KENTECH
Korea Institute of Energy Technology

# Advanced Computer Vision
# Week 09

Nov. 1, 2022

Seokju Lee

Parts of slides are by Prof. In So Kweon and Prof. Shree Nayar

KENTECH
Korea Institute of Energy Technology

# Structure-from-Motion (SfM)

# Contents of Structure-from-Motion

**Multi-view 3D reconstruction**

- Pinhole camera model
- Two-view geometry
- Epipolar geometry
- Essential matrix
- Fundamental matrix
- Bundle adjustment
- ...

# Pinhole Camera Model

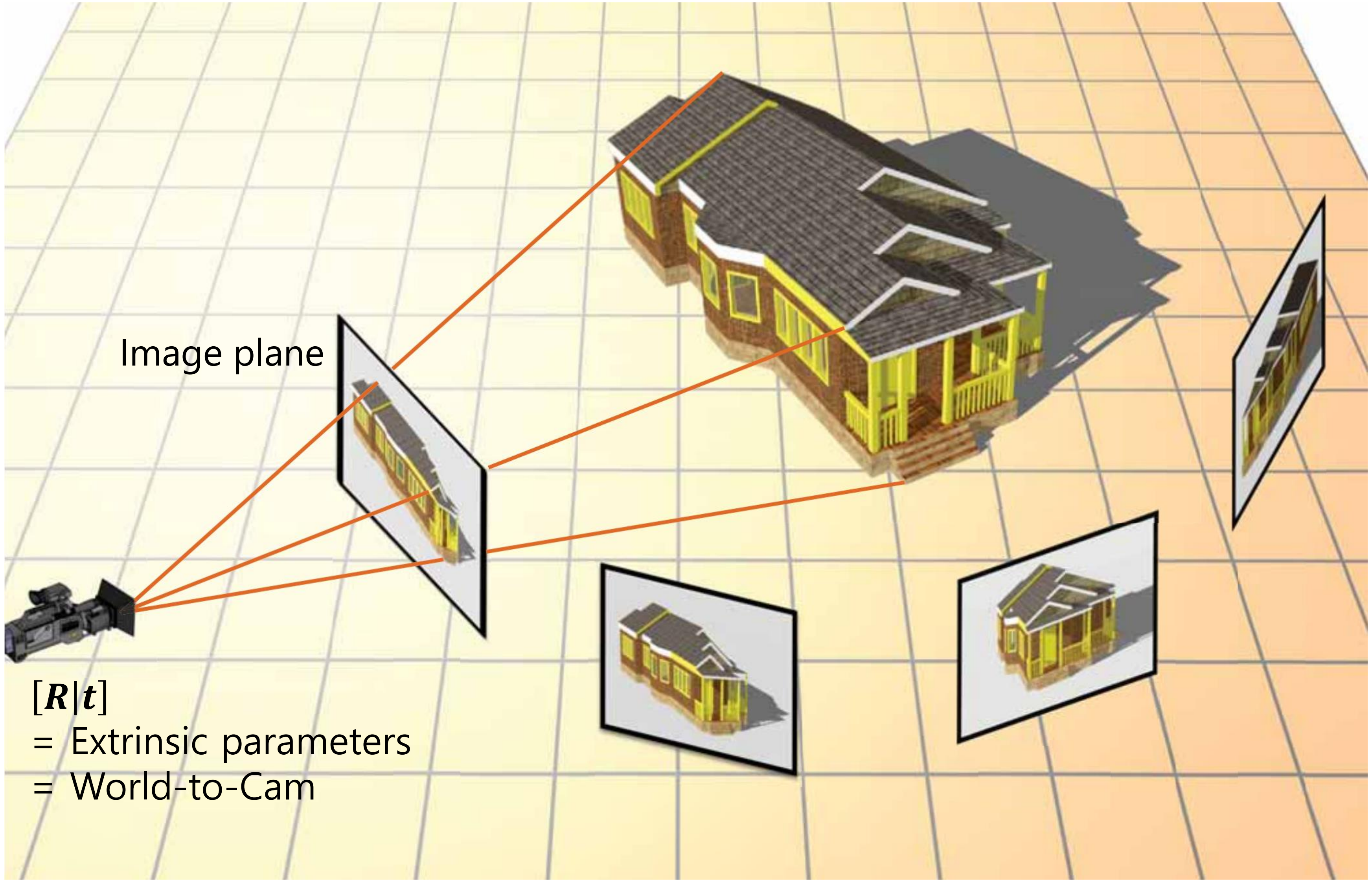
$$x = K[R|t]\mathbf{X} = \mathbf{PX}$$

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

$$= K[R|t]\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

- $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$: 3D point in the world coordinate
- $[\boldsymbol{R}|\boldsymbol{t}]$: extrinsic parameters to convert the world coordinate into the camera coordinate
- $\boldsymbol{K}$: intrinsic parameters to represent the camera characteristics
- $\boldsymbol{K}[\boldsymbol{R}|\boldsymbol{t}]$: camera projection matrix
- $(\boldsymbol{x}, \boldsymbol{y})$: 2D pixel location in the image plane
- $\boldsymbol{s}$: scale factor

# Multi-View 3D Reconstruction




http://www.cs.cmu.edu/~yaser/TutorialSlides-Integrated-small.pdf

# Structure-from-Motion Pipeline

Given only the 2D multi-view images of a scene,

recover the underlying 3D **structure** and the camera **motion**.


Schonberger, et al., “Structure-from-Motion Revisited”, CVPR 2016.

# Structure-from-Motion Pipeline

Q1. How to find **2D-3D points** to reconstruct?

Q2. How to find an **optimal** 3D structure and camera poses for multiple view?


Schonberger, et al., “Structure-from-Motion Revisited”, CVPR 2016.

# Structure-from-Motion Pipeline

Q1. How to find **2D-3D points** to reconstruct?

Q2. How to find an **optimal** 3D structure and camera poses for multiple view?

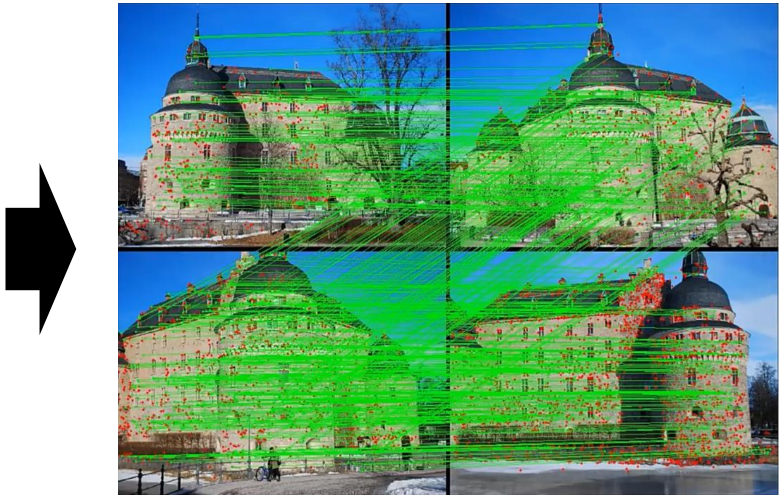

Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.

# Structure-from-Motion (1): Correpondence Search

## Correspondence search

- Feature extraction
- Matching
- Geometric verification

*A set of images*



*Feature extraction & matching (graph)*


Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.

