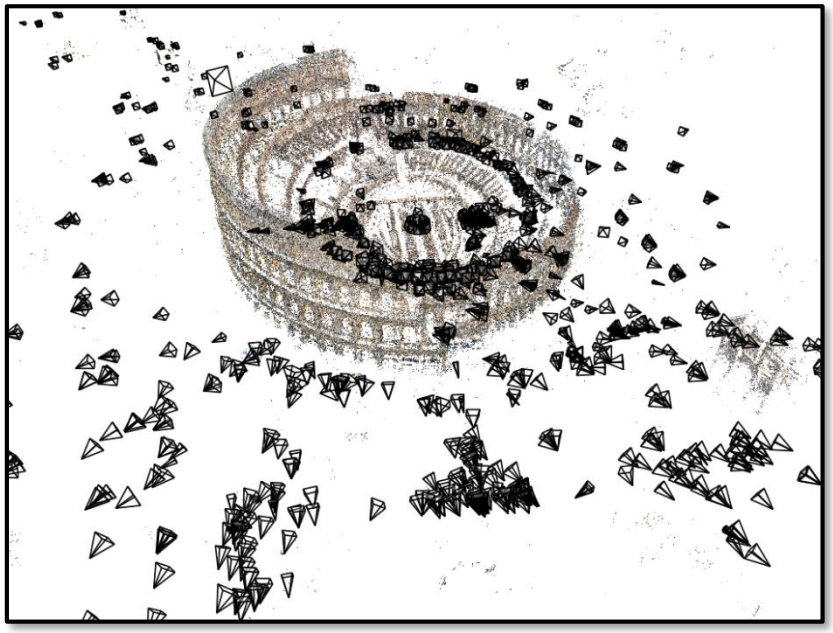
Structure from motion

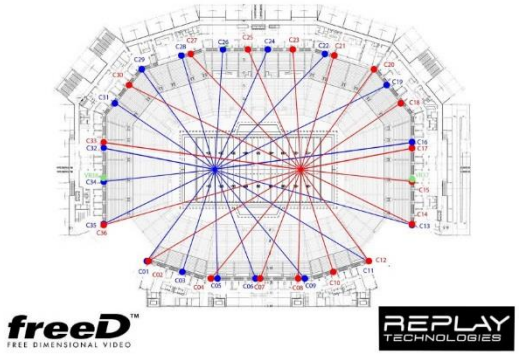


What is Point Cloud

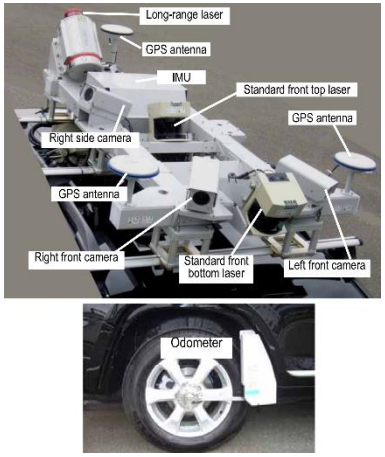
- A collection of Un-ordered points with
  - Geometry: expressed as [x, y, z]
  - Color Attributes: [r g b], or [y u v]
  - Additional info: normal, timestamp, …etc.
- Key difference from mesh: no order info

