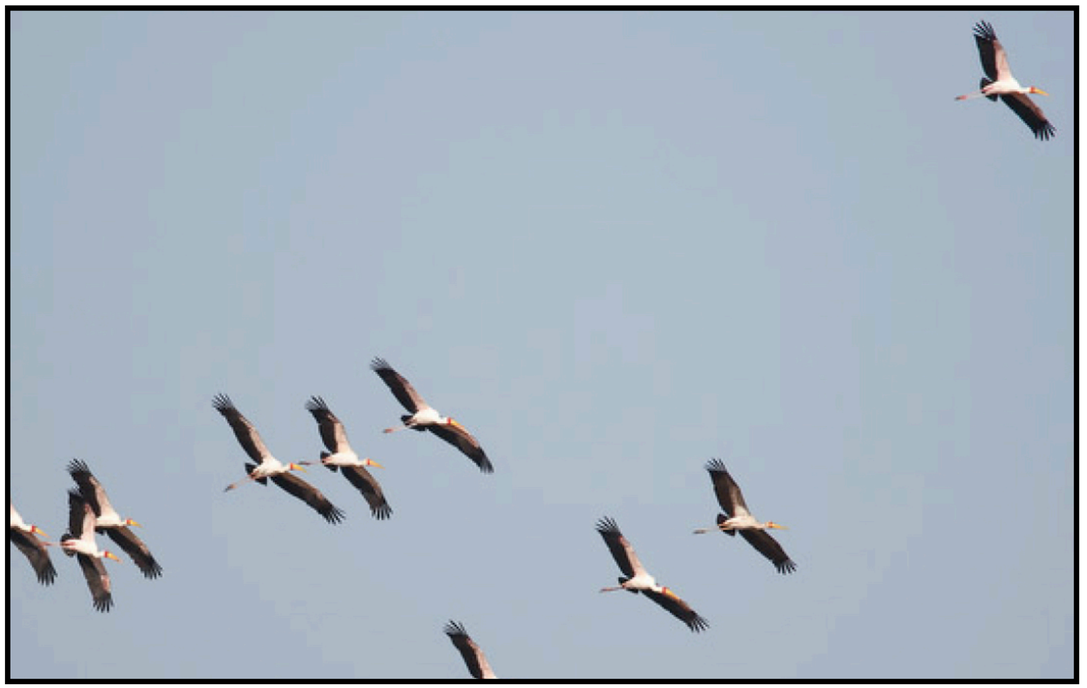


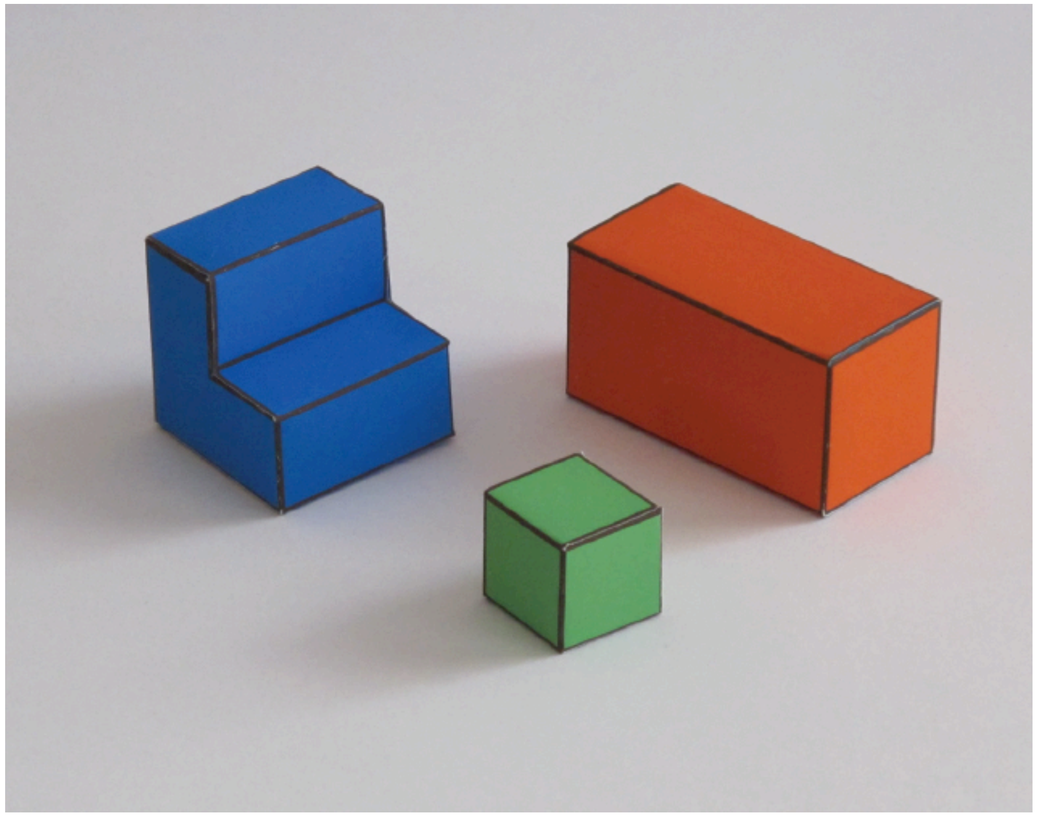
Object and scene understanding

Bill Freeman, Antonio Torralba, Phillip Isola

6.819 / 6.869

A simple goal

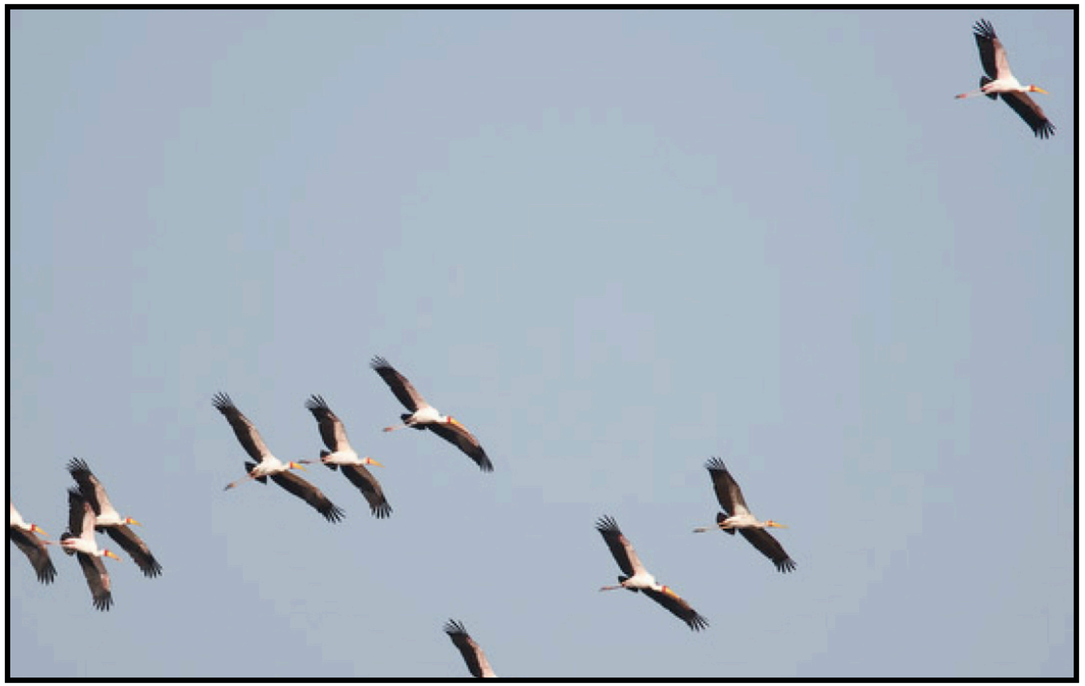
To recover the 3D structure of the world



We want to recover X(x,y), Y(x,y), Z(x,y) using as input I(x,y)

A tour of “core problems” in high-level computer vision

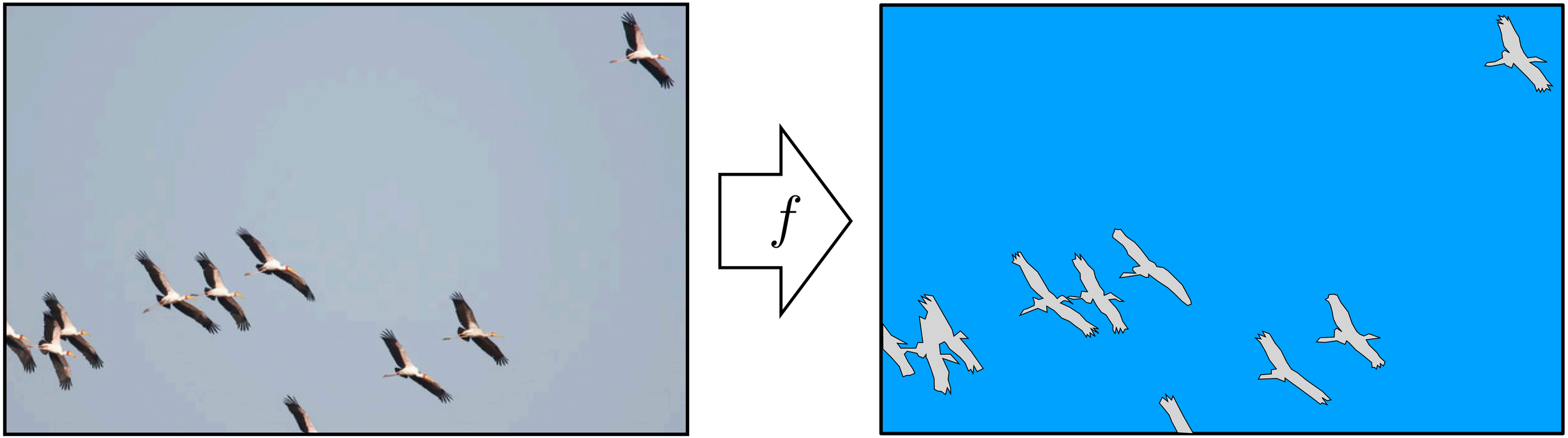
Image classification

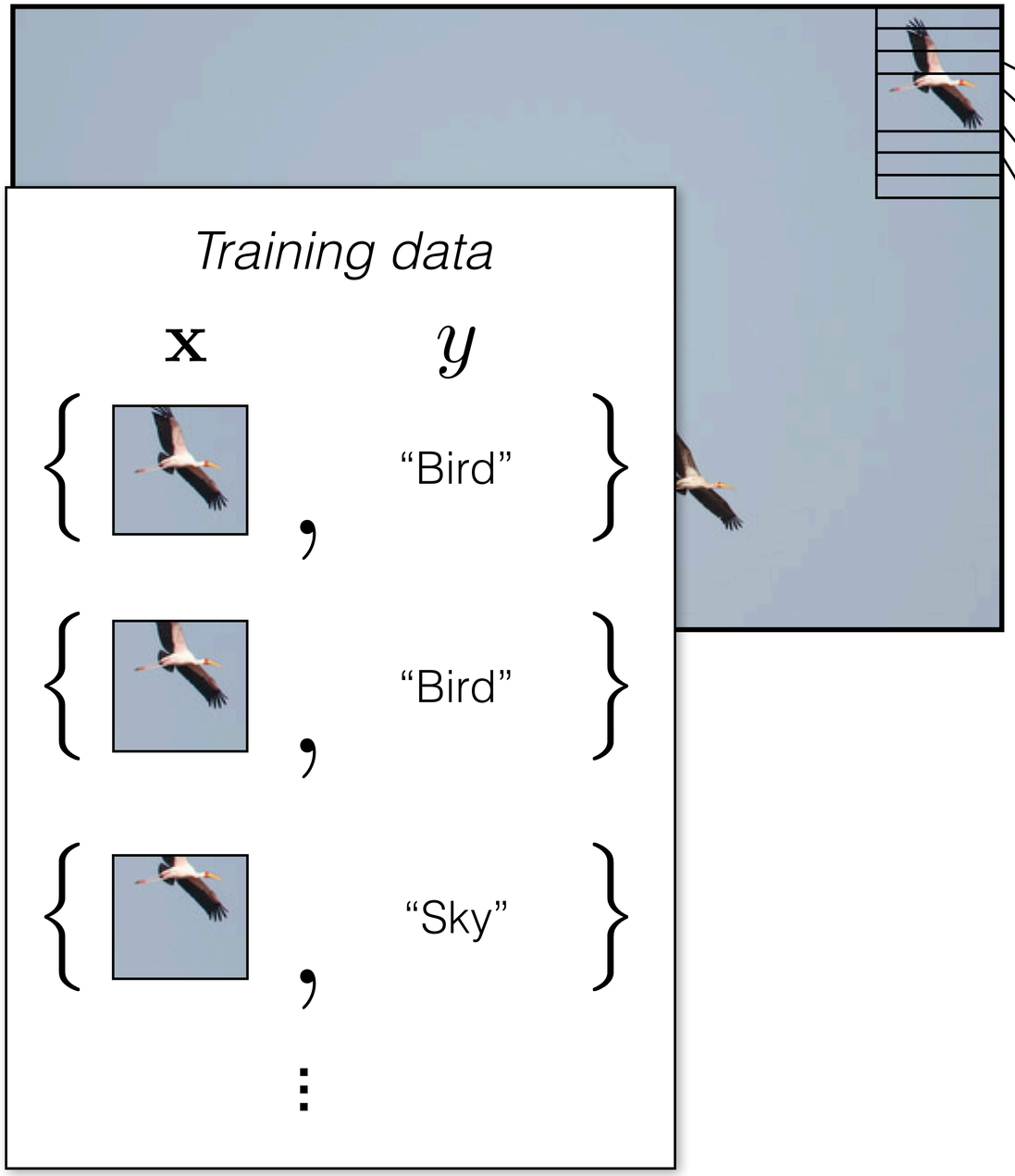


f

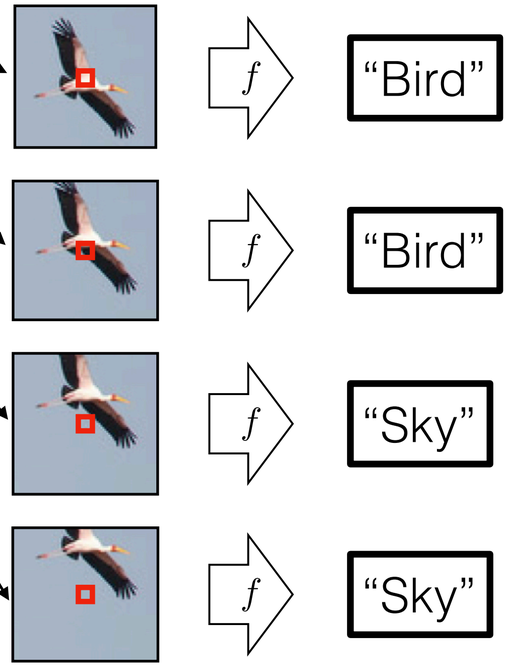
“Birds”

Semantic segmentation





What’s the object class of the center pixel?

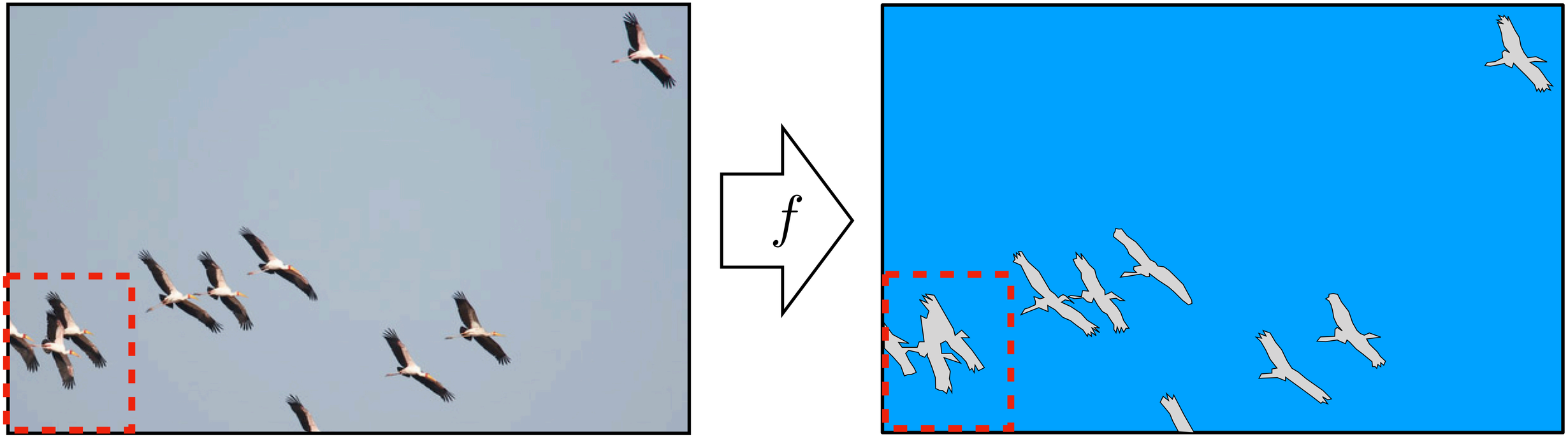


K-way classification problem

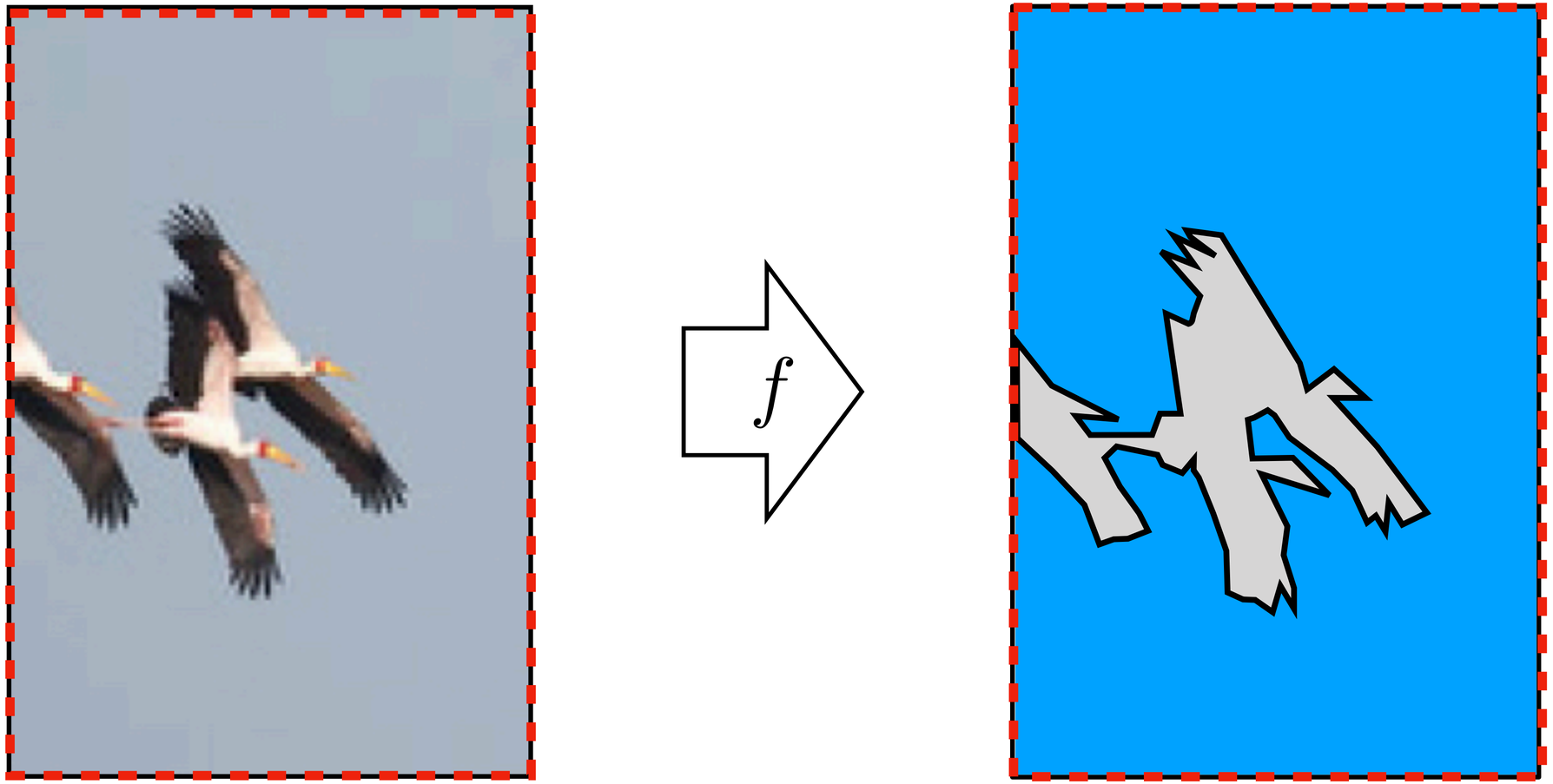
Solve with K-dimensional softmax regression:

