Multi-view Geometry

Angjoo Kanazawa

CS280 Spring 25

Feb 3, 2025

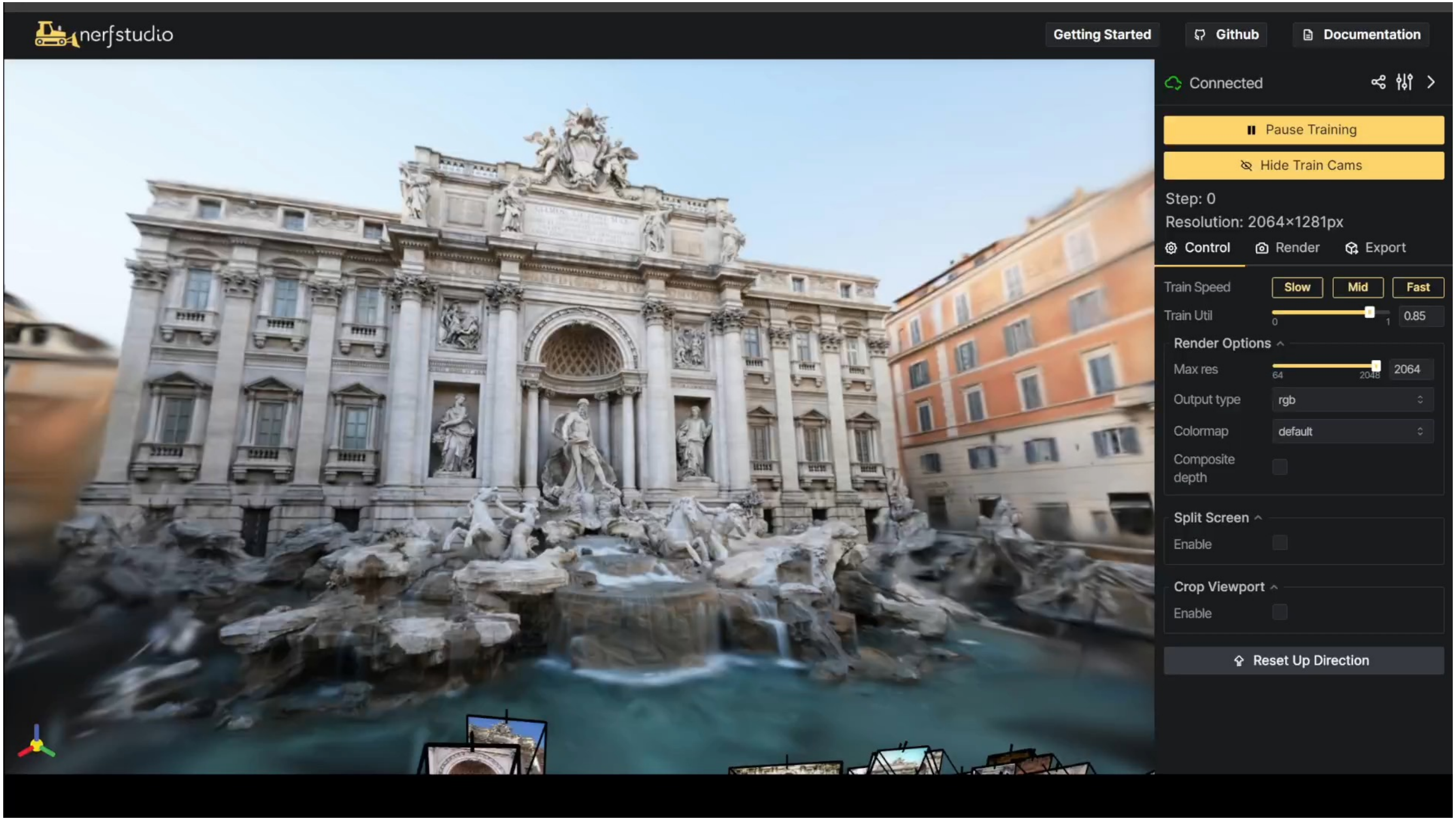
So far

- Calibration
  - Unknown: Camera K, {R, T}
  - Known: 3D geometry and 2D correspondences
- Refining Calibrated cameras
  - Minimize Reprojection loss
  - PnP: Solve for {R, T} with known 3D and 2D points
- Triangulation
  - Unknown: 3D points
  - Known: Camera, 2D correspondences
  - Special case: parallel optical axis

Now

- General camera
- **Don’t know camera AND 3D shape**

Application:
Reconstructing Internet Images



Problem Statement

- General camera
- Unknown: camera and 3D shape
- Known: N Correspondences
- **Goal:** Solve for camera and the depth of those points

How?

- Define the relationship between cameras and points → "Epipolar Geometry"
- Get camera from points using Epipolar Geometry
- Solve for depth via triangulation
- Refine everything, aka "Bundle Adjustment"

Structure from Motion (SfM)

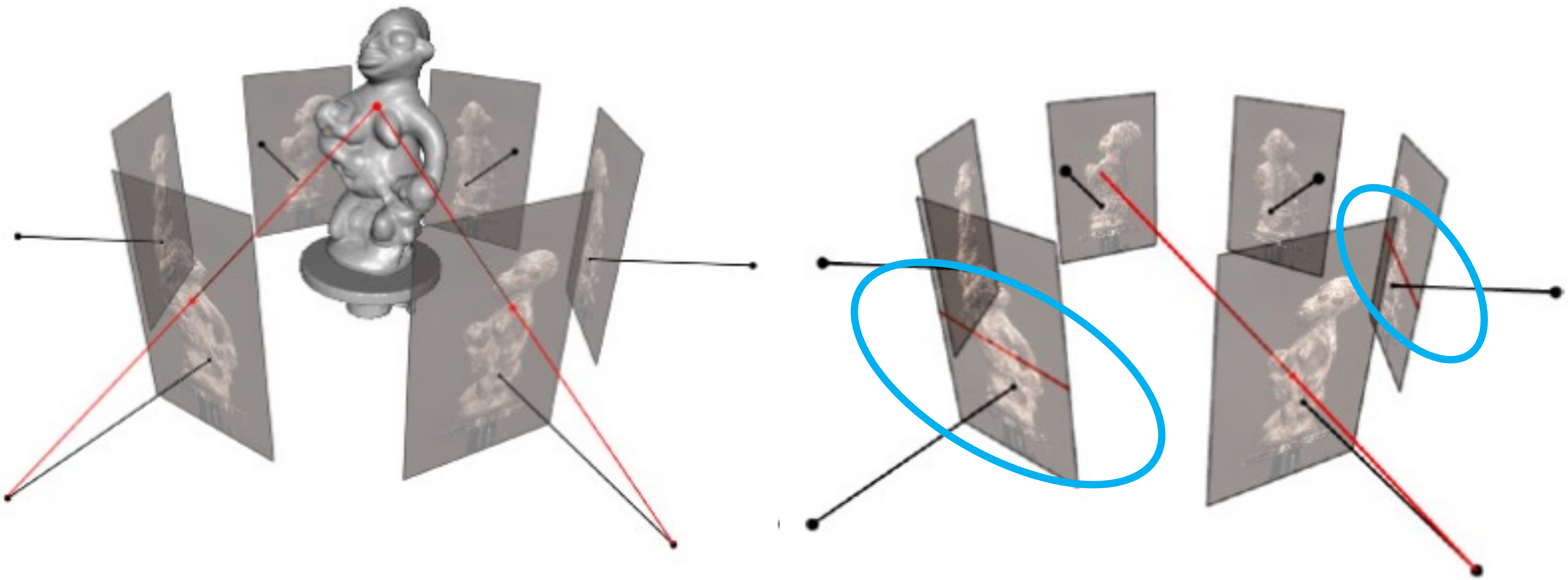
Simultaenous Localization and Mapping SLAM (online version)

SfM Steps

- Define the relationship between cameras and points → “Epipolar Geometry”
- Get camera from points using Epipolar Geometry
- Solve for depth via triangulation
- Refine everything, aka “Bundle Adjustment”

Epipolar Geometry

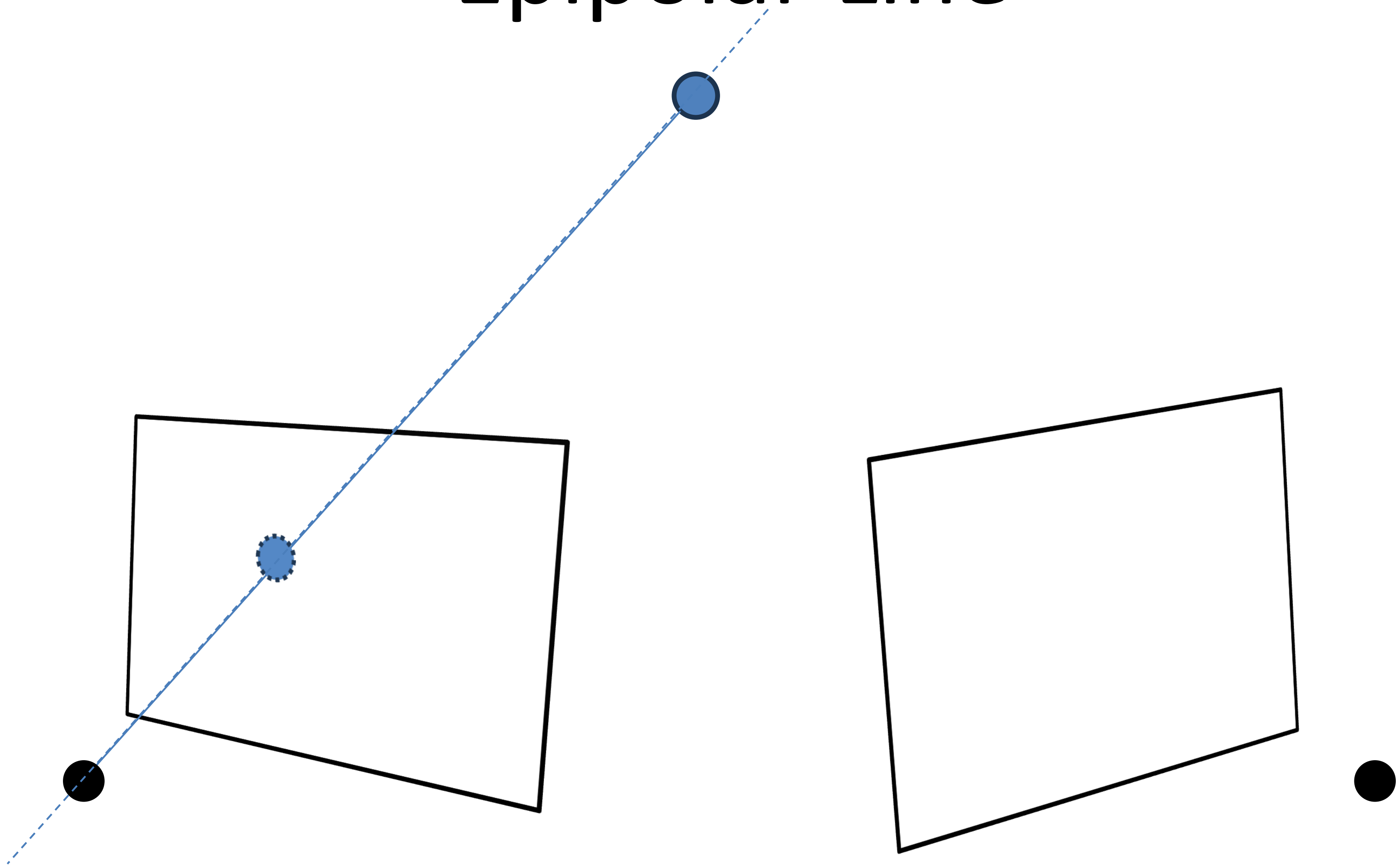
Intuitive Picture



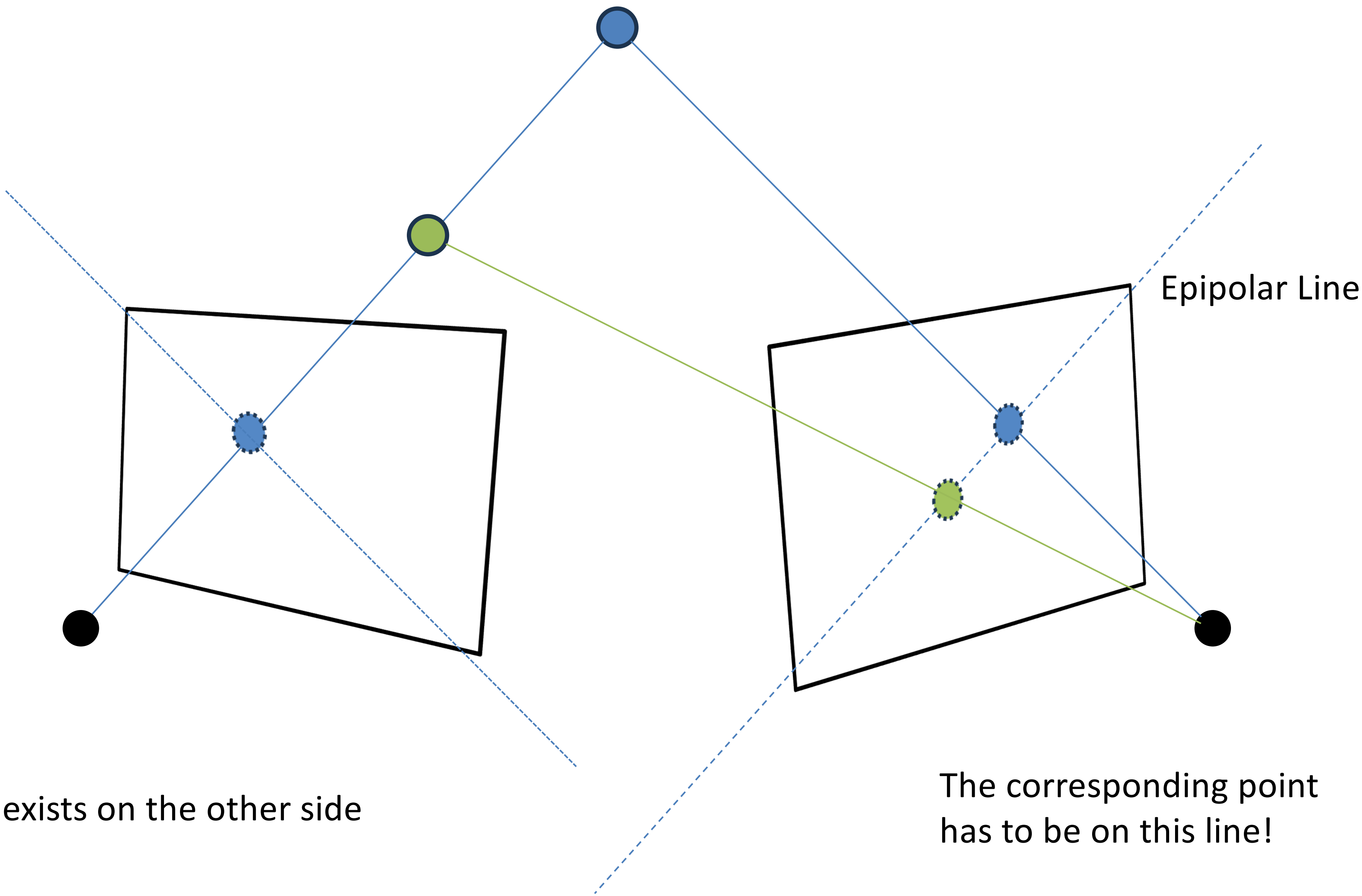
Figures by Carlos Hernandez

If you get confused with the following math, look at this picture again, it just describes this.

