVS)
- Classic MVS
- Learning-based MVS: a first pipeline
- Learning-based MVS: Improvements
  - Adaptive Space Sampling
  - ***Depth-Normal Consistency***

Depth Supervision Alone Does Not Give Smooth Surface



Prediction

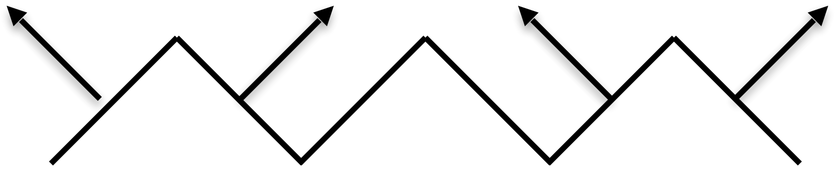


Ground truth

Slides credit: Hao Su

How to Improve Surface Smoothness?

- **Key observation**: Surface smoothness is reflected by surface normal.

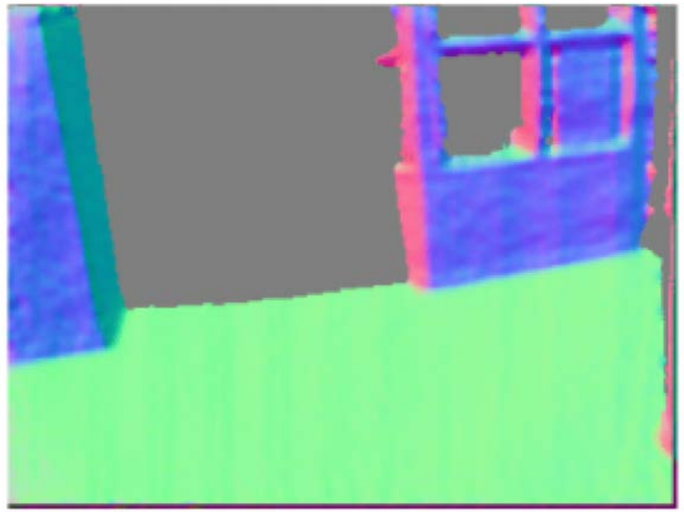


Rough surface

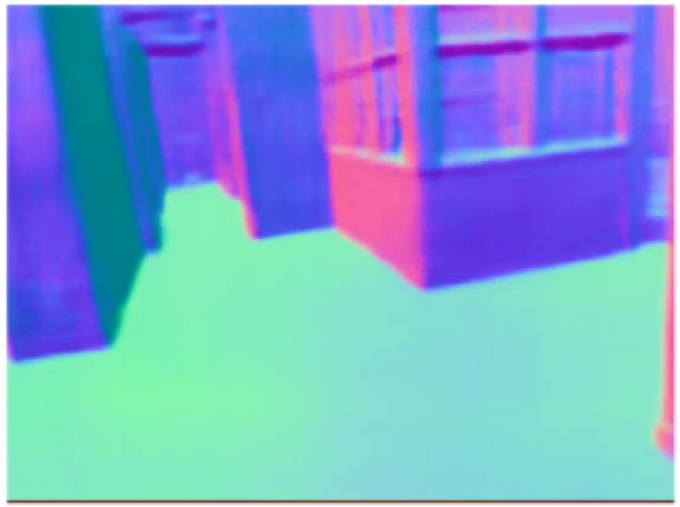
Plane surface

Slides credit: Hao Su

Observation: Normal Prediction is Easier



GT Normal

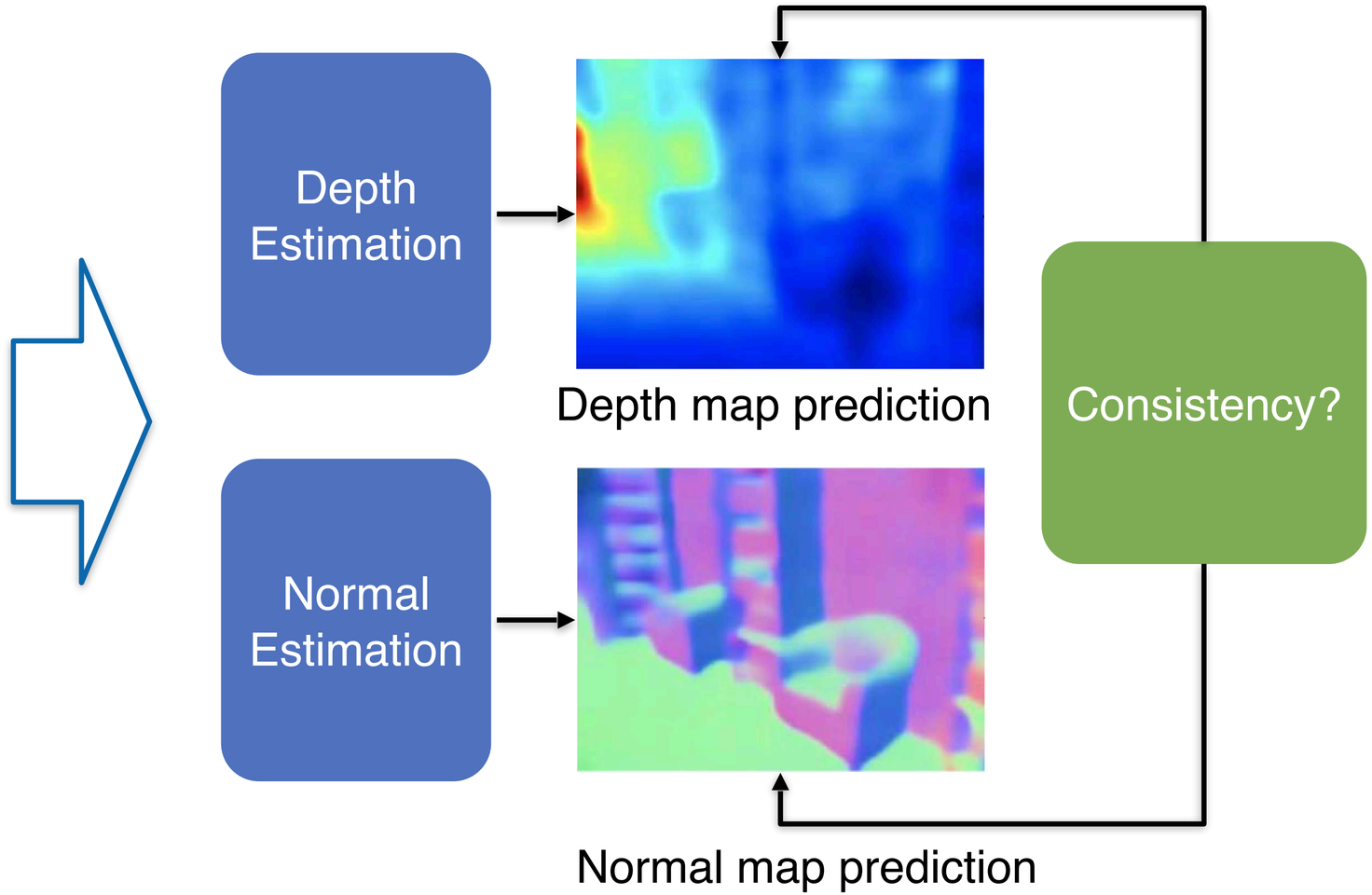


Predicted Normal

Slides credit: Hao Su

Depth Normal Consistency

- Estimate normal along with depth map.
- Regularize depth by normals.



