



NYU

TANDON SCHOOL  
OF ENGINEERING



# Robot Perception

## Multi-View Geometry

Dr. Chen Feng

[cfeng@nyu.edu](mailto:cfeng@nyu.edu)

ROB-GY 6203, Fall 2024



# Overview

- \* Hands-on: AprilTag & camera calibration

- + Epipolar geometry

- ++ Fundamental matrix

- + Essential matrix

- + Planar Homography

- + PnP problem

- ++ Hand-eye calibration

\*: know how to code

++: know how to derive

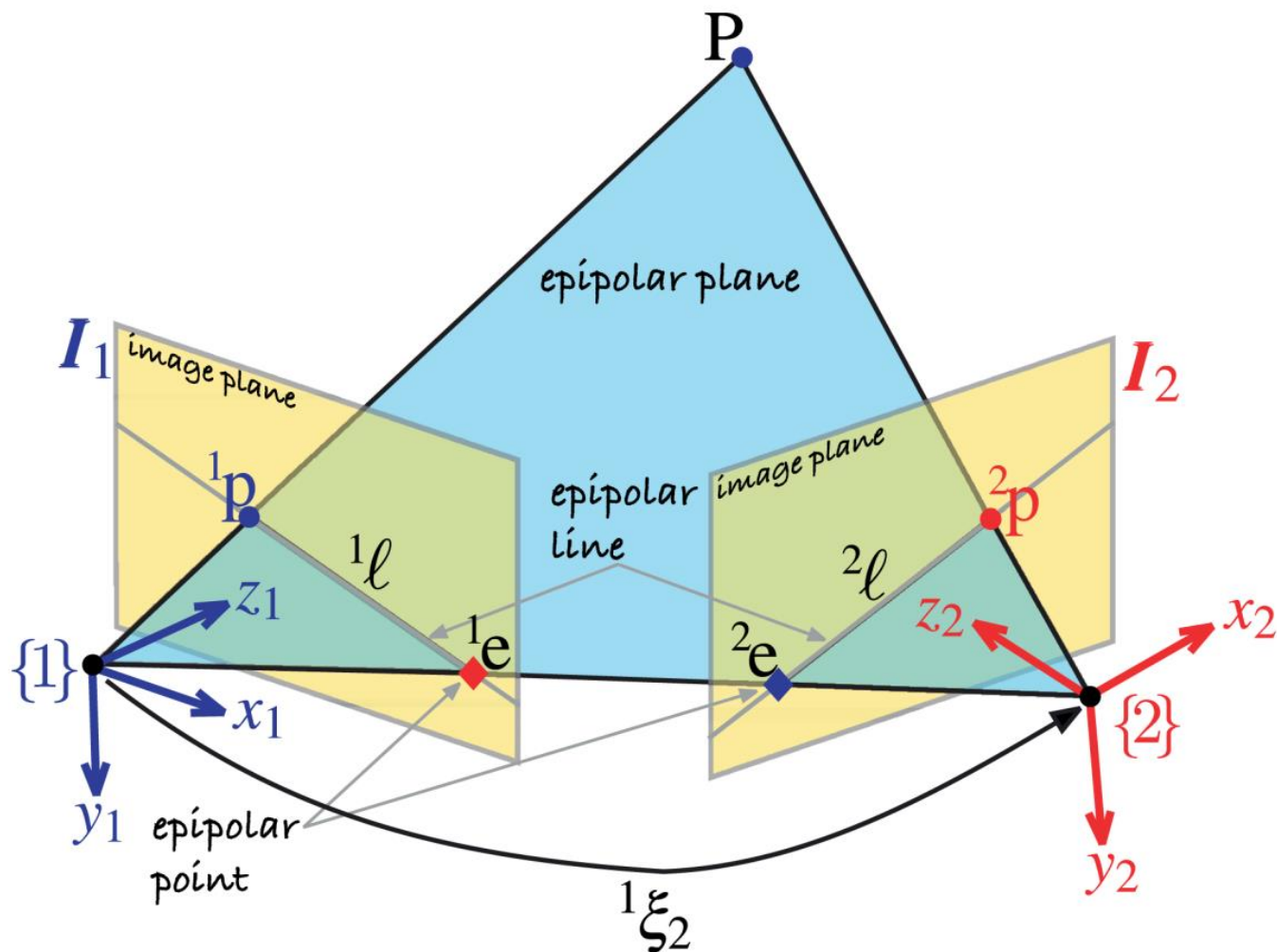
+: know the concept



# References

- HZ2003:
  - Section 9.1, 9.2, 9.3, 9.5, 9.6, 11.1, 11.2, 11.7
- Co2017:
  - Section 14.2, 11.2.3
- Sz2022:
  - Section 11.3, 11.2, 12.1
- FP2011:
  - Section 7.1, 8.1.2
- Radu Horaud, Fadi Dornaika. Hand-eye Calibration. International Journal of Robotics Research, SAGE Publications, 1995, 14 (3), pp.195–210.

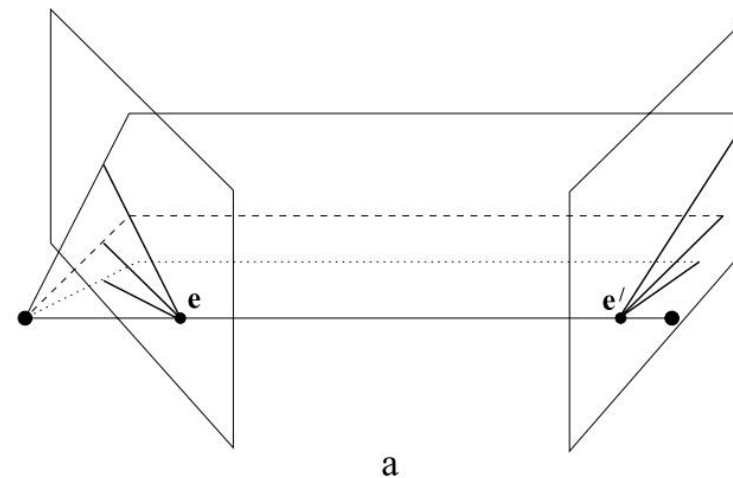
# Two-view Geometry





# Epipolar Geometry

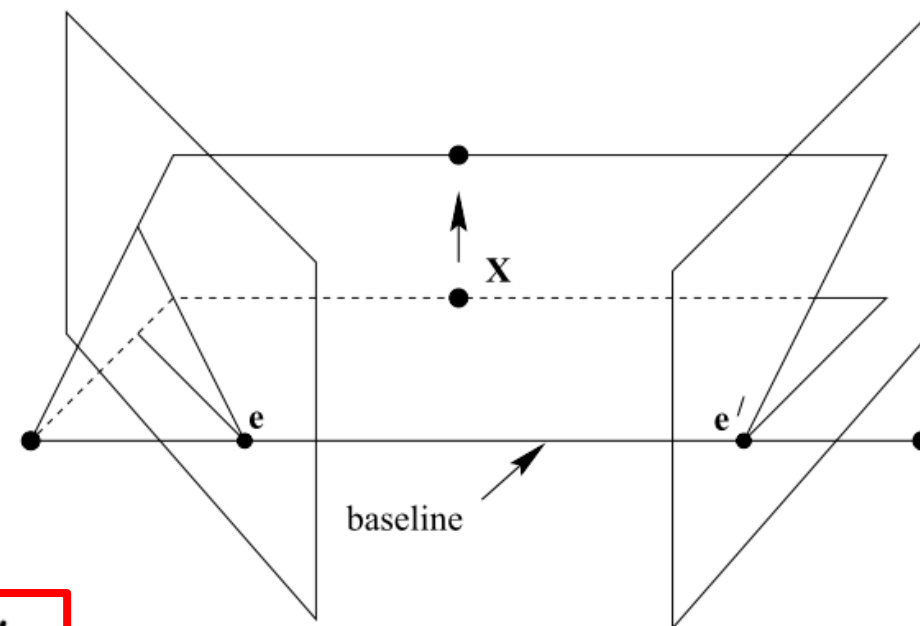
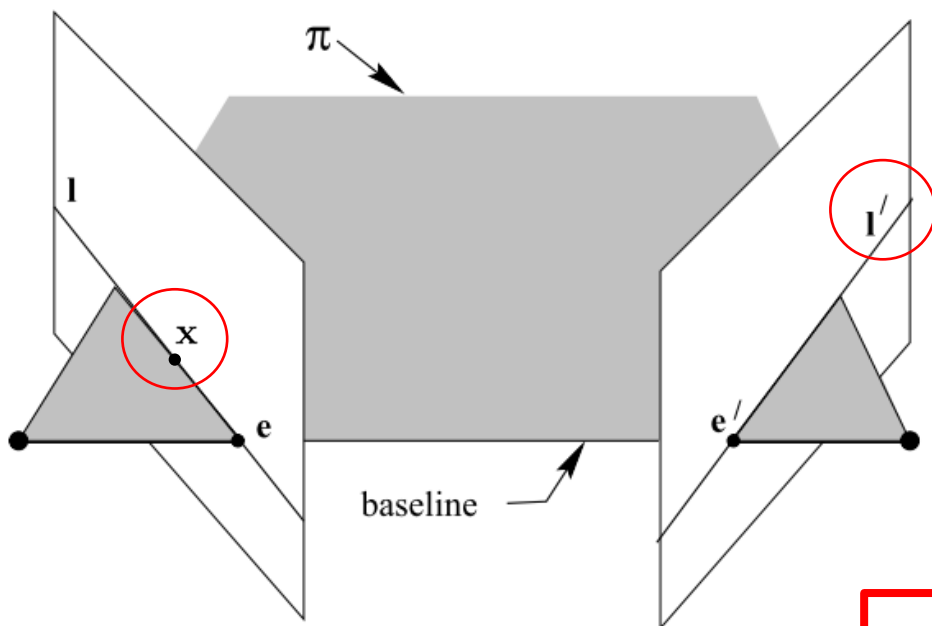
- Independent of scene structure
- Only depends on the **cameras' internal parameters and relative pose**