Epipolar Line



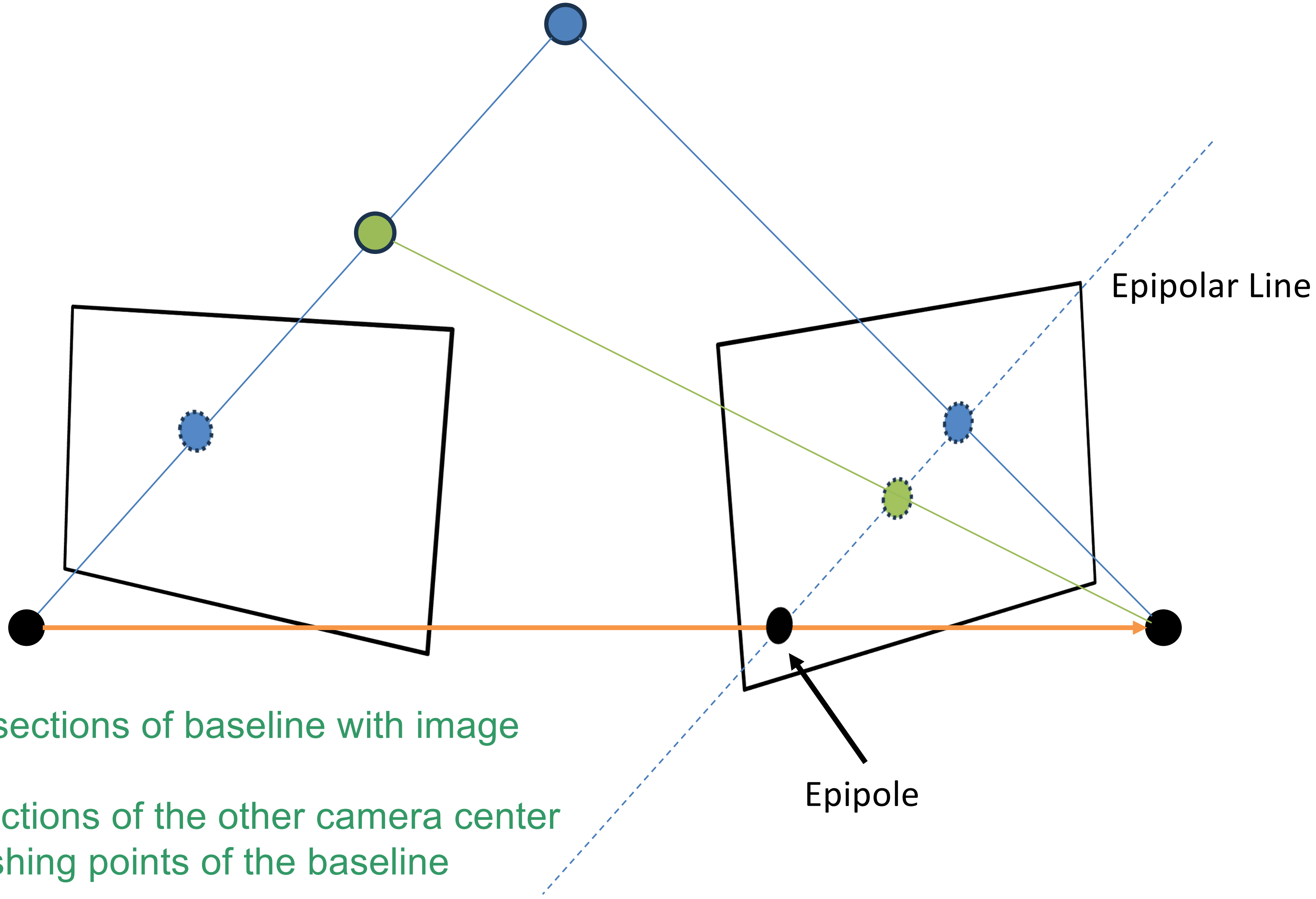
Epipolar Line



Also exists on the other side

The corresponding point has to be on this line!

Epipole

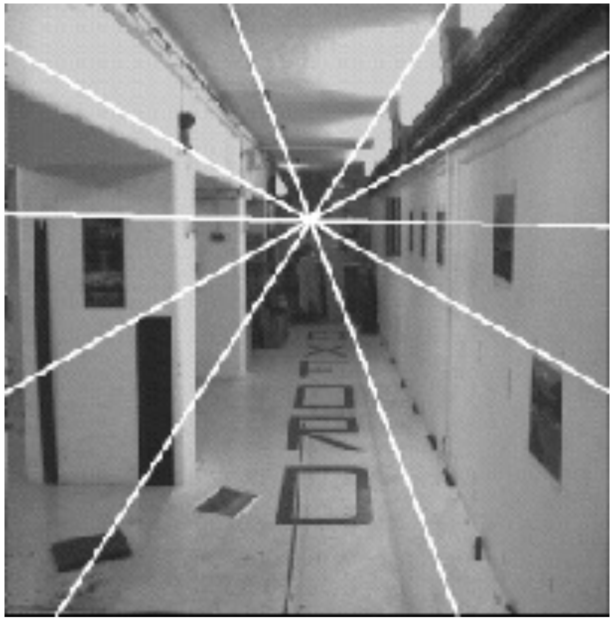


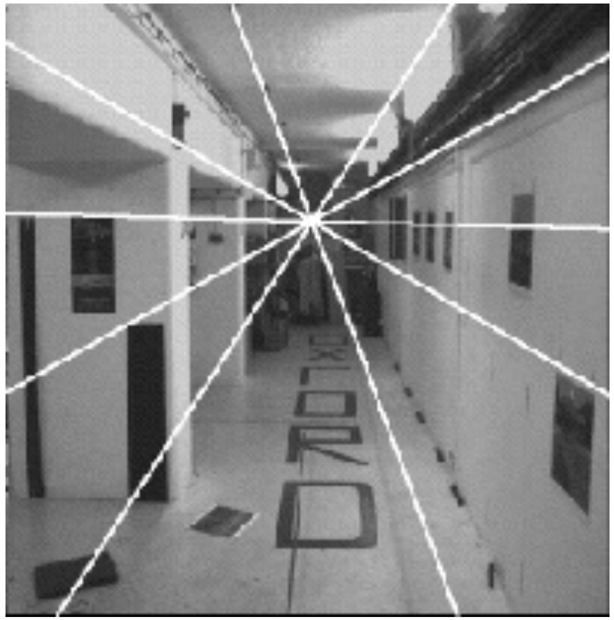
= intersections of baseline with image planes
= projections of the other camera center
= vanishing points of the baseline

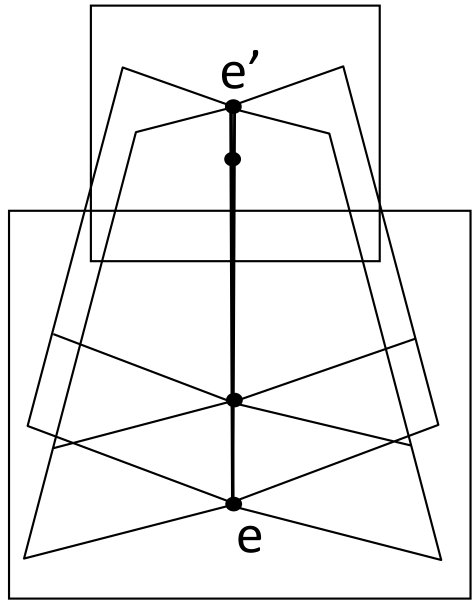
The Epipole

Photo by Frank Dellaert

Example: forward motion



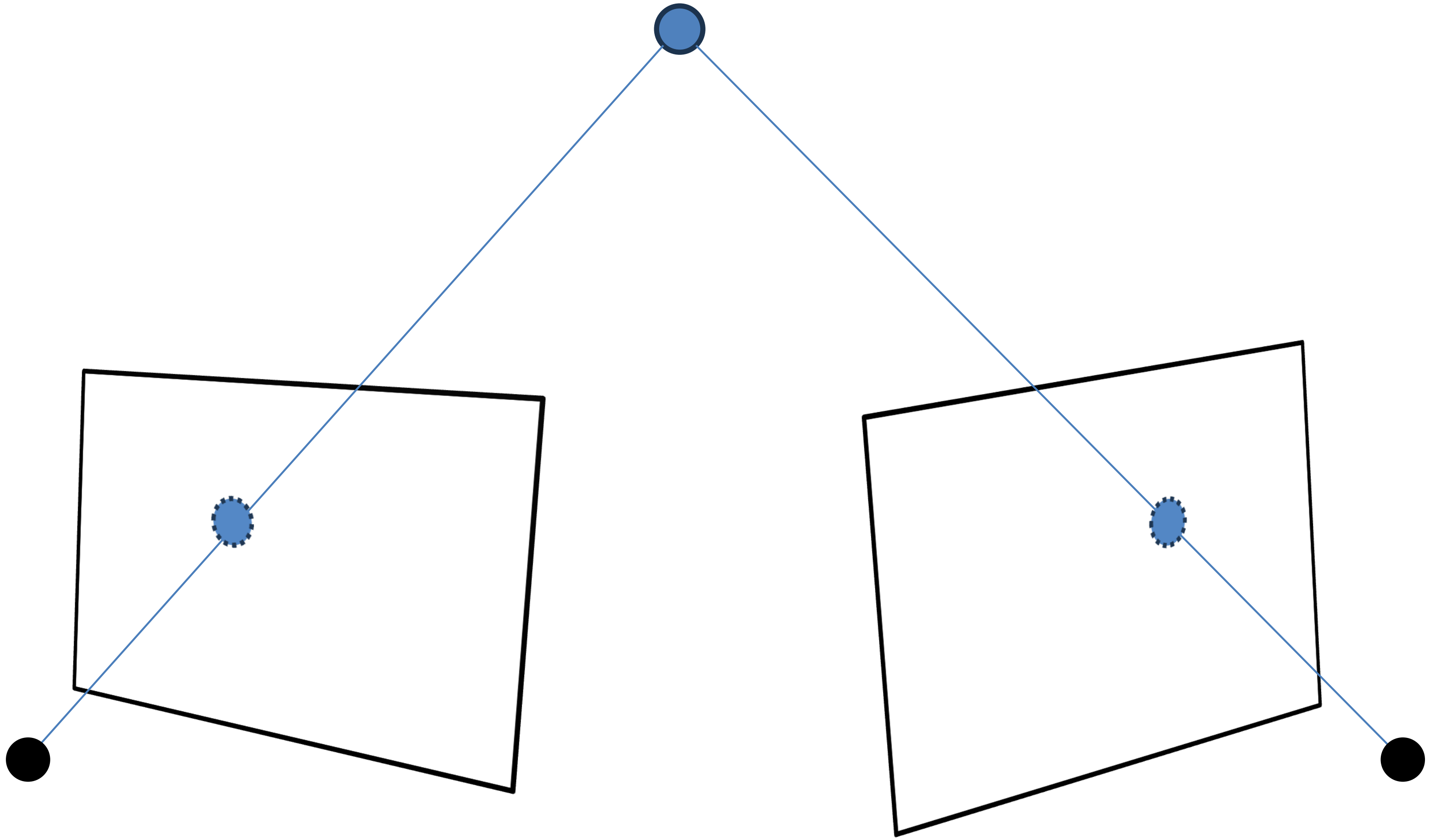




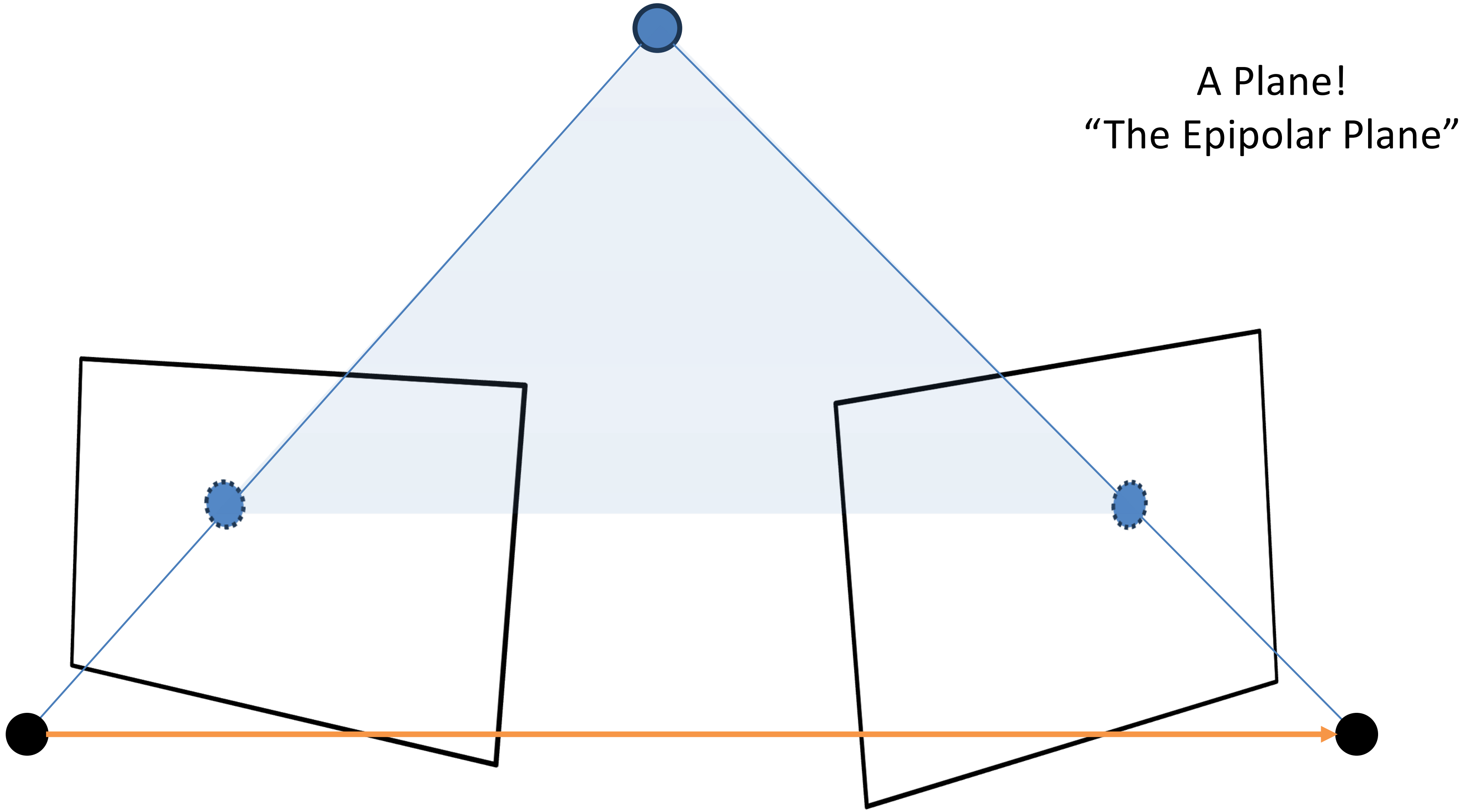
Epipole has same coordinates in both images.
Points move along lines radiating from e: “Focus of expansion”