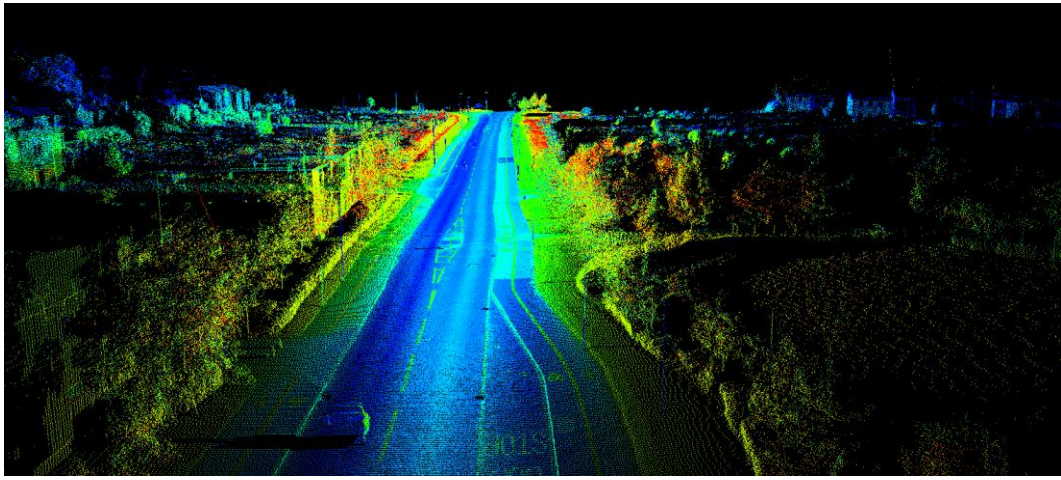
Point Cloud Capture

- Passive: Camera array stereo depth sensor



- Active: LiDAR, mmWave, TOF sensors





Credit: Zhu Li

Shape From X

- Recovery of 3D (shape) from one or two 2D images

Structure from motion

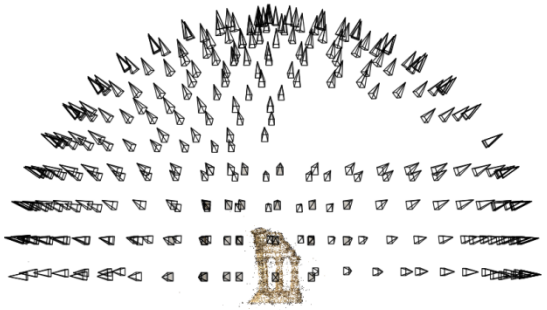
- Given many images, how can we
  a) figure out where they were all taken from?
  b) build a 3D model of the scene?

This is (roughly) the **structure from motion** problem

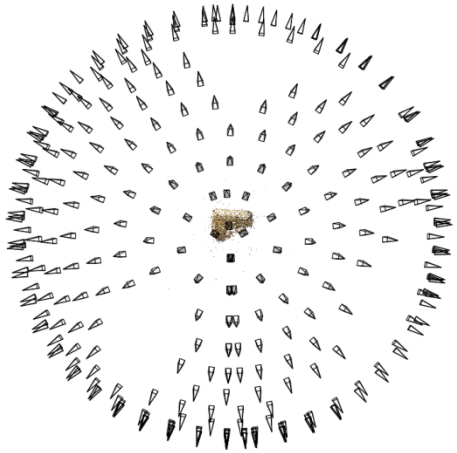
Applications

- Object Recognition
- Robotics
- Computer Graphics
- Image Retrieval
- Localization