# Structure-from-Motion (1): Correpondence Search

*c.f.) Homography matrix*

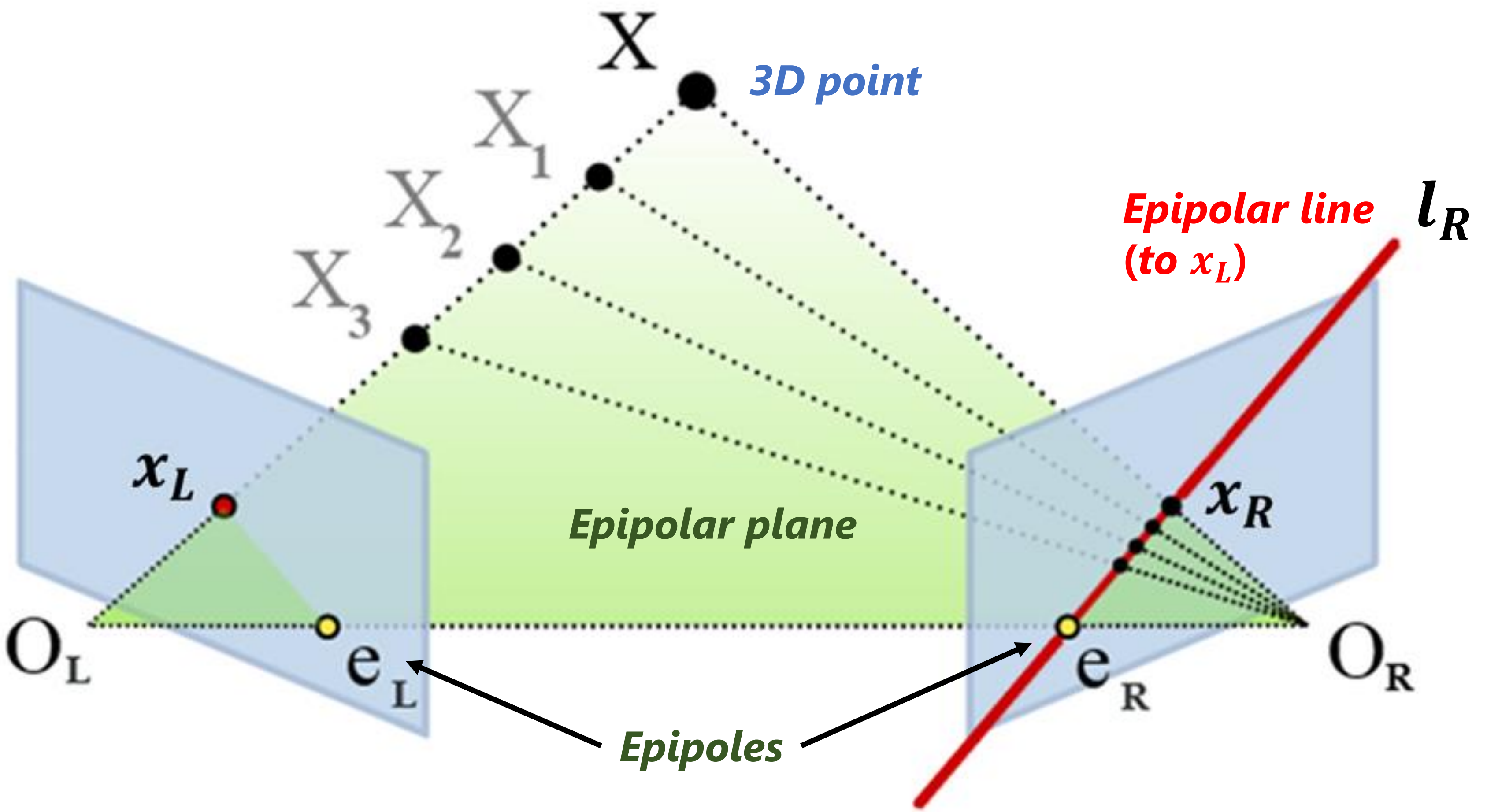
## Correpondence search

- Feature extraction
- Matching
- **Geometric verification**

- Epipolar geometry describes the relation of moving cameras **(5 or 8-point algorithm)** through the essential matrix, $\mathbf{E} \in \mathbb{R}^{3\times3}$, or the fundamental matrix, $\mathbf{F} \in \mathbb{R}^{3\times3}$
- If estimated $\mathbf{E}$ projects a sufficient number of features between the images, it is verified! **(RANSAC)**


Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.