# Epipolar Geometry

- From a **point  $x$**  on one image, epipolar geometry allows us to find on the other image a **corresponding line  $l'$**  that **MUST** contain the **corresponding point  $x'$** , without knowing where  $x'$  is



$$x \mapsto l'$$



# Fundamental Matrix: the Algebra of Epipolar Geometry

**Result 9.3.** *The fundamental matrix satisfies the condition that for any pair of corresponding points  $\mathbf{x} \leftrightarrow \mathbf{x}'$  in the two images*

$$\mathbf{x}'^T \mathbf{F} \mathbf{x} = 0.$$

[LonguetHiggins-81] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, September 1981.

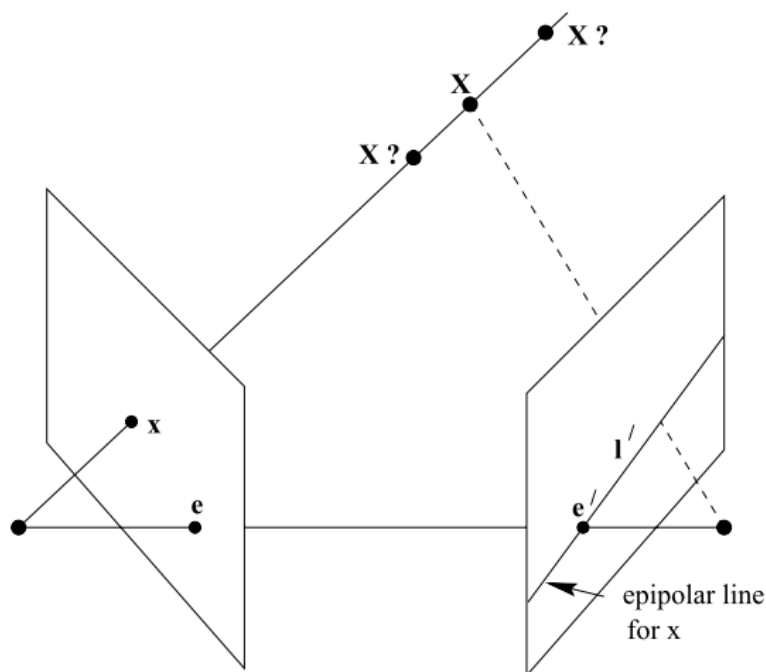


# Fundamental Matrix: the Algebra of Epipolar Geometry



**Result 9.3.** *The fundamental matrix satisfies the condition that for any pair of corresponding points  $\mathbf{x} \leftrightarrow \mathbf{x}'$  in the two images*

$$\mathbf{x}'^T \mathbf{F} \mathbf{x} = 0.$$



$$\mathbf{x} \mapsto \mathbf{l}'$$






# Important Properties of Fundamental Matrix

- $F$  matrix is homogeneous


- $rank(F) == 2$  

Geometrically,  $F$  represents a mapping from the 2-dimensional projective plane  $\mathbb{P}^2$  of the first image to the pencil of epipolar lines through the epipole  $e'$ . Thus, it represents a mapping from a 2-dimensional onto a 1-dimensional projective space, and hence must have rank 2.

- $F$  matrix has **only 7** degrees-of-freedom (DOF)
  - 8 ratios – 1 rank deficiency

- $x \Rightarrow l' = Fx$
- $x' \Rightarrow l = F^T x'$  

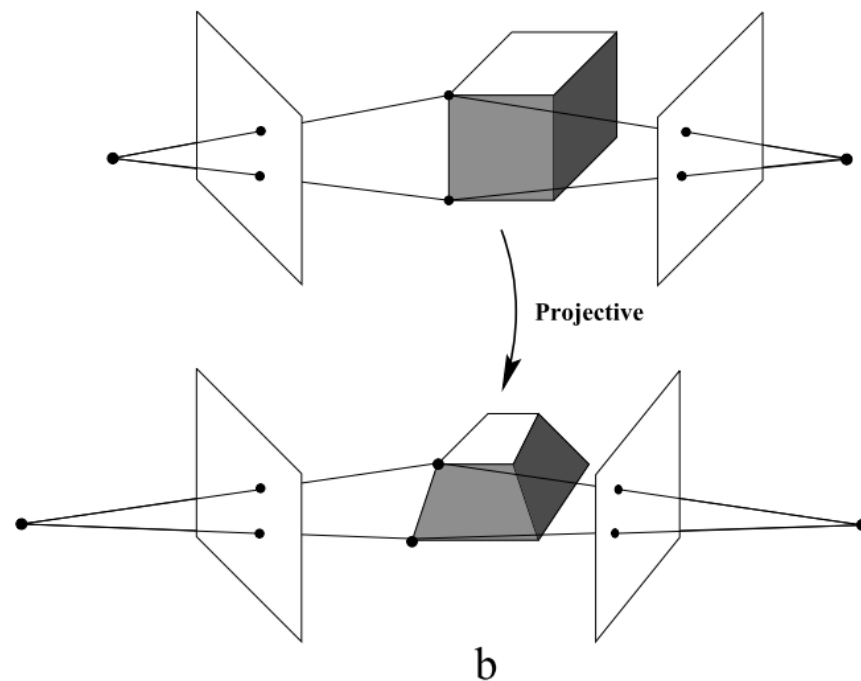
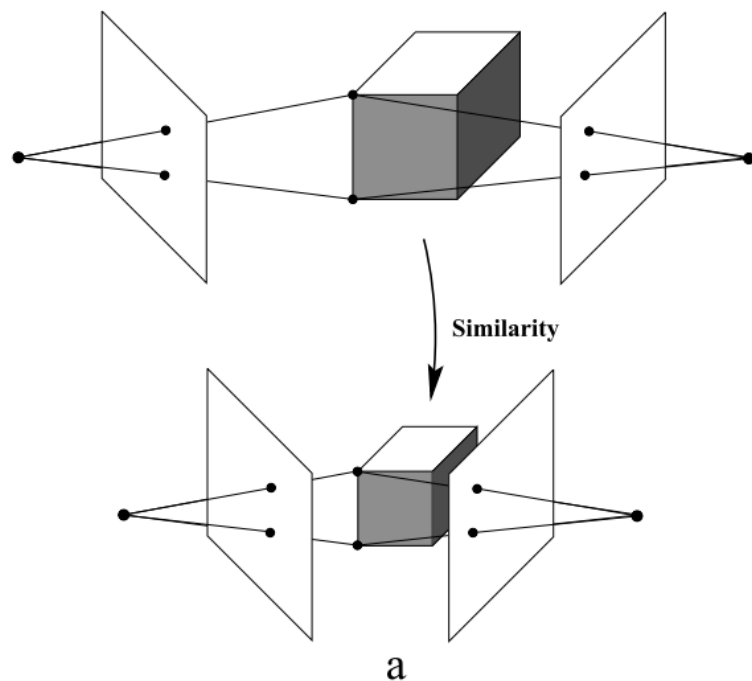
- From  $x$  in image 1, we can use  $F$  to compute the corresponding epipolar line  $l'$  in image 2.
- And vice versa.

- $Fe = 0$
- $F^T e' = 0$  

- Given  $F$ , we can compute the epipole  $e$  in image 1 as the null space of  $F$ .
- And the epipole  $e'$  in image 2 is the left null space of  $F$ .



