CSE 252D: Advanced Computer Vision

Manmohan Chandraker

Lecture 1: Overview

Virtual classrooms

- Virtual lectures on Zoom
  - Only host shares the screen
  - Keep video off and microphone muted
  - But please do speak up (remember to unmute!)
  - Slides uploaded on webpage just before class
- Virtual interactions on Zoom
  - Ask and answer plenty of questions
  - “Raise hand” feature on Zoom when you wish to speak
  - Post questions on chat window
  - Happy to try other suggestions!
- Lectures recorded and upload on Kaltura
  - Available under “My Media” on Canvas

Enrollment logistics

- To enroll if you are on the waitlist
  - Send **"Request to enroll"** email to instructor if on waitlist
  - Include CV, courses, project experience relevant to computer vision
- While on the waitlist
  - You are welcome to attend lectures even if on waitlist
  - To limit TA workload, we can grade only enrolled students
  - Most should be able to enroll eventually
- Canvas
  - All enrolled and waitlisted students should have access
- All announcements will be posted on Piazza
  - Send email to TA (CC instructor) if cannot access Piazza

Course Details

Course details

- Each class will cover topics in computer vision
- Examples of topics
    - Correspondence
    - Optical flow
    - Structure from motion
    - Face recognition
    - Human pose estimation
    - Material and lighting
    - Semantic segmentation
    - Object detection
    - Action recognition
    - Domain adaptation

Course details

- Topics structured into a few modules
  - Background
  - Structure and Motion
  - Faces and Humans
  - Objects and Stuff
  - Material and Lighting
  - Bias and Adaptation

Course details

- "Lightning" presentations
  - Provide a broad view of the field
  - An important skill to digest and present literature
  - Four students to present in one class
  - Papers to be assigned by instructor
  - Order of presentation: alphabetic (Googledoc will be posted)
- Send recorded presentation video 3 days before class
  - Will share PPT template
  - Well-practiced and fluent presentation
  - Incorporate feedback from instructor or TA
  - Include question to class
  - Ask and answer questions after presentation

Course details

- Presentation format:
    1. Motivation and problem description
    2. Prior work
    3. Method overview
    4. Method analysis
    5. Experiments
    6. Future work and discussion

Course details

- Class webpage:
  - http://cseweb.ucsd.edu/~mkchandraker/classes/CSE252D/Spring2021/
- Instructor email: mkchandraker@eng.ucsd.edu
- TA: Yu-Ying Yeh (Email: yuyeh@eng.ucsd.edu)
- Grading
  - 10% presentation
  - 60% assignments
  - 30% final exam
  - Ungraded quizzes
- Aim is to learn together, discuss and have fun!

Overall goals for the course

- Introduce the state-of-the-art in computer vision
- Study principles that make them possible
- Get understanding of tools that drive computer vision
- Enable one or all of several such outcomes
  - Pursue higher studies in computer vision
  - Join industry to do cutting-edge work in computer vision
  - Gain appreciation of modern computer vision technologies
- This is a great time to study computer vision!

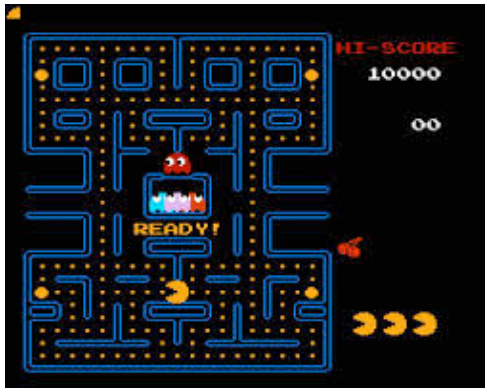
Deep Neural Networks

Deep learning is revolutionizing AI

Tic-tac-toe (1952)

Checkers (1994)

Chess (1997)



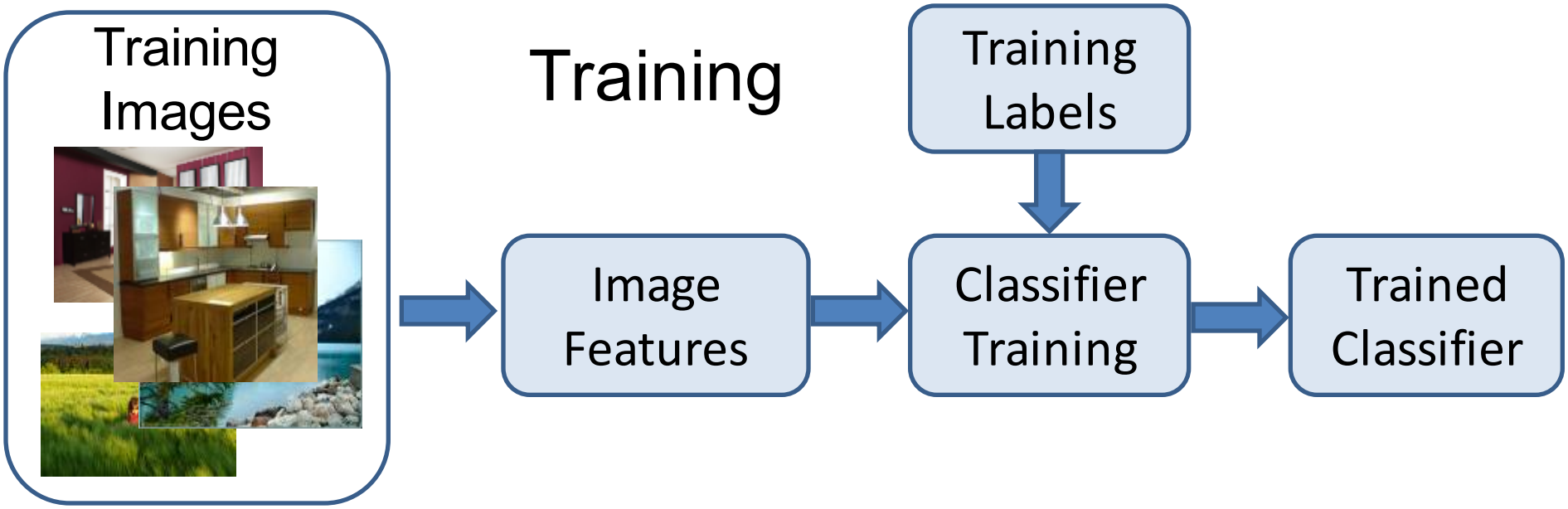
Atari (2015)

Go (2016)

Computer vision is also riding the wave

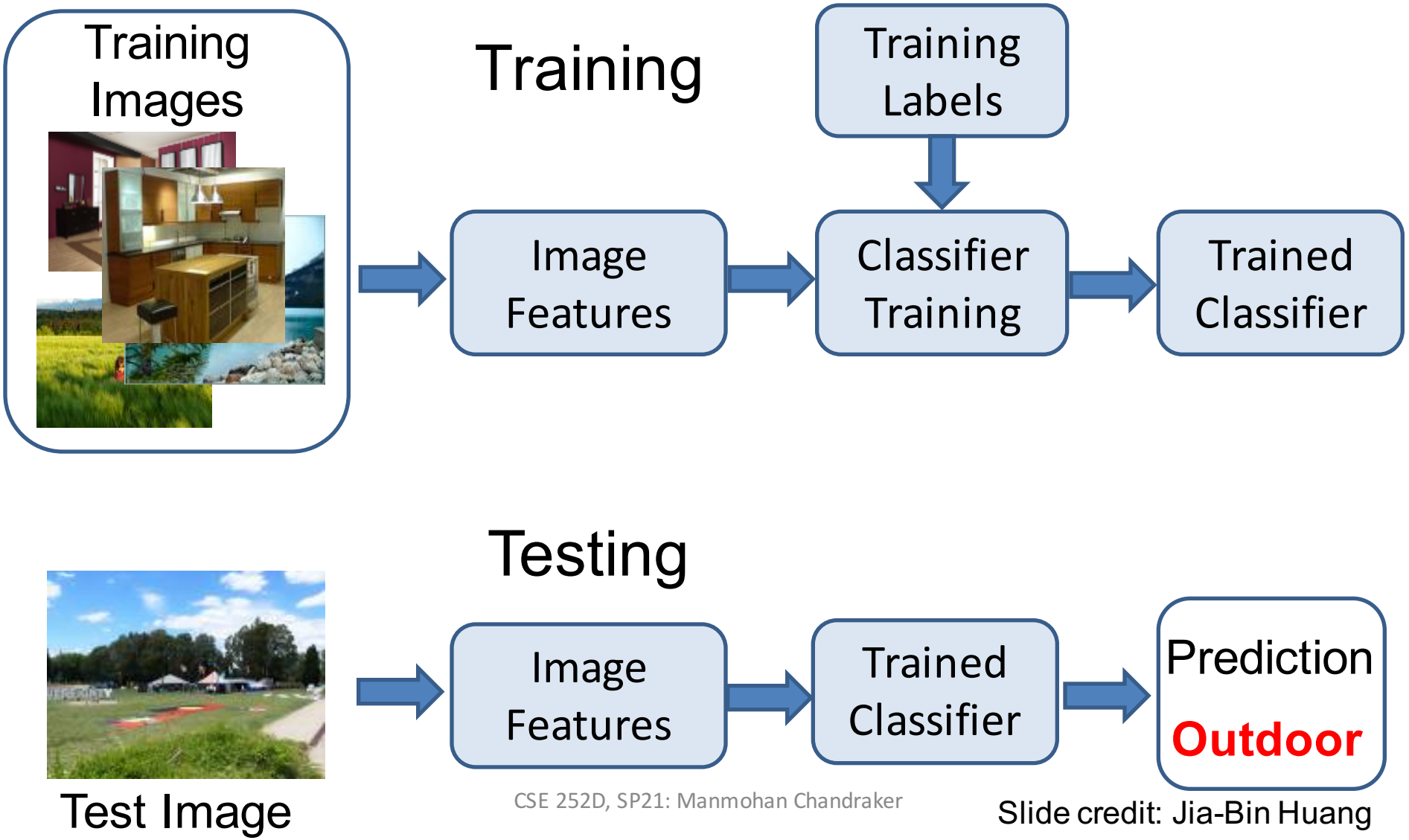
- Autonomous driving (Google, Tesla, Mobileye, ...)
- Augmented reality (HoloLens, Oculus, MagicLeap)
- Social networks (Google, Facebook, ...)
- Mobile applications
- Surveillance

Traditional Image Categorization: Training phase



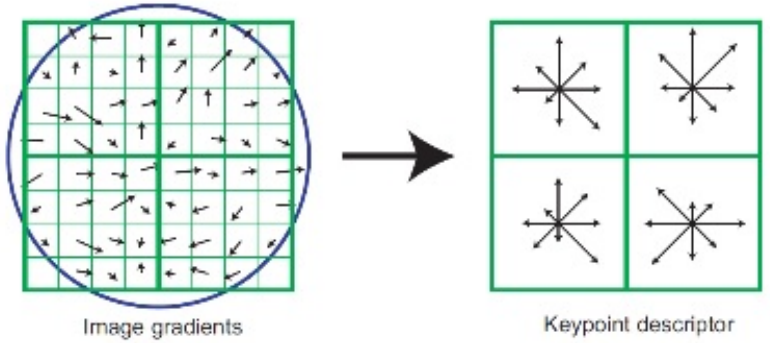
Slide credit: Jia-Bin Huang

Traditional Image Categorization: Testing phase

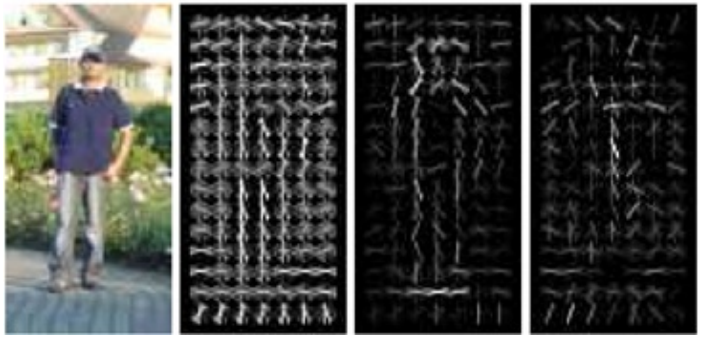


Slide credit: Jia-Bin Huang

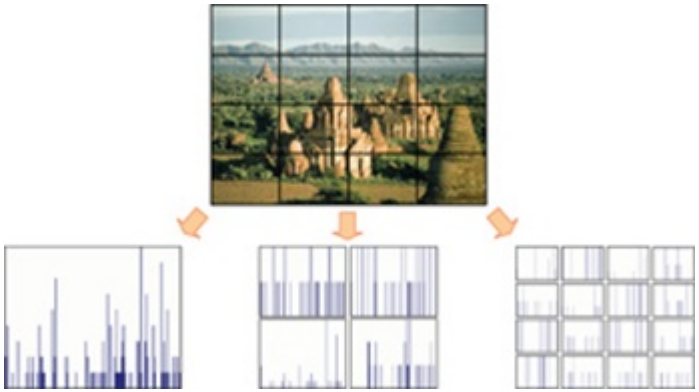
Features have been key



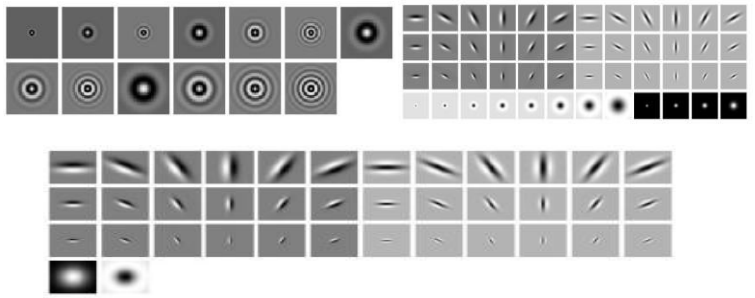
SIFT [Lowe IJCV 04]



HOG [Dalal and Triggs CVPR 05]



SPM [Lazebnik et al. CVPR 06]

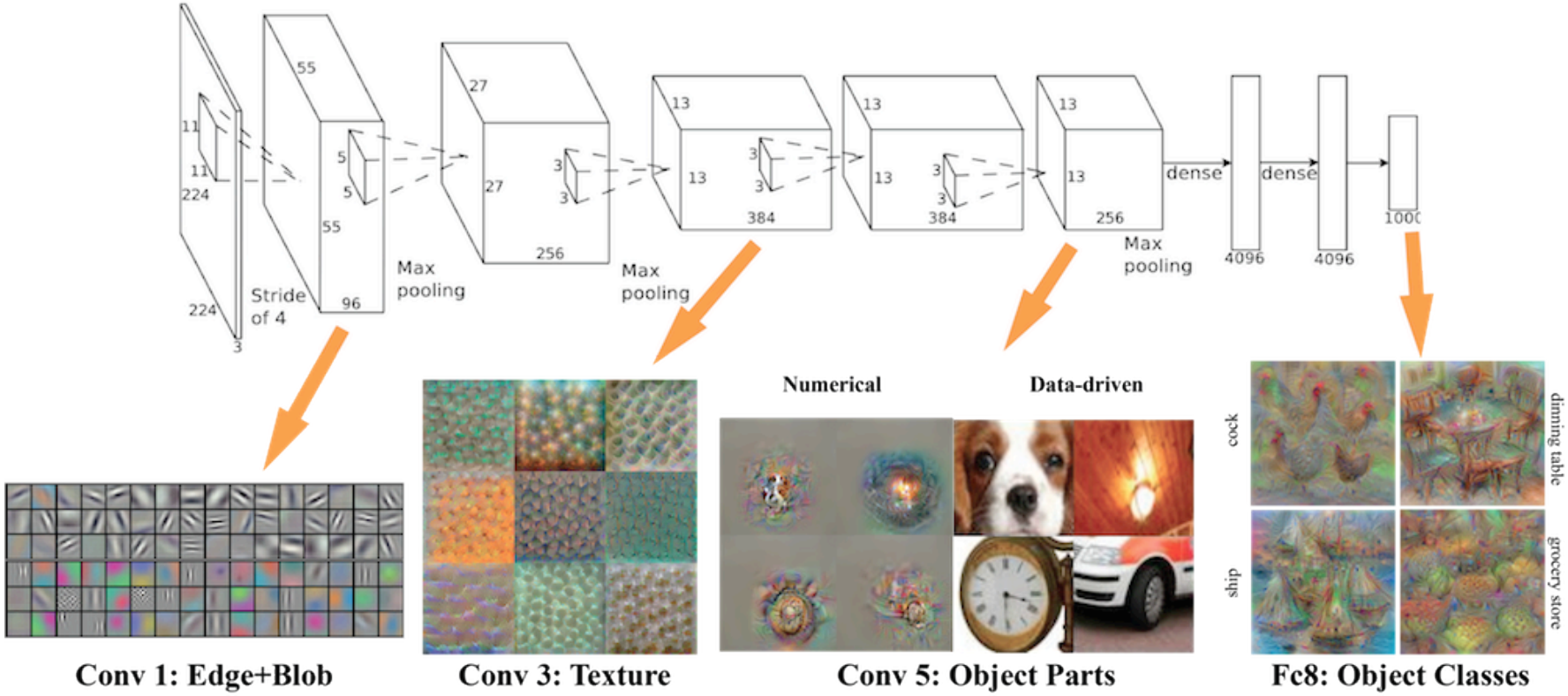


Textons

and many others:

SURF, MSER, LBP, GLOH, ….

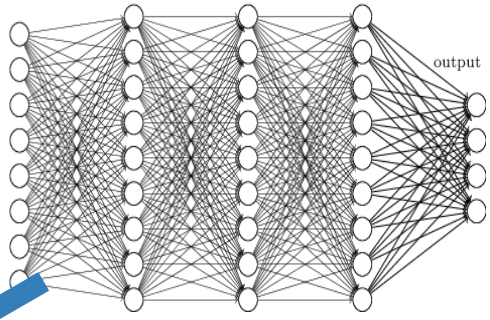
Learning a Hierarchy of Feature Extractors



- Hierarchical and expressive feature representations
- Trained end-to-end, rather than hand-crafted for each task
- Remarkable in transferring knowledge across tasks

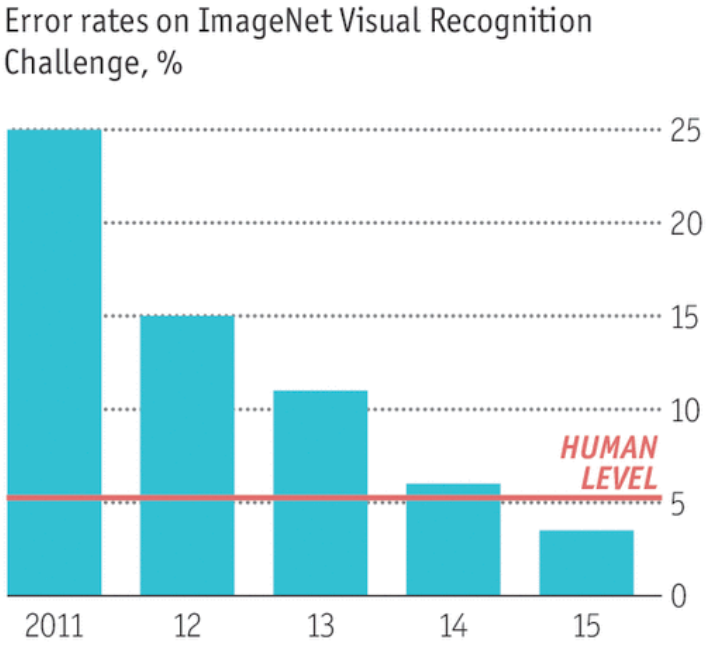
Significant recent impact on the field

Big labeled datasets



Deep learning

GPU technology



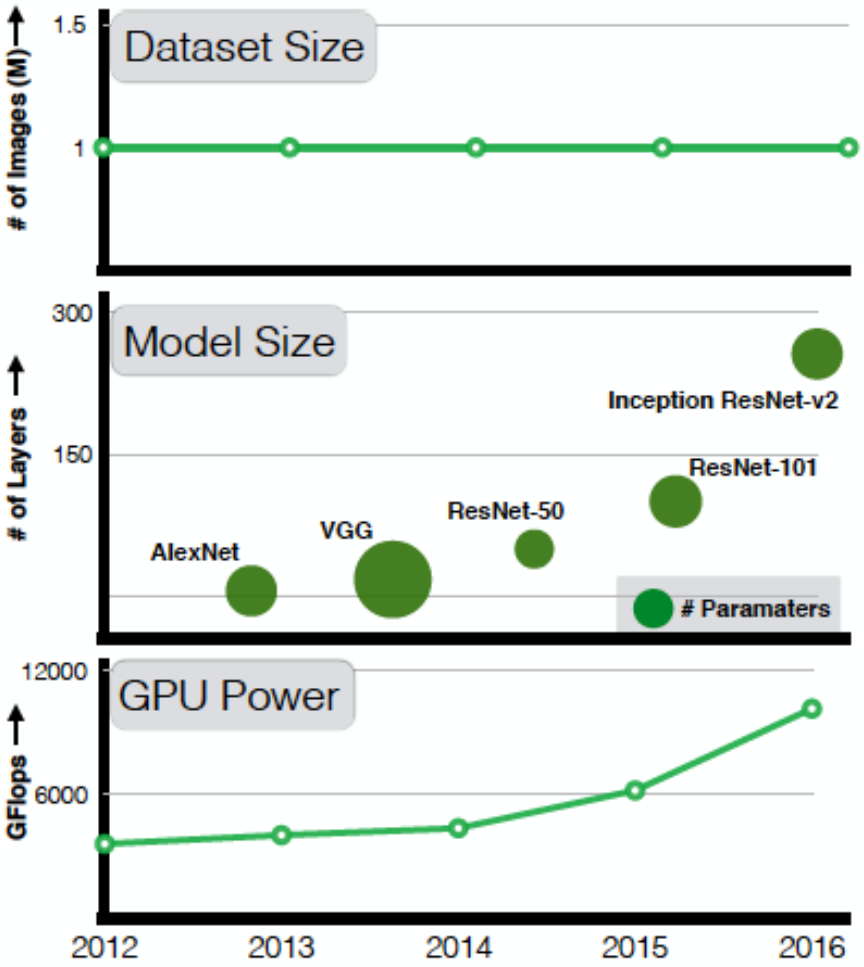
Deep learning has opened new areas

- Availability of large-scale image and video data
- Availability of computational power
  - Better and cheaper GPUs
  - Cloud computing resources
- Better understanding of how to train deep neural networks
- Advantages available for many areas of computer vision
  - Recognize objects across shape and appearance variations
  - Data-driven priors for 3D reconstruction
  - Predict long-term future behaviors in complex scenes
  - End-to-end training rather than expensive feature design.

