Semantic & Panoptic Segmentation
and
Image Processing

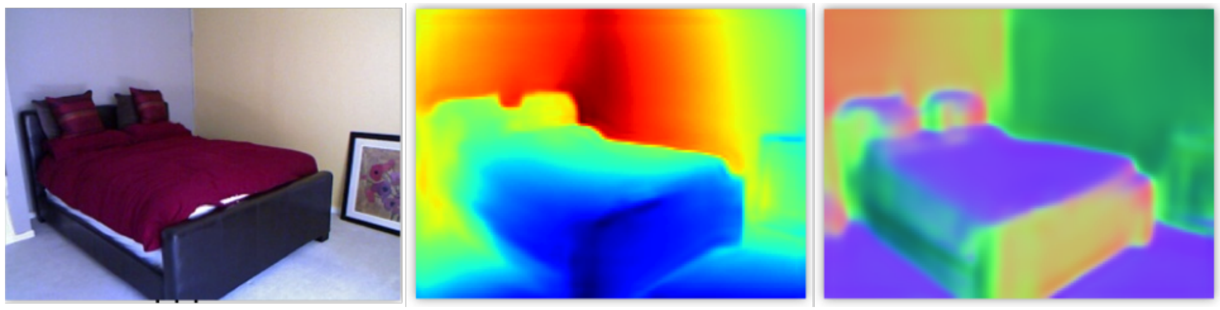
with Convnets

Lecture 11
2022

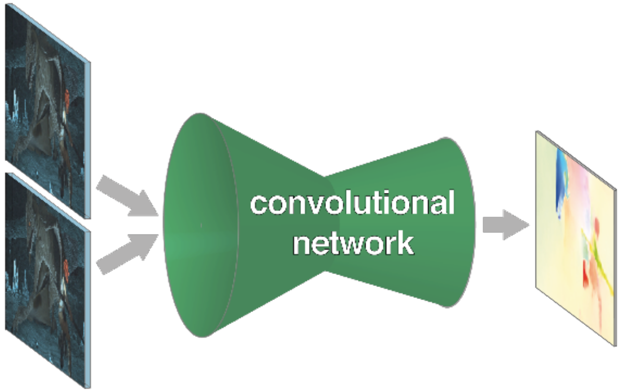
pixels in, pixels out

semantic
Segmentation
Long et al. 2015

monocular depth + normals Eigen & Fergus 2015



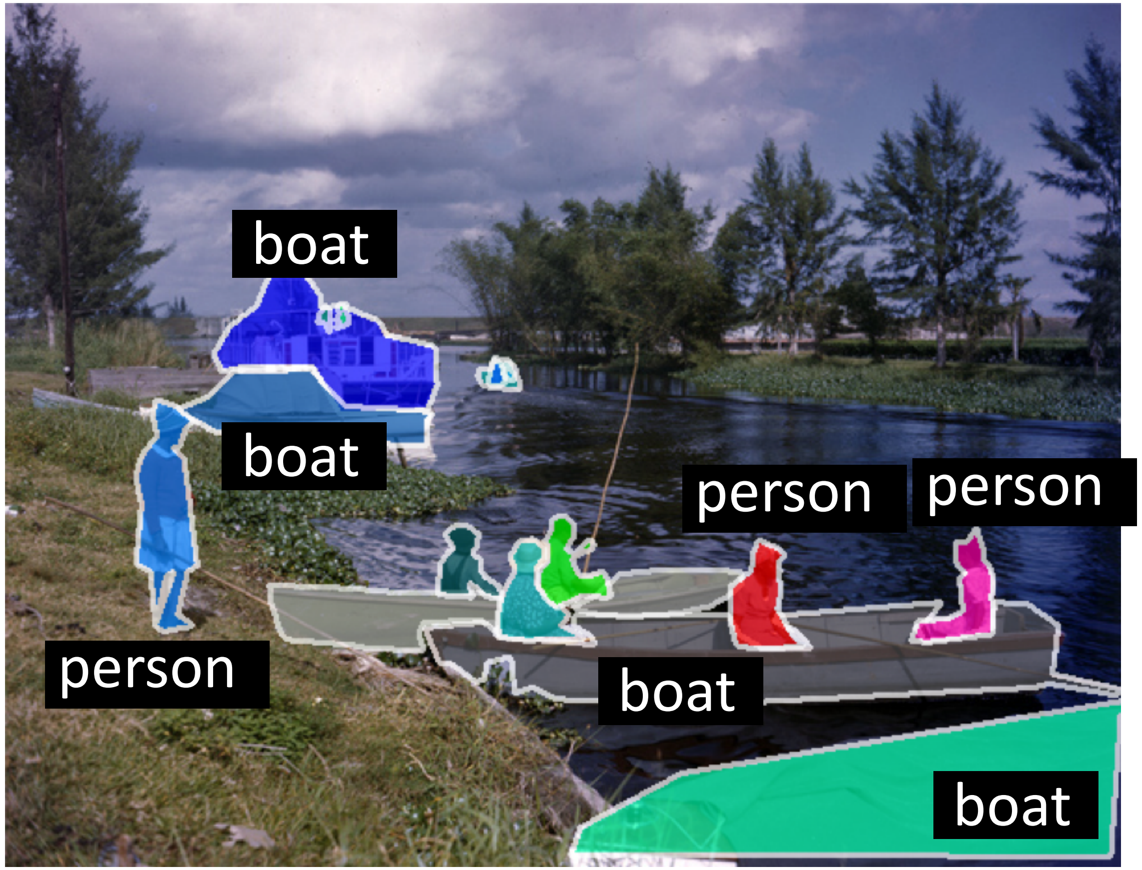
colorization
Zhang et al.2016



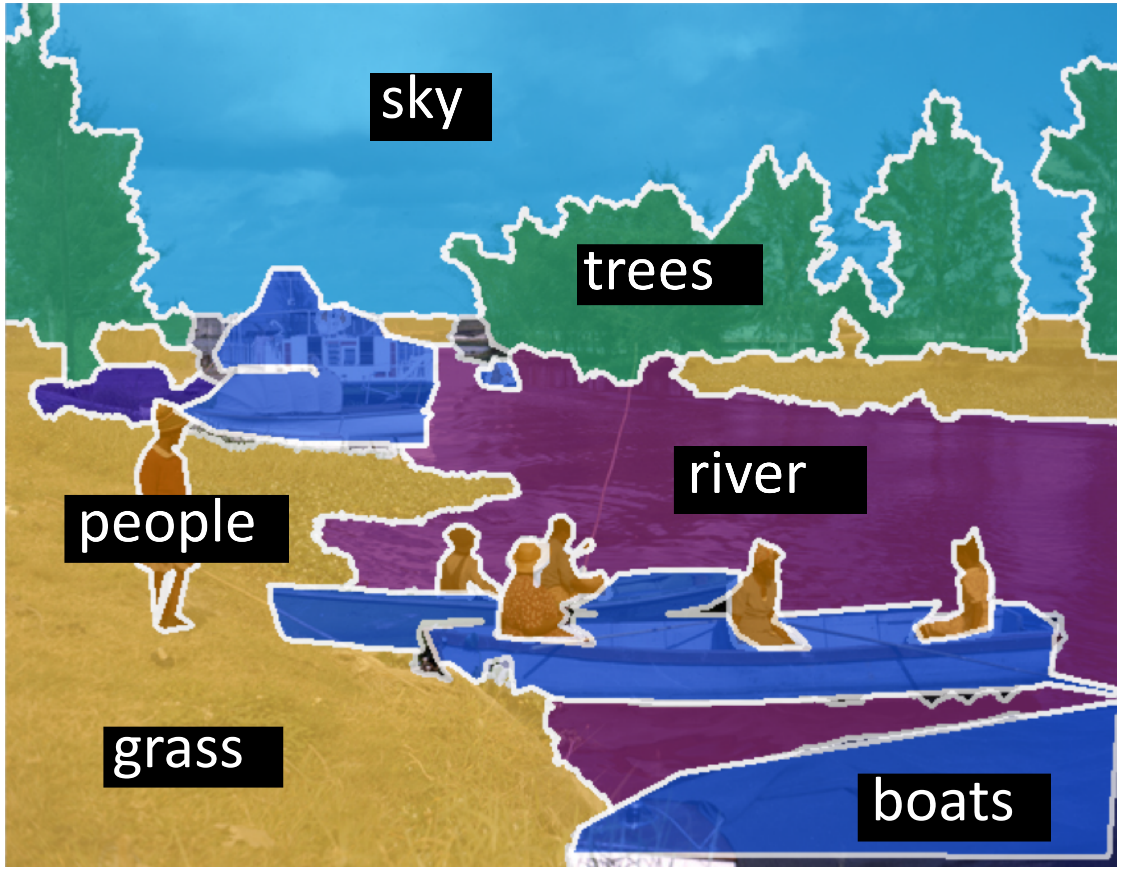
optical flow Fischer et al. 2015

boundary prediction Xie & Tu 2015

Image segmentation tasks over last 10 years



instance segmentation



semantic segmentation

[Slide: A. Kirillov]

Overview

- Semantic Segmentation
  - Fully Convolutional Nets [Shelhamer et al. 2016] https://arxiv.org/abs/1605.06211
- Panoptic Segmentation
- Image processing with Convnets