Figure from Hartley & Zisserman

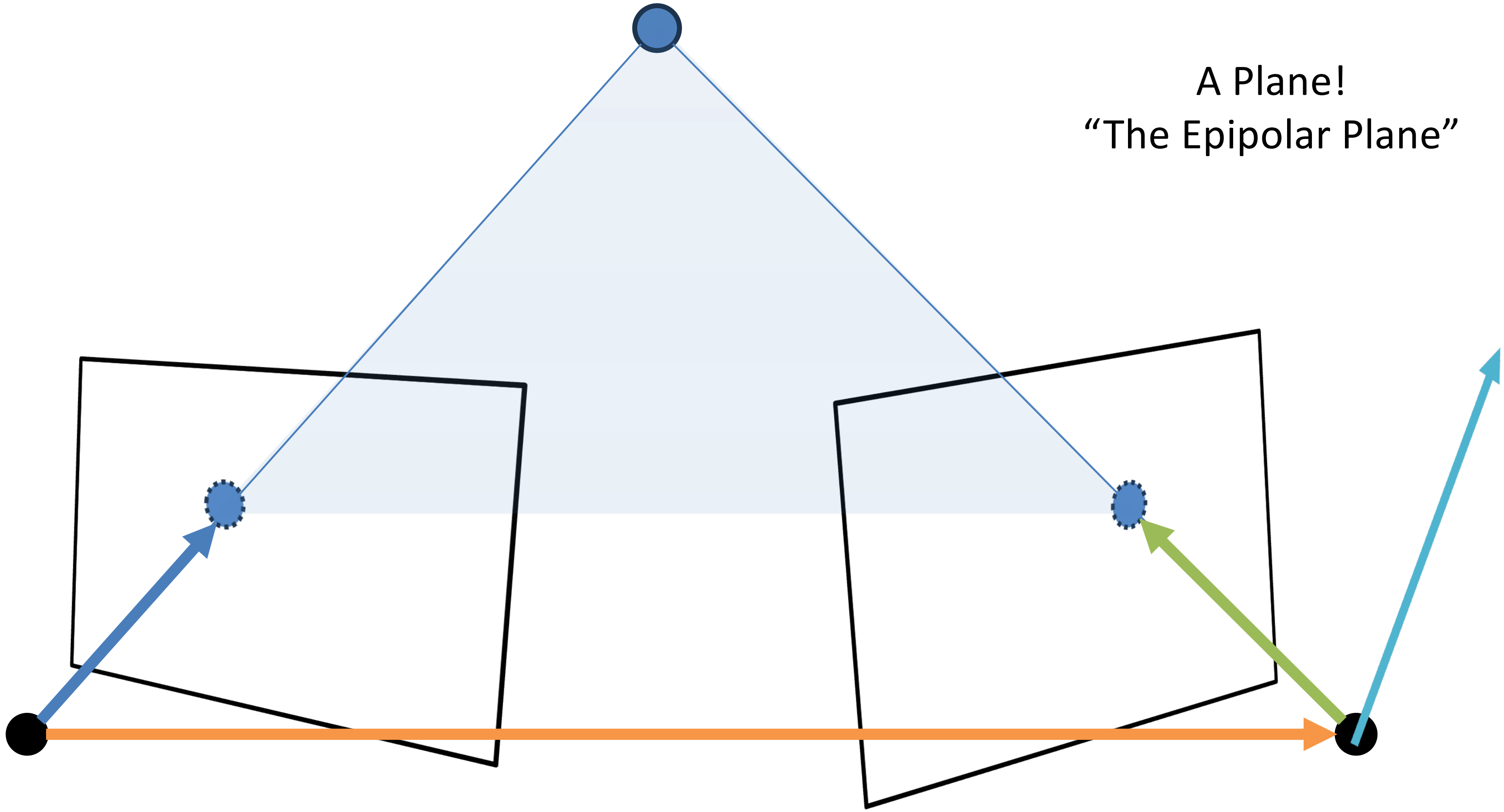
What is the relationship?



What is the relationship?



What can you say?

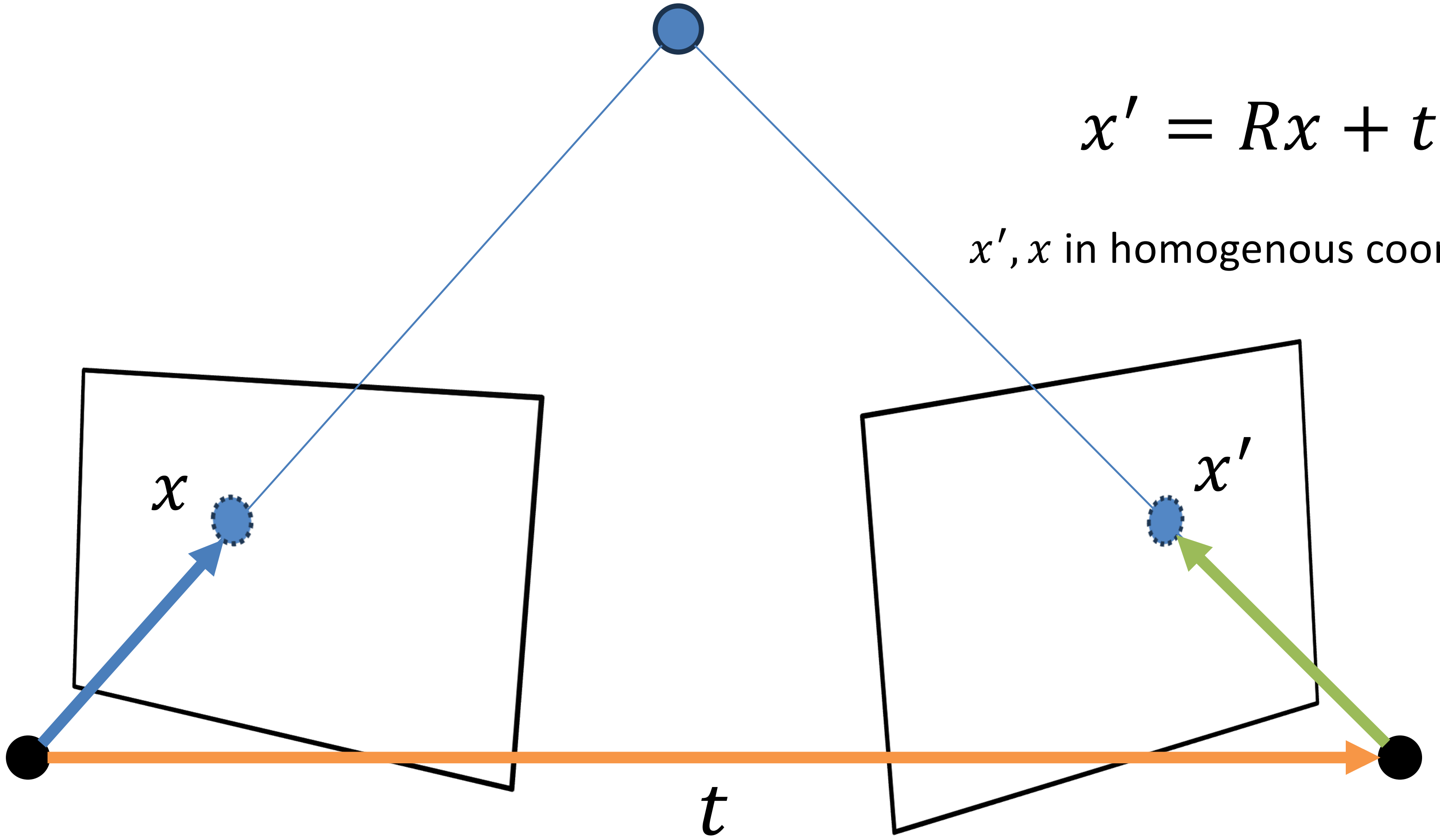


Now we are going to use this relationship to solve for camera R, t!!
Then, using the camera, the depth of the corresponding points

SfM Steps

- Define the relationship between cameras and points → "Epipolar Geometry"
- Get camera from points using Epipolar Geometry
- Solve for depth via triangulation
- Refine everything, aka "Bundle Adjustment"