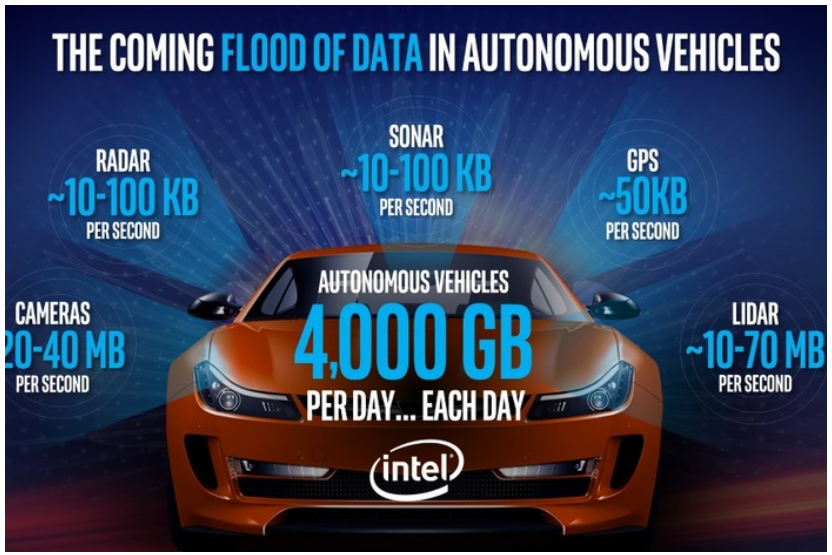
Limits of deep learning

- Deep networks are powerful
- But we must be aware of their limitations

Data: hardware and models scale more than labels

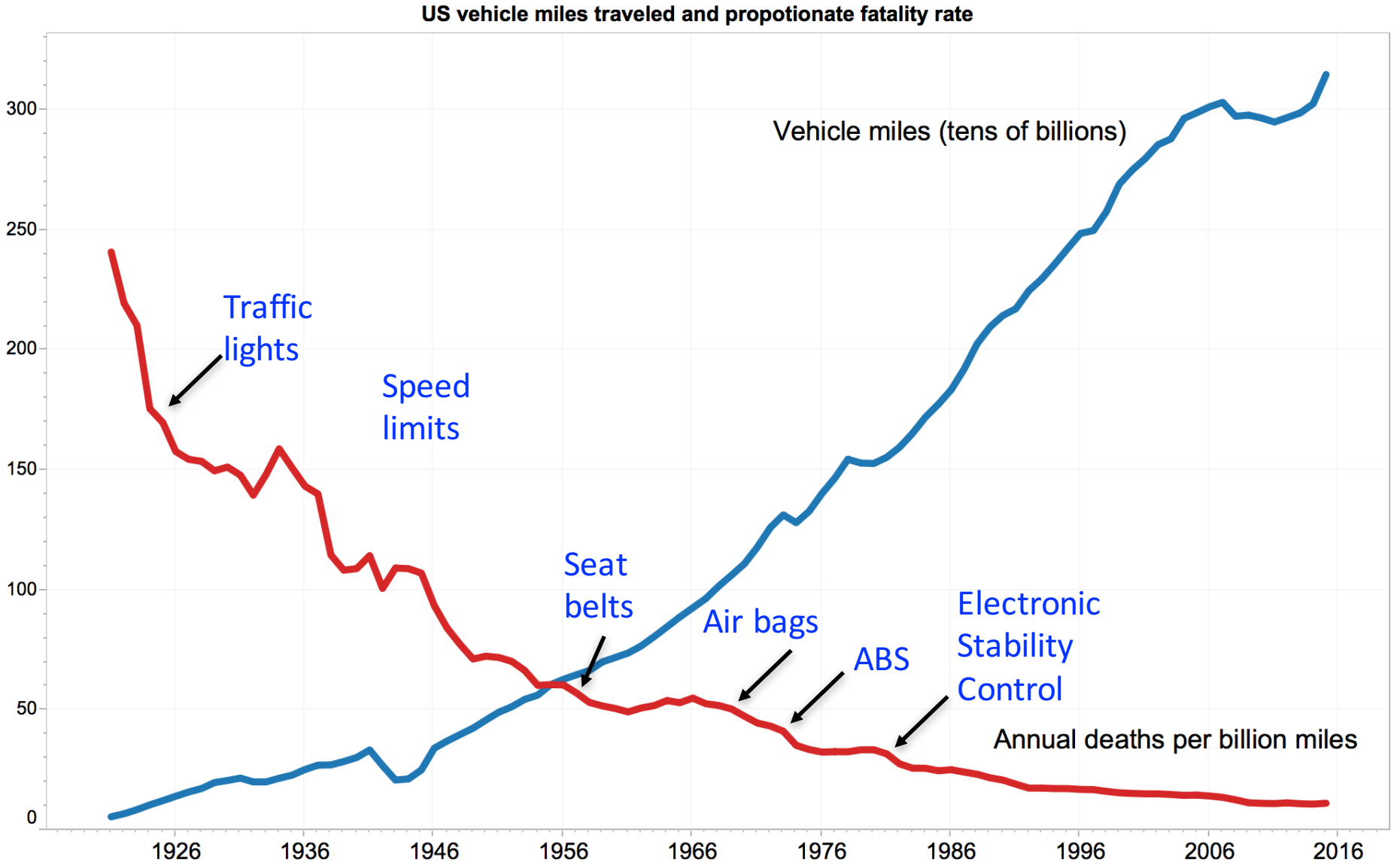


[Sun et al. 2017]

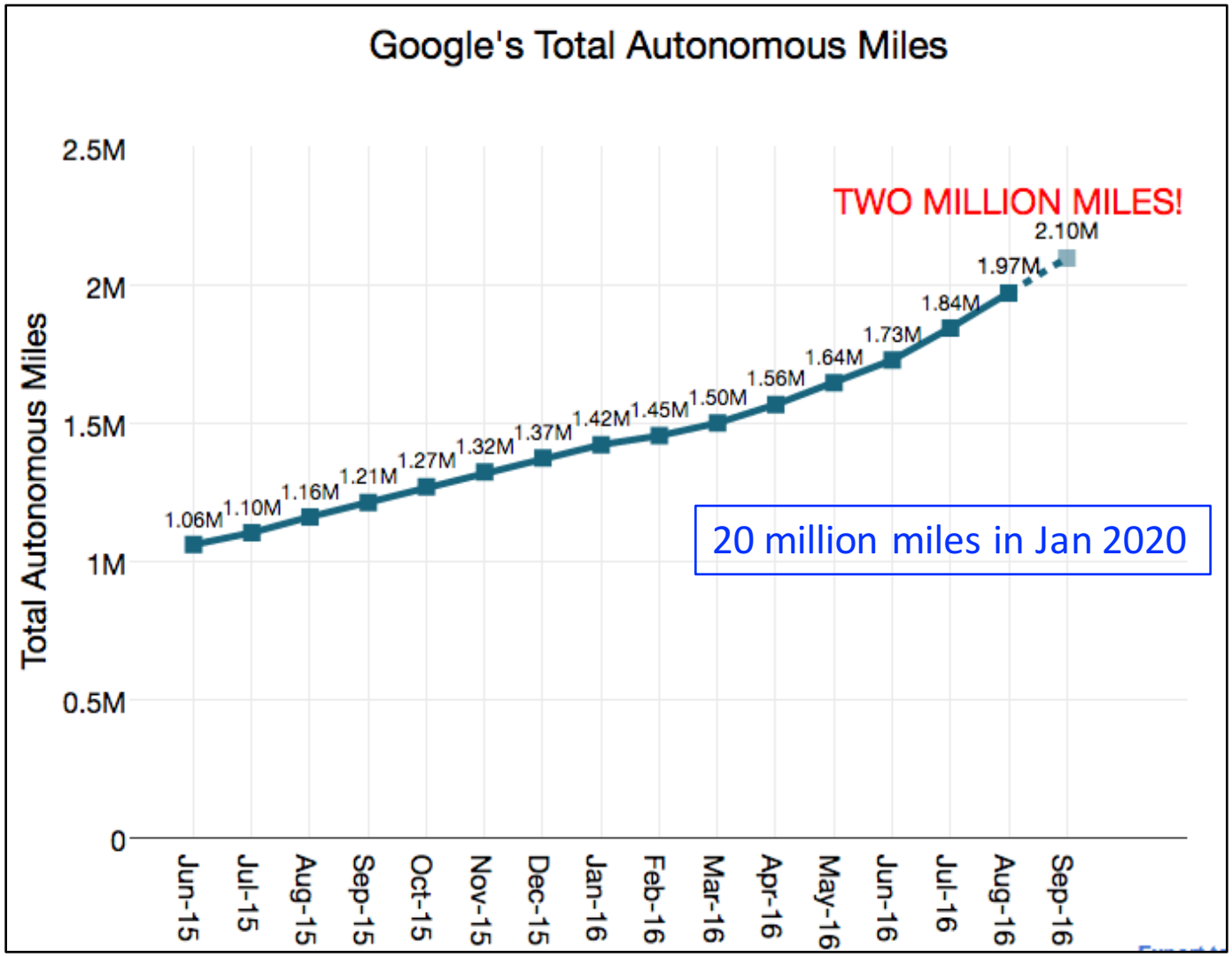


- More data helps
- 4 TB of data per day from a car
- Training effort
- Rare events matter more
- Purely supervised methods not scalable

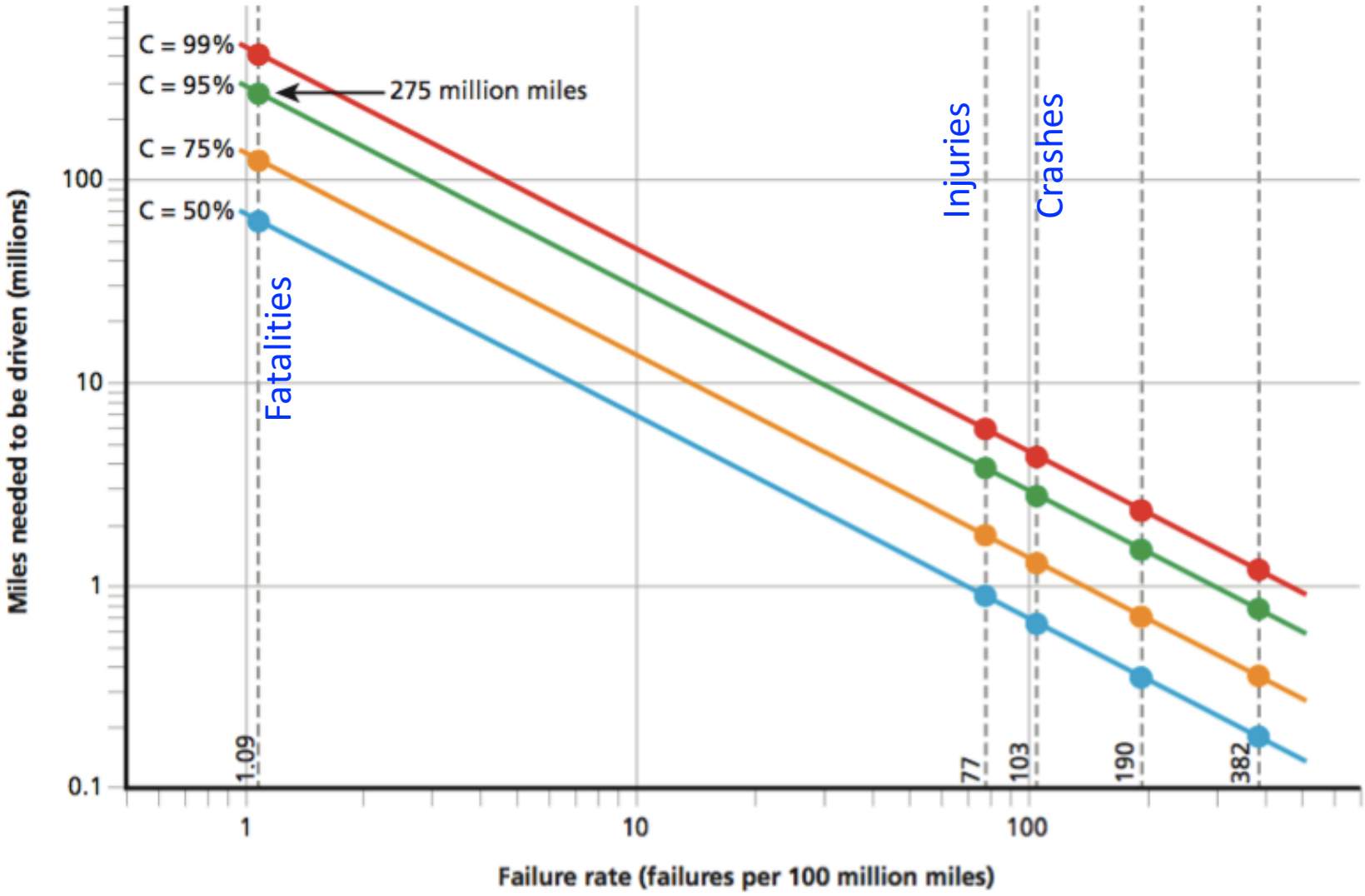
Miles to go before ....



Miles to go before ....



Miles to go before ....



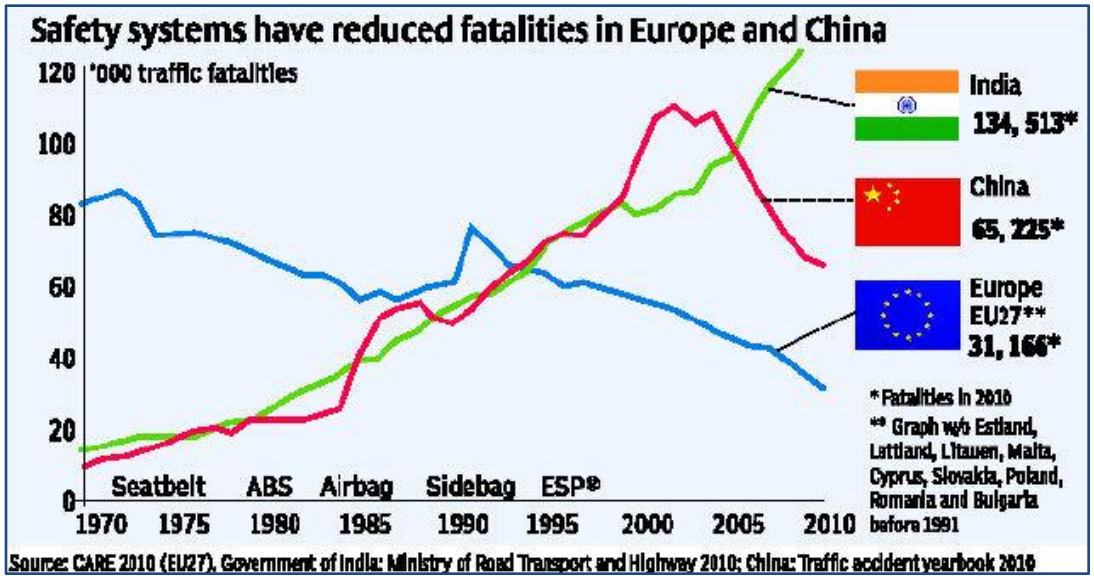
Miles to go before ....

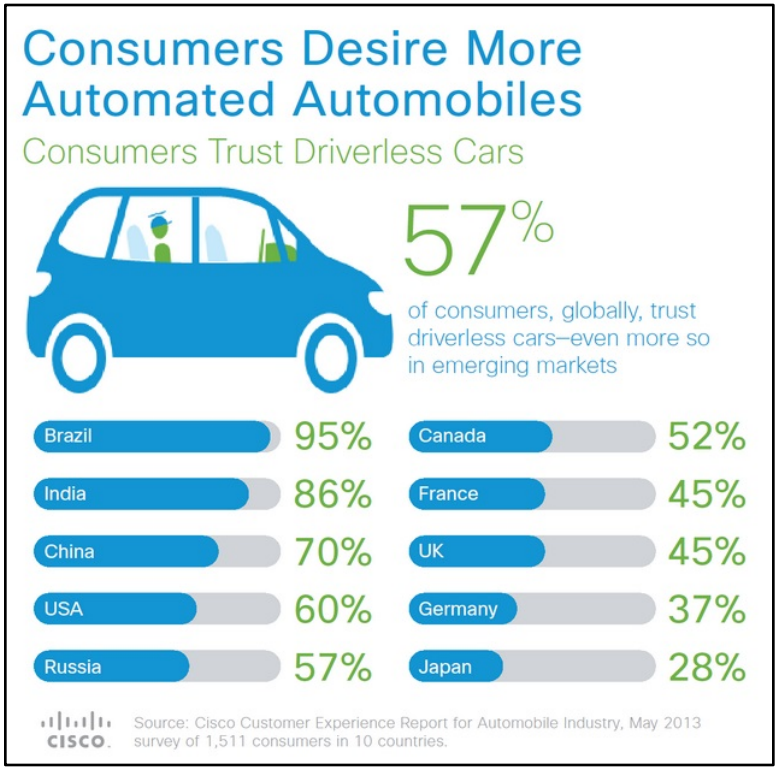
Object detection for an auto rickshaw

Getting self-driving to where it is needed the most

Number of people who will die on Indian roads:

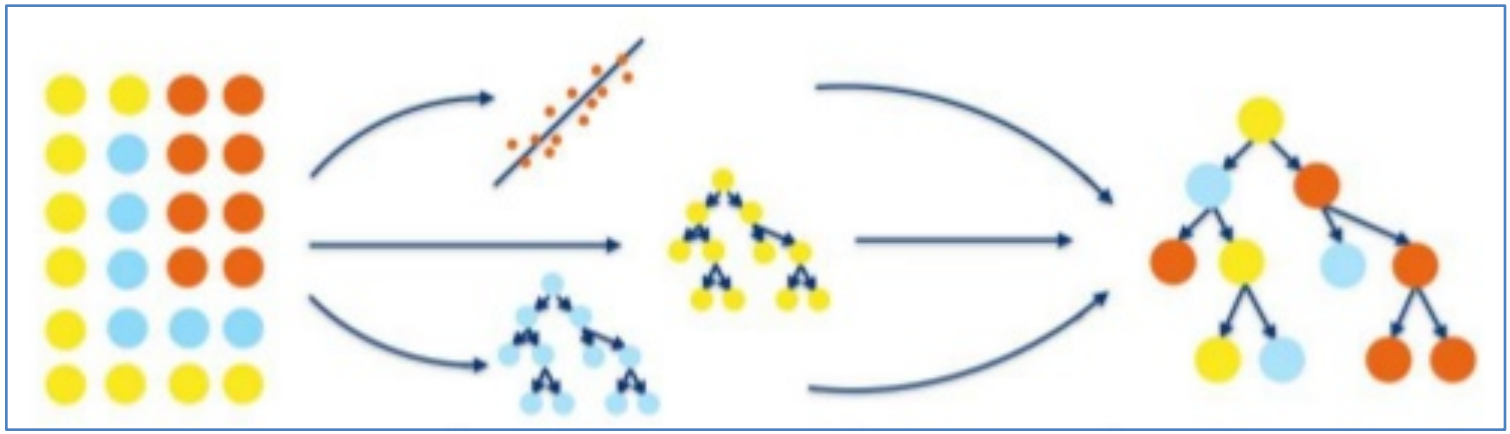
- 20 : During this lecture
- 400 : By this time tomorrow
- 145000 : By next year



