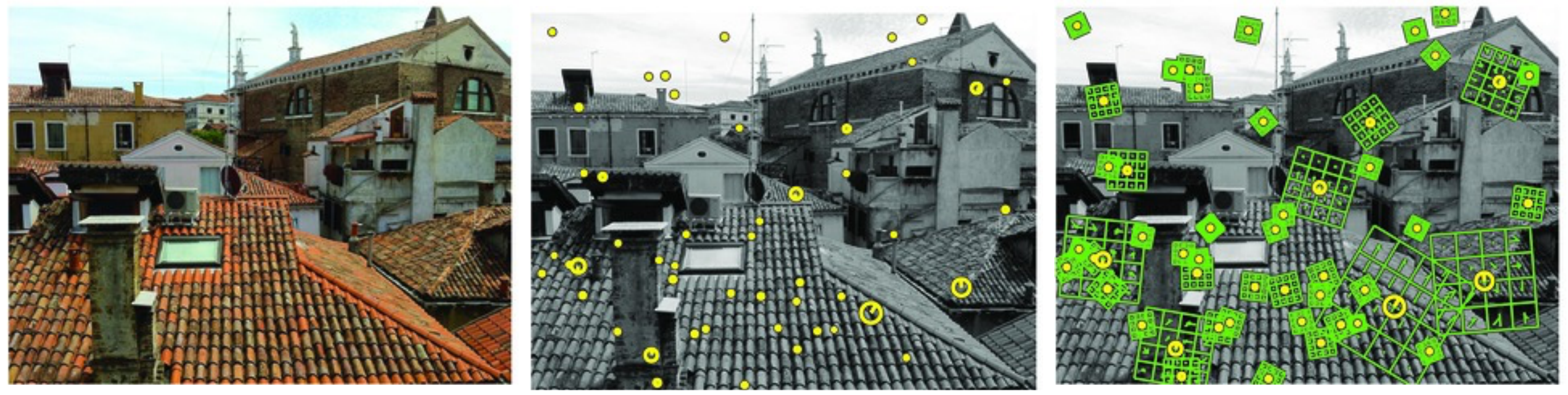
THE UNIVERSITY OF BRITISH COLUMBIA

CPSC 425: Computer Vision



Lecture 16: SIFT cont., HOG, SURF, Object Recognition

Menu for Today (**March 5, 2019**)

Topics:

— SIFT continued
— HOG, SURF descriptors
— Object detection with SIFT
— RANSAC intro

Redings:

— **Today's** Lecture: Forsyth & Ponce (2nd ed.) 5.4, 10.4.2
"Distinctive Image Features for Scale-Invariant Keypoints

— **Today's** & **Next** Lecture: Forsyth & Ponce (2nd ed.) 10.1, 10.2

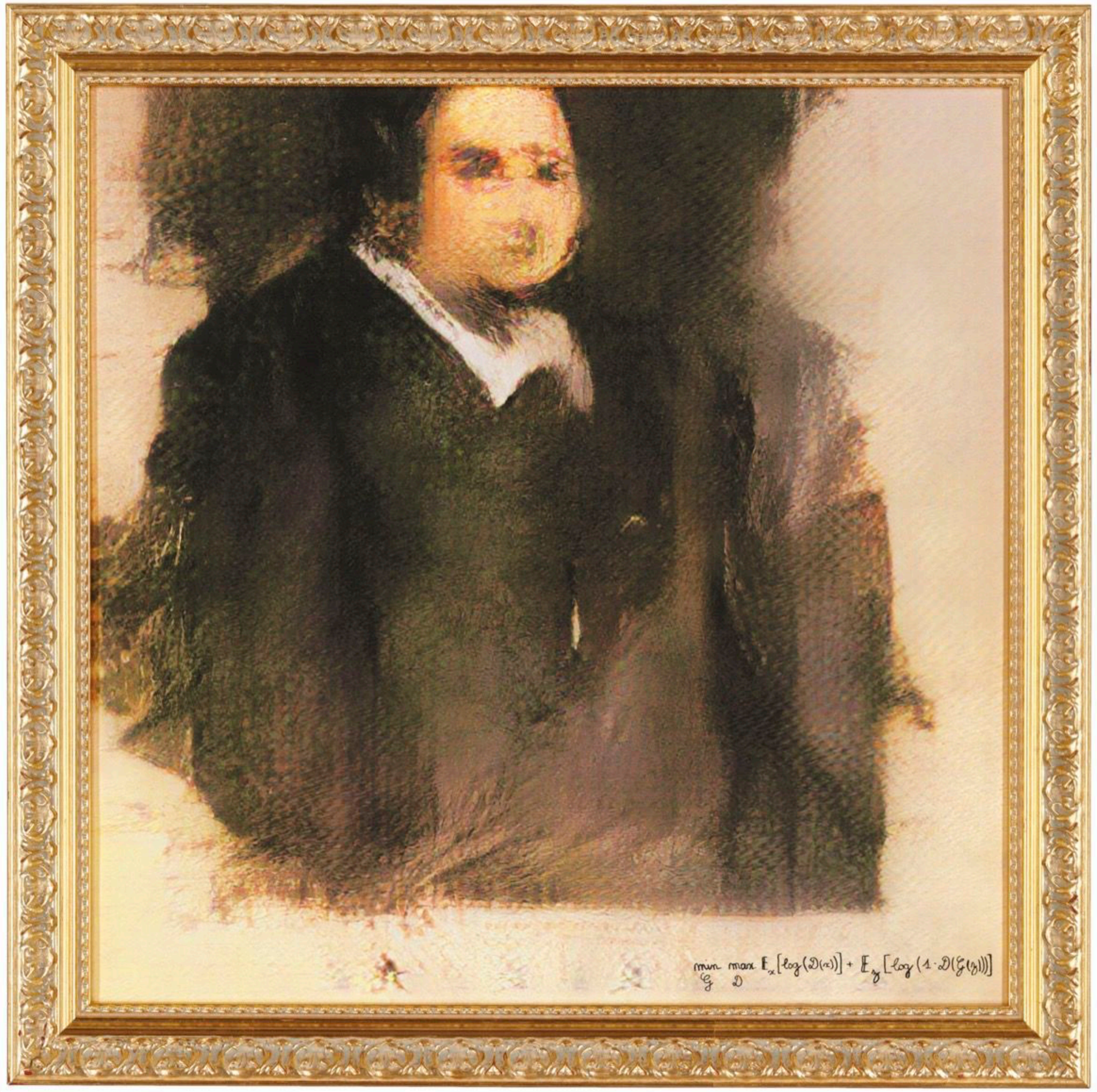
Reminders:

— **Assignment 4**: will be out **today**

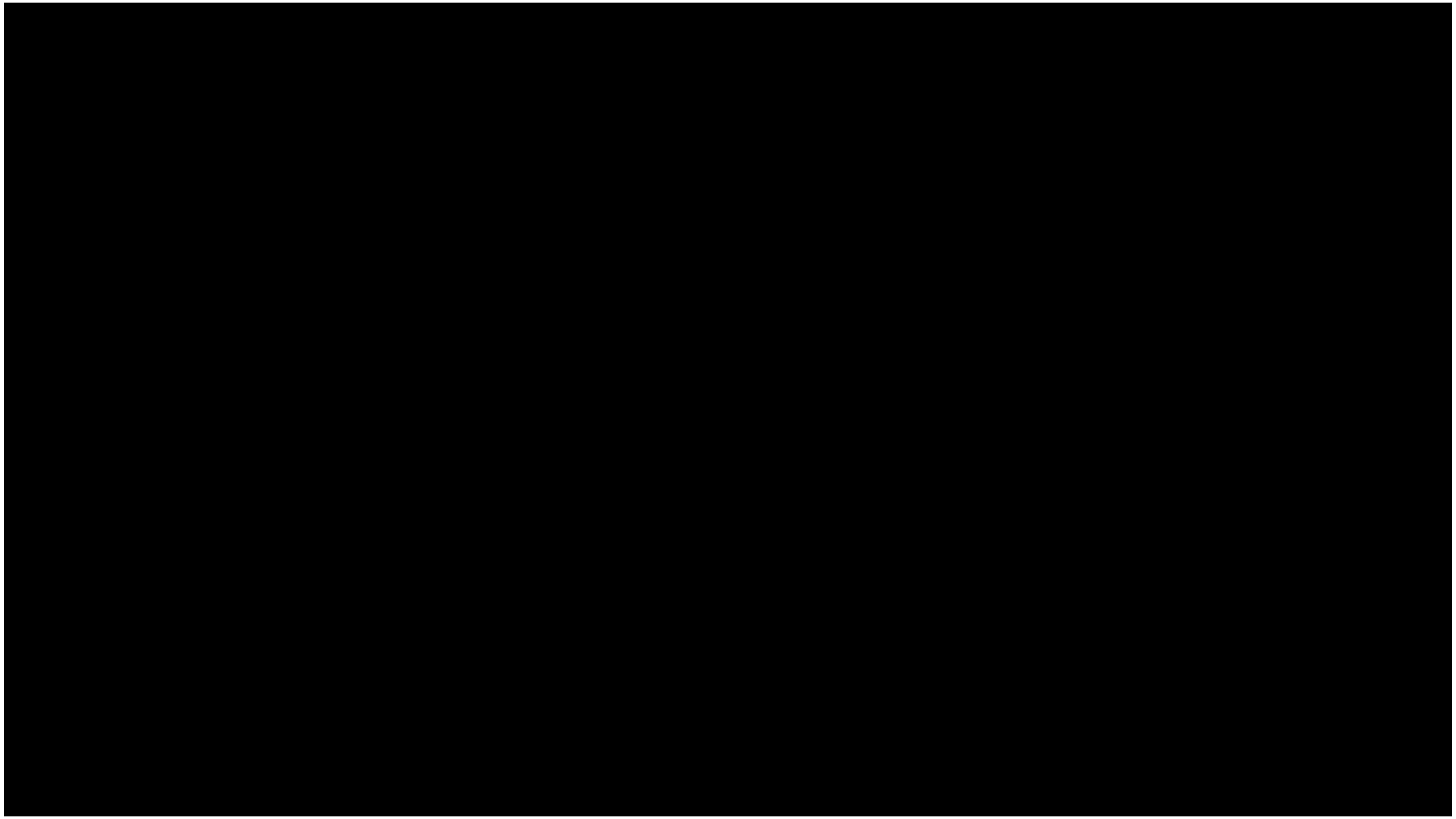
— **Midterm** is almost graded (will return Thursday/Friday)

Today's "**fun**" Example: AI Generated Portrait

Sold 5 months ago for $432,500 at British auction house



Today's "**fun**" Example: Sunspring



Lecture 18: Re-Cap

— We motivated SIFT for identifying locally distinct keypoints in an image (**detection**)

— SIFT features (**description**) are invariant to translation, rotation, and scale; robust to 3D pose and illumination

1. Multi-scale extrema detection
2. Keypoint localization
3. Orientation assignment
4. Keypoint descriptor

Lecture 18: Re-Cap

Keypoint is an image location at which a descriptor is computed

— Locally distinct points

— Easily localizable and identifiable

The feature **descriptor** summarizes the local structure around the key point

— Allows us to (hopefully) unique matching of keypoints in presence of object pose variations, image and photometric deformations

Note, for repetitive structure this would still not give us unique matches.

Lecture 18: Re-Cap

Keypoint is an image location at which a descriptor is computed

— Locally distinct points

— Easily localizable and identifiable

The feature **descriptor** summarizes the local structure around the key point

— Allows us to (hopefully) unique matching of keypoints in presence of object pose variations, image and photometric deformations

Note, for repetitive structure this would still not give us unique matches.