Structure from motion



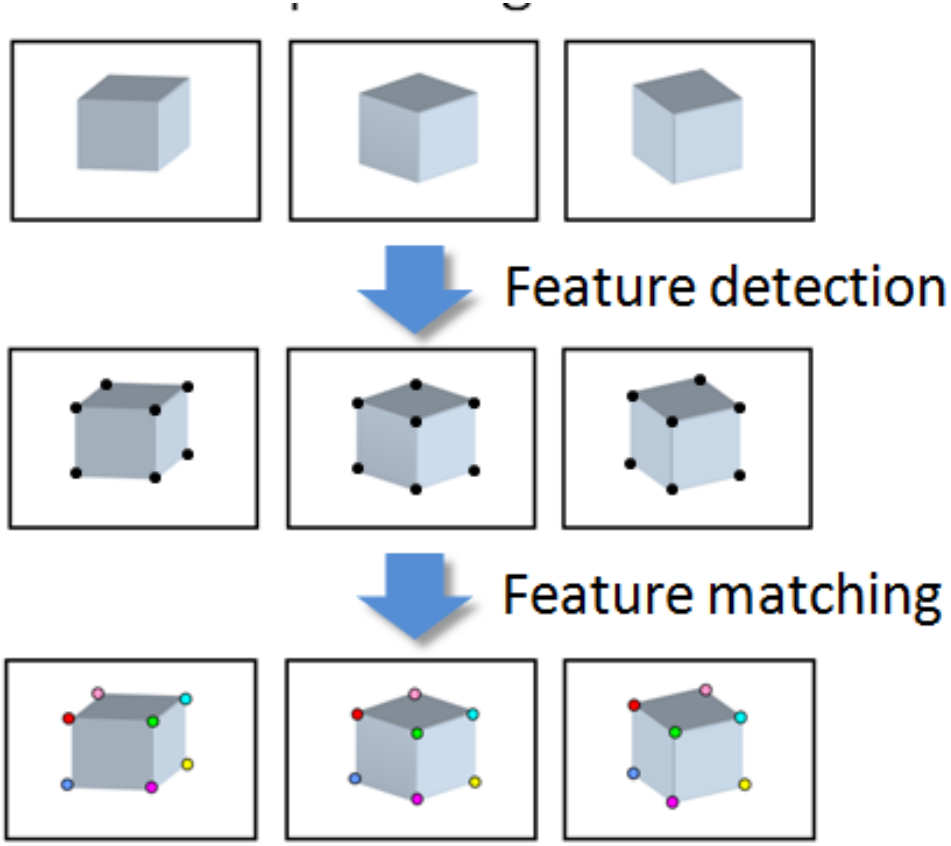
Reconstruction (side)



(top)

- Input: images with points in correspondence
  $p_{i,j} = (u_{i,j}, v_{i,j})$
- Output
  - structure: 3D location $\mathbf{x}_i$ for each point p_i
  - motion: camera parameters $\mathbf{R}_j$, $\mathbf{t}_j$
- Objective function: minimize *reprojection error*

Input



Camera calibration & triangulation

- Suppose we know 3D points
  - and have matches between these points and an image
  - How can we compute the camera parameters?

- Suppose we have known camera parameters, each of which observes a point
  - How can we compute the 3D location of that point?

Structure from motion

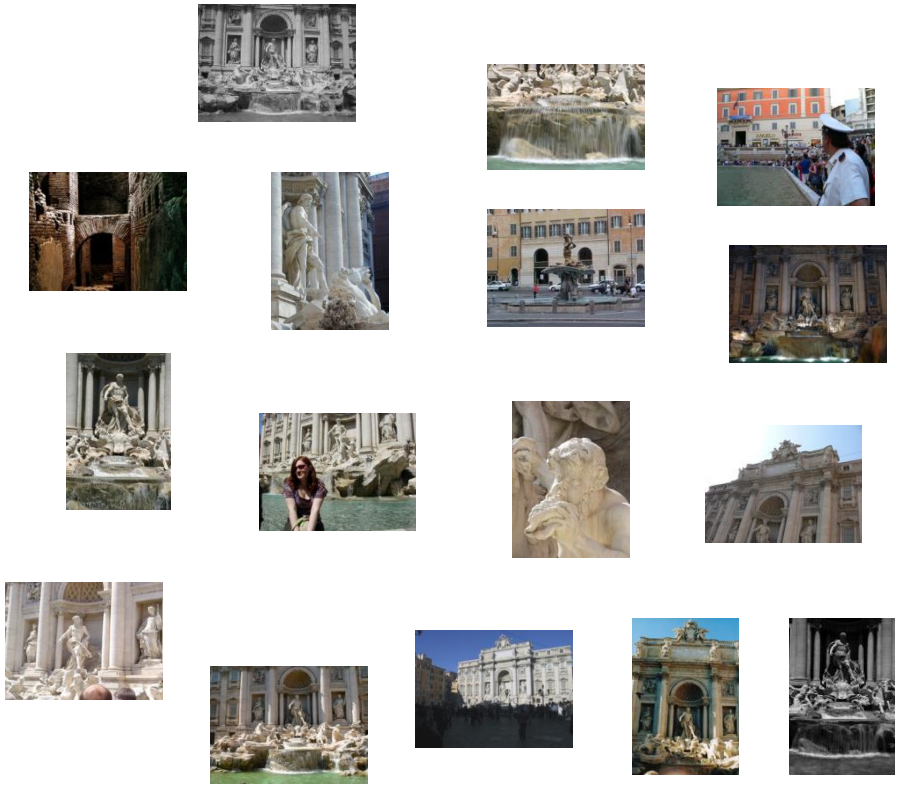
- SfM solves both of these problems *at once*
- A kind of chicken-and-egg problem
  - (but solvable)