Slides credit: Hao Su

Depth Normal Consistency

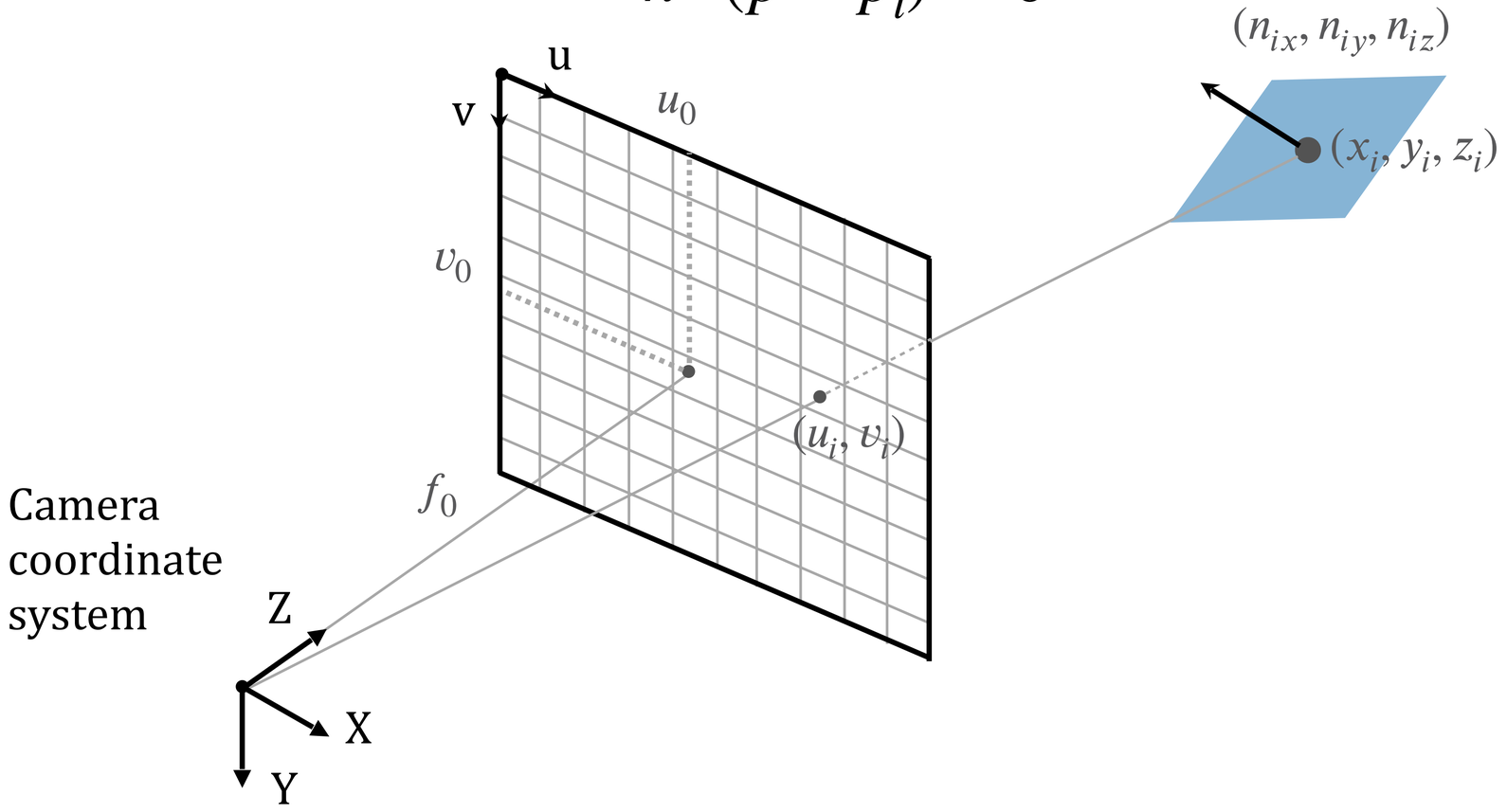
- Practice 1: Normal estimation as an auxiliary loss
  - Already quite effective
- Practice 2: Use normal estimation to correct depth

Slides credit: Hao Su

Refine Depth from Normal

- **Key assumption**: pixels within a local neighborhood lie on the same tangent plane.