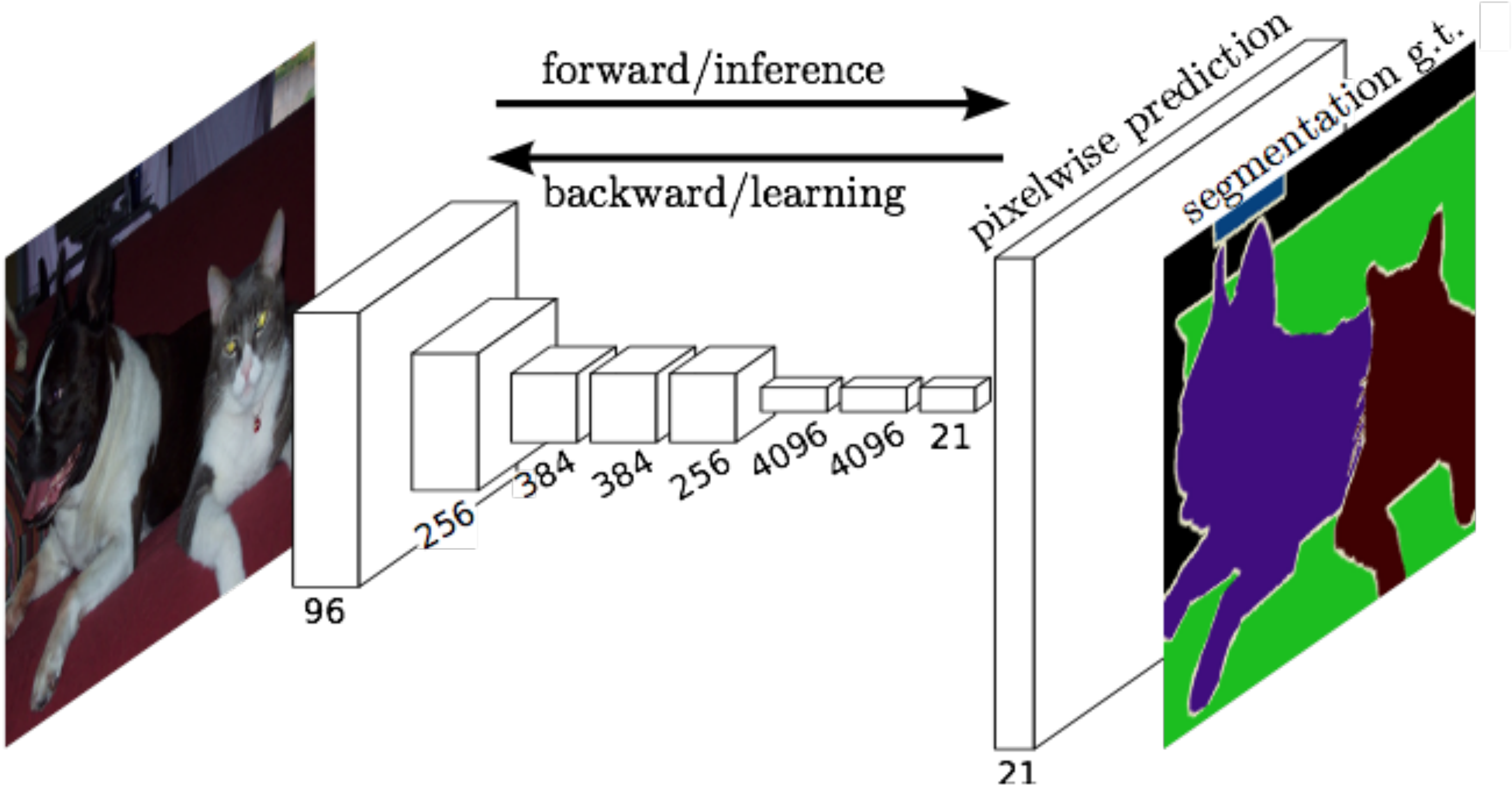
Overview

- Semantic Segmentation
  - Fully Convolutional Nets [Shelhamer et al. 2016]
    https://arxiv.org/abs/1605.06211