First step: how to get correspondence?

- Feature detection and matching

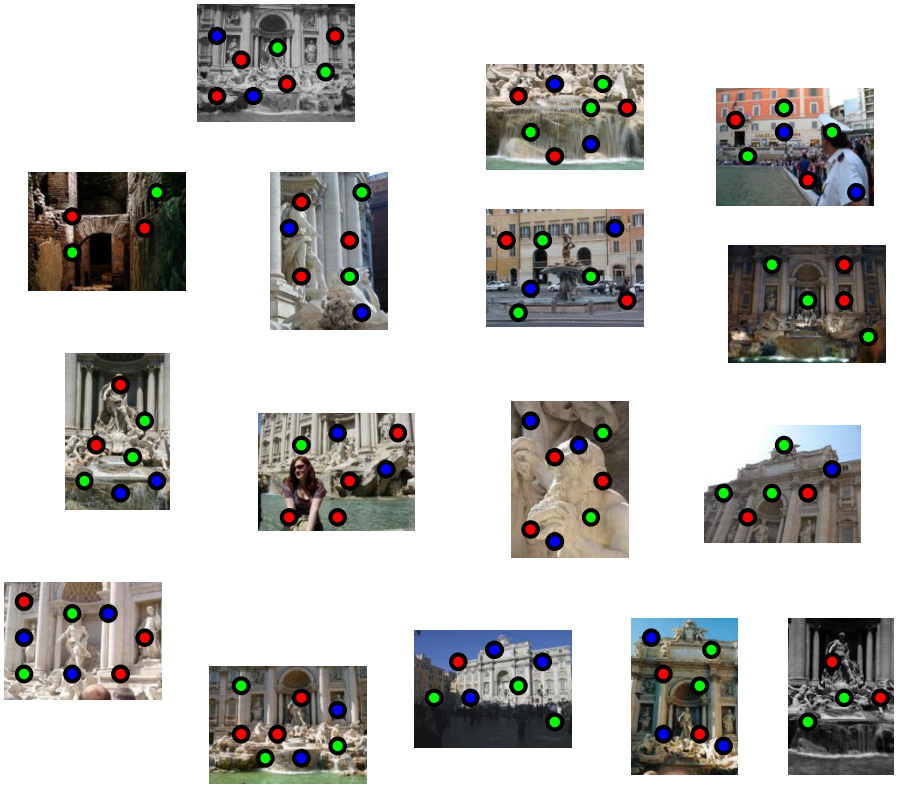
Feature detection

Detect features using SIFT [Lowe, IJCV 2004]