$$f_\theta : X \to \mathbb{R}^K$$

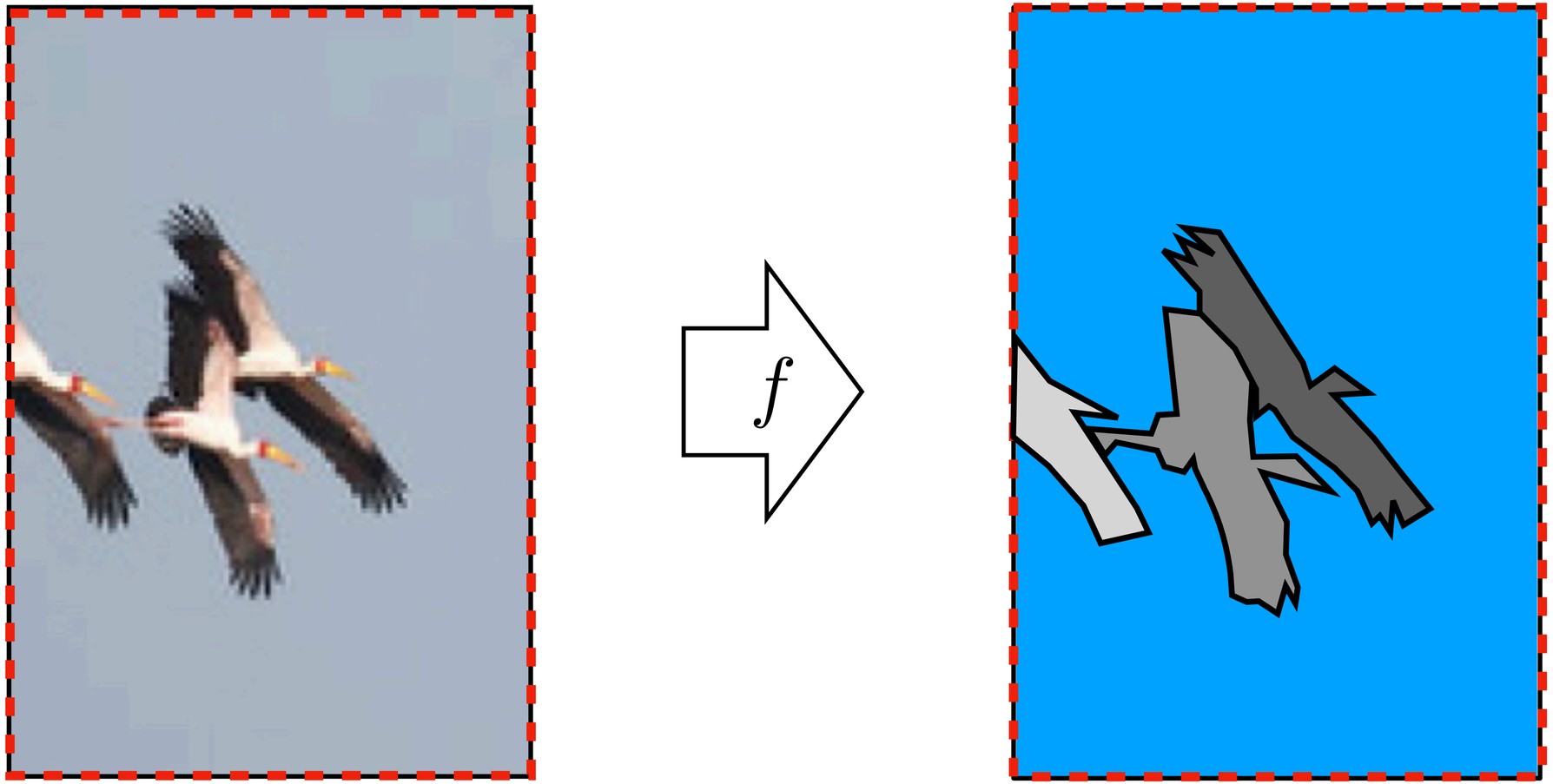
Semantic segmentation



Semantic segmentation



Instance segmentation

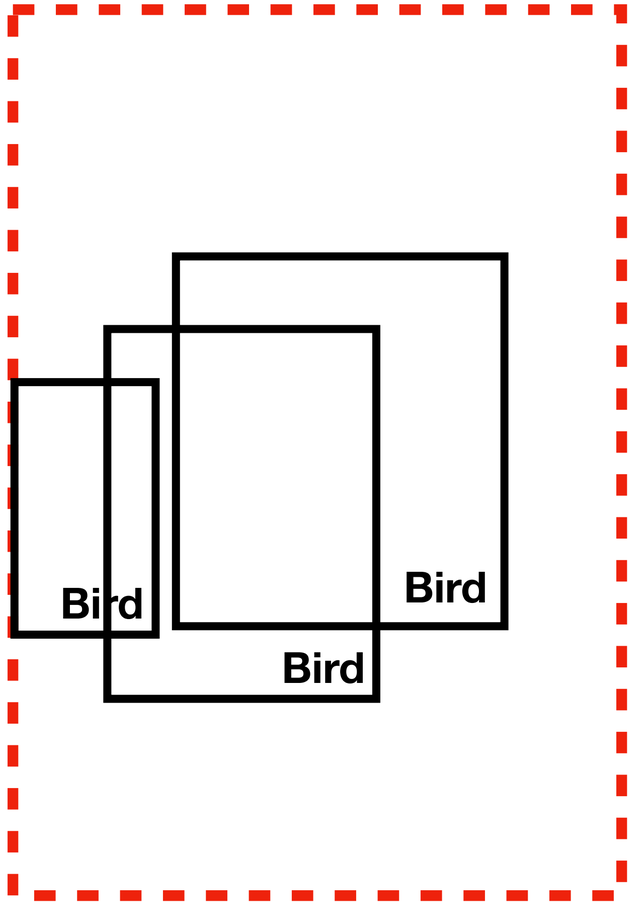


Challenge: unbounded number of output instances

(can’t just do K-way classification)

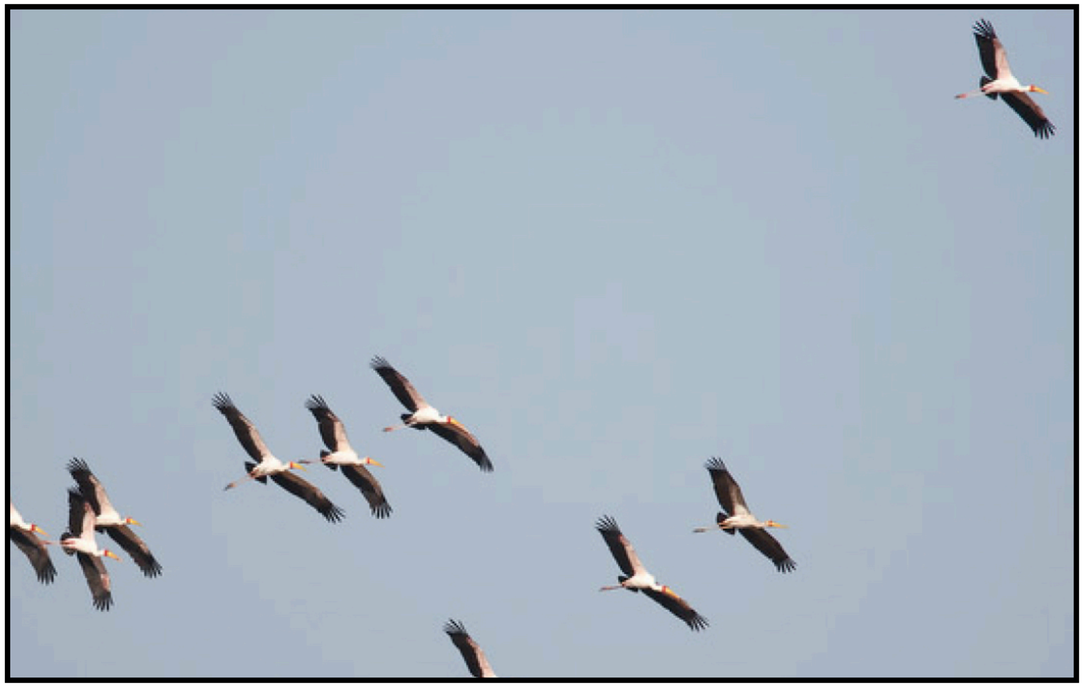
Object detection

f



Challenge: unbounded number of detections, possibly multiple detections per pixel

Attribute prediction



f

“Nature”

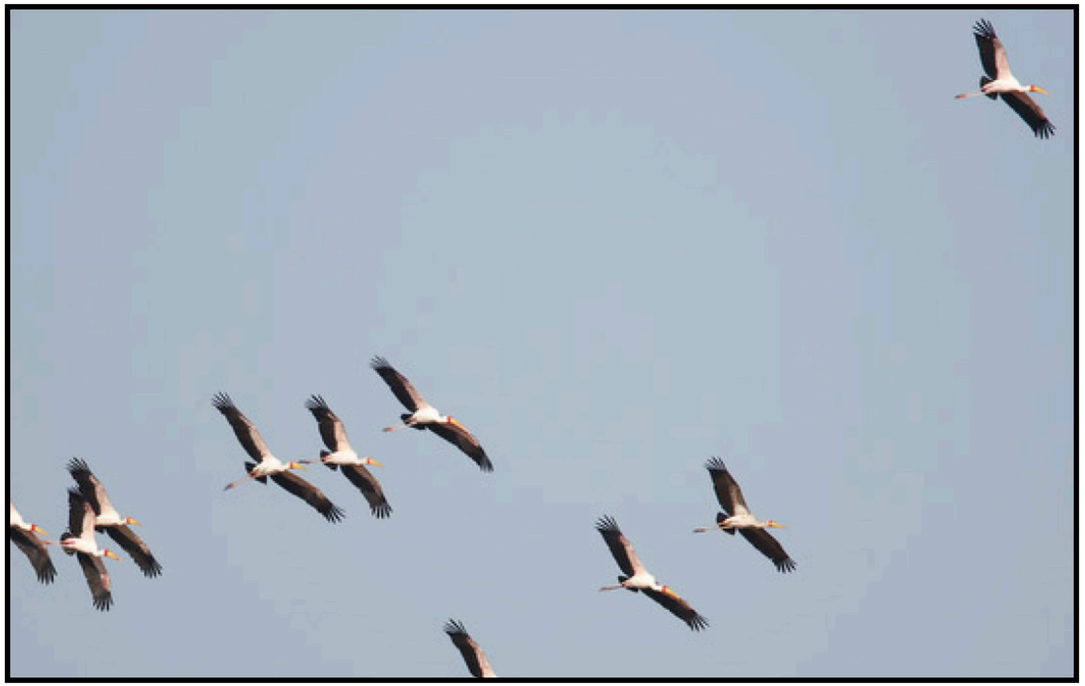
“Daytime”

“Serene”

…

Instead of one K-way classification problem,
K binary classification problems

Captioning



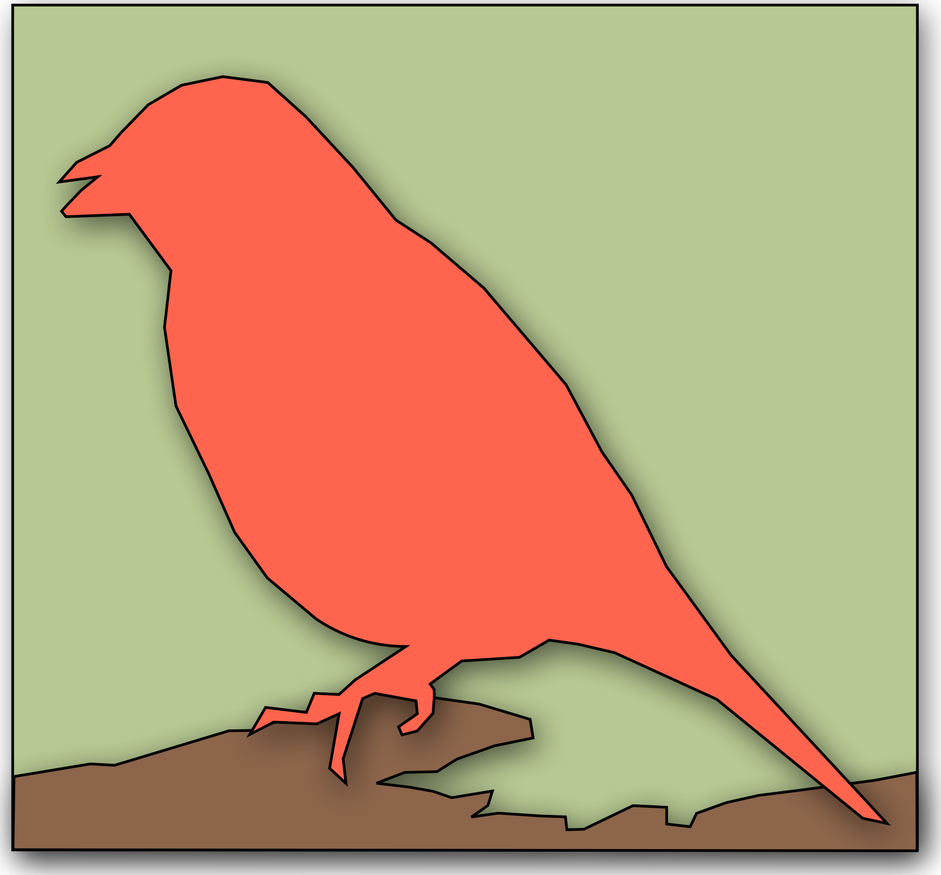
f

"A flock of birds against a gray sky"

We will study this next lecture

3D scene understanding

f



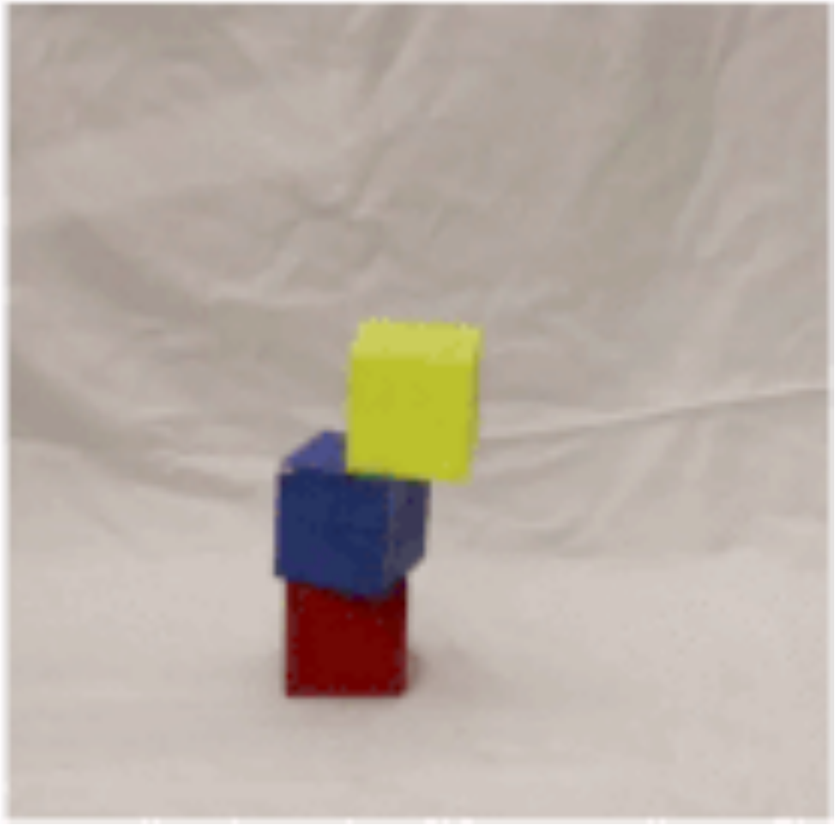
layer world representation

3D scene understanding

f

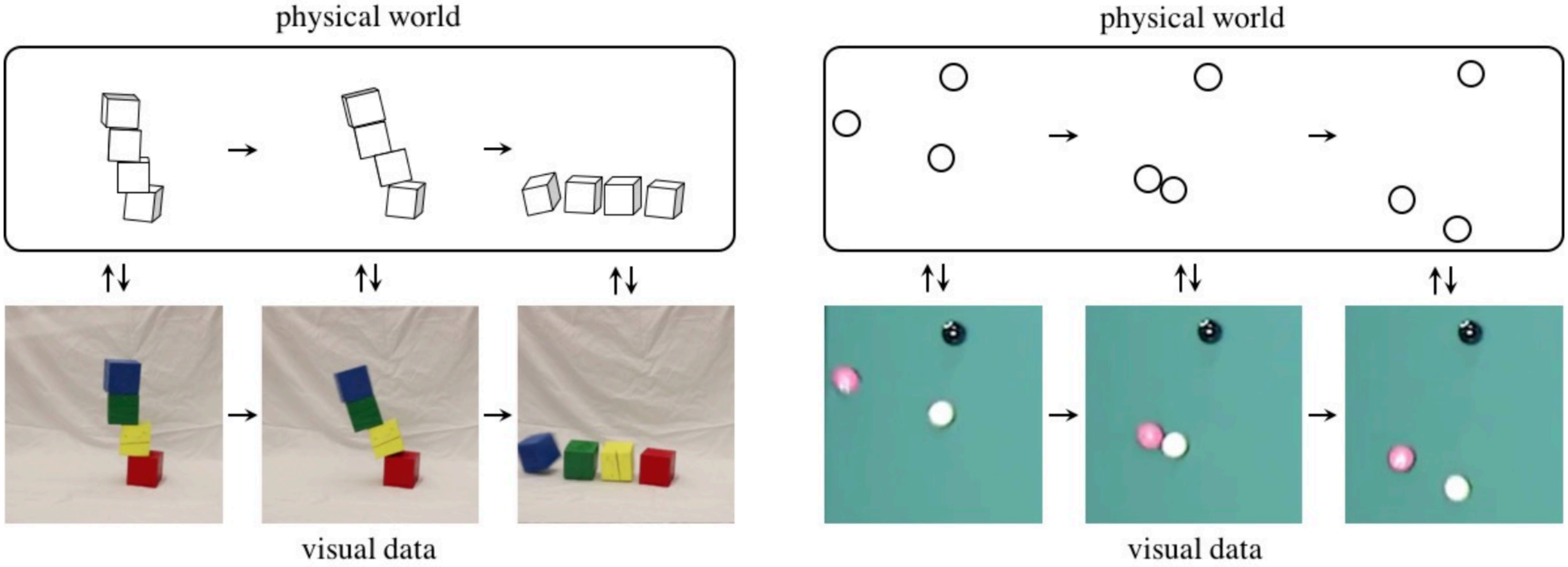
[Kanazawa, Tulsiani, et al., ECCV 2018]

Intuitive physics



["Learning to See Physics via Visual De-animation", Wu et al., NIPS 2017]