# Projective Ambiguity of Fundamental Matrix



**Result 9.8.** *If  $H$  is a  $4 \times 4$  matrix representing a projective transformation of 3-space, then the fundamental matrices corresponding to the pairs of camera matrices  $(P, P')$  and  $(PH, P'H)$  are the same.*

**How can we reduce such ambiguities?**



# Normalized Coordinate and Essential Matrix

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

$$\mathbf{P} = \mathbf{K}[\mathbf{R} \mid \mathbf{t}]$$

Normalized coordinate  $\hat{\mathbf{x}} = \mathbf{K}^{-1}\mathbf{x} = [\mathbf{R} \mid \mathbf{t}]\mathbf{X}$

$$\hat{\mathbf{x}}'^T \mathbf{E} \hat{\mathbf{x}} = 0$$

**Result 9.17.** *A  $3 \times 3$  matrix is an essential matrix if and only if two of its singular values are equal, and the third is zero.*

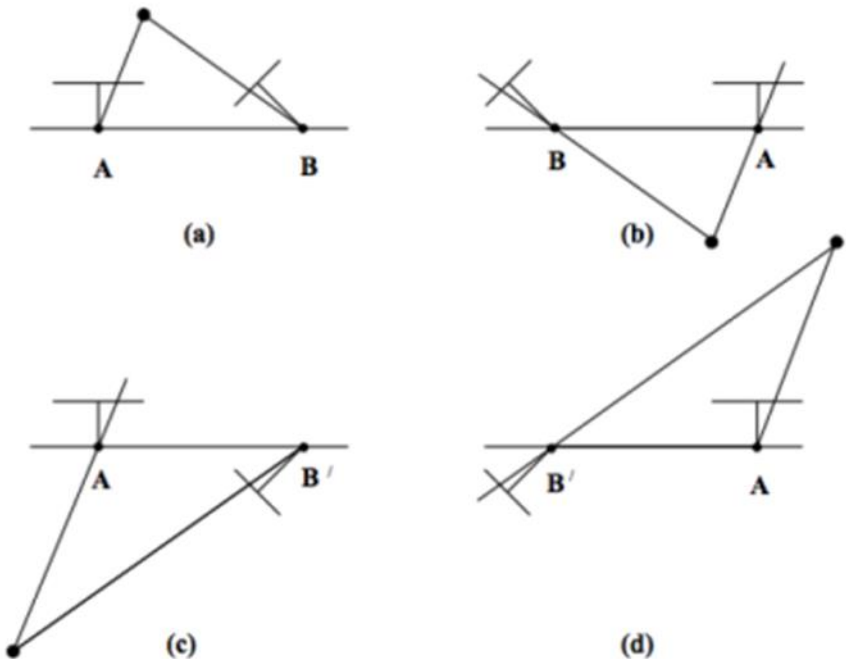
$$\mathbf{E} = \mathbf{K}'^T \mathbf{F} \mathbf{K}$$

$$\mathbf{E} = [\mathbf{t}]_{\times} \mathbf{R}$$



# Extracting Relative Camera Pose from E-matrix

- There are 4 potential  $(R, t)$  solutions that would give the same E-matrix
  - Two potential solutions for  $R$
  - Two potential solutions for  $t$
- But only one solution is practical!
- How do we eliminate the impossible?
  - Using our prior knowledge
- All 3D points should be in front of the two cameras
  - Reconstruction 3D points for each potential solution
  - Count #points ( $=N$ ) in front of the camera
  - Choose the potential solution with the max  $N$



2D illustration of the 4 potential solutions



## Estimating F-matrix (the 8-point algorithm)



- Find multiple  $\mathbf{X} \leftrightarrow \mathbf{X}'$  correspondences ( $\geq 8$ ) between two images

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

$$x'x f_{11} + x'y f_{12} + x' f_{13} + y'x f_{21} + y'y f_{22} + y' f_{23} + x f_{31} + y f_{32} + f_{33} = 0$$

$$\mathbf{A} \mathbf{f} = \begin{bmatrix} x'_1 x_1 & x'_1 y_1 & x'_1 & y'_1 x_1 & y'_1 y_1 & y'_1 & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x'_n x_n & x'_n y_n & x'_n & y'_n x_n & y'_n y_n & y'_n & x_n & y_n & 1 \end{bmatrix} \mathbf{f} = \mathbf{0}$$

- Enforce rank-2 constraint by SVD



# Normalized 8-point Algorithm

## Objective

Given  $n \geq 8$  image point correspondences  $\{\mathbf{x}_i \leftrightarrow \mathbf{x}'_i\}$ , determine the fundamental matrix  $F$  such that  $\mathbf{x}'_i{}^T F \mathbf{x}_i = 0$ .

## Algorithm

- (i) **Normalization:** Transform the image coordinates according to  $\hat{\mathbf{x}}_i = T\mathbf{x}_i$  and  $\hat{\mathbf{x}}'_i = T'\mathbf{x}'_i$ , where  $T$  and  $T'$  are normalizing transformations consisting of a translation and scaling.
- (ii) Find the fundamental matrix  $\hat{F}'$  corresponding to the matches  $\hat{\mathbf{x}}_i \leftrightarrow \hat{\mathbf{x}}'_i$  by
  - (a) **Linear solution:** Determine  $\hat{F}$  from the singular vector corresponding to the smallest singular value of  $\hat{A}$ , where  $\hat{A}$  is composed from the matches  $\hat{\mathbf{x}}_i \leftrightarrow \hat{\mathbf{x}}'_i$  as defined in (11.3).
  - (b) **Constraint enforcement:** Replace  $\hat{F}$  by  $\hat{F}'$  such that  $\det \hat{F}' = 0$  using the SVD (see section 11.1.1).
- (iii) **Denormalization:** Set  $F = T'^T \hat{F}' T$ . Matrix  $F$  is the fundamental matrix corresponding to the original data  $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$ .



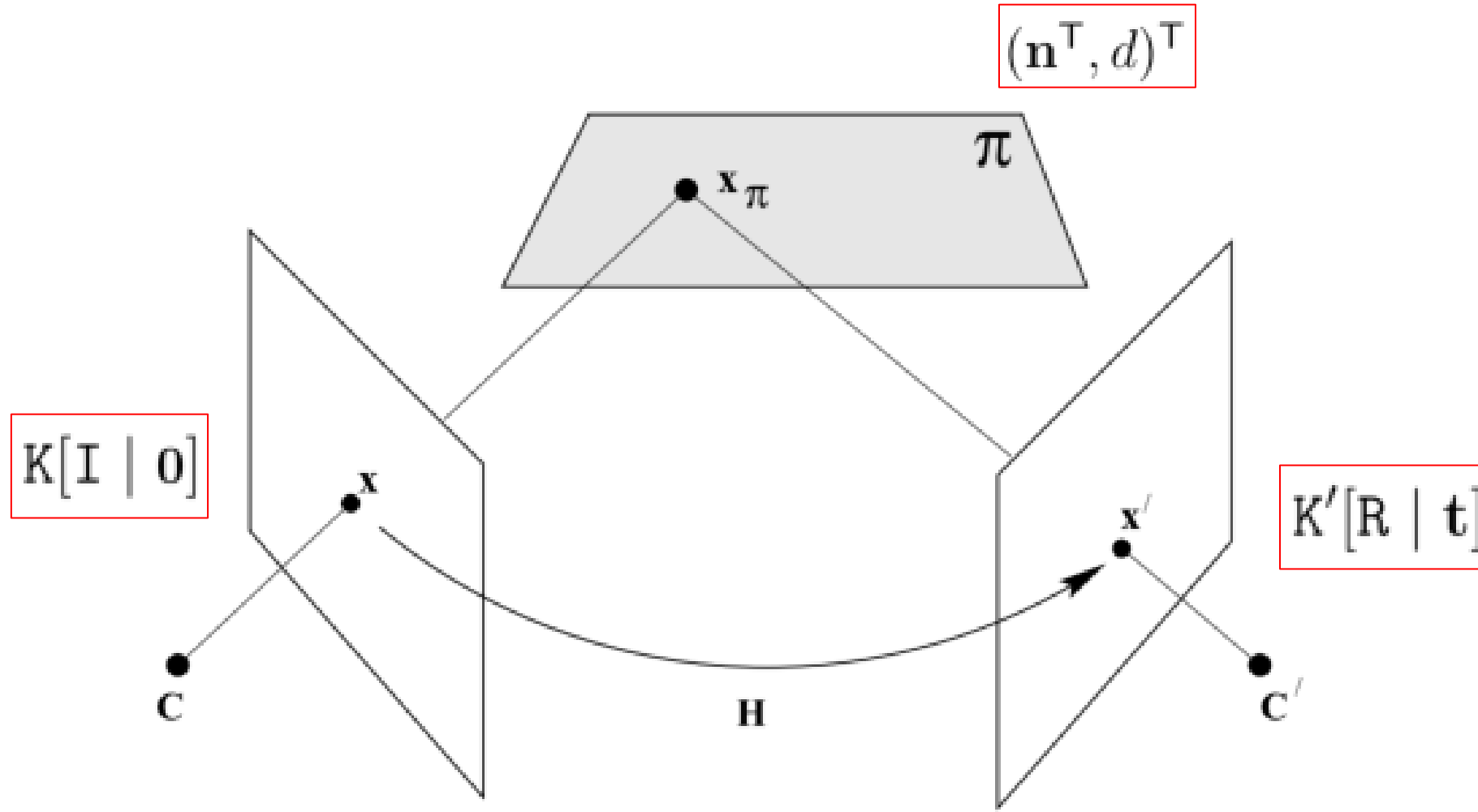
# The Fundamental Matrix Song – Daniel Wedge