# Epipolar Geometry

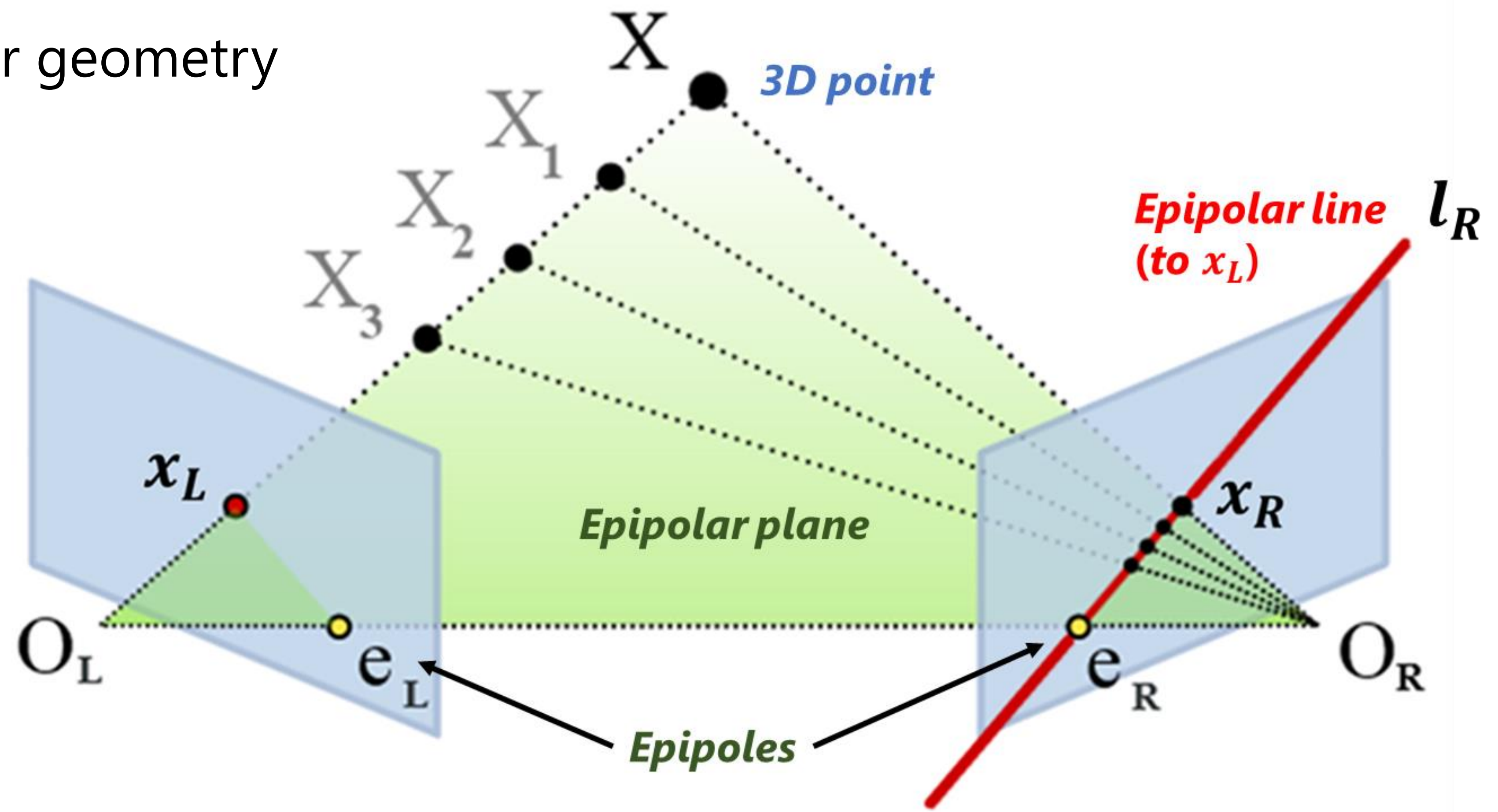



Potential matches for $\boldsymbol{x_L}$ are on the epipolar line $\boldsymbol{l_R}$

Schonberger, et al., “Structure-from-Motion Revisited”, CVPR 2016.

# Essential Matrix

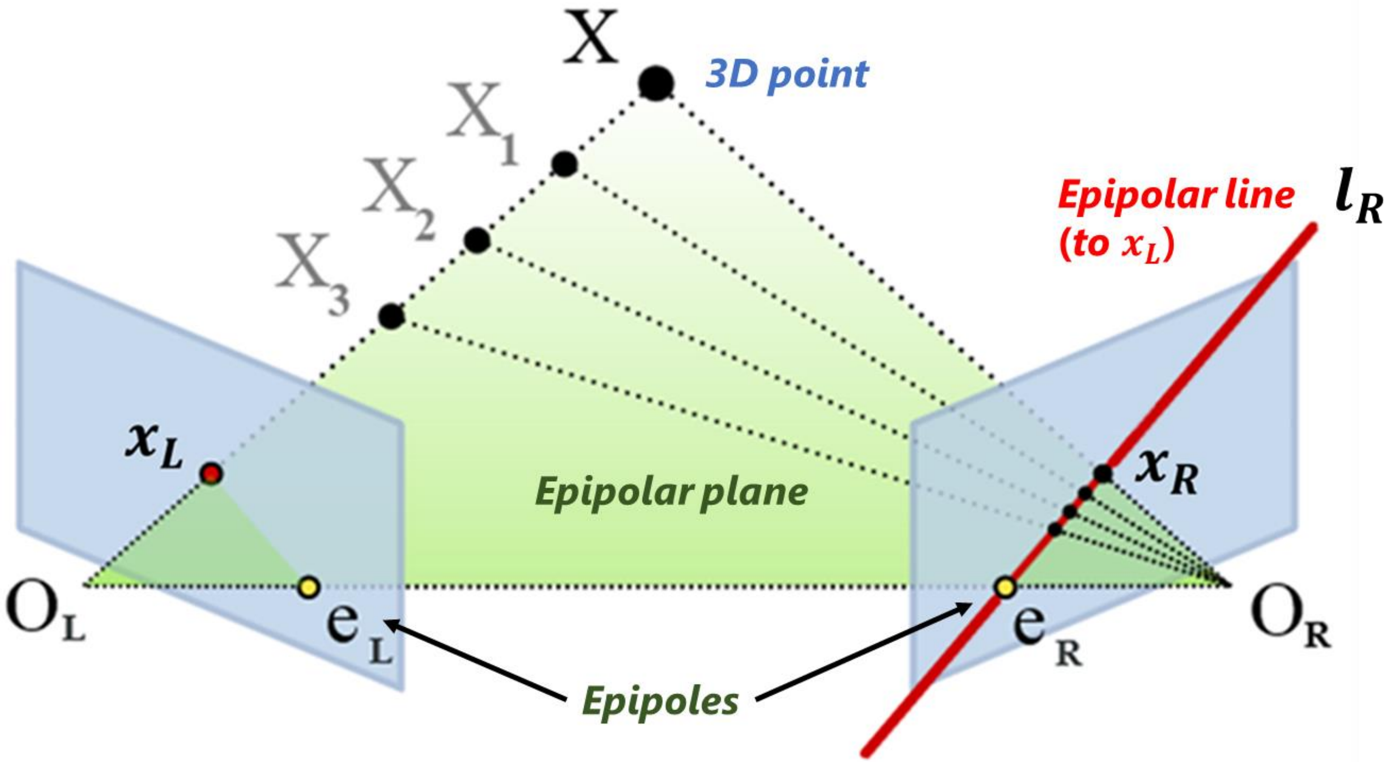
: Encodes epipolar geometry




Given a point $\boldsymbol{x_L}$ in one image, multiplying by the **essential matrix** $\mathbf{E} \in \mathbb{R}^{3\times3}$ will make the epipolar line in the right view: $\mathbf{E}\boldsymbol{x_L} = \boldsymbol{l_R}$

Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.

# Epipolar Constraint




For a **linear** equation,

$ax + by + c = 0$ in vector form $l = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$.

If the point $\boldsymbol{x_R}$ is on the epipolar line $\boldsymbol{l_R}$,

$$\boldsymbol{x_R}^\top \boldsymbol{l_R} = 0$$

or, $\boldsymbol{l_R}^\top \boldsymbol{x_R} = 0$

Since $\mathbf{E}\boldsymbol{x_L} = \boldsymbol{l_R}$,

$$\boldsymbol{x_R}^\top \mathbf{E} \boldsymbol{x_L} = 0$$

$\boldsymbol{l_L}^\top$

Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.

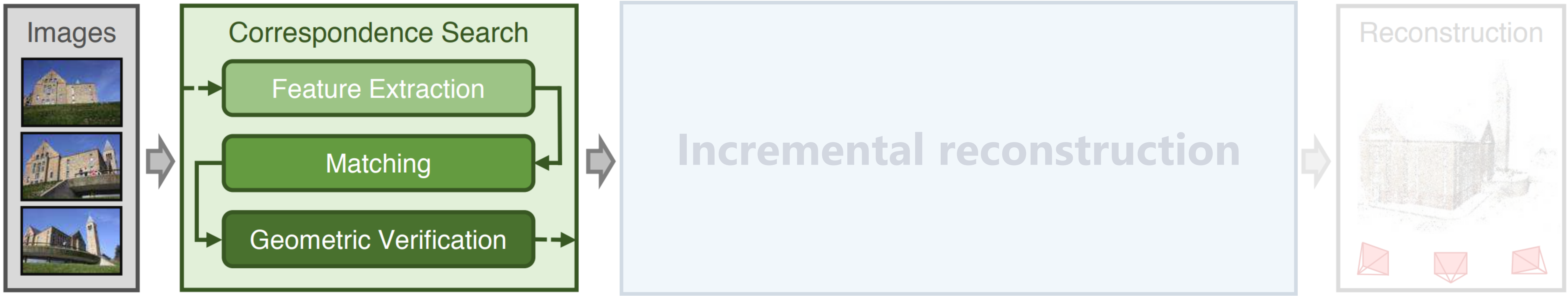
# Structure-from-Motion (1): Correpondence Search

## Correpondence search: Geometric verification

- **5-point** / **8-point** algorithm:

  Estimating the **essential matrix** from the feature point correspondences.

