

Object Recognition

Seminar

Rita Osadchy

So what does object recognition involve?



Verification: is that a bus?



Detection: locate the cars in the image



Identification: is that a picture of Mao?



Object categorization



sky

building

flag

face

banner

wall

street lamp

bus

bus

cars

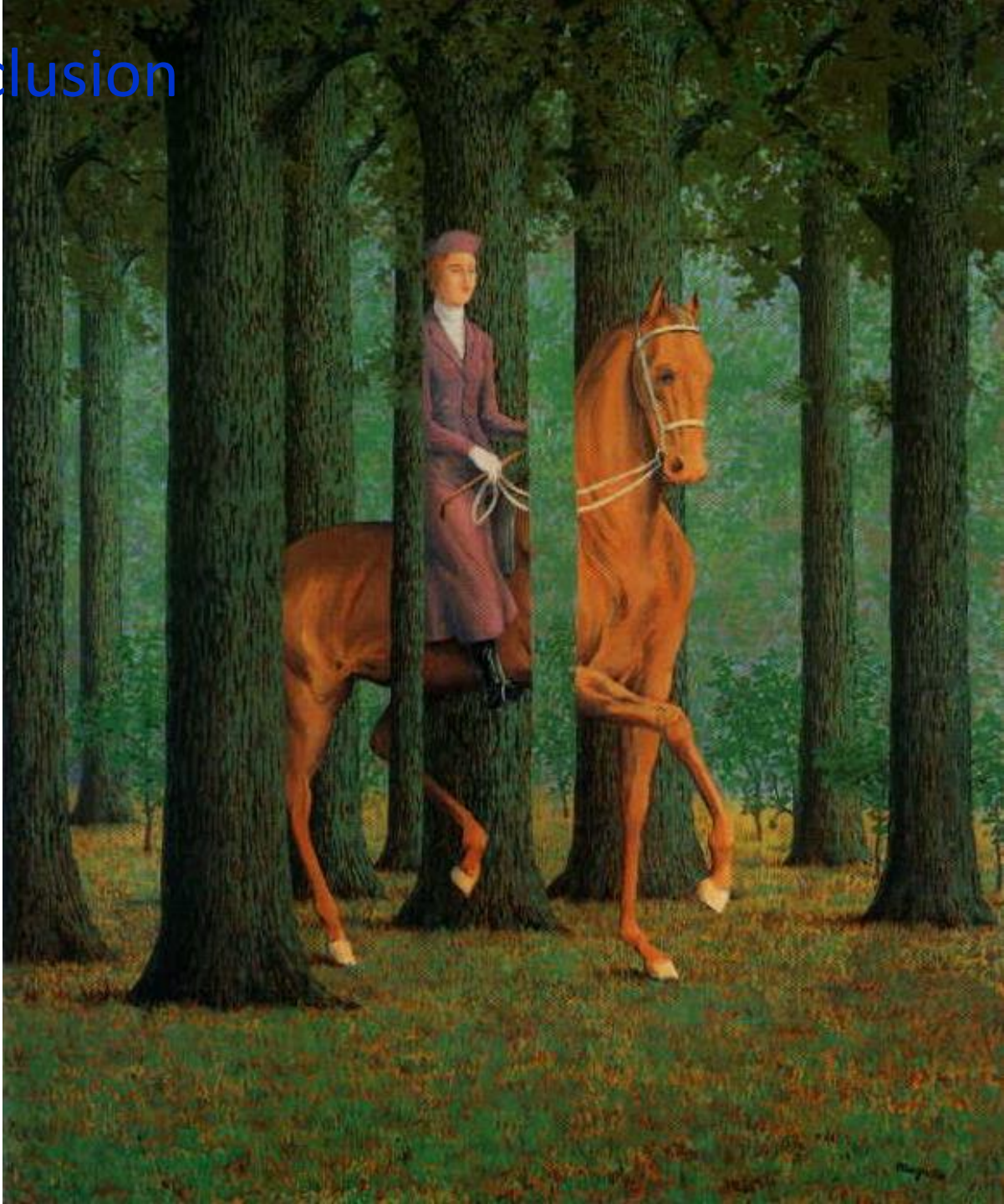
Challenges 1: view point variation



Challenges 2: illumination



Challenges 3: occlusion

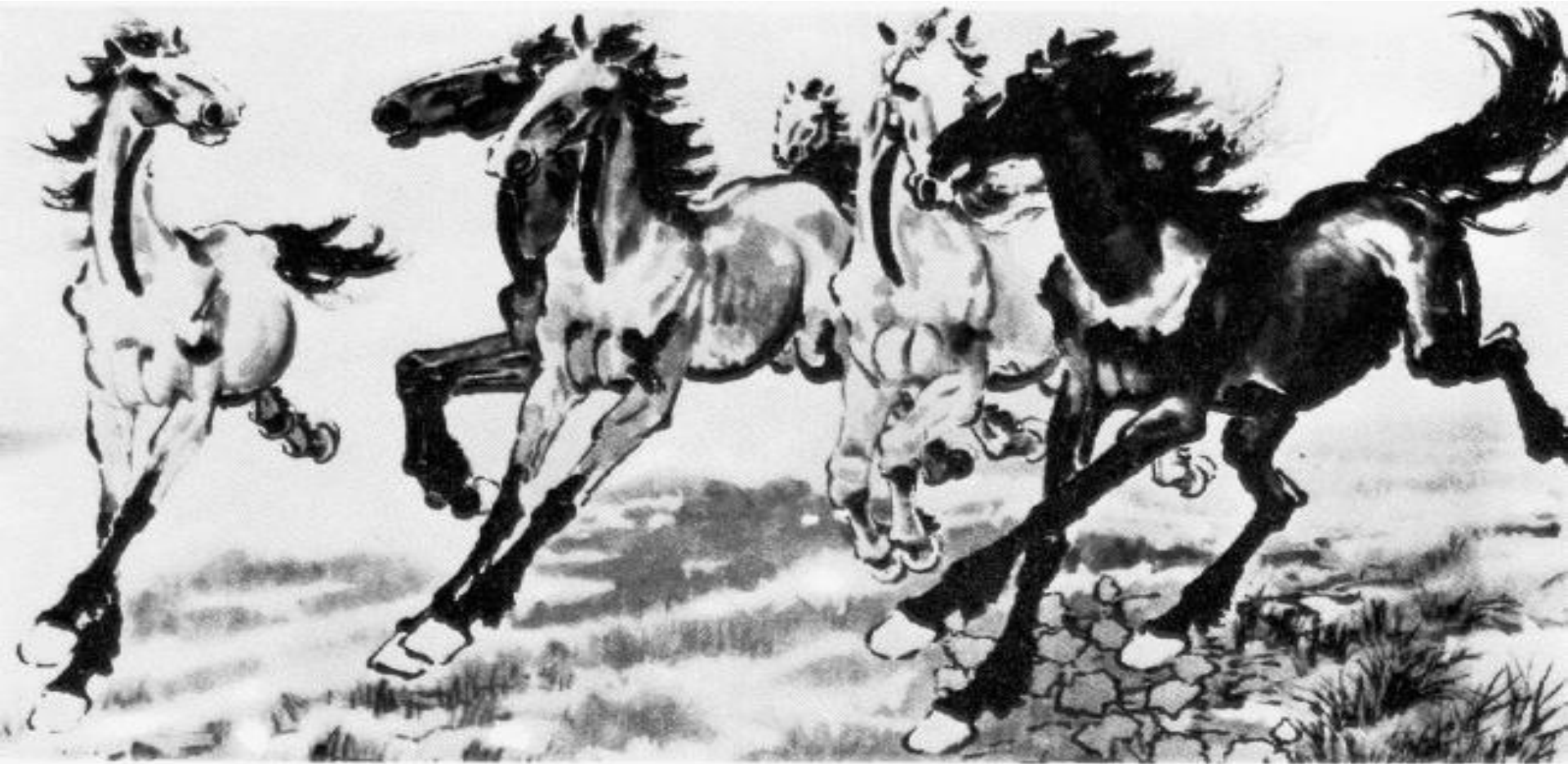


Magritte, 1957

Challenges 4: scale

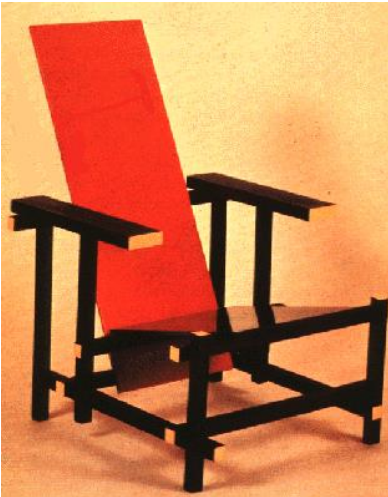


Challenges 5: deformation



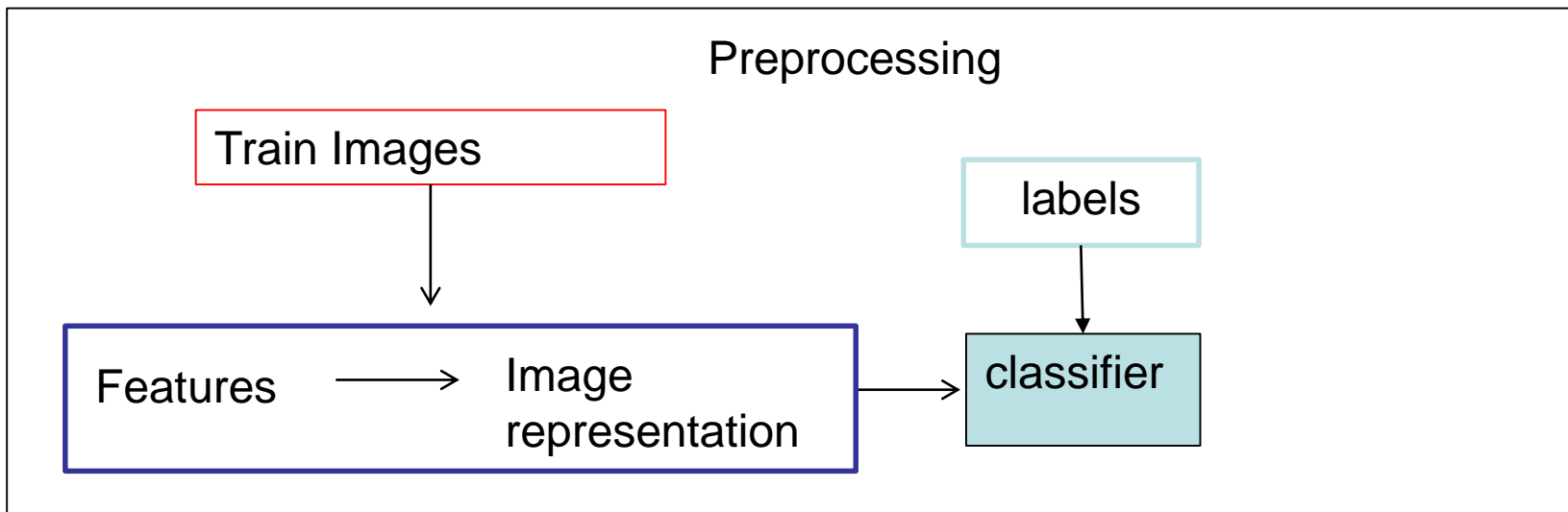
Xu, Beihong 1943

Challenges 7: intra-class variation

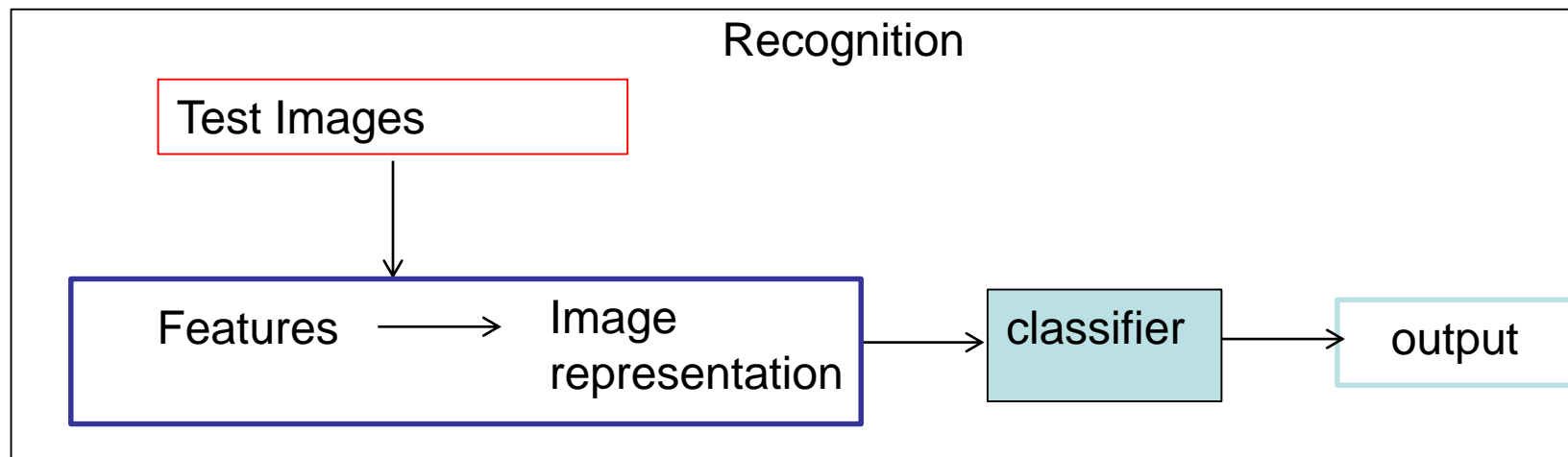


Recognition Steps

Preprocessing



Recognition

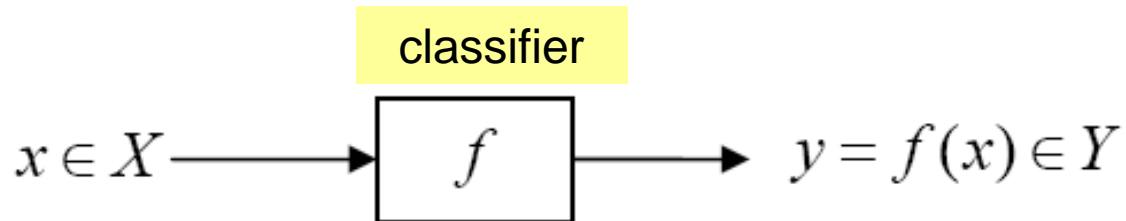