Lecture 18: Re-Cap

Keypoint is an image location at which a descriptor is computed

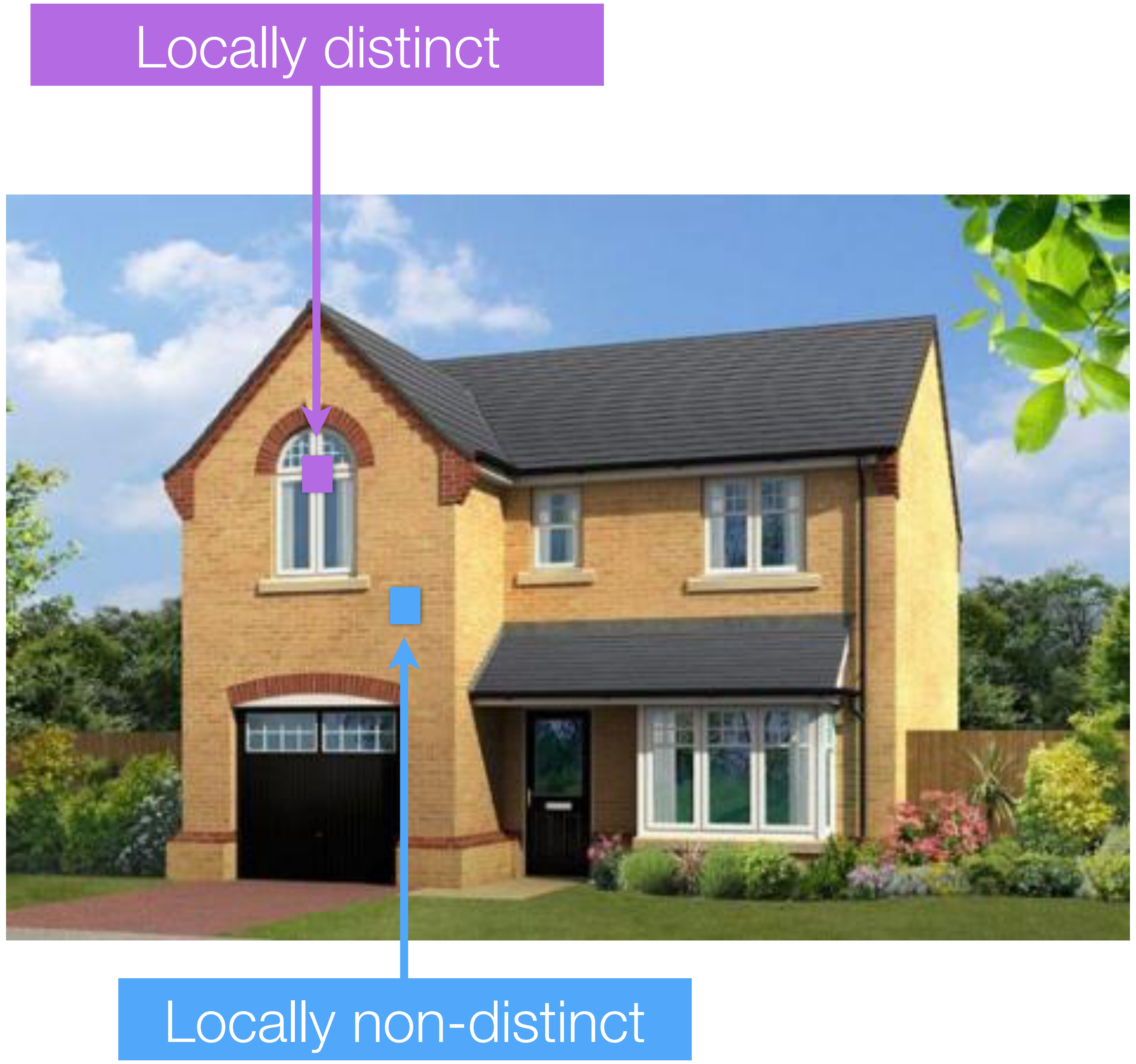
— Locally distinct points

— Easily localizable and identifiable

The feature **descriptor** summarizes the local structure around the key point

— Allows us to (hopefully) unique matching of keypoints in presence of object pose variations, image and photometric deformations

Note, for repetitive structure this would still not give us unique matches.



Lecture 18: Re-Cap

Keypoint is an image location at which a descriptor is computed

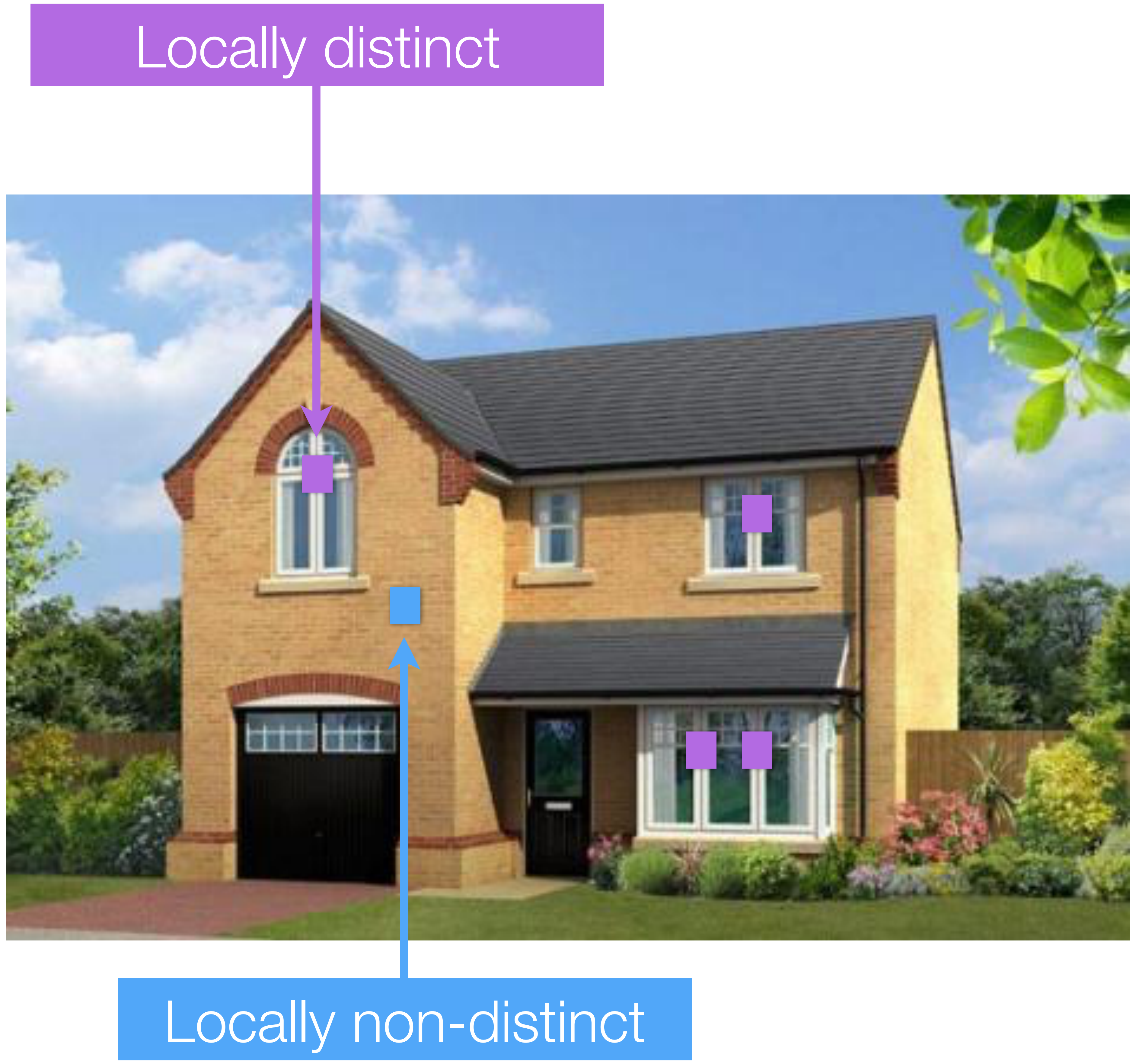
— Locally distinct points

— Easily localizable and identifiable

The feature **descriptor** summarizes the local structure around the key point

— Allows us to (hopefully) unique matching of keypoints in presence of object pose variations, image and photometric deformations

Note, for repetitive structure this would still not give us unique matches.



Lecture 18: Re-Cap

Keypoint is an image location at which a descriptor is computed

— Locally distinct points

— Easily localizable and identifiable

The feature **descriptor** summarizes the local structure around the key point

— Allows us to (hopefully) unique matching of keypoints in presence of object pose variations, image and photometric deformations

Note, for repetitive structure this would still not give us unique matches.

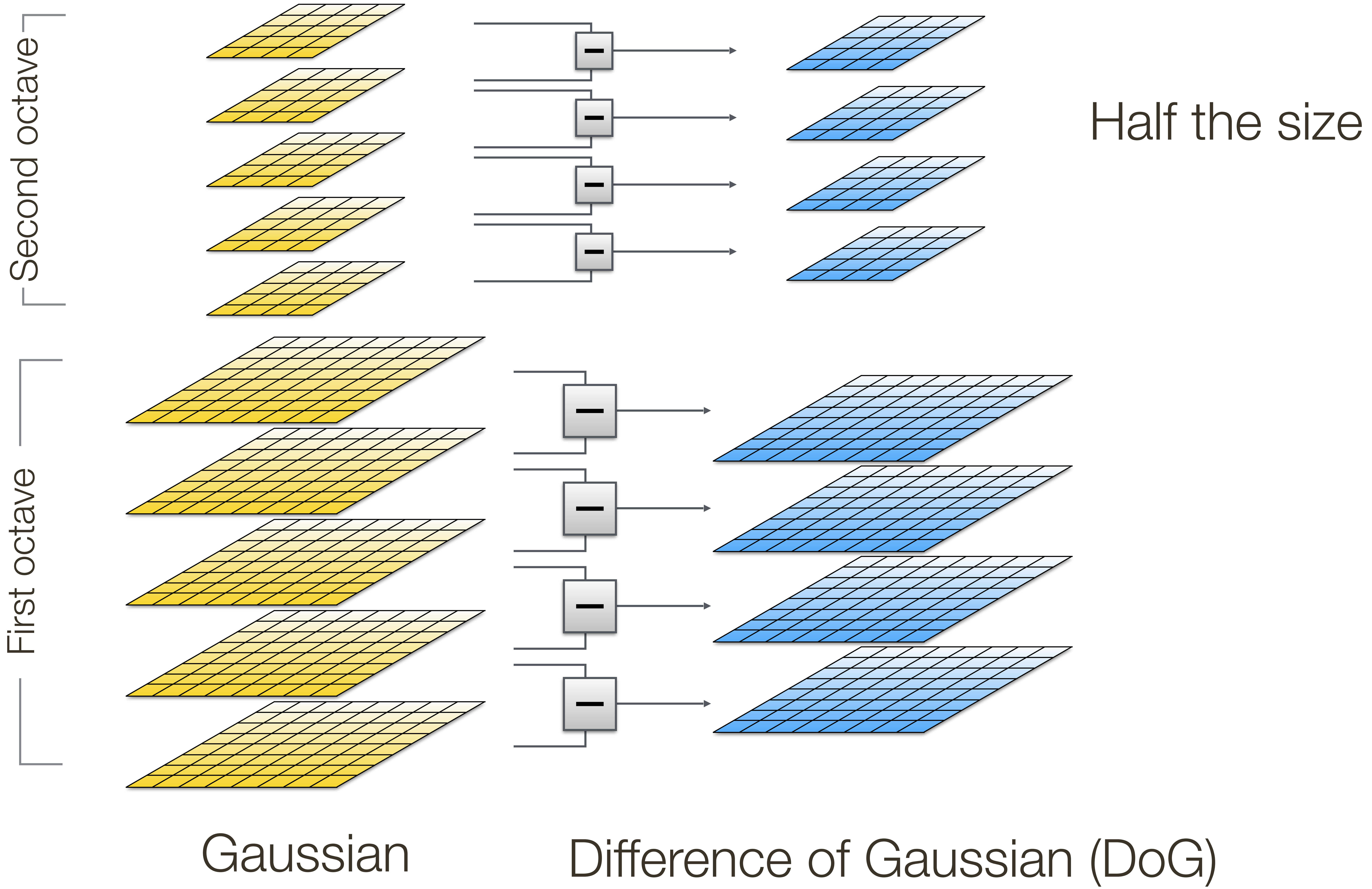
Lecture 18: Re-Cap

— We motivated SIFT for identifying locally distinct keypoints in an image (**detection**)

— SIFT features (**description**) are invariant to translation, rotation, and scale; robust to 3D pose and illumination