Video from: <https://youtu.be/DgGV3l82NTk>



# Planar Homography





# Planar Homography

- Any point  $X$  on the plane

- Plane equation:

$$n^T X + d = 0$$

- Homogeneous coordinate:

$$(X, \frac{-n^T X}{d})$$

- Image on camera 1:

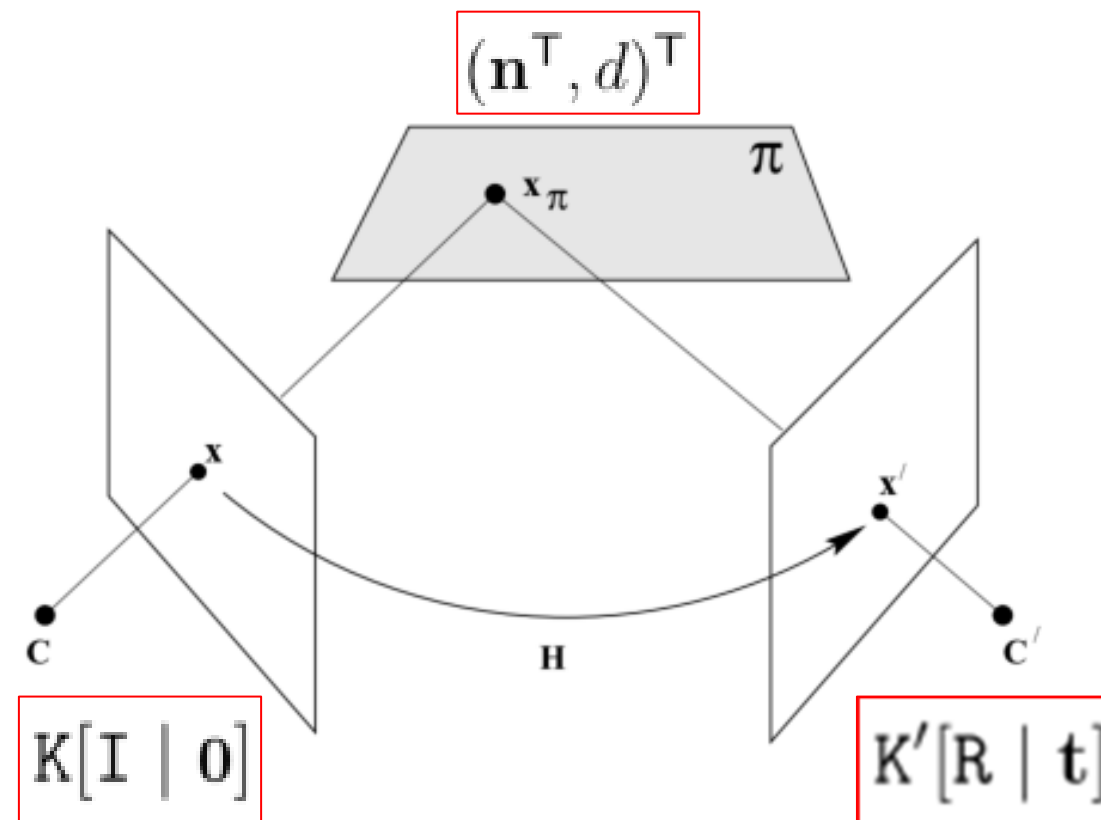
$$x_1 = KX$$

- Image on camera 2:

$$x_2 = K'[RX - \frac{tn^T X}{d}]$$

- Relate two images by  $H$

$$x_2 = Hx_1 = HKX = K'[RX - \frac{tn^T X}{d}]$$



Malis, Ezio, and Manuel Vargas. "Deeper understanding of the homography decomposition for vision-based control." (2007).

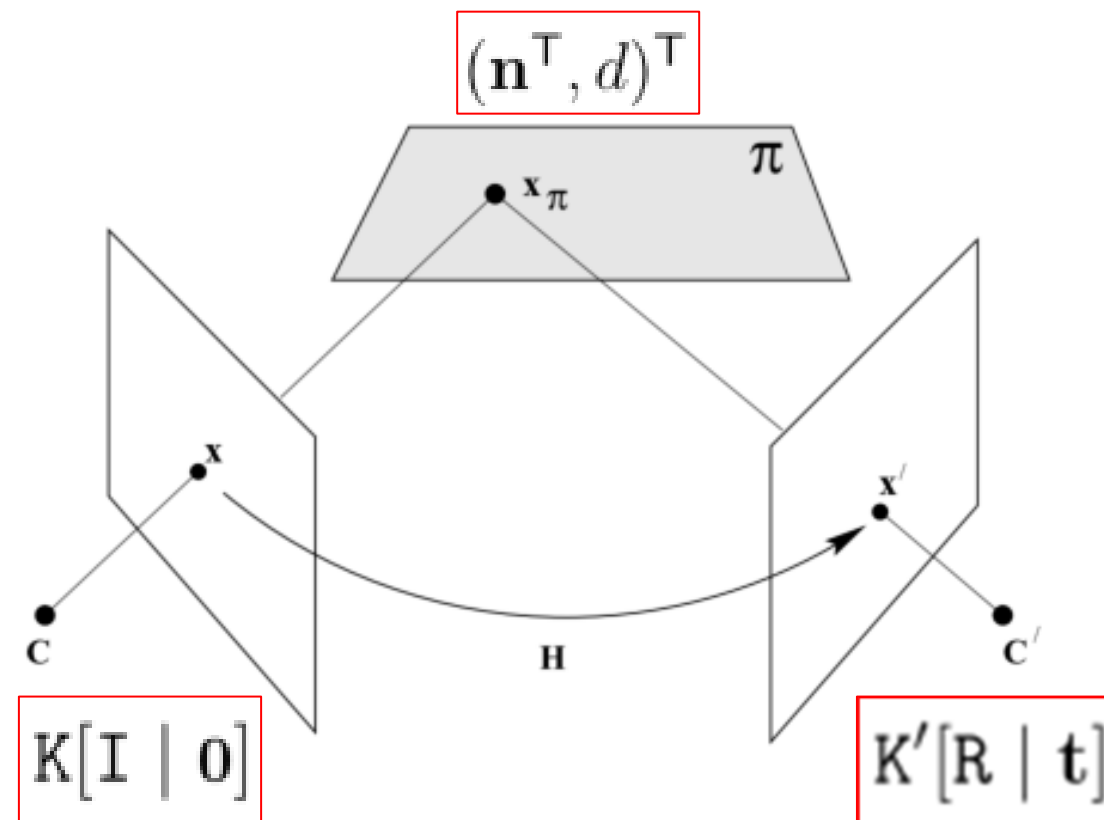
# Planar Homography

- Above equation holds for any  $X$

$$H = K' \left[ RX - \frac{tn^T X}{d} \right] K^{-1}$$

$$K'^{-1} H K = R - \frac{tn^T}{d} = R - t'n^T$$

Decompose  $K'^{-1} H K$  to retrieve  $R$ ,  $t'$ , and  $n$



Malis, Ezio, and Manuel Vargas. "Deeper understanding of the homography decomposition for vision-based control." (2007).