- Panoptic Segmentation

- Image processing with Convnets

A Fuller Understanding of Fully Convolutional Networks

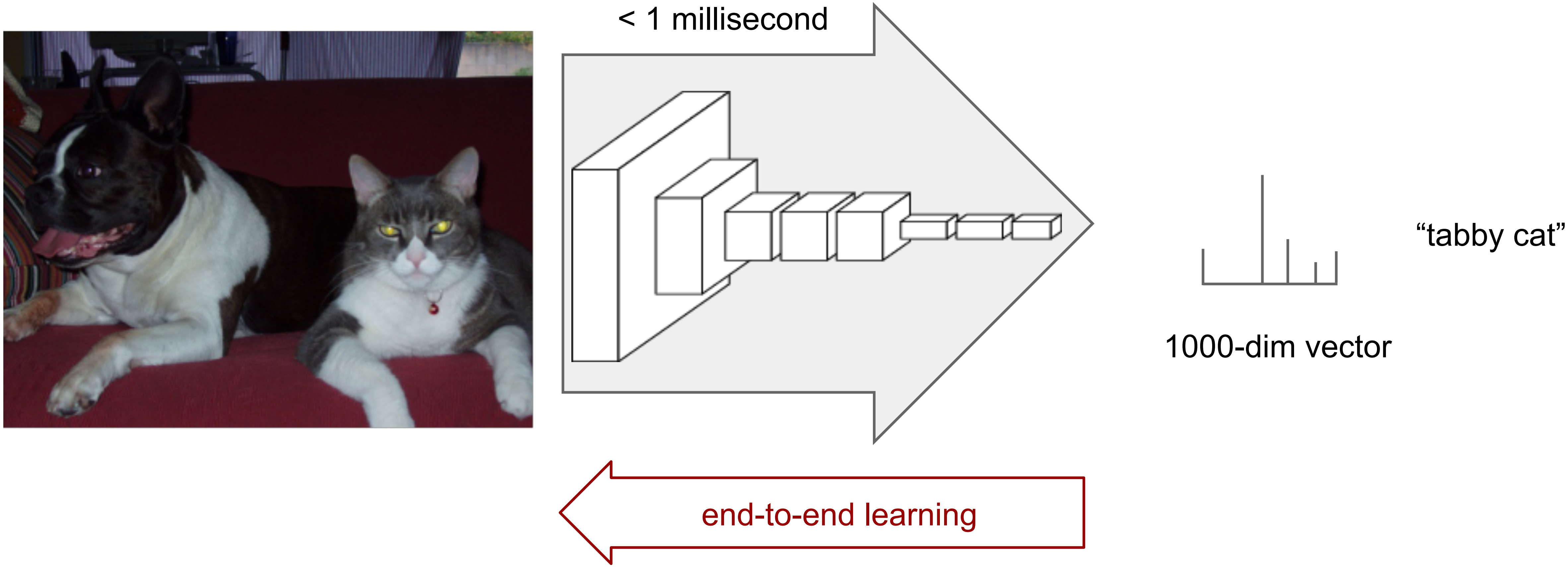


Evan Shelhamer* Jonathan Long* Trevor Darrell