Intuitive physics

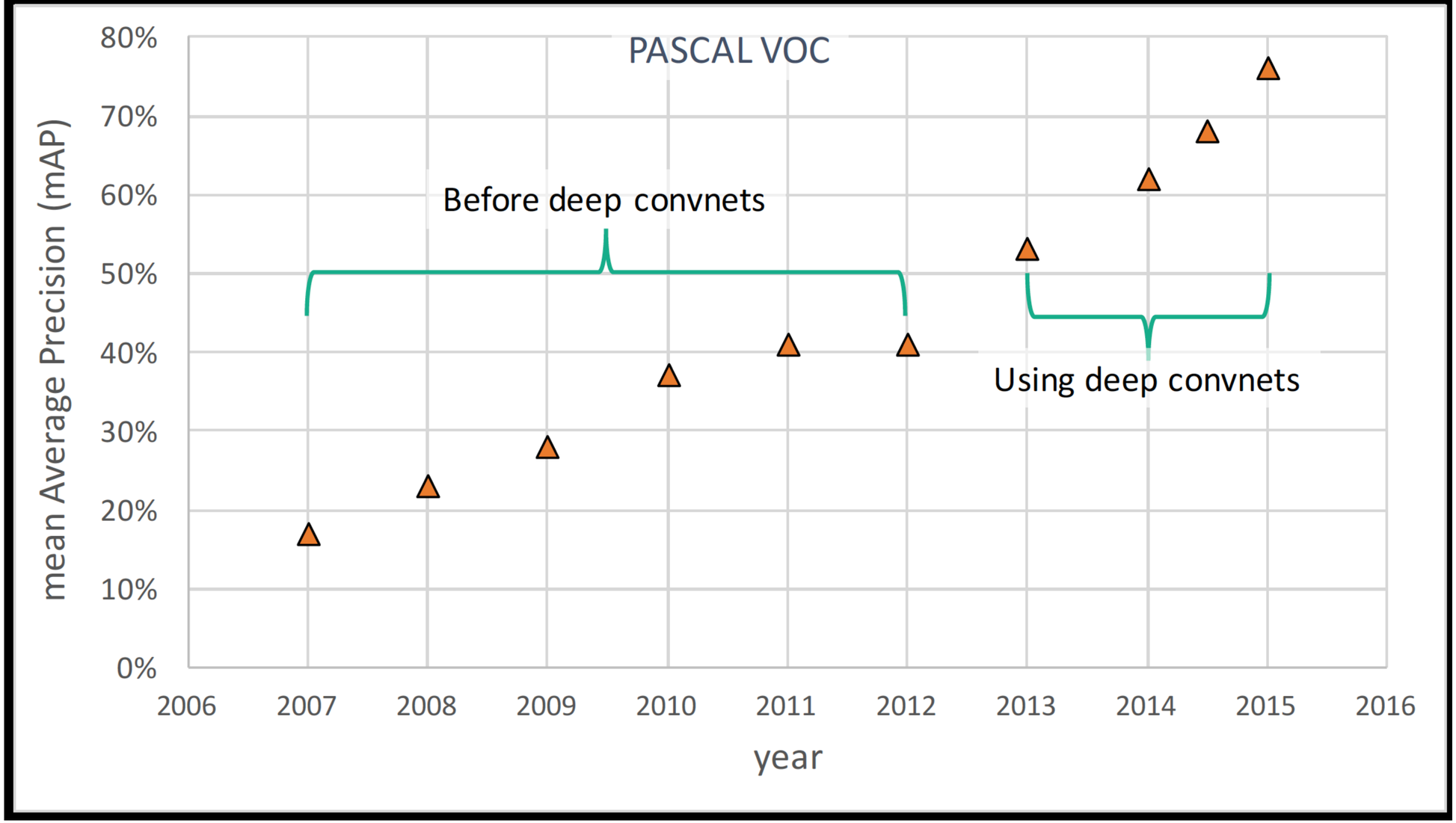


["Learning to See Physics via Visual De-animation", Wu et al., NIPS 2017]

Complete object/scene understanding?

f

Object detection renaissance (2013-present)



Slide credit: Ross Girshick

Why do we care about recognition?

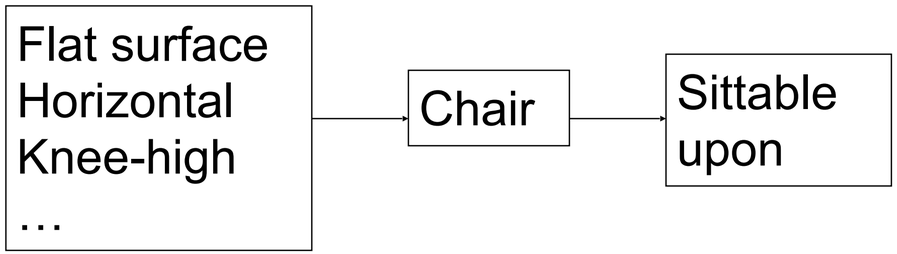
Perception of function: We can perceive the 3D shape, texture, material properties, without knowing about objects. But, the concept of category encapsulates also information about what can we do with those objects.

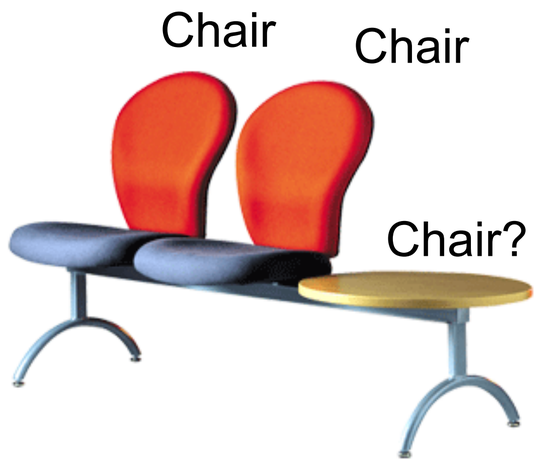
"We therefore include the perception of function as a proper –indeed, crucial- subject for vision science", *from Vision Science, chapter 9, Palmer*.

The perception of function

- Direct perception (affordances): Gibson

- Mediated perception (Categorization)





Direct perception

Some aspects of an object function can be perceived directly

- Functional form: Some forms clearly indicate to a function (“sittable-upon”, container, cutting device, …)

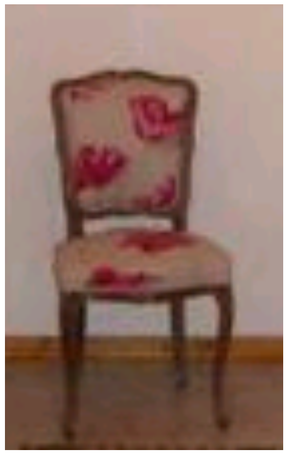
Sittable-upon Sittable-upon

Sittable-upon

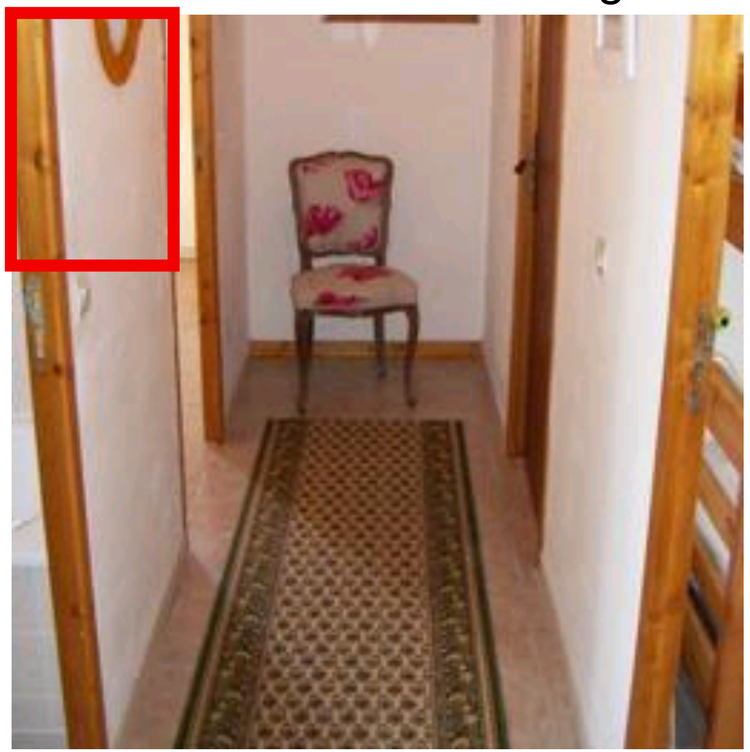
It does not seem easy to sit-upon this…

Object recognition
Is it really so hard?

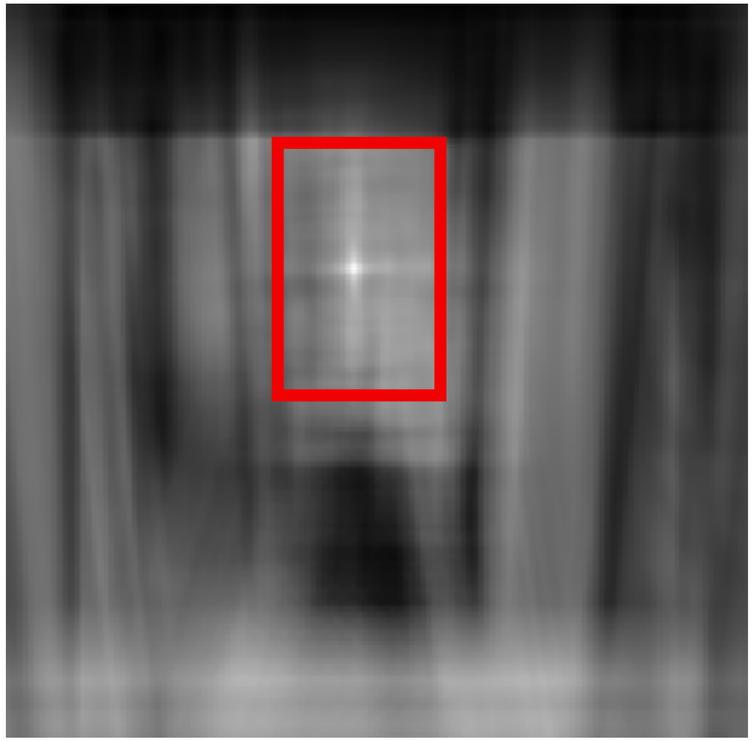
This is a chair

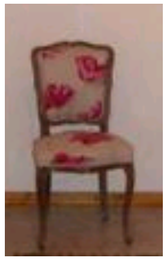


Find the chair in this image



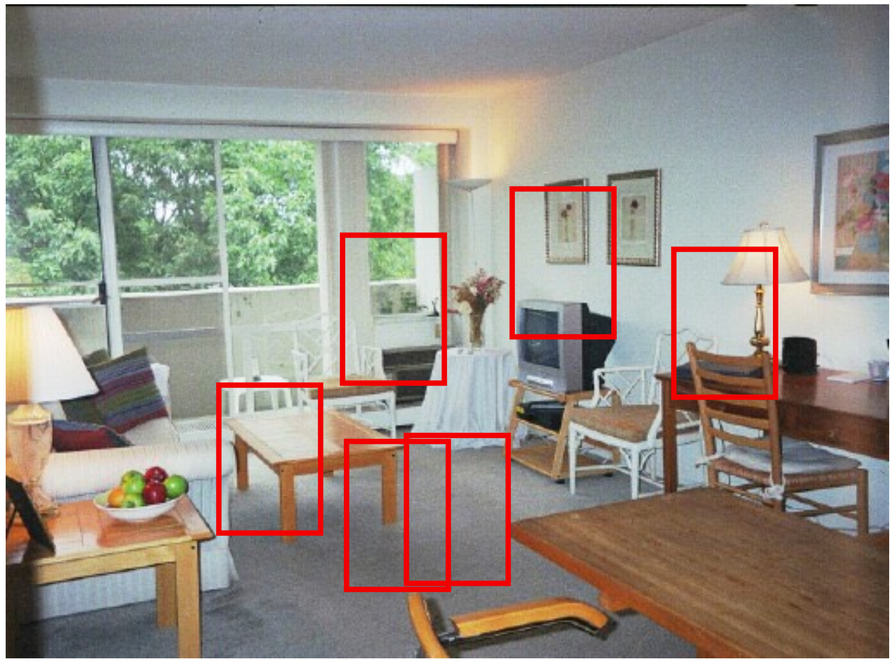
Output of normalized correlation

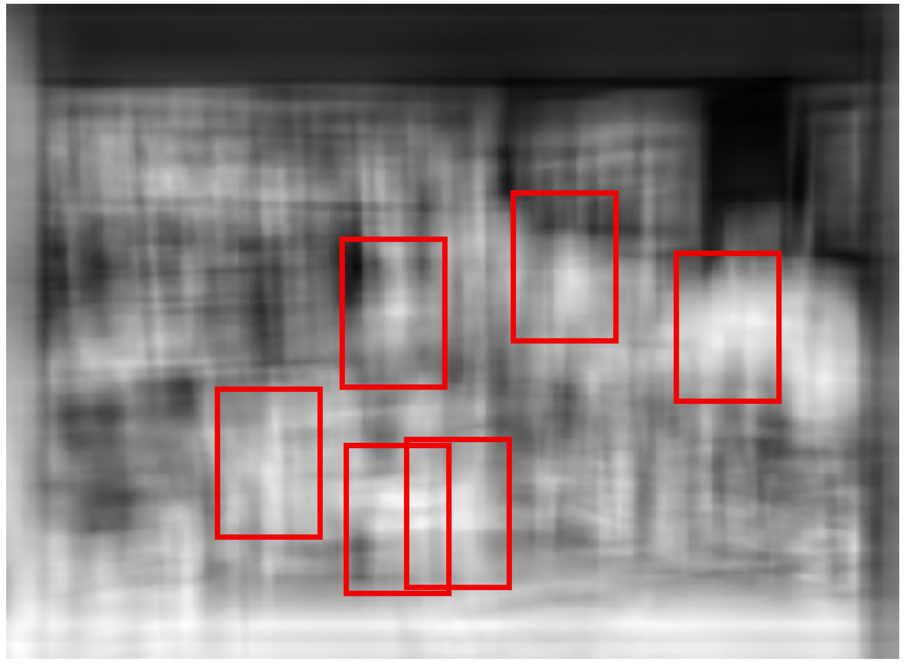




Object recognition
Is it really so hard?

Find the chair in this image



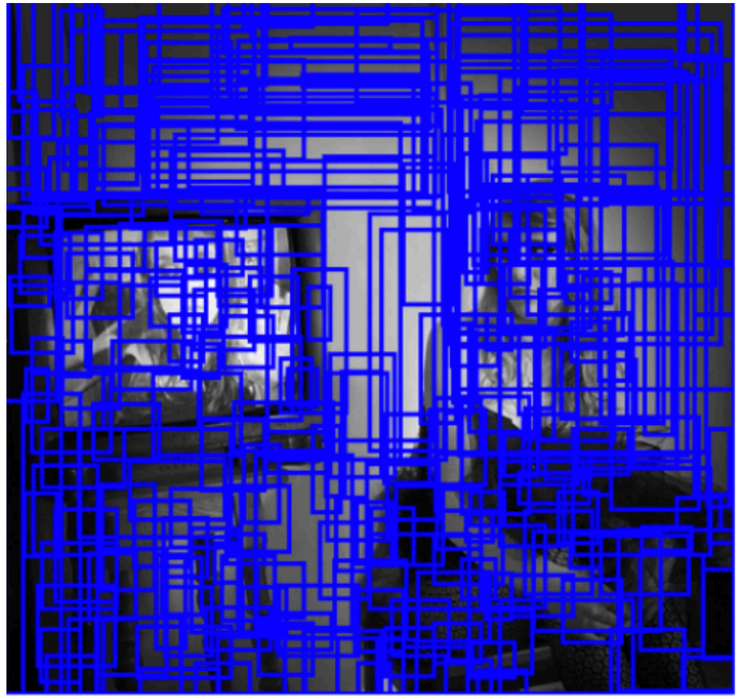


Pretty much garbage
Simple template matching is not going to make it

My biggest concern while making this slide was:
how do I justify 50 years of research, and this course, if this experiment did work?

Object detection via selective search

Input image



Candidate bounding boxes

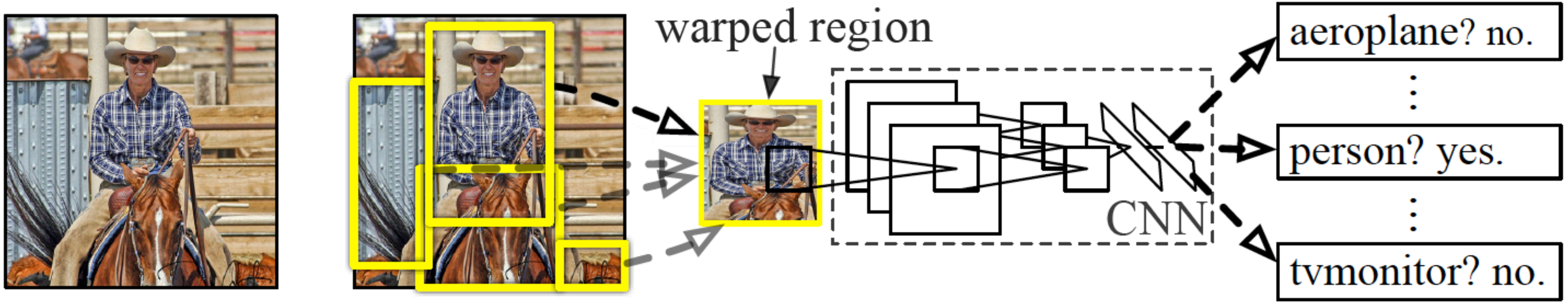
Detected objects
(by applying classifier on
candidate bounding boxes)

[Uijlings et al. 2013]

Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik
UC Berkeley
{rbg,jdonahue,trevor,malik}@eecs.berkeley.edu

CVPR 2014



Slow R-CNN

Input image

Girshick et al. CVPR14.