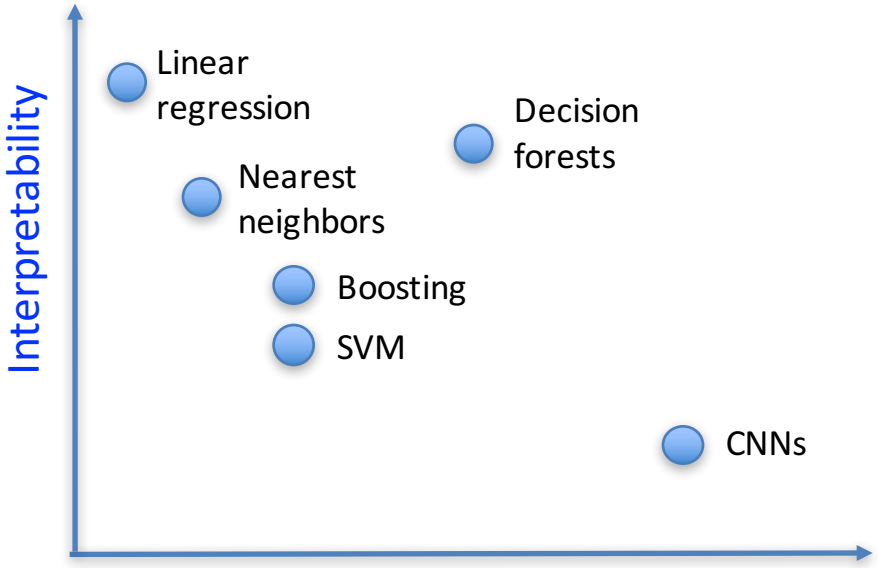


Interpretability of outputs

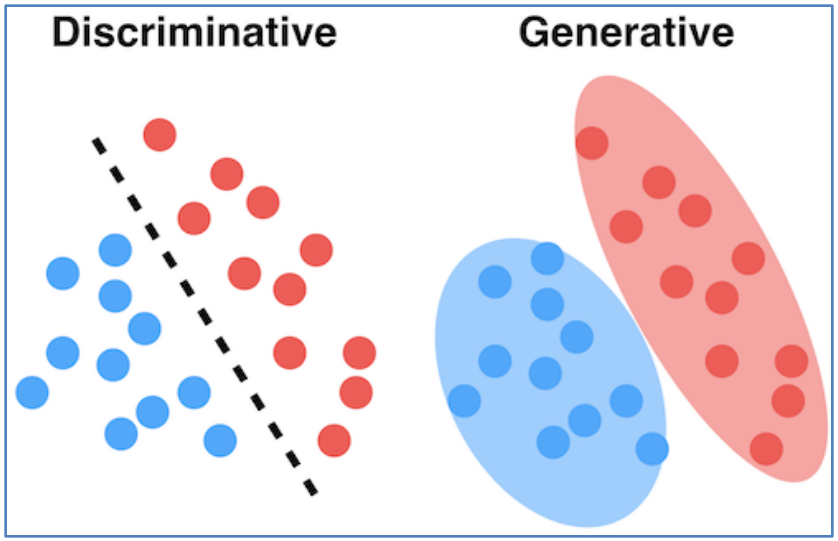
Automobile industry wants models built by combining validated components



Trade-offs for various learning approaches

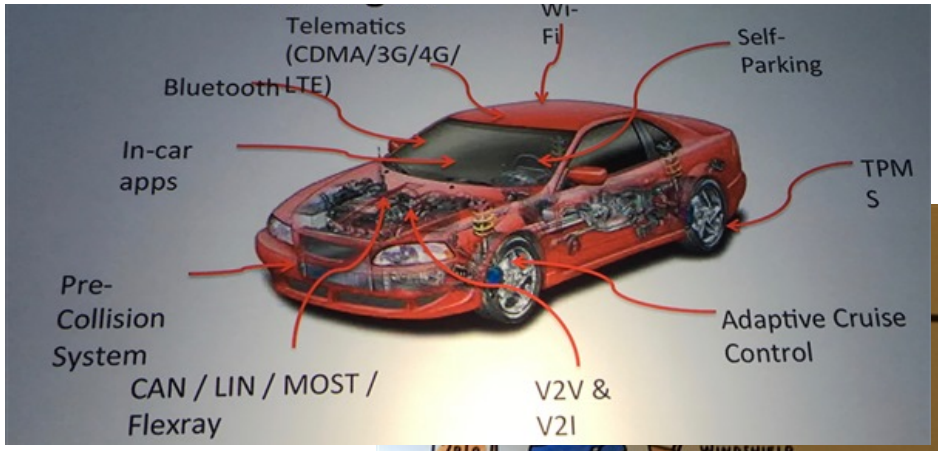


Generative and discriminative methods

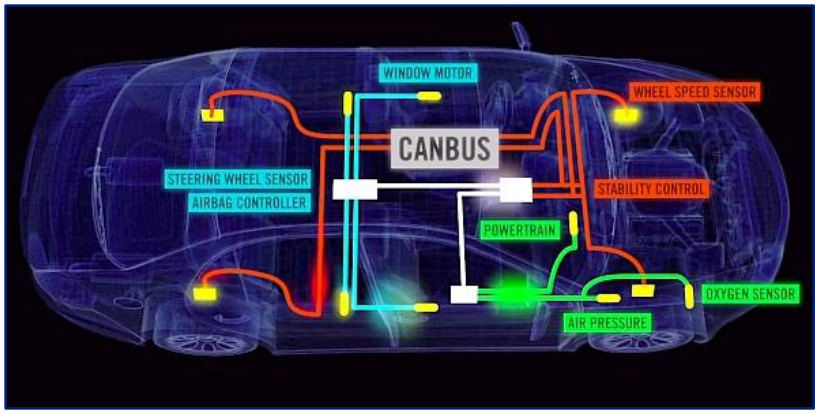


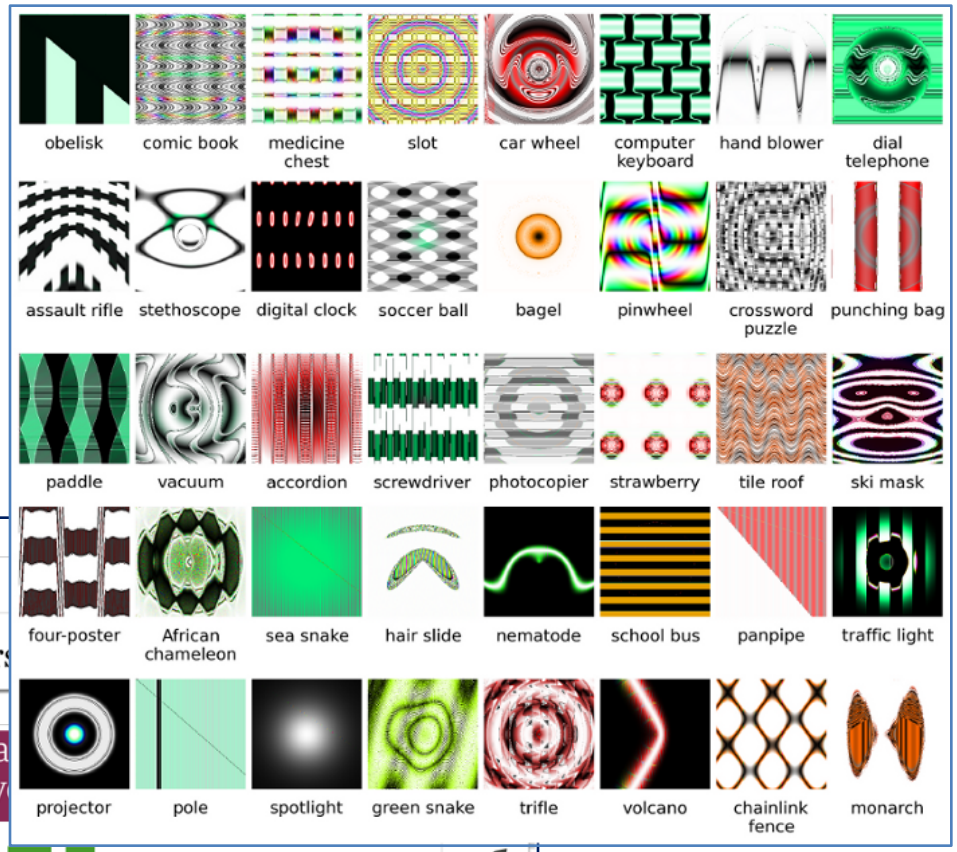
Security: all features are potential targets

Automobile manufacturer view



Security: algorithms may face adversarial attacks





Limits of deep learning

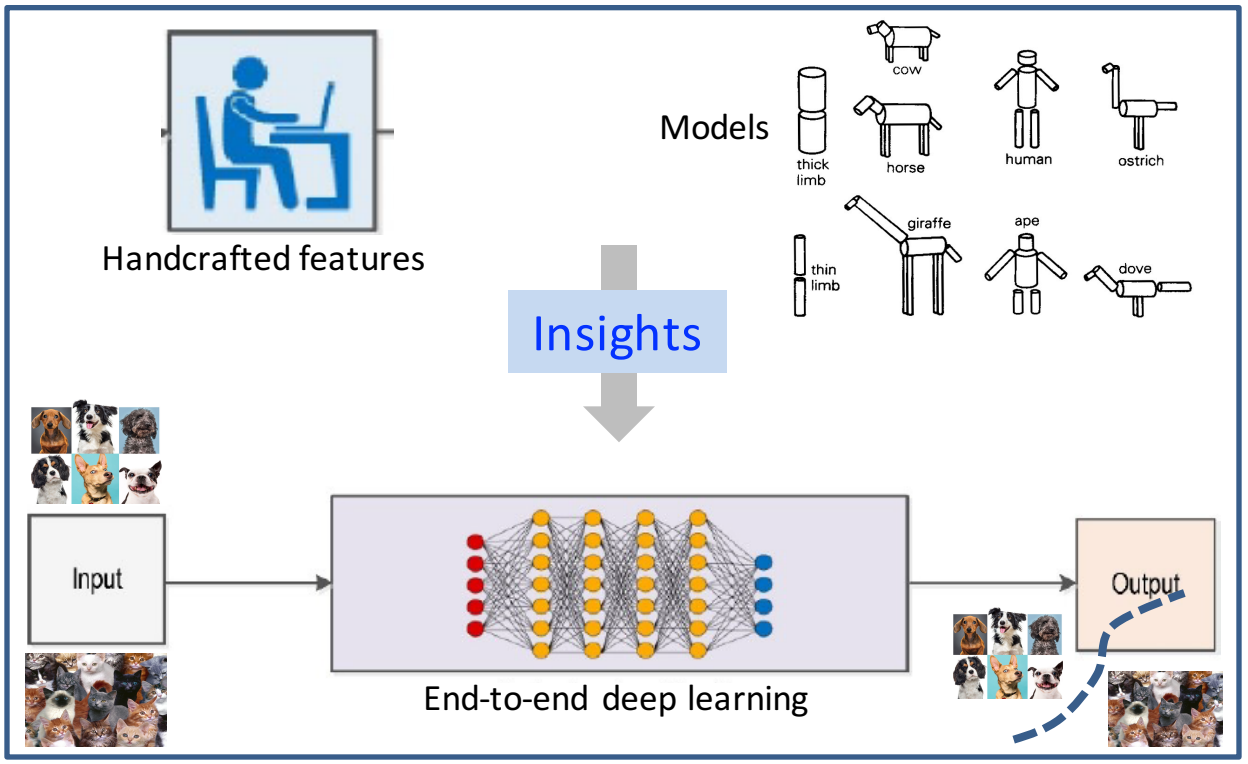
- Large scale labeled data is not always available
- Lack of generalization to unseen domains
- Good at narrow “classification”, not at broad “reasoning”
- Lack of interpretability
- Lack of reliability, security or privacy guarantees

New approaches to overcome limits

- Weak supervision
- Semi-supervision
- Self-supervision
- Domain adaptation
- Physical modeling
- Privacy-preservation

Take-home message

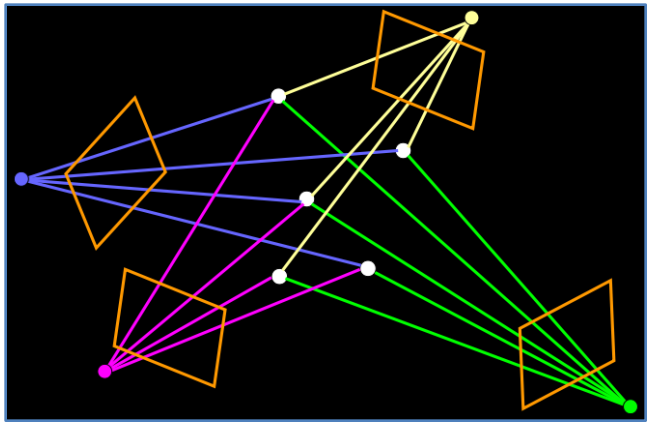
- Powerful new tools allow rapid progress
- Must gain insights and understand limitations
- Understand the old to inform the new

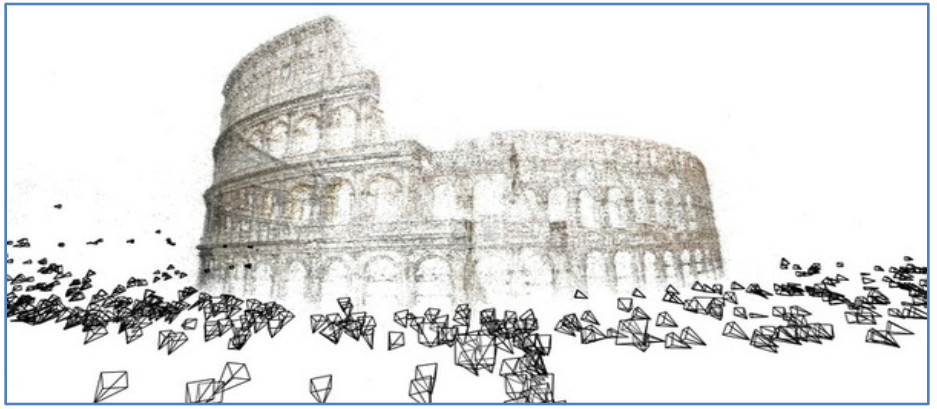


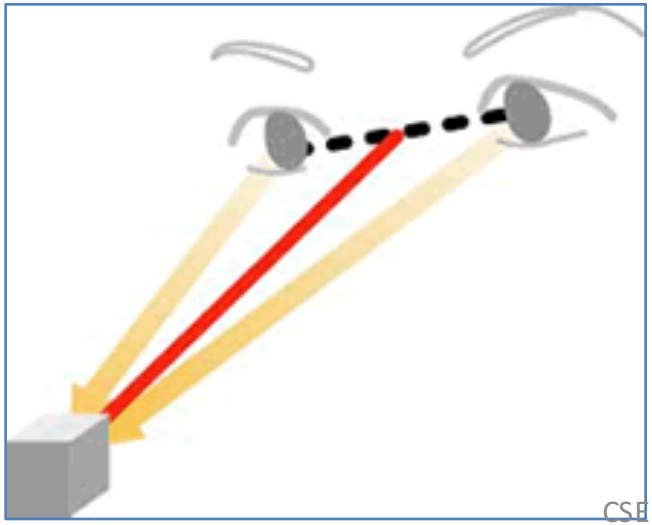
Preview
Module 1: Structure and Motion

Correspondence

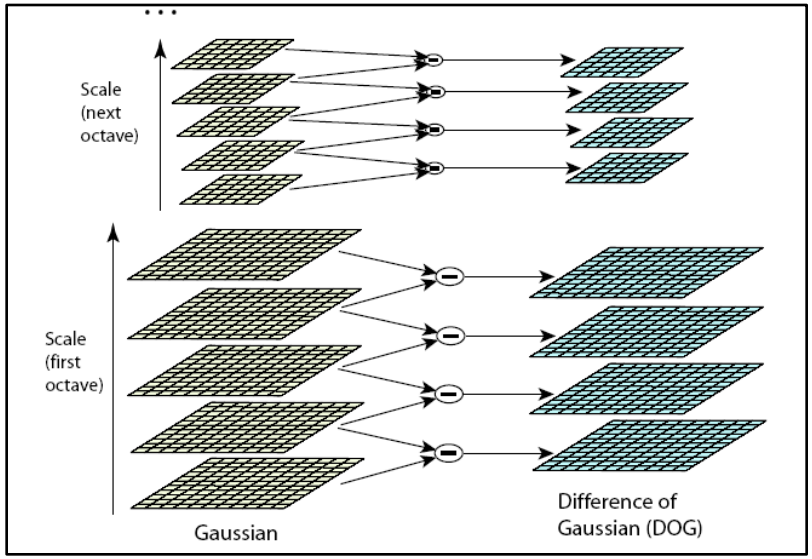
Relate projections of the same point in two or more images of the scene.