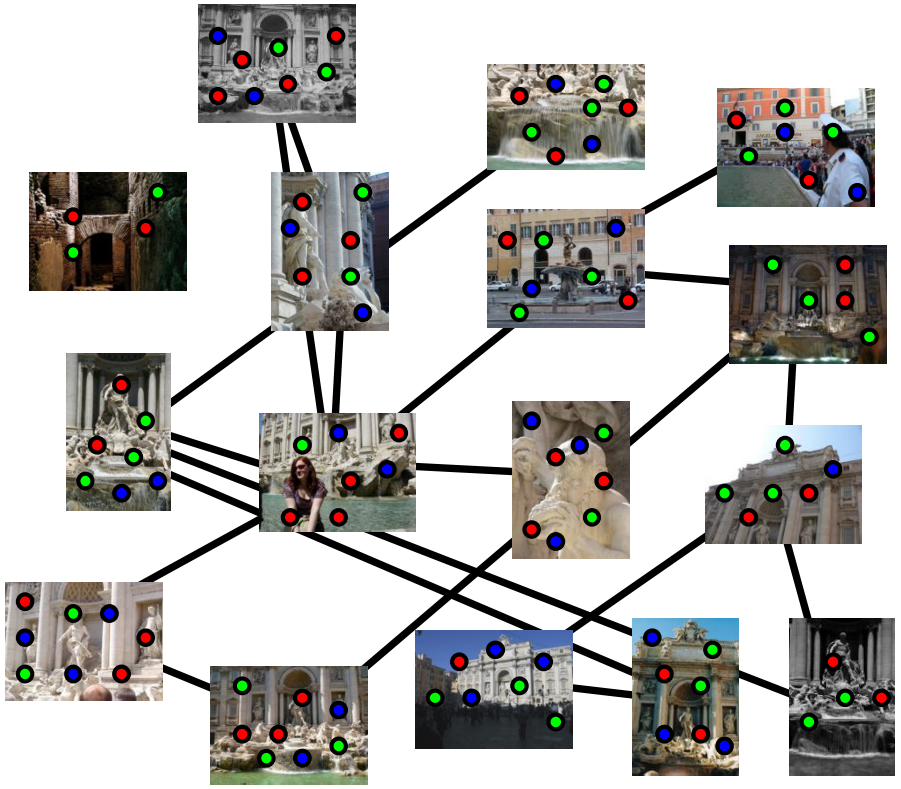
Feature detection

Detect features using SIFT [Lowe, IJCV 2004]



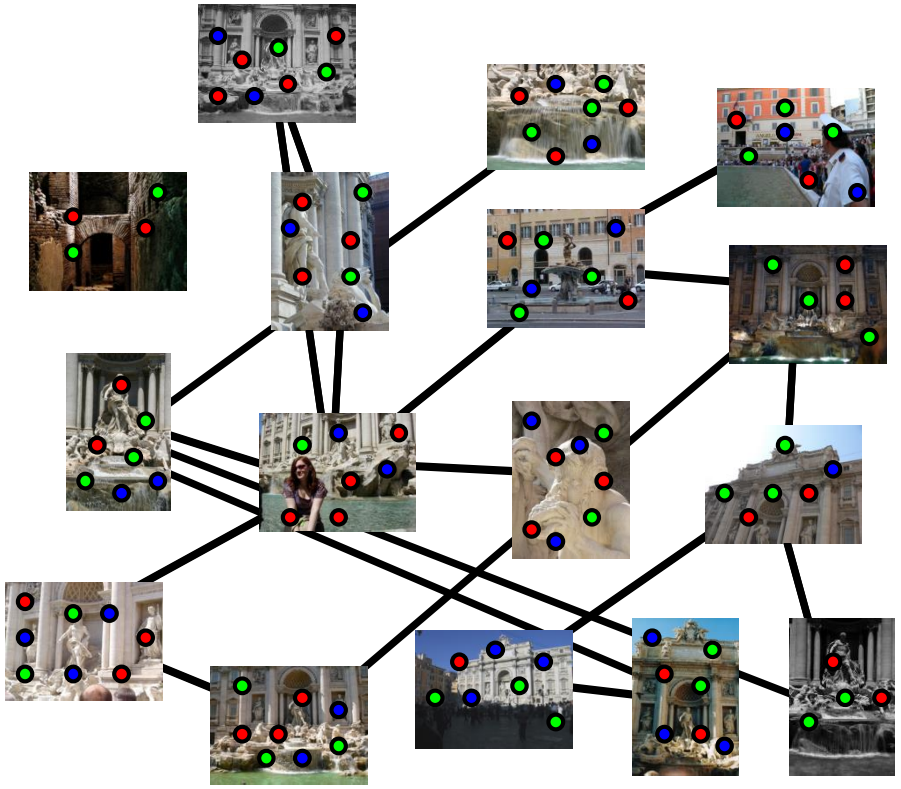
Feature matching

Match features between each pair of images



Feature matching

Estimate fundamental matrix between each pair (with RANSAC)