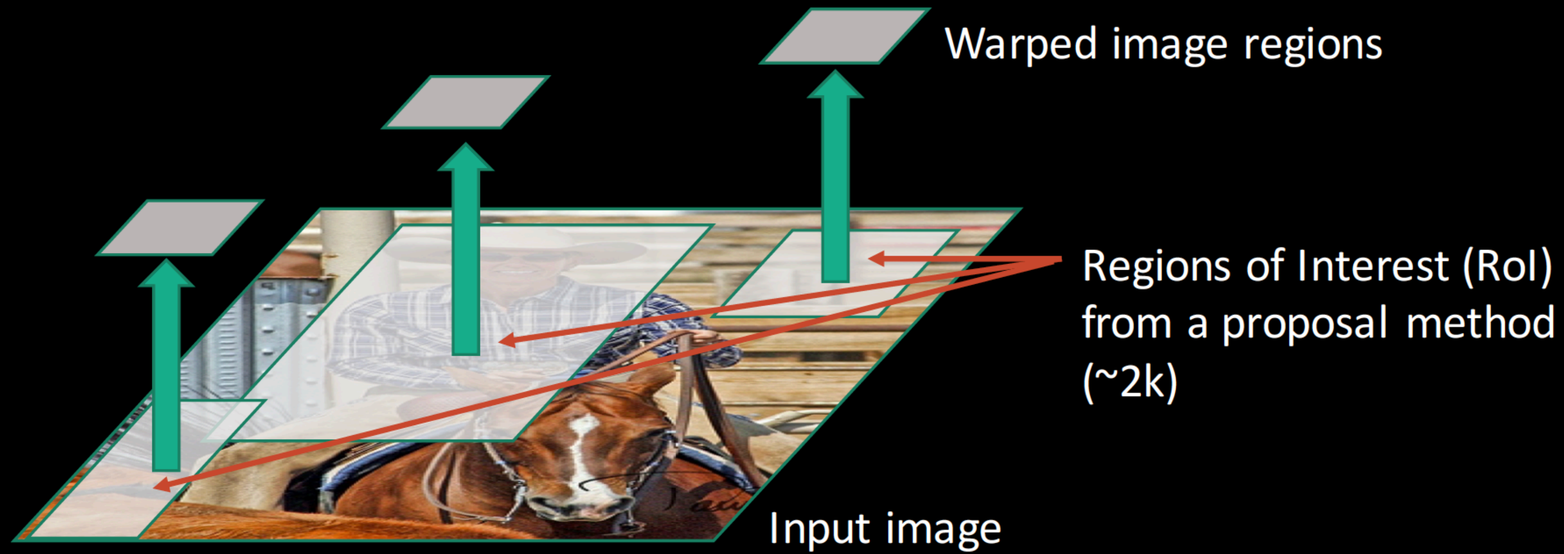
Slow R-CNN

Regions of Interest (RoI) from a proposal method (~2k)

Input image

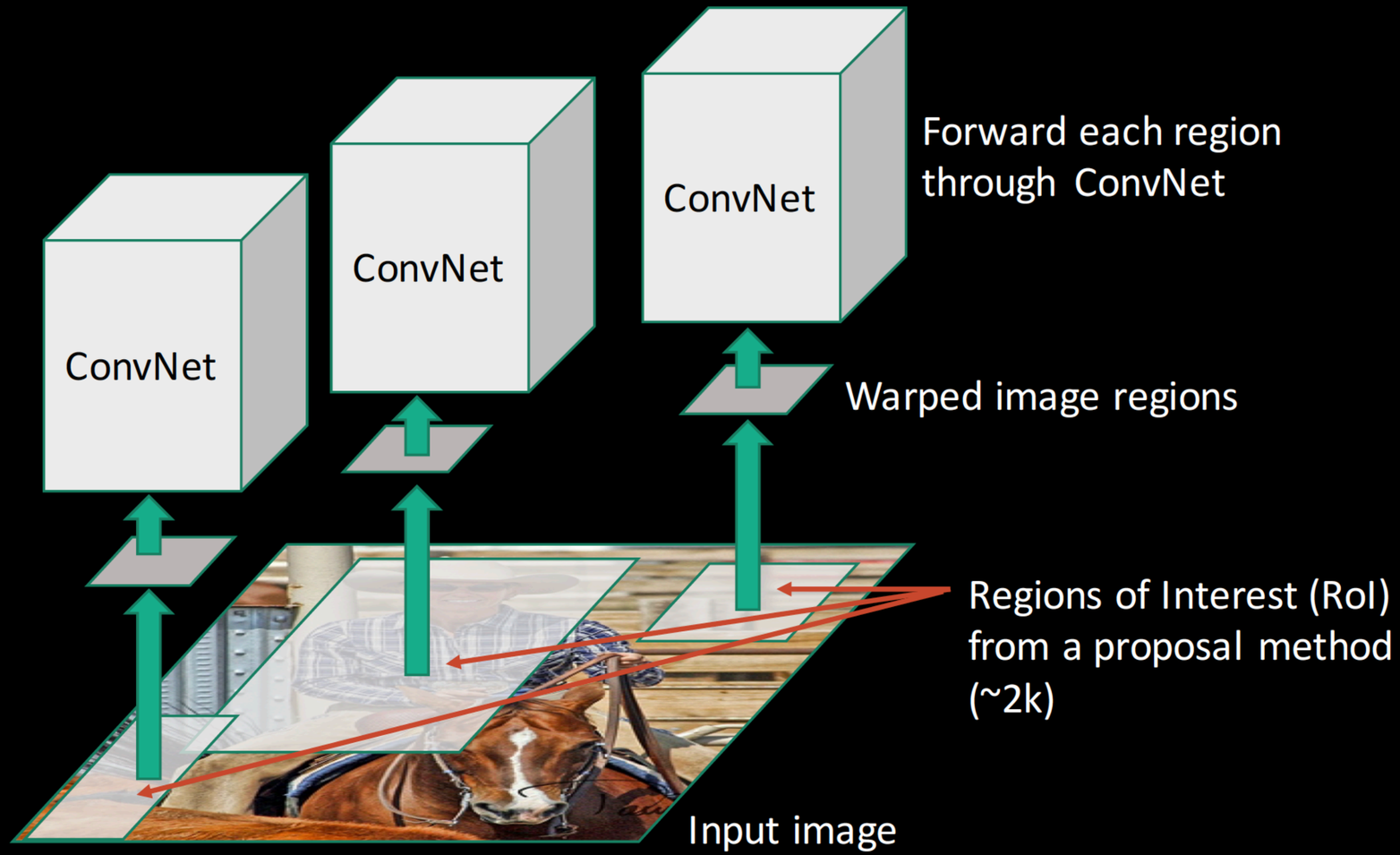
Girshick et al. CVPR14.

Slow R-CNN



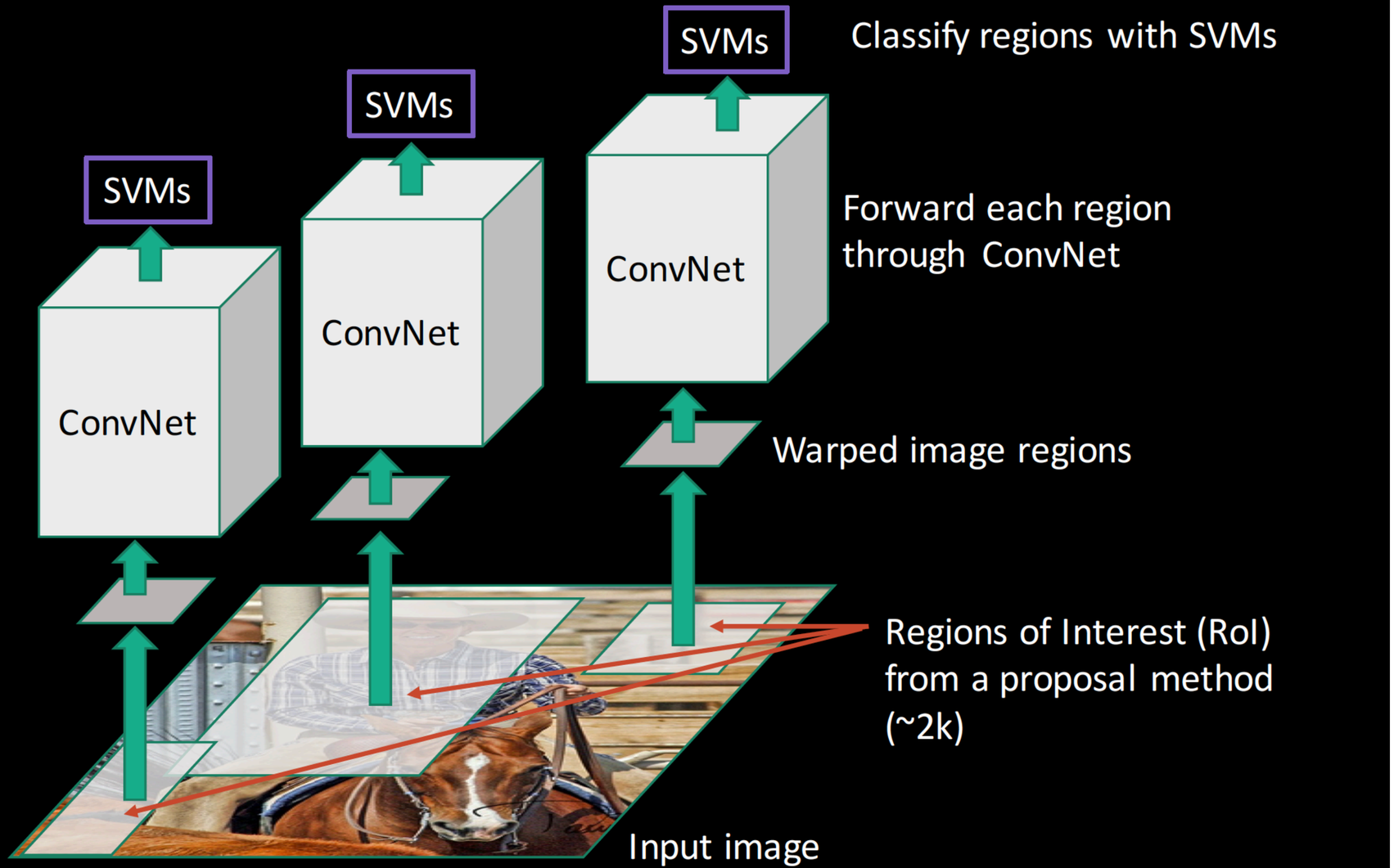
Girshick et al. CVPR14.

Slow R-CNN



Girshick et al. CVPR14.

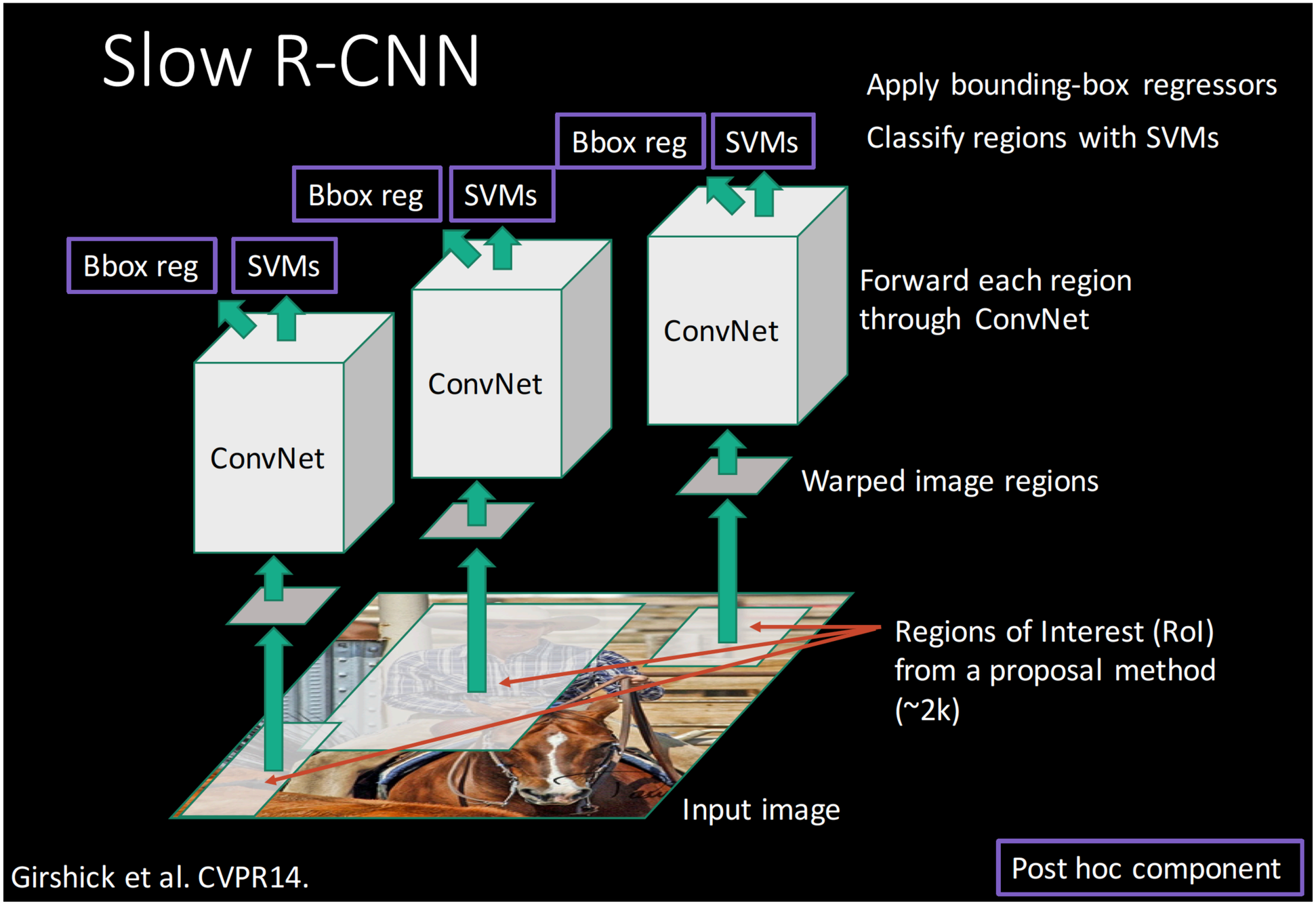
Slow R-CNN



Girshick et al. CVPR14.

Post hoc component

Slow R-CNN



Girshick et al. CVPR14.

Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun

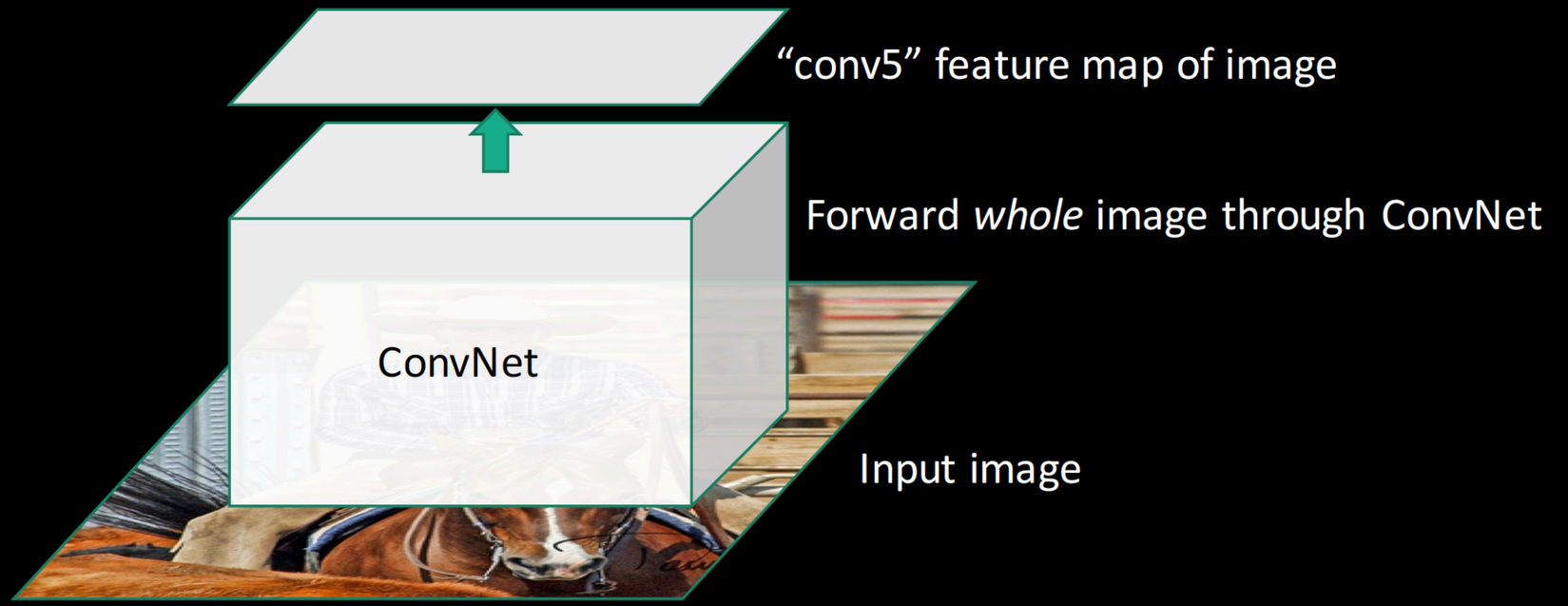
https://arxiv.org/pdf/1406.4729.pdf

SPP-net

Input image

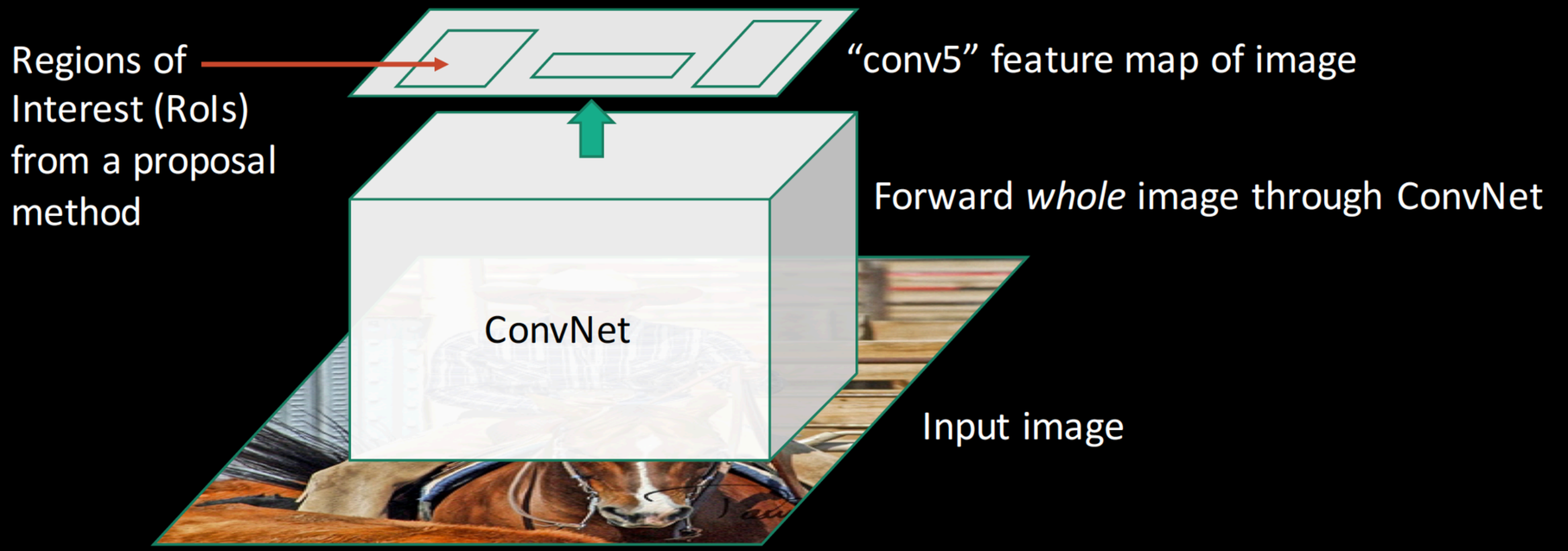
He et al. ECCV14.

SPP-net



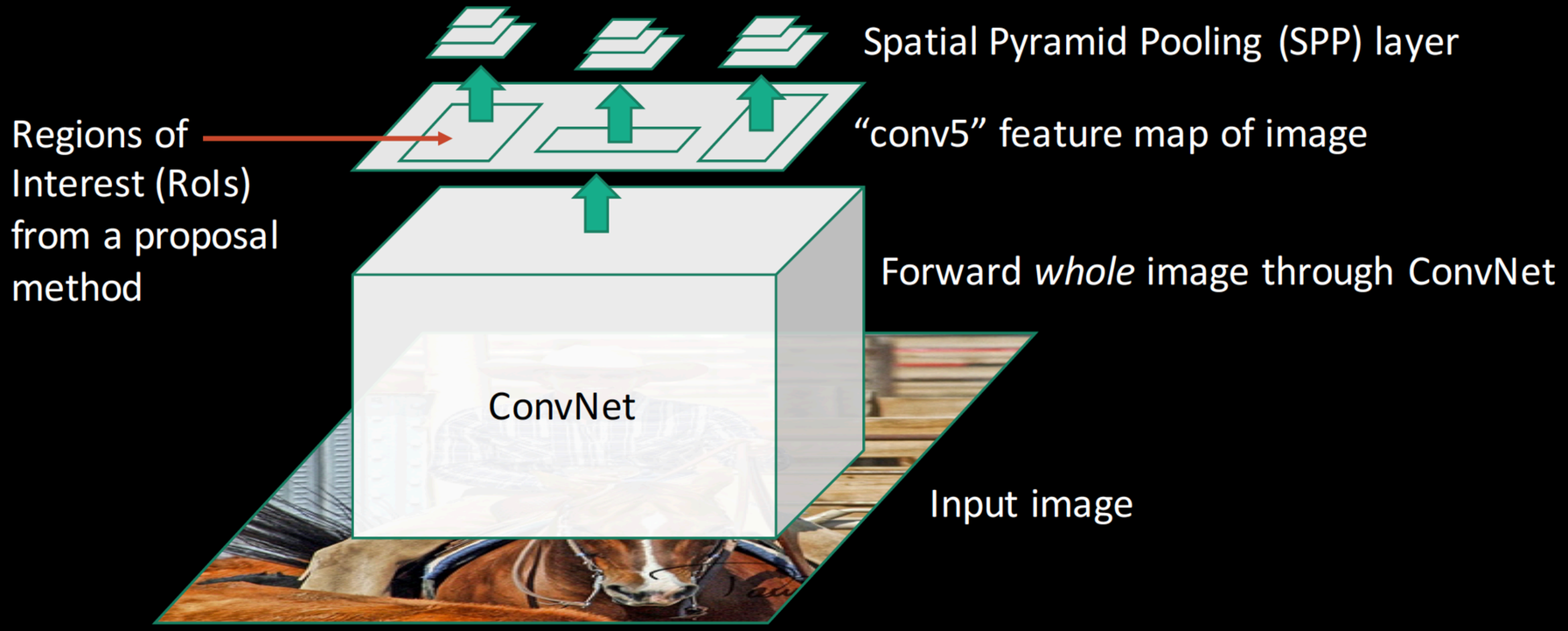
He et al. ECCV14.