  → Direct linear transform: $\boldsymbol{x_R}^\top \mathbf{E} \boldsymbol{x_L} = 0 \rightarrow A\mathbf{p} = 0$ → Use SVD to solve!




Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.

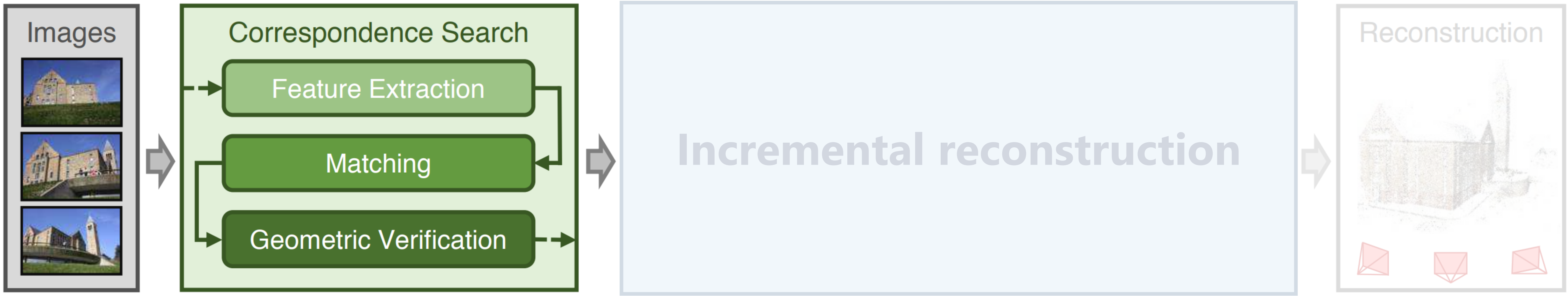
# Structure-from-Motion (1): Correpondence Search

## Correpondence search: Geometric verification

- **5-point** / **8-point** algorithm: $\boxed{\boldsymbol{x_R}^{\top}\mathbf{E}\boldsymbol{x_L} = 0 \rightarrow A\mathbf{p} = 0}$

  Sample at least **5** or **8** points and compute the **essential matrix**.

  → **RANSAC** to discriminate **inliers/outliers** and the **best** essential matrix!

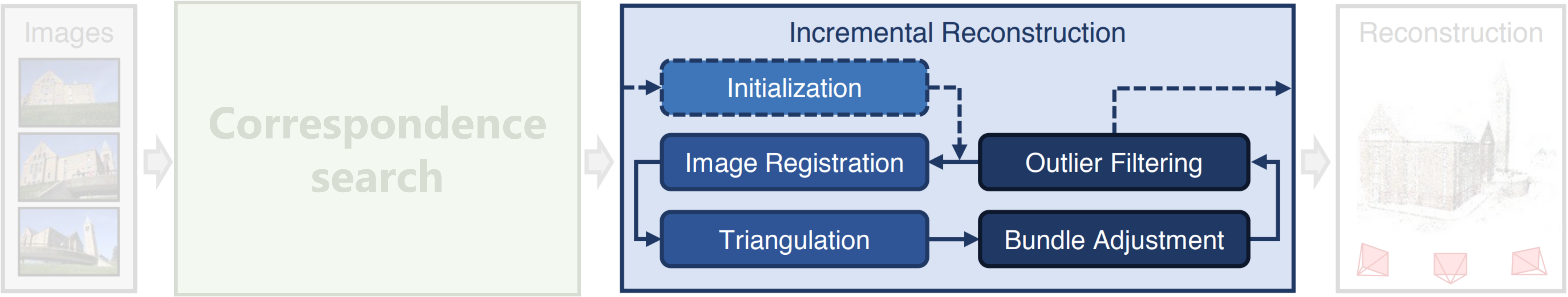



Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.

# Structure-from-Motion (2): Incremental Reconstruction

## Incremental reconstruction

- Camera initialization
- Triangulation
- Bundle adjustment (refinement)

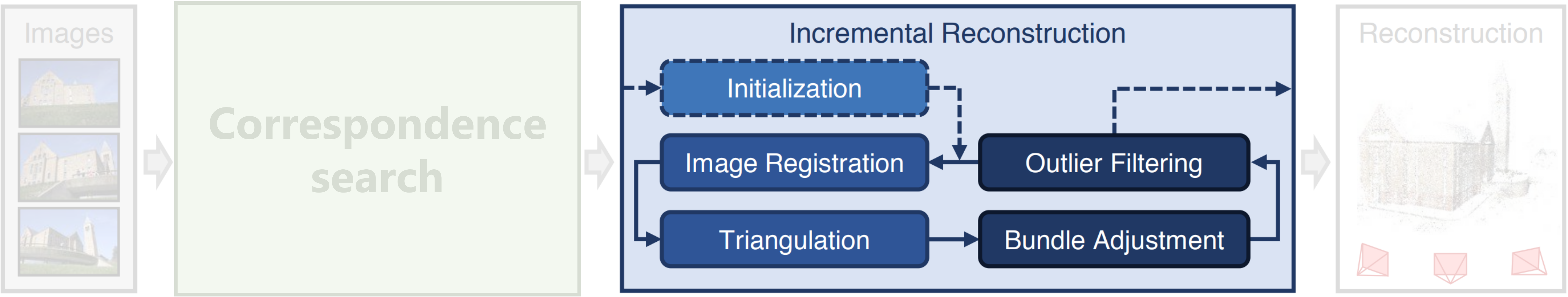



Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.

# Structure-from-Motion (2): Incremental Reconstruction

**Starting from two views, aggregate more views to refine the estimation**

- Camera initialization → Estimate camera **projection** matrices from $\mathbf{E} \in \mathbb{R}^{3\times3}$
- Triangulation → **Lift** 2D points to 3D spaces using camera projection matrices
- Bundle adjustment (refinement)



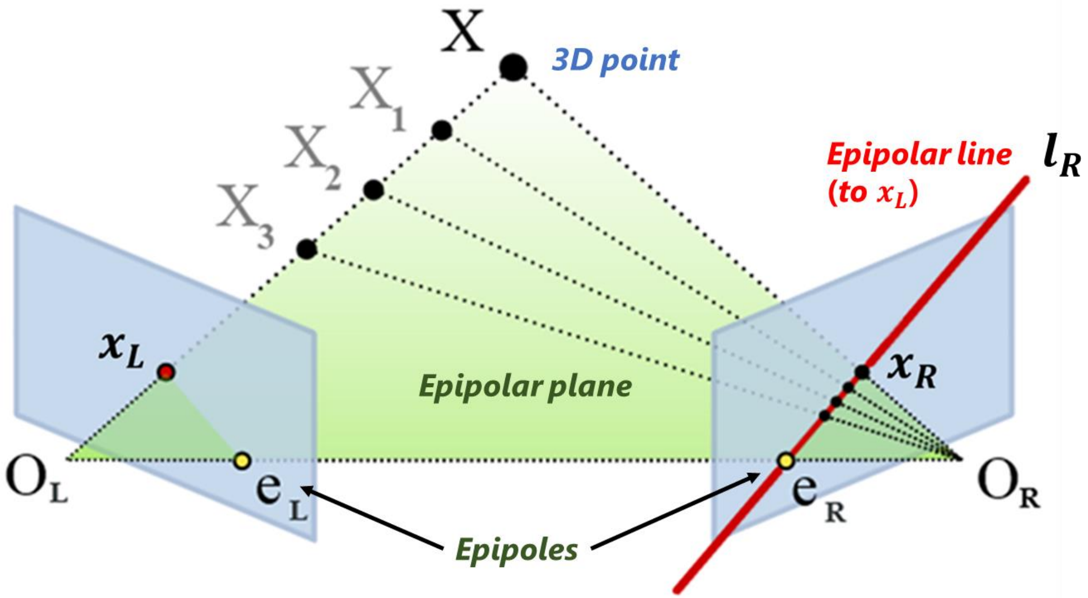

Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.