Classification in Machine Learning

- A **classifier** is a function or an algorithm that maps every possible input (from a legal set of inputs) to a finite set of categories.
- X – **input space**, $x \in X$ **sample** from an input space.
- A typical input space is high-dimensional, for example $x = \{x_1, \dots, x_d\} \in R^d$, $d > 1$. We also call x a **feature vector**.
- Ω is a **finite set of categories** to which the input samples belong: $\Omega = \{1, 2, \dots, C\}$.
- $w_i \in \Omega$ are called **labels**.

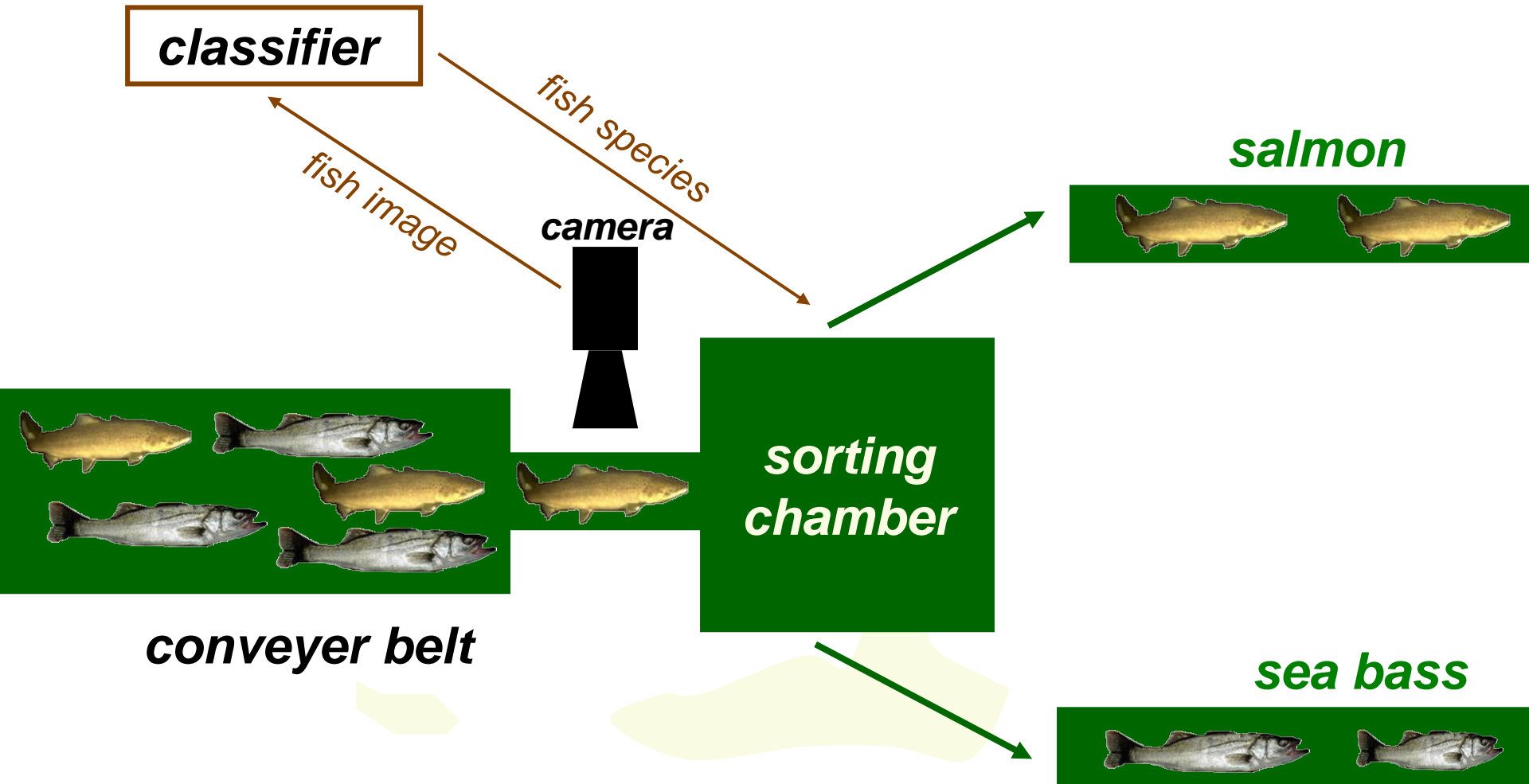
Definition of Classification

- Y is a finite **set of decisions** – the **output set** of the classifier.
- A classifier is a function $f : X \rightarrow Y$



- Classification is also called **Pattern Recognition**.

Toy Application: fish sorting



How to design a PR system?