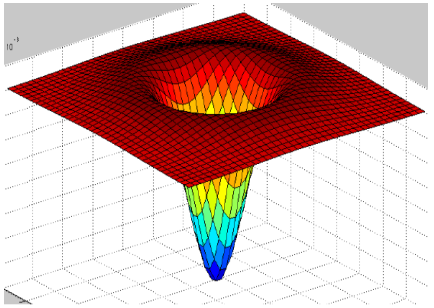


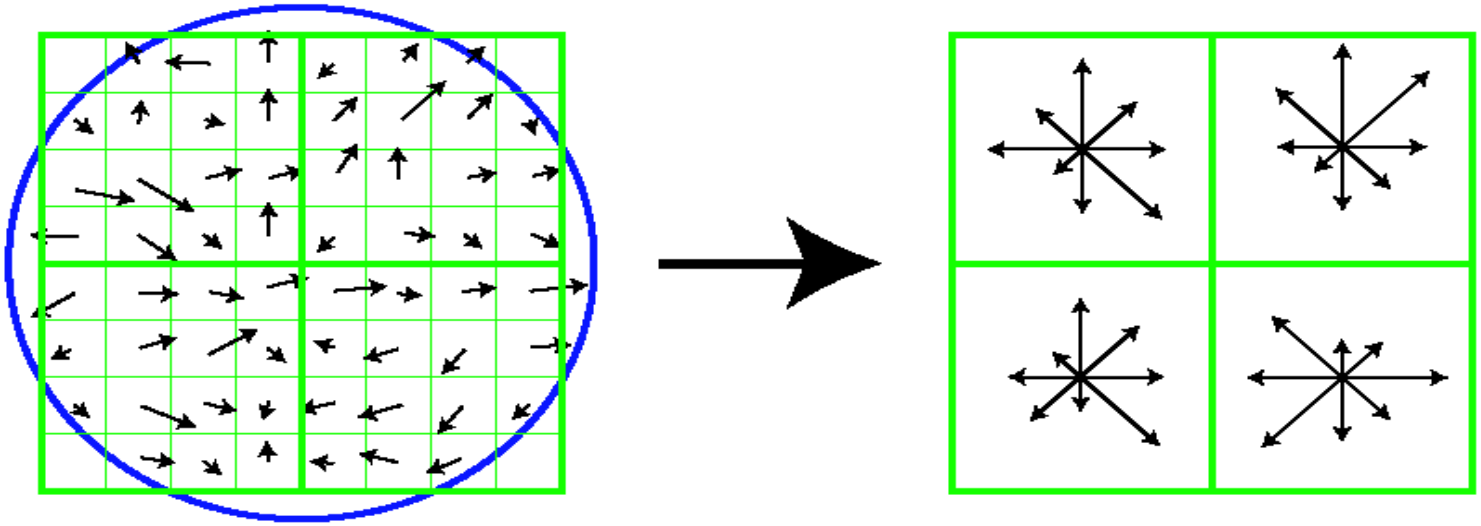
Correspondence



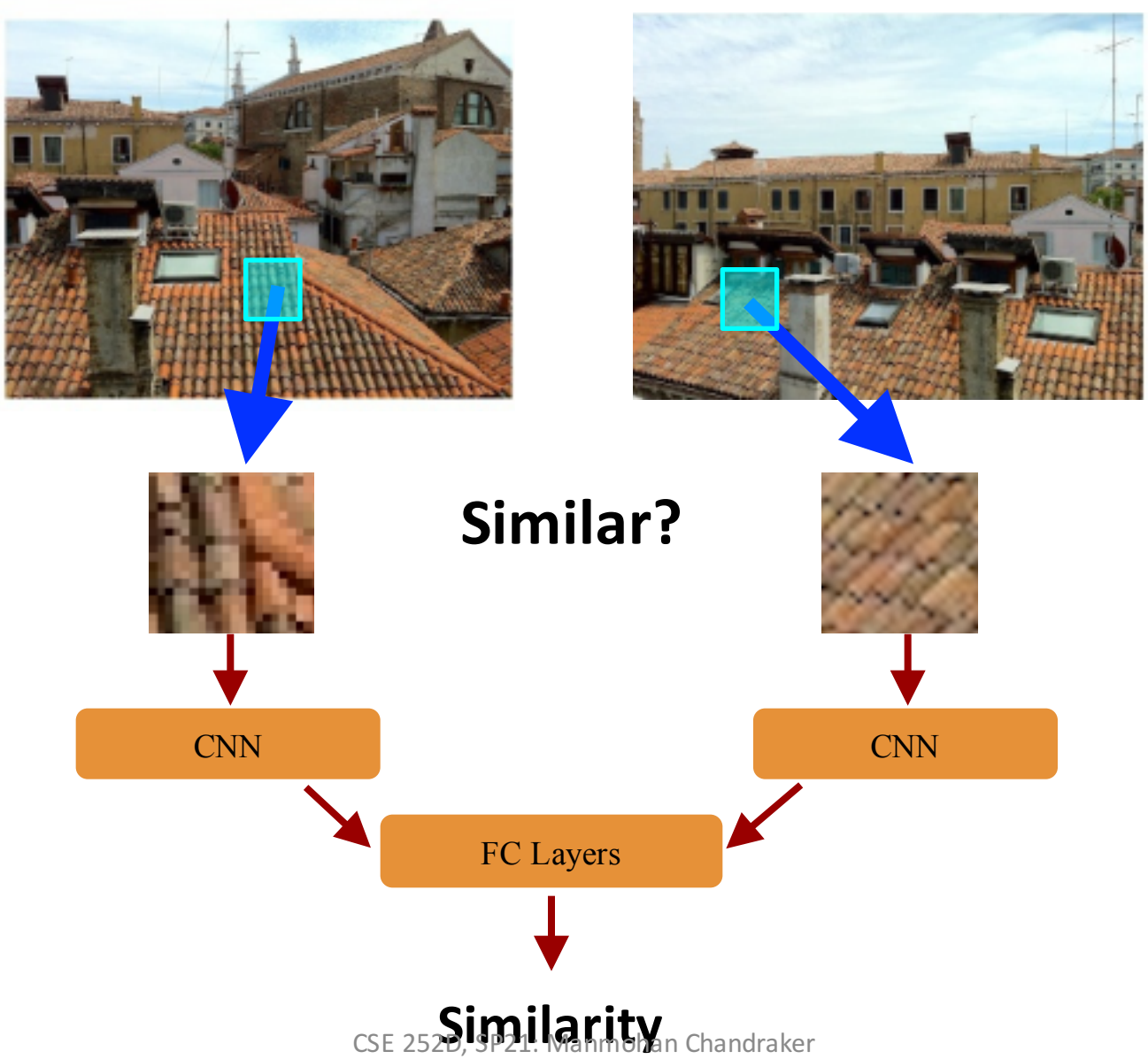
$$G(x, y, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{x^2+y^2}{\sigma^2}}$$

$$G(x, y, k\sigma) - G(x, y, \sigma) \approx (k-1)\sigma^2 \nabla^2 G.$$

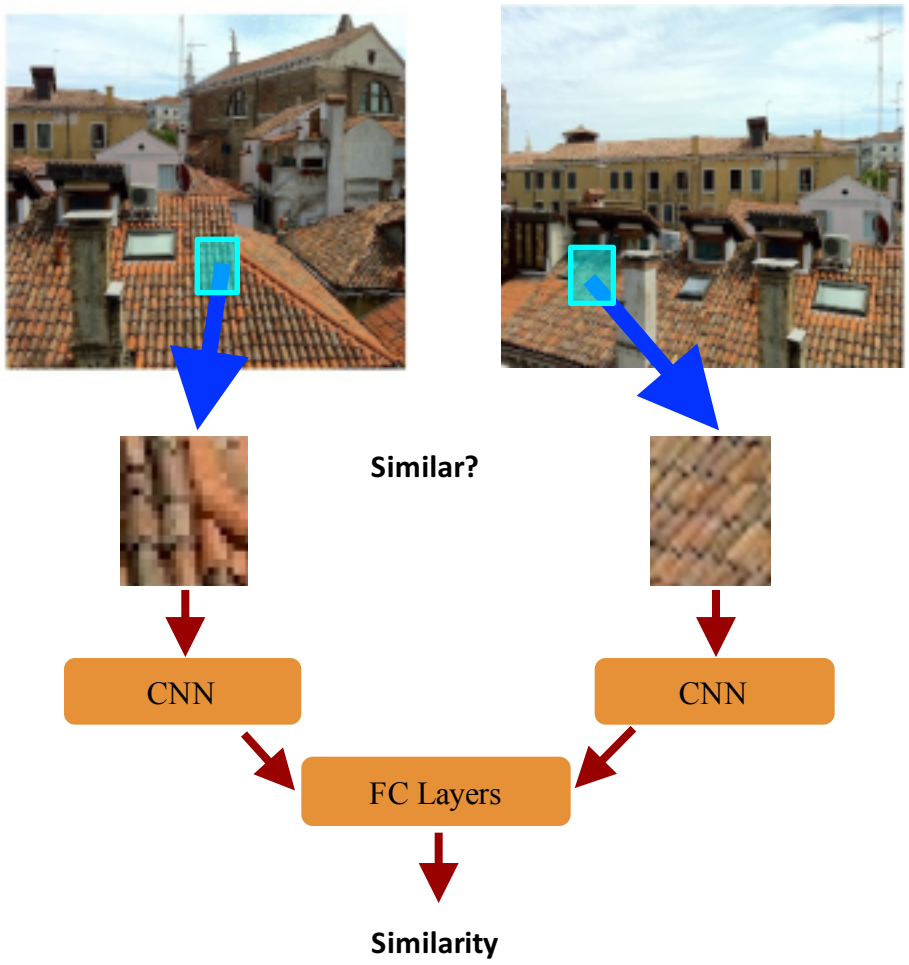




Correspondence using CNNs

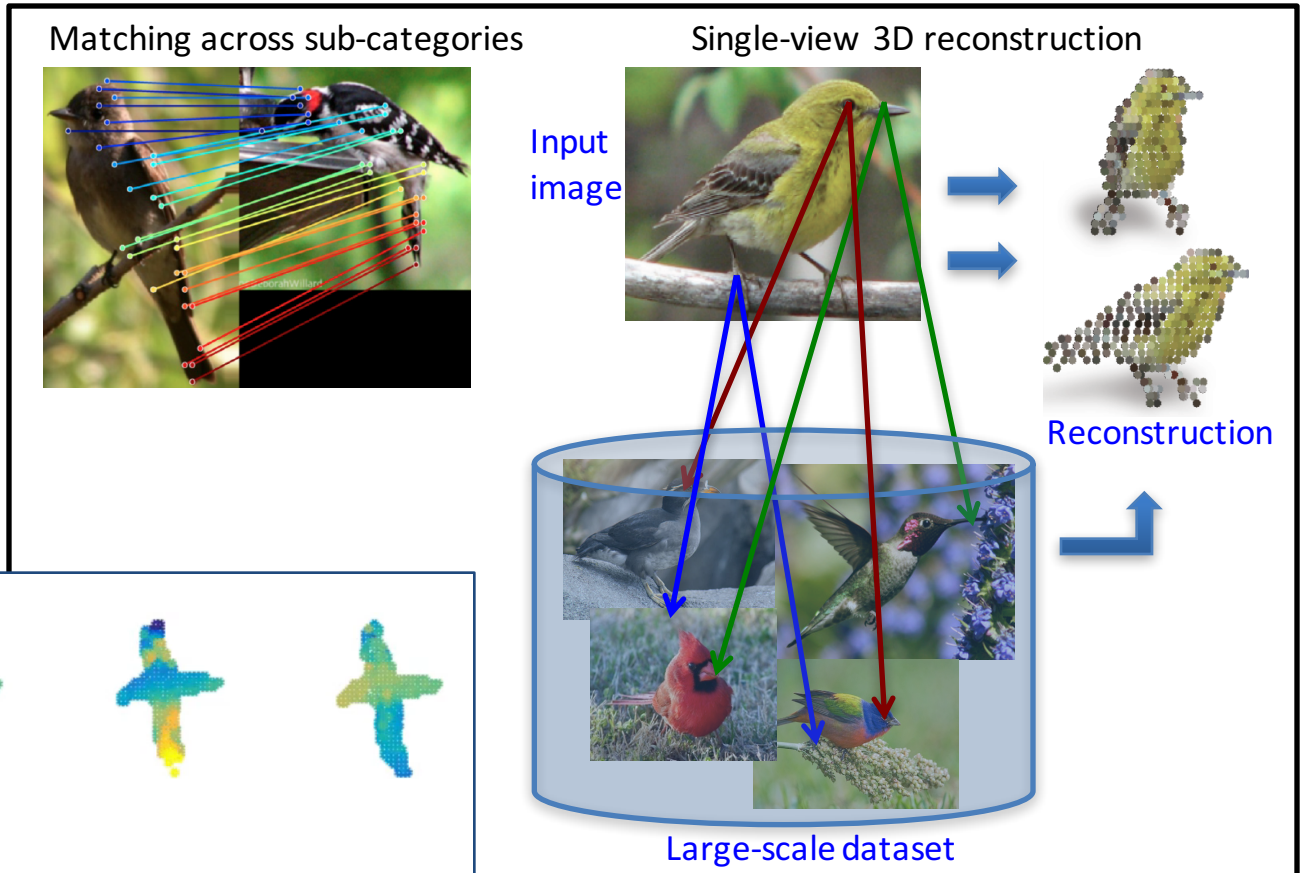


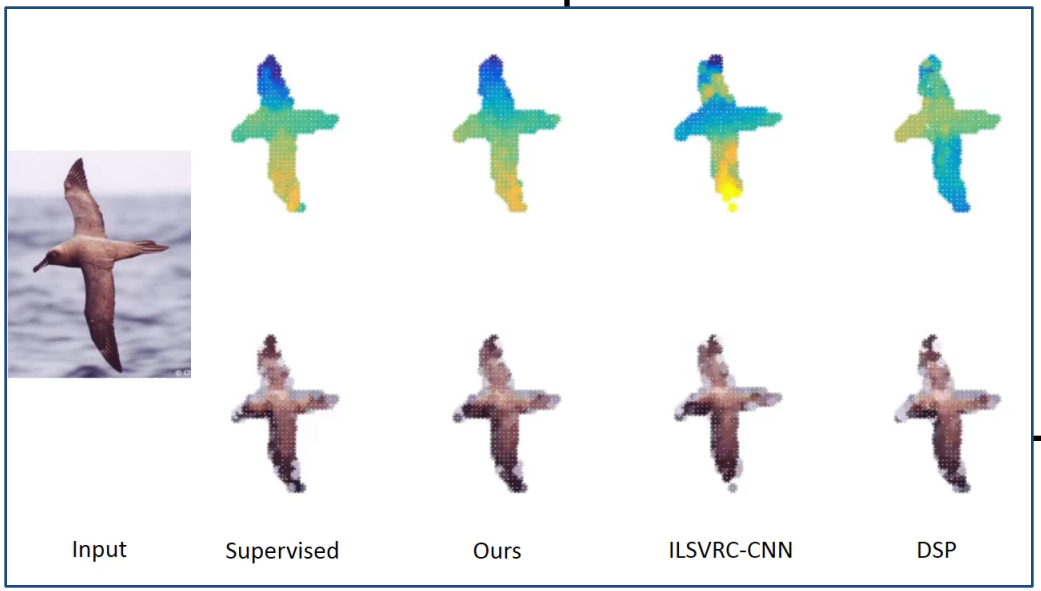
Correspondence beyond similarity CNN



- Detection of interest points
- Normalization of patches
- Multiscale information
- Obtaining training data
- Efficient training and testing

Semantic correspondence





[Kanazawa et al., "WarpNet"]

Metric learning

- A metric quantifies distance between any two members of a set
- Goal: pull positives closer, push negatives farther away

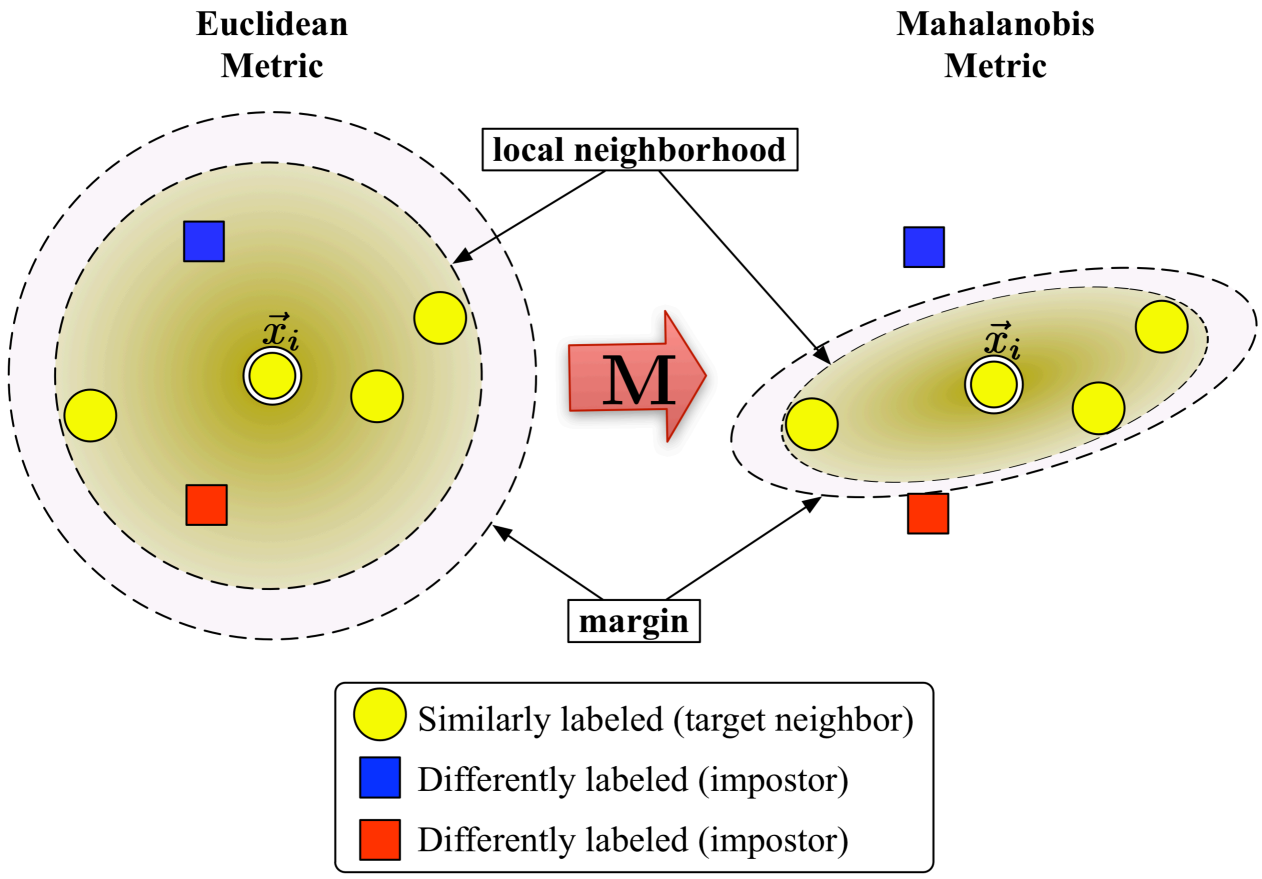
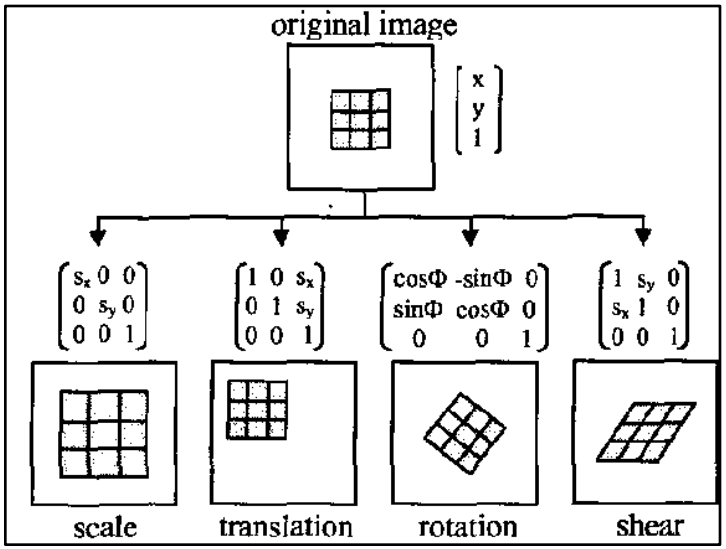
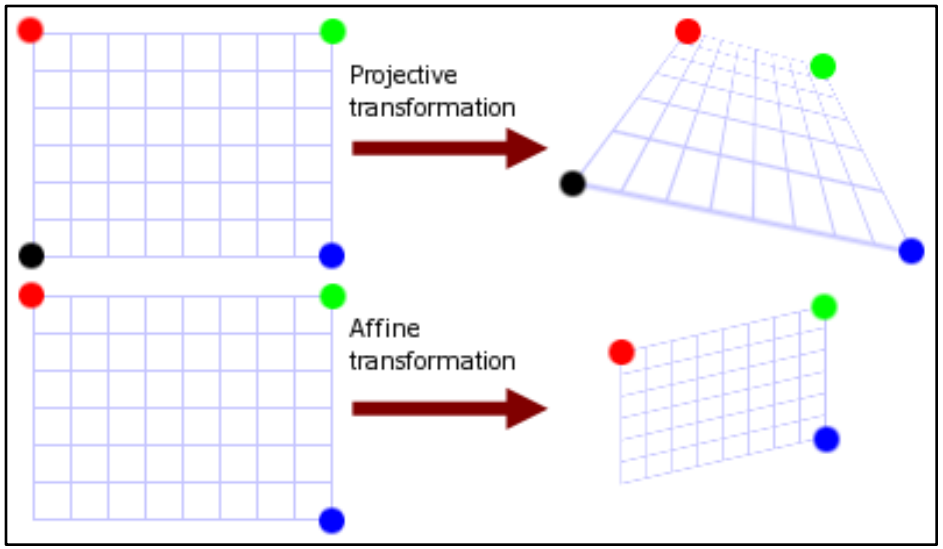
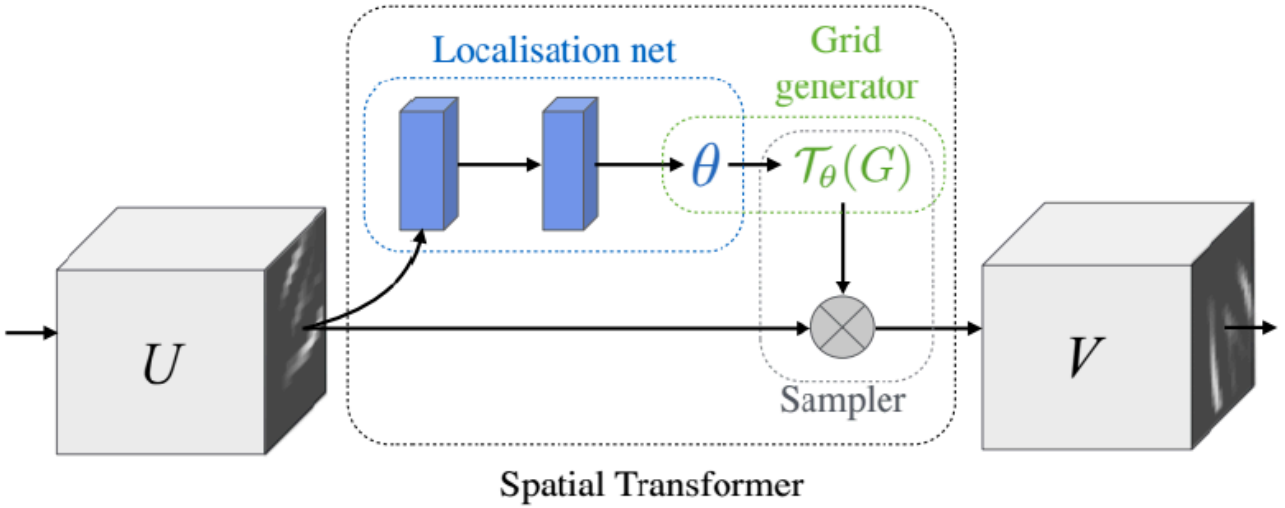