Lets define the plane



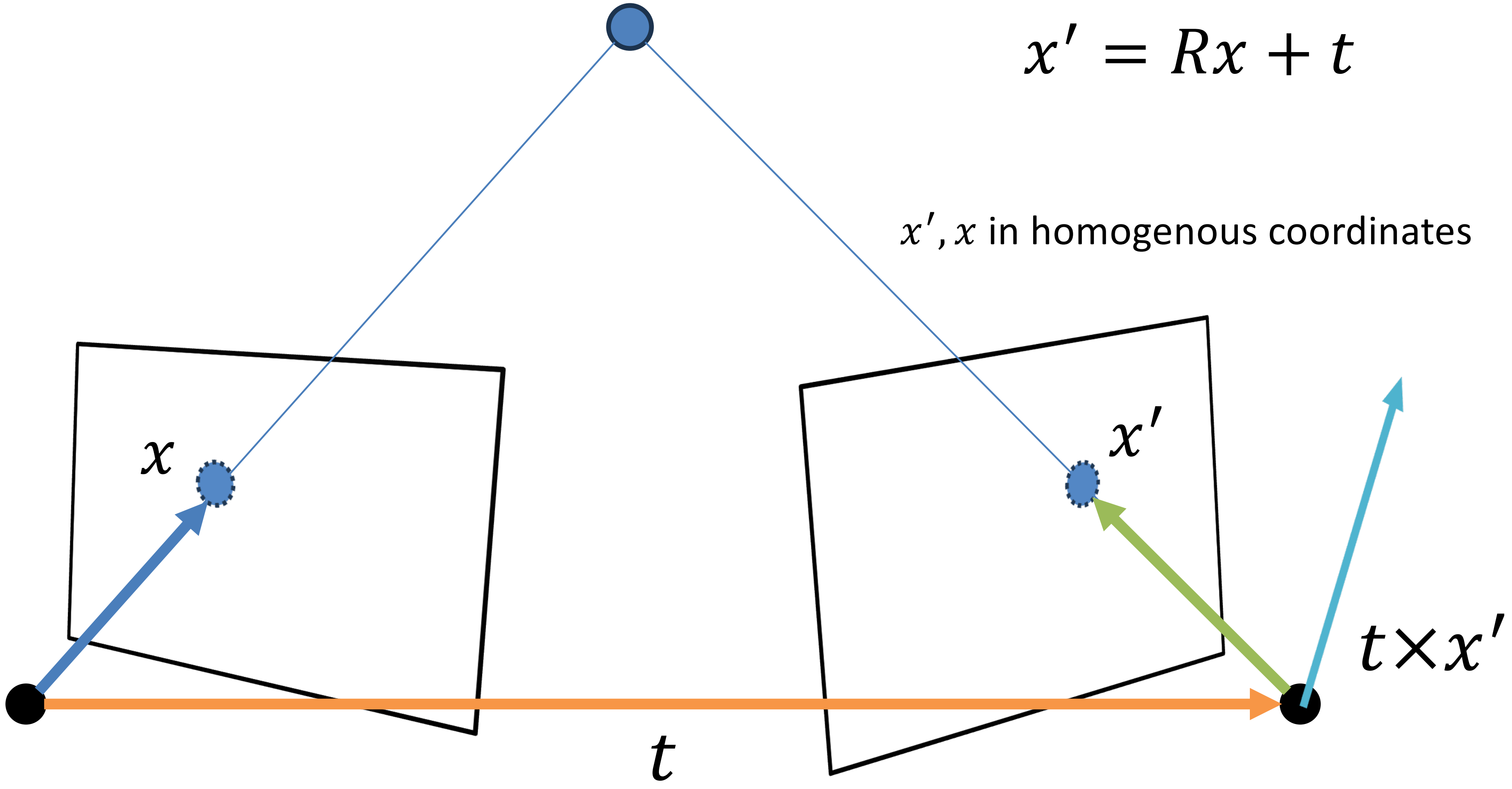
$$x' = Rx + t$$

x', x in homogenous coordinates

Lets define the plane

$$x' = Rx + t$$

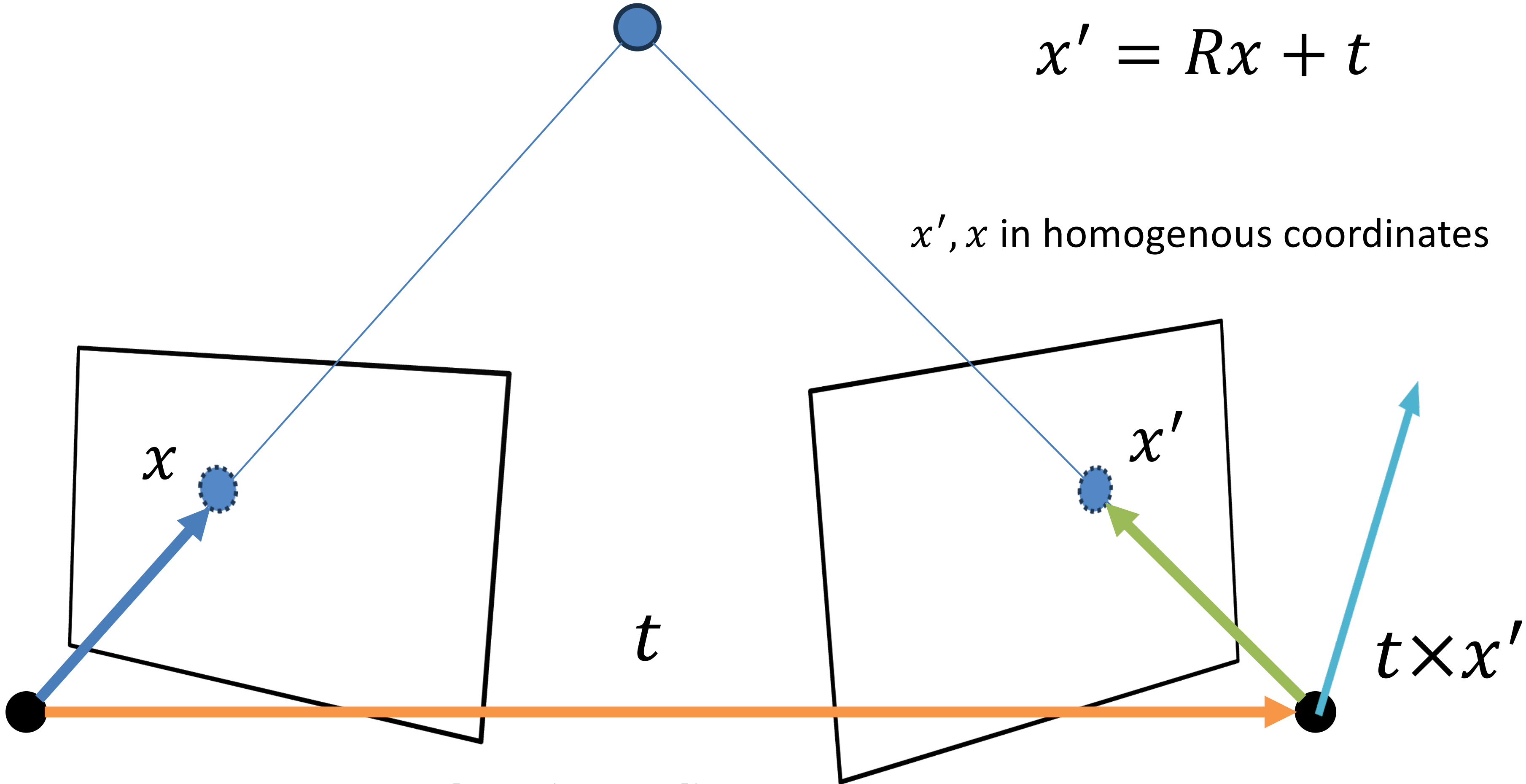
x', x in homogenous coordinates



Equation of plane

$$x' = Rx + t$$

x', x in homogenous coordinates



$$x' \cdot (t \times x') = 0$$

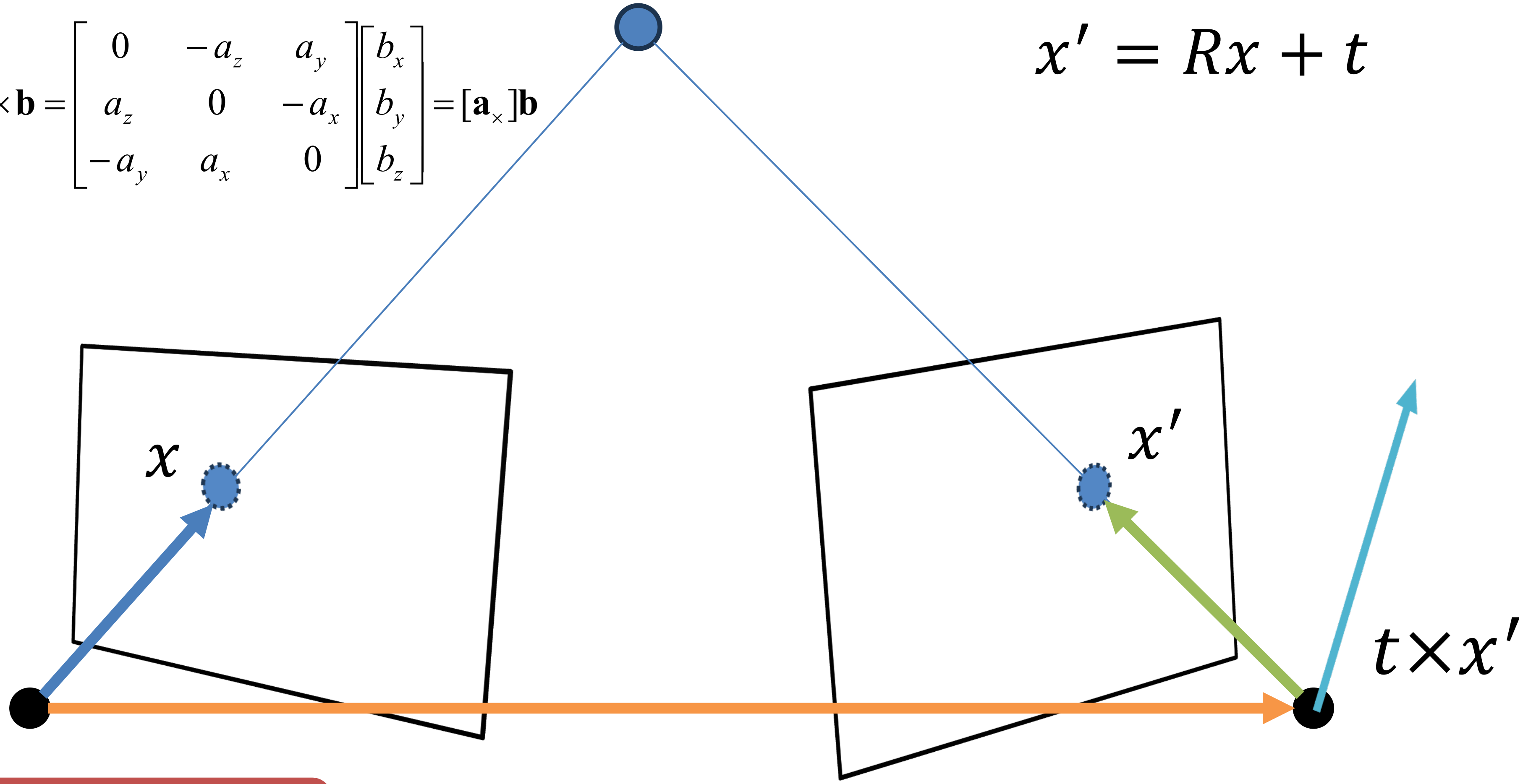
$$x' \cdot (\mathrm{t} \times (\mathrm{Rx} + \mathrm{t})) = 0$$

$$x' \cdot (\mathrm{t} \times \mathrm{Rx} + \mathrm{t} \times \mathrm{t})) = \boxed{x' \cdot (\mathrm{t} \times \mathrm{Rx}) = 0}$$

Equation of plane

Recall: $\mathbf{a} \times \mathbf{b} = \begin{bmatrix} 0 & -a_z & a_y \\ a_z & 0 & -a_x \\ -a_y & a_x & 0 \end{bmatrix} \begin{bmatrix} b_x \\ b_y \\ b_z \end{bmatrix} = [\mathbf{a}_\times]\mathbf{b}$

$x' = Rx + t$



$\boxed{x' \cdot (\mathrm{t} \times \mathrm{Rx}) = 0}$

$\boldsymbol{x}'^T \underbrace{[\boldsymbol{t}_\times]\boldsymbol{R}}_{E} \boldsymbol{x} = 0 \quad \Rightarrow \quad \boldsymbol{x}'^T E \boldsymbol{x} = 0$

Essential Matrix

(Longuet-Higgins, 1981)

Epipolar constraint: Uncalibrated case

- We normalized the coordinates

$$x = K^{-1}\hat{x} \quad x' = K'^{-1}\hat{x}'$$

$$\hat{x} = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

where $\hat{x}$ is the image coordinates

- But in the *uncalibrated* case, ***K*** and ***K'*** are unknown!
- We can write the epipolar constraint in terms of *unknown* normalized coordinates:

$$x'^T E x = 0$$

$$(K'^{-1}\hat{x}')'^T E(K^{-1}\hat{x}) = 0$$

$$\hat{x}'^T \underbrace{K'^{-T} E (K^{-1}}\hat{x}) = 0$$

$$\hat{x}'^T F \hat{x} = 0$$

$$F = K'^{-T} E K^{-1}$$

Fundamental Matrix
(Faugeras and Luong, 1992)

Coplanarity of 3-vectors implies triple product is zero

v_1, v_2, v_3 are coplanar $\Rightarrow$

$v_1 \cdot (v_2 \wedge v_3) = 0$

$v_1^T \hat{v}_2 v_3 = 0$

Now we can solve for E

- Same as DLT

Longuet-Higgins 8 point algorithm

(1981)

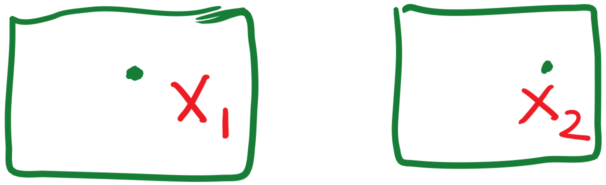
- Find n (≥ 8) corresponding points in the 2 views
- Estimate the E matrix (= $\hat{T}R$) from these point correspondences.
- Extract (R, t).
- Recover depth by triangulation.

$E \to \lambda E$

Given projections x_1, x_2

$$x_2^T E x_1 = 0$$

$$\underbrace{\begin{bmatrix} -, & -, & - \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} & e_{13} \\ e_{21} & e_{22} & e_{23} \\ e_{31} & e_{32} & e_{33} \end{bmatrix} \begin{bmatrix} - \\ - \\ - \end{bmatrix}}$$



measured in each camera's coordinates

Each point gives a linear equation for entries of E

Essential matrix can be decomposed

- $E = T_x R$

- $$\begin{bmatrix} e_{11} & e_{12} & e_{13} \\ e_{21} & e_{22} & e_{23} \\ e_{31} & e_{32} & e_{33} \end{bmatrix} = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$$

  Given that $T_\times$ is a Skew-Symmetric matrix (a_{ij} = - a_{ji}) and R is an Orthonormal matrix, it is possible to "decouple" $T_\times$ and R from their product using "Singular Value Decomposition".

Slide from Shree Nayar

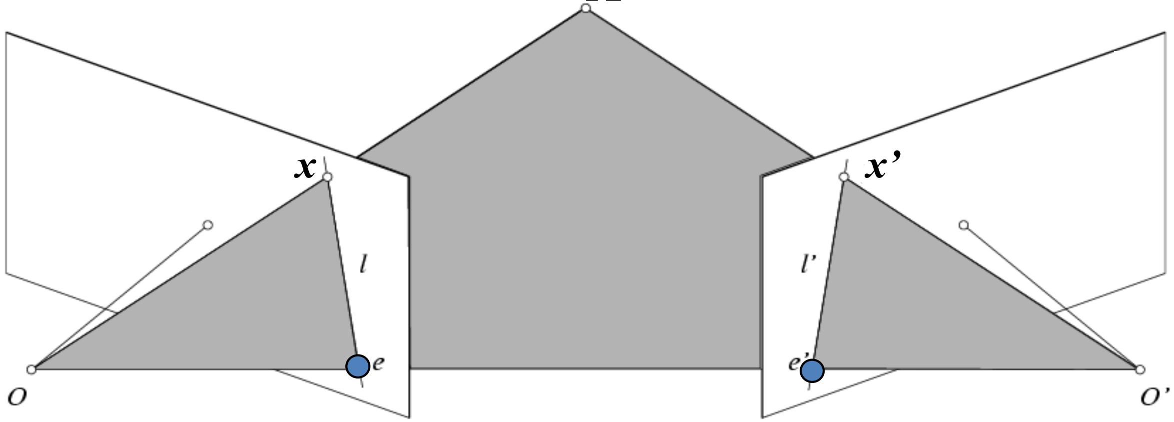
SfM Steps

- Define the relationship between cameras and points → "Epipolar Geometry"
- Get camera from points using Epipolar Geometry
- Solve for depth via triangulation
- Refine everything, aka "Bundle Adjustment"

Depth by triangulation

- We know about the camera, K_1, K_2 and [R t]:

$$X = X' = RX + T$$



And know the corresponding points: $x \leftrightarrow x'$

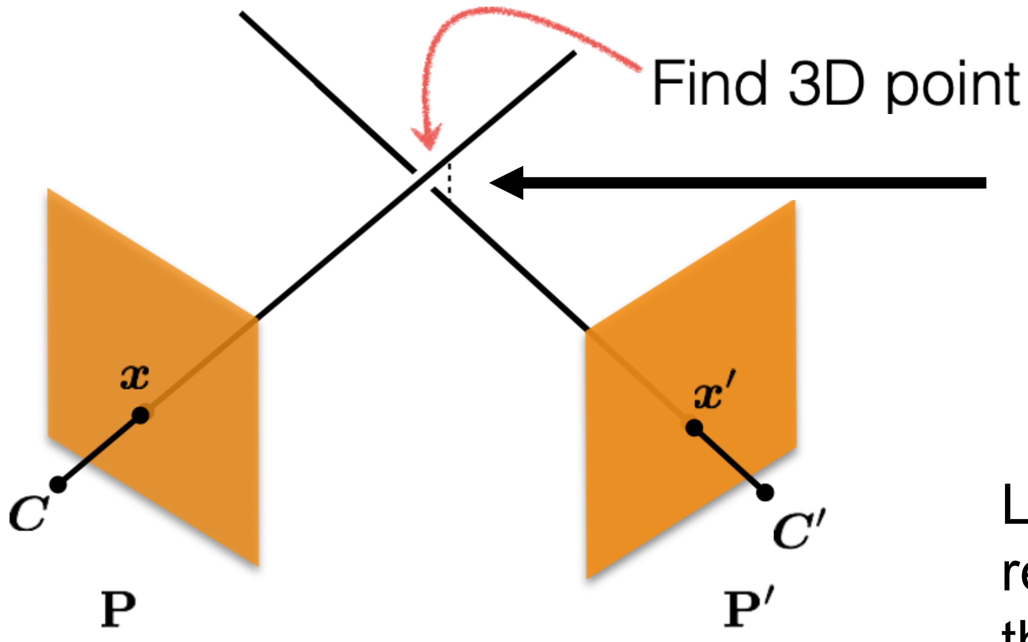
$$x = KX \qquad x' = K'X'$$
$$= K'(RX + T)$$

only unknowns!