# Fundamental Matrix

(focal length = 1)

**Essential matrix** ($\mathbf{E} \in \mathbb{R}^{3\times3}$ ): Relation between normalized pixel correpondences

**Fundamental matrix** ($\mathbf{F} \in \mathbb{R}^{3\times3}$ ): Relation between camera pixel correpondences




real pixel    normalized

$$\hat{\boldsymbol{x}}_L = \boldsymbol{K}_L \boldsymbol{x}_L \qquad \hat{\boldsymbol{x}}_R = \boldsymbol{K}_R \boldsymbol{x}_R$$

$$\boldsymbol{x}_L = \boldsymbol{K}_L^{-1} \hat{\boldsymbol{x}}_L \qquad \boldsymbol{x}_R = \boldsymbol{K}_R^{-1} \hat{\boldsymbol{x}}_R$$

$$\boldsymbol{x}_R{}^\top \mathbf{E} \boldsymbol{x}_L = \left(\boldsymbol{K}_R^{-1} \hat{\boldsymbol{x}}_R\right)^\top \mathbf{E}\left(\boldsymbol{K}_L^{-1} \hat{\boldsymbol{x}}_L\right)$$

$$= \boldsymbol{x}_R{}^\top \underbrace{\boxed{\left(\boldsymbol{K}_R{}^\top\right)^{-1} \mathbf{E}\left(\boldsymbol{K}_L^{-1}\right)}}_{=\mathbf{F}} \hat{\boldsymbol{x}}_L = 0$$

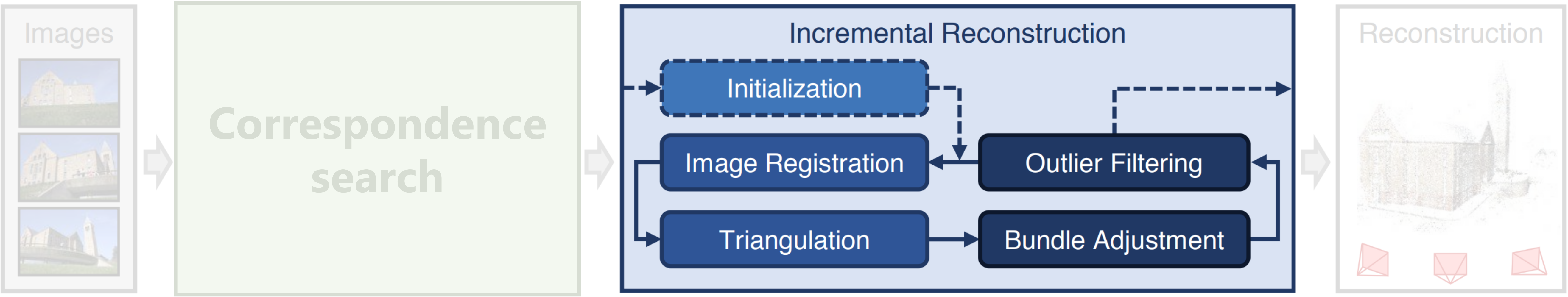
Schonberger, et al., “Structure-from-Motion Revisited”, CVPR 2016.

# Structure-from-Motion (2): Incremental Reconstruction

**Starting from two views, aggregate more views to refine the estimation**

- Camera initialization → Estimate camera **projection** matrices from $\mathbf{E} \in \mathbb{R}^{3\times3}$
- If we capture images with the **same** camera,

$$p_{img}'^{\top}\mathbf{F}p_{img} = \mathbf{0} \qquad \mathbf{F} = \left(K^{\top}\right)^{-1}\mathbf{E}K^{-1} \qquad \mathbf{E} = K^{\top}\mathbf{F}K$$

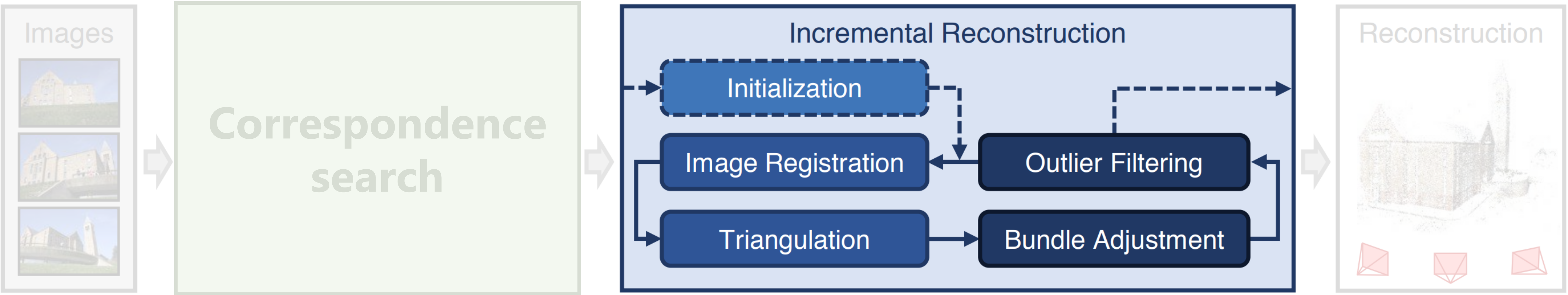



Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.