Image Coordinate Transformations





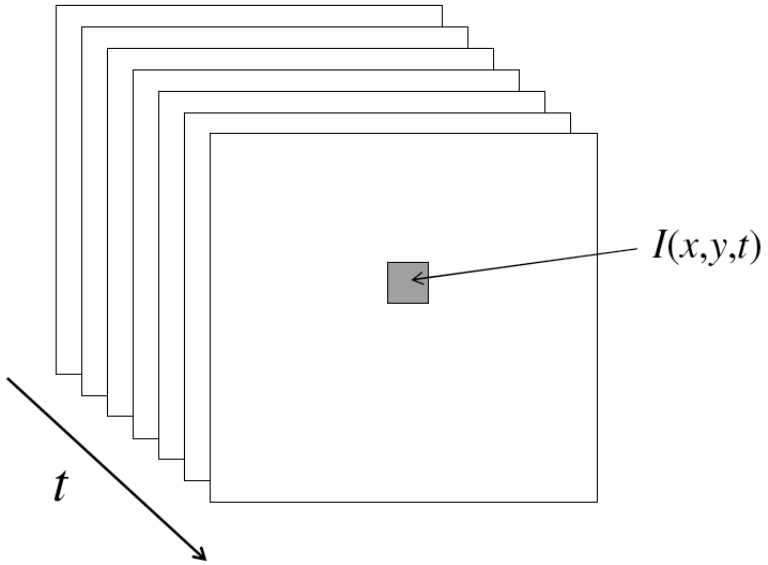


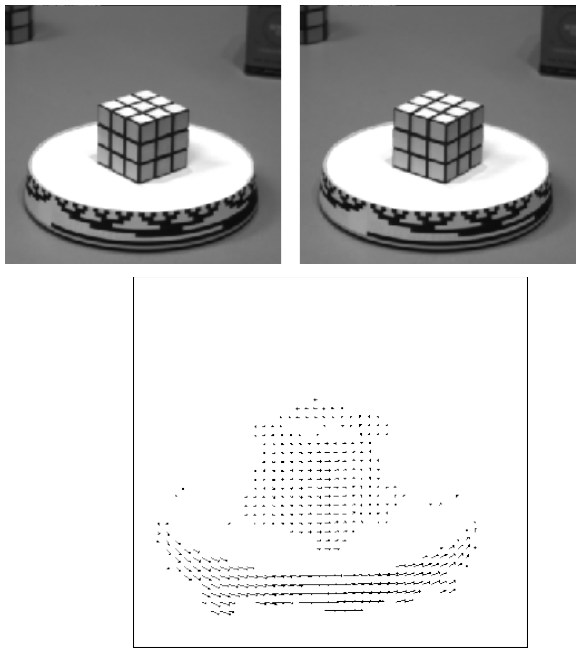
Optical flow

Brightness constancy constraint equation

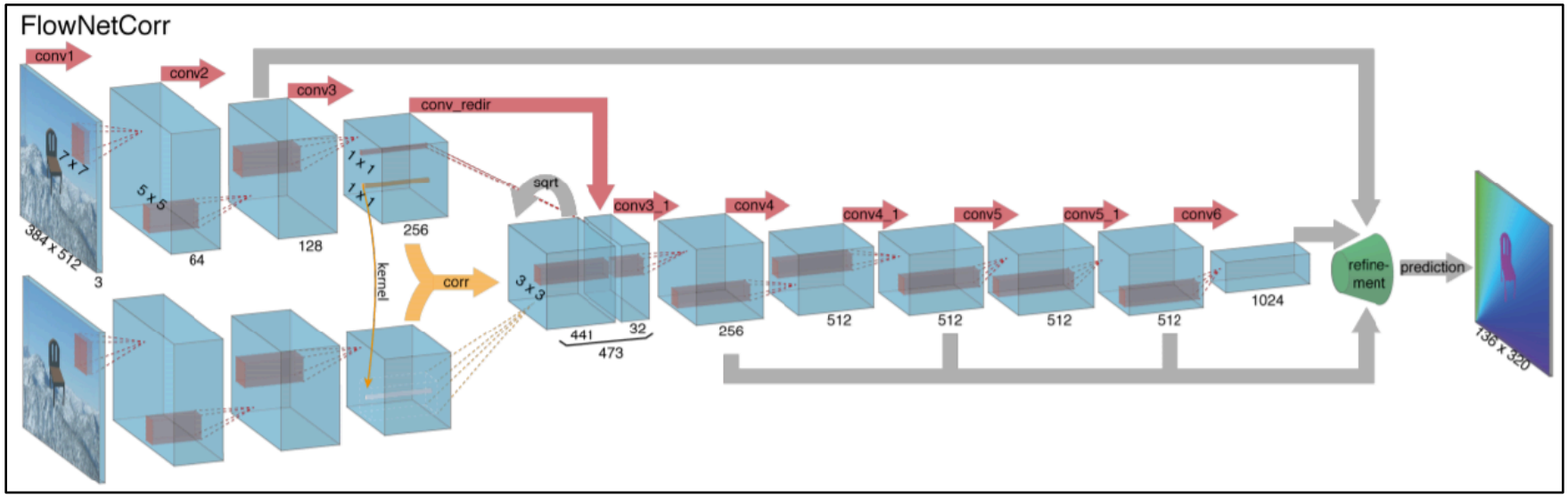
$$I_x\, u + I_y\, v + I_t = 0$$

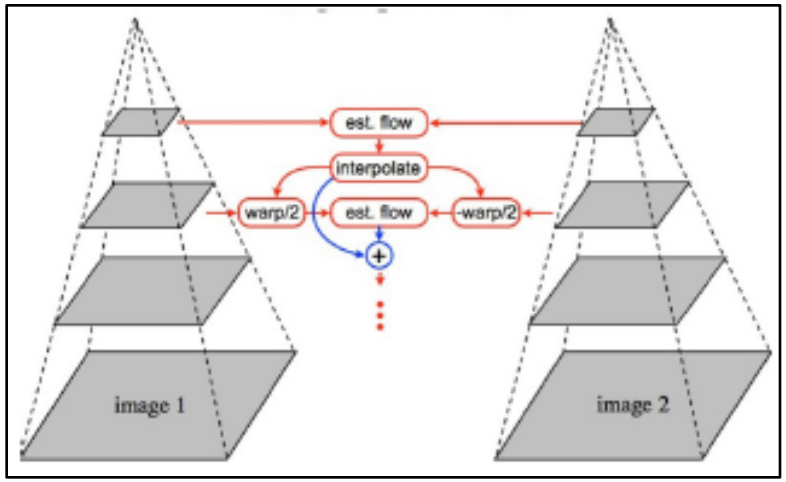
- Number of equations and unknowns per pixel:
  - One equation, two unknowns (u, v)

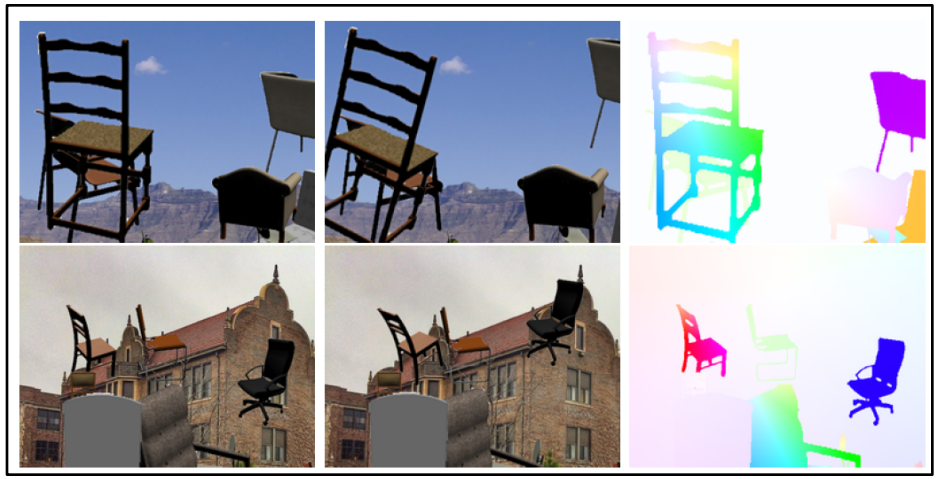




Optical flow: CNNs

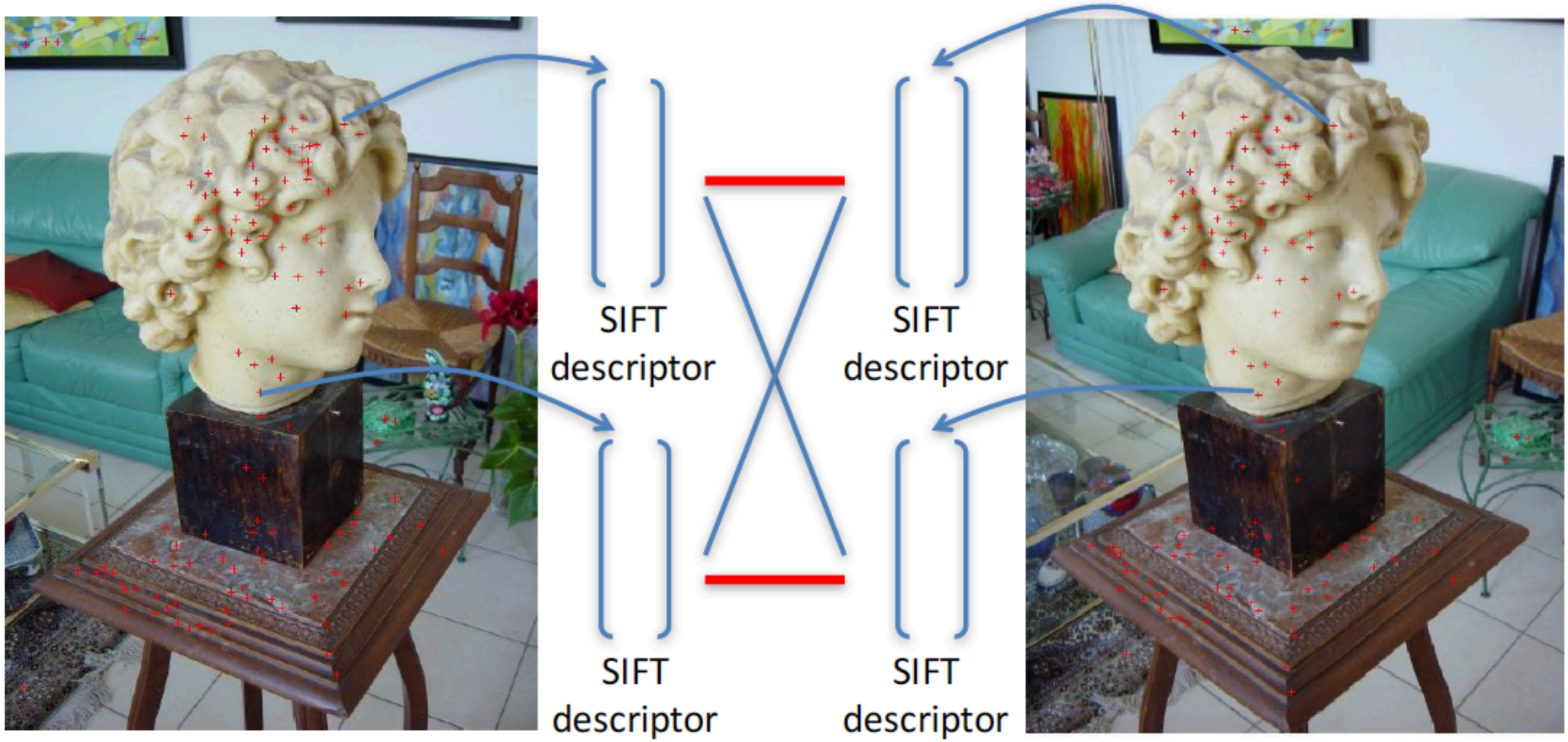


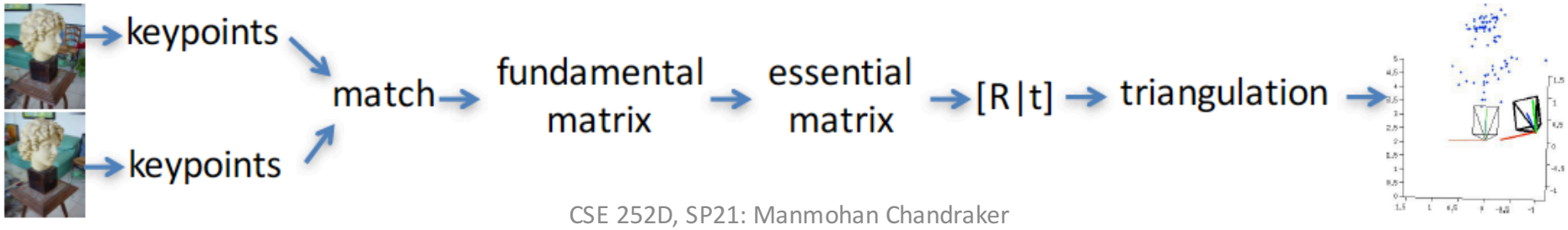




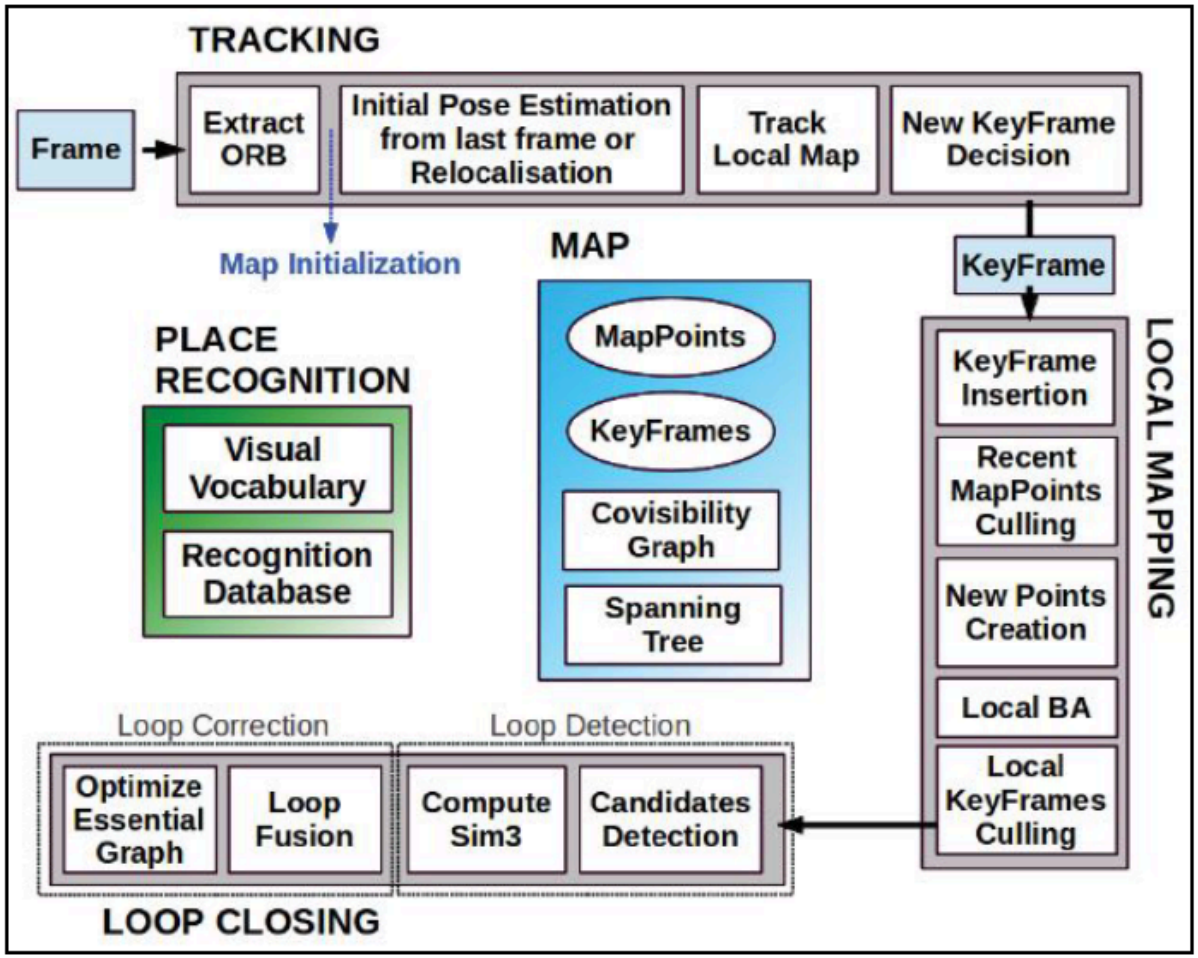
[Dosovitskiy et al., "FlowNet"]