# Perspective-n-point (PnP) Problem

## • Given/known variables

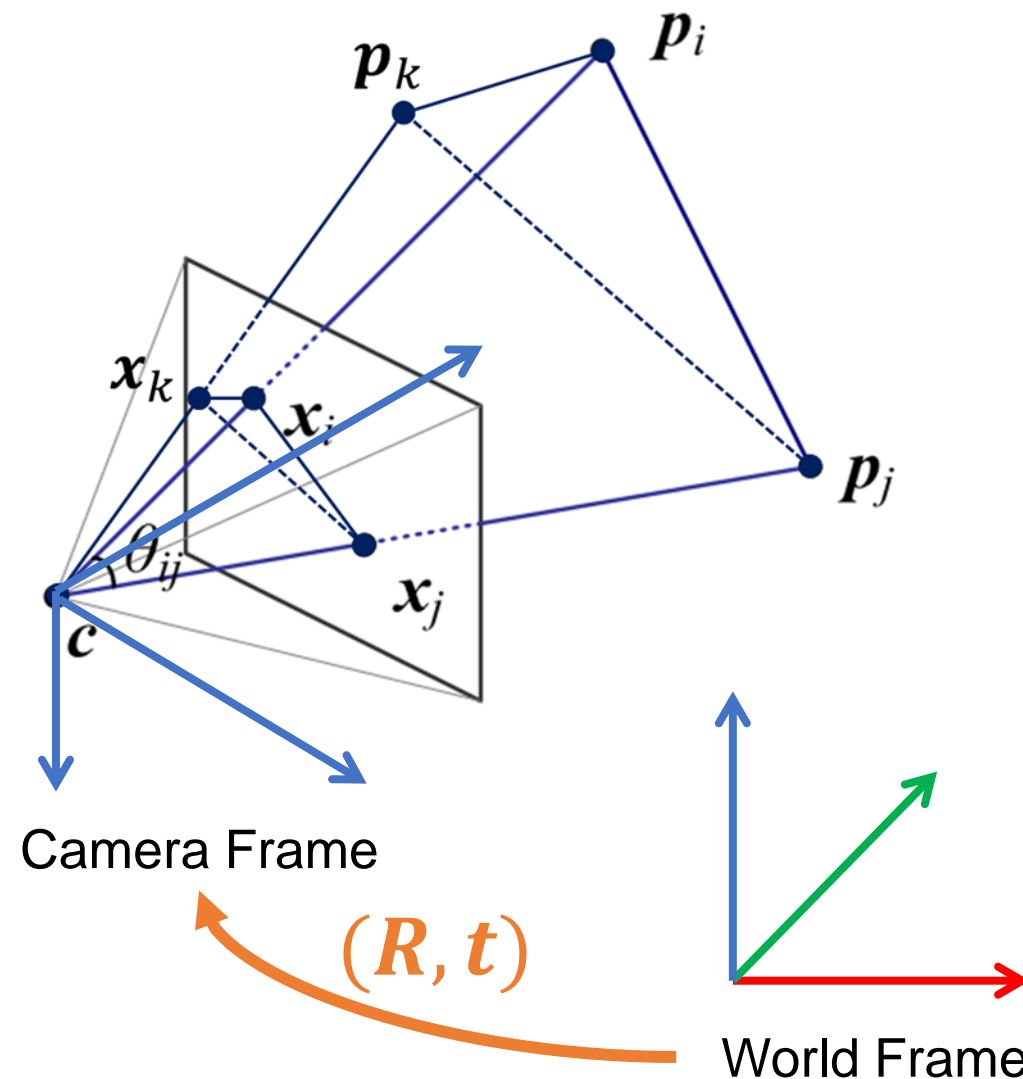
- $p_i \in \mathbb{R}^3$ , a set of  $n$  3D points in the world frame
- $x_i \in \mathbb{P}^2$ , a set of homogeneous image points
  - **corresponding to**  $p_i$  (with the same index  $i$ ) through camera projection
- $K$ , camera intrinsics

## • Unknown Variables

- $R$ , camera rotation
- $t$ , camera translation
  - pose of the world frame w.r.t. the camera frame

- $s_i x_i = K(Rp_i + t)$

- Note: we can write “=” sign here instead of “ $\propto$ ”
- Because we write the unknown scale  $s_i$  explicitly



# Perspective-n-point (PnP) Problem

- A calibrated camera is an angular sensor

- $\hat{x}_i = K^{-1}x_i / \|K^{-1}x_i\|$

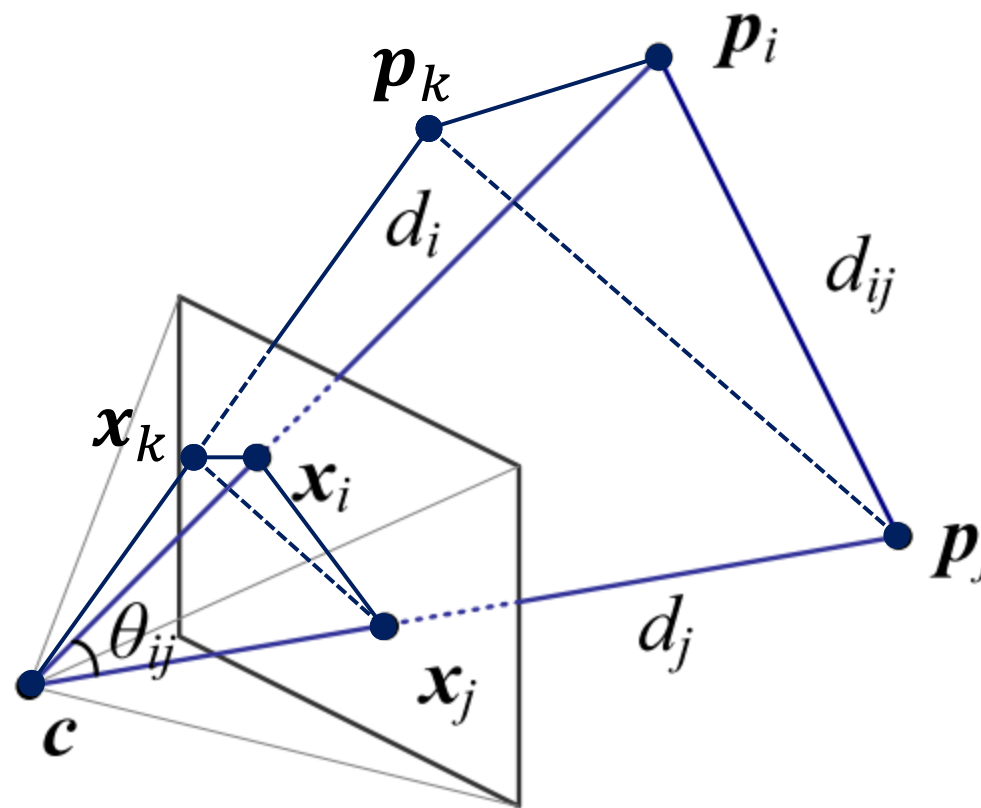
- Visual angles between any pair of image points must be the same as the angle between corresponding 3D points