SPP-net



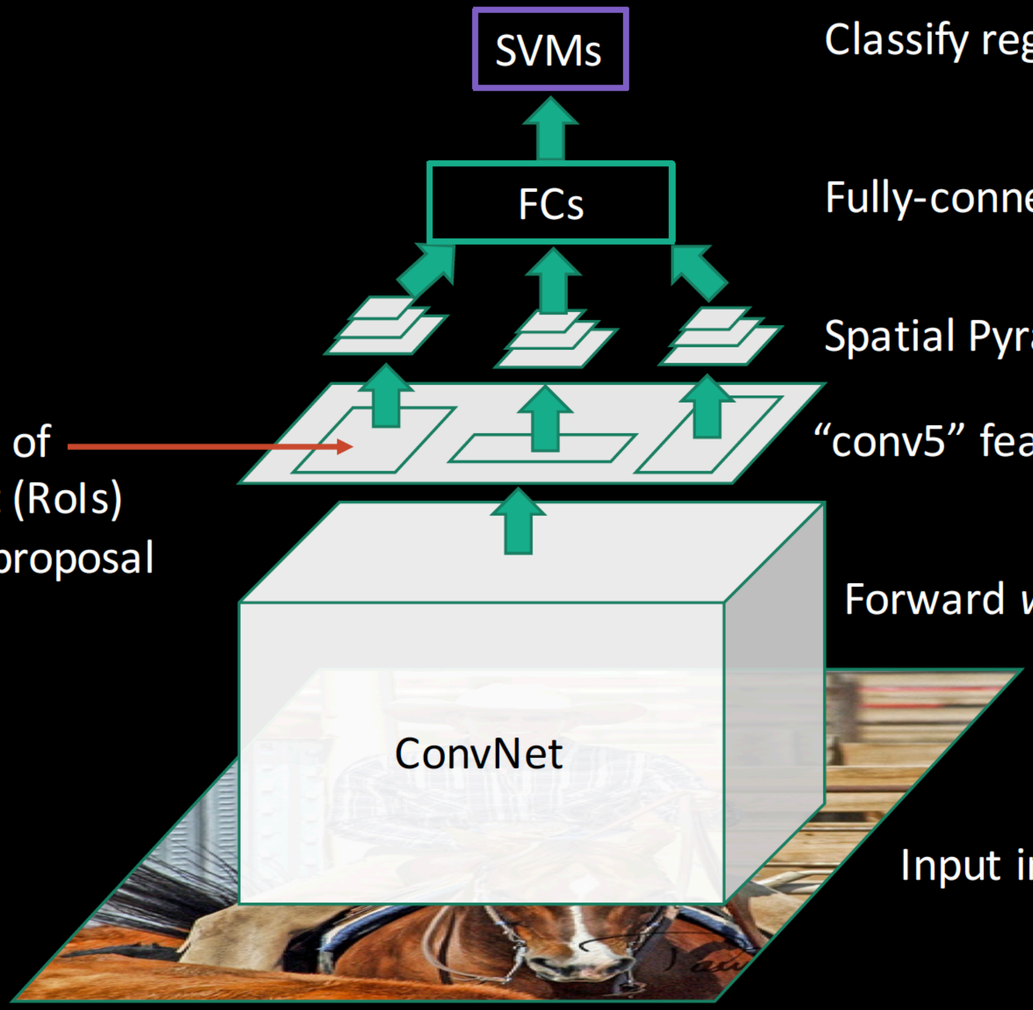
He et al. ECCV14.

SPP-net



He et al. ECCV14.

SPP-net



Classify regions with SVMs

Fully-connected layers

Spatial Pyramid Pooling (SPP) layer

"conv5" feature map of image

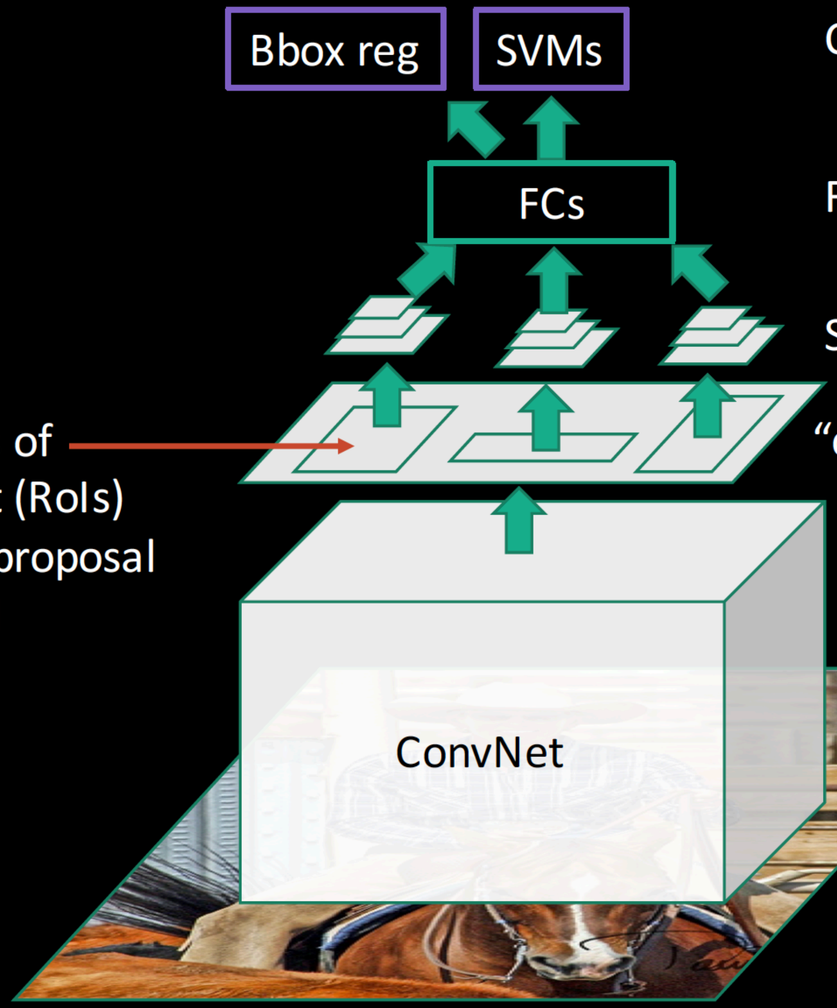
Forward *whole* image through ConvNet

Input image

He et al. ECCV14.

Post hoc component

SPP-net



Apply bounding-box regressors

Classify regions with SVMs

Fully-connected layers

Spatial Pyramid Pooling (SPP) layer

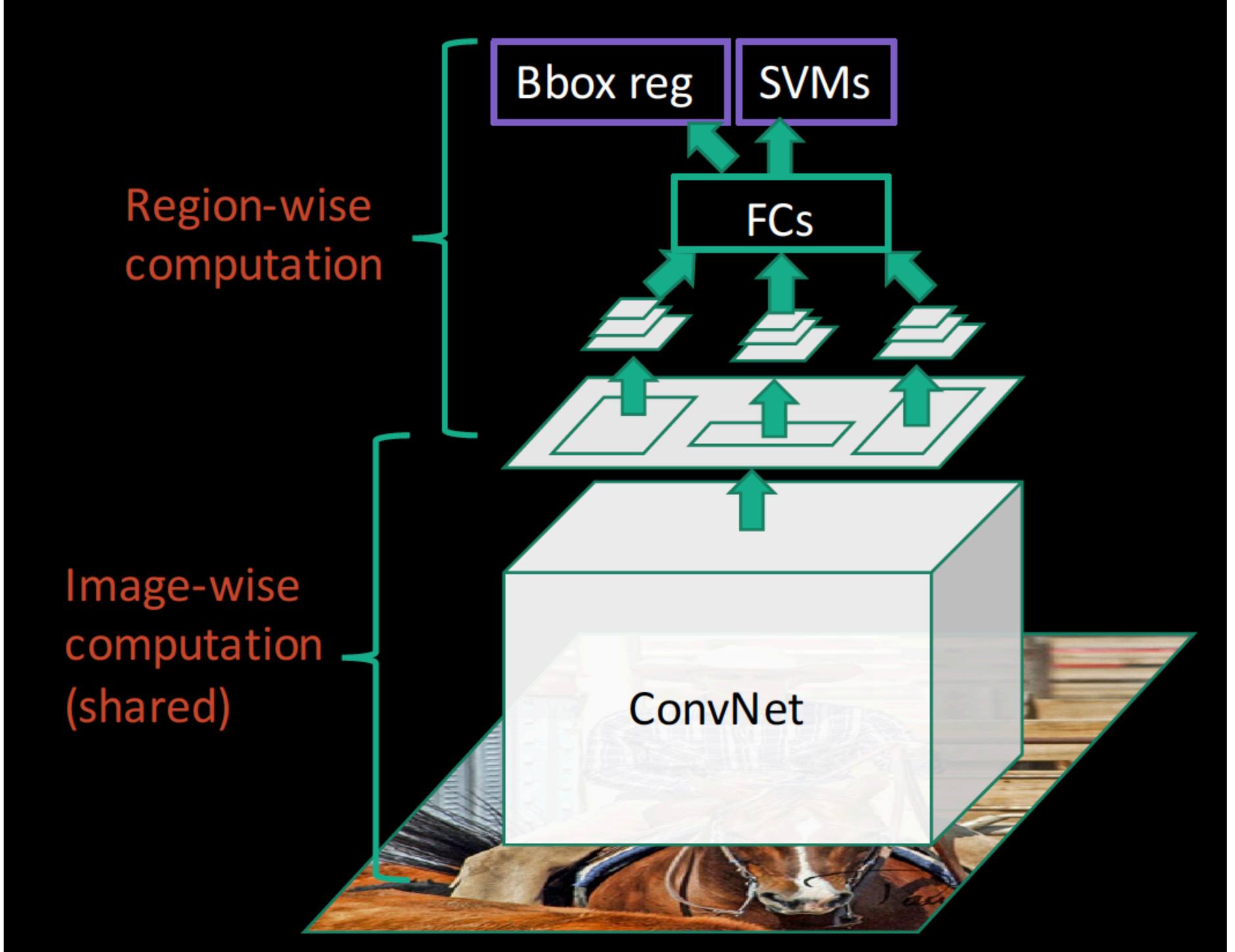
"conv5" feature map of image

Forward *whole* image through ConvNet

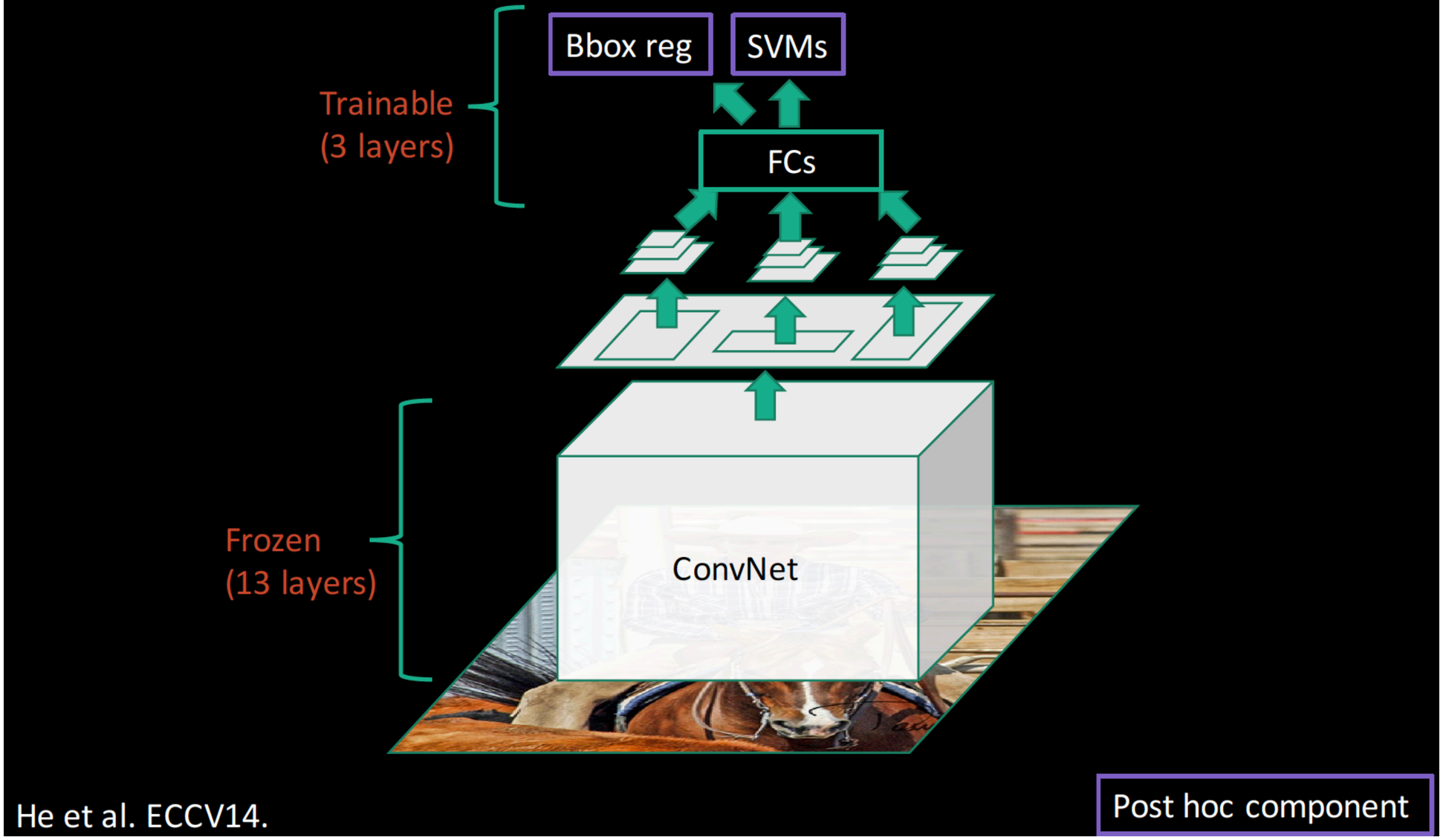
Input image

He et al. ECCV14.

Post hoc component

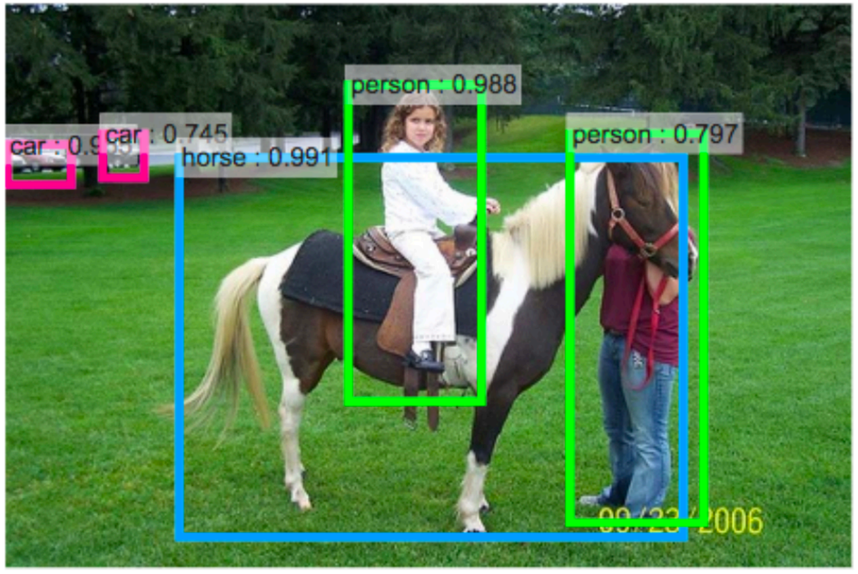

Bbox reg
SVMs
FCs
Region-wise computation
Image-wise computation (shared)
ConvNet

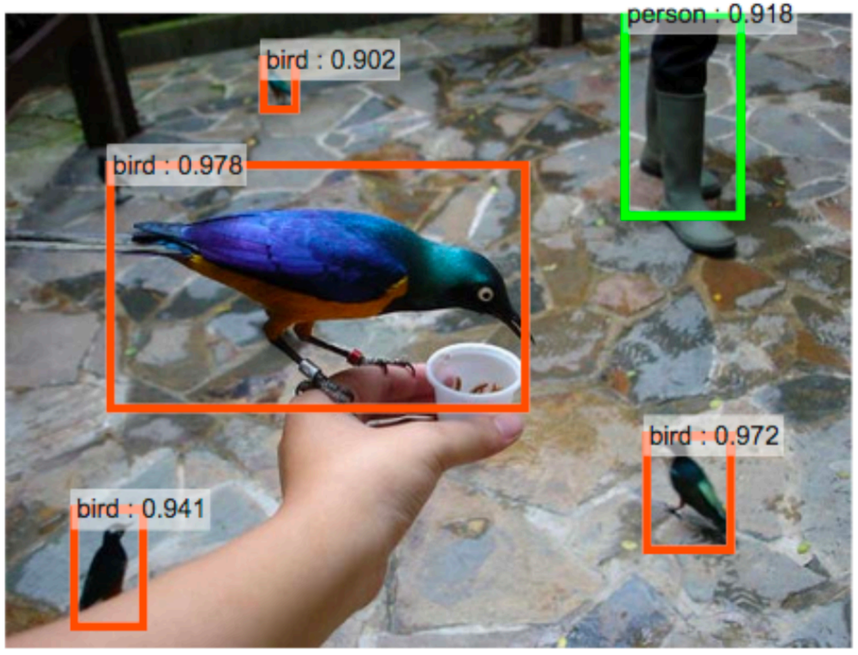
SPP-net: the main limitation



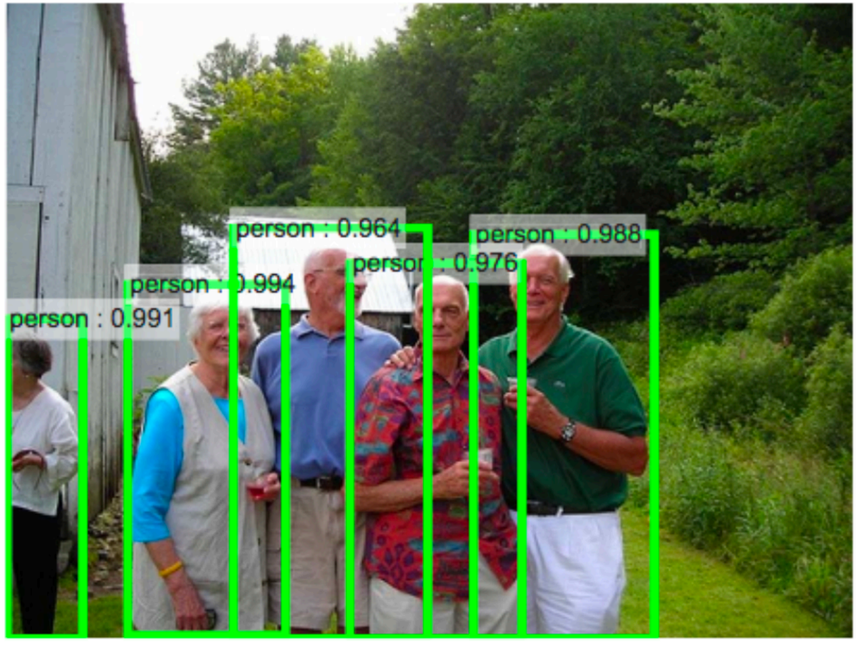
He et al. ECCV14.