# Structure-from-Motion (2): Incremental Reconstruction

**Camera initialization: Estimate camera projection matrices $\mathrm{P}$, $\mathrm{P}'$ from $\mathrm{E} \in \mathbb{R}^{3\times 3}$**

- Suppose $\mathbf{P} = [\mathbf{I}|\mathbf{0}] \in \mathbb{R}^{3\times 4}$ (center of the world-coordinate)
- From derivation, there are 4 possible candidates for $\mathbf{P}'$
- Triangulate 2D point correspondences from two views, and physically verify $\mathbf{P}'$



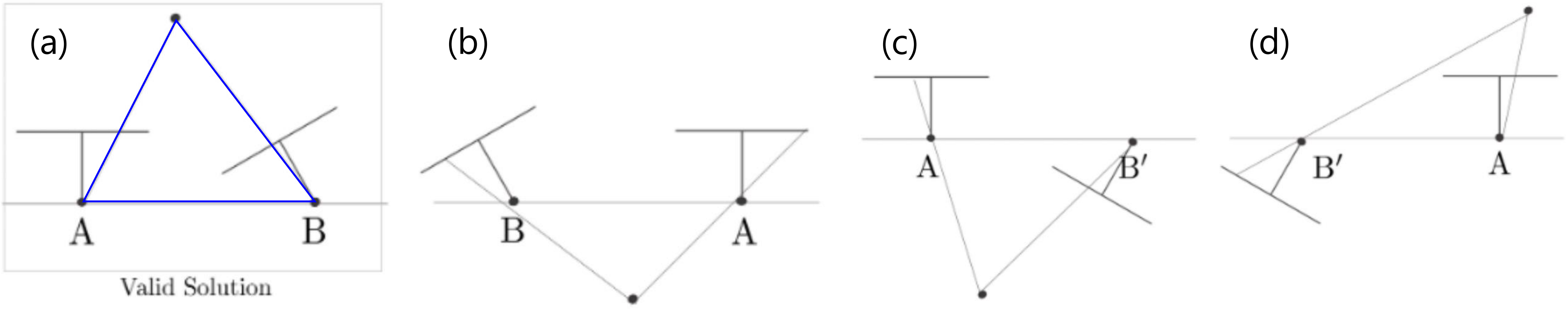

Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.

# Structure-from-Motion (2): Incremental Reconstruction

**Camera initialization: Estimate camera projection matrices $\mathrm{P}$, $\mathrm{P}'$ from $\mathbf{E} \in \mathbb{R}^{3\times3}$**

- Suppose $\mathbf{P} = [\mathbf{I}|\mathbf{0}] \in \mathbb{R}^{3\times4}$ (center of the world-coordinate)
- From derivation, there are 4 possible candidates for $\mathbf{P}'$
- Triangulate 2D point correspondences from two views, and physically verify $\mathbf{P}'$

Target 3D point is located in front of the two camera views



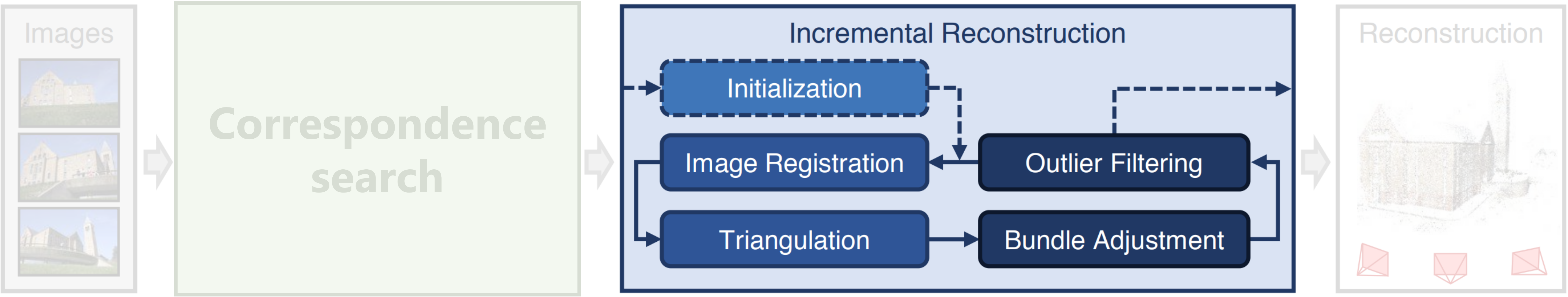

**Four** solutions of essential matrix

Schonberger, et al., “Structure-from-Motion Revisited”, CVPR 2016.

# Structure-from-Motion (2): Incremental Reconstruction

## Incremental reconstruction

- Camera initialization → Estimate camera **projection** matrices from $\mathbf{E} \in \mathbb{R}^{3\times3}$
- Triangulation → **Lift** 2D points to 3D spaces using camera projection matrices
- Bundle adjustment (refinement)

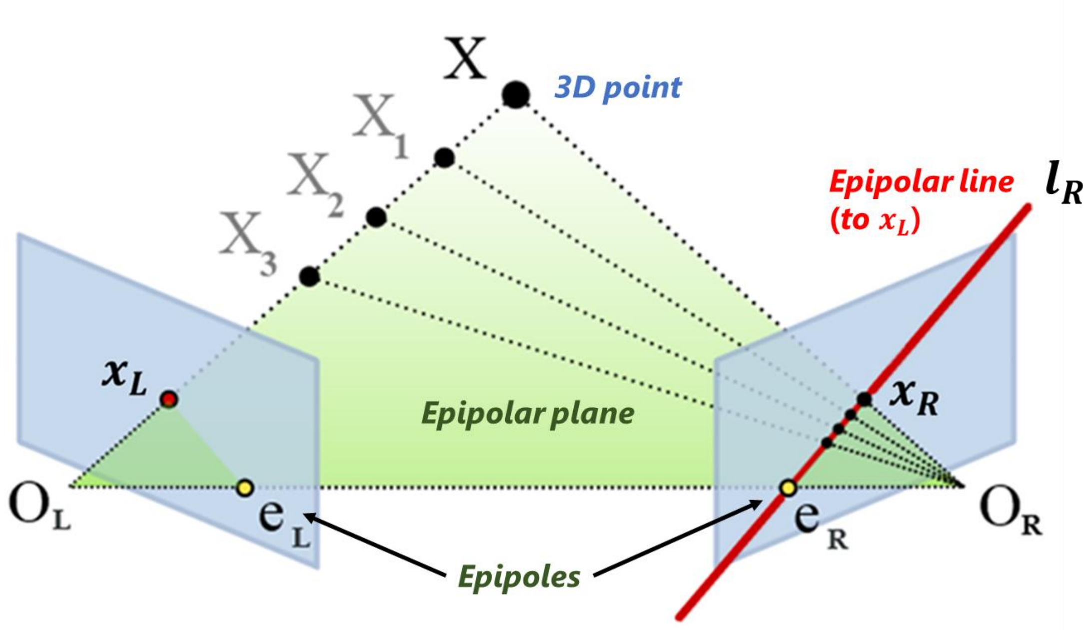



Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.

# Triangulation

Given a set of corresponding points $\overset{\text{(known)}}{\boldsymbol{x}_L, \boldsymbol{x}_R}$ and camera matrices $\overset{\text{(known)}}{\mathbf{P}, \mathbf{P}'}$, estimate $\overset{\text{(unknown)}}{\mathbf{X}}$

**Fundamental matrix** ($\mathbf{F} \in \mathbb{R}^{3\times3}$ ): Relation between camera pixel correpondences




$$\boldsymbol{x} = \mathbf{P}\mathbf{X} = \alpha\mathbf{P}\mathbf{X}$$

homogeneous coordinate

same direction but differs by a scale factor

$$\overbrace{\boldsymbol{x} \times \mathbf{P}}^{\text{(known)}}\overbrace{\mathbf{X}}^{\text{(unknown)}} = \mathbf{0}$$

cross product is zero (this equality removes scale factor)

$$\begin{aligned} x(\mathbf{p}_{3,row}^{\top}\mathbf{X}) - (\mathbf{p}_{1,row}^{\top}\mathbf{X}) &= 0 \\ y(\mathbf{p}_{3,row}^{\top}\mathbf{X}) - (\mathbf{p}_{2,row}^{\top}\mathbf{X}) &= 0 \\ x(\mathbf{p}_{2,row}^{\top}\mathbf{X}) - y(\mathbf{p}_{1,row}^{\top}\mathbf{X}) &= 0 \end{aligned}$$

One 2D to 3D point correspondence gives **two** equations

$$\underbrace{\begin{bmatrix} x\mathbf{p}_{3,row}^{\top} - \mathbf{p}_{1,row}^{\top} \\ y\mathbf{p}_{3,row}^{\top} - \mathbf{p}_{2,row}^{\top} \\ x'\mathbf{p}'^{3\top} - \mathbf{p}'^{1\top} \\ y'\mathbf{p}'^{3\top} - \mathbf{p}'^{2\top} \end{bmatrix}}_{\mathbf{A}} \mathbf{X} = 0$$

Now we can make system equations. Use SVD to solve!