- $\angle x_i c x_j \equiv \angle p_i c p_j$

- 3 pair of 2D-3D correspondences leads to 4 possible solutions

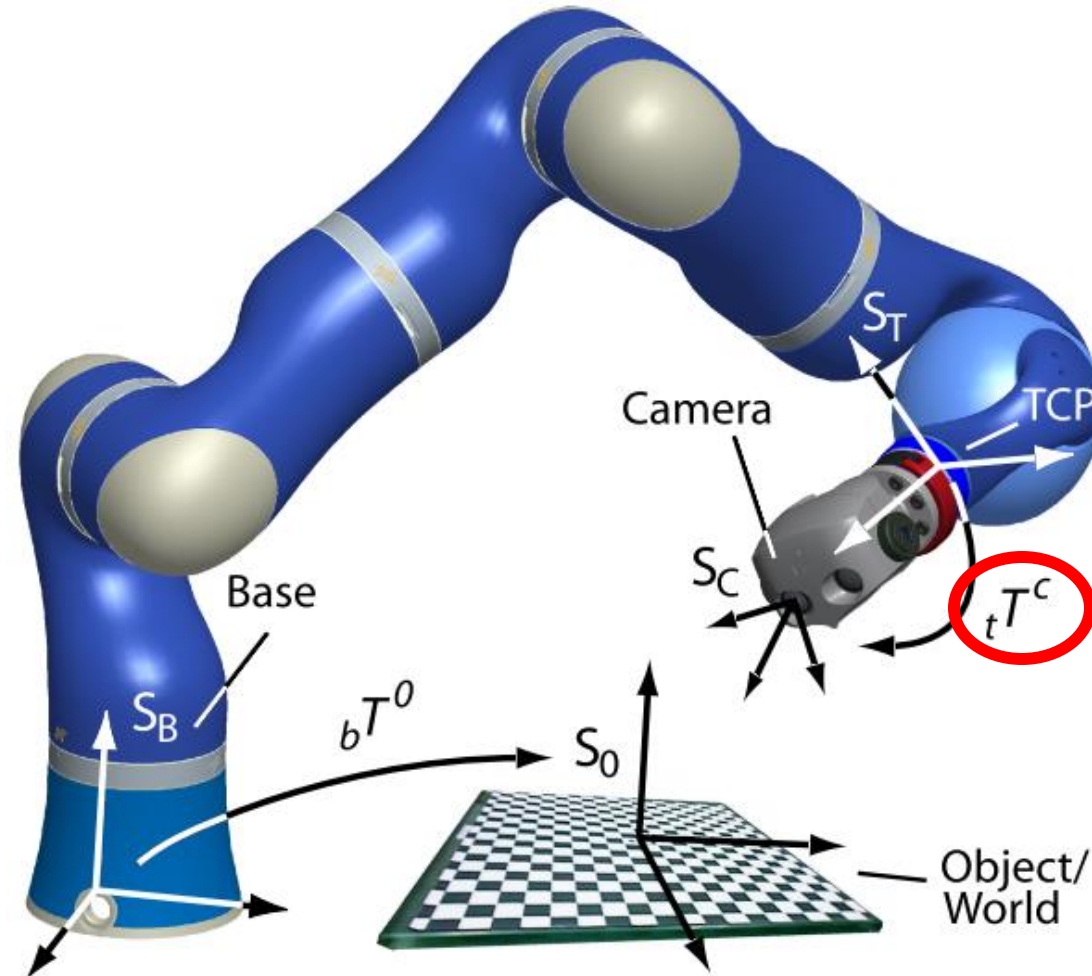
- P3P problem

- Useful for estimating object pose





# Hand-eye Calibration





# Hand-eye Calibration

- For any static point  $P$  in the calibration frame

$$B_1 X^{-1} A_1 P = B_2 X^{-1} A_2 P, \forall P$$

$$B_2^{-1} B_1 X^{-1} = X^{-1} A_2 A_1^{-1}$$

$$X B_1^{-1} B_2 = A_1 A_2^{-1} X$$

- $AX = XB$**

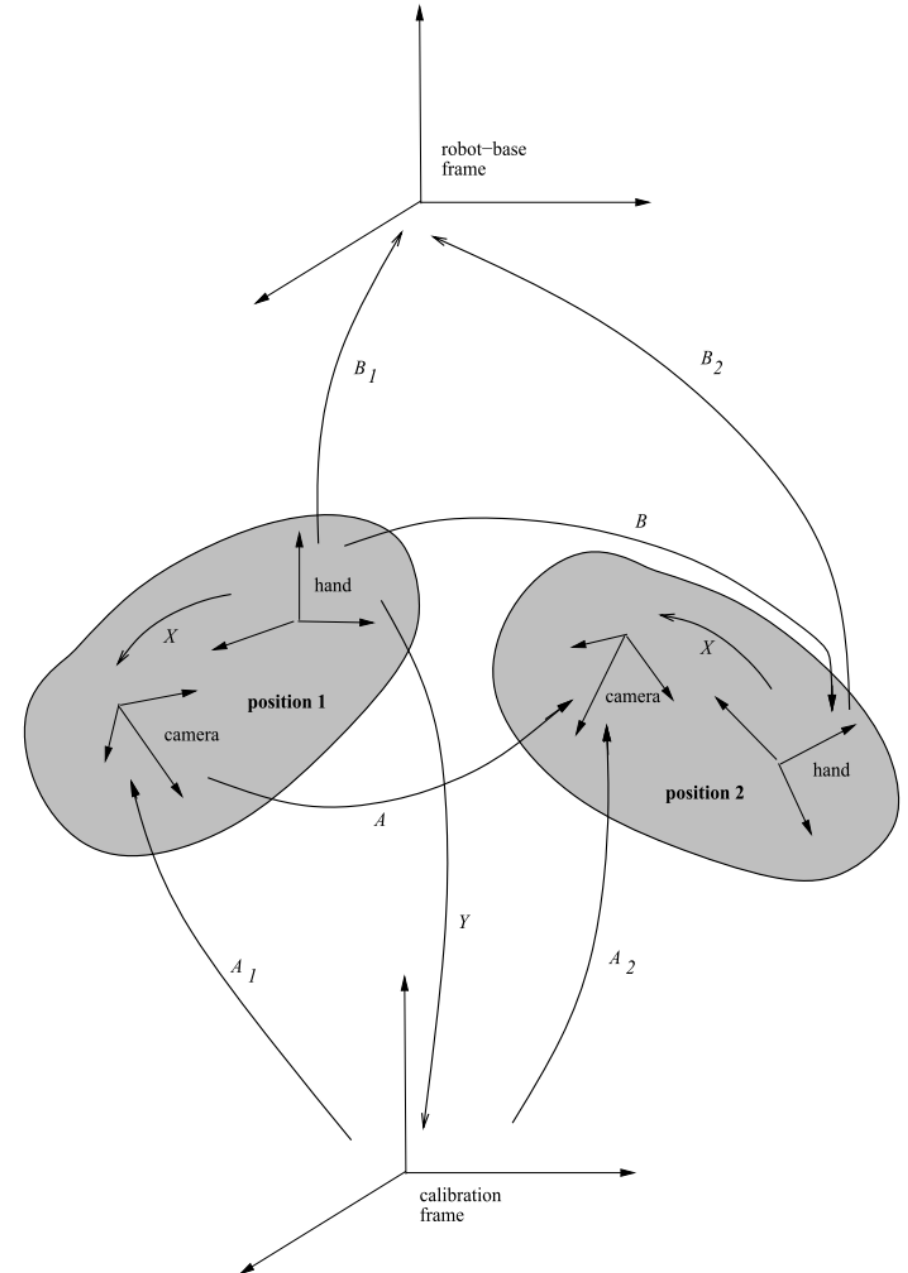
$$A = A_1 A_2^{-1}$$

$$B = B_1^{-1} B_2$$

- Solve  $X$  by observing multiple  $(A, B)$

$$\begin{cases} A_{12}X = XB_{12} \\ \vdots \\ A_{i-1\ i}X = XB_{i-1\ i} \\ \vdots \\ A_{n-1\ n}X = XB_{n-1\ n} \end{cases}$$

Image from Horaud 1995





## Solve $AX=XB$

- $A = \begin{bmatrix} R_A & t_A \\ 0 & 1 \end{bmatrix}$ , and so on for B and X
- $\begin{bmatrix} R_A & t_A \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R_X & t_X \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} R_X & t_X \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R_B & t_B \\ 0 & 1 \end{bmatrix}$
- $R_A R_X = R_X R_B$
- $R_A t_X + t_A = R_X t_B + t_X \Rightarrow (R_A - I)t_X = R_X t_B - t_A$
- Solve rotation first!



## Solve $R_A R_X = R_X R_B$

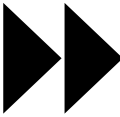


- Let  $n_B$  be  $R_B$ 's rotation axis
  - $n_B = R_B n_B$
- Multiply both sides by  $R_X$ 
  - $R_X n_B = R_X R_B n_B$
  - $\quad \quad = R_A R_X n_B$
- Consider  $R_A$ 's rotation axis:  $n_A = R_A n_A$
- So  $R_X n_B$  turns out to be  $R_A$ 's rotation axis
  - $n_A = R_X n_B$
- $R_X$  can be solved with multiple ( $\geq 2$ ) pairs of (A, B)!
  - Polar decomposition on a covariance matrix  $N_A N_B^T$
  - Orthogonal Procrustes problem





# Orthogonal Procrustes Problem



$$R = \arg \min_{\Omega} \|\Omega A - B\|_F \quad \text{subject to} \quad \Omega^T \Omega = I,$$

$$A: 3 \times n, B: 3 \times n, \Omega: 3 \times 3$$

$$= \arg \min_{\Omega} \langle \Omega A - B, \Omega A - B \rangle$$

$$\langle \mathbf{A}, \mathbf{B} \rangle_F = \sum_{i,j} \overline{A_{ij}} B_{ij}, = \text{tr}(\overline{\mathbf{A}^T} \mathbf{B})$$

$$= \arg \min_{\Omega} \|\Omega A\|_F^2 + \|B\|_F^2 - 2\langle \Omega A, B \rangle$$