1. Multi-scale extrema detection
2. Keypoint localization
3. Orientation assignment
4. Keypoint descriptor

1. Multi-scale Extrema Detection

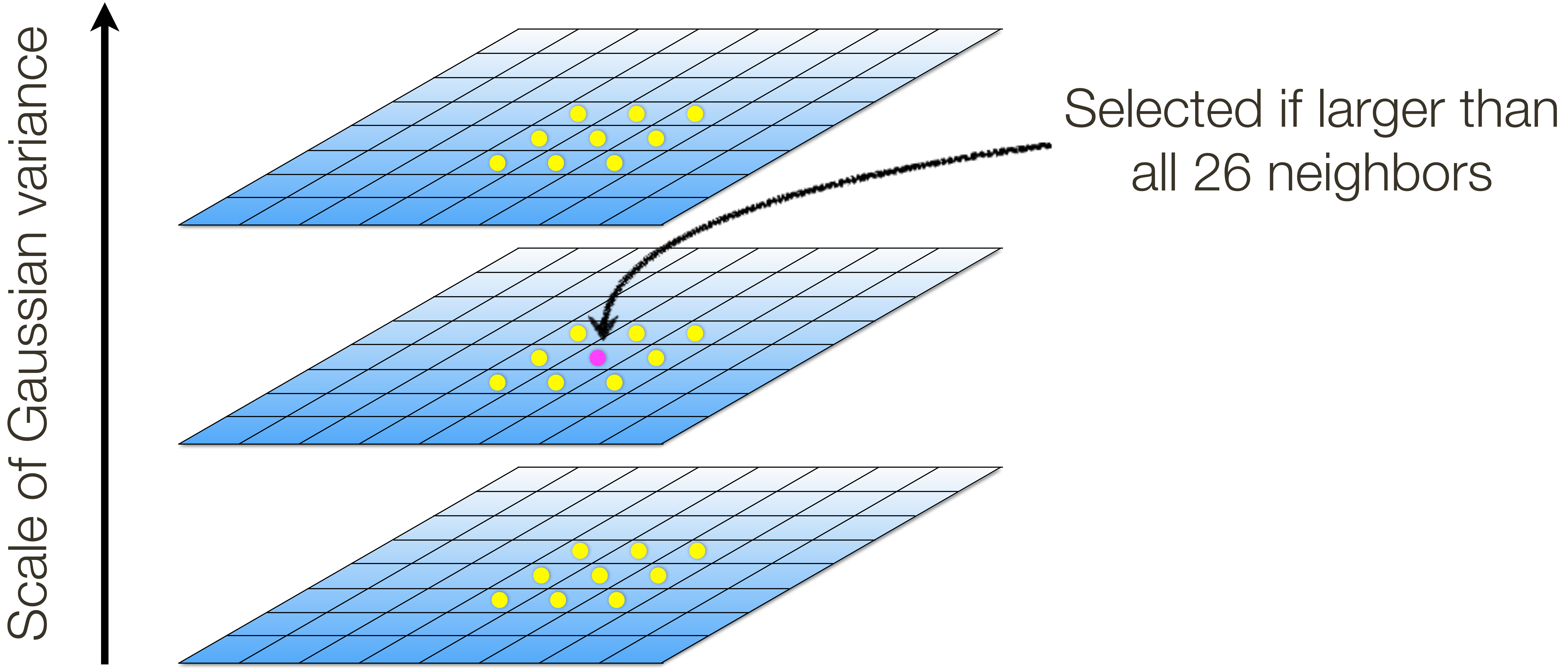


Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

1. Multi-scale Extrema Detection

Detect maxima and minima of Difference of Gaussian in scale space

$$\sigma = 2^{1/s}$$



Difference of Gaussian (DoG)

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

1. Multi-scale Extrema Detection

Detect maxima and minima of Difference of Gaussian in scale space

— Responds to blob-line and corner-like structues

— Could also give strong responses at edges

2. Keypoint Localization

— After keypoints are detected, we remove those that have **low contrast** or are **poorly localized** along an edge

How do we decide whether a keypoint is poorly localized, say along an edge, vs. well-localized?

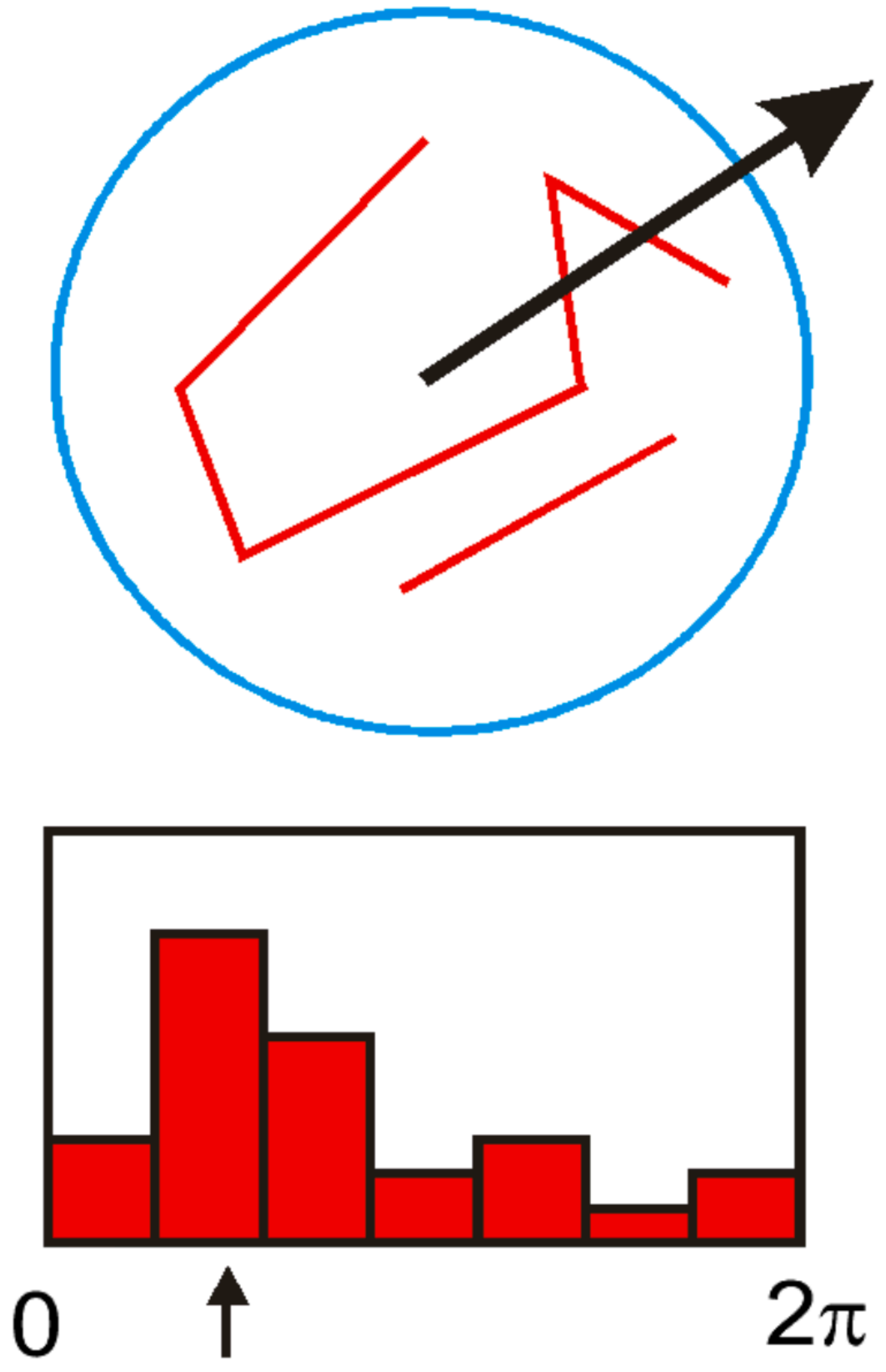
$$C = \begin{bmatrix} \sum_{p \in P} I_x I_x & \sum_{p \in P} I_x I_y \\ \sum_{p \in P} I_y I_x & \sum_{p \in P} I_y I_y \end{bmatrix}$$

3. Orientation Assignment

— Create **histogram** of local gradient directions computed at selected scale

— Assign **canonical orientation** at peak of smoothed histogram

— Each key specifies stable 2D coordinates (x , y , scale, orientation)

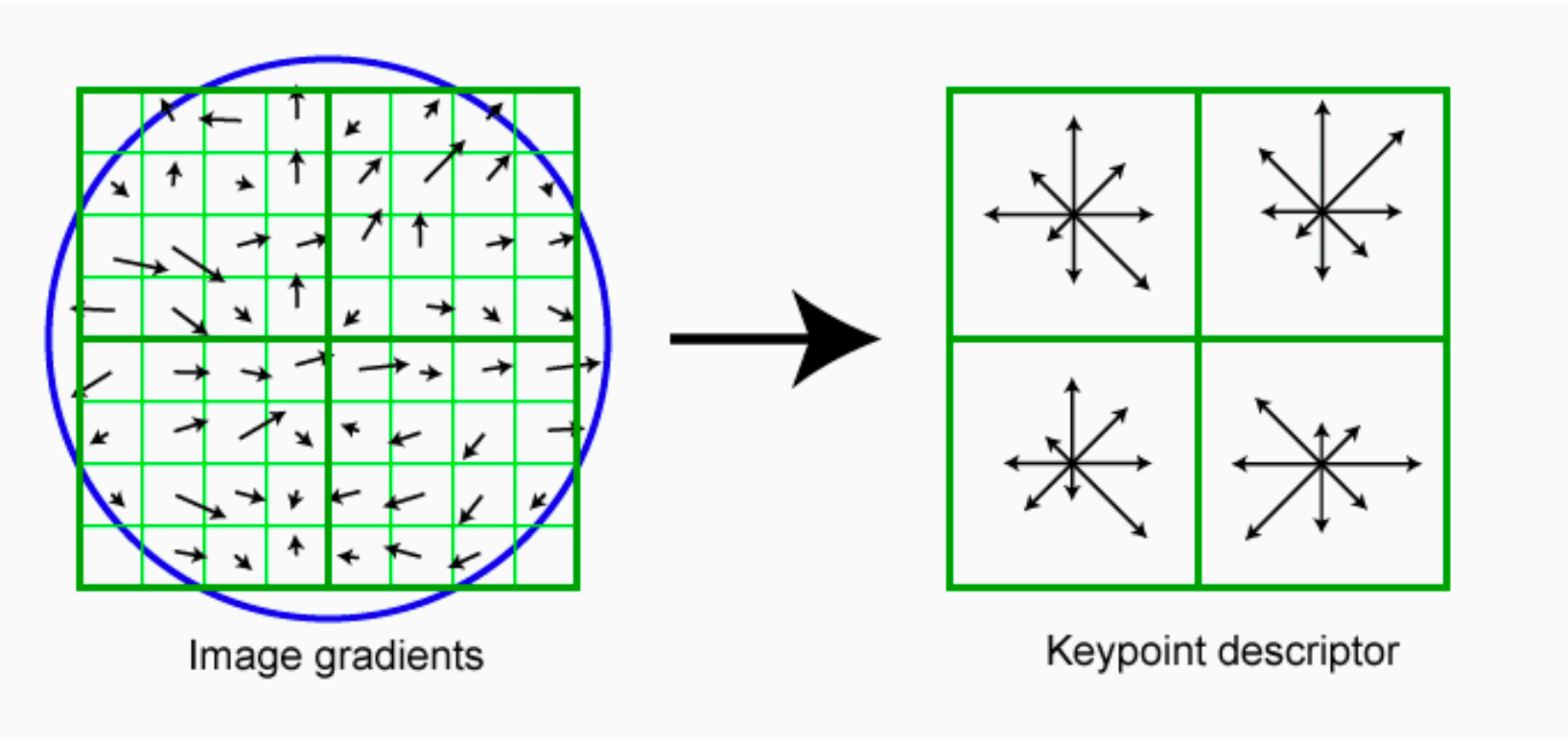


4. SIFT Descriptor

— Thresholded image gradients are sampled over 16 × 16 array of locations in scale space (weighted by a Gaussian with sigma half the size of the window)

— Create array of orientation histograms

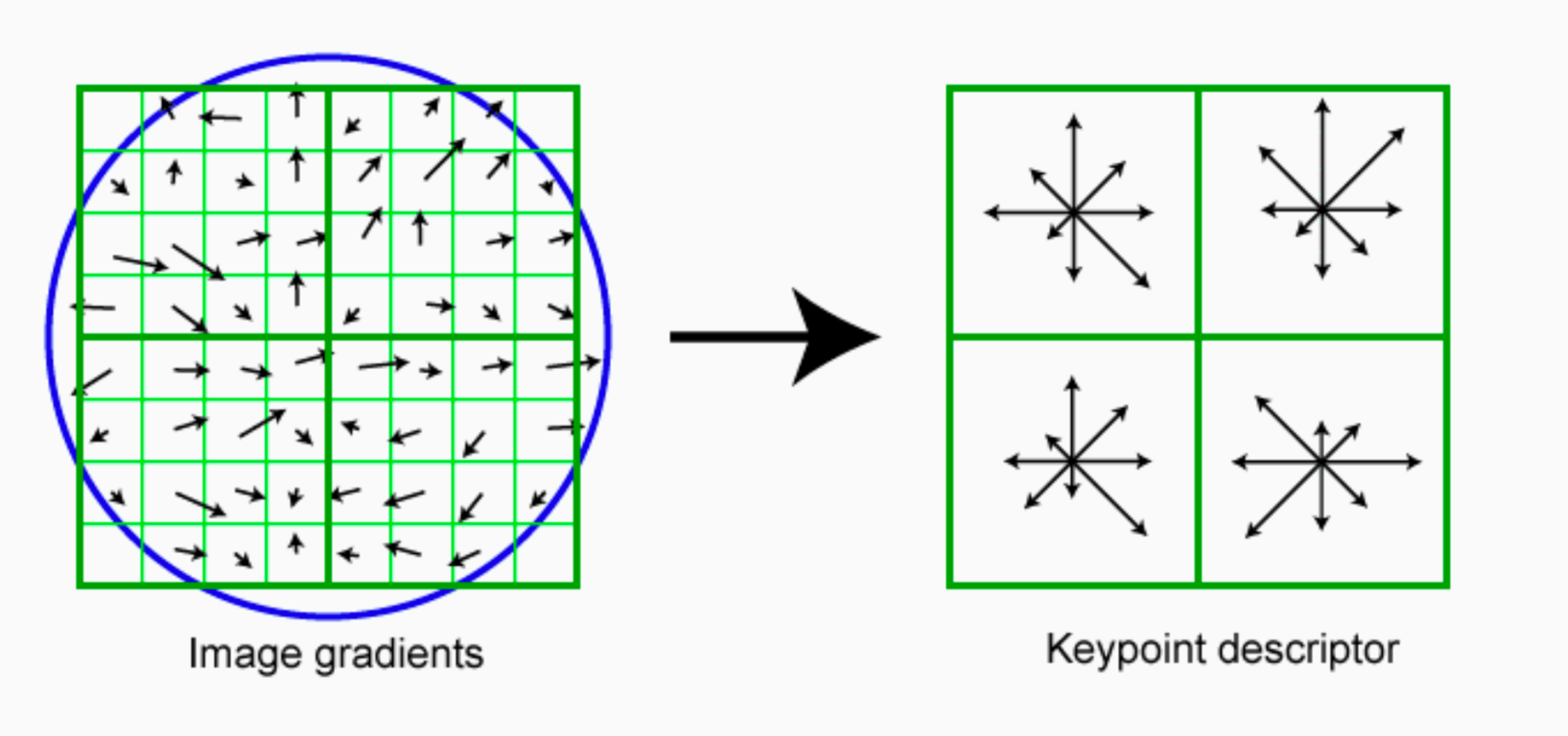
— 8 orientations × 4 × 4 histogram array



4. SIFT Descriptor

How many dimensions are there in a SIFT descriptor?