- **Collect data** and classify by hand



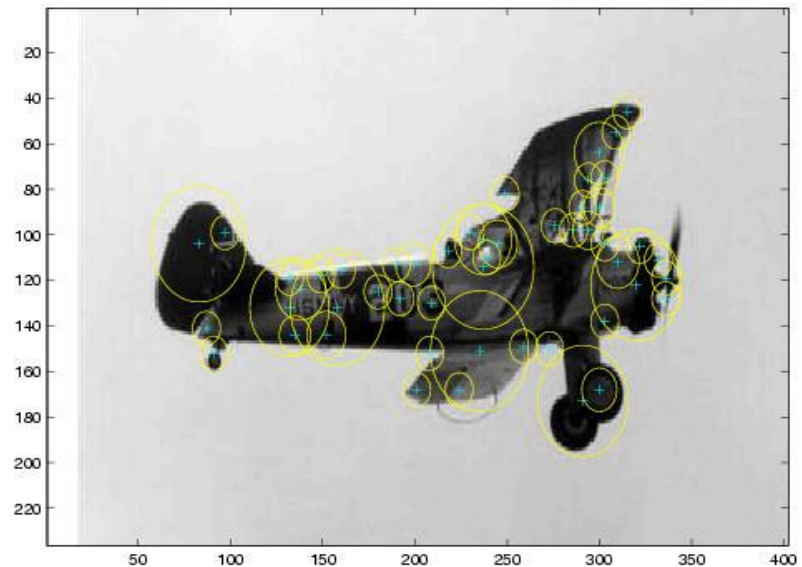
- **Preprocess** by segmenting fish from background



- **Extract** possibly discriminating **features**
 - length, lightness, width, number of fins, etc.
- **Classifier design**
 - Choose model
 - **Train classifier** on part of collected data (**training** data)
- **Test classifier** on the rest of collected data (**test** data)
i.e. the data not used for training
 - Should classify **new** data (new fish images) well

Interest Point Detectors

- Basic requirements:
 - Sparse
 - Informative
 - Repeatable
- Invariance
 - Rotation
 - Scale (Similarity)
 - Affine

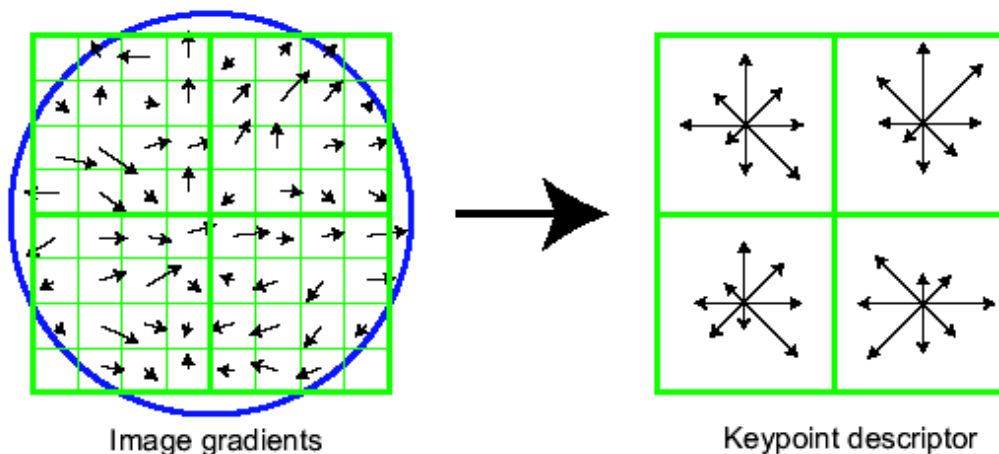


Representation of appearance: Local Descriptors

- Invariance
 - Rotation
 - Scale
 - Affine
- Insensitive to small deformations
- Illumination invariance
 - Normalize out

SIFT – Scale Invariant Feature Transform

- Descriptor overview:
 - Determine **scale** (by maximizing DoG in scale and in space), **local orientation** (as the dominant gradient direction). Use this scale and orientation to make all further computations invariant to scale and rotation.
 - Compute **gradient orientation histograms** of several small windows (128 values for each point)
 - Normalize the descriptor to make it invariant to intensity change

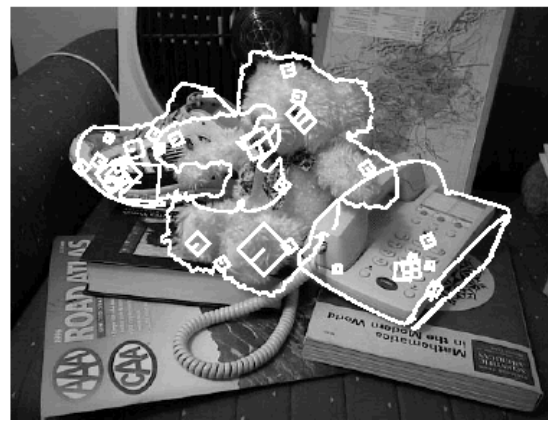
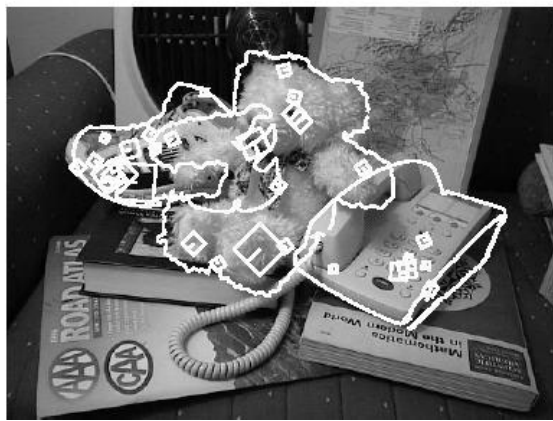


Recognizing Specific Objects

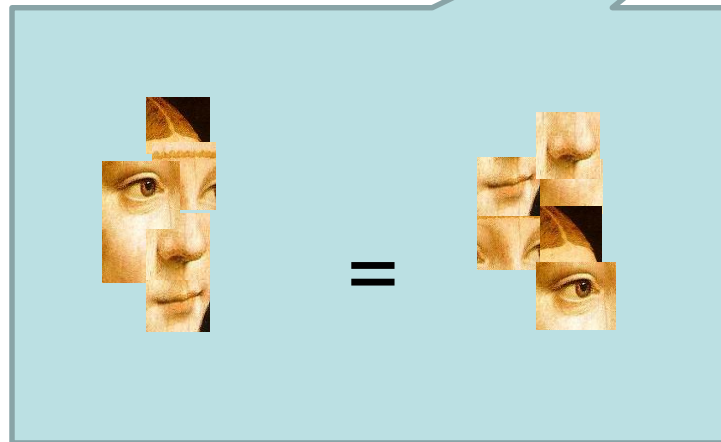
Learned models of local features, and got object outline from



Objects may then be found under occlusion and 3D rotation



Bag of Features



Bag of Features



Pros: fairly flexible and computationally efficient

Cons: problems with large clutter



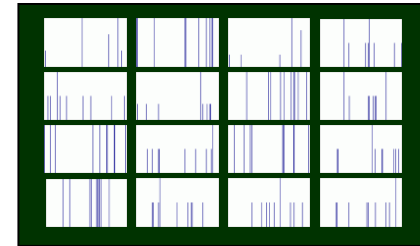
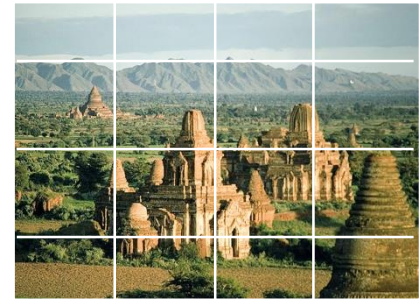
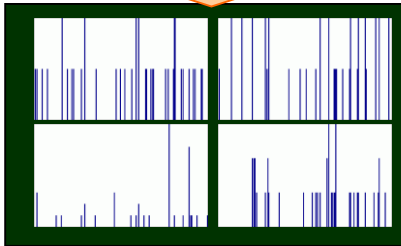
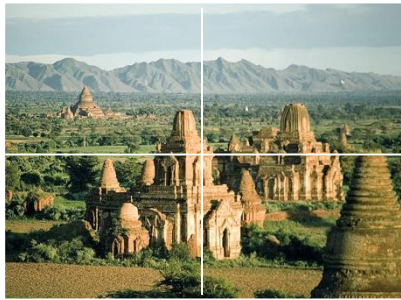
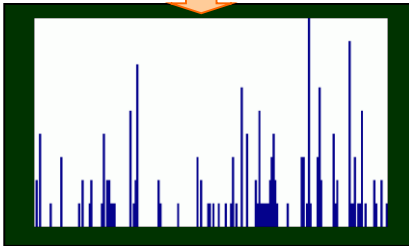
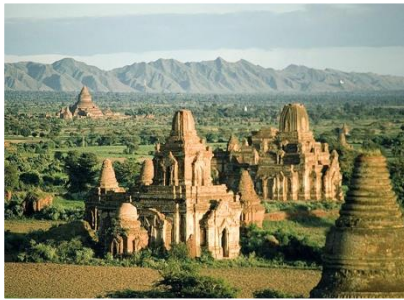
Different objects, but Similar representations;



Similar objects, different representations;

Beyond Bags of Features

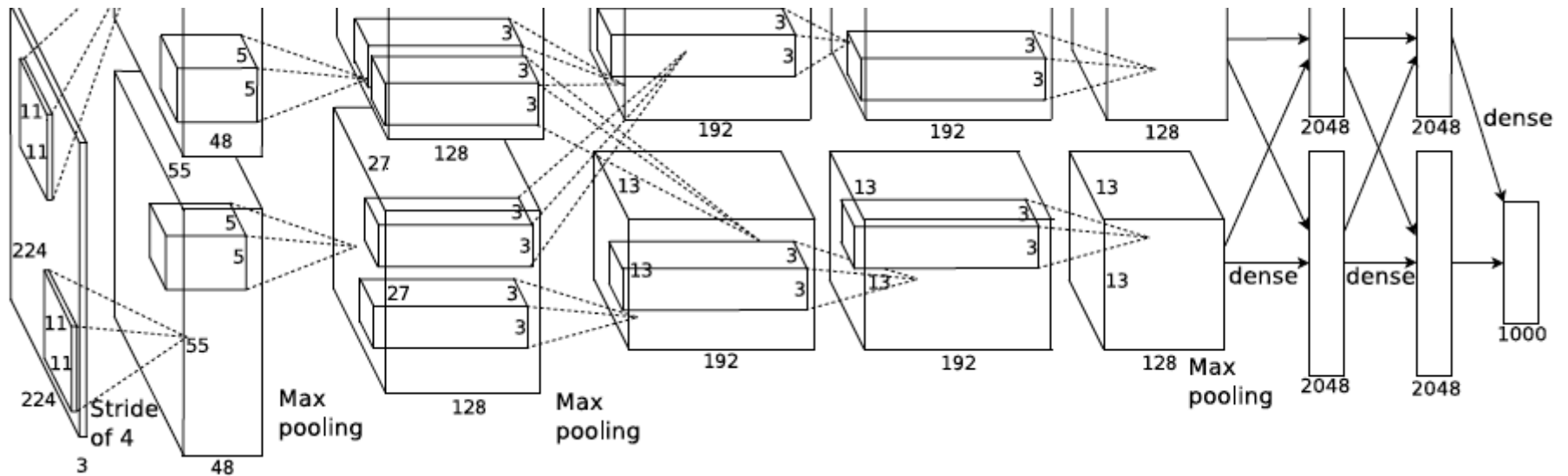
- Computing bags of features on sub-windows of the whole image.