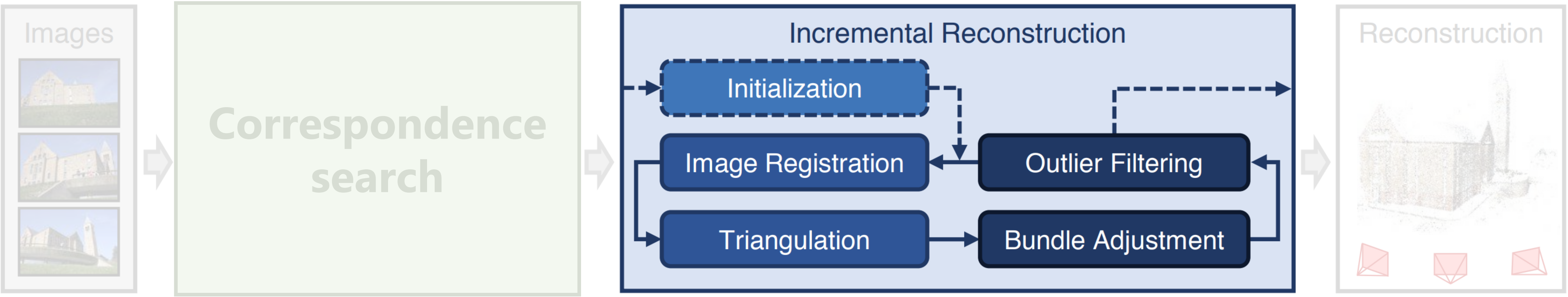
→ Two-view geometry

Schonberger, et al., “Structure-from-Motion Revisited”, CVPR 2016.

# Structure-from-Motion (2): Incremental Reconstruction

## Incremental reconstruction

- Camera initialization → Estimate camera **projection** matrices from $\mathbf{E} \in \mathbb{R}^{3\times3}$
- Triangulation → **Lift** 2D points to 3D spaces using camera projection matrices
- Bundle adjustment (refinement)

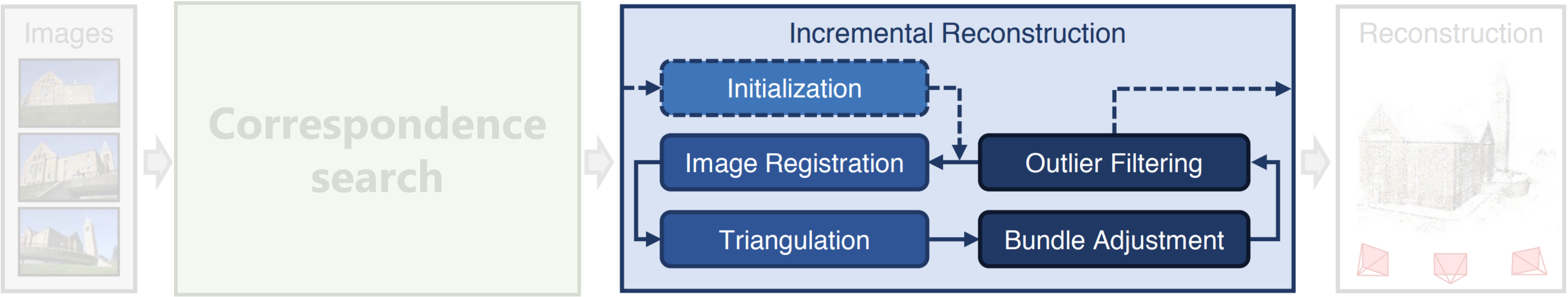



Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.

# Structure-from-Motion (2): Incremental Reconstruction

## Bundle adjustment: Non-linear method for jointly refining SfM

- **Bundle** of light rays leaving each 3D points and **converging** on each camera center
- Refining **3D points** and the camera **poses** by minimizing **reprojection error**
- Usually optimized with Levenberg-Marquardt optimization



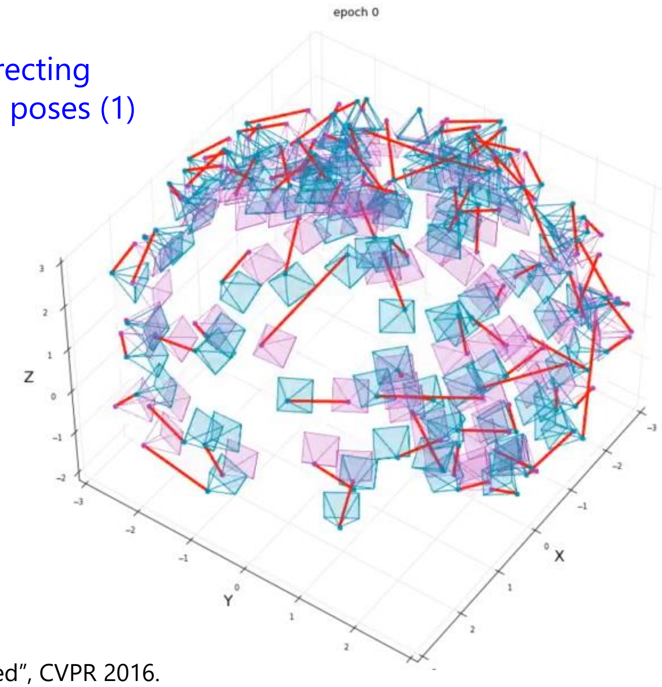

Schonberger, et al., “Structure-from-Motion Revisited”, CVPR 2016.