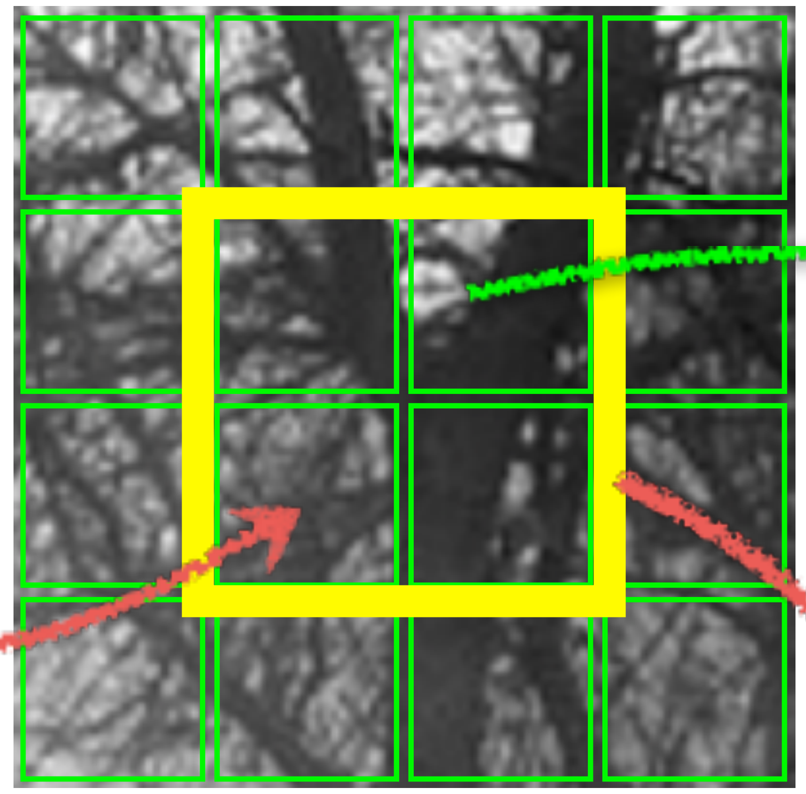
(**Hint**: This diagram shows a 2 x 2 histogram array but the actual descriptor uses a 4 x 4 histogram array)



Histogram of Oriented Gradients (**HOG**) Features

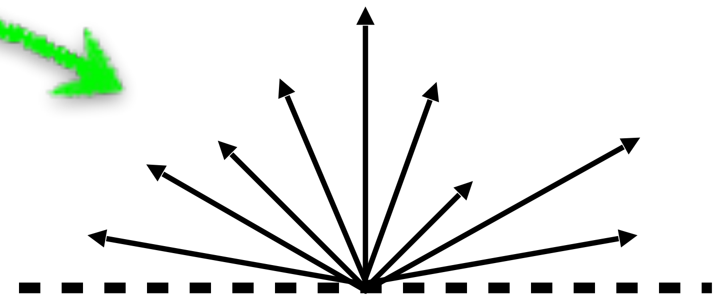
Dalal, Triggs. Histograms of Oriented Gradients for Human Detection. CVPR, 2005

Cell
(8x8 pixels)



Block
(2x2 cells)

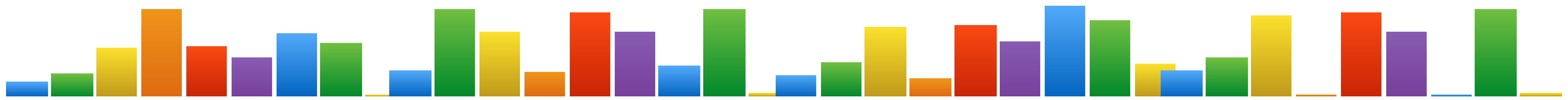
histogram of
'unsigned'
gradients



soft binning

gradient magnitude histogram
(one for each cell)

Concatenate and L-2 normalization



Single scale, no dominant orientation

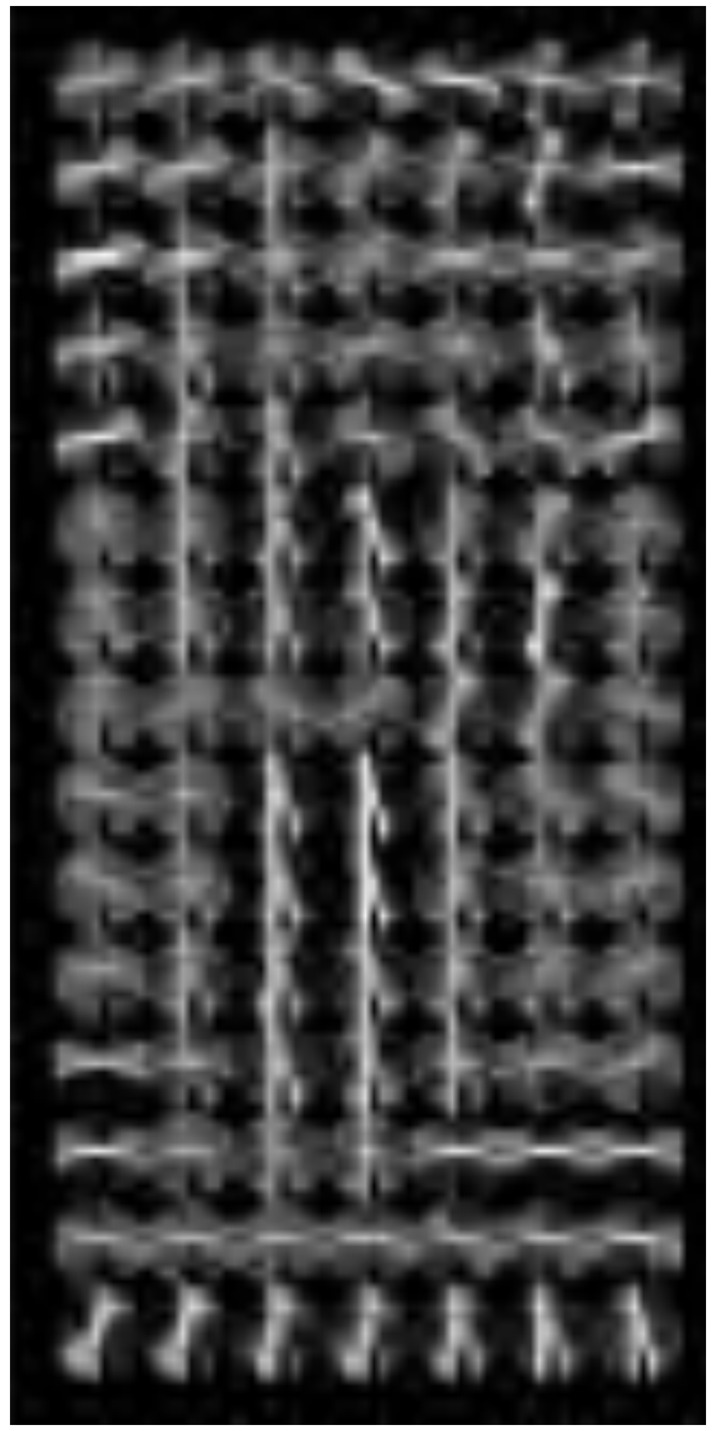
Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Histogram of Oriented Gradients (**HOG**) Features

1 cell step size

visualization

Pedestrian detection



128 pixels
16 cells
15 blocks

15 x 7 x 4 x 36 =
3780

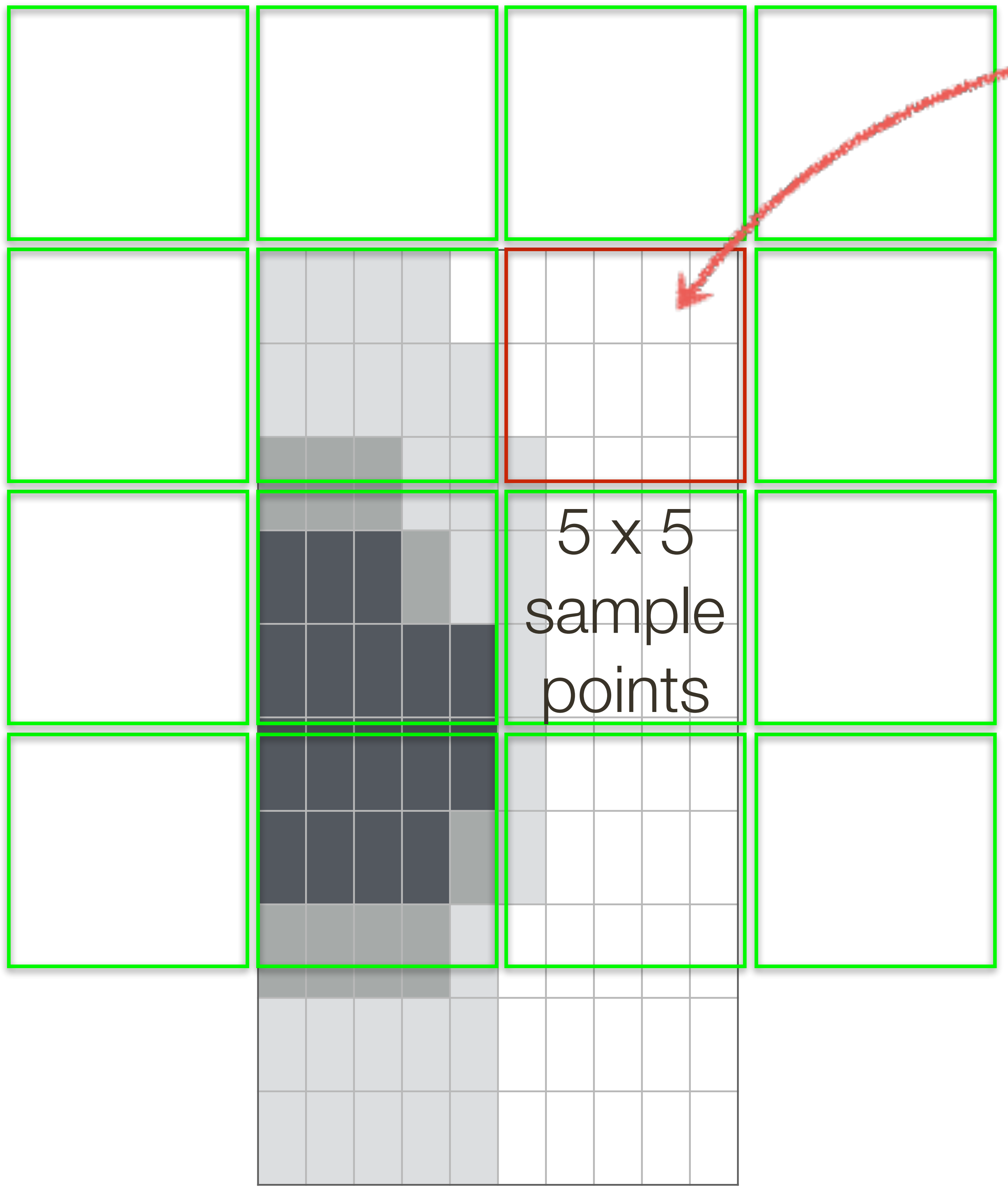
64 pixels
8 cells
7 blocks

Redundant representation due to overlapping blocks

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

‘Speeded’ Up Robust Features (**SURF**)