$$= \arg \min_{\Omega} \|A\|_F^2 + \|B\|_F^2 - 2\langle \Omega A, B \rangle$$

$$= \arg \max_{\Omega} \langle \Omega, BA^T \rangle$$

$$= \arg \max_{\Omega} \langle \Omega, U \Sigma V^T \rangle$$

More generally, the trace is *invariant under cyclic permutations*, i.e.,

$$\text{tr}(ABCD) = \text{tr}(BCDA) = \text{tr}(CDAB) = \text{tr}(DABC).$$

$$= \arg \max_{\Omega} \langle U^T \Omega V, \Sigma \rangle$$

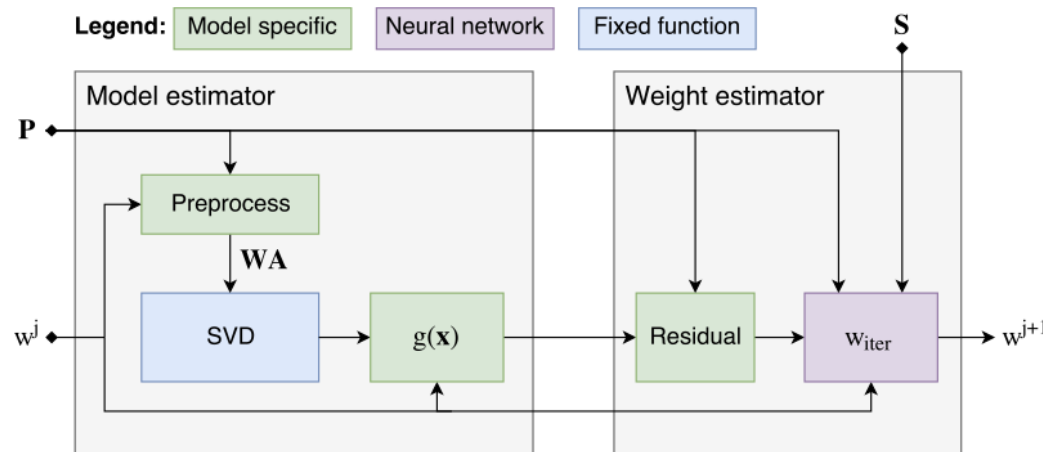
$$= \arg \max_{\Omega} \langle S, \Sigma \rangle \quad \text{where } S = U^T \Omega V$$

$$S^* = I \\ \implies$$

$$I = U^T R V \\ R = U V^T$$



# Recent Works on Multiview Geometry with Deep Learning

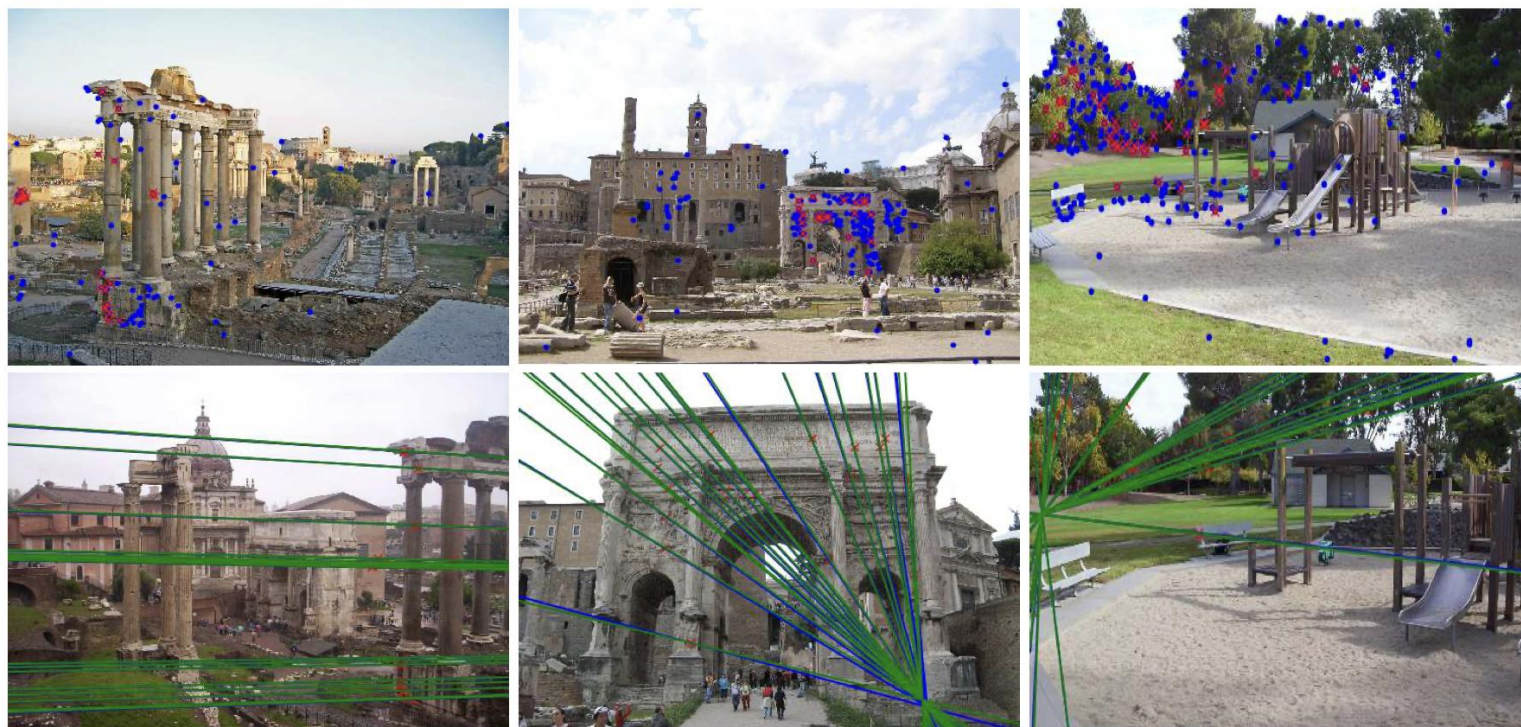


Layer	# in	# out	L-ReLU+IN
1	–	64	✓
2	64	128	✓
3	128	1024	✓
4	1024	512	✓
5	512	256	✓
6	256	1	✗

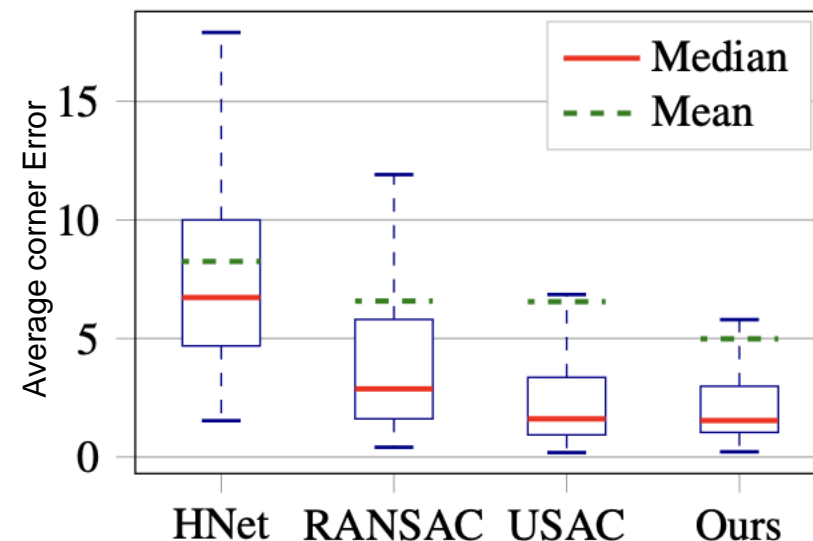
**Figure 1 & Table 1.** Estimation module and network architecture. Left: The estimation module is composed of two parts. Given input points and a weighting, a model is estimated using weighted least-squares. In the second stage a new set of weights is generated given the points, their residuals with respect to the previously estimated model, and possibly side information. Right: The network architecture of  $w_{init}$  and  $w_{iter}$ . A checkmark in column L-ReLU+IN indicates that a leaky ReLU followed by instance normalization is applied to the output of the layer.

**The Challenge:** Estimating Fundamental Matrix parameters via reweighted least-squares algorithm (IRLS) with a complex, learned reweighting function (parameterized by a NN basically)

# Recent Works on Multiview Geometry with Deep Learning



**Figure 3.** Image pairs from *Roman Forum* (first and second column) and *Tanks and Temples* (last column). Top row: First image with inliers (red) and outliers (blue). Bottom row: Epipolar lines of a random subset of inliers in the second image. We show the epipolar lines of our estimate (green) and of the groundtruth (blue). Images have been scaled for visualization.