Correspondence estimation

- Link up pairwise matches to form connected components of matches across several images

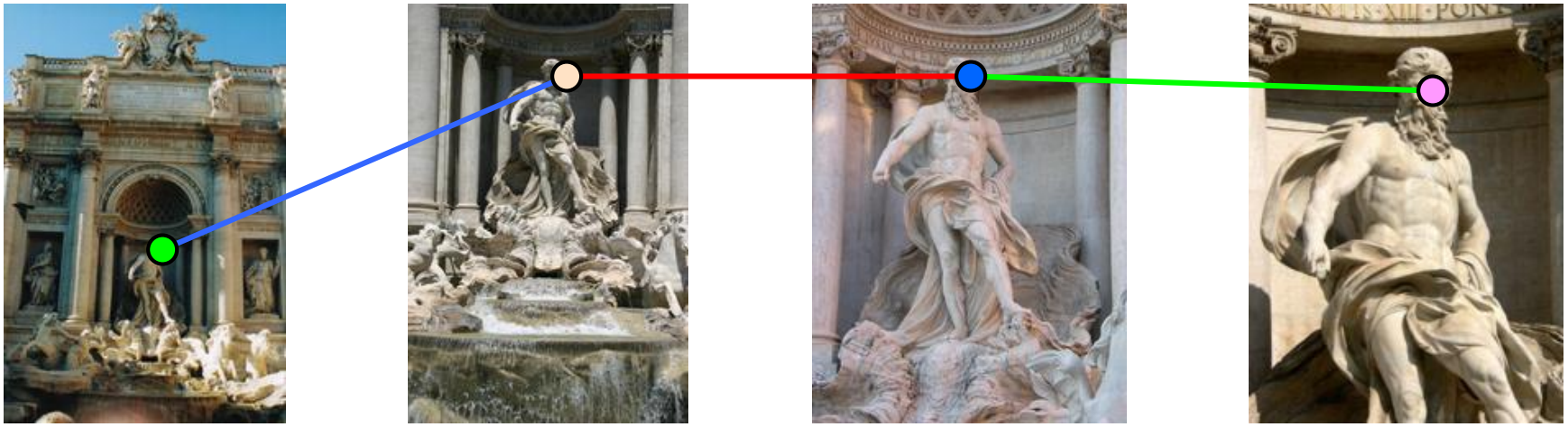
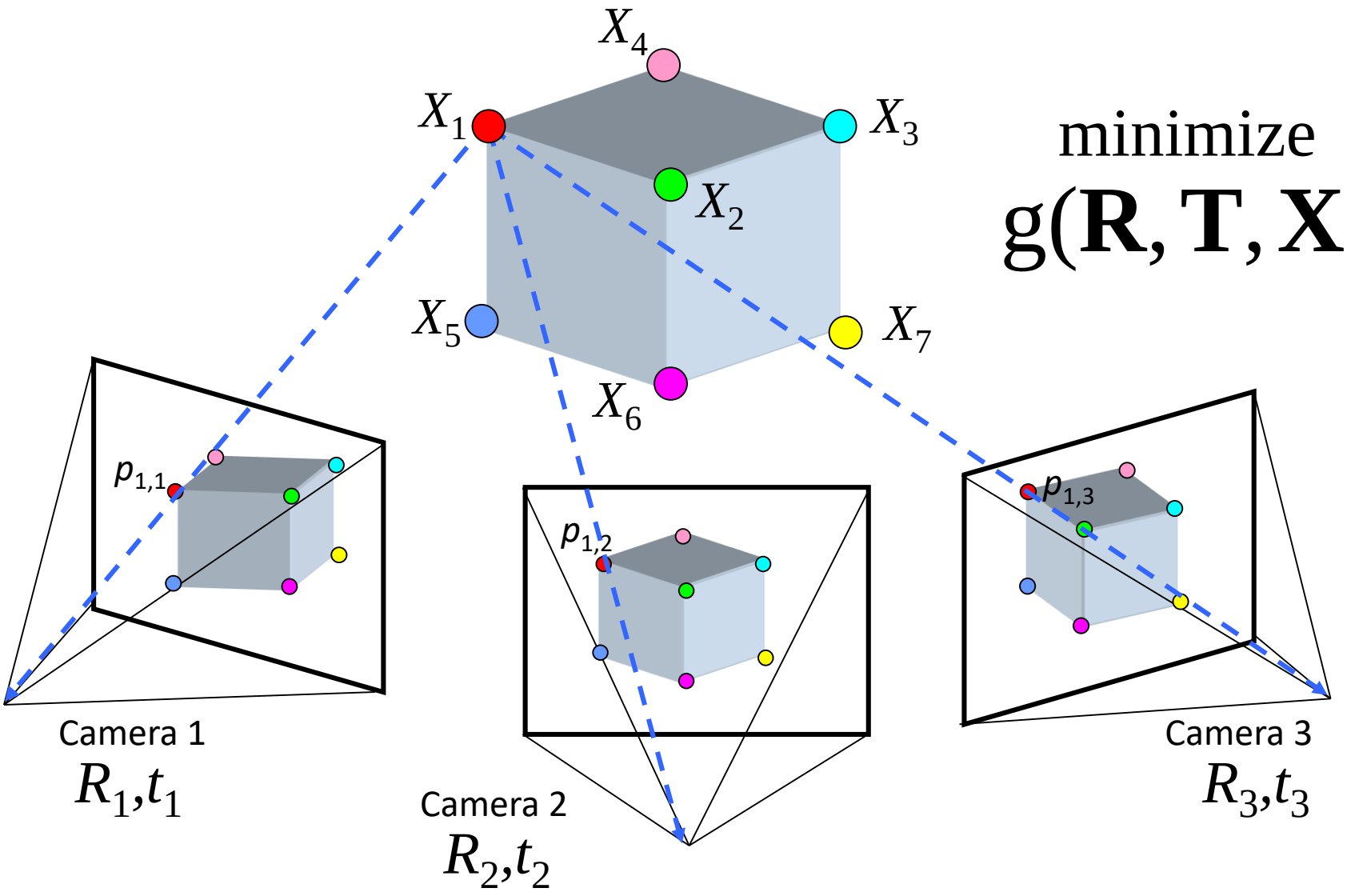