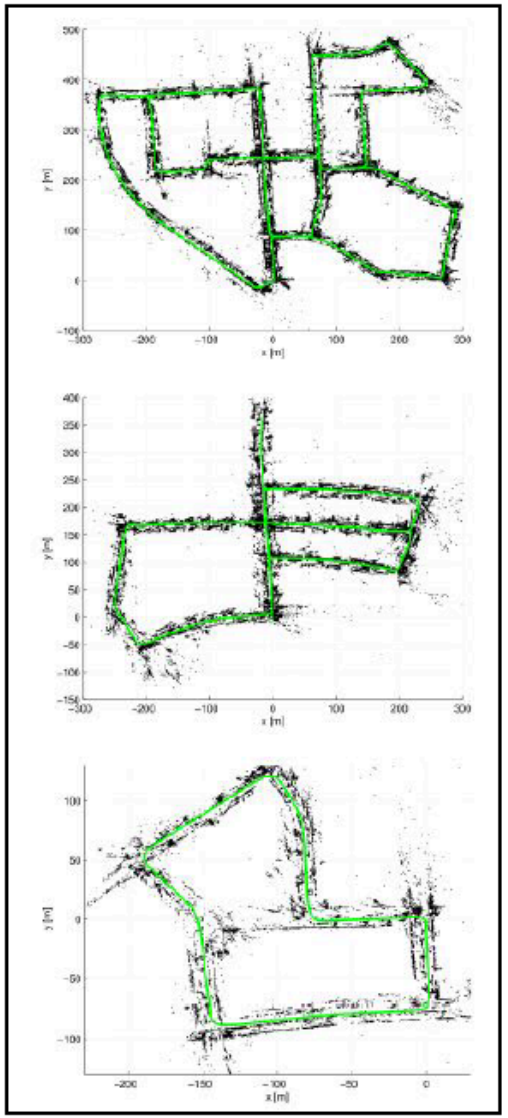
Structure from Motion





Structure from Motion

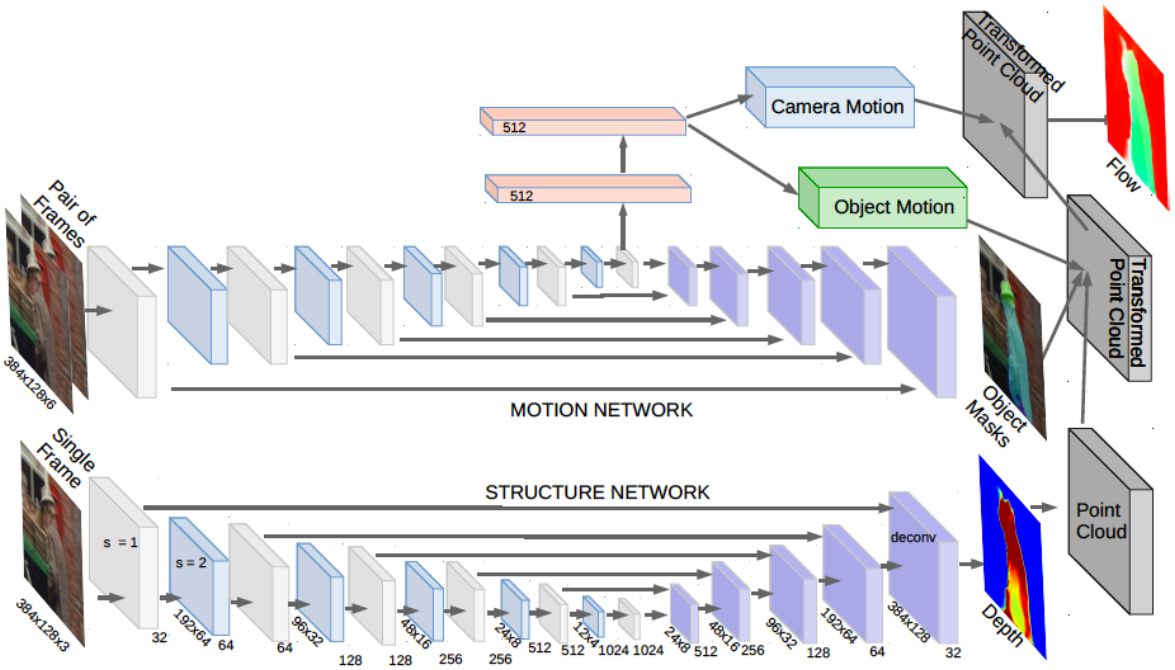




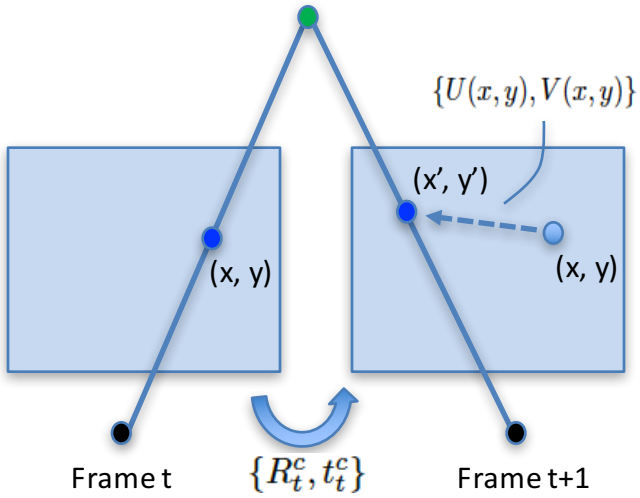
[Mur-Artal et al., "ORB-SLAM"]

Deep Networks for SFM

Structure and motion sub-networks

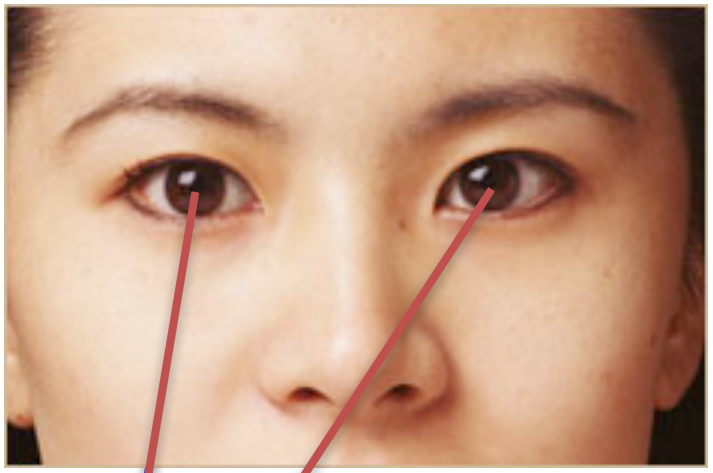


Photoconsistency



Correspondence

Why Do We Have Two Eyes?

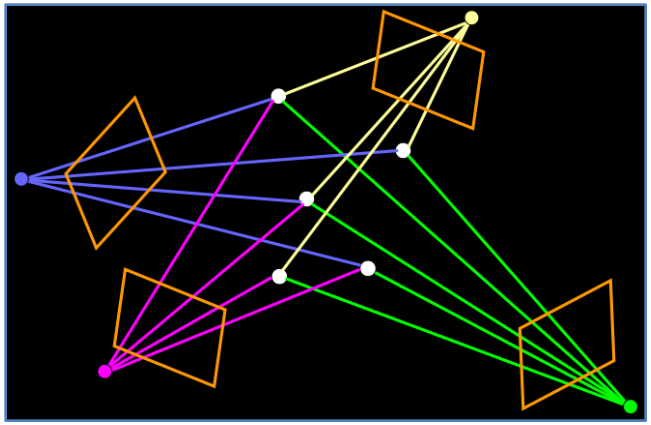


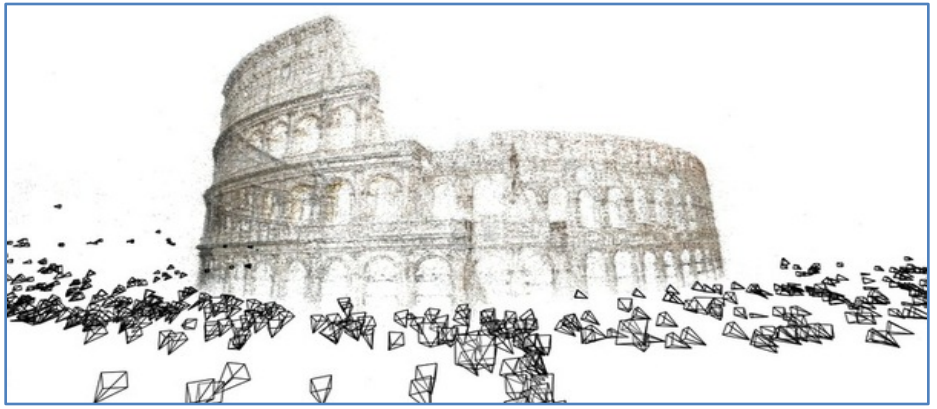
Binocular (stereo) vision
enables depth estimation

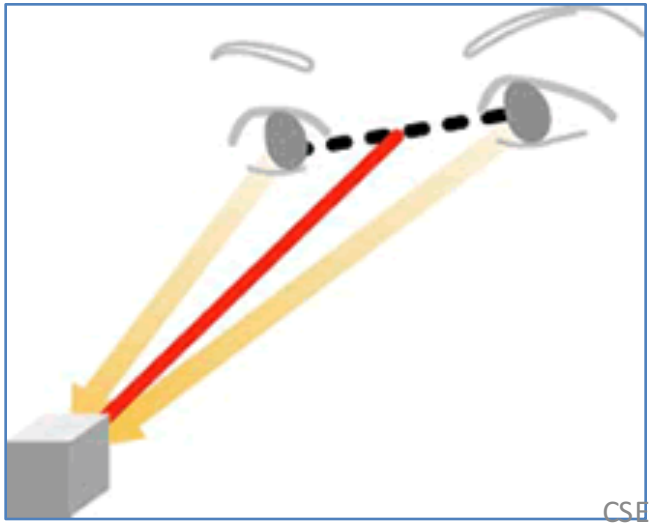
Depth information lost
in image formation

How do we perceive depth in images?

Relate projections of the same point in two or more images of the scene.

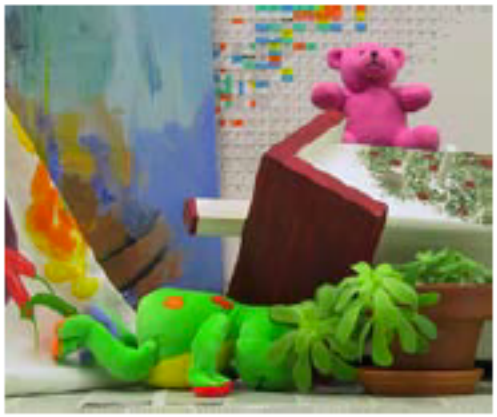
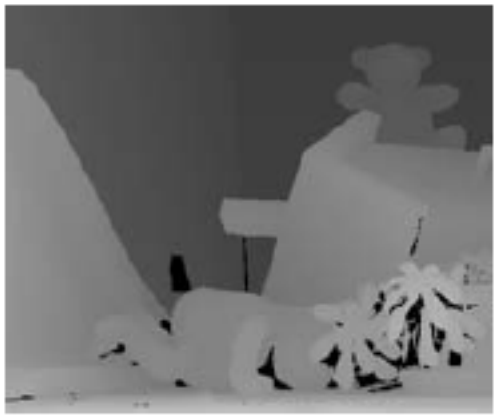


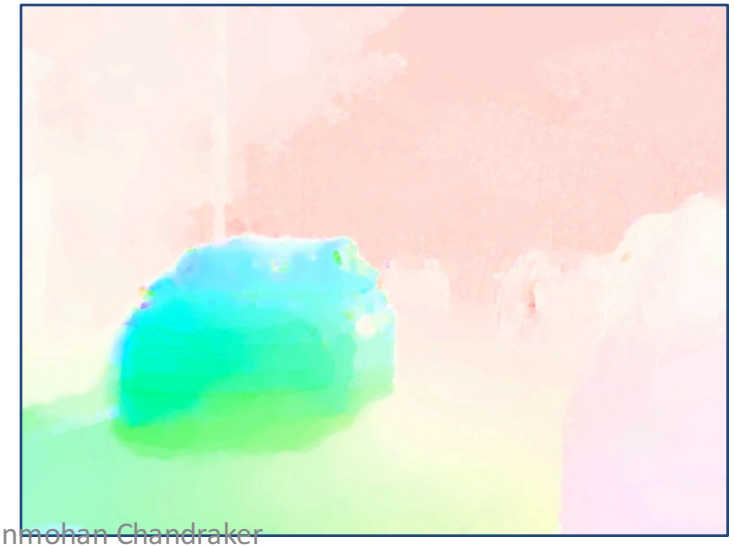




How do we perceive depth in images?

Relate intensities of image points in two or more images.



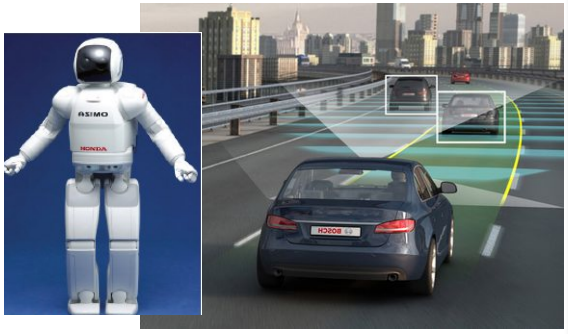
How do we perceive depth in images?

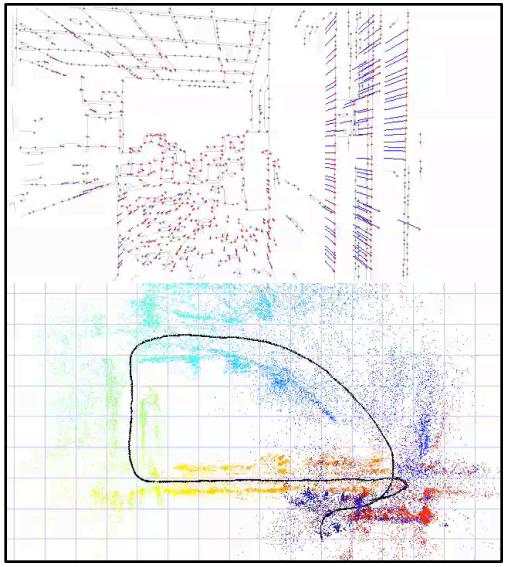
Use prior information in the form of semantics, function, affordance.

Visual correspondence aids 3D reconstruction

A key problem in 3D reconstruction: relate similar elements across a set of images

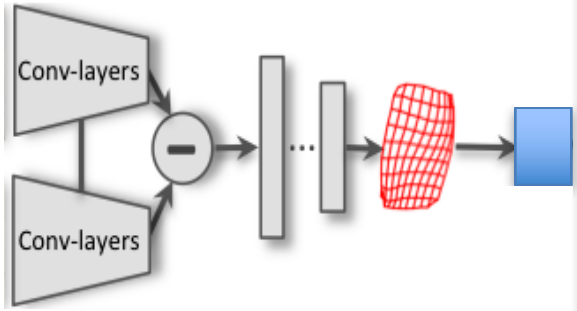
Geometric

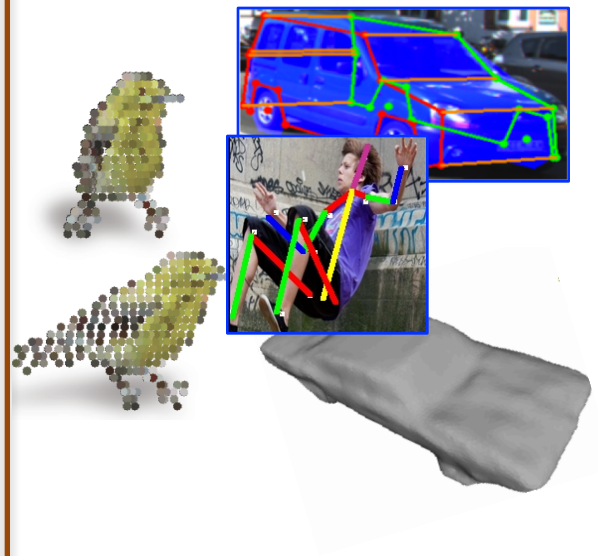




Photometric

Semantic





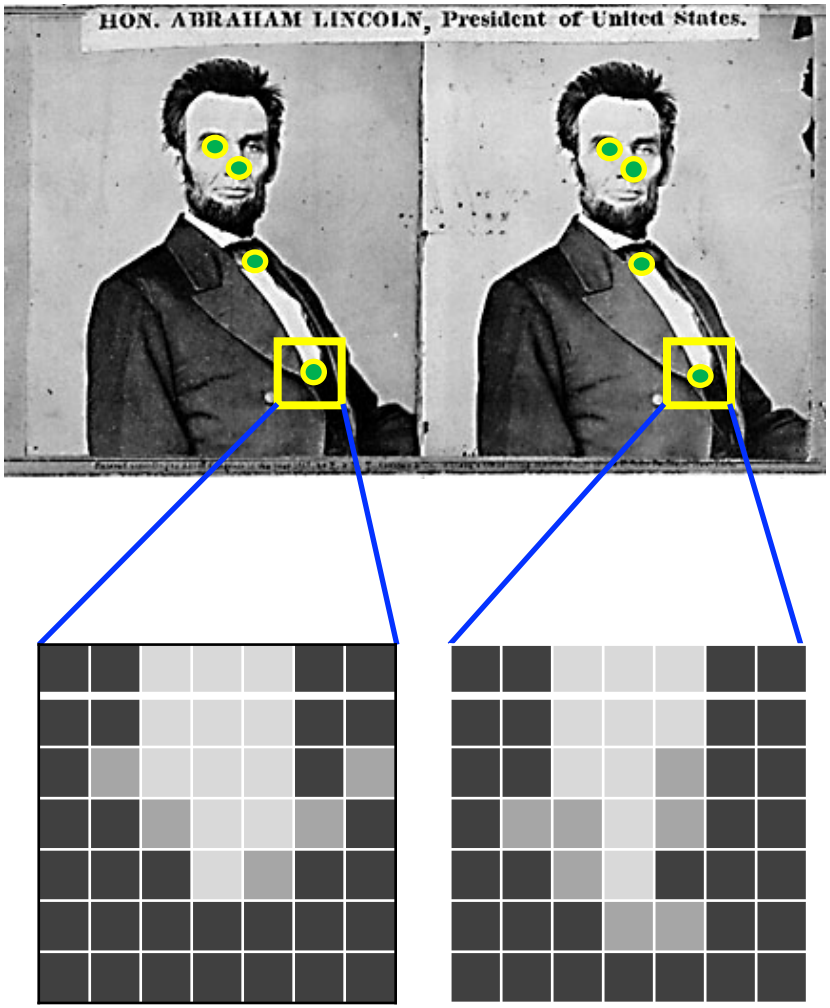
Simple matching methods

Interest point:

- Localized position
- Informative about image content
- Repeatable under variations

Descriptor:

- *Function* applied on W_1 and W_2, to enable *comparing* them



$W_1(x, y)$: k x k pixel patch in image 1

$W_2(x, y)$: k x k pixel patch in image 2

Simple matching methods

- SSD (Sum of Squared Differences)

$$\sum_{x,y} |W_1(x,y) - W_2(x,y)|^2$$

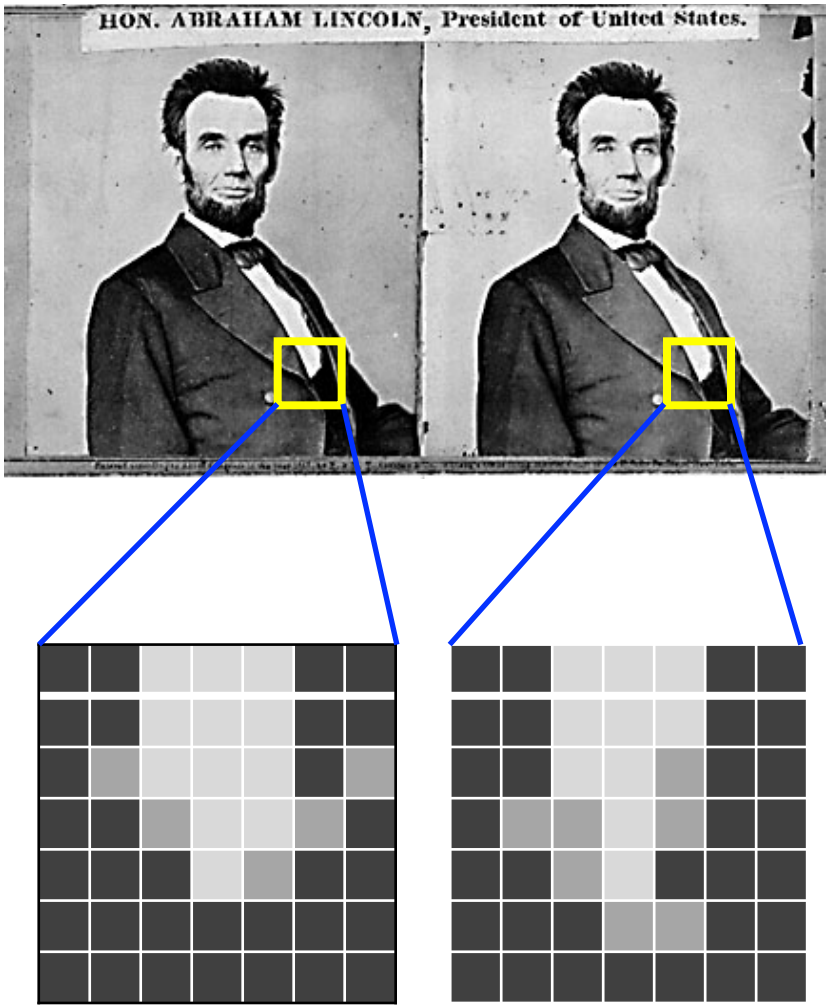
- NCC (Normalized Cross Correlation)

$$\sum_{x,y} \frac{(W_1(x,y) - \overline{W_1})(W_2(x,y) - \overline{W_2})}{\sigma_{W_1}\sigma_{W_2}}$$

$$\overline{W_i} = \frac{1}{n}\sum_{x,y} W_i \,, \quad \sigma_{W_i} = \sqrt{\frac{1}{n}\sum_{x,y}(W_i - \overline{W_i})^2}$$

(Mean) (Standard deviation)

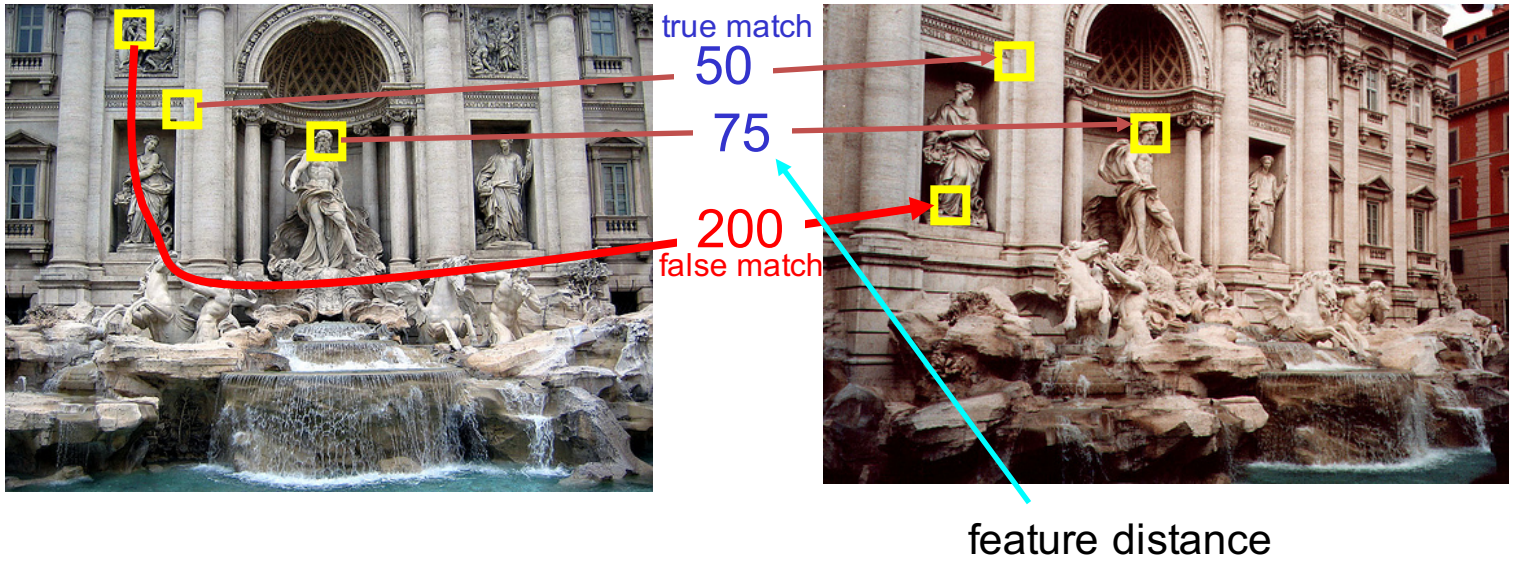
- What advantages might NCC have over SSD?