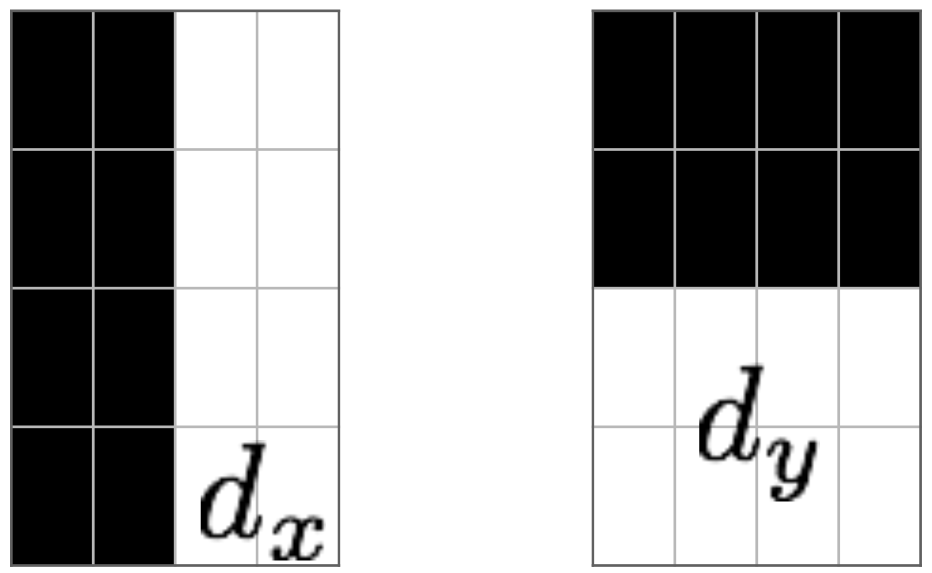
4 x 4 cell grid



Each cell is represented by 4 values:

$$\left[\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|\right]$$

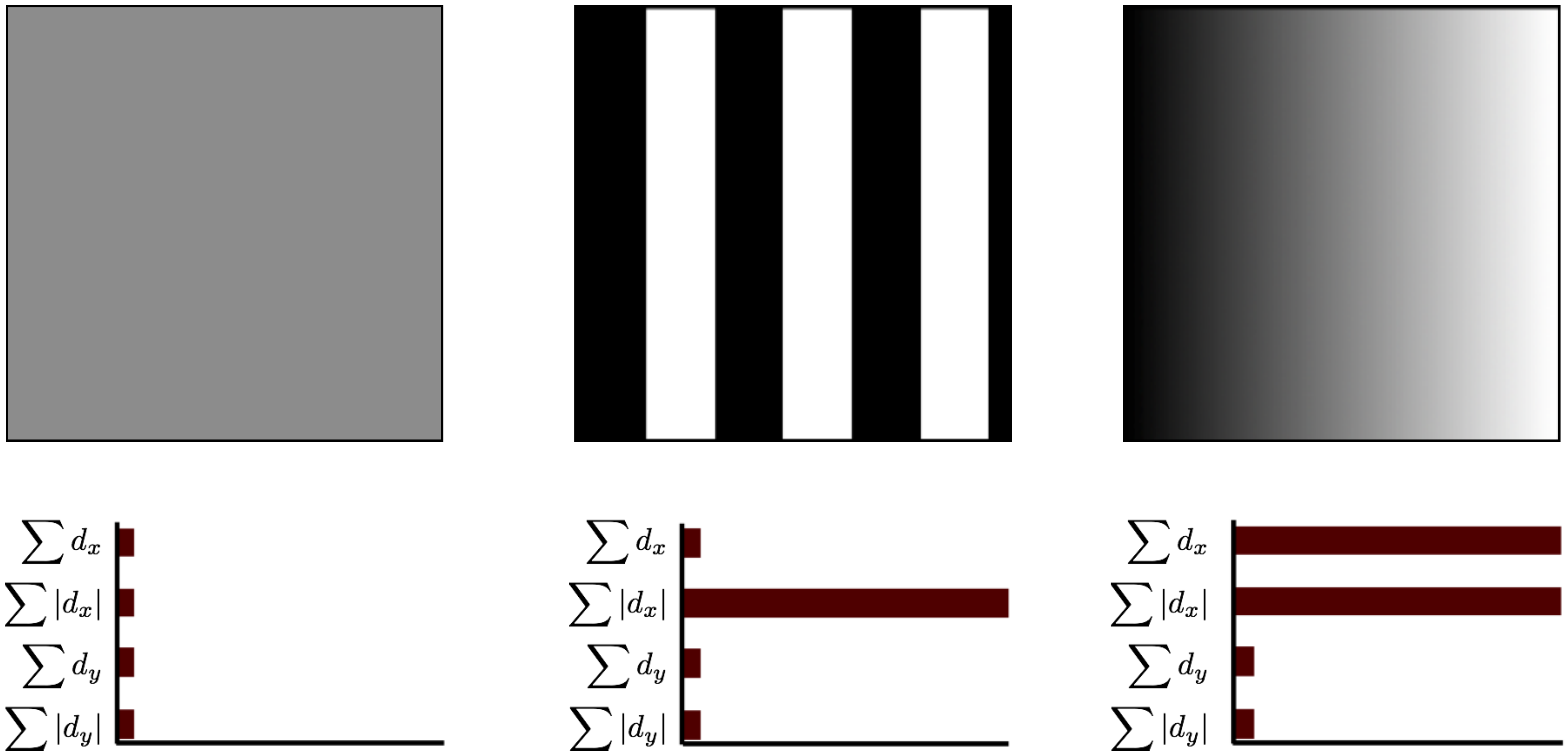
Haar wavelets filters
(Gaussian weighted from center)



How big is the SURF descriptor?

64 dimensions

‘Speeded’ Up Robust Features (**SURF**)



Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

SIFT and **Object Recognition**

Object recognition requires us to first match each keypoint independently to the database of keypoints

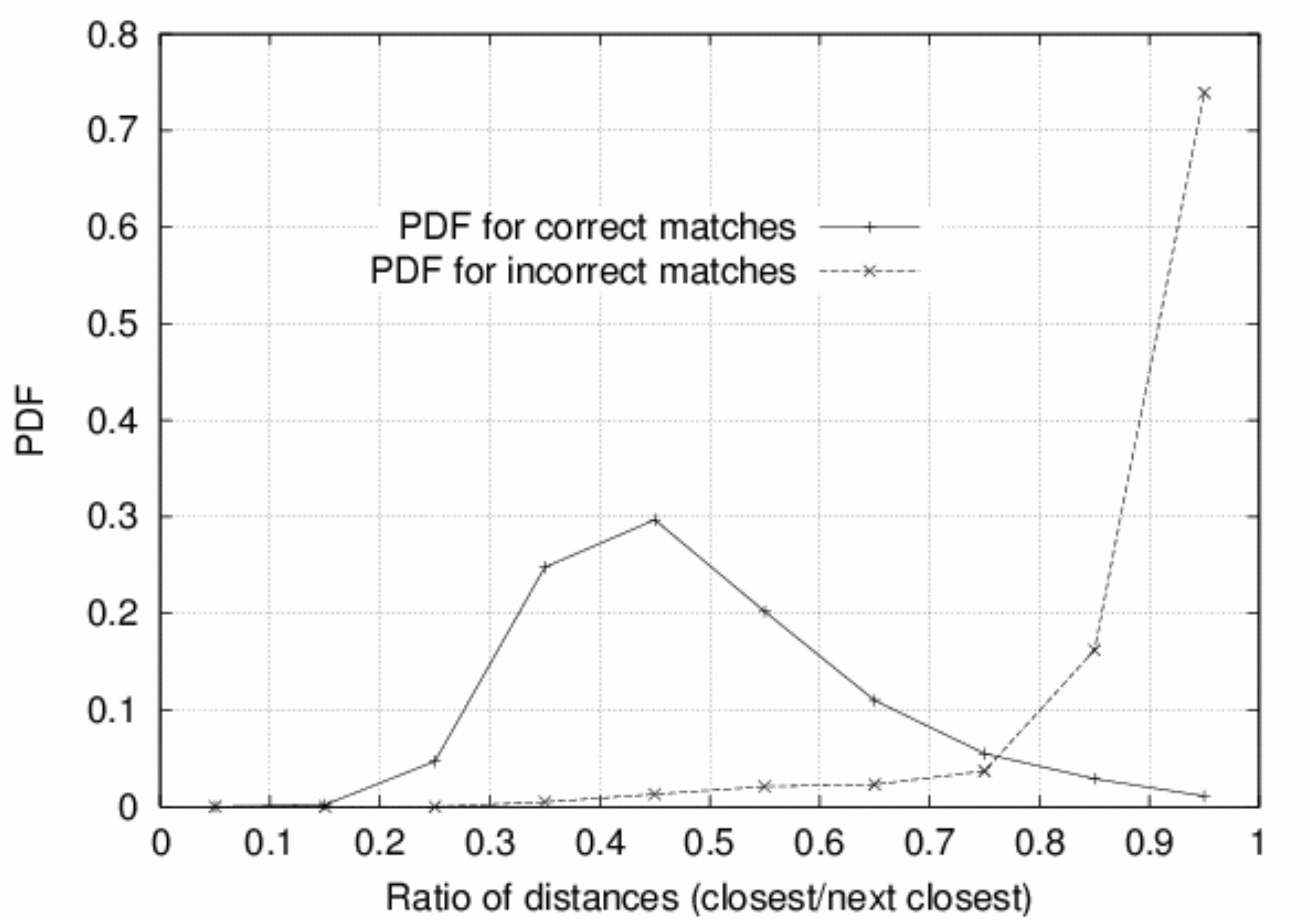
Many features will not have any correct match in the database because they arise from background clutter

It would be useful to have a way to **discard features** that do not have any good match

Probability of **Correct** Match

Compare ratio of distance of **nearest** neighbour to **second** nearest neighbour (from different object)

Threshold of 0.8 provides excellent separation



Nearest-Neighbor Matching to Feature Database

Hypotheses are generated by **approximate nearest neighbour** matching of each feature to vectors in the database

— Use best–bin–first (Beis & Lowe, 97) modification to k-d tree algorithm

— Use heap data structure to identify bins in order by their distance from query point

Result: Can give speedup by factor of 1,000 while finding nearest neighbour (of interest) 95% of the time

Identifying **Consistent** Features

We have matched keypoints to a database of known keypoints extracted from training images

Next we identify **clusters of at least 3 features** that agree on an object and its pose

— a typical image contains 2,000+ features → detecting less than 1% inliers among 99% outliers!

Lowe's solution uses the generalized **Hough transform**

— vote for each potential match according to model ID and pose

— insert into multiple bins to allow for error in similarity approximation

— (more on Hough transforms later)

Model **Verification**

1. Examine all clusters with at least 3 features

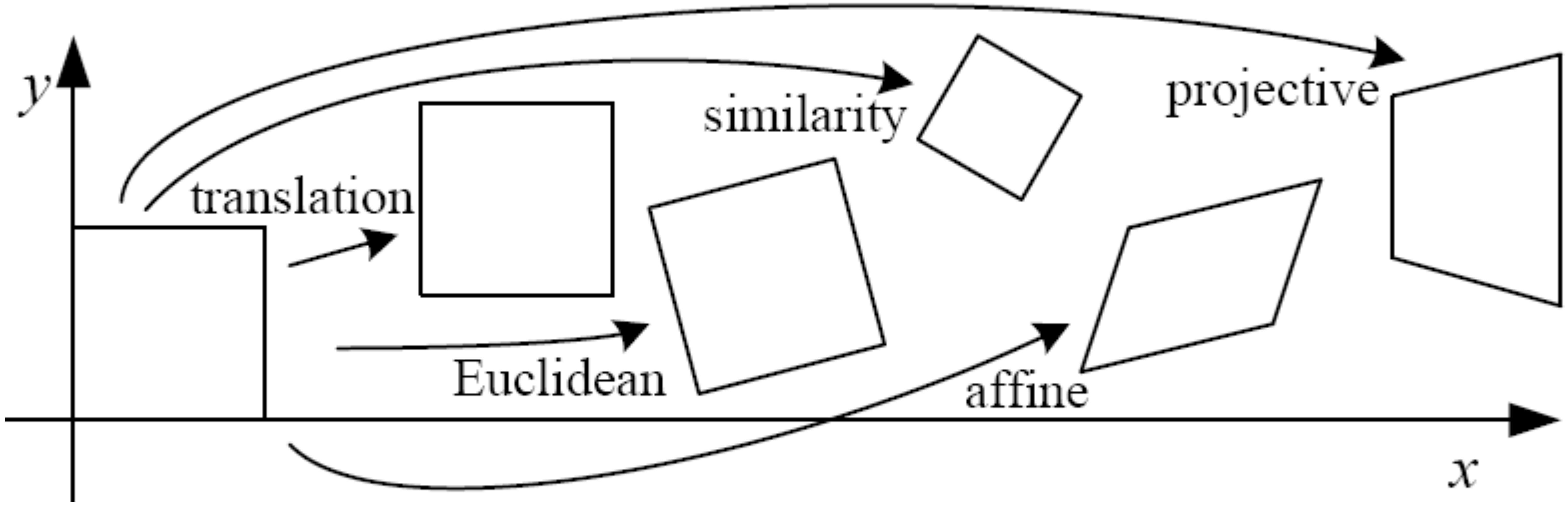
2. Perform least-squares affine **fit to model**