Convolutional Neural Networks

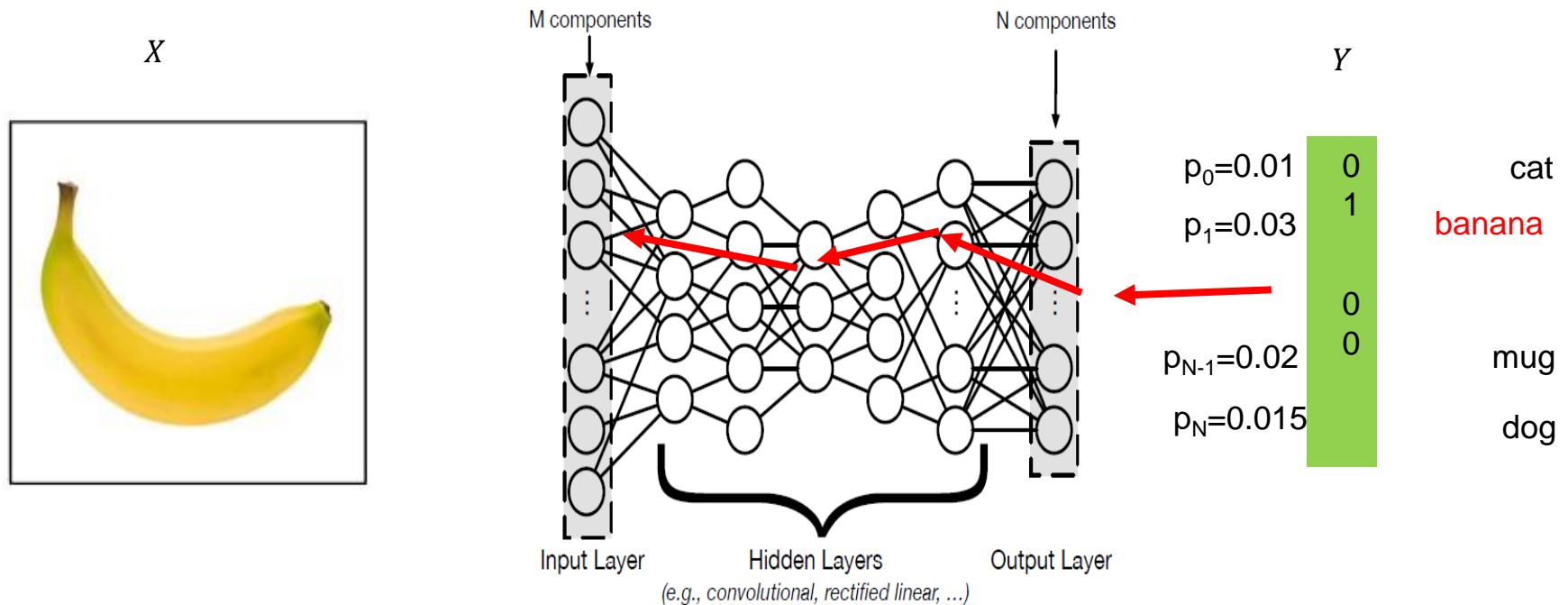
- Learn all in one deep architecture:
 - low level features
 - high level representations
 - context
 - classifiers
- Efficient Classification
- Efficient Detection
- Scalable to very large sets and large number of categories