$W_1(x, y)$: k x k pixel patch in image 1

$W_2(x, y)$: k x k pixel patch in image 2

Feature distance: threshold

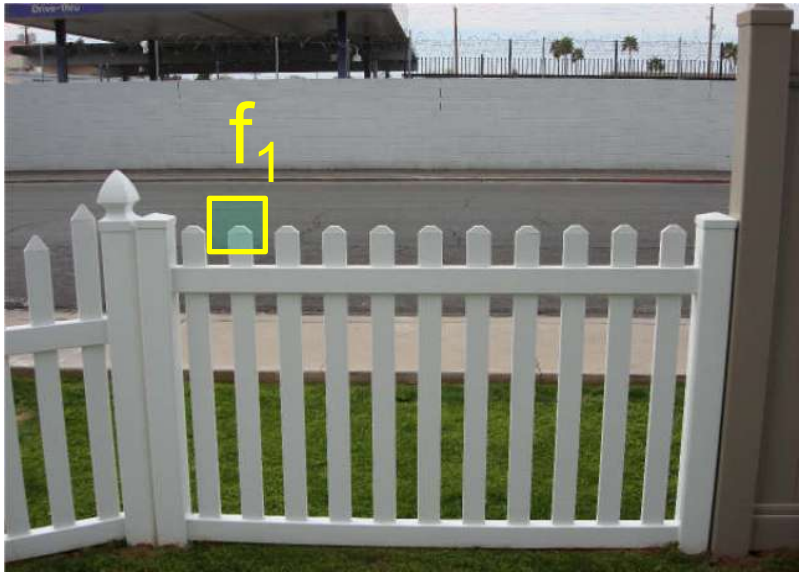


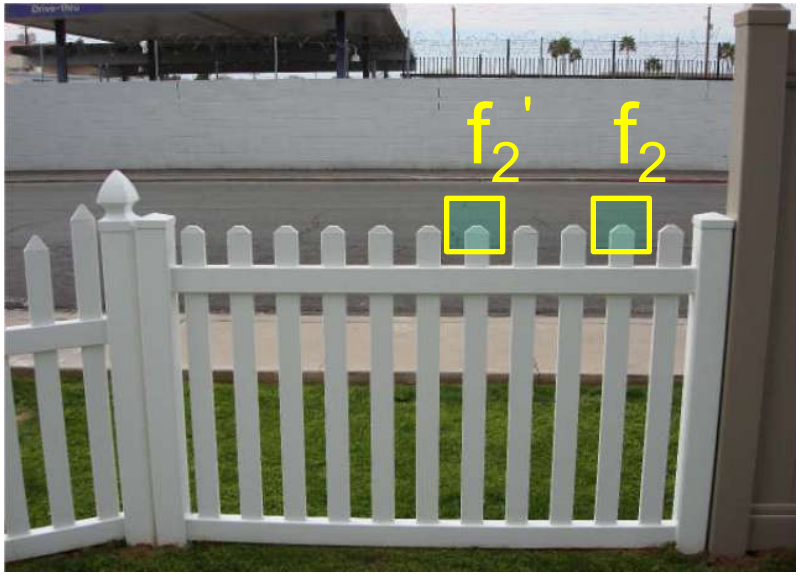
- The distance threshold affects performance
- Only matches with distance less than threshold are allowed

- True positives = number of detected matches that are correct
  - Suppose we want to maximize these — how to choose threshold?
- False positives = number of detected matches that are incorrect
  - Suppose we want to minimize these — how to choose threshold?

Choosing a match: ratio test

- First approach: use SSD = $\|f_2 - f_1\|$
- Better approach: ratio distance = $\dfrac{\|f_2 - f_1\|}{\|f_2' - f_1\|}$
  - f_2 is best SSD match to f_1 in I_2
  - f_2' is second best SSD match to f_1 in I_2
  - Gives large values (close to 1) for ambiguous matches.





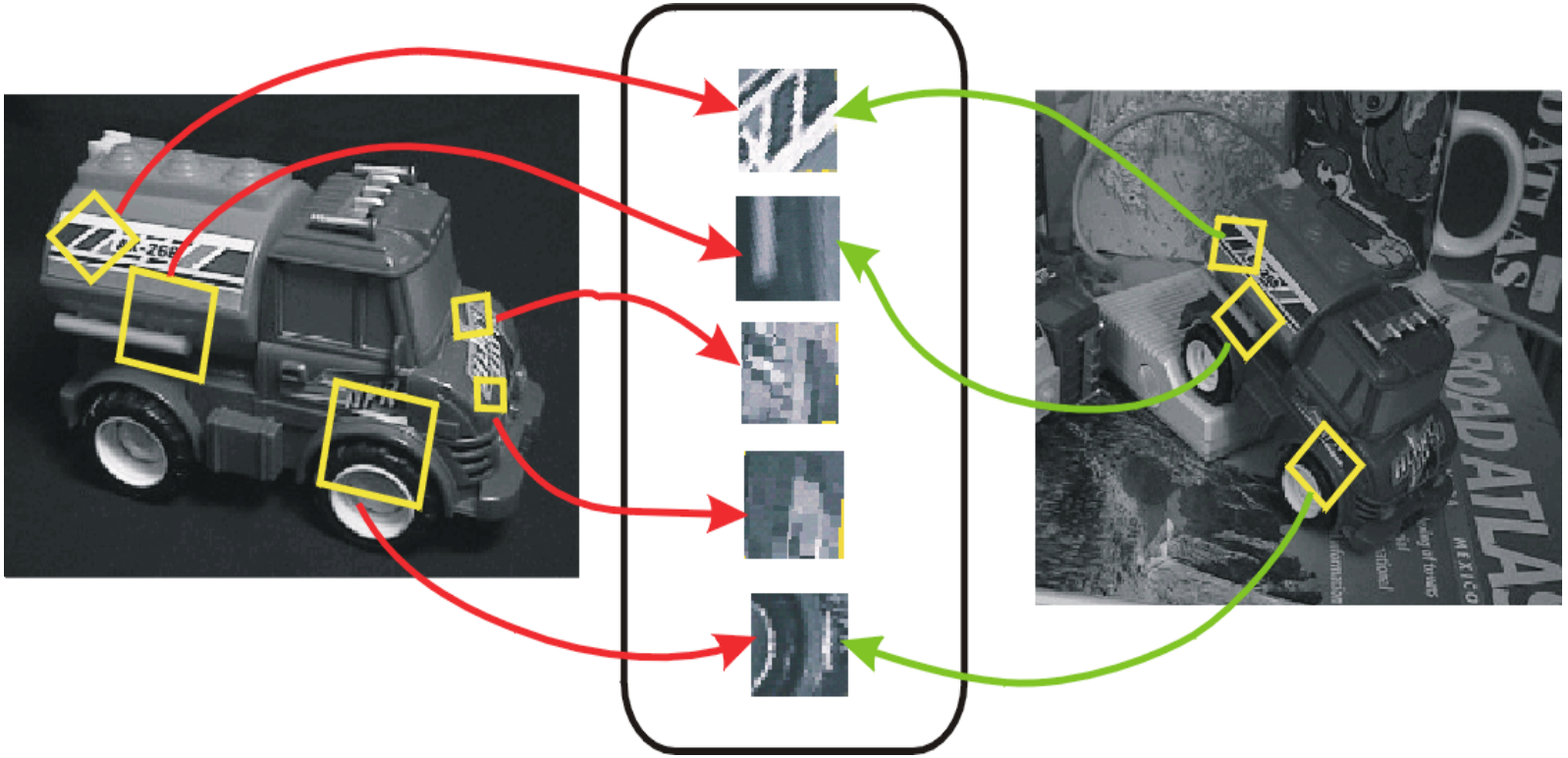
I_1 I_2

SIFT

Desirable property: invariance

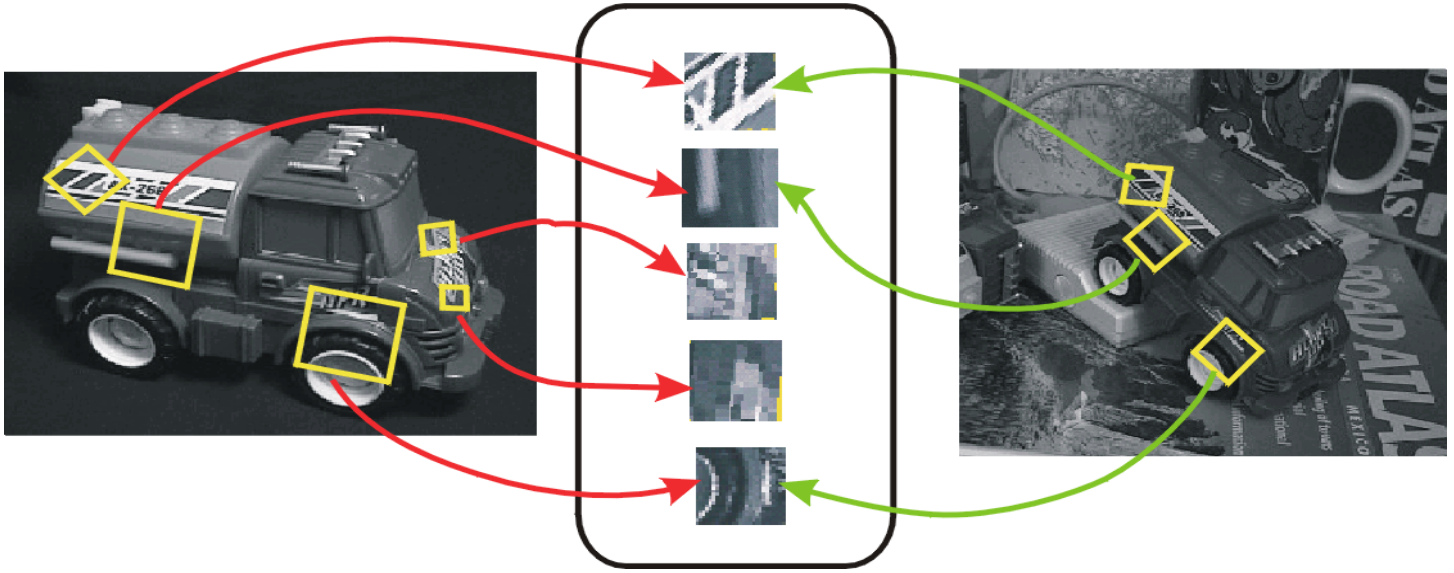
Find features that are invariant to transformations

- geometric invariance: translation, rotation, scale
- photometric invariance: brightness, exposure, …



Idea of SIFT

- For better image matching, need to develop an interest operator invariant to scale and rotation.
- Also, need a descriptor robust to typical variations. The descriptor is the most-used part of SIFT.



[Adapted from: D. Lowe]

Overall Procedure at a High Level

1. Scale-space extrema detection
   Search over multiple scales and image locations.

2. Keypoint localization
   Fit a model to determine location and scale.
   Select keypoints based on a measure of stability.

3. Orientation assignment
   Compute best orientation(s) for each keypoint region.

4. Keypoint description
   Use local image gradients at selected scale and rotation to describe each keypoint region.

[Adapted from: D. Lowe]

1. Scale-space extrema detection

- Goal: Identify locations and scales that can be reliably assigned under different views of the same scene or object.
- Method: search for stable features across multiple scales using a continuous function of scale.
- The scale space of an image is a function $L(x,y,\sigma)$ that is produced from the convolution of a Gaussian kernel (at different scales) with the input image.

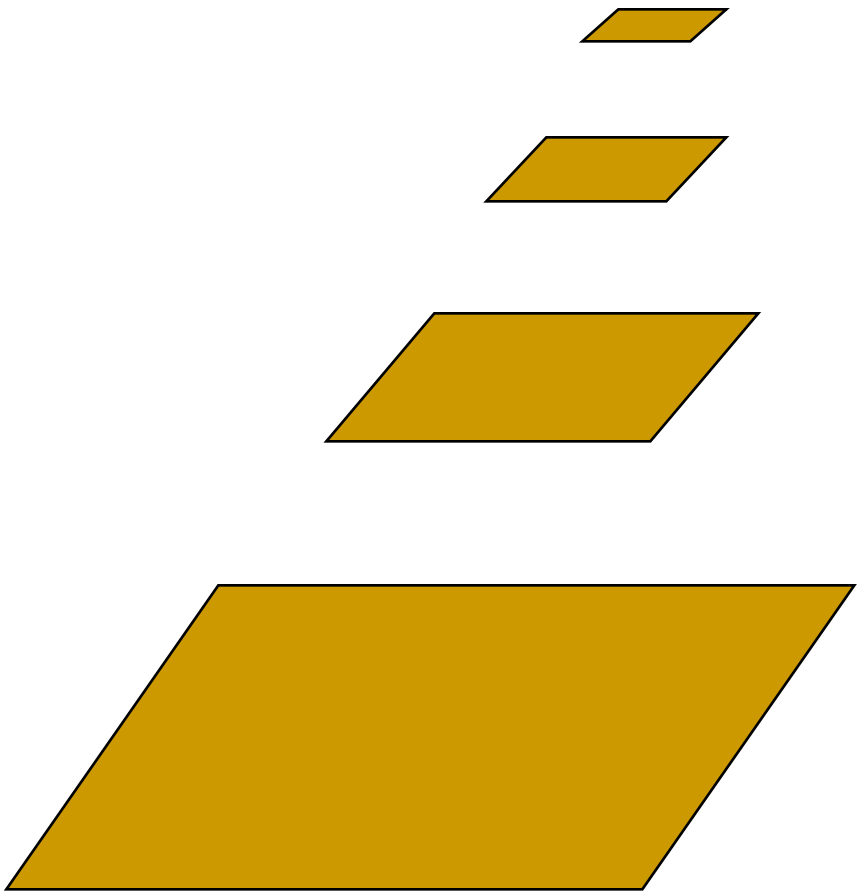
[Adapted from: D. Lowe]

Example: Gaussian Smoothed Scale Space

- Scale space axioms: linearity, shift invariant, rotation invariant, no spurious extrema
- Gaussian filter uniquely satisfies axioms

Aside: Image Pyramids

A concept that arises all across computer vision



And so on.

3rd level derived from 2nd level according to same function

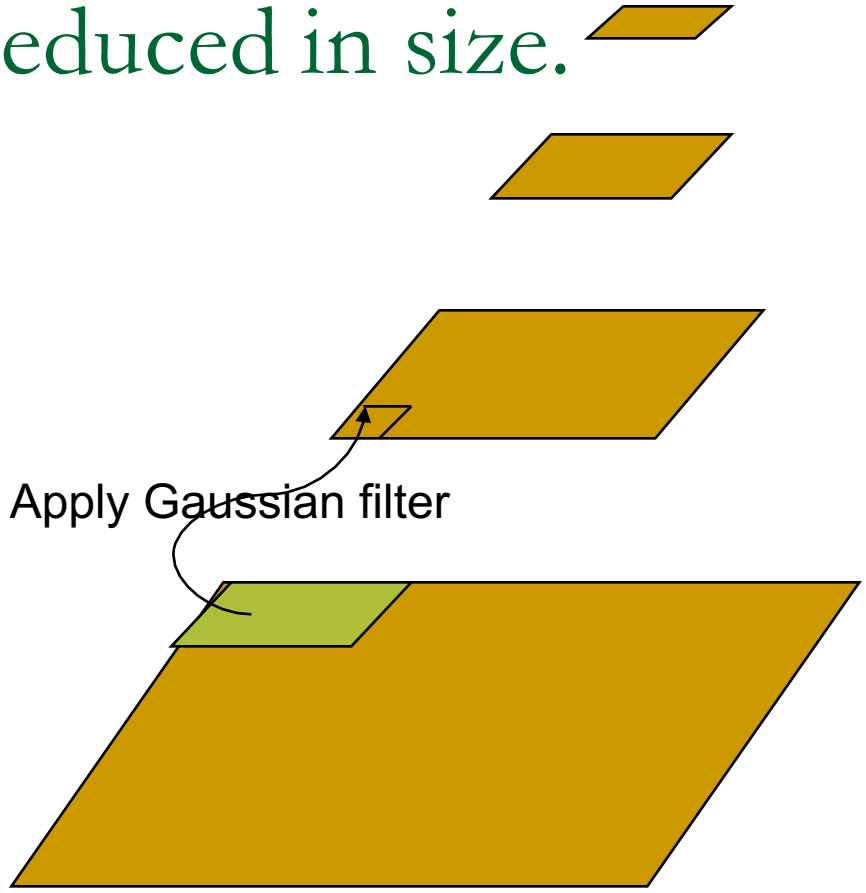
2nd level is derived from the original image according to some function

Bottom level is the original image.

[Adapted from: D. Lowe]

Aside: Gaussian Pyramid

At each level, image is smoothed and reduced in size.



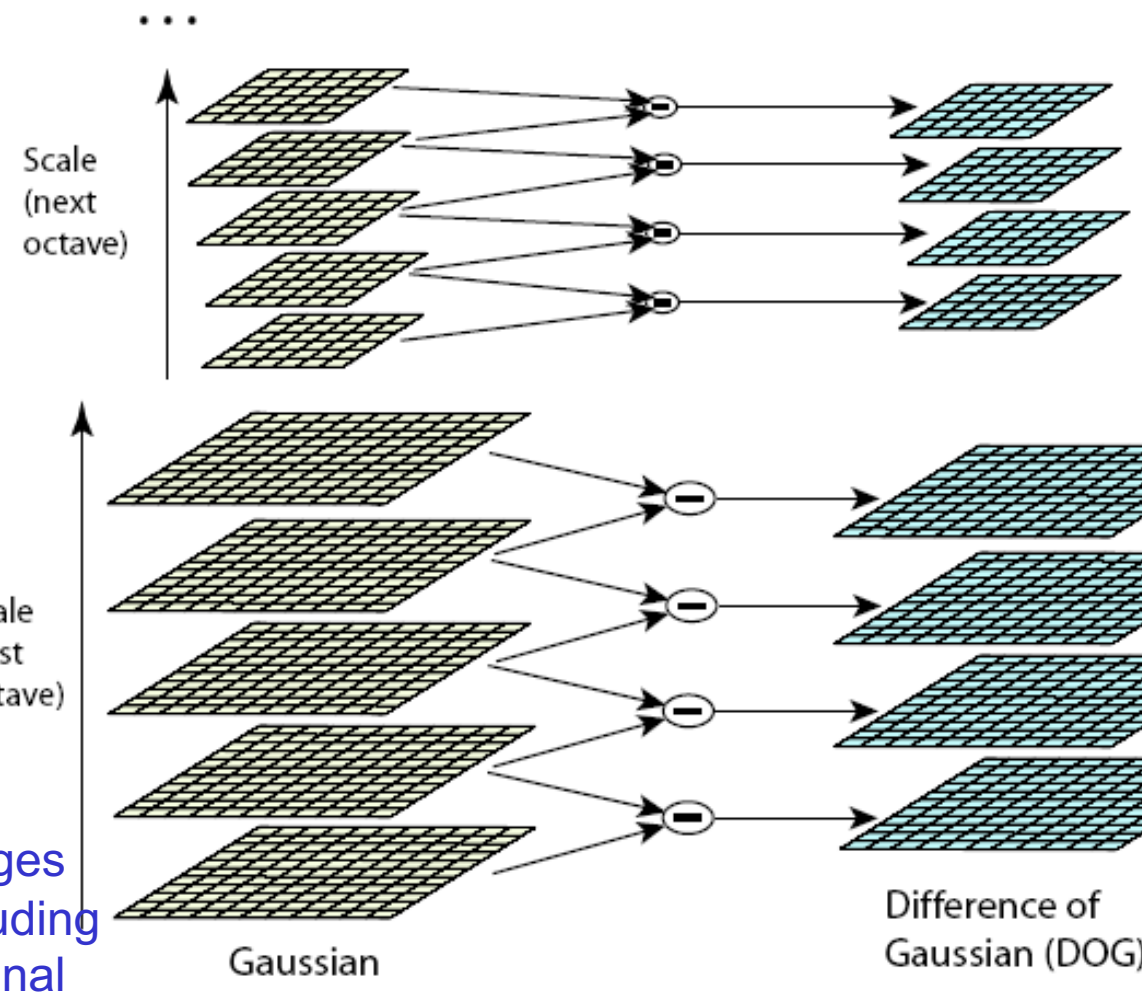
And so on.