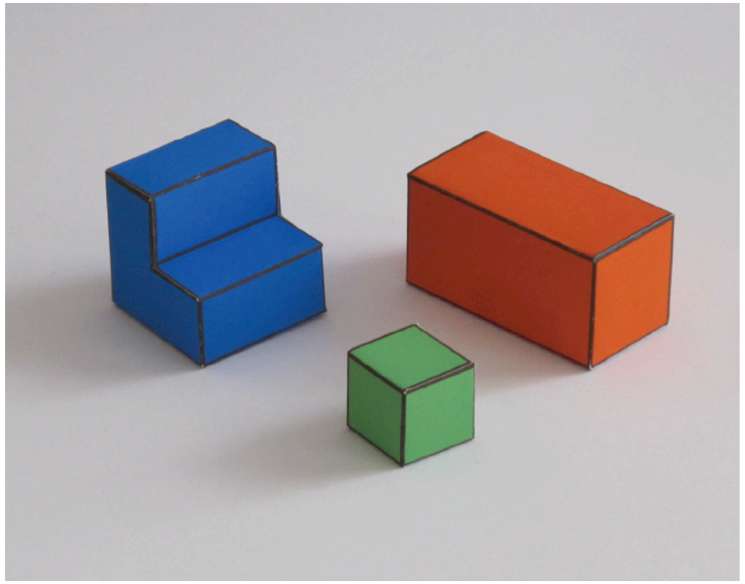
Post hoc component

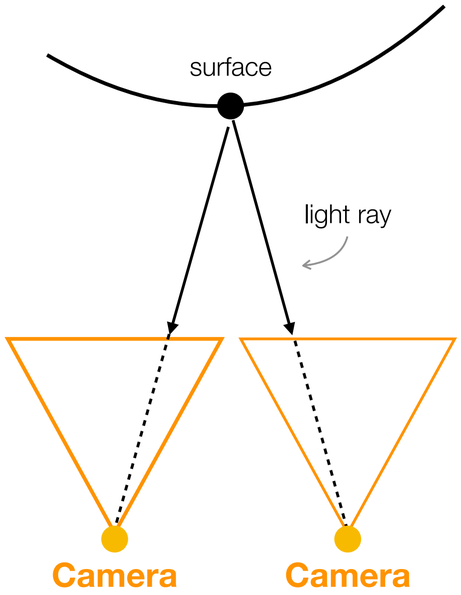

person : 0.988
car : 0.745
horse : 0.991
person : 0.797


person : 0.918
bird : 0.902
bird : 0.978
bird : 0.972
bird : 0.941


car : 0.999
person : 0.929


person : 0.964
person : 0.988
person : 0.976
person : 0.994
person : 0.991





3D scene understanding in the deep net era

Single-view 3D Prediction

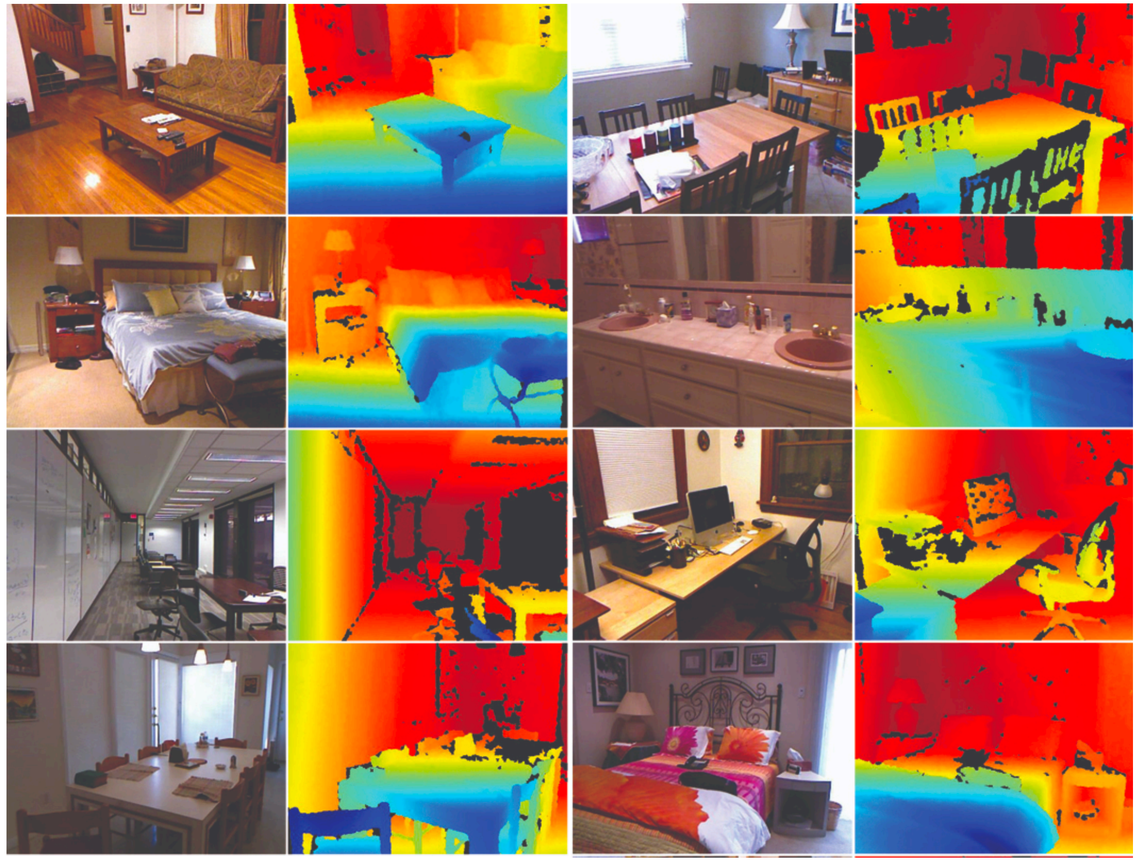
Learned Predictor

3D Representation

- Step 1: Collect training data.
- Step 2: Learn a predictor.
  - Step 2a: Wait for a few days.
- Step 3: Use the predictor!

[Slide credit: Shubham Tulsiani]

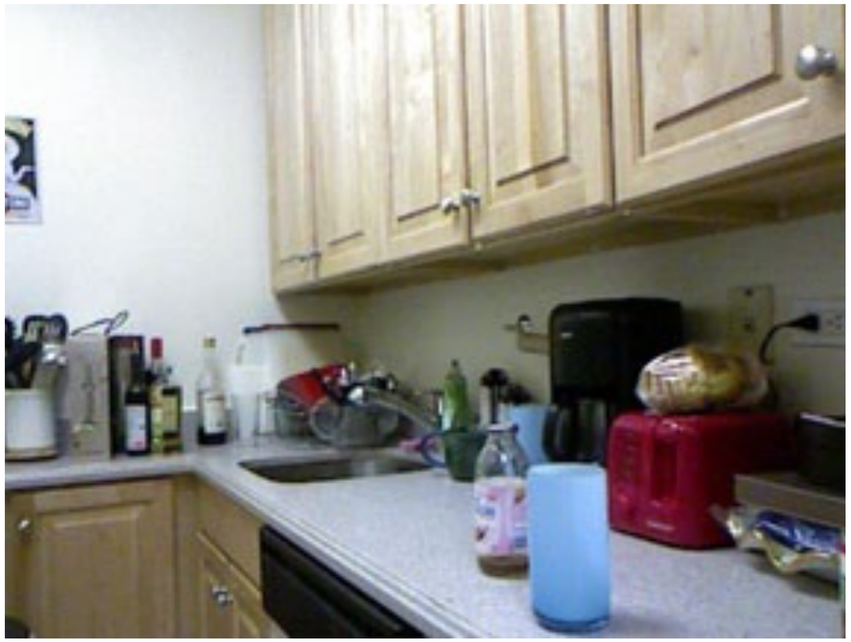
Depth from a Single Image



Collecting Training Data

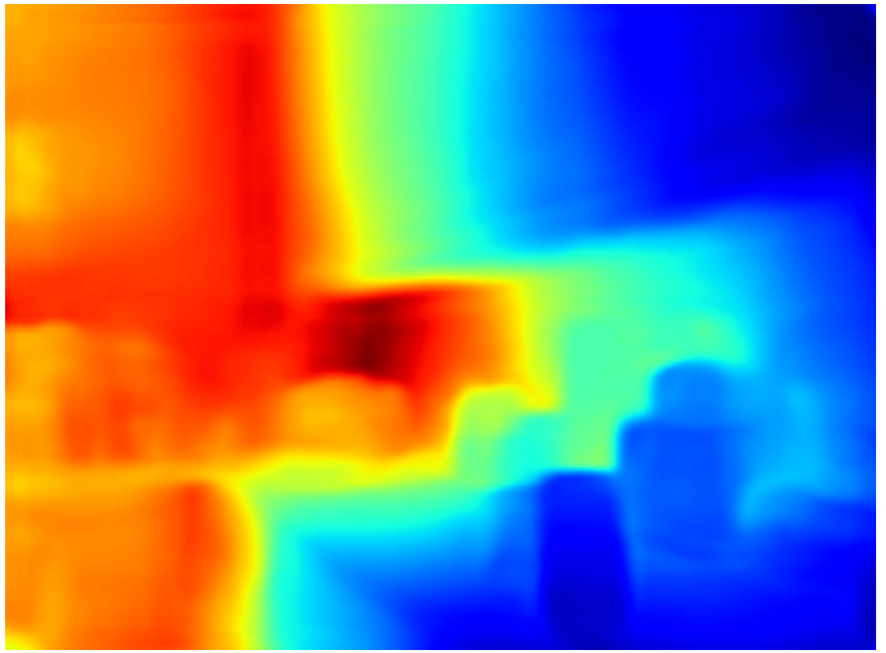
[Slide credit: Shubham Tulsiani]

Input image



f

Predicted depth map



[Result of Eigen et al., NIPS, 2014]

[Eigen, Puhrsch, and Fergus, NIPS 2014]

Data

$\{x_i, y_i\}_{i=1}^N \rightarrow$

Learner

Objective

scale invariant MSE in log space

Hypothesis space

Multiscale CNN

Optimizer

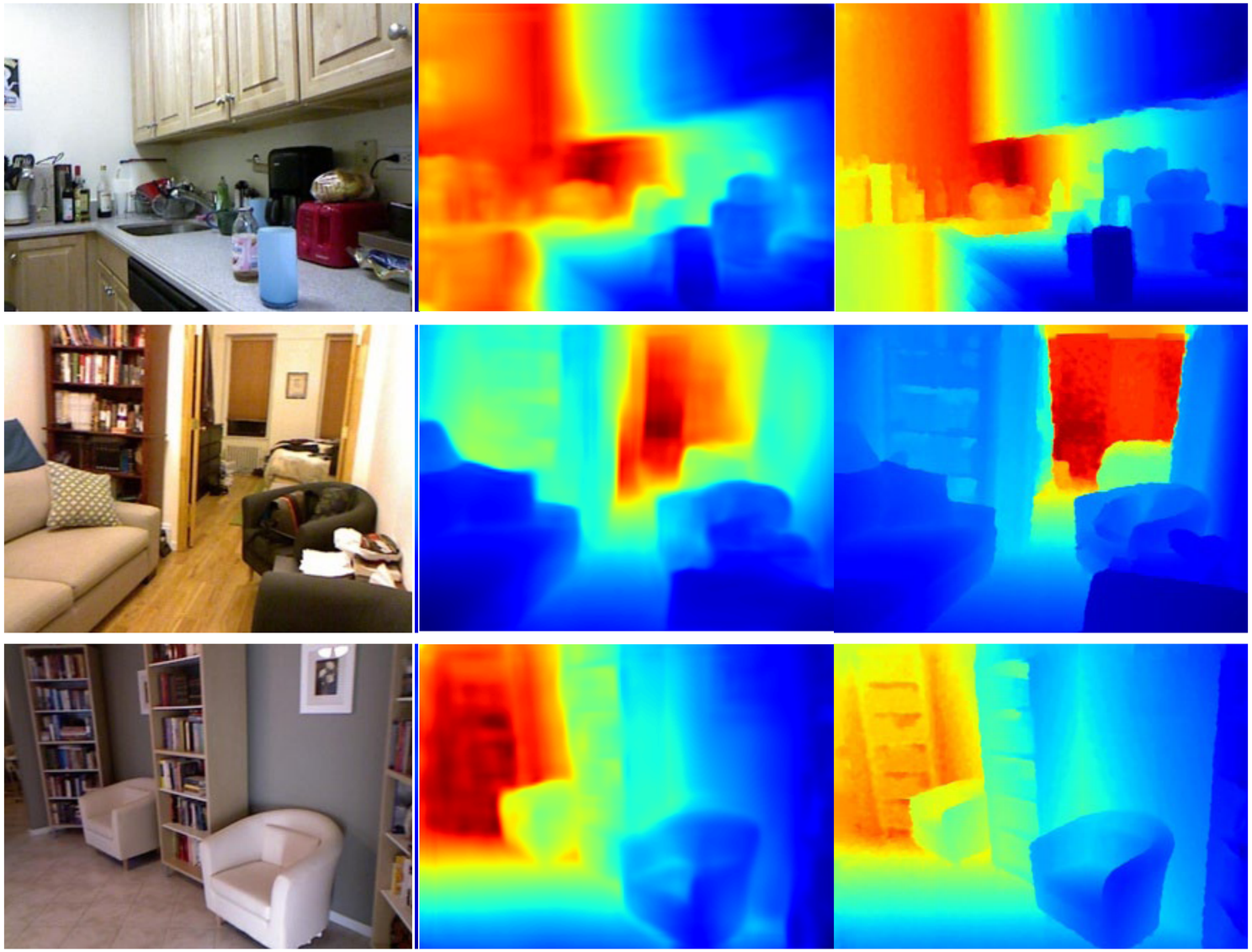
SGD

$\rightarrow f$

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i)$$

Regular old supervised learning!

Depth from a Single Image



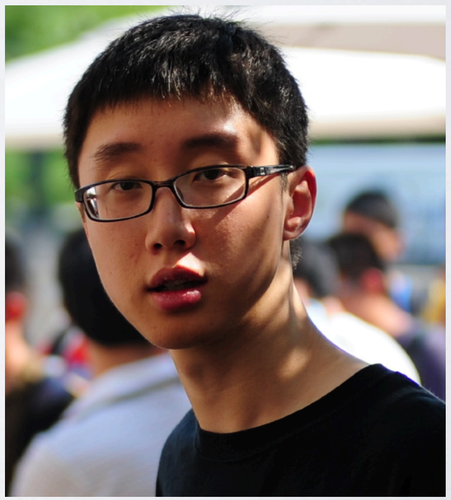
Input Prediction Ground-truth

Figure credits:
David Eigen

Unsupervised Learning of Depth and Ego-motion from Video

Tinghui Zhou[1], Matthew Brown[2], Noah Snavely[2], David Lowe[2]

UC Berkeley[1], Google[2]



[Slides credit: Tinghui Zhou]