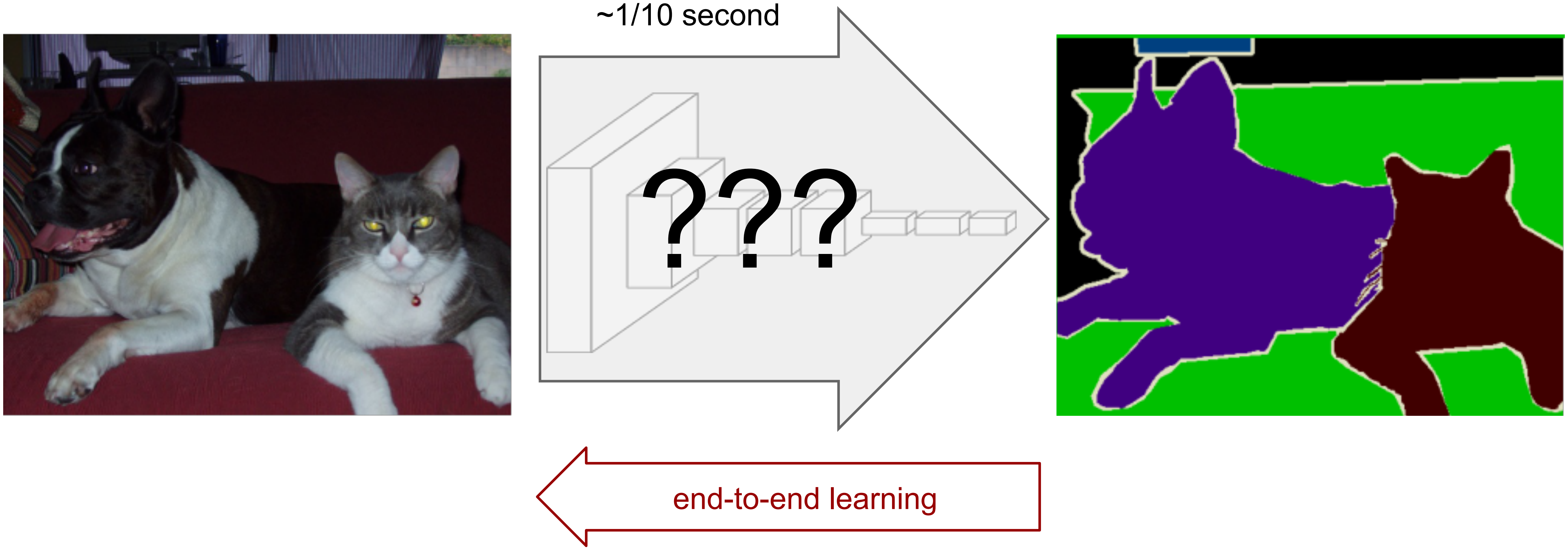
UC Berkeley

in CVPR'15, PAMI'16

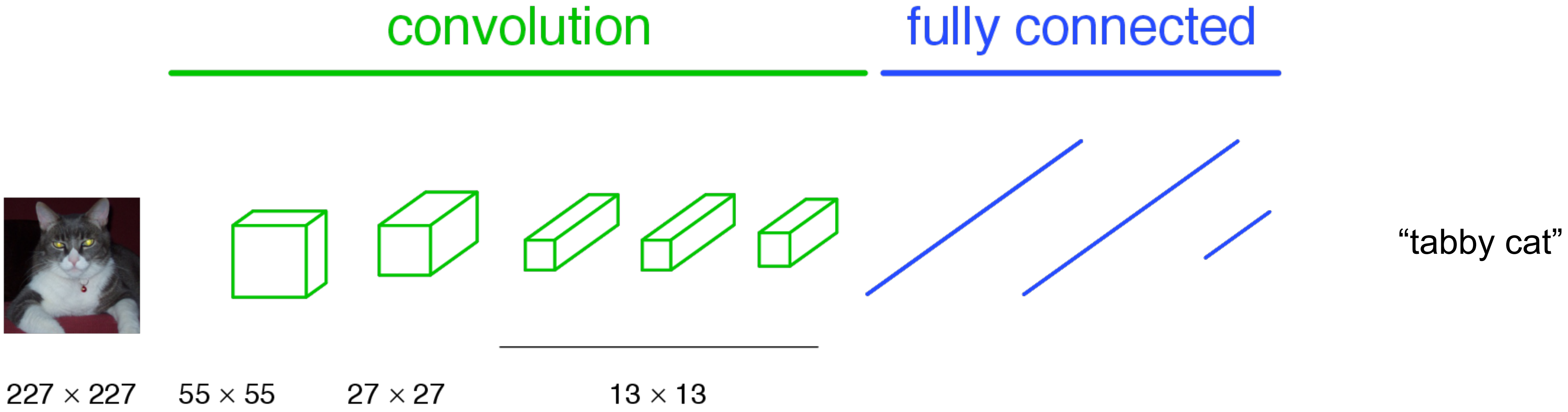
convnets perform classification



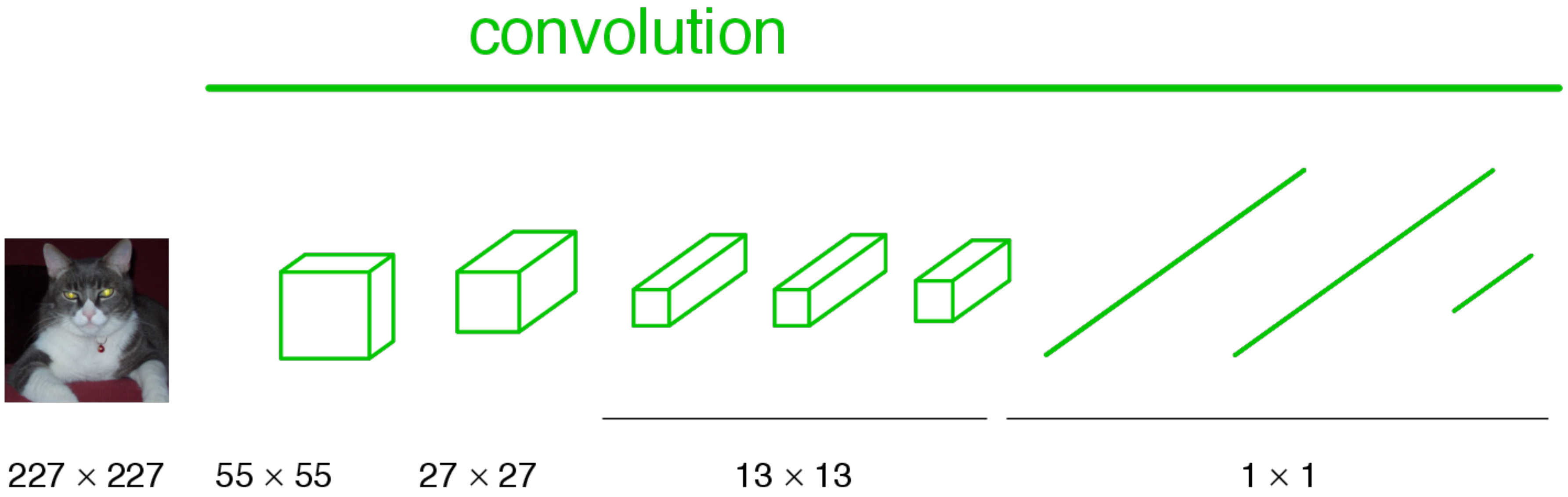
lots of pixels, little time?