$$\overrightarrow{n}^T(p - p_i) = 0$$

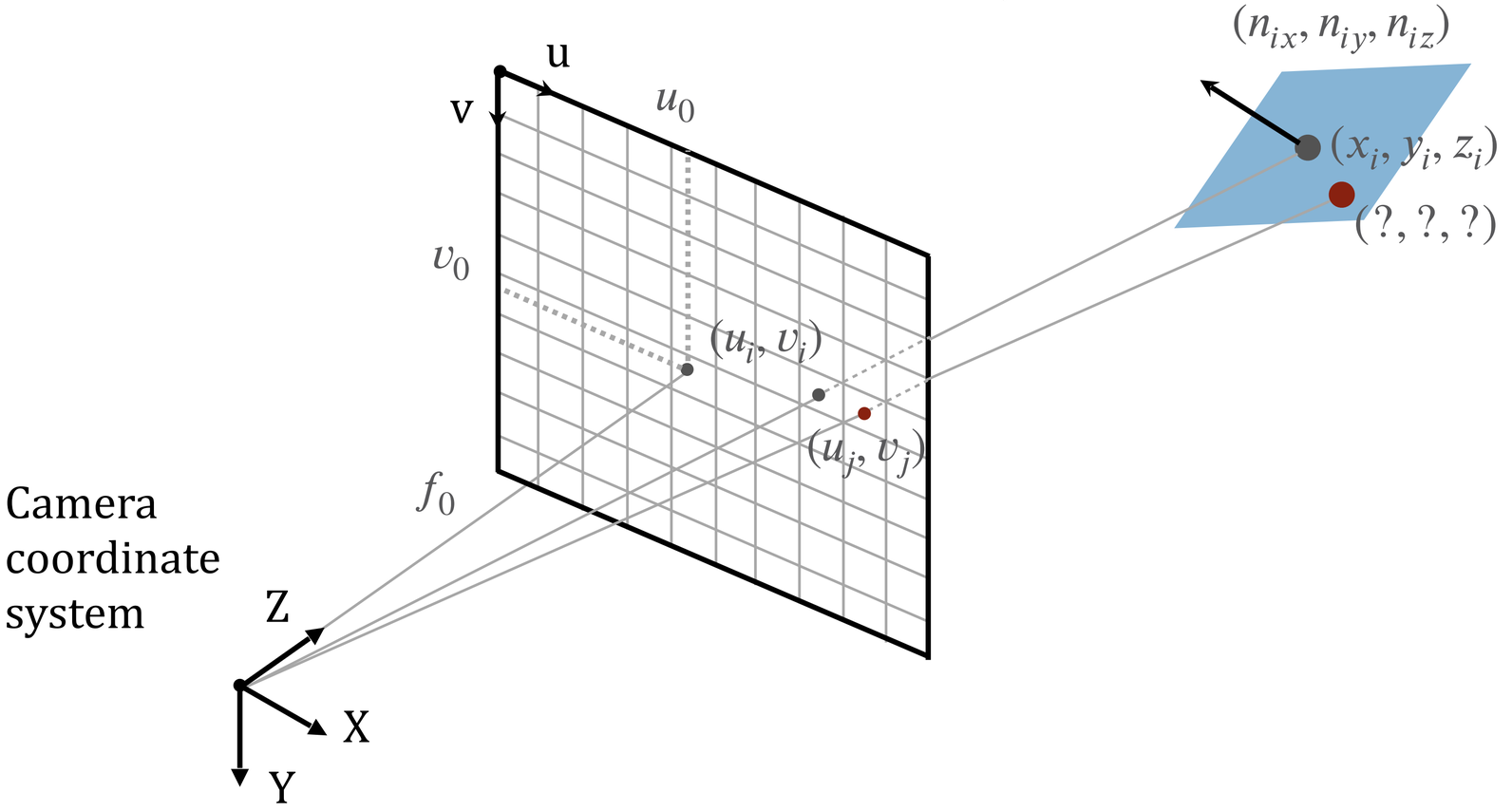


Slides credit: Hao Su