Image 1 Image 2 Image 3 Image 4

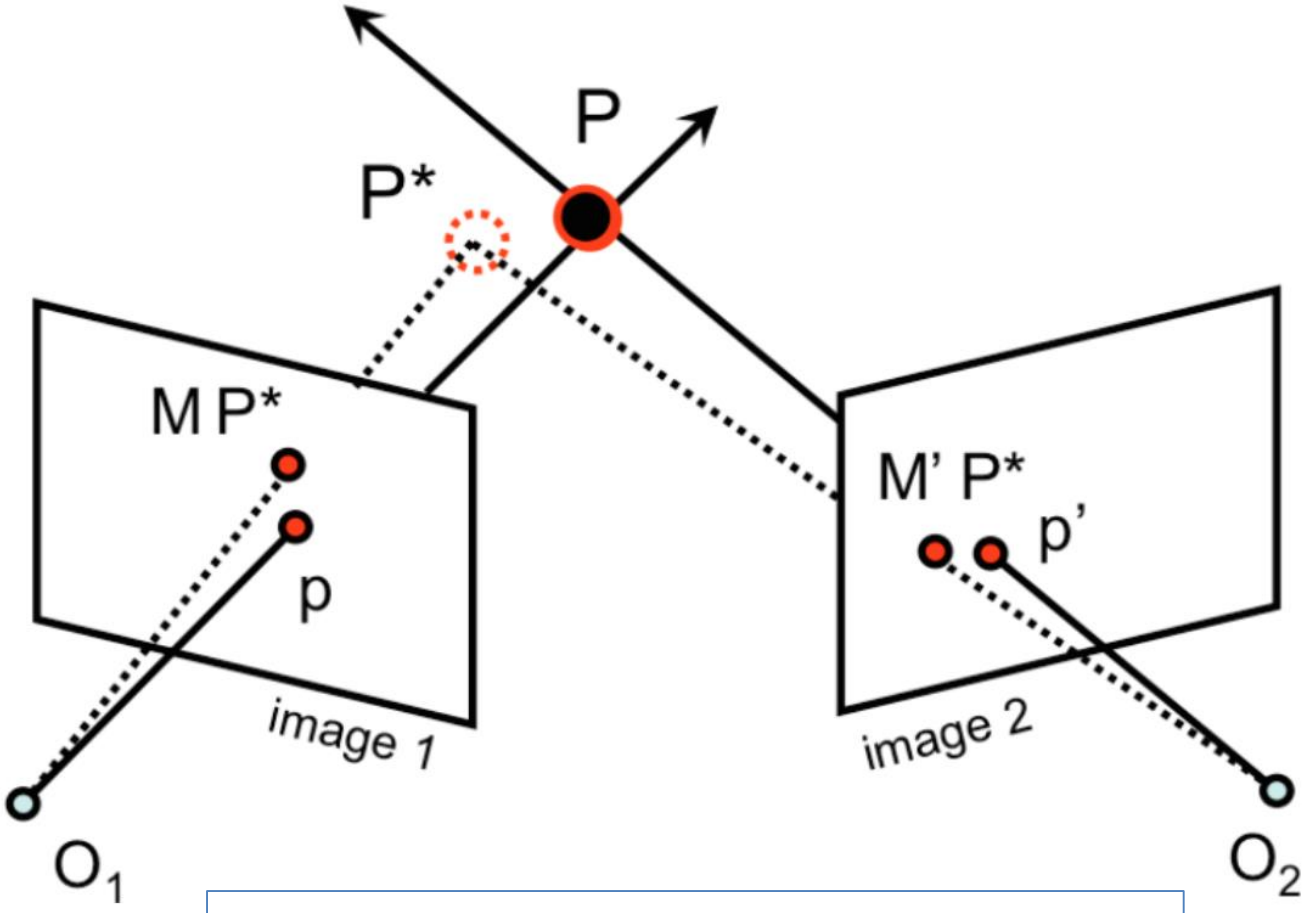
Structure from motion



minimize

$$g(\mathbf{R}, \mathbf{T}, \mathbf{X})$$

Re-projection Error



$$\min \|M\hat{P} - p\|^2 + \|M'\hat{P} - p'\|^2$$

Problem size

- What are the variables need to be solved? R t P

- Trevi Fountain collection

  466 input photos

  \+ > 100,000 3D points

  = very large optimization problem

Constraints vs #Unknowns

$$\arg\min_{\{P_i\},K,\{R_j\},\{T_j\}} \sum_{j=1}^{M}\sum_{i=1}^{N} (u_i^j - f(K,R_j,T_j,P_i))^2 + (v_i^j - g(K,R_j,T_j,P_i))^2$$

M camera poses

N points

2MN point constraints

Structure from motion

- Minimize sum of squared reprojection errors:

$$g(\mathbf{X}, \mathbf{R}, \mathbf{T}) = \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} \cdot \left\| \mathbf{P}(\mathbf{x}_i, \mathbf{R}_j, \mathbf{t}_j) - \begin{bmatrix} u_{i,j} \\ v_{i,j} \end{bmatrix} \right\|^2$$

predicted image location

observed image location

indicator variable: is point *i* visible in image *j* ?

- Minimizing this function is called *bundle adjustment*