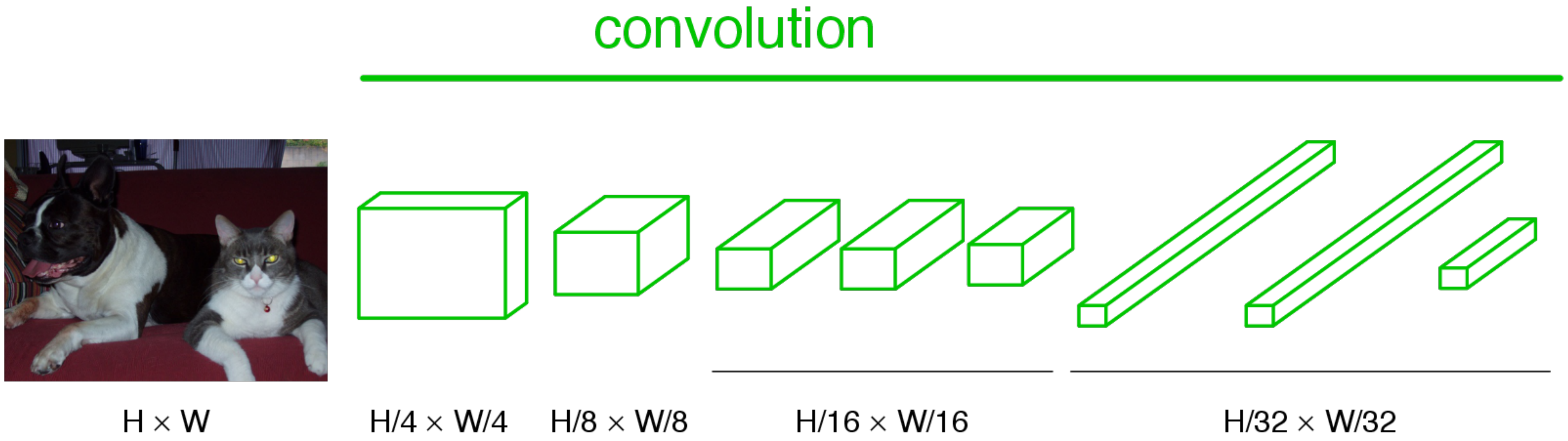
a classification network



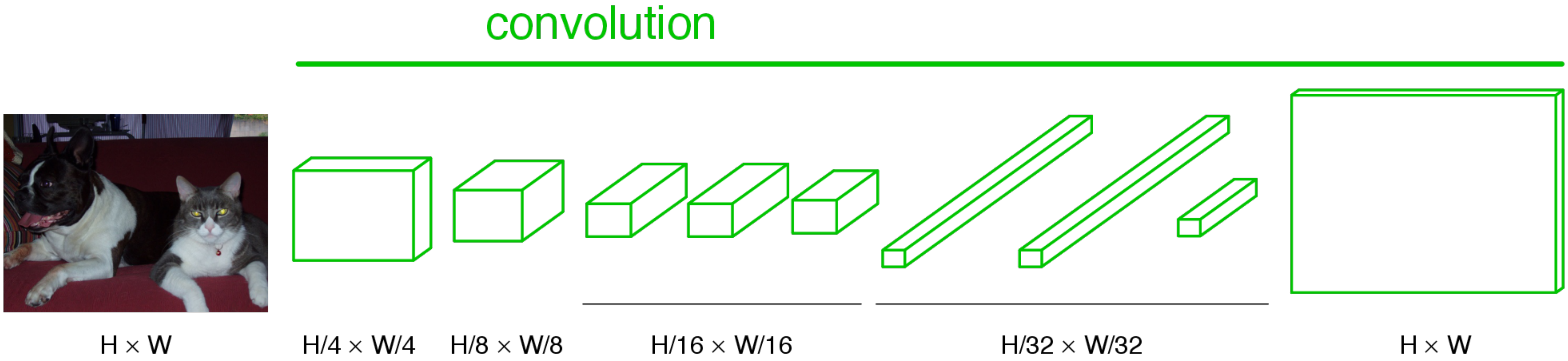
becoming fully convolutional



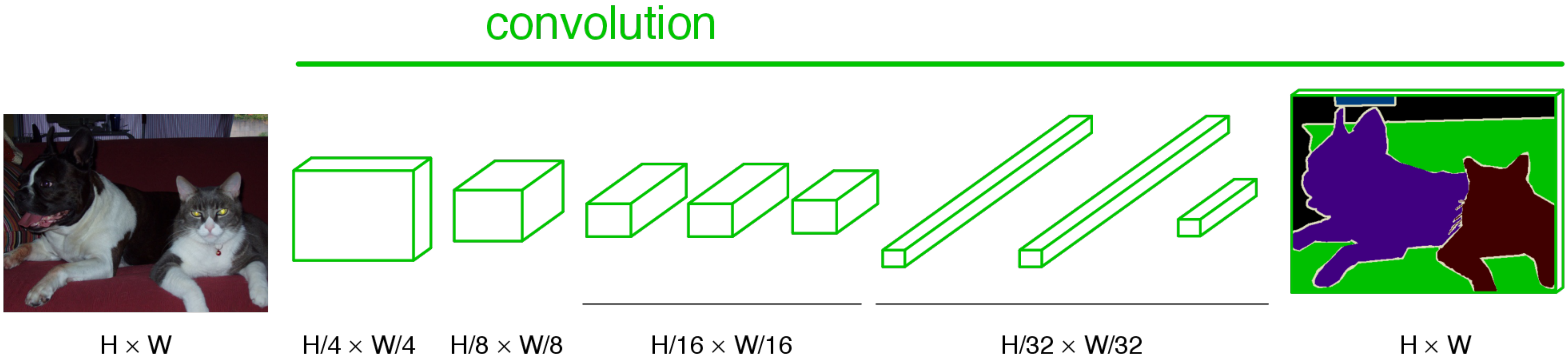
becoming fully convolutional