Refine Depth from Normal

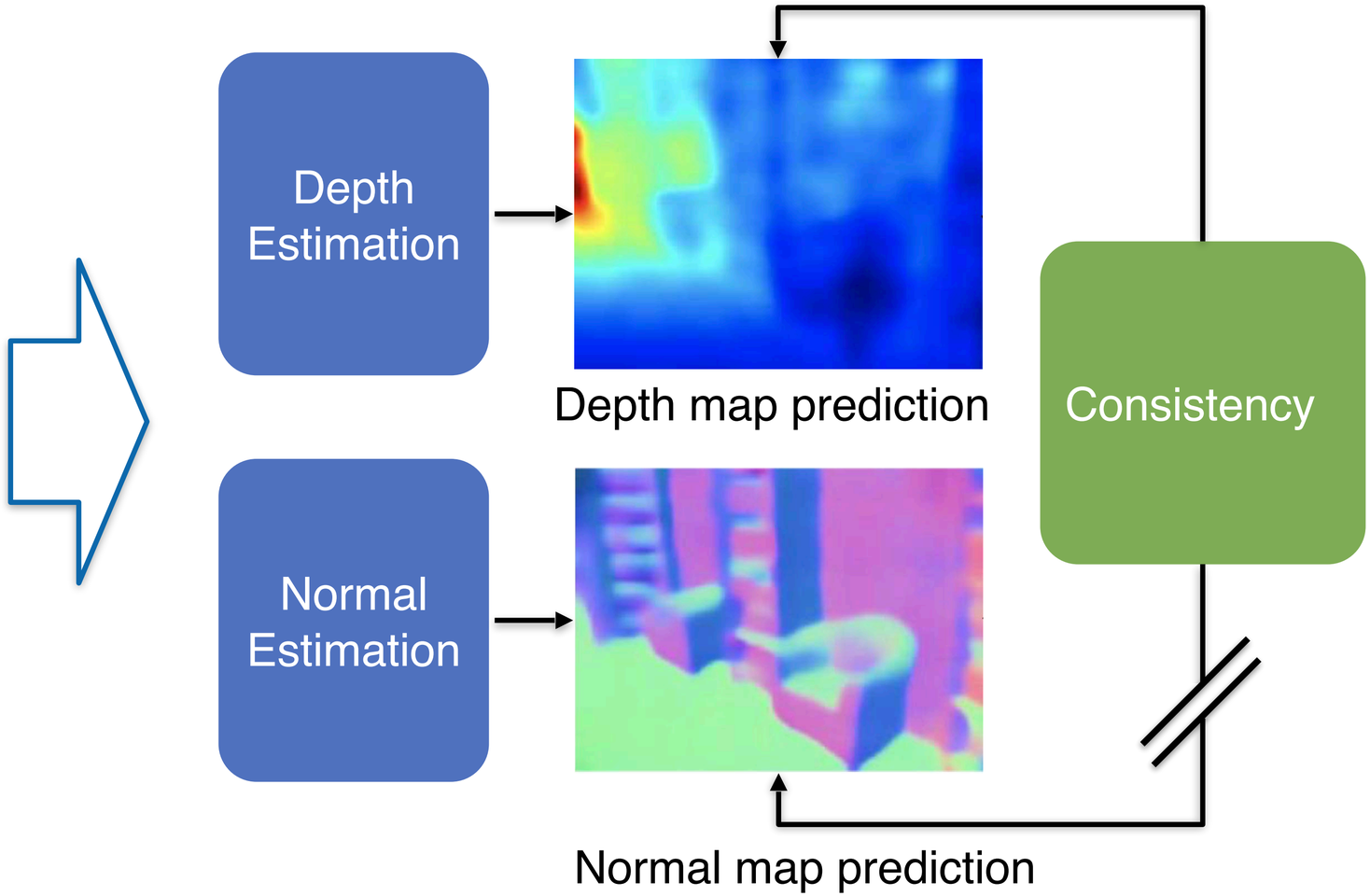
- Derive neighbor pixel depth from current pixel normal.