Learning 3D from 2D Views

Training

Multi-views

Testing

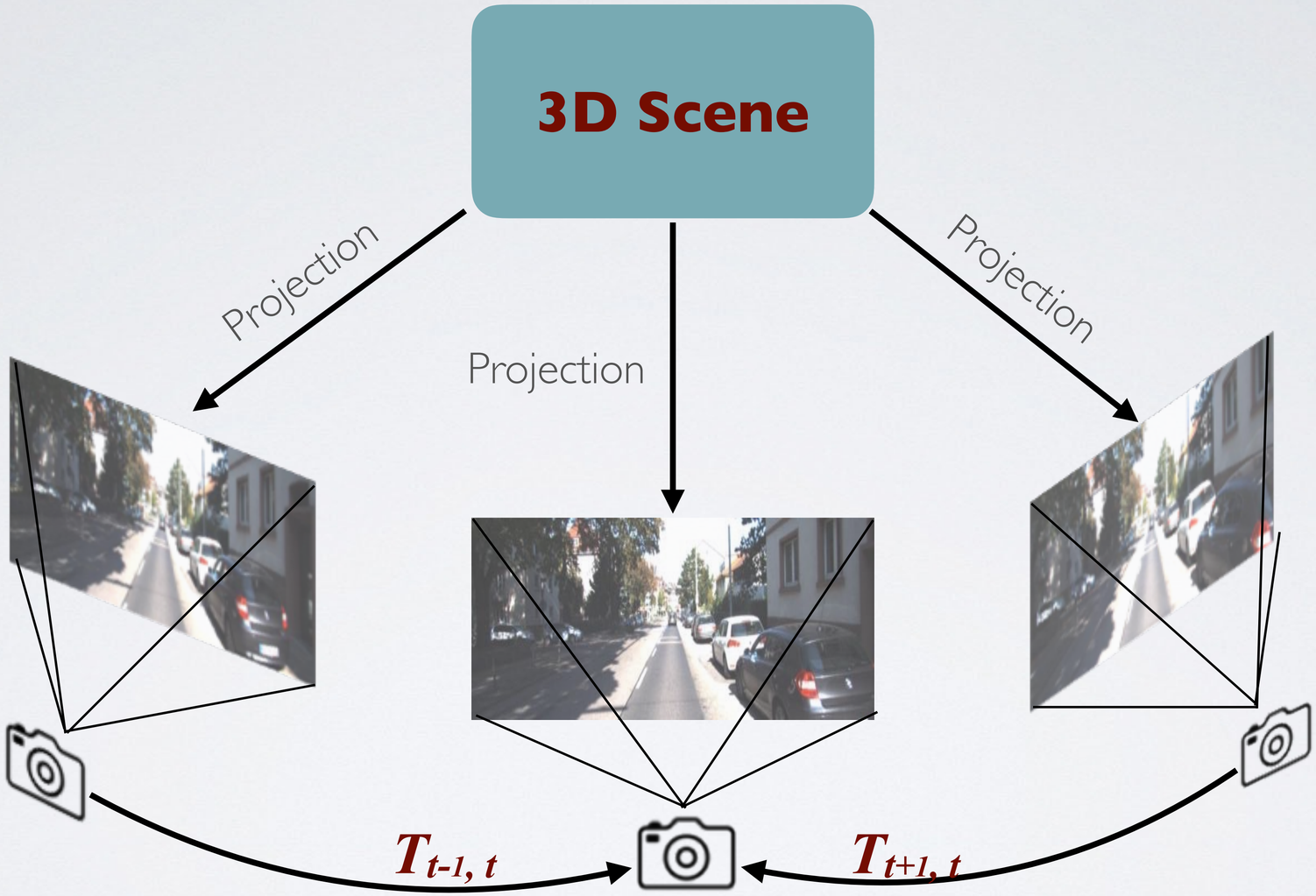
Single-view

View Synthesis

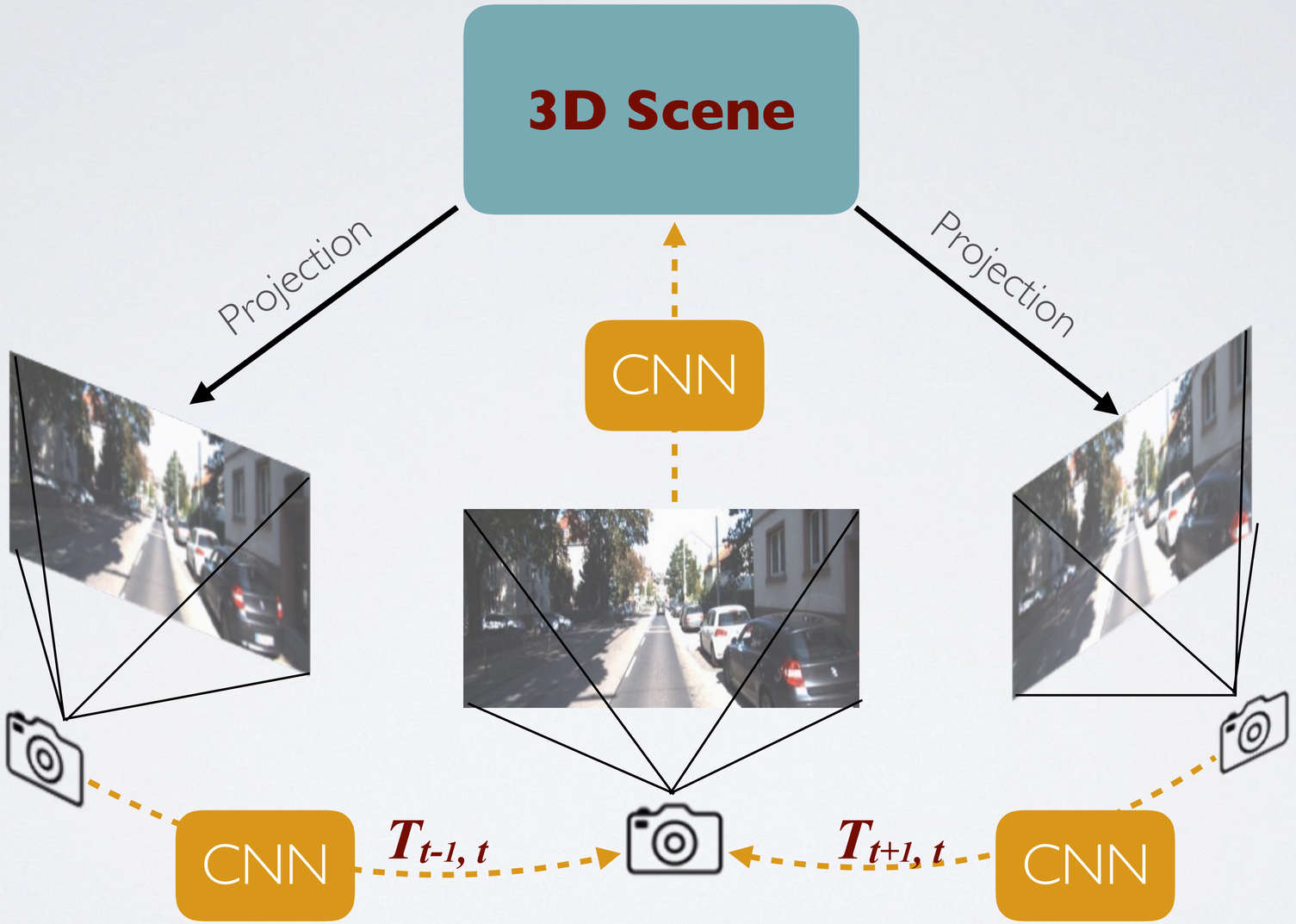
View Synthesis



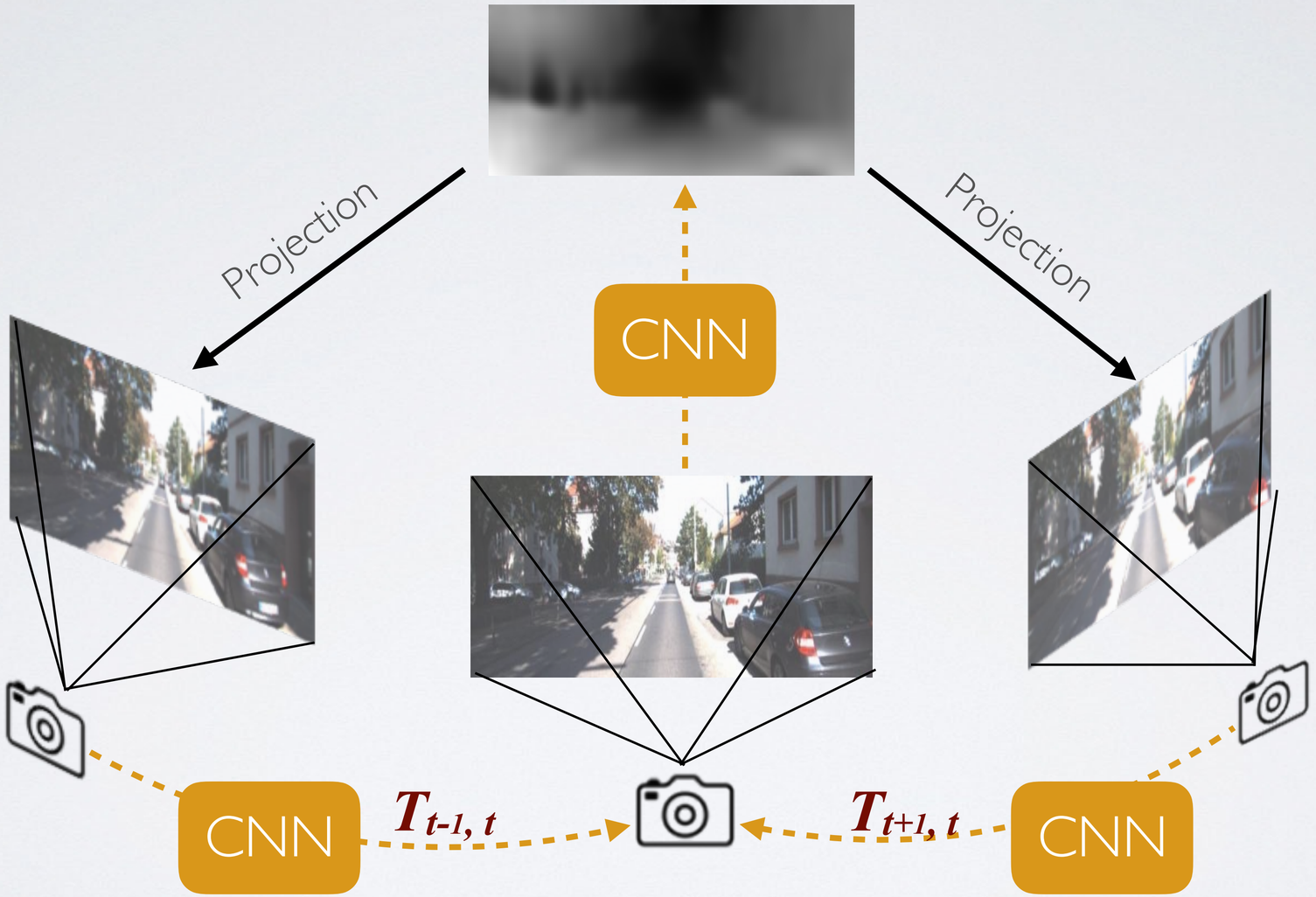
View Synthesis



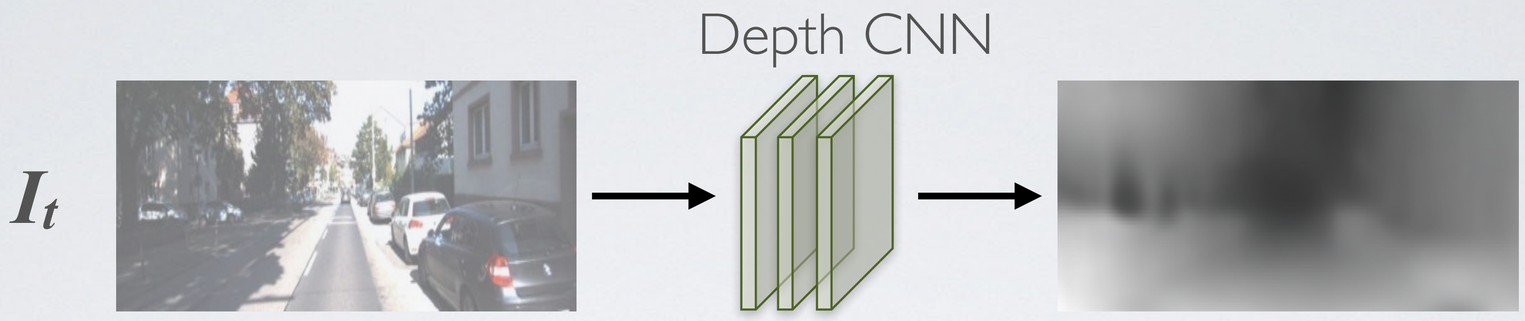
View Synthesis as Supervision



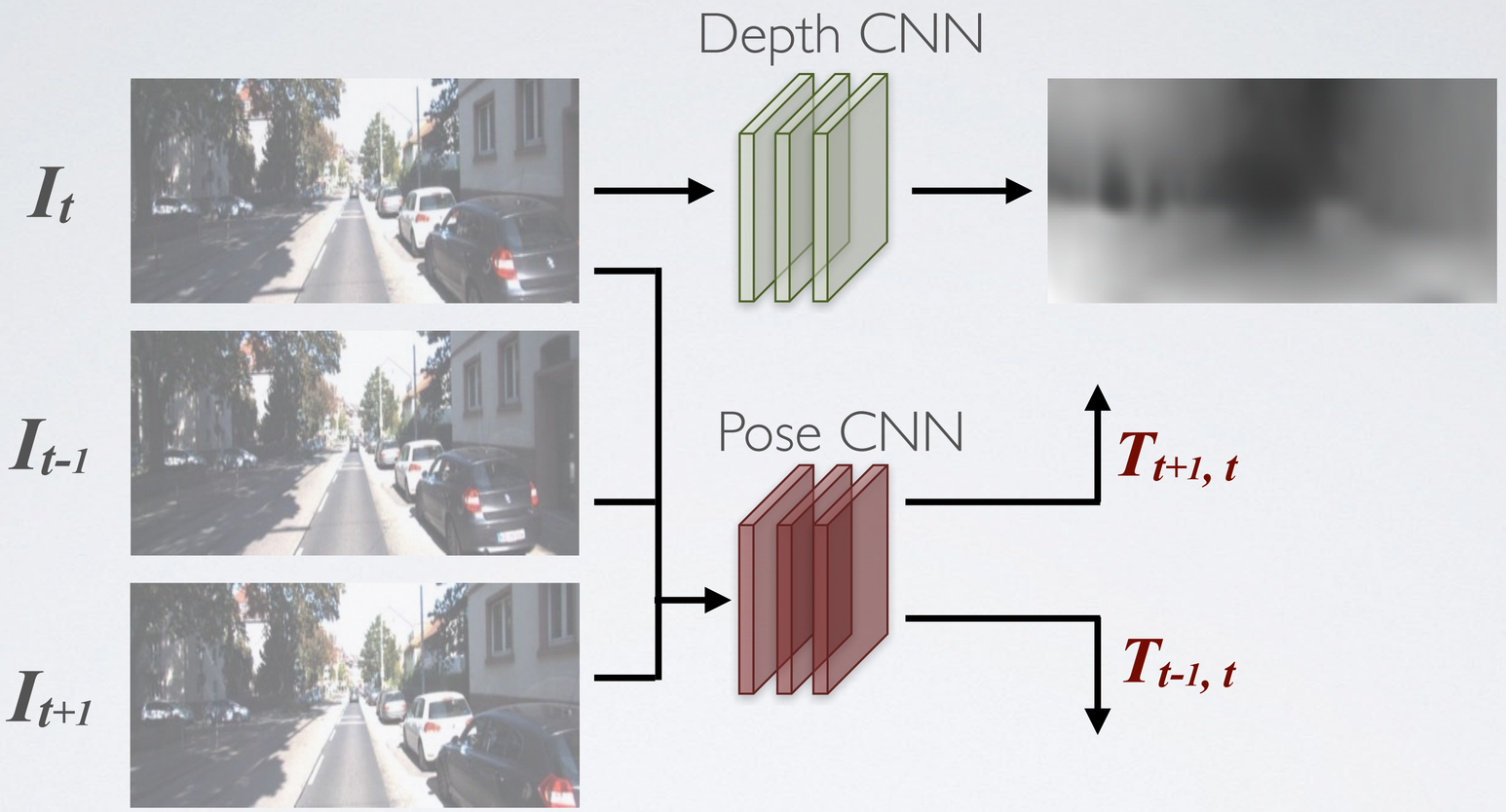
View Synthesis as Supervision



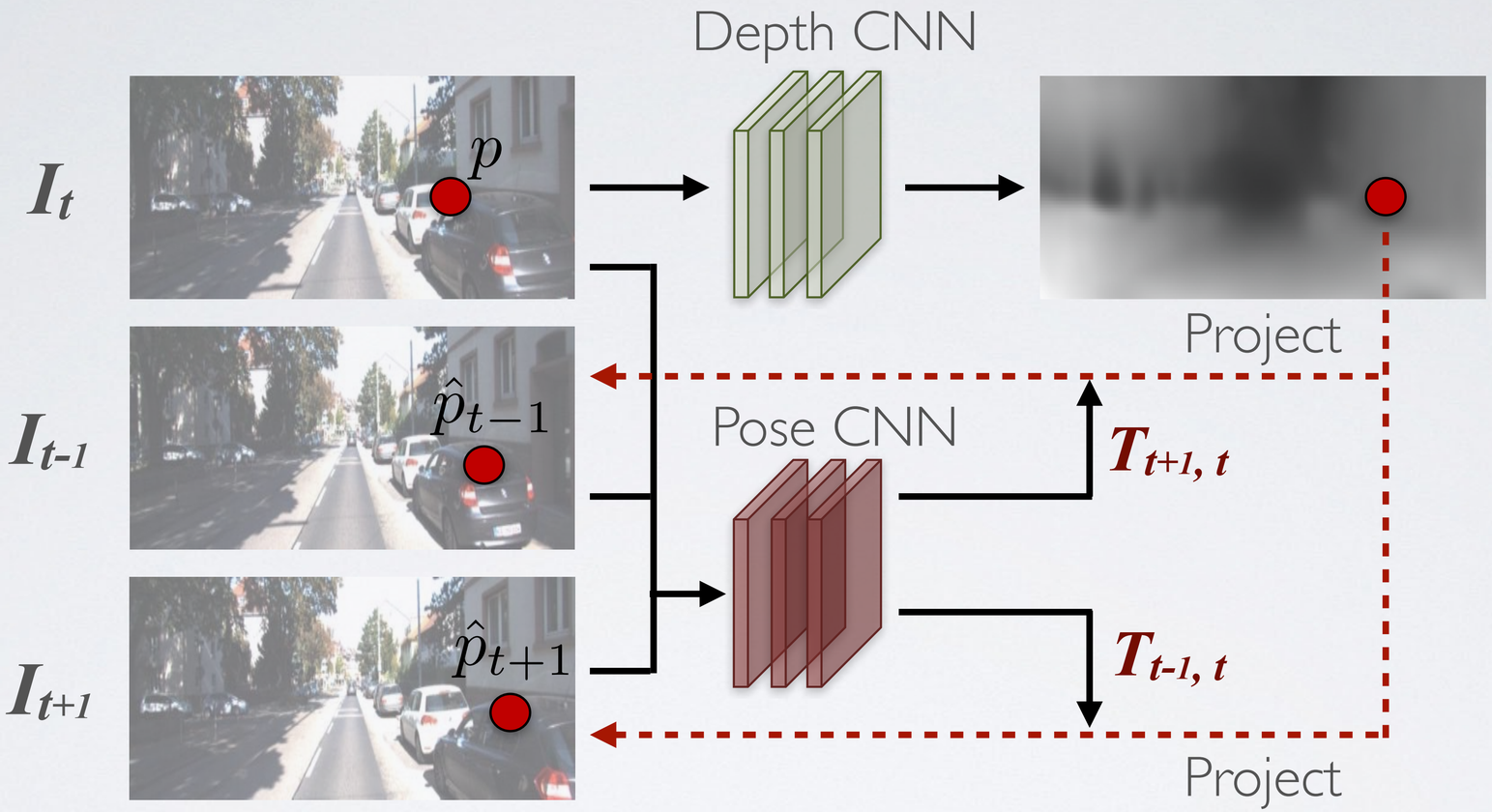
View Synthesis as Supervision



View Synthesis as Supervision



View Synthesis as Supervision



$$\mathcal{L}_{vs} = \sum_{s \in \{\text{nearby frames}\}} \sum_{p} |I_t(p_t) - I_s(\hat{p}_s)|$$

?

Results

Datasets

KITTI

Cityscapes

"Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite", *Geiger et al.*, CVPR'12
"The Cityscapes Dataset for Semantic Urban Scene Understanding", *Cordts et al.*, CVPR'16

Results (Single-View Depth)

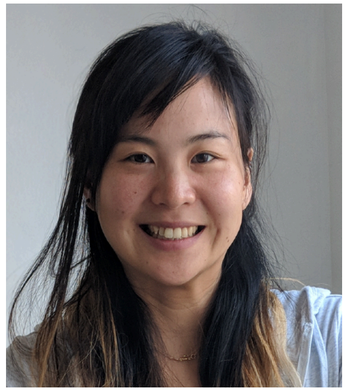
Cityscapes

KITTI

Learning Category-Specific Mesh Reconstruction from Image Collections

Angjoo Kanazawa*, Shubham Tulsiani* , Alyosha Efros, Jitendra Malik

ECCV 2018



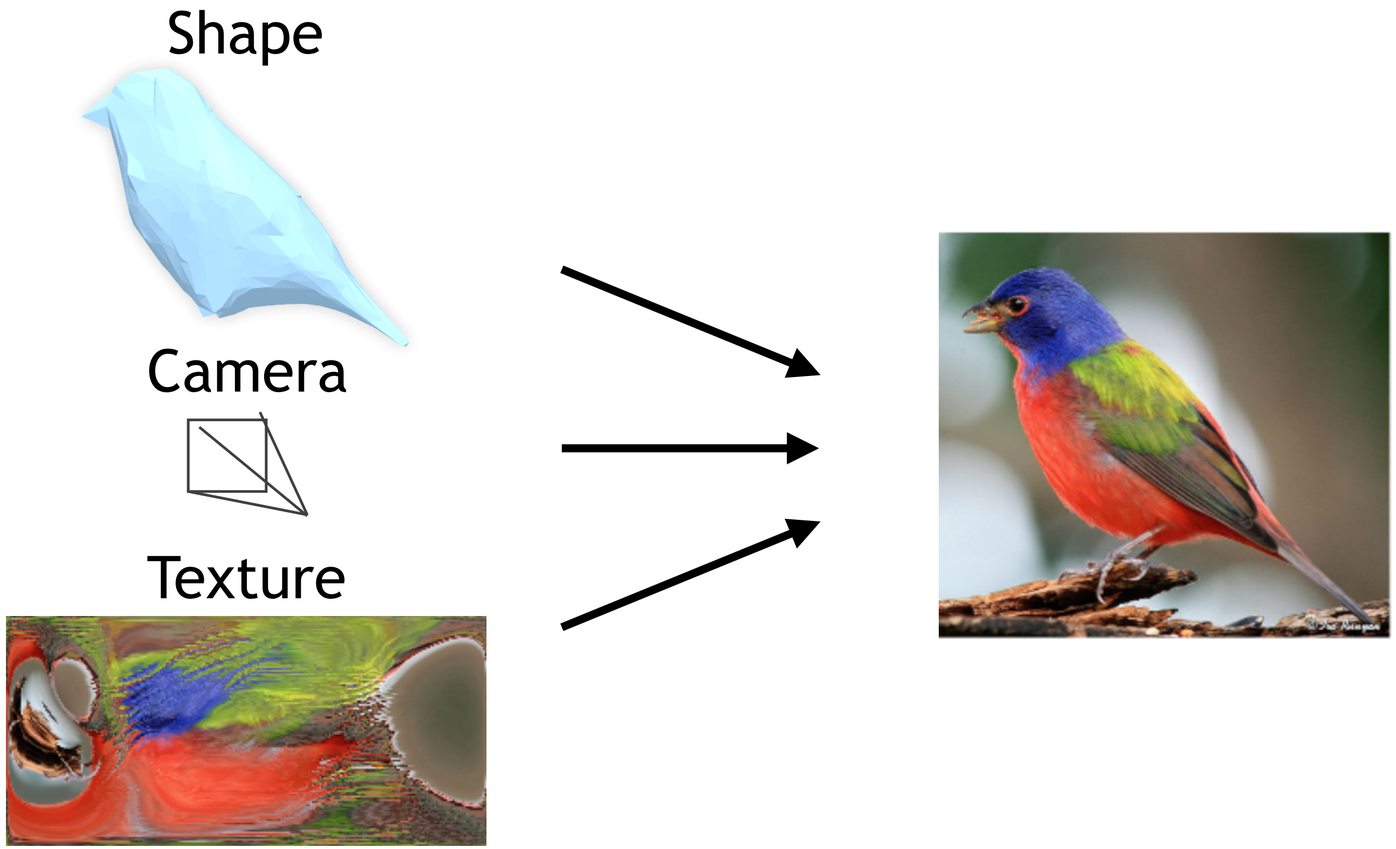
[Slides credit: Angjoo Kanazawa]

* Equal contribution

Goal:
Learn 3D only from image-based annotations

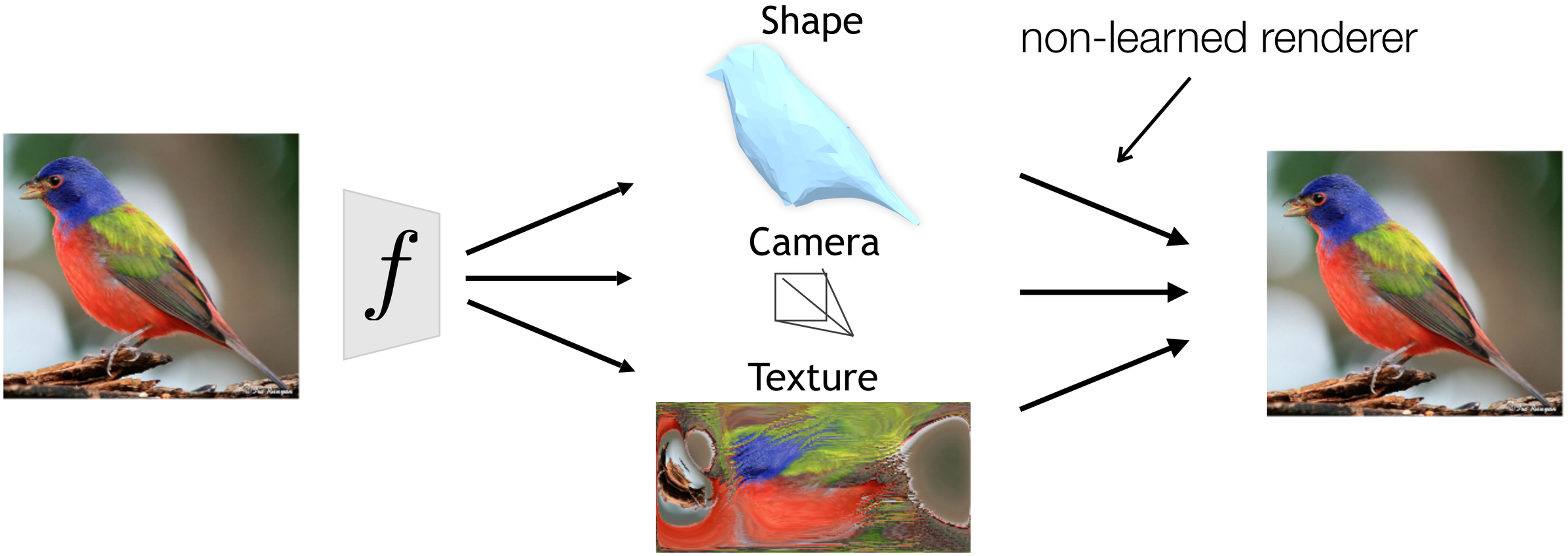
Analysis by synthesis

Find a [shape, camera, texture] combination (analysis) that renders to the image (synthesis).