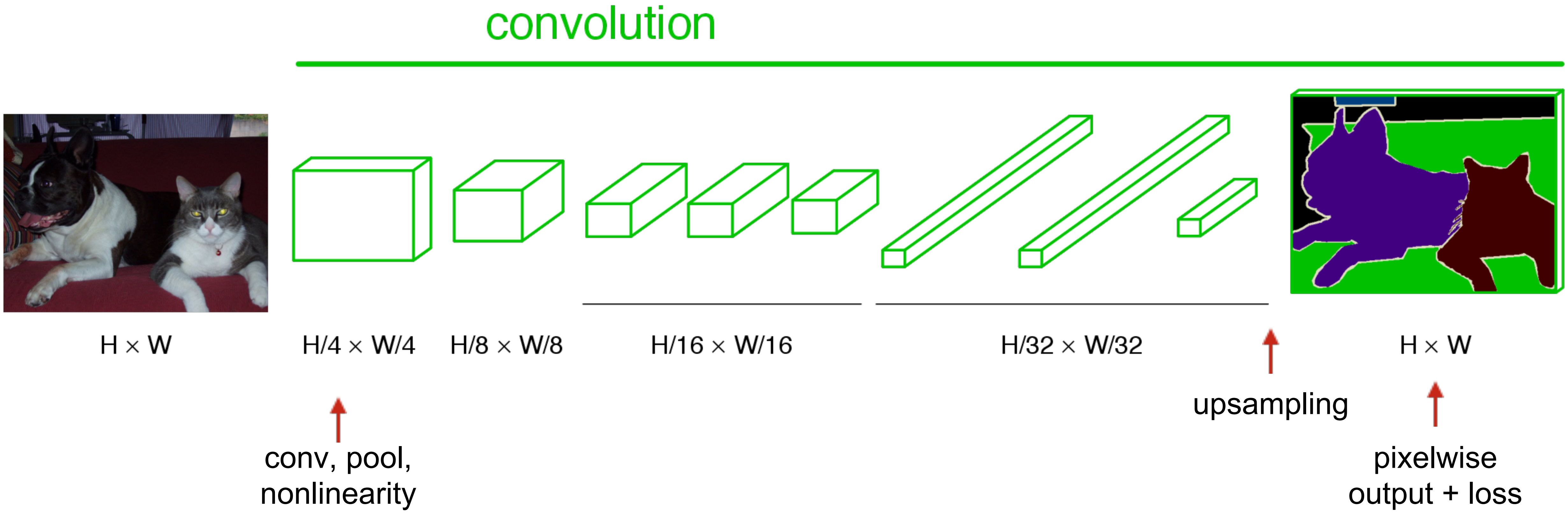
upsampling output



end-to-end, pixels-to-pixels network



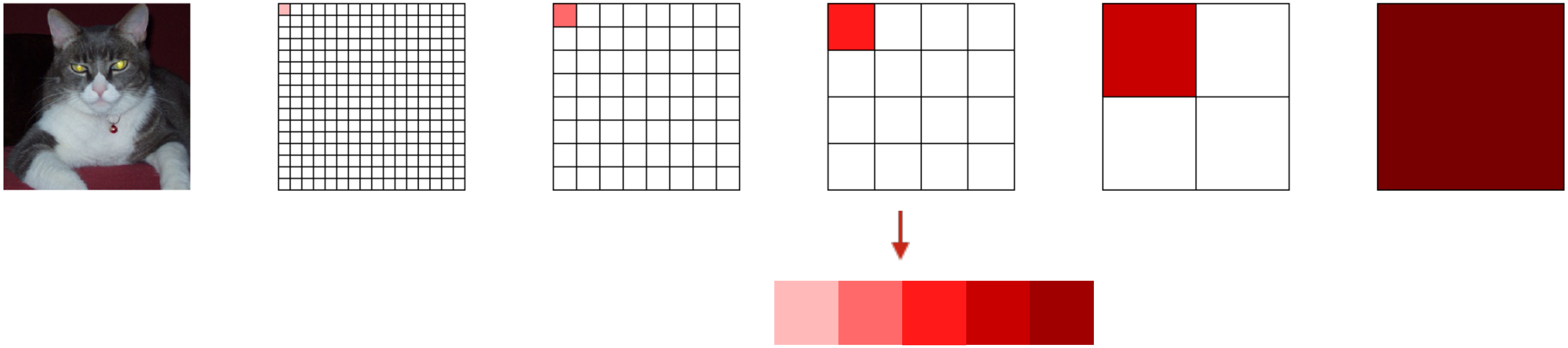
end-to-end, pixels-to-pixels network



spectrum of deep features