How many unknowns + how many equations do we have?

Solve by least squares

Triangulation Issue: Noise



Ray's don't always intersect
because of noise!!!

$\mathbf{X}$ s.t.

$$\mathbf{x} = \mathbf{P}\mathbf{X}, \ \mathbf{x}' = \mathbf{P}'\mathbf{X}$$

Least squares get you to a reasonable solution but it's not the actual geometric error (it's how far away the solution is from Ax = 0)

In practice with noise, you do non-linear least squares, or **"bundle adjustment" against reprojection loss**

Slide credit: Shubham Tulsiani

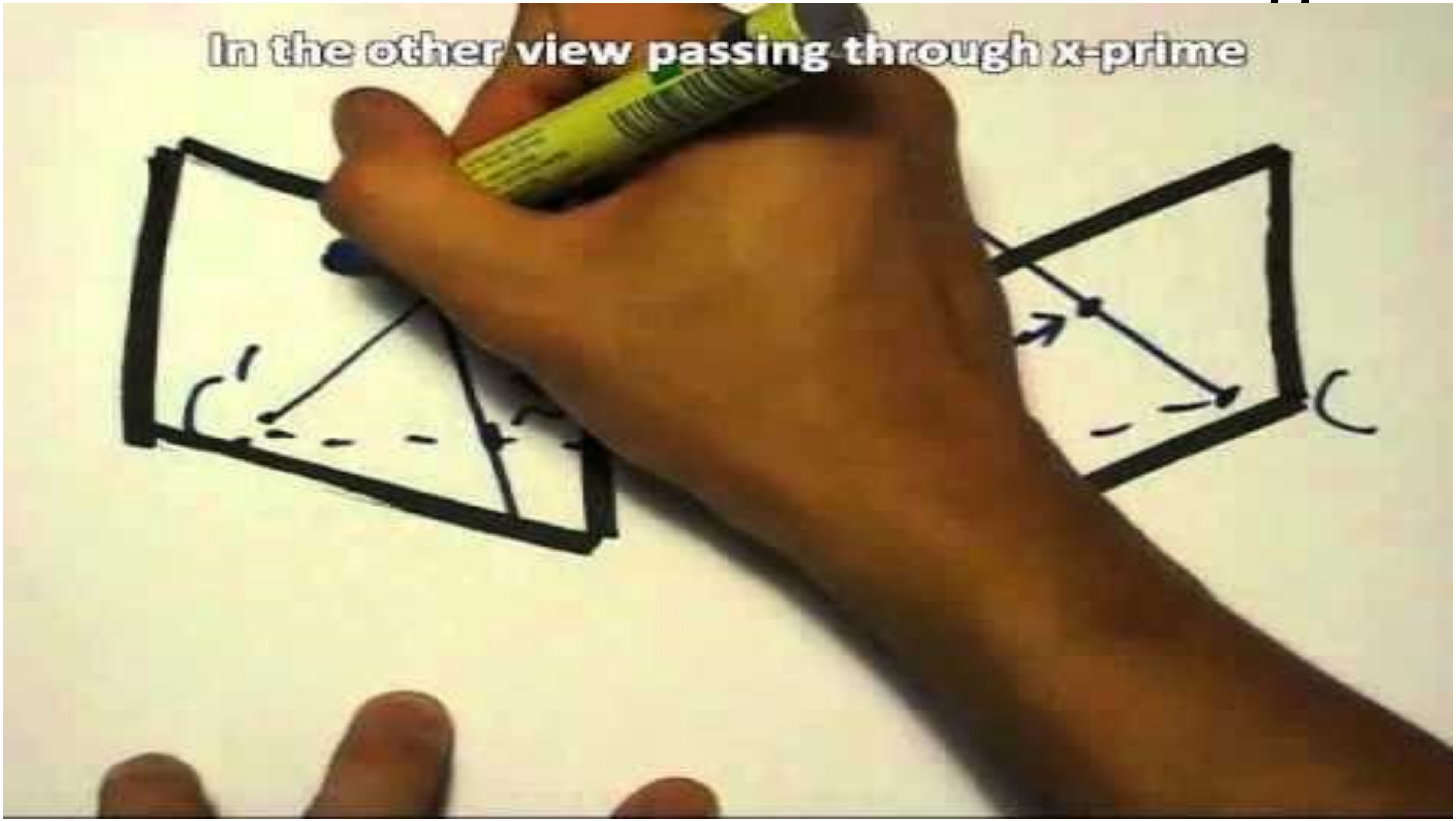
The approach

- The basic module is recovering 3D structure from 2 views with relative orientation (R, t) of cameras unknown. This has several steps:
  - Finding n corresponding points in the 2 views, i.e. image points which are the projections of the same point in the scene.
  - Estimate the E matrix (= $\hat{T}R$) from these point correspondences.
  - Extract (R, t).
  - Recover depth by triangulation.
- The outer loop combines information from all the cameras in a global coordinate system. Note that not all points will be seen by all cameras. This process is a nonlinear least squares optimization, called bundle adjustment. The error that is minimized is the re-projection error.
- For example, the 3D reconstruction of the Colosseum in Rome was based on 2 K images, and 800 K points.

Summary

- The basic module of recovering 3D structure from 2 views with relative orientation (R, t) of cameras unknown can be implemented using the Longuet-Higgins 8 point algorithm.
- The outer loop combines information from all the cameras in a global coordinate system using bundle adjustment. The error that is minimized is the re-projection error. The big idea is that given the guessed 3d positions of a point, one can predict image plane 2d positions in any camera where it is visible. We wish to minimize the squared error between this predicted position and the actual position, summed over all cameras and over all points.
- Lots of engineering has gone into making these approaches work. Read Szeliski's book, Chapter 7, for more.

The Fundamental Matrix Song

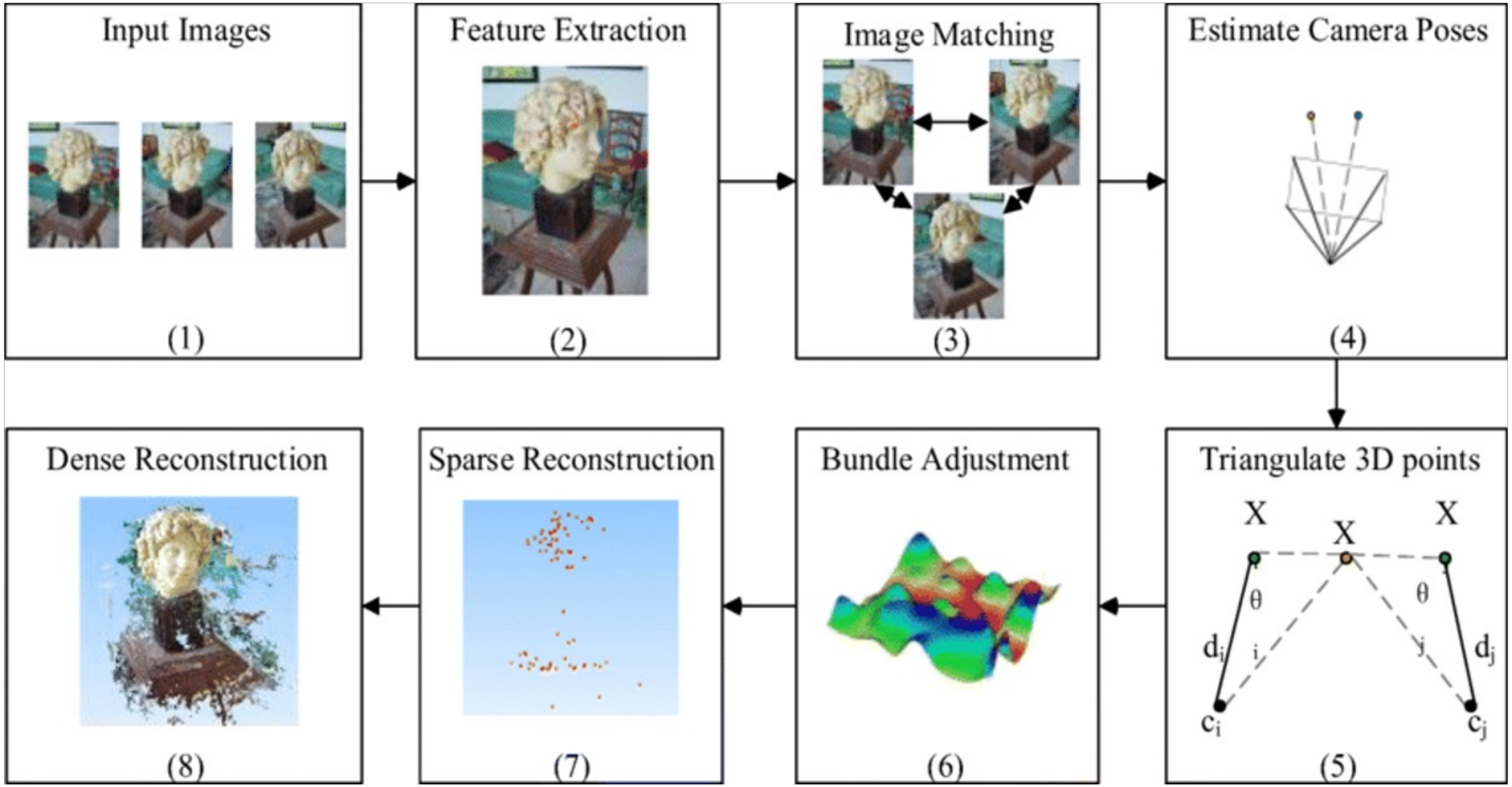


http://danielwedge.com/fmatrix/
https://www.youtube.com/watch?time_continue=8&v=DgGV3l82NTk&feature=emb_title

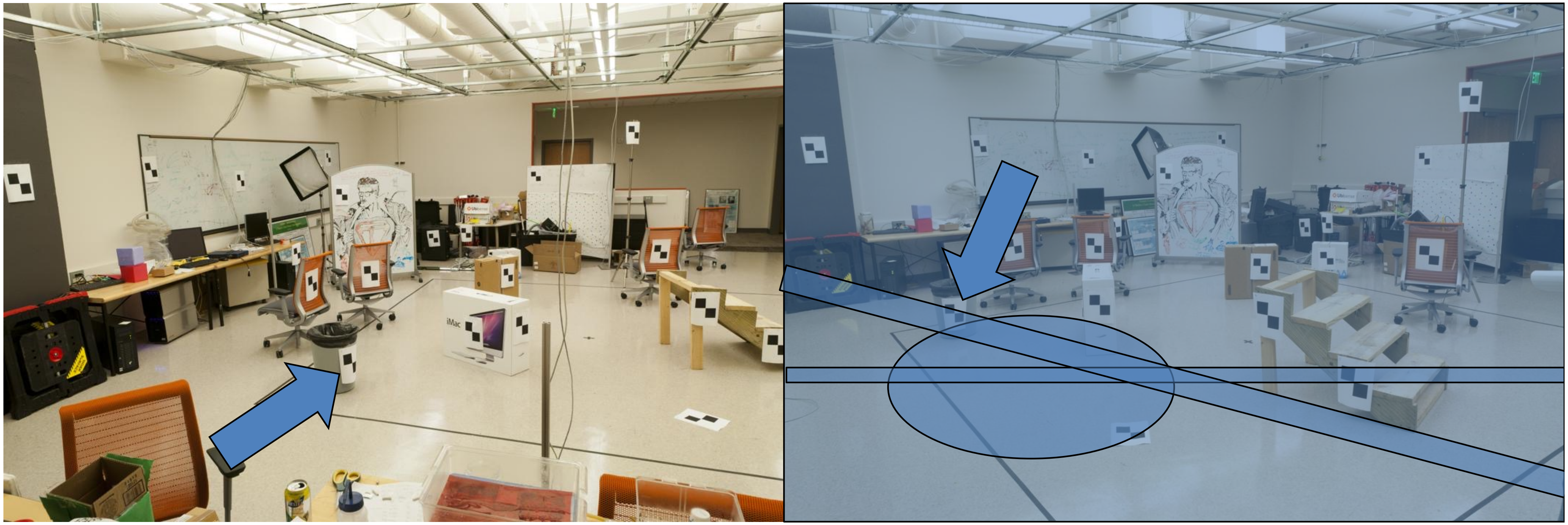
Break

In practice..