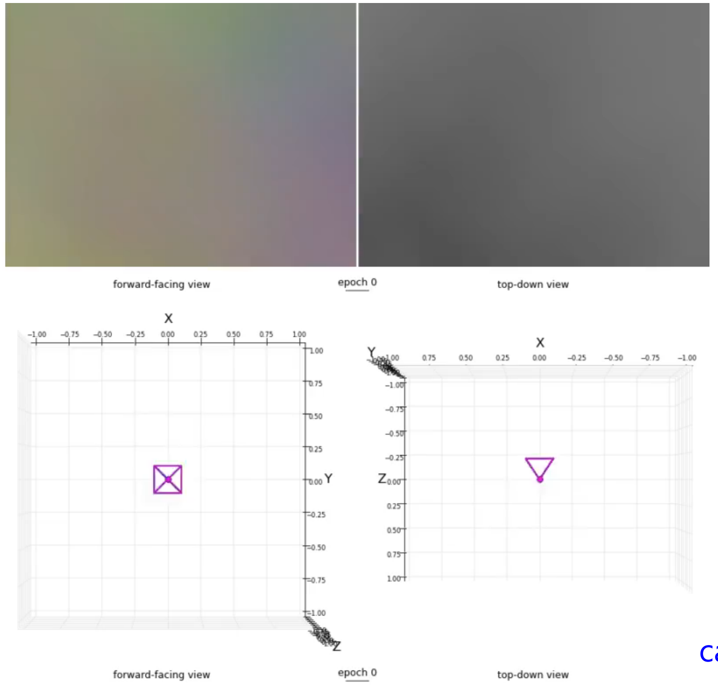
# Structure-from-Motion (2): Incremental Reconstruction

## Bundle adjustment: Non-linear method for jointly refining SfM

- **Bundle** of light rays leaving each 3D points and **converging** on each camera center
- Refining **3D points** and the camera **poses** by minimizing **reprojection error**
- Usually optimized with Levenberg-Marquardt optimization

Correcting
camera poses (1)




Correcting
camera poses (2)

[1] Schonberger, et al., “Structure-from-Motion Revisited”, CVPR 2016.
[2] Lin, et al., “BARF : Bundle-Adjusting Neural Radiance Fields”, ICCV 2021.

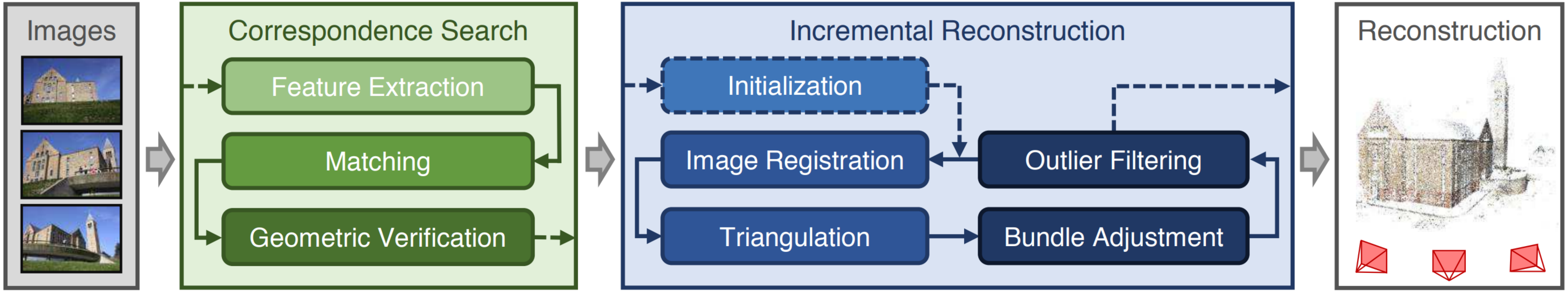
# Structure-from-Motion Pipeline

## Correpondence search

- Feature extraction
- Matching
- Geometric verification

## Incremental reconstruction

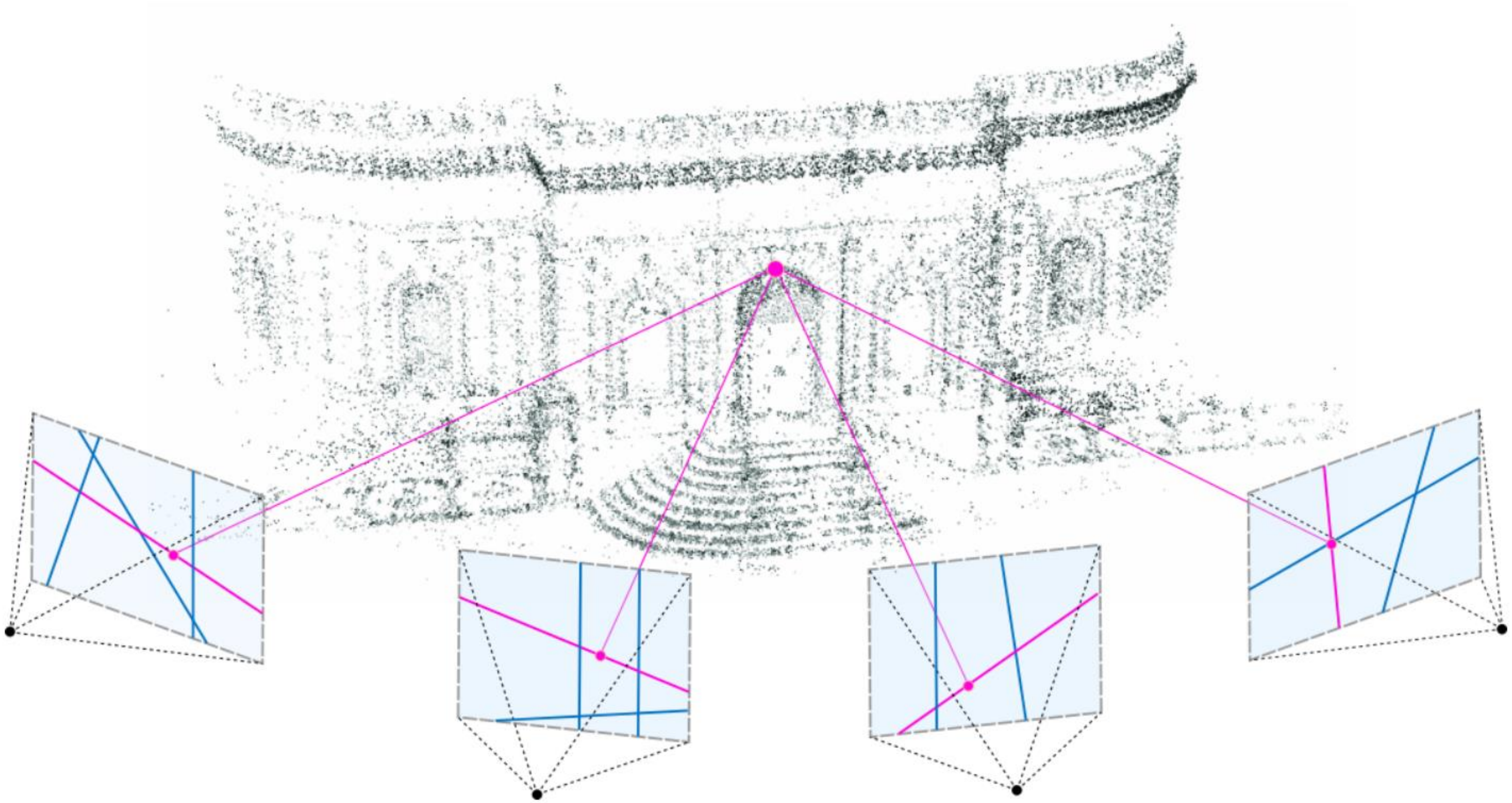
- Camera initialization
- Triangulation
- Bundle adjustment (refinement)




Schonberger, et al., “Structure-from-Motion Revisited”, CVPR 2016.

# Experiment: Build Your Own 3D Structure

https://view.kentech.ac.kr/f088fa7f-874e-44bc-bd6d-6084b42dfdf7



Schonberger, et al., "Structure-from-Motion Revisited", CVPR 2016.