Convolutional Neural Networks



Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks

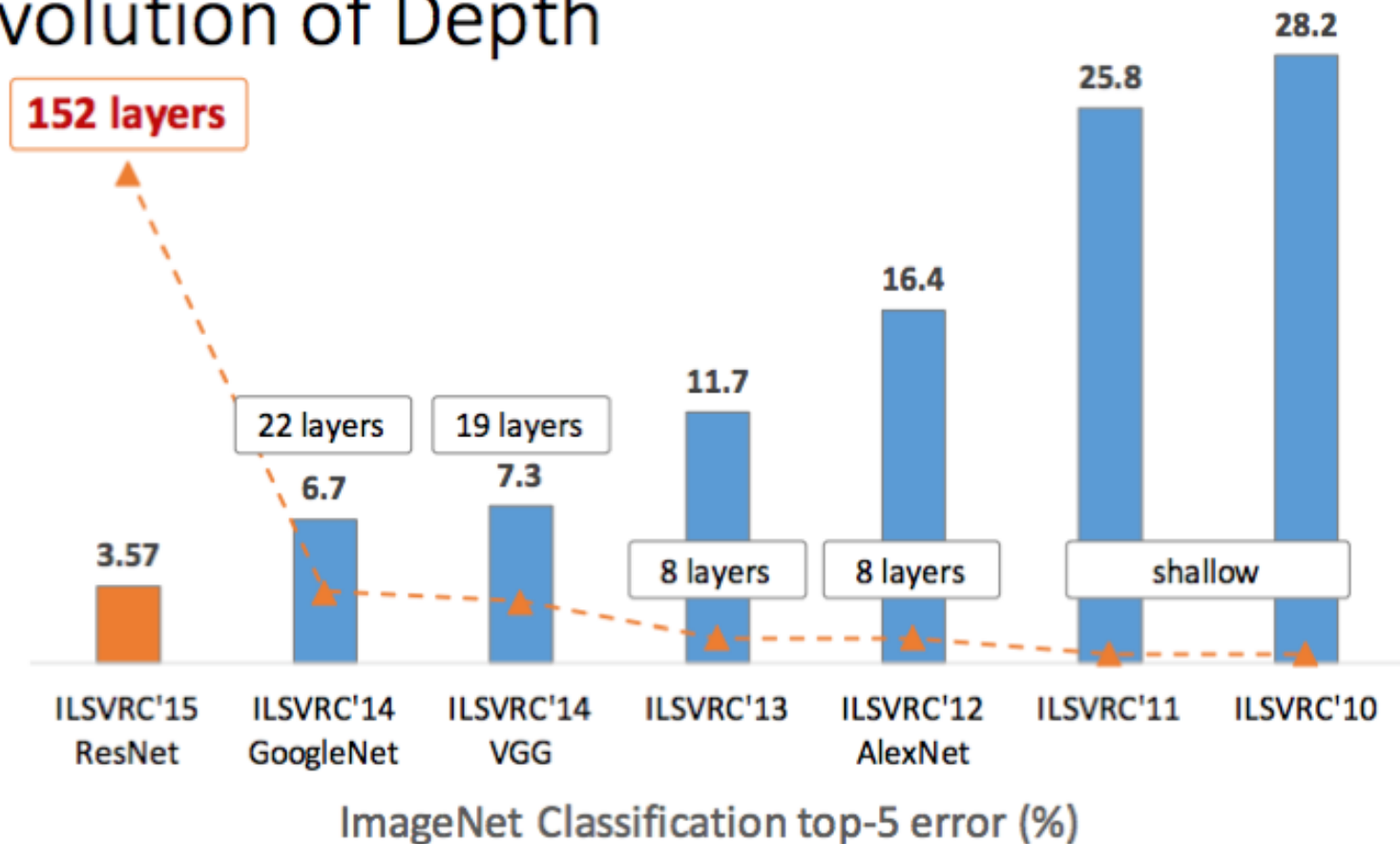
Training the Network



Compute the gradient with respect to the **parameters of the network.**

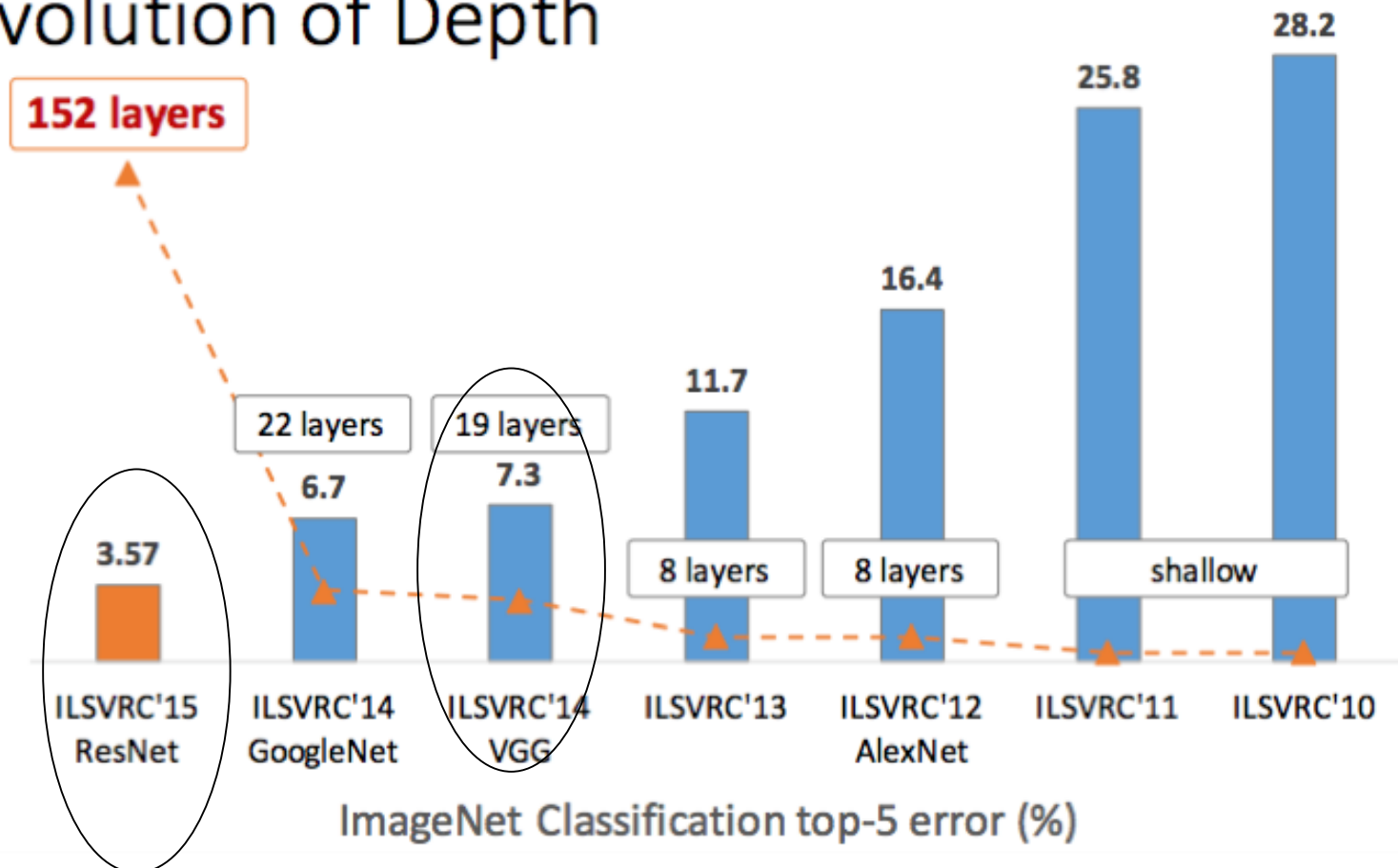
Very Deep Networks

Revolution of Depth



Very Deep Networks

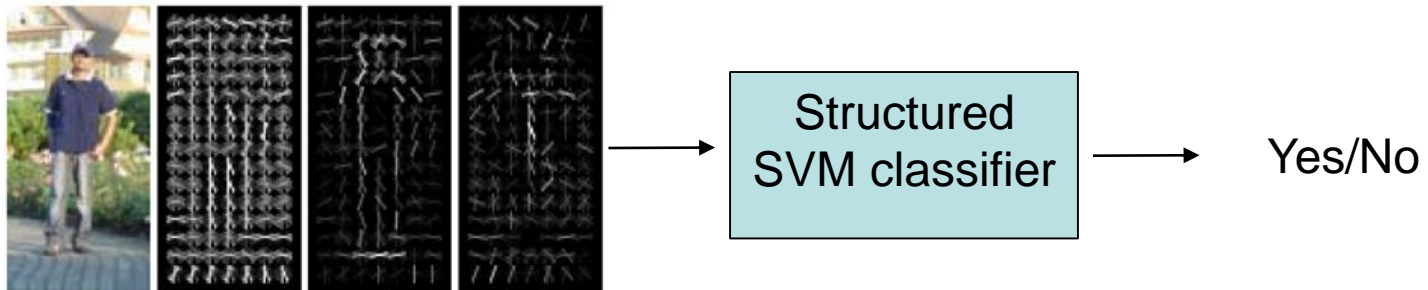
Revolution of Depth



- K. Simonyan and A. Zisserman Very Deep Convolutional Networks for Large-Scale Image Recognition
- K. He, X. Zhang, S. Ren, and J. Sun: Deep Residual Learning for Image Recognition

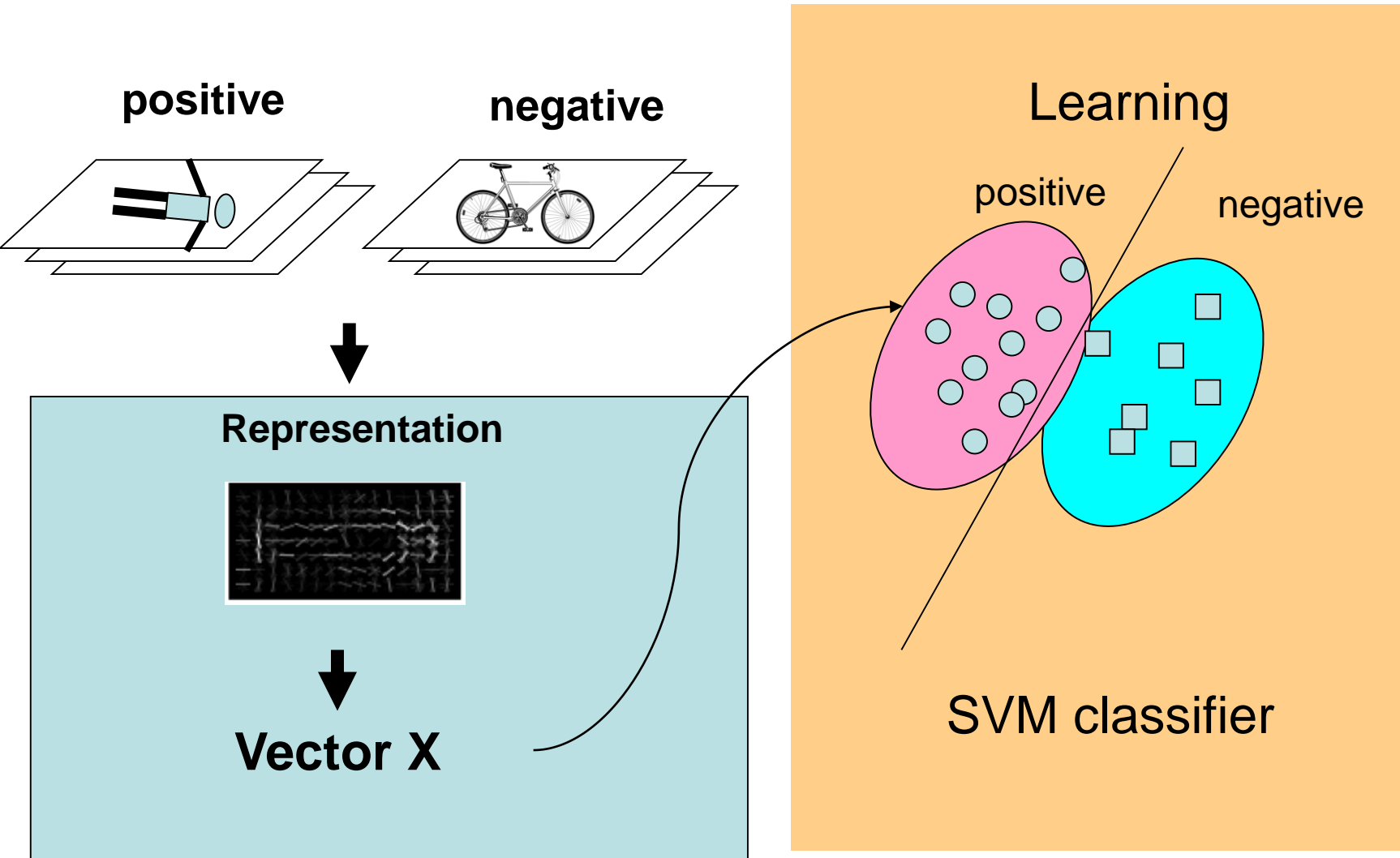
Detection as a binary decision

- Sliding window detection, detection as a binary decision problem.



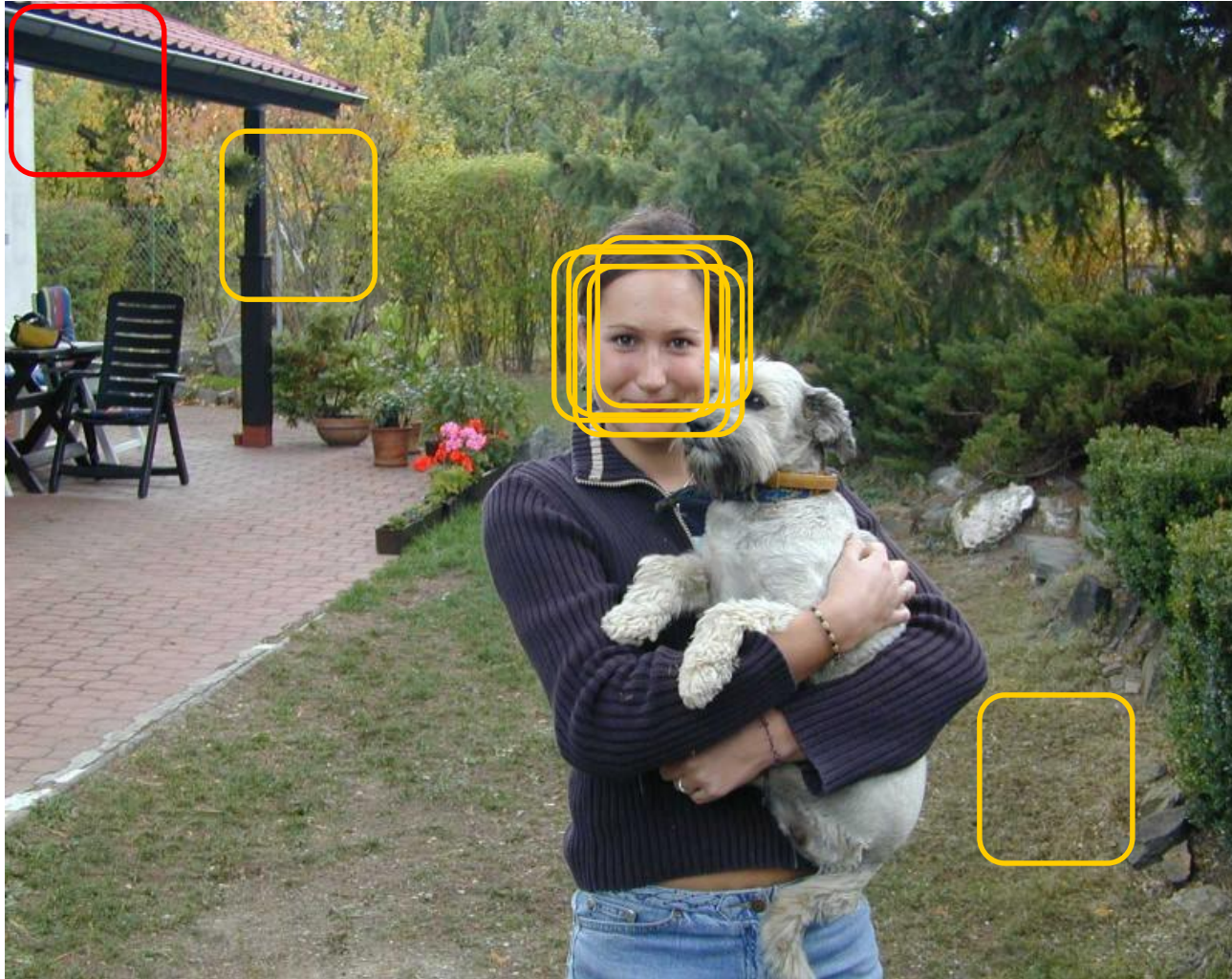
Histograms of Oriented
Gradients for Human
Detection