At 2nd level, each pixel is the result of applying a Gaussian mask to the first level and then subsampling to reduce the size.

Bottom level is the original image.

[Adapted from: D. Lowe]

Scale Space Pyramid



s+2 filters

$\sigma_{s+1}=2^{(s+1)/s}\sigma_0$

.

.

$\sigma_i=2^{i/s}\sigma_0$

.

.

$\sigma_2=2^{2/s}\sigma_0$

$\sigma_1=2^{1/s}\sigma_0$

σ_0

s+3 images including original

s+2 difference images

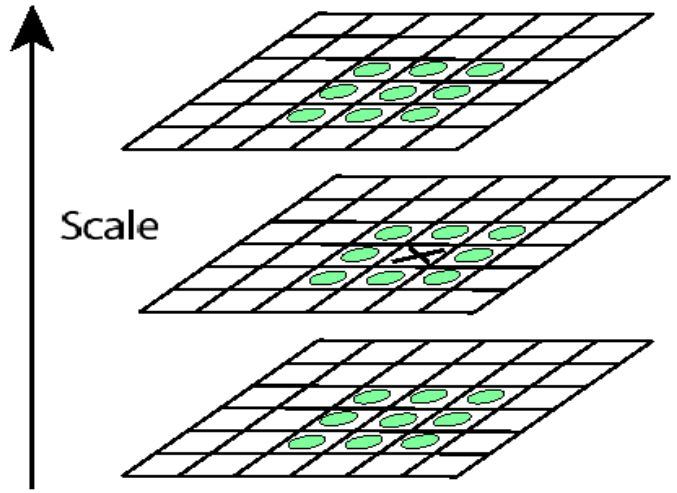
The parameter **s** determines the number of images per octave.

[Adapted from: D. Lowe]

2. Key point localization

s+2 difference images.
top and bottom ignored.
s planes searched.

- Detect maxima and minima of difference-of-Gaussian in scale space
- Each point is compared to its 8 neighbors in the current image and 9 neighbors each in the scales above and below

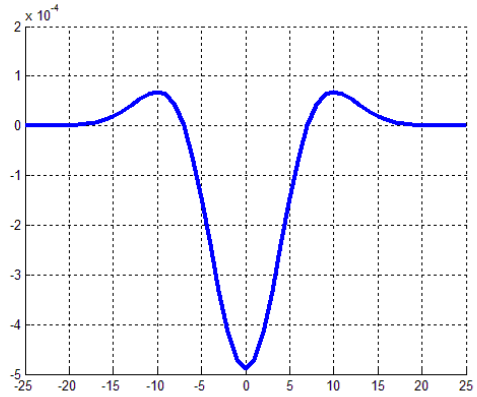


For each max or min found, output is the **location** and the **scale**.

[Adapted from: D. Lowe]

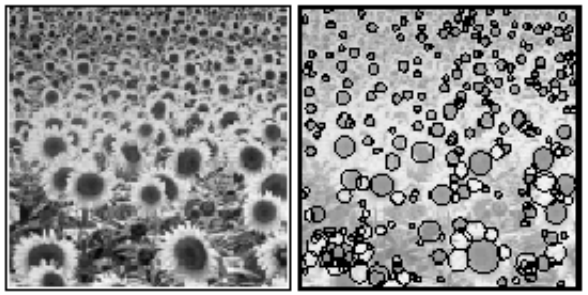
Difference of Gaussians

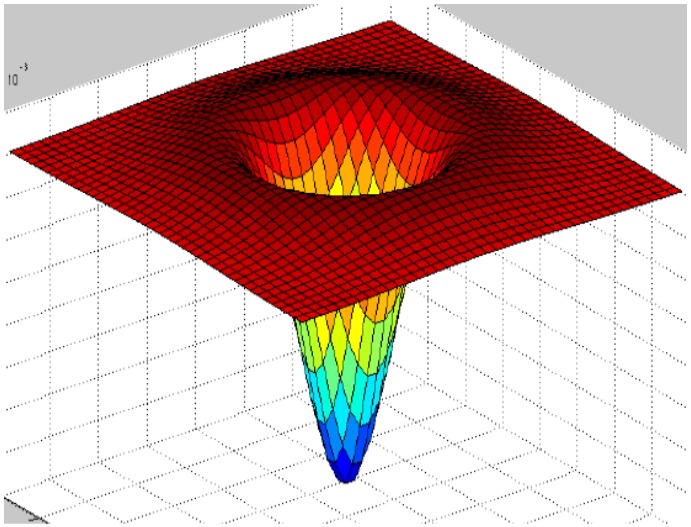
- Scale-space detection
  - Find local maxima across scale/space
  - A good “blob” detector



$$G(x, y, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{x^2+y^2}{\sigma^2}}$$

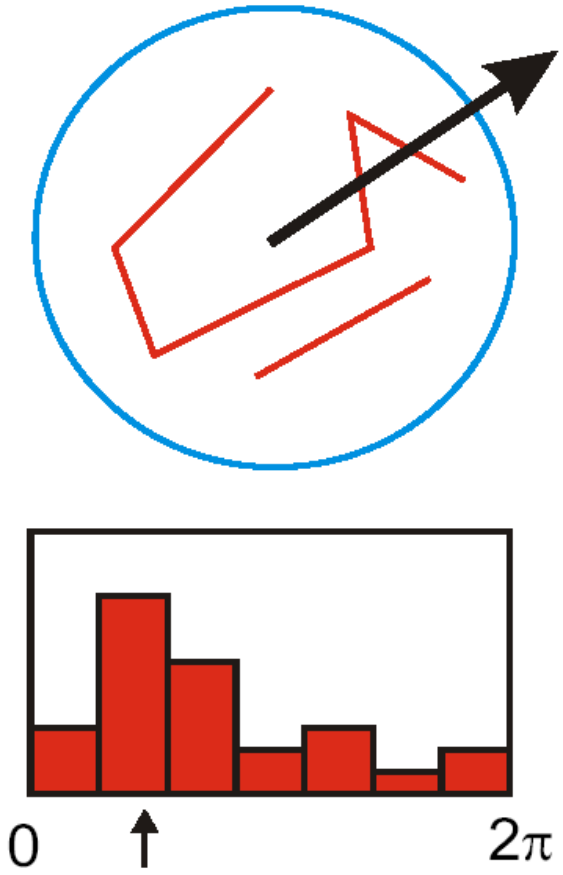
$$G(x, y, k\sigma) - G(x, y, \sigma) \approx (k-1)\sigma^2 \nabla^2 G.$$





[T. Lindeberg IJCV 1998]

3. Orientation assignment



- Create histogram of local gradient directions at selected scale
- Assign canonical orientation at peak of smoothed histogram

If 2 major orientations, use both.

[Adapted from: D. Lowe]

4. Keypoint Descriptors

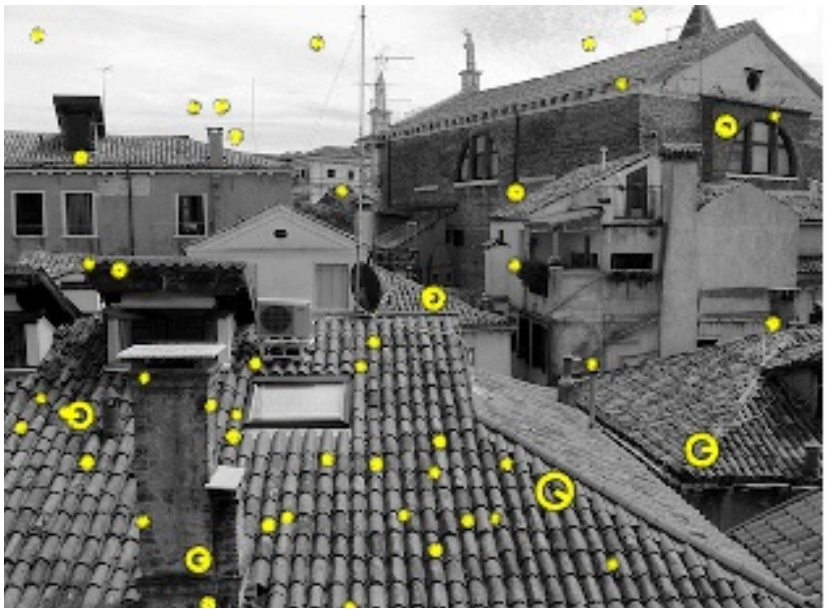
- At this point, each keypoint has
  - location
  - scale
  - orientation

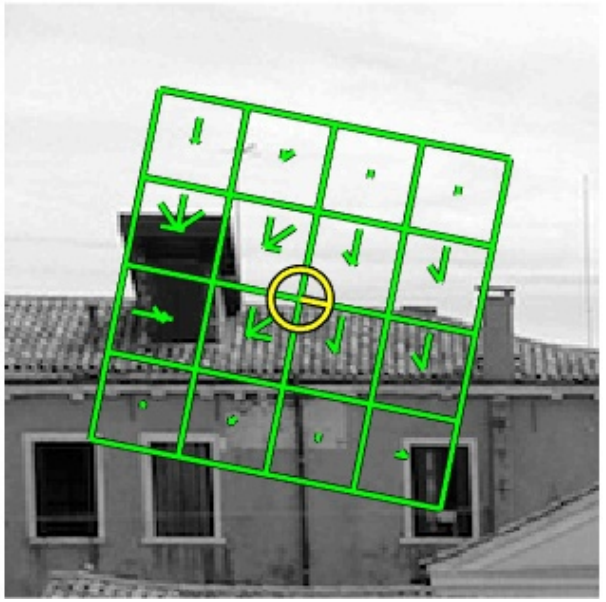
- Next is to compute a descriptor for the local image region about each keypoint that is
  - highly distinctive
  - as invariant as possible to variations such as changes in viewpoint and illumination

[Adapted from: D. Lowe]

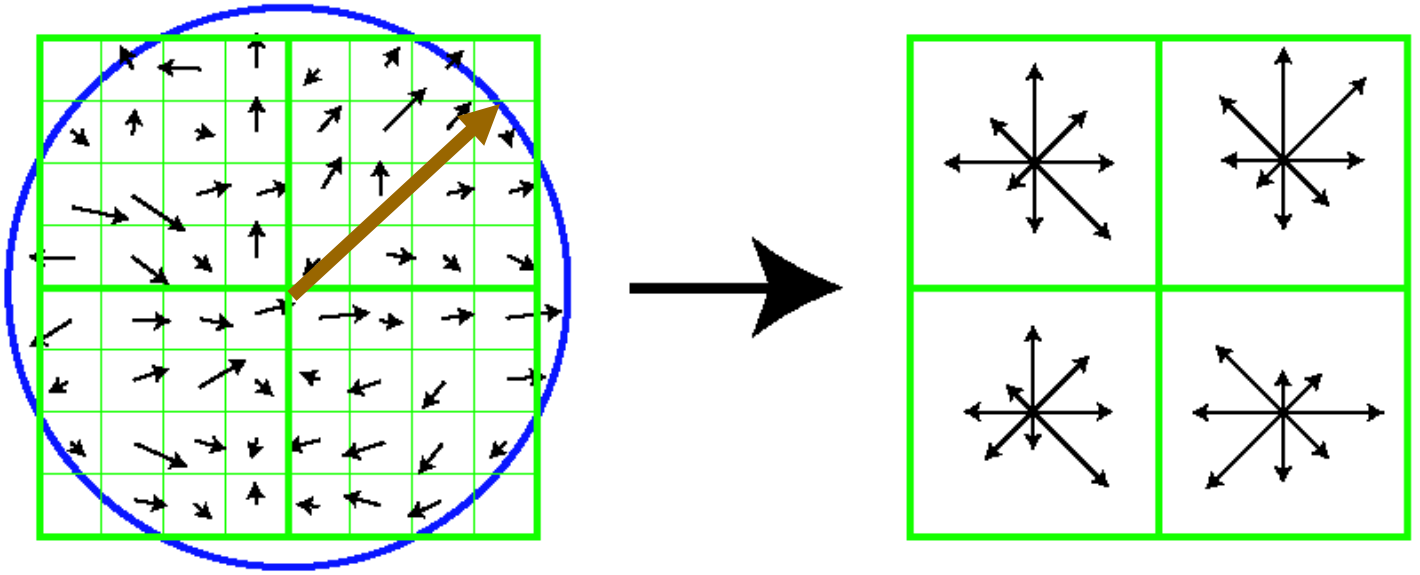
Normalization

- Rotate the window to standard orientation
- Scale the window size based on the scale at which the point was found.





SIFT Keypoint Descriptor (shown with 2 X 2 descriptors)

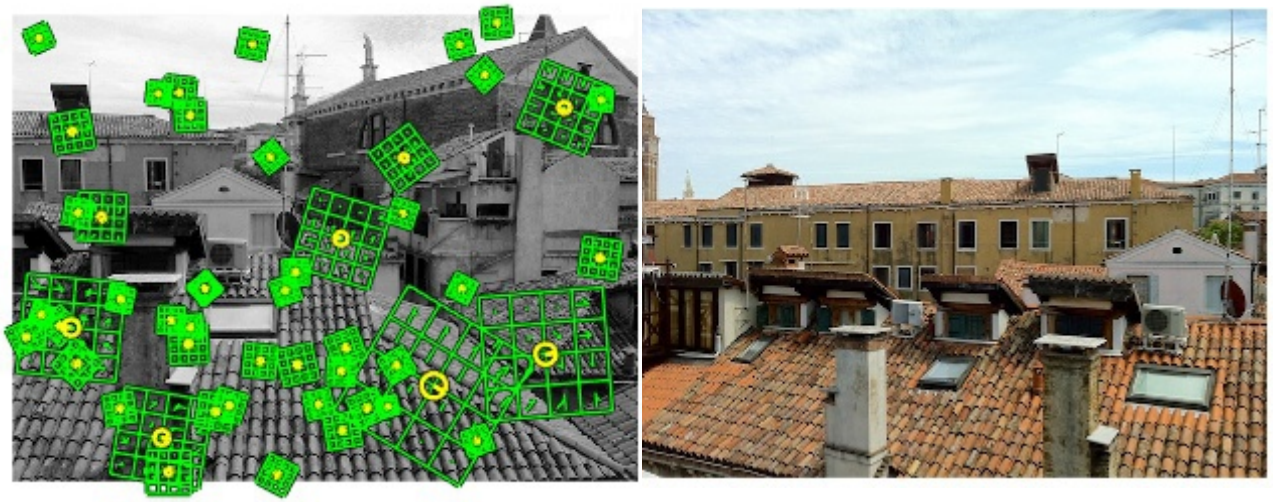


In implementation, 4x4 arrays of 8 bin histogram are used, a total of 128 features for one keypoint.

SIFT for Correspondence

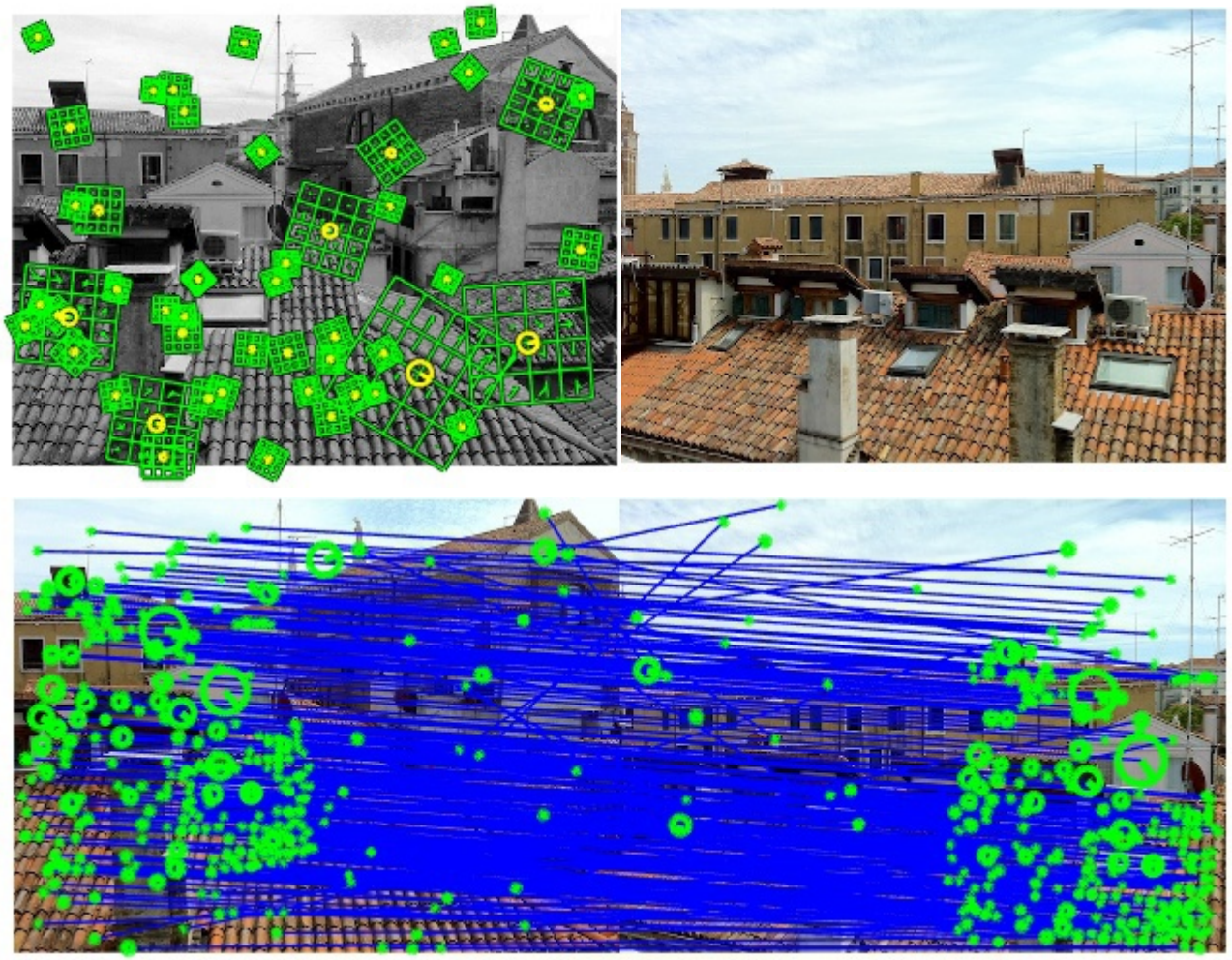
[Code and tutorial: https://www.vlfeat.org/overview/sift.html]

SIFT for Correspondence



[Code and tutorial: https://www.vlfeat.org/overview/sift.html]

SIFT for Correspondence



[Code and tutorial: https://www.vlfeat.org/overview/sift.html]

Cases where SIFT does not work

- Strong illumination changes
- Large out-of-plane rotations
- Non-rigid deformations or articulations
- Semantic correspondence