- Many images and lots of engineering



Epipolar Geometry helps you search correspondences

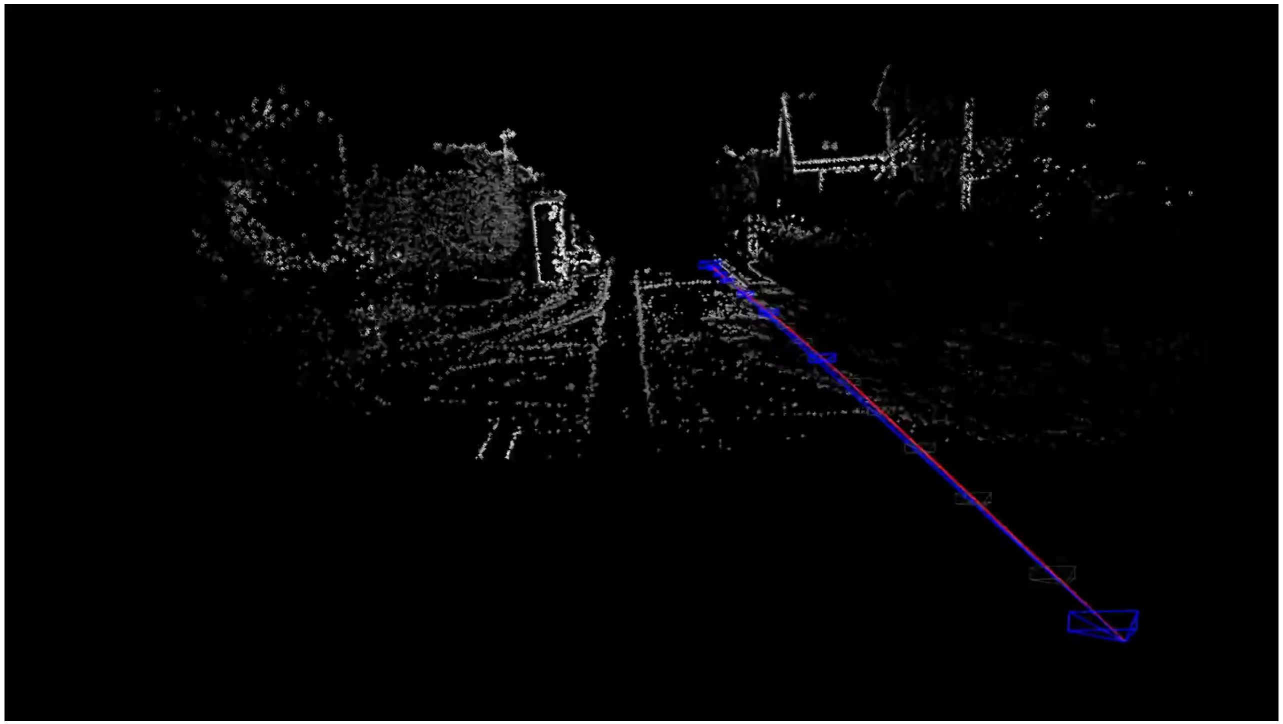


Knowing camera helps you find the right corepondences, bc they have to be on the epipolar line.
In practice you do RANSAC with Essential matrix (using current inliners)

Slide credit: James Hays

Visual Simultaneous Localization and Mapping (V-SLAM)

- Main differences with SfM:
  - Continuous visual input from sensor(s) over time
  - Gives rise to problems such as loop closure
  - Often the goal is to be online / real-time



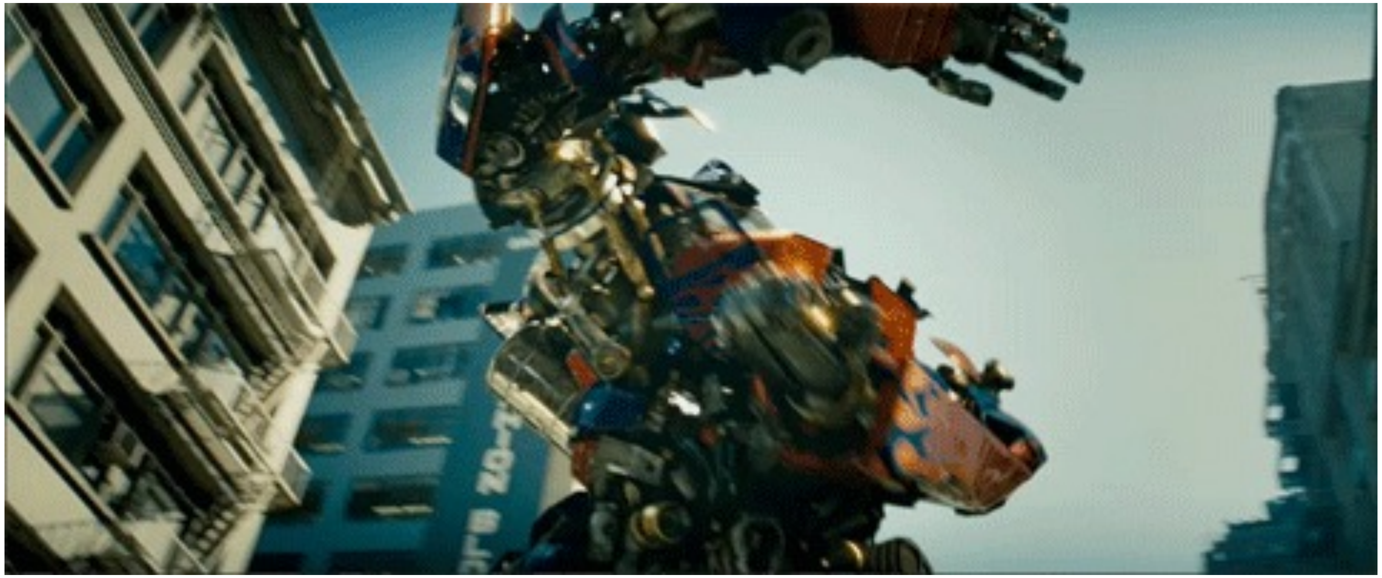
Video from Daniel Cremer’s Lab

Applications: Match Moving

Or Motion tracking, solving for camera trajectory

Integral for visual effects (VFX)

Why?



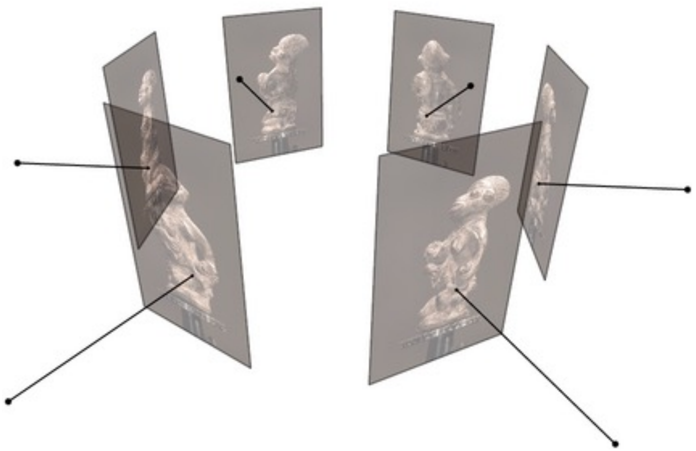
What if we want solid models?

- Up until now we only have points

Slide credit: Noah Snavely

Multi-view Stereo (Lots of calibrated images)

- Input: calibrated images from several viewpoints (known camera: intrinsics and extrinsics)
- Output: Dense 3D Model



Figures by Carlos Hernandez

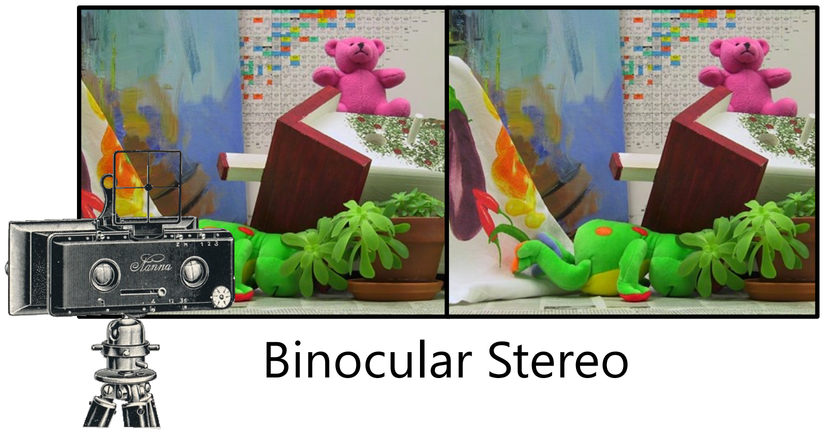
Slide credit: Noah Snavely

In general, conducted in a controlled environment with multi-camera setup that are all calibrated