3. **Discard outliers** and perform top-down check for additional features

4. Evaluate probability that match is correct

   — Use Bayesian model, with probability that features would arise by chance if object was not present (Lowe, CVPR 01)

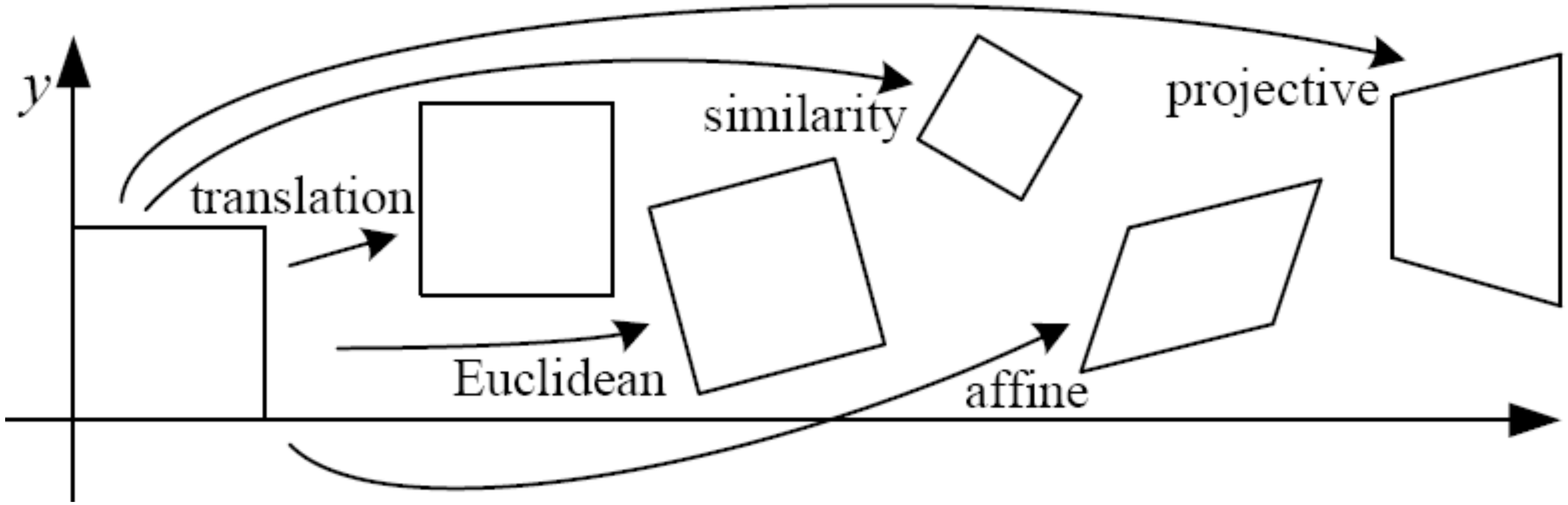
Aside: Classification of 2D Transformations



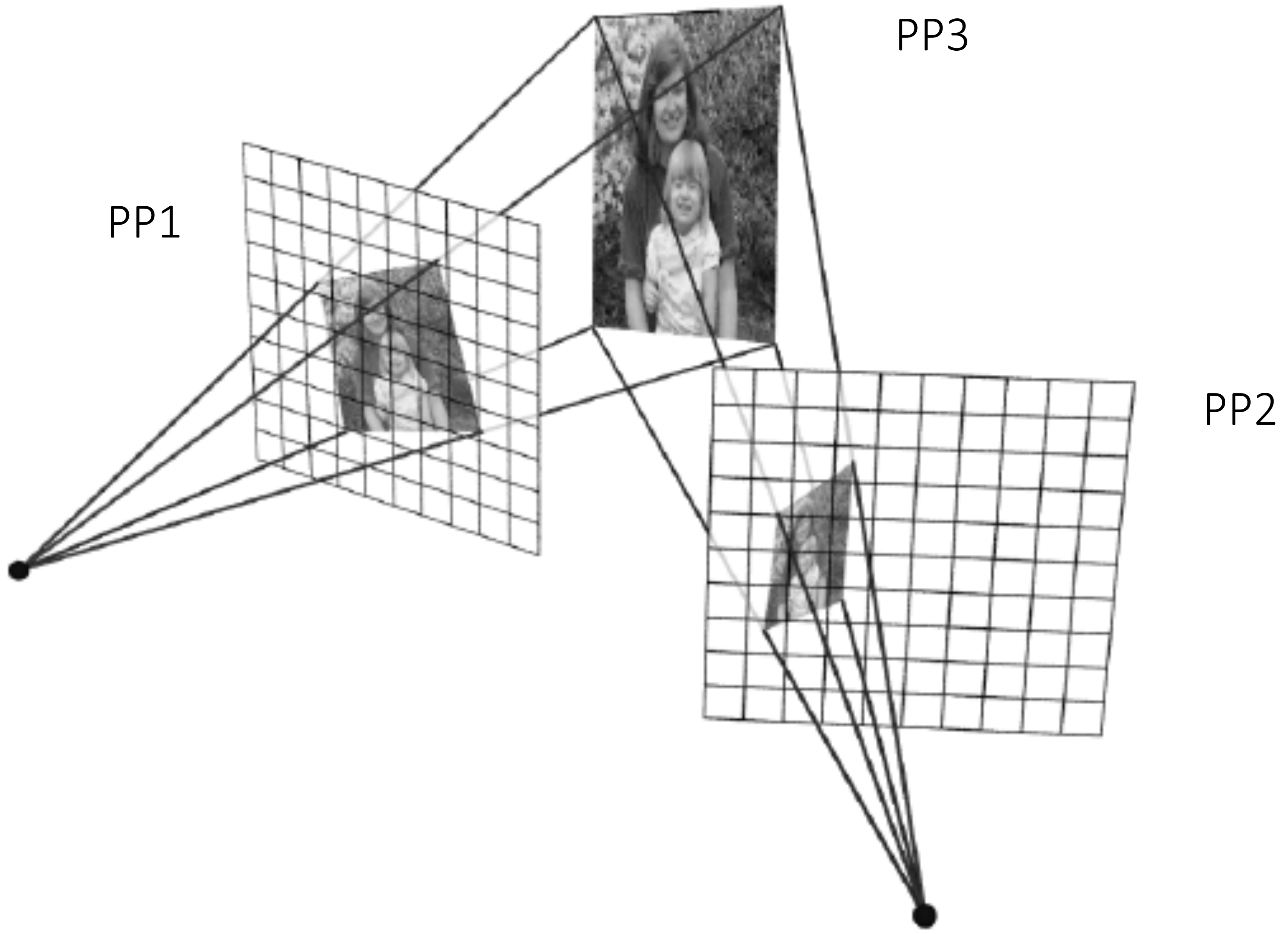
Name	Matrix	# D.O.F.
translation	$\begin{bmatrix} \boldsymbol{I} \mid \boldsymbol{t} \end{bmatrix}_{2\times3}$	2
rigid (Euclidean)	$\begin{bmatrix} \boldsymbol{R} \mid \boldsymbol{t} \end{bmatrix}_{2\times3}$	3
similarity	$\begin{bmatrix} s\boldsymbol{R} \mid \boldsymbol{t} \end{bmatrix}_{2\times3}$	4
affine	$\begin{bmatrix} \boldsymbol{A} \end{bmatrix}_{2\times3}$	6
projective	$\begin{bmatrix} \tilde{\boldsymbol{H}} \end{bmatrix}_{3\times3}$	8

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Aside: Classification of 2D Transformations

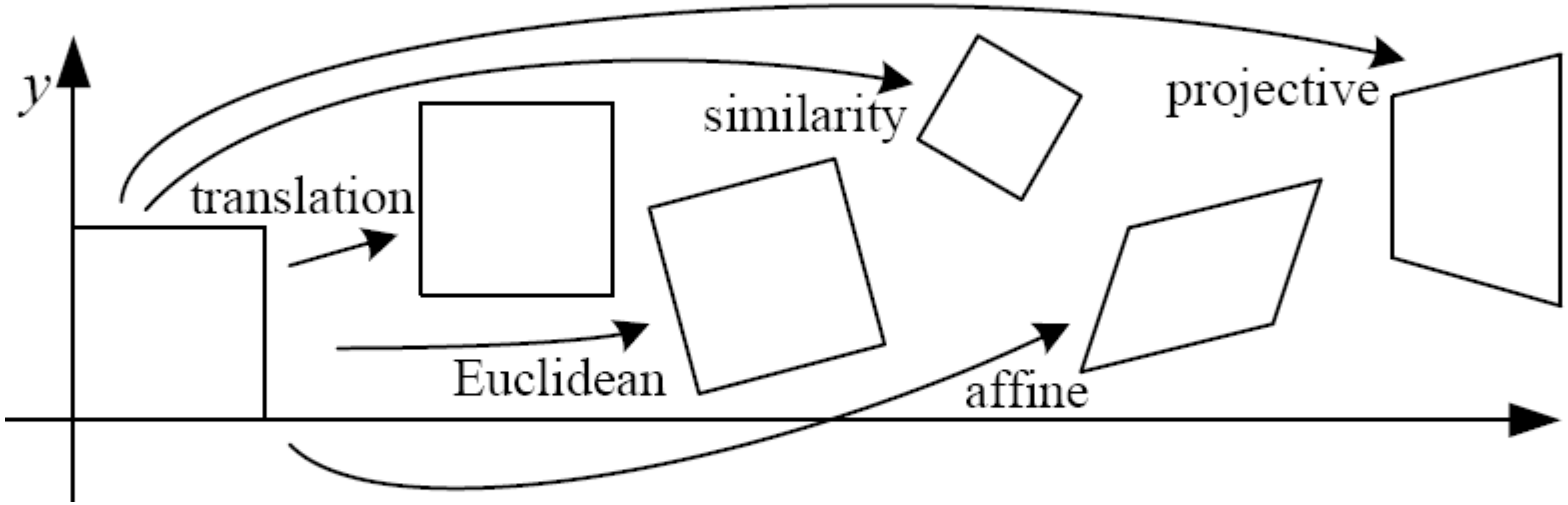


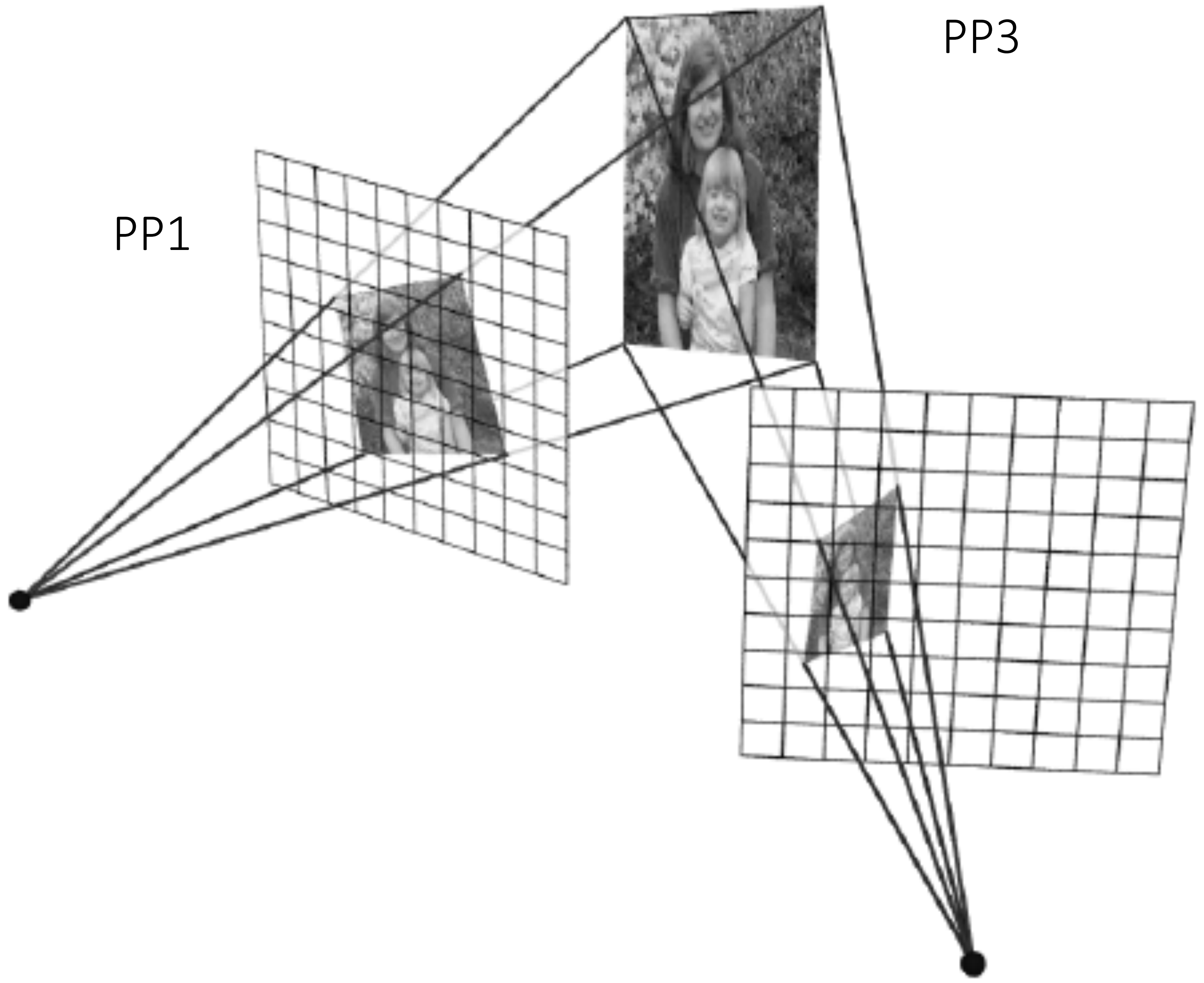
Which kind **transformation** is needed to warp projective plane 1 into projective plane 2?



Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Aside: Classification of 2D Transformations





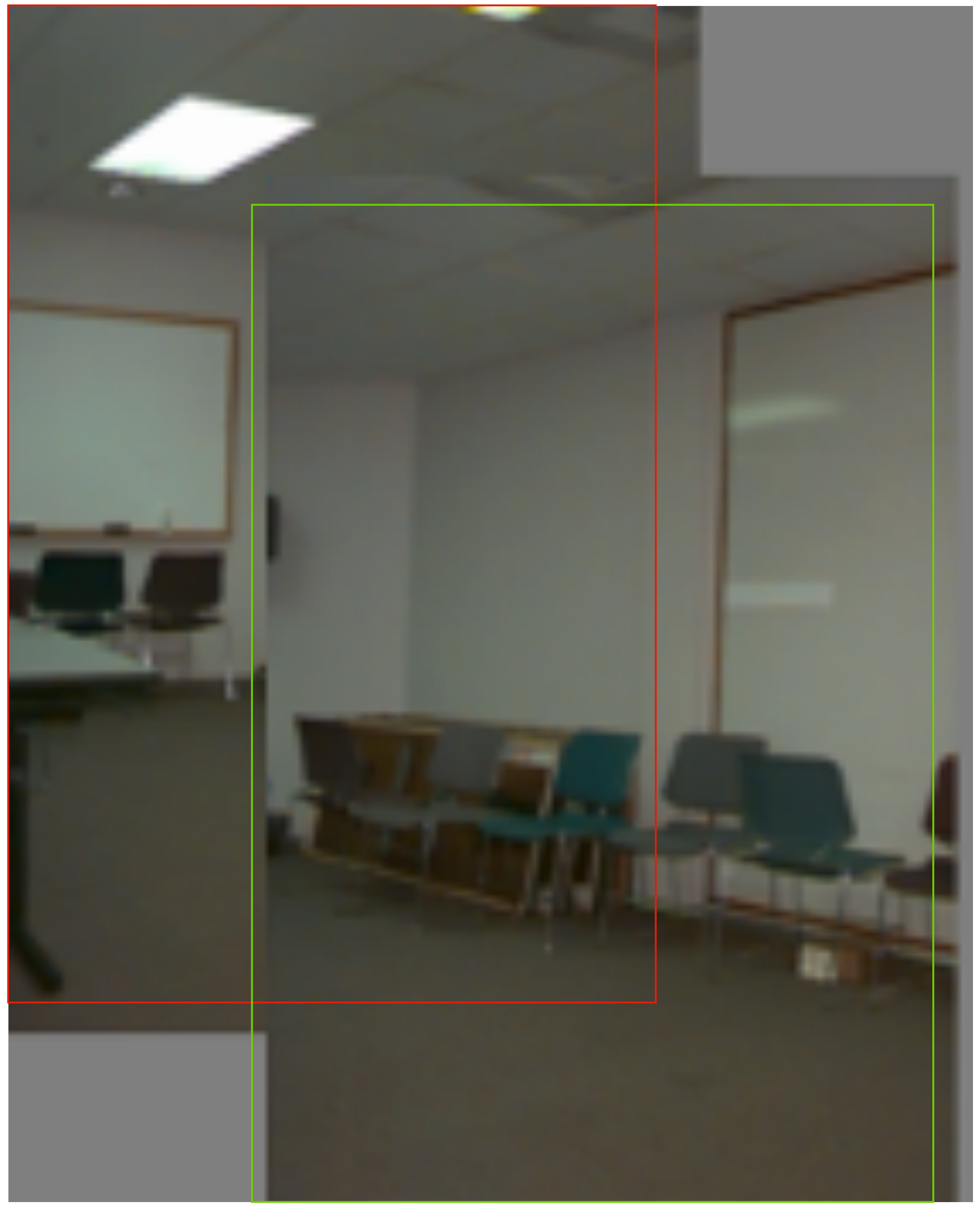
Which kind **transformation** is needed to warp projective plane 1 into projective plane 2?

— A **projective** transformation (a.k.a. a homography).

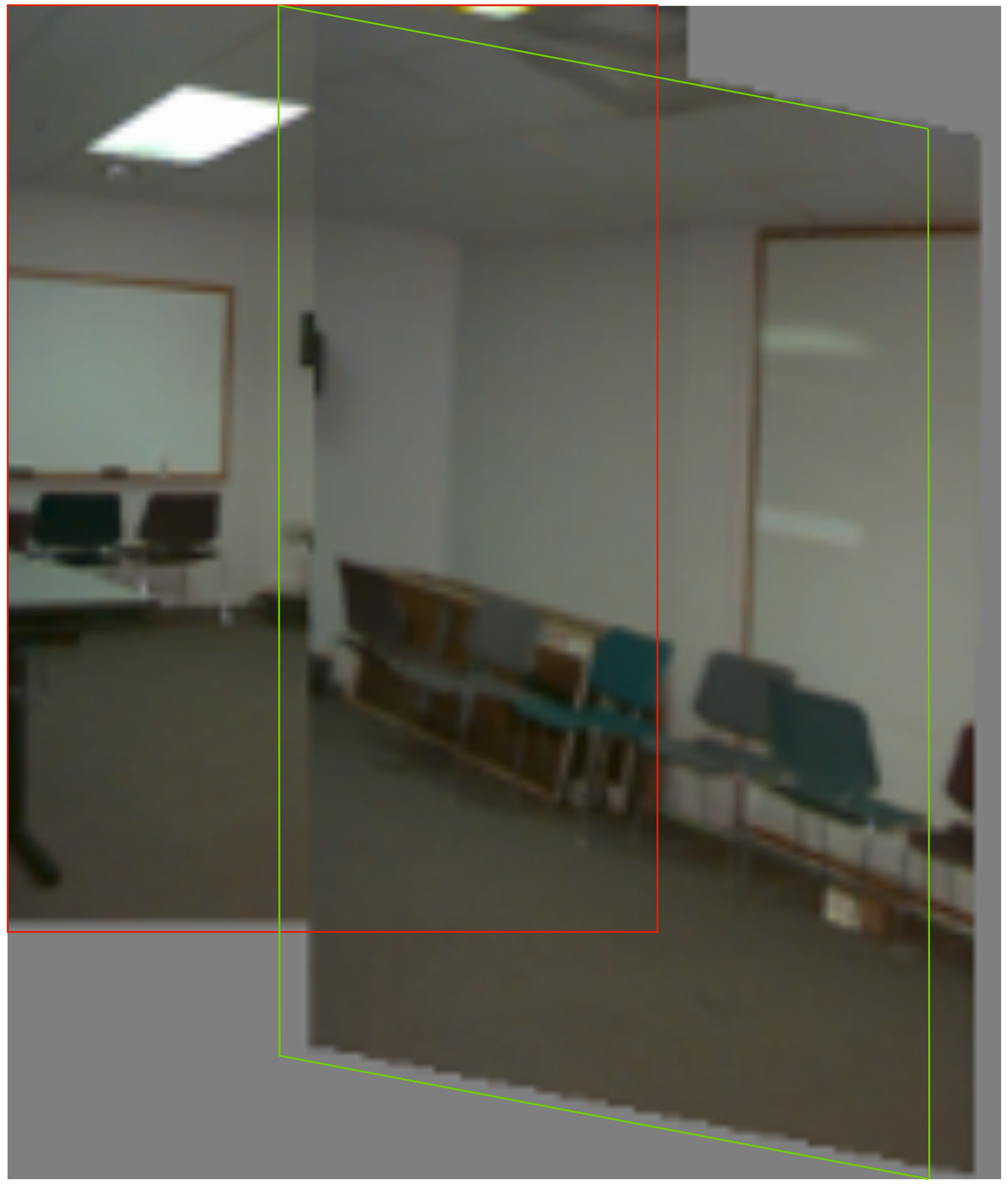
Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Aside: Warping with Different Transformations

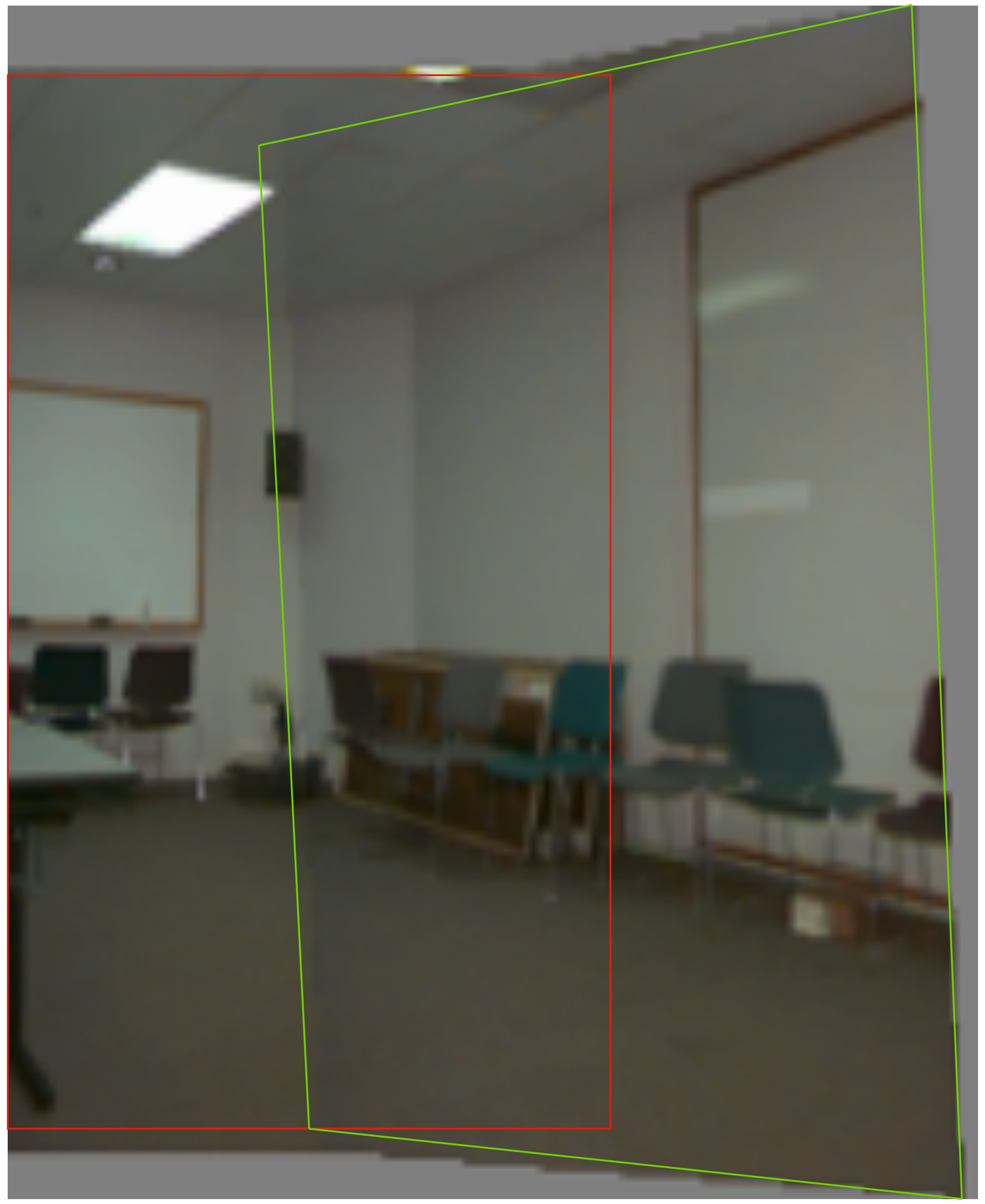
Translation



Affine



Projective
(homography)



Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Aside: We can use homographies when …

1.… the scene is planar; or

2.… the scene is very far or has small (relative) depth variation → scene is approximately planar

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Aside: We can use homographies when …

3.… the scene is captured under camera rotation only (no translation or pose change)

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Solution for **Affine** Parameters