$$z'_{i\to j} = \frac{n_{ix}x_j + n_{iy}y_j + n_{iz}z_j}{(u_i - u_0)n_{ix}/f_0 + (v_i - v_0)n_{iy}/f_0 + n_{iz}}$$



Slides credit: Hao Su

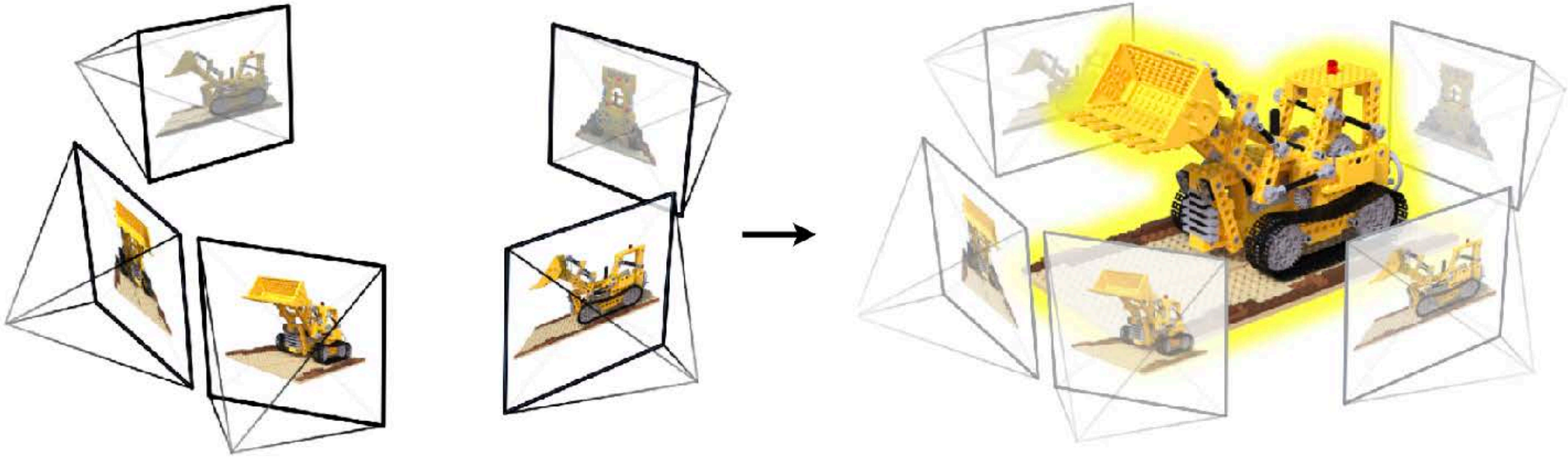
Depth Normal Consistency



Summary

- Deep volumetric stereo can lead to more robust matching and more complete reconstruction
- But volume-based methods are NOT computationally efficient, since the 3D target scene is sparse
- Adaptive sampling can improve computation efficiency and reconstruction quality
- Normal prediction is easier than depth, and can help improve depth accuracy and smoothness

Next Time



input images + camera poses

3D scene representation

NeRF