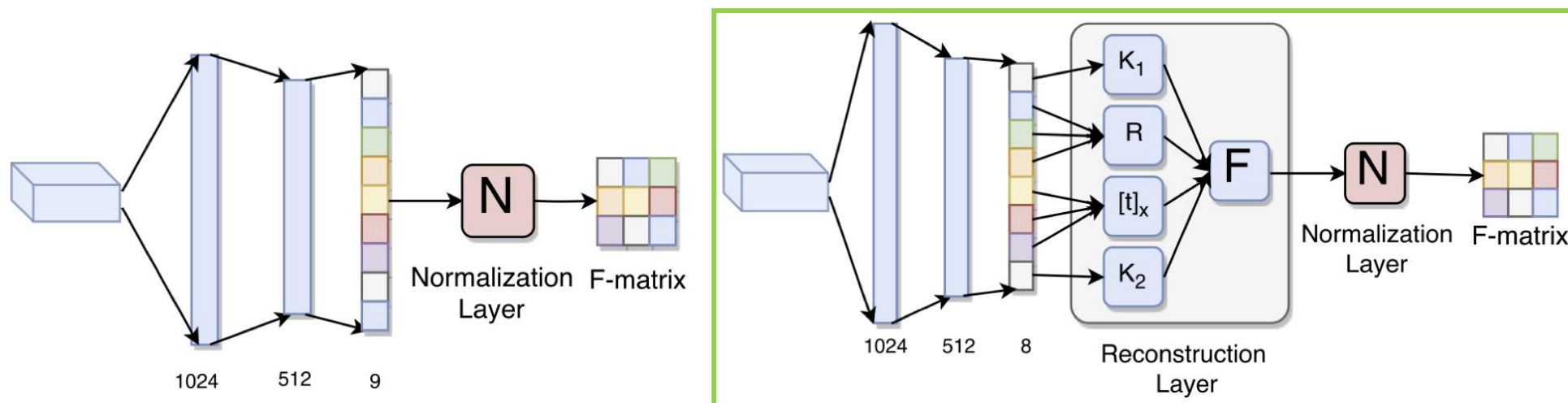


**Notes:** It does quite well against traditional parameters estimation method. We can see some example of the Epipolar lines across corresponding views on the left.

Can we do better still? Yes!



# Recent Works on Multiview Geometry with Deep Learning



**Fig. 3.** Different regression methods for predicting F-matrix entries from the features. The architecture to directly regress the entries of the F-matrix is shown on the left. The network with the reconstruction and normalization layers is shown on the right, and is able to estimate homogeneous F-matrices with rank two and seven degrees of freedom.

## The Idea:

1. Here we don't need correspondences to estimate the matrix, we just need the RGB input of Multiview images!
2. And instead of simple direct linear regression, they perform multiple regression via MLP layers specifically to reconstruct the individual components of the F-matrix to ensure proper structure – e.g. F matrix is of Rank 2.



# Recent Works on Multiview Geometry with Deep Learning

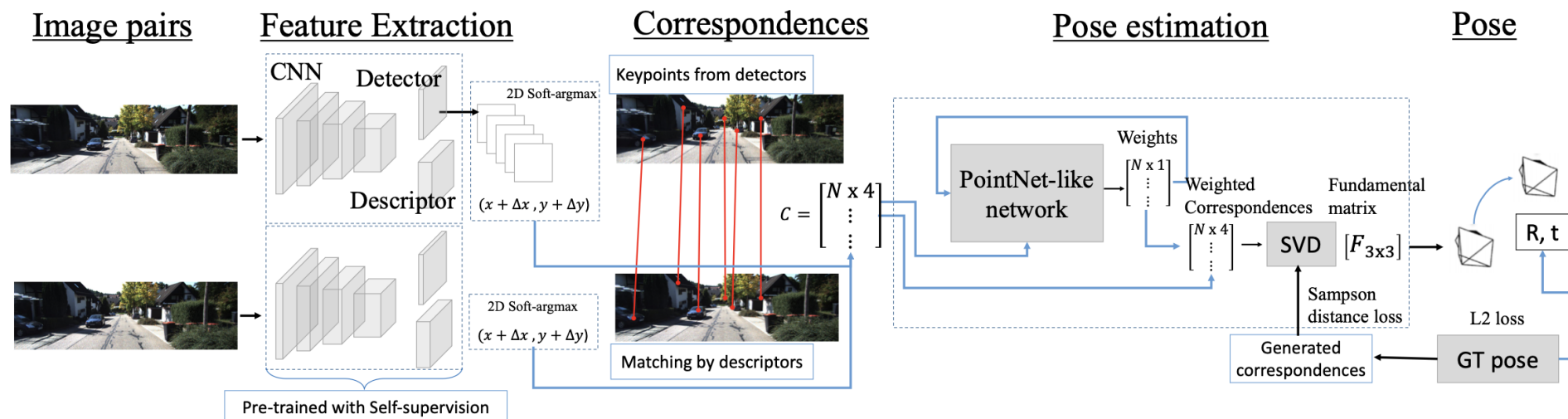


Fig. 1: **Overview of the system.** A pair of images is fed into the pipeline to predict the relative camera pose. Feature extraction predicts detection heatmaps and descriptors for finding sparse correspondences. Local 2D Softargmax is used as a bridge to get subpixel prediction with gradients. Matrix  $C$  of size  $N \times 4$  is formed from correspondences.  $C$  is the input for pose estimation, where the PointNet-like network predicts weights for all correspondences. Weighted correspondences are passed through SVD to find fundamental matrix  $F$ , which is further decomposed into poses. Ground truth poses (GT poses) are used to compute L2 loss between rotation and translation (**pose-loss**). Correspondences generated from GT poses are used to compute fundamental matrix loss (**F-loss**). See more details in Sec. III.

## The Idea:

1. Kind of a combination of prior methods – **Deep Feature Extraction** via SuperPoint + **Pose / F Matrix estimation** liken to "Deep Fundamental Matrix Estimation" from slide 26.
2. End-to-end framework that performs feature extraction, correspondence matching and parameter estimation.

# Recent Works on Multiview Geometry with Deep Learning

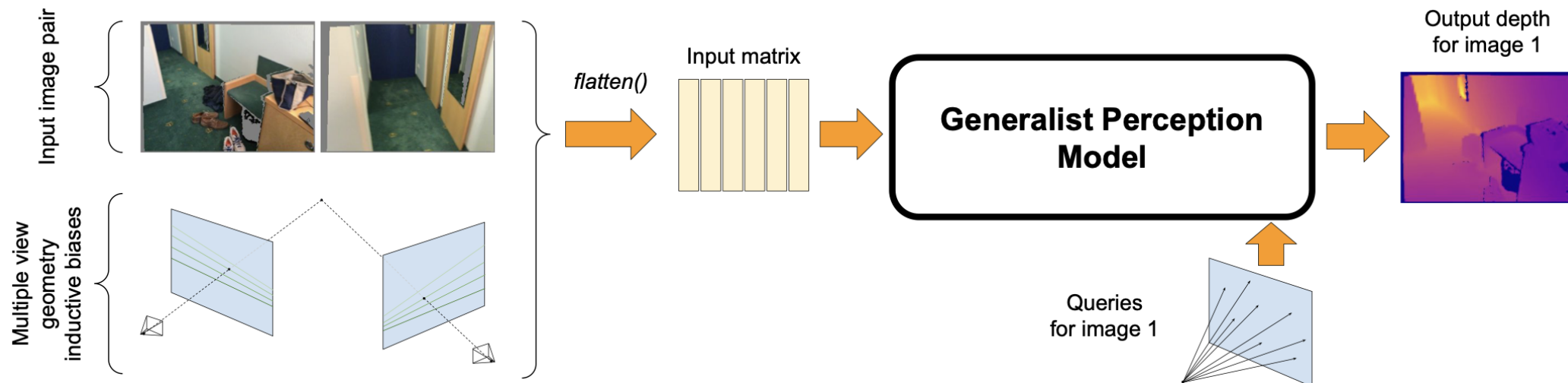


Figure 1. **Input-level inductive biases.** We explore 3D reconstruction using a generalist perception model, the recent Perceiver IO [23] which ingests a matrix of unordered and flattened inputs (*e.g.* pixels). The model is interrogated using a query matrix and generates an output for every query – in this paper the outputs are depth values for all pixels of the input image pair. We incorporate inductive biases useful for multiple view geometry into this generalist model without having to touch its architecture, by instead encoding them directly as additional inputs.

## The Idea:

1. Here instead of estimating the intrinsic, the paper proposes a way to supply existing DL based vision models with handy information such as the aforementioned camera intrinsic as priors! (basically, the reverse problem from previous slides)
2. E.g. for Depth estimation with [Perceiver IO](#) (multipurpose vision model based with Transformers)



# Overview

- + Feature detection

  - Harris/FAST/DoG

- + Feature description & matching

  - SIFT/SURF

- ++ Linear & total least square

- \* RANSAC

  - Intuitions behind RANSAC

  - How RANSAC works

  - Why minimal solution is important

  - More example problems

  - Variations

\*: know how to code

++: know how to derive

+: know the concept



## References

- HZ2003:
  - Section 4.7, 4.8, 11.6
- Corke 2011:
  - Section 14.2.3
- Sz2022:
  - Section 7.1, 7.2, 8.1.4
- DeTone, D., Malisiewicz, T. and Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. In *CVPR workshops* (pp. 224-236).
- Sarlin, P.E., DeTone, D., Malisiewicz, T. and Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. In *CVPR* (pp. 4938-4947).