combine where (local, shallow) with what (global, deep)

image intermediate layers



fuse features into deep jet

(cf. Hariharan et al. CVPR15 “hypercolumn”)

skip layers

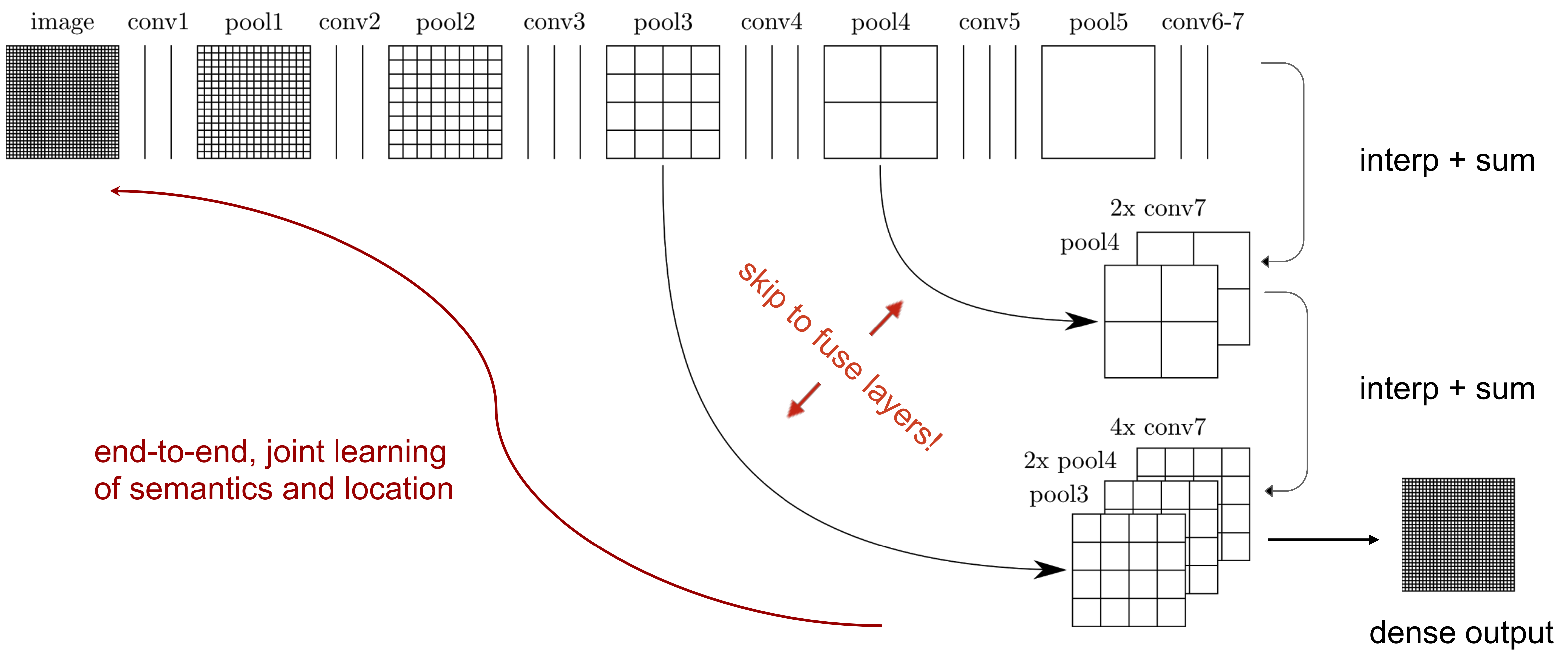