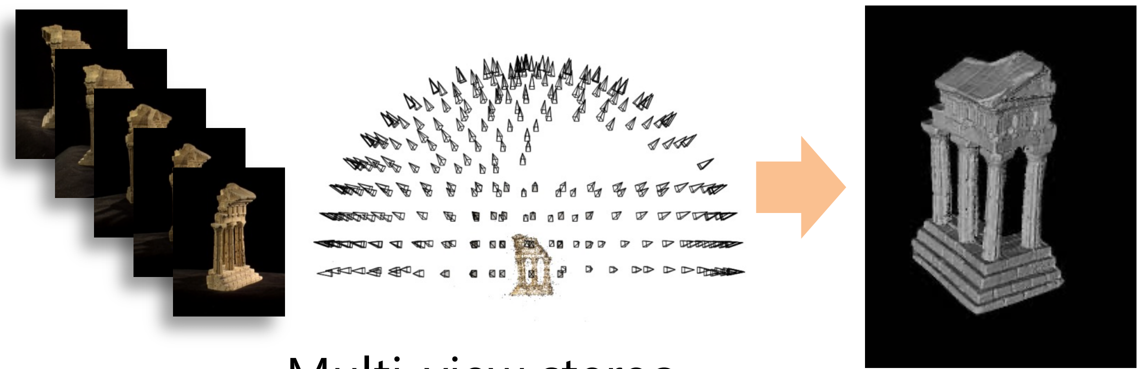
Slide credit: Noah Snavely

Multi-view Stereo

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



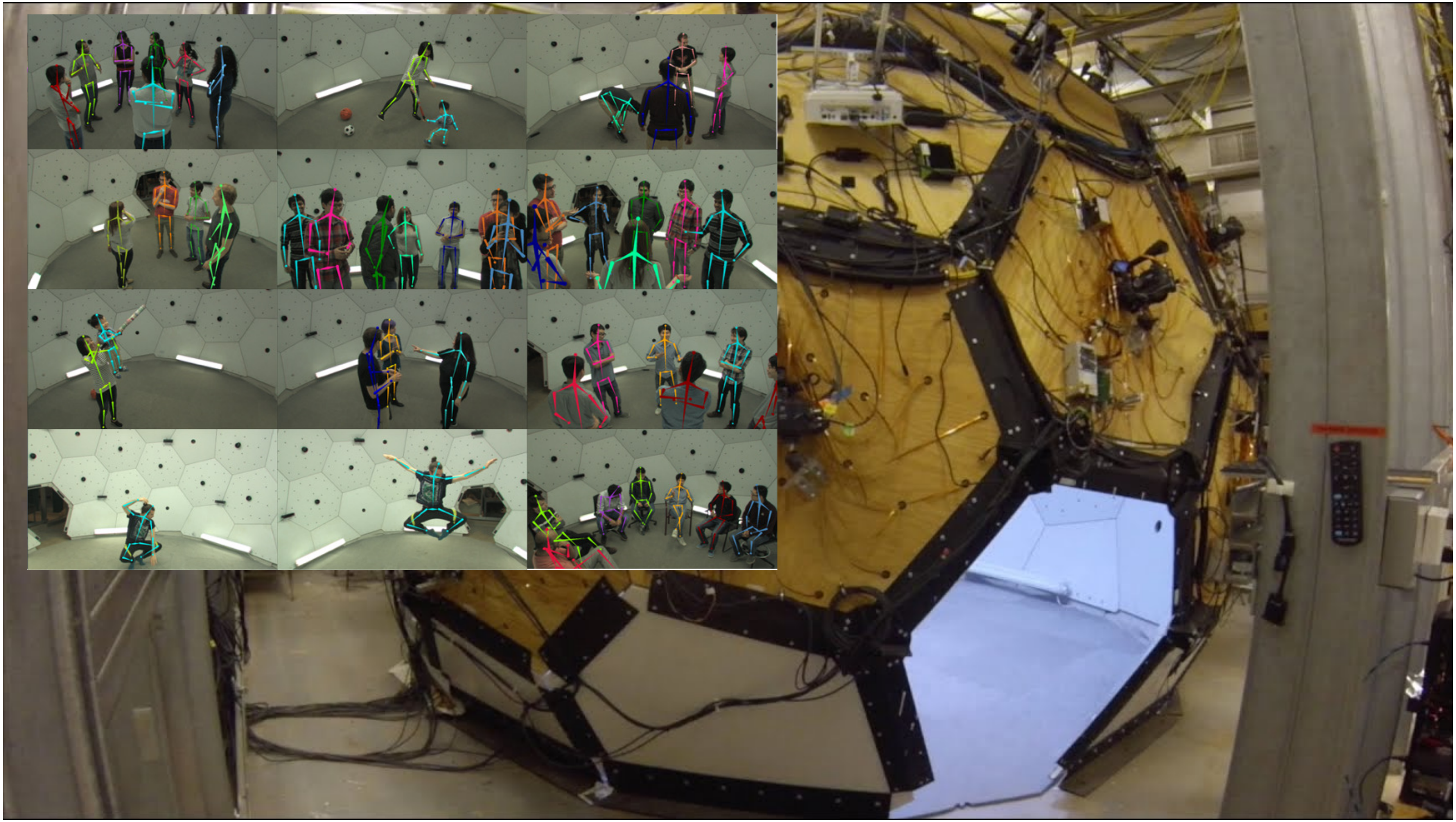
Binocular Stereo



Multi-view stereo

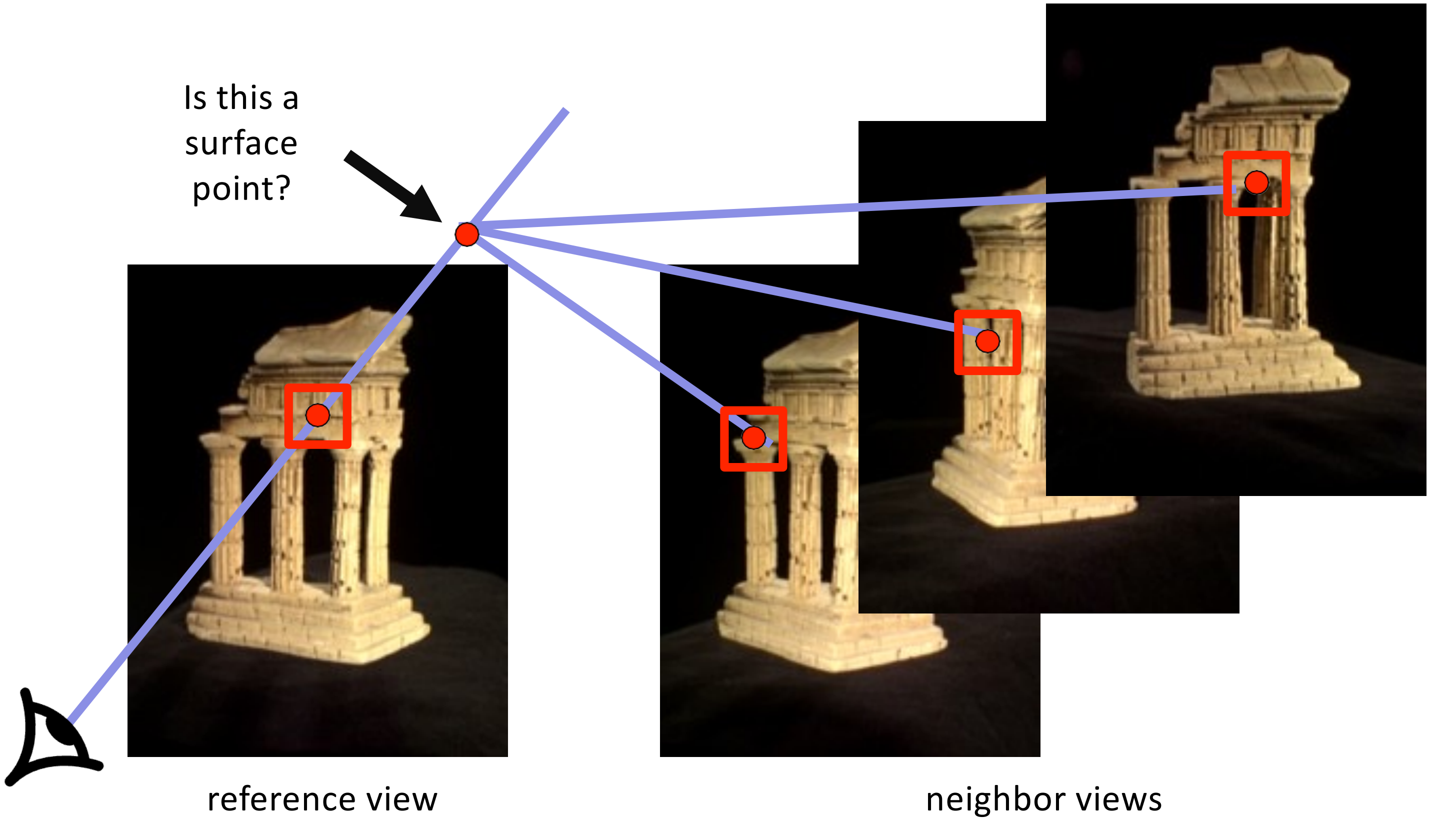
Slide credit: Noah Snavely

Examples: Panoptic studio



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

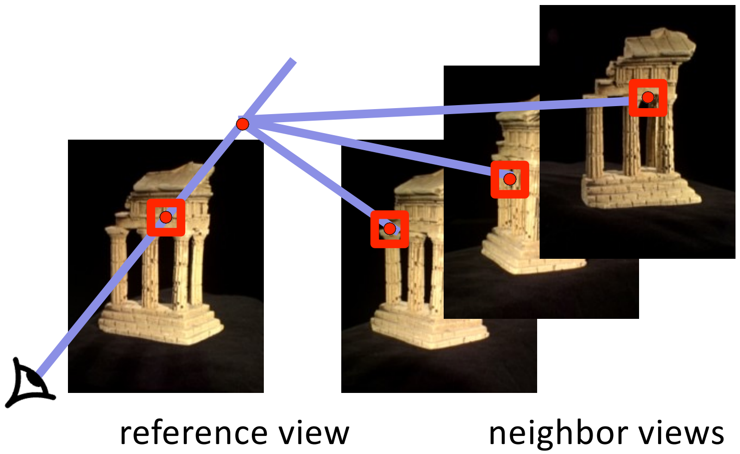
Multi-view stereo: Basic idea

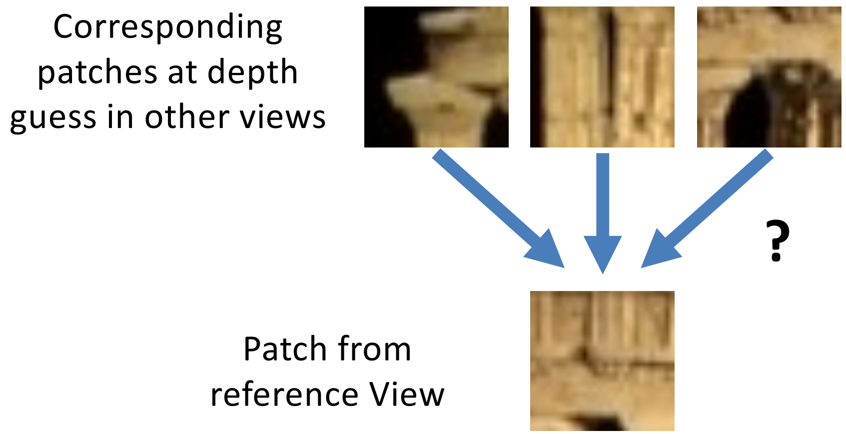


Source: Y. Furukawa

Multi-view stereo: Basic idea

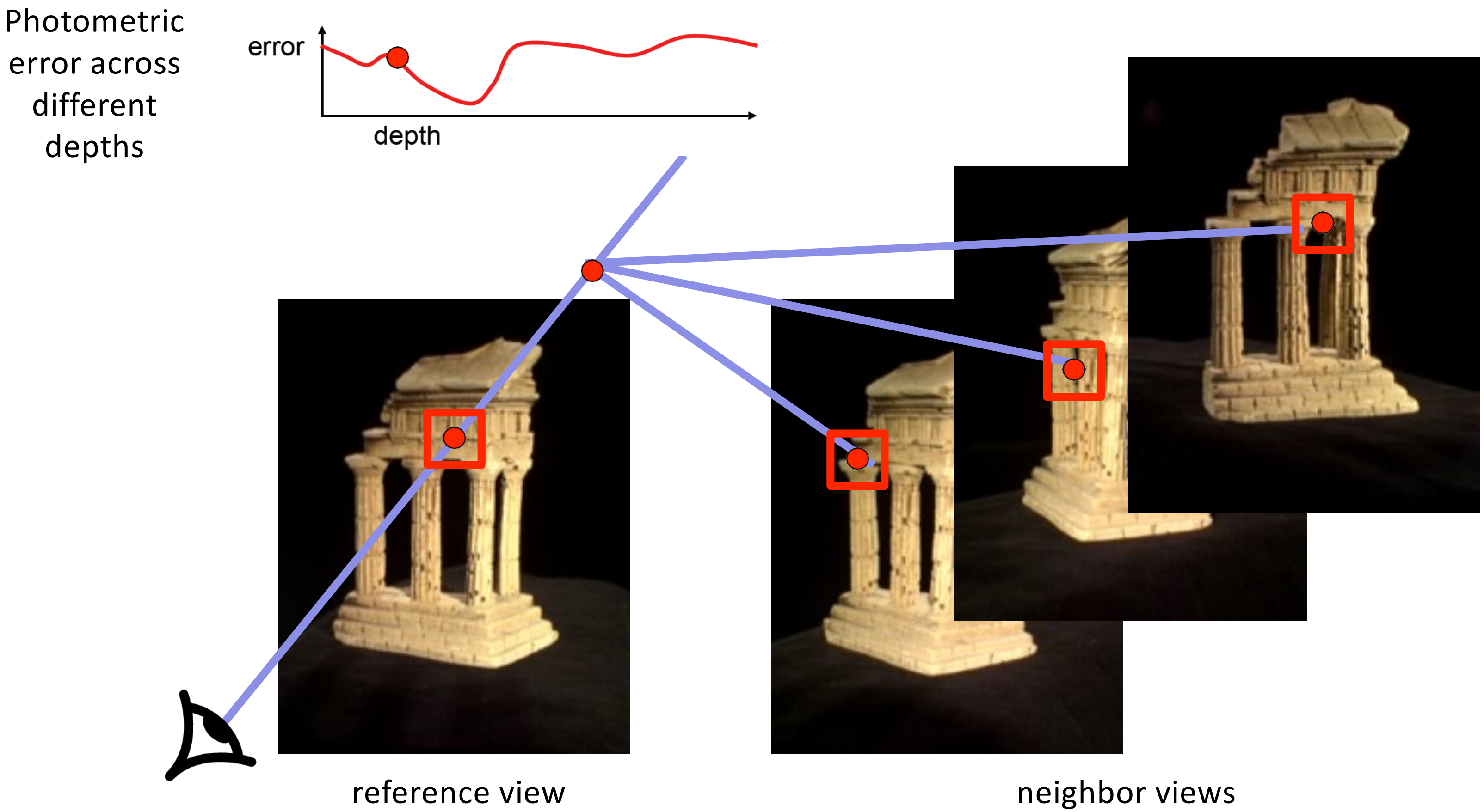
Evaluate the likelihood of geometry at a particular depth for a particular reference patch:





Source: Y. Furukawa

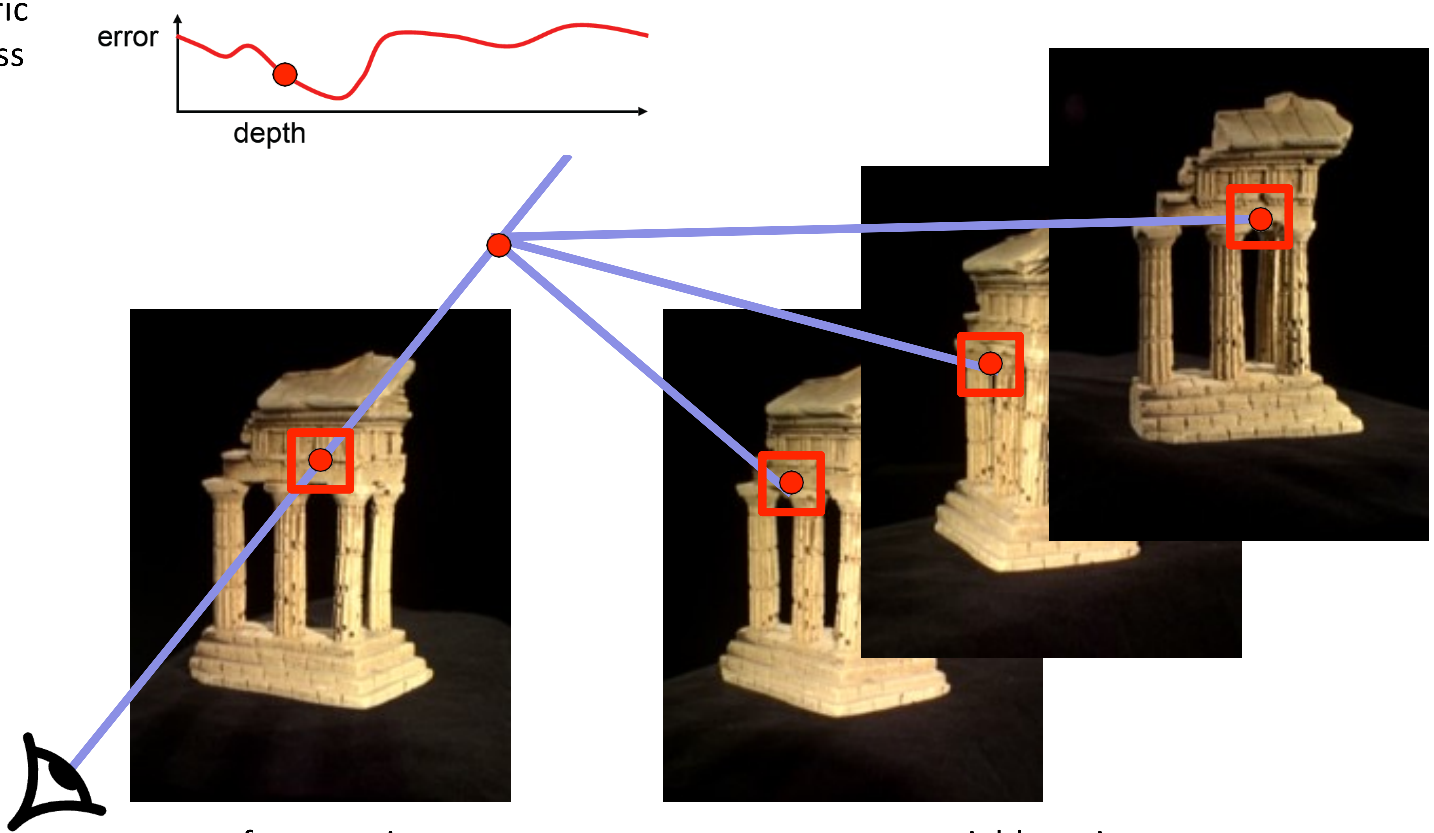
Multi-view stereo: Basic idea



Source: Y. Furukawa

Multi-view stereo: Basic idea

Photometric error across different depths



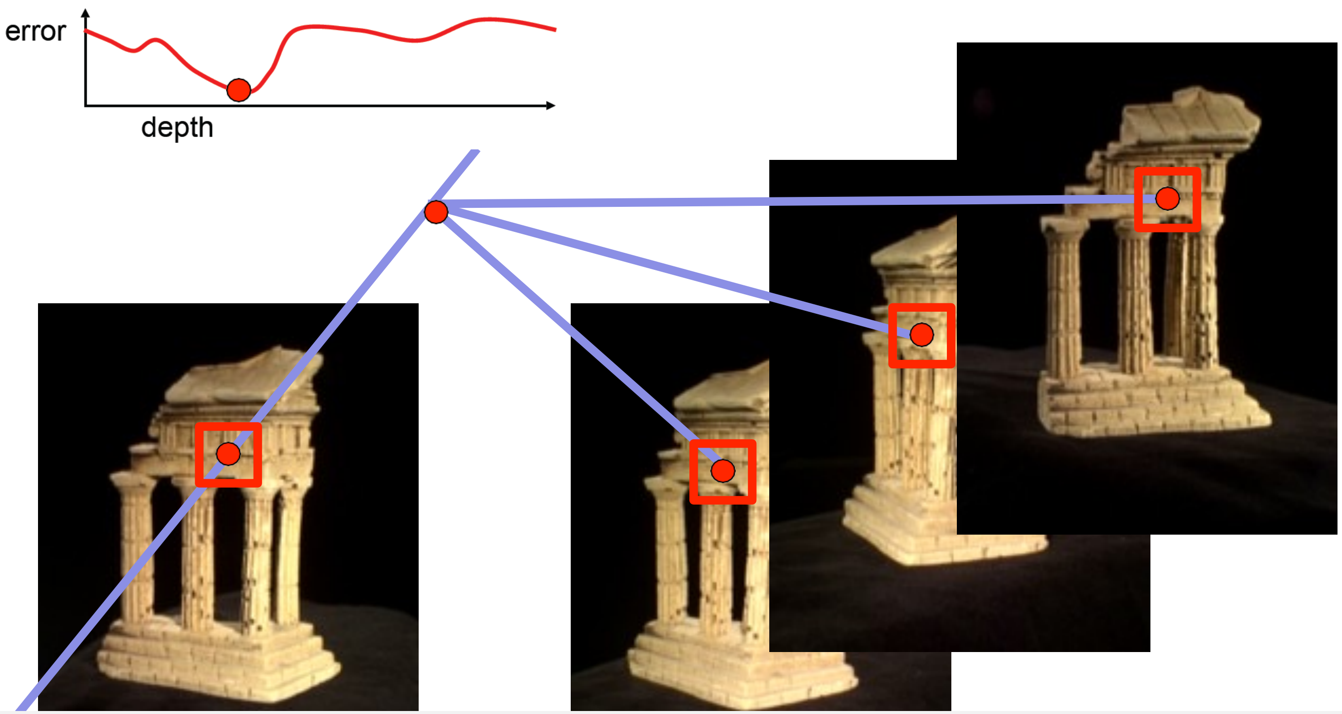
reference view

neighbor views

Source: Y. Furukawa

Multi-view stereo: Basic idea

Photometric error across different depths



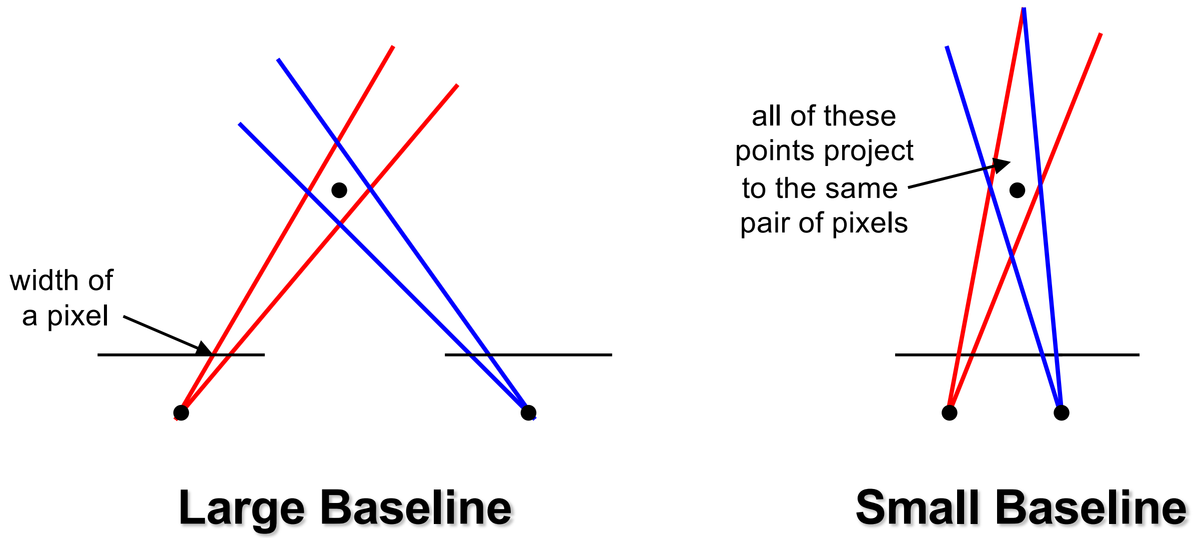
In this manner, solve for a depth map over the whole reference view

Multi-view stereo: advantages

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

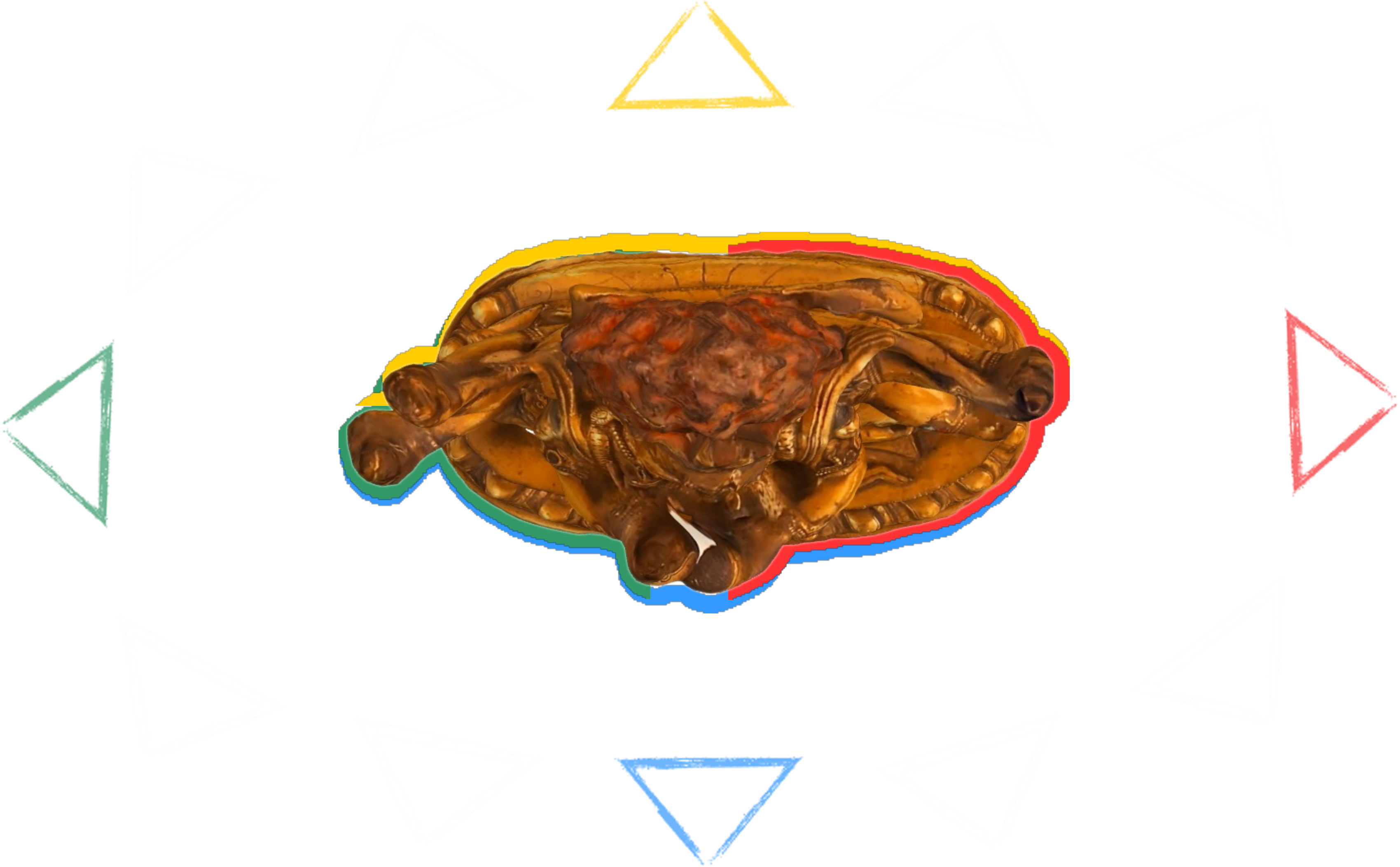
Source: Y. Furukawa

Choosing the baseline



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

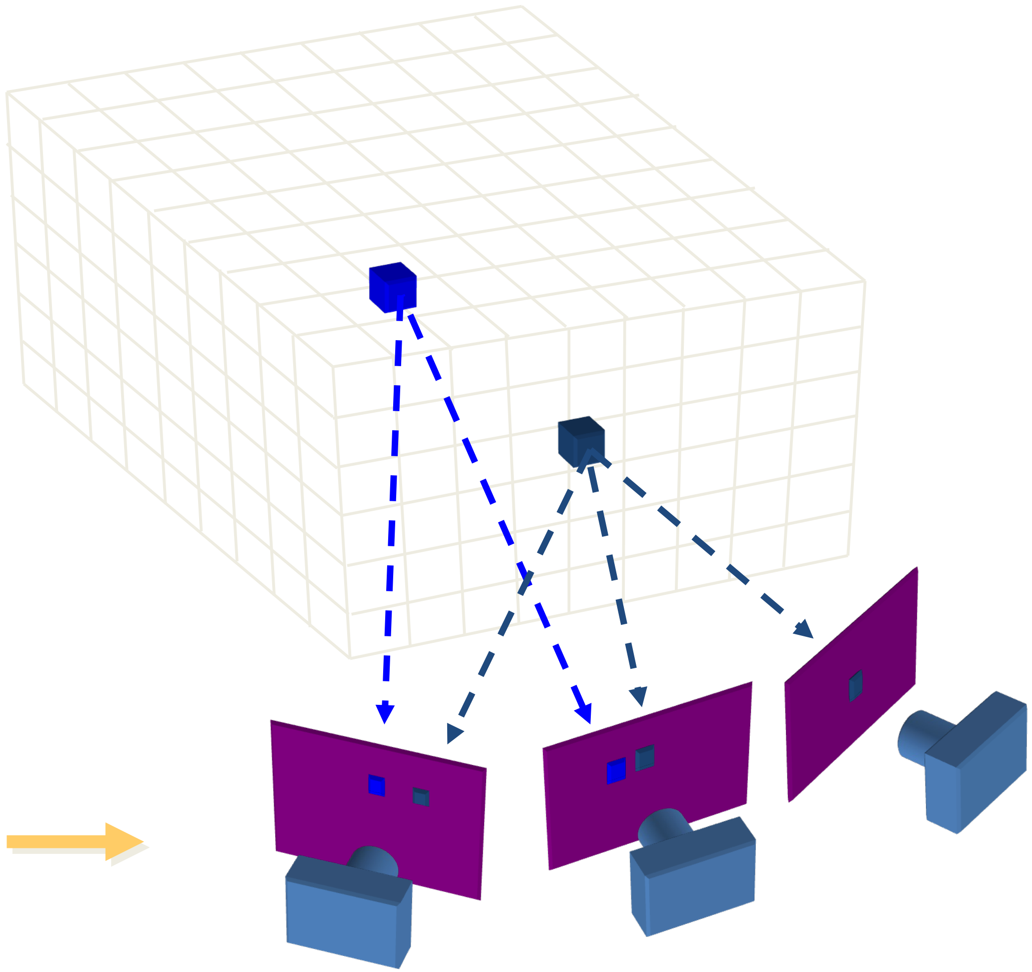
Slide credit: Noah Snavely



Volumetric stereo

Discretized Scene Volume

Input Images (Calibrated)



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

Space Carving

- Space Carving Algorithm
  - Initialize to a volume V containing the true scene
  - Choose a voxel on the outside of the volume
  - Project to visible input images
  - Carve if not photo-consistent
  - Repeat until convergence

K. N. Kutulakos and S. M. Seitz, **[A Theory of Shape by Space Carving]**, *ICCV* 1999

Space Carving Results

Input Image (1 of 45)

Reconstruction

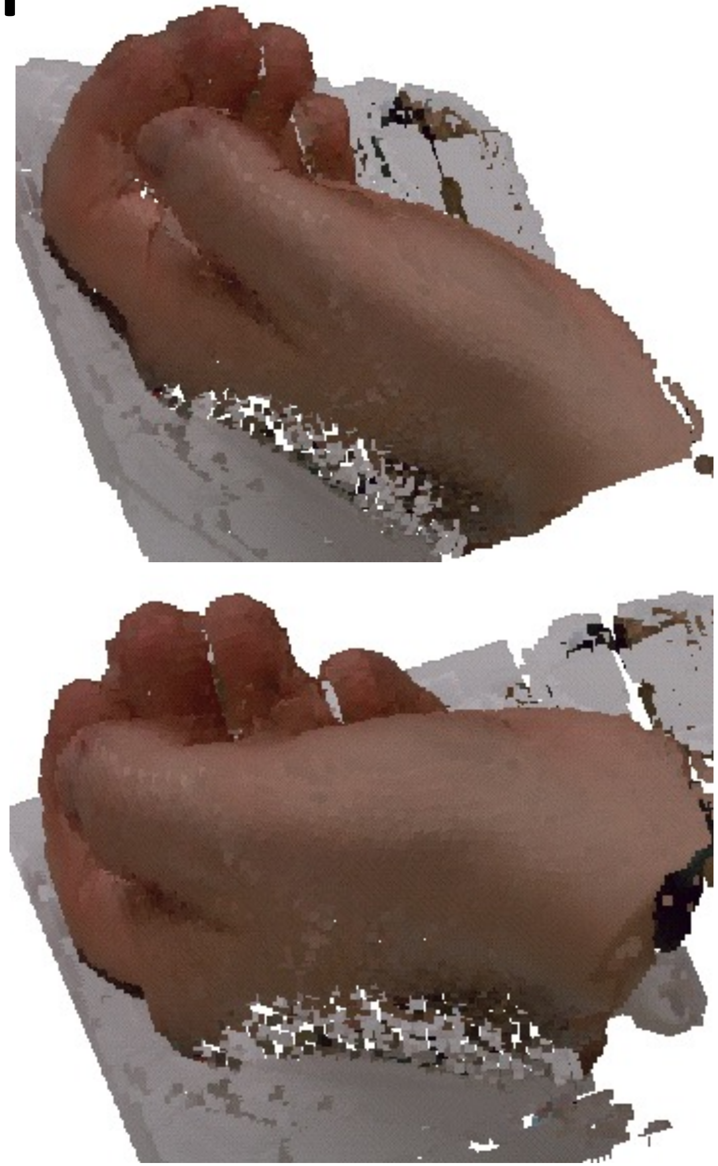
Reconstruction

Reconstruction

Source: S. Seitz

Space Carving Results

Input Image (1 of 100)



Reconstruction

Source: S. Seitz

Tool for you: COLMAP

https://github.com/colmap/colmap

A general SfM + MVS pipeline