SVM classification

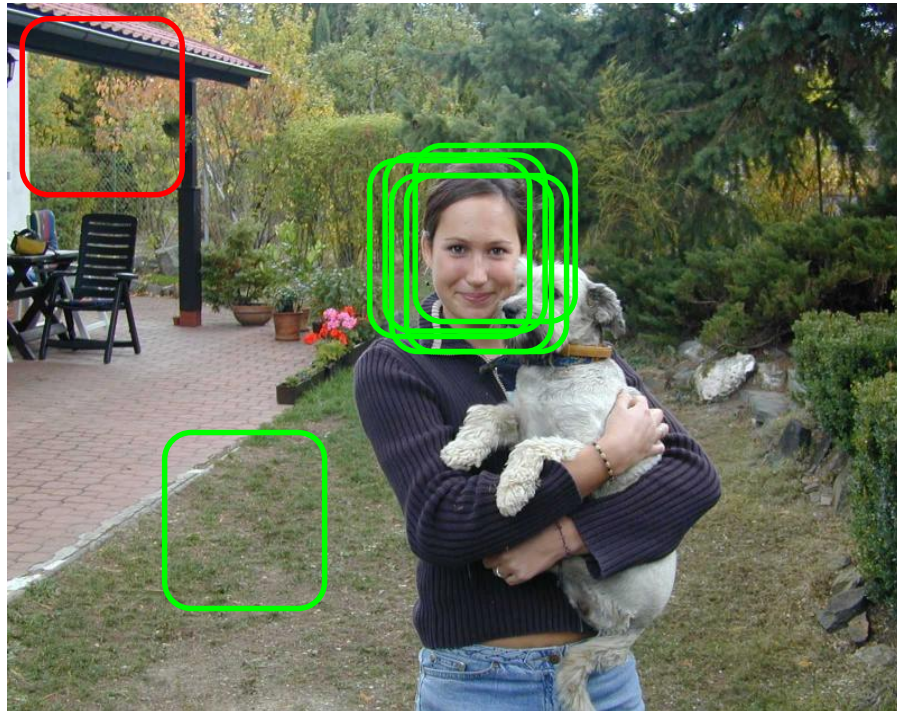


Detection

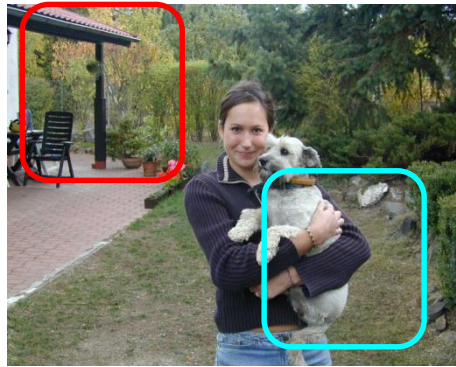
Apply classifier at Scale / position range to search over



Detection

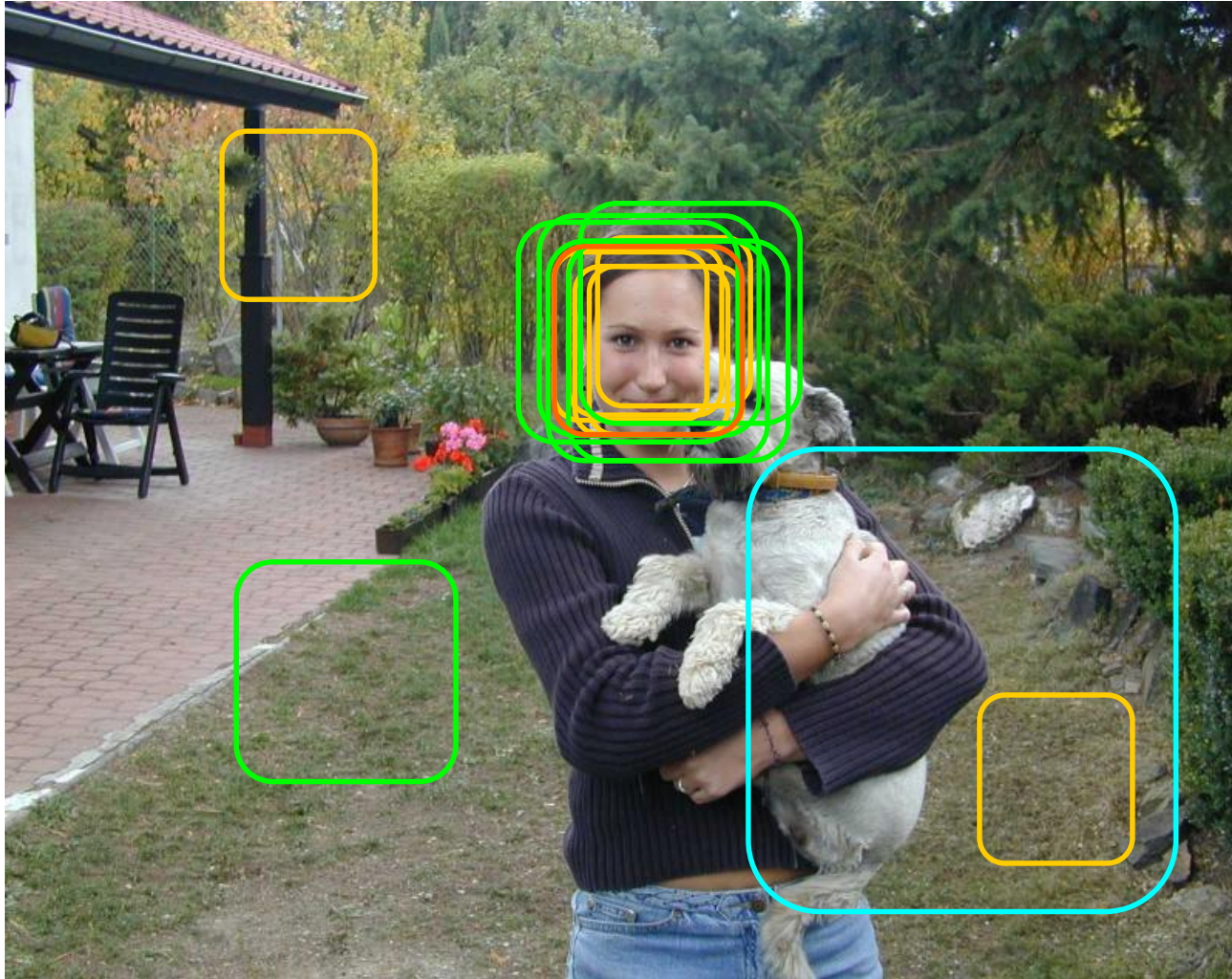


Detection



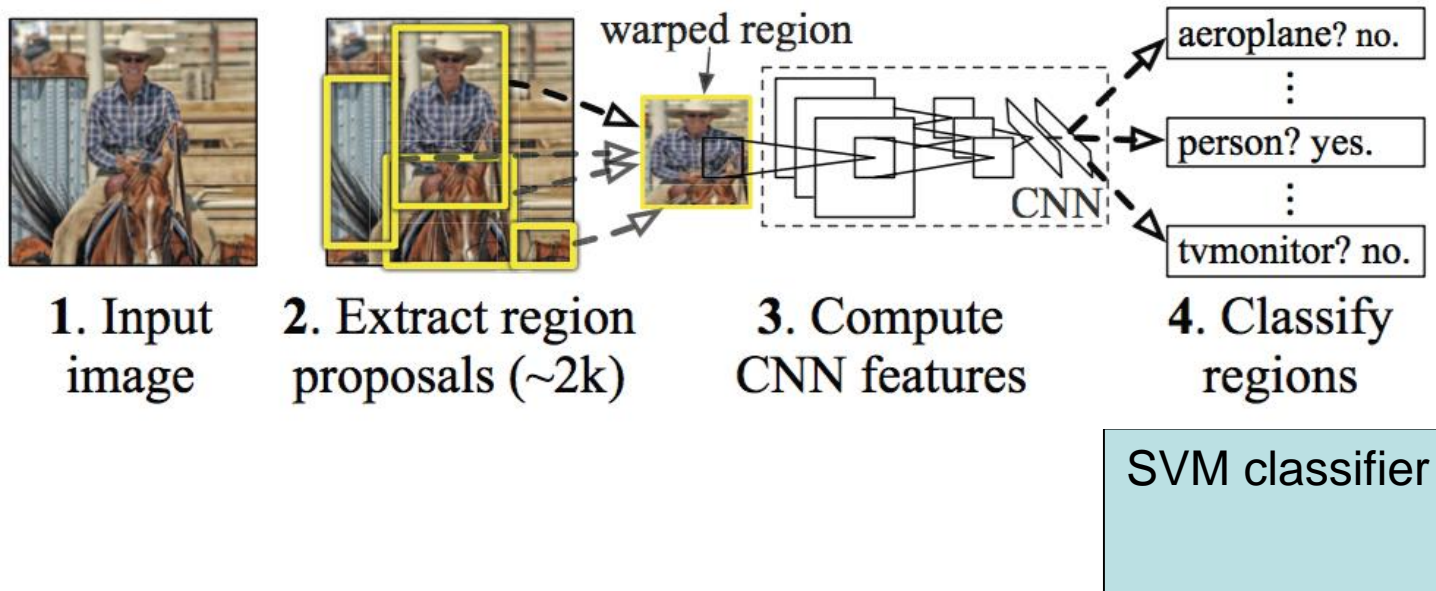
Detection

- Combine detection over space and scale.



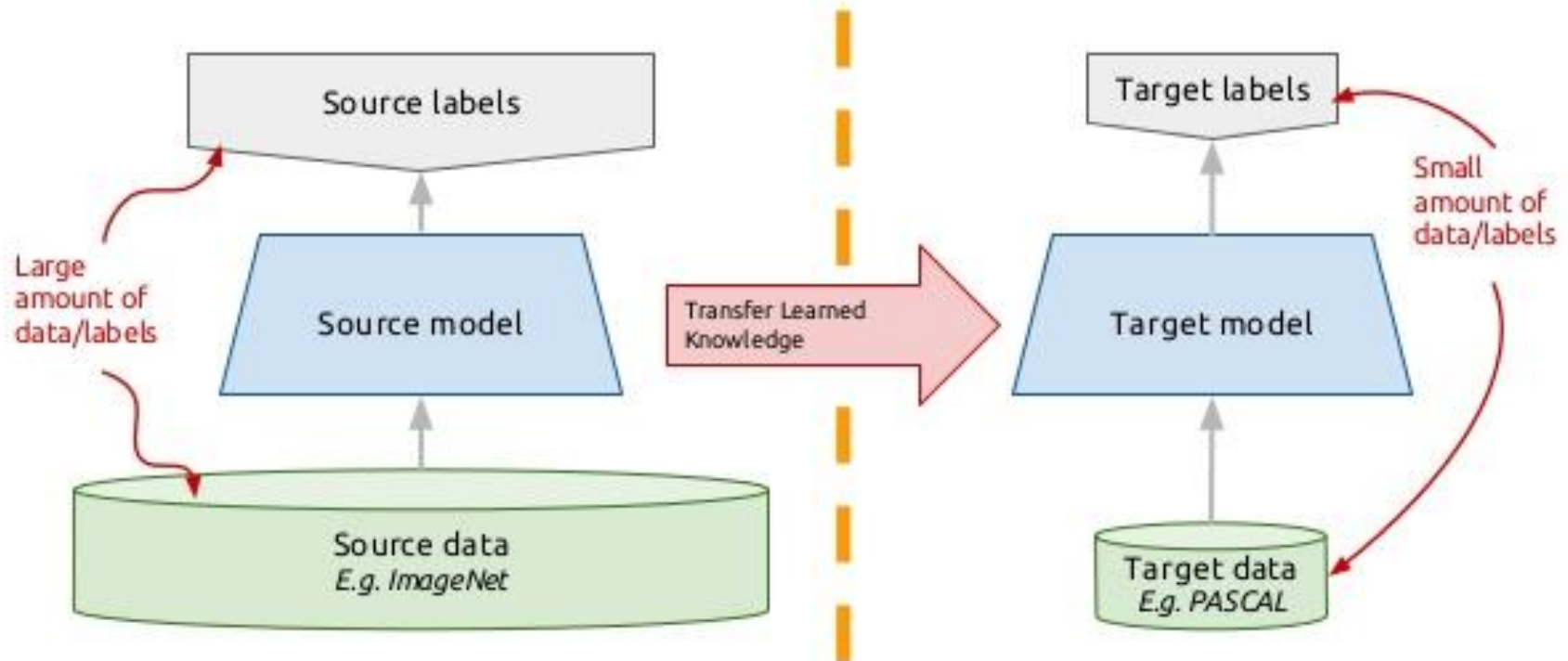
Deep Learning in Object Detection

R-CNN: *Regions with CNN features*









Transfer Learning

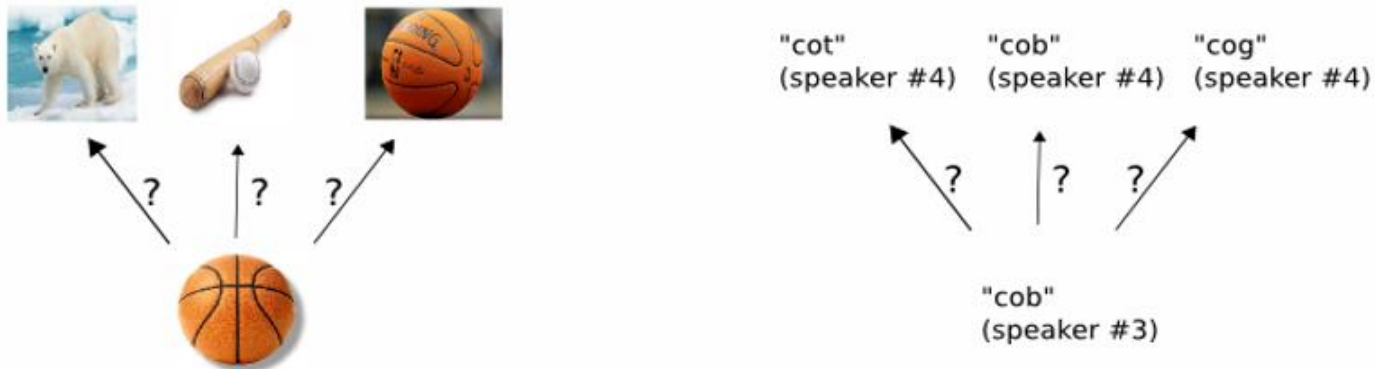
Transfer learning: idea



One Shot Learning

		same	"cow" (speaker #1)	"cow" (speaker #2)	same
		different	"cow" (speaker #1)	"cat" (speaker #2)	different
		same	"can" (speaker #1)	"can" (speaker #2)	same
		different	"can" (speaker #1)	"cab" (speaker #2)	different

Verification tasks (training)



One-shot tasks (test)

Describing Objects with Attributes

- Shift the goal of recognition from naming to describing

otter

black:	yes
white:	no
brown:	yes
stripes:	no
water:	yes
eats fish:	yes



Discover/detect new categories

polar bear

black:	no
white:	yes
brown:	no
stripes:	no
water:	yes
eats fish:	yes



Improvement



Attributes	Presence		Rating	
	walrus	polar bear	walrus	polar bear
Spot	no	no	less relevant	irrelevant
Blue	no	no	irrelevant	less relevant
Swim	yes	yes	highly relevant	relevant
Coastal	yes	yes	relevant	highly relevant

Face Verification Progress

Eigenface	0.6002
DeepFace	0.9130
Best automatic	0.9938
Best Human	0.9920

Deep Learning for Faces Survey

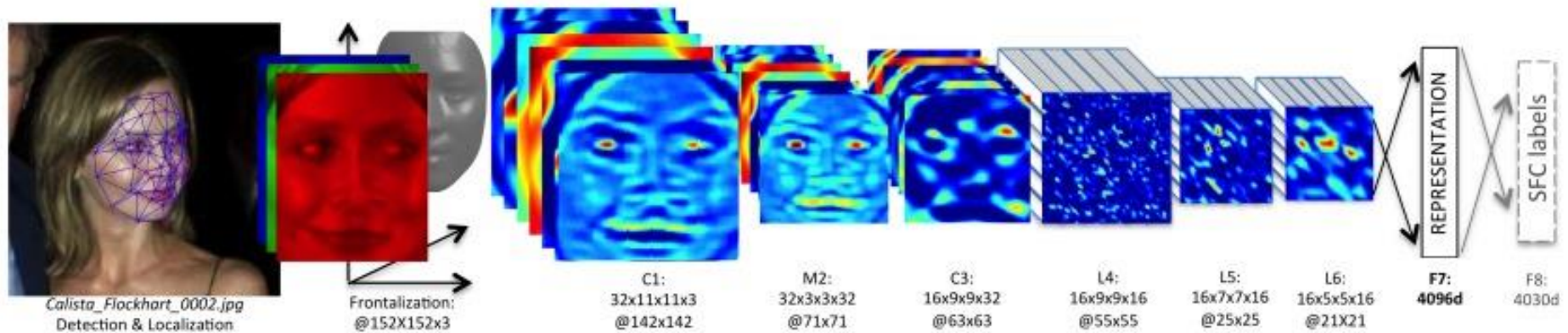
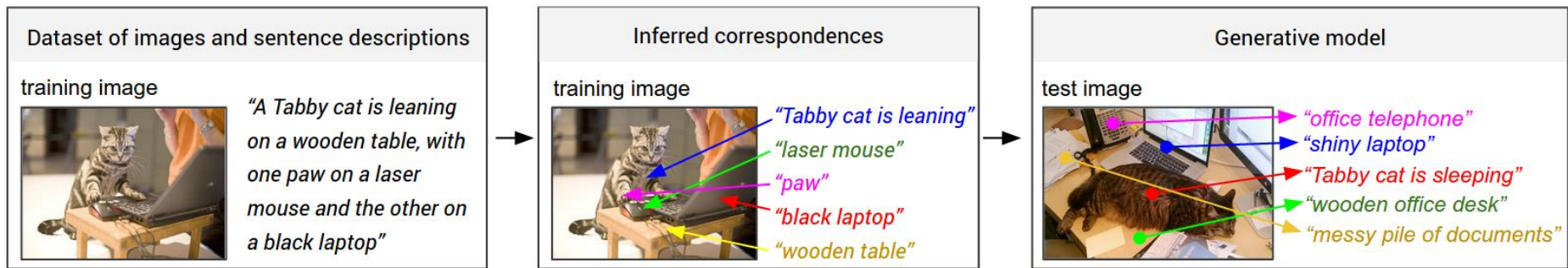


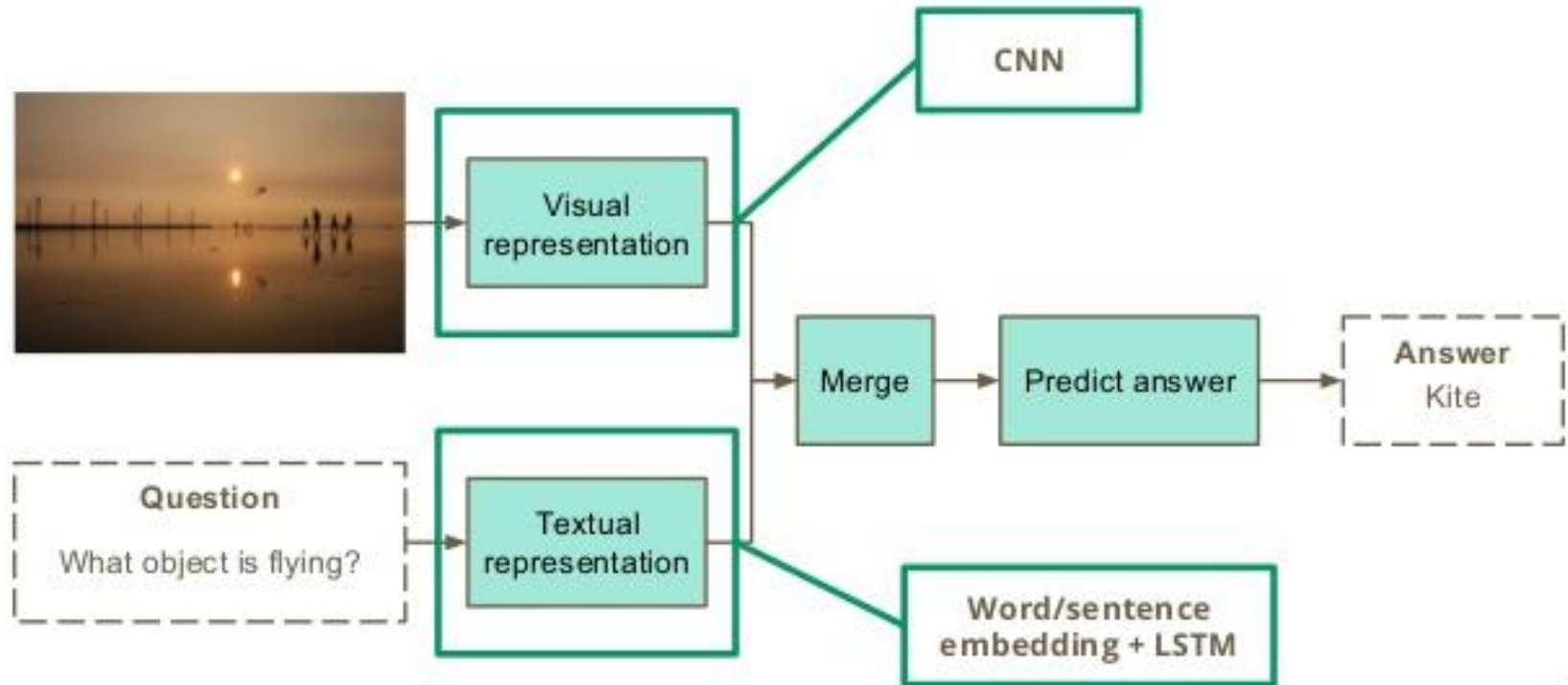
Image Descriptions (Captioning)



Uses CNNs and RNNs

VQA: Visual Question Answering

VQA: Common approach



Very Cool Results

Where is the child sitting?

fridge

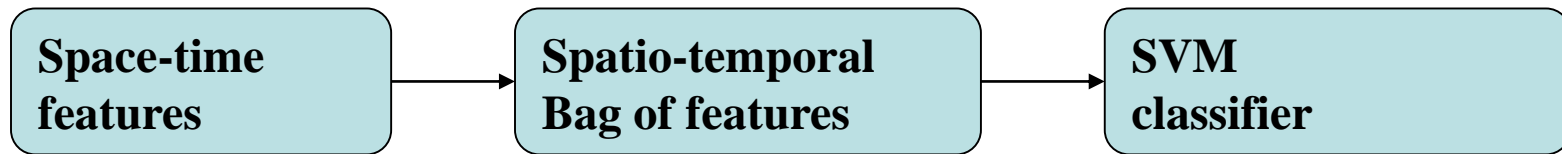


arms



Video Classification

- Older Approaches



AnswerPhone



GetOurCar



HandShake



HugPerson



Kiss



SitDown



SitUp



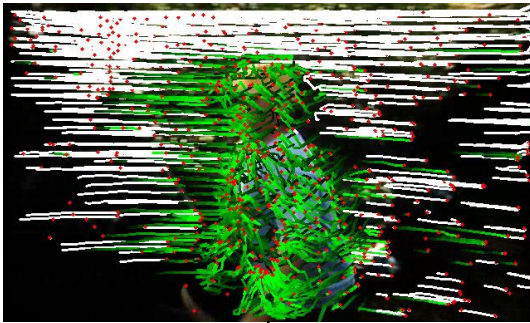
StandUp



Video Classification

- Shallow Approaches

trajectories
removed due to camera
motion in white

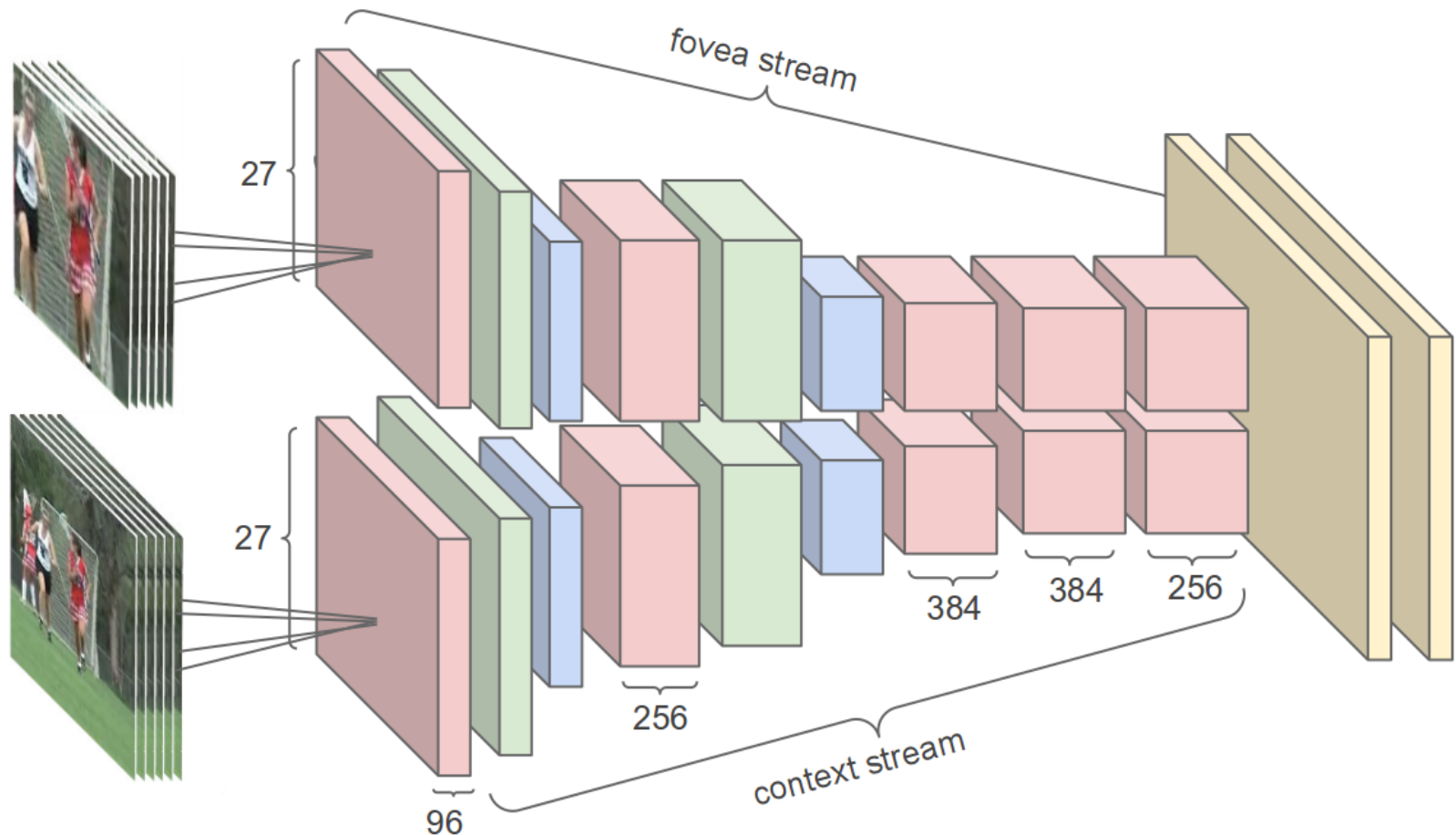


images of two consecutive
frames overlaid



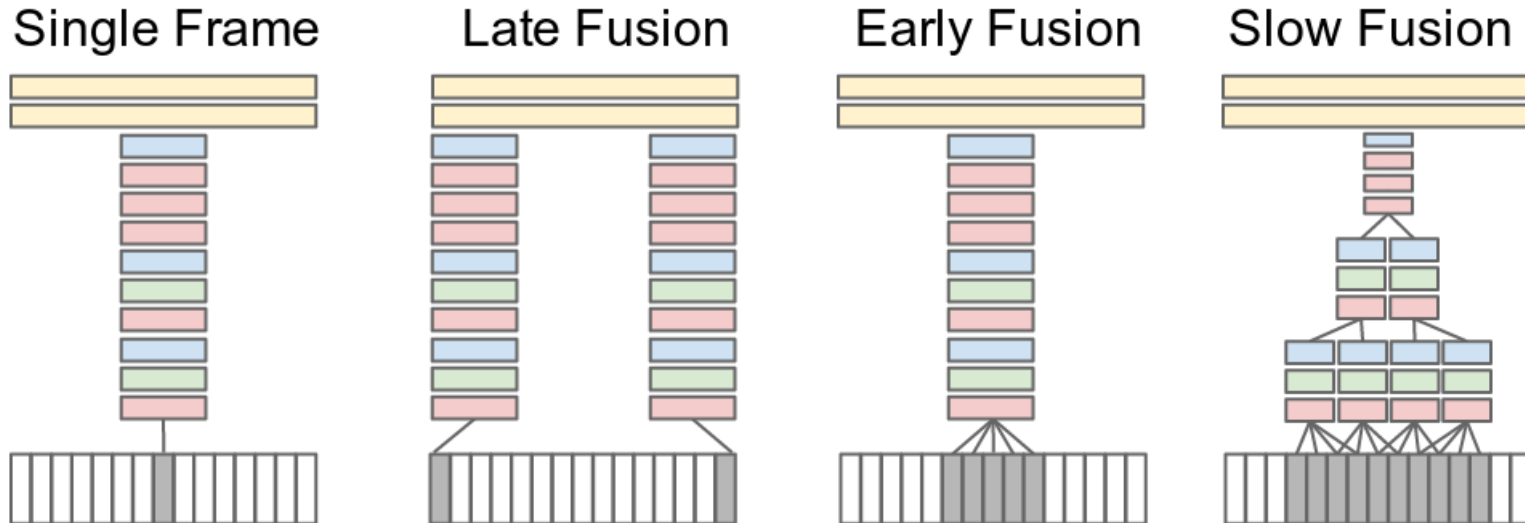
Video Classification

- Using CNN – Naïve Approach



Video Classification

- Using CNN – Naïve Approach



Temporal fusion

Video Classification

- Modern Approaches