Affine transform of $[x, y]$ to $[u, v]$

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}$$

Rewrite to solve for **transformation** parameters:

$$\begin{bmatrix} x_1 & y_1 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_1 & y_1 & 0 & 1 \\ x_2 & y_2 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_2 & y_2 & 0 & 1 \\ & & \dots & \dots & & \\ & & \dots & \dots & & \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ \dots \\ \dots \end{bmatrix}$$

(6 equations 6 unknowns)

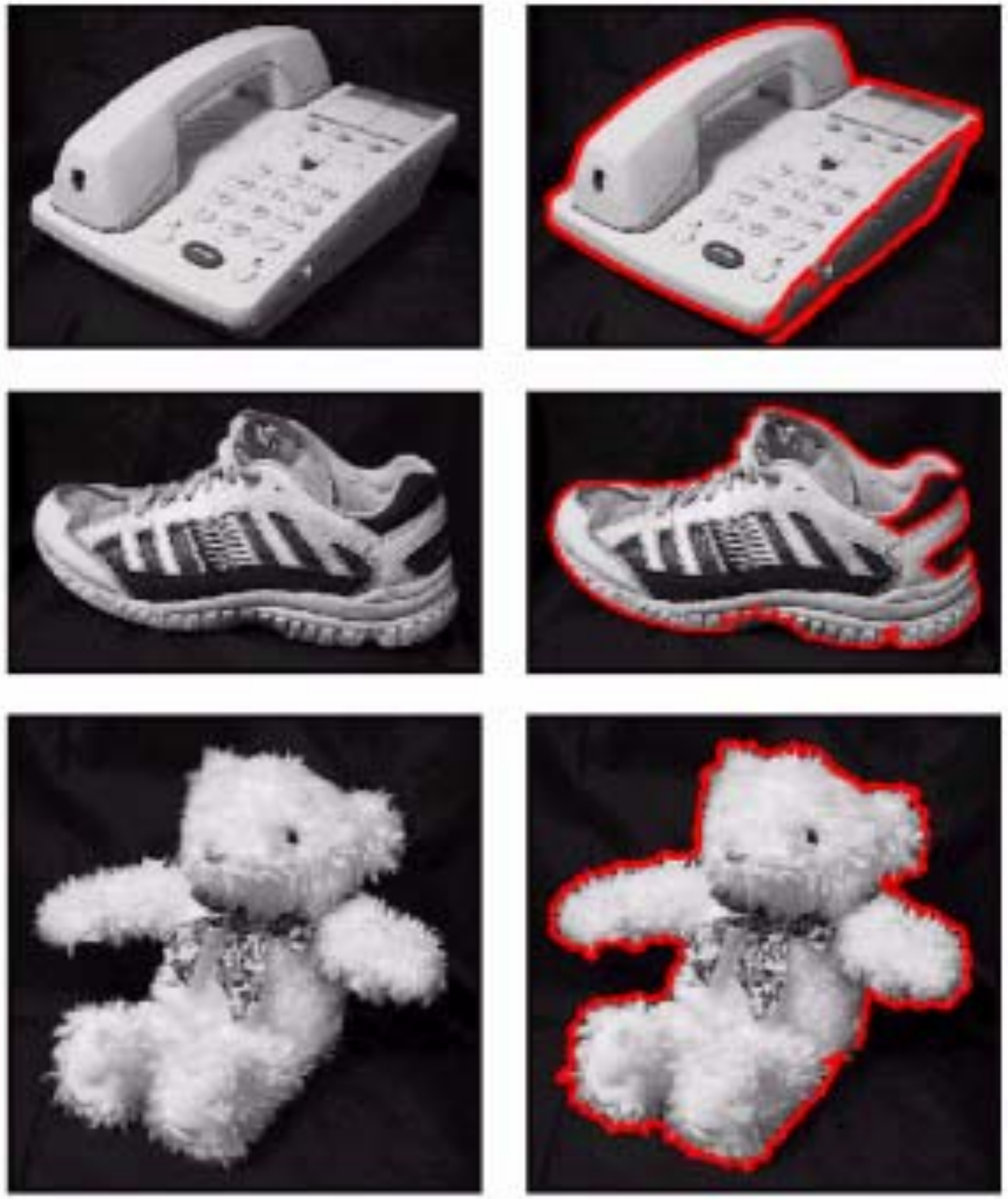
Solution for **Affine** Parameters

Suppose we have $k \geq 3$ matches, $[x_i, y_i]$ to $[u_i, v_i]$, $i = 1, 2, \cdots, k$

Then,

$$\begin{bmatrix} x_1 & y_1 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_1 & y_1 & 0 & 1 \\ x_2 & y_2 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_2 & y_2 & 0 & 1 \\ & & \cdots & \cdots & & \\ & & \cdots & \cdots & & \\ x_k & y_k & 0 & 0 & 1 & 0 \\ 0 & 0 & x_k & y_k & 0 & 1 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ \cdots \\ \cdots \\ u_k \\ v_k \end{bmatrix}$$

3D **Object Recognition**

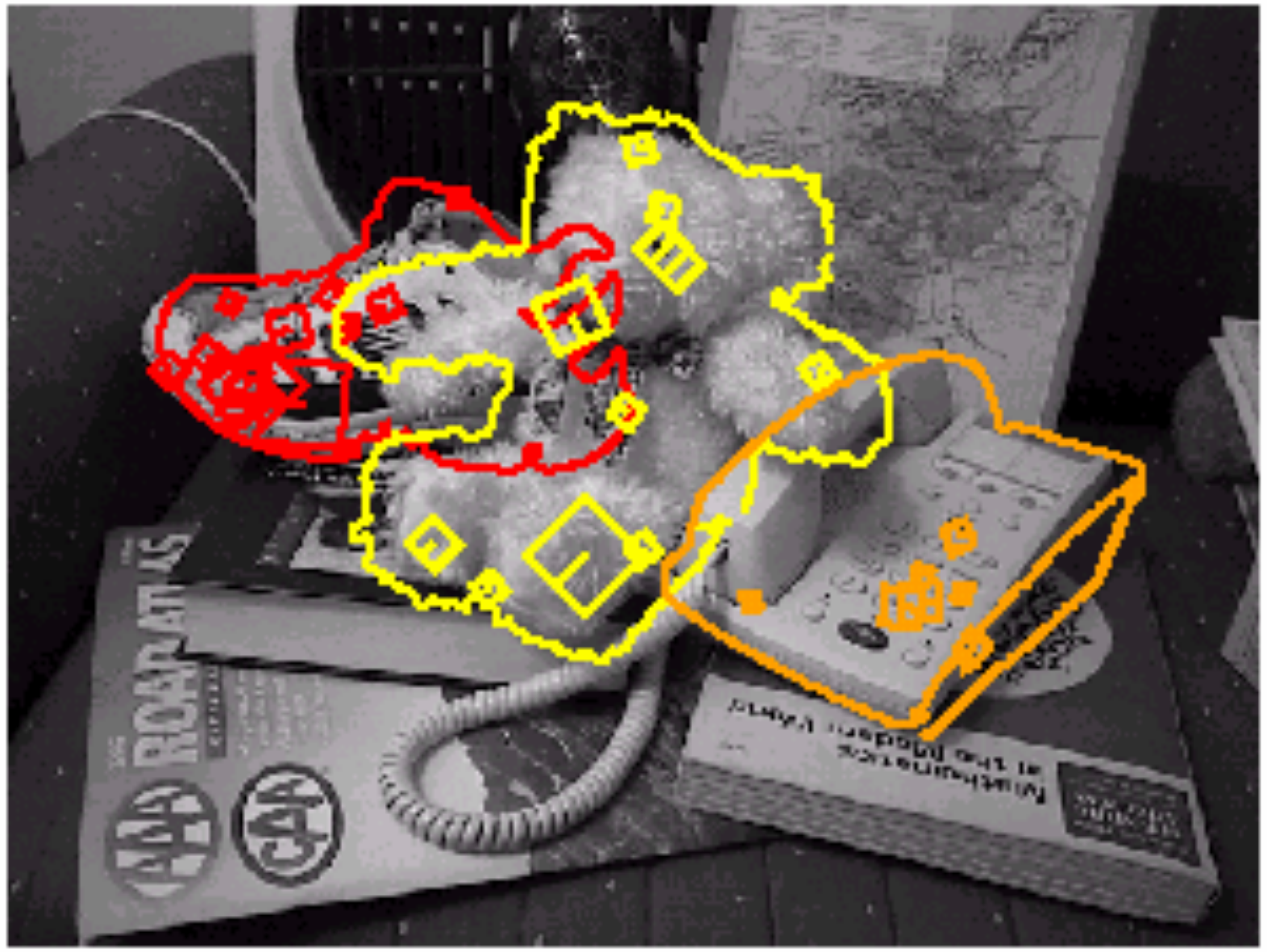


Extract outlines with background subtraction

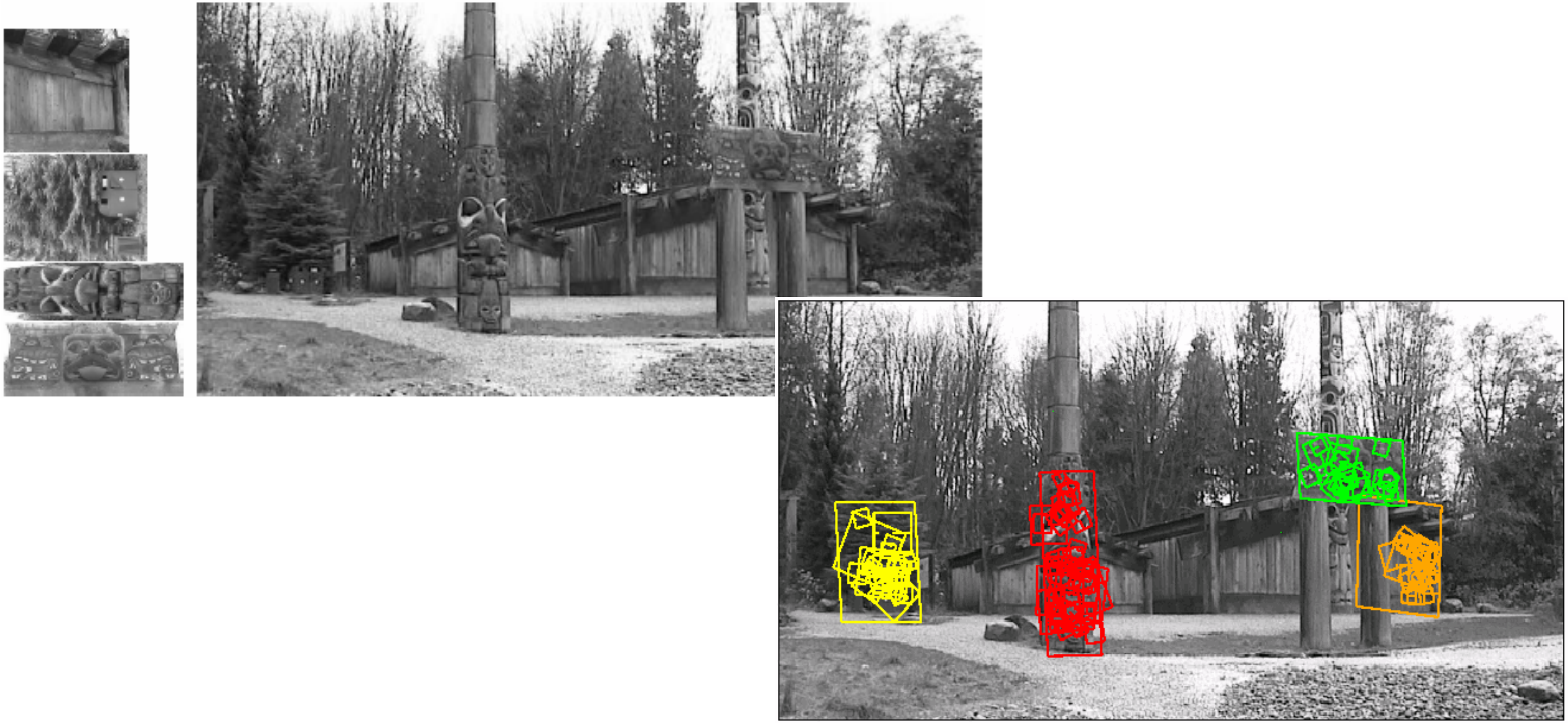
3D **Object Recognition**

Only 3 keys are needed for recognition, so extra keys provide robustness

Recognition Under **Occlusion**



Location Recognition



Example 1: Sony Aibo

SIFT Usage

— Recognize charging station

— Communicate with visual cards

Summary of Object Recognition with SIFT

Match each keypoint independently to database of known keypoints extracted from "training" examples

— use fast (approximate) nearest neighbour matching

— threshold based on ratio of distances to best and to second best match

Identify **clusters of (at least) 3 matches** that agree on an object and a similarity pose

— use generalized Hough transform

Check each cluster found by performing detailed geometric fit of affine transformation to the model

— accept/reject interpretation accordingly