Funny looking autoencoder

Find a [shape, camera, texture] combination that renders to the image.

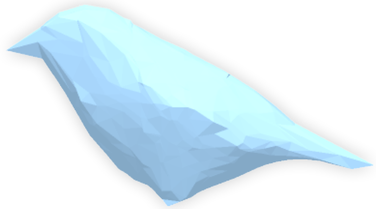


Approach

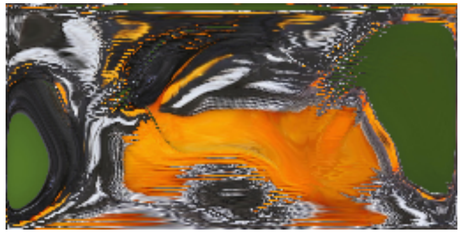
Encoder

Camera π

Shape

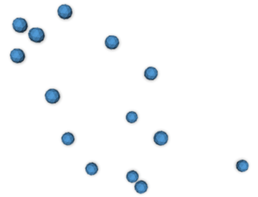


Texture



3D Keypoints

A



Losses:

Predicted, GT

Texture:

$\| \quad - \quad \|$

Mask:

$\| \quad - \quad \|$

SfM Camera:

$$\| \quad \pi \quad - \pi^{\mathrm{sfm}} \quad \|$$

Keypoints:

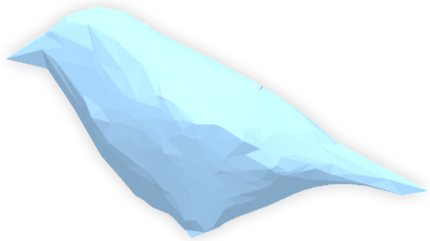
$$\left\| \pi^{\mathrm{sfm}}\left(\quad \right) - x \right\|$$

Shape Representation

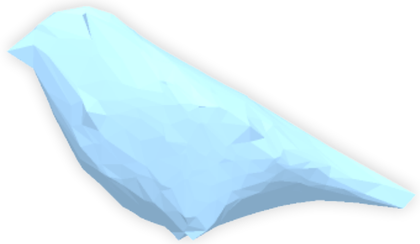
Predicted Shape

Learned Mean Shape

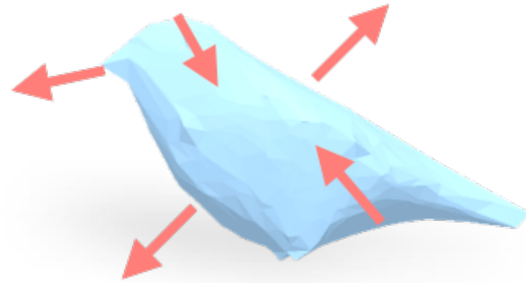
Shape Deformation



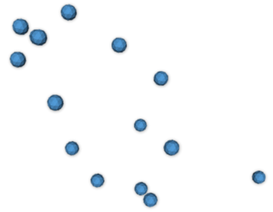
=



+

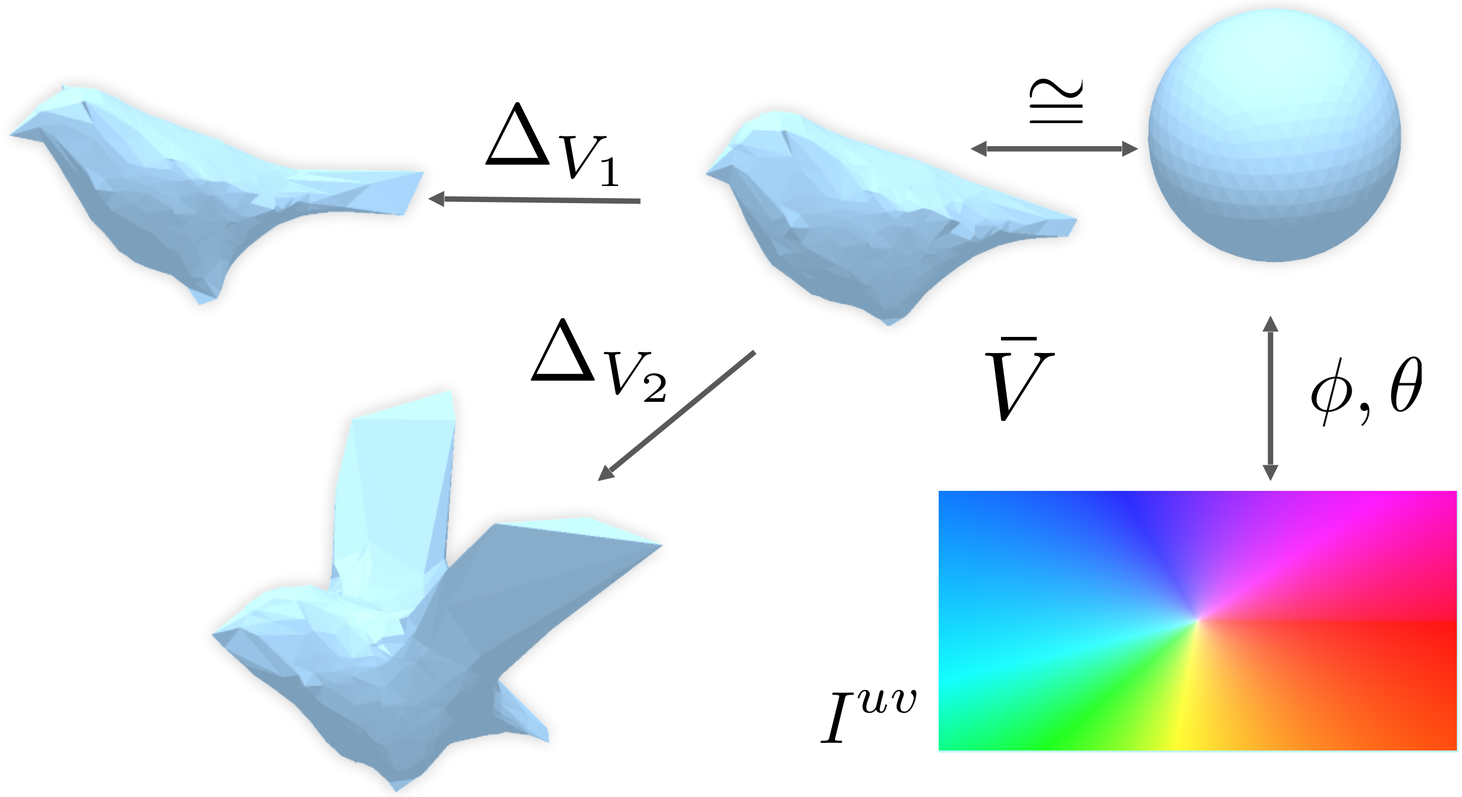


A

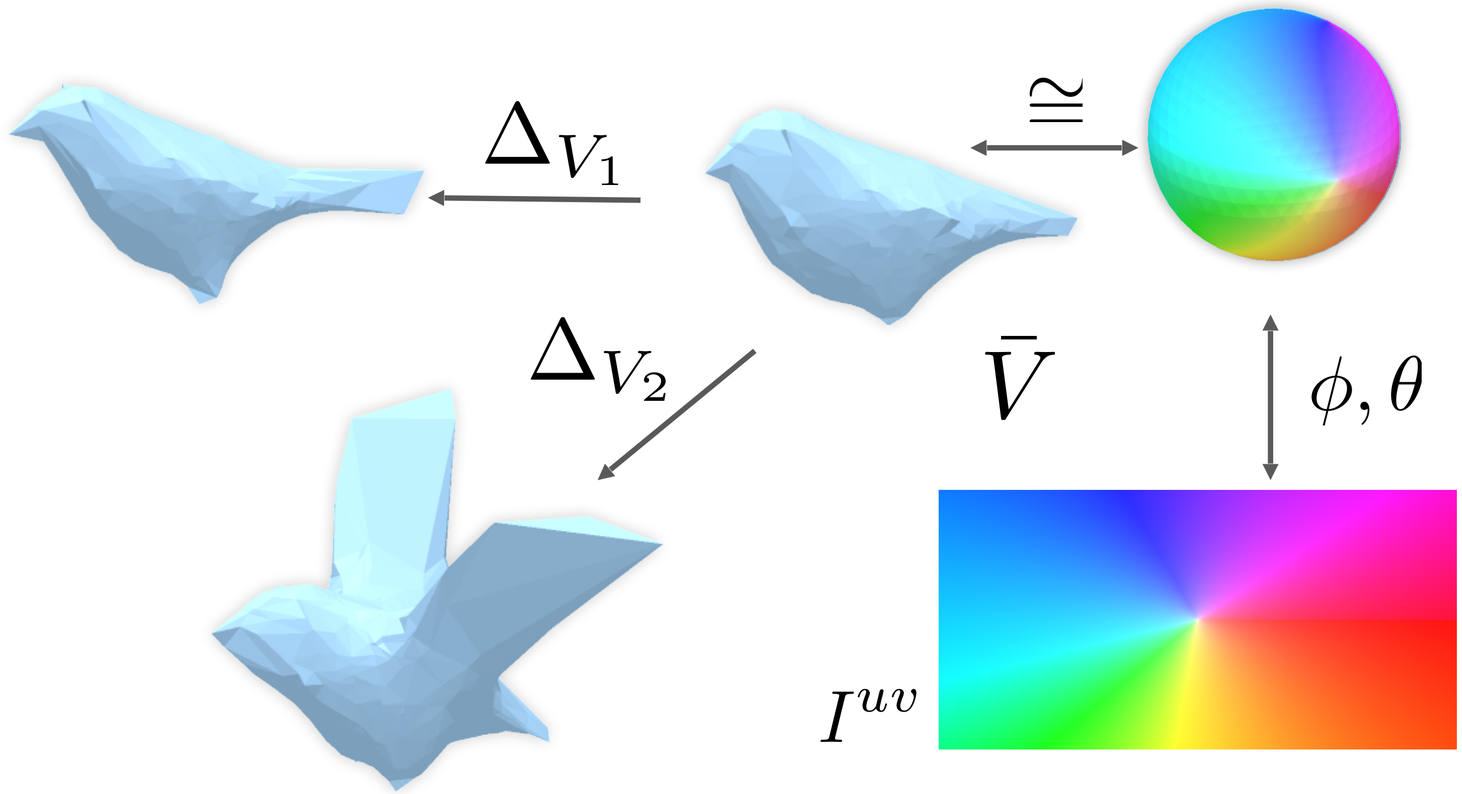


3D keypoints

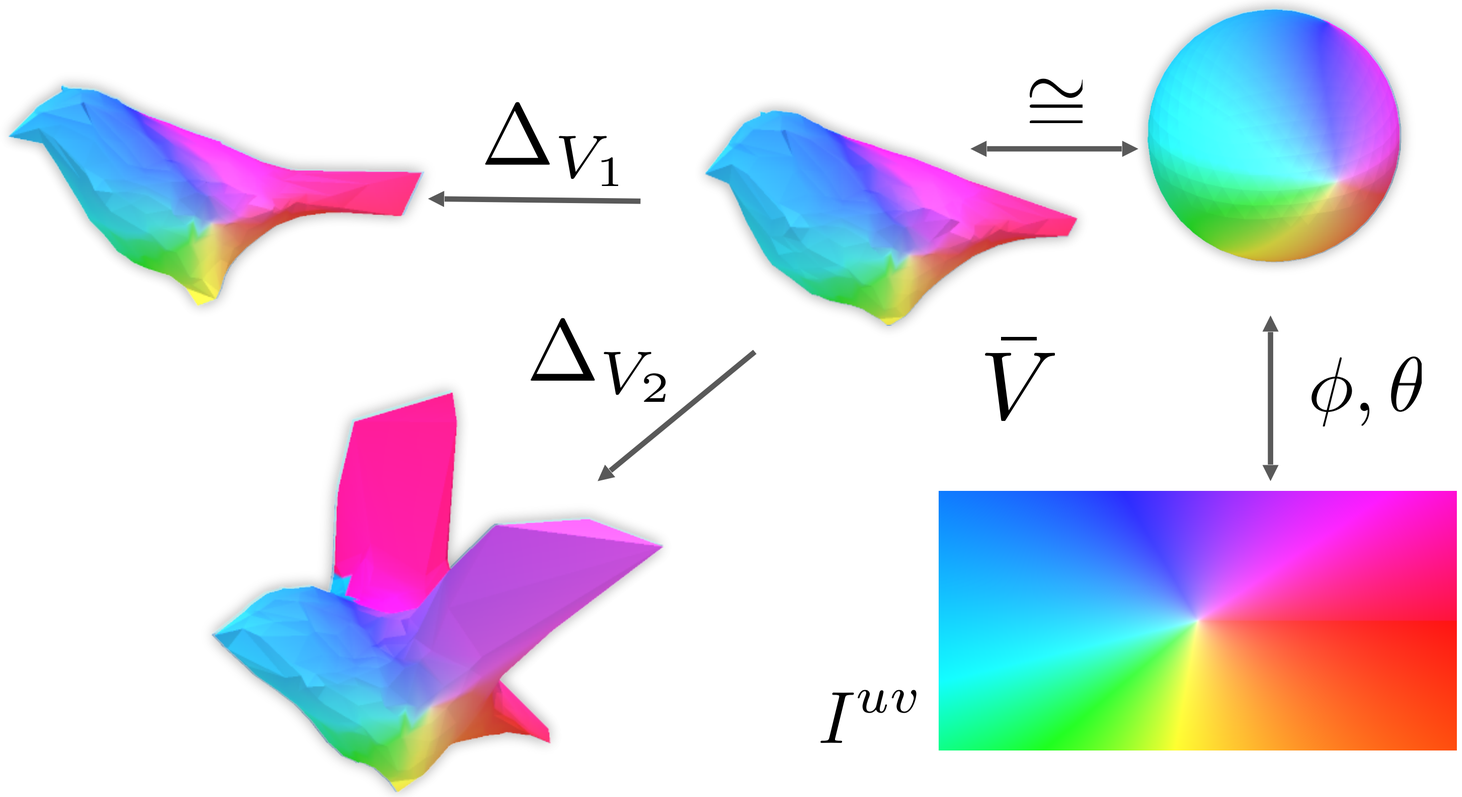
Texture Representation



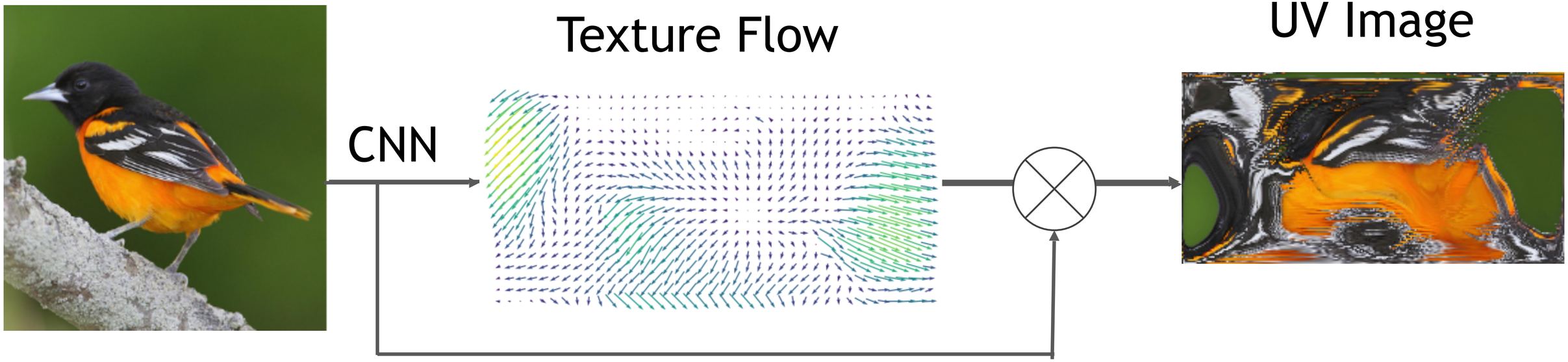
Texture Representation



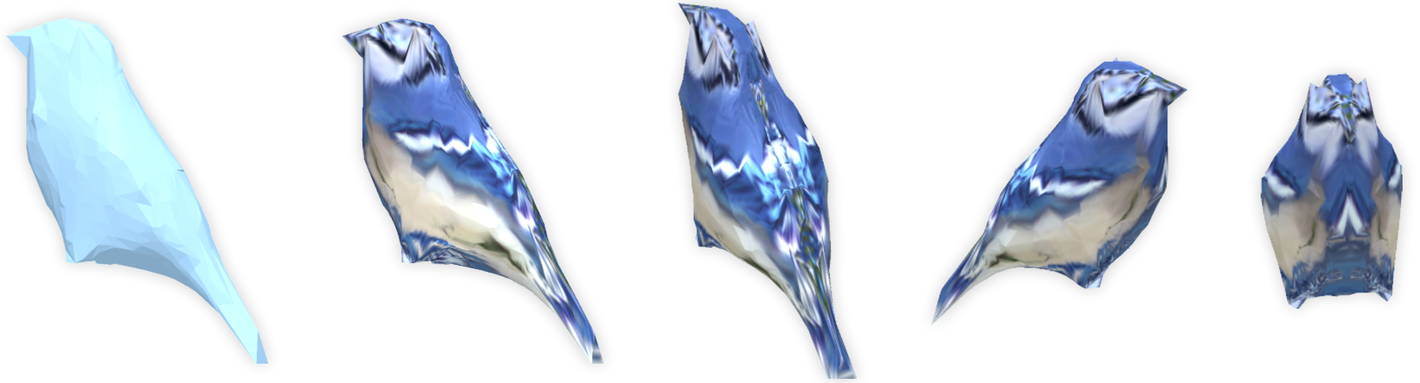
Texture Representation



Texture as UV Image Prediction



Results



Texture Transfer

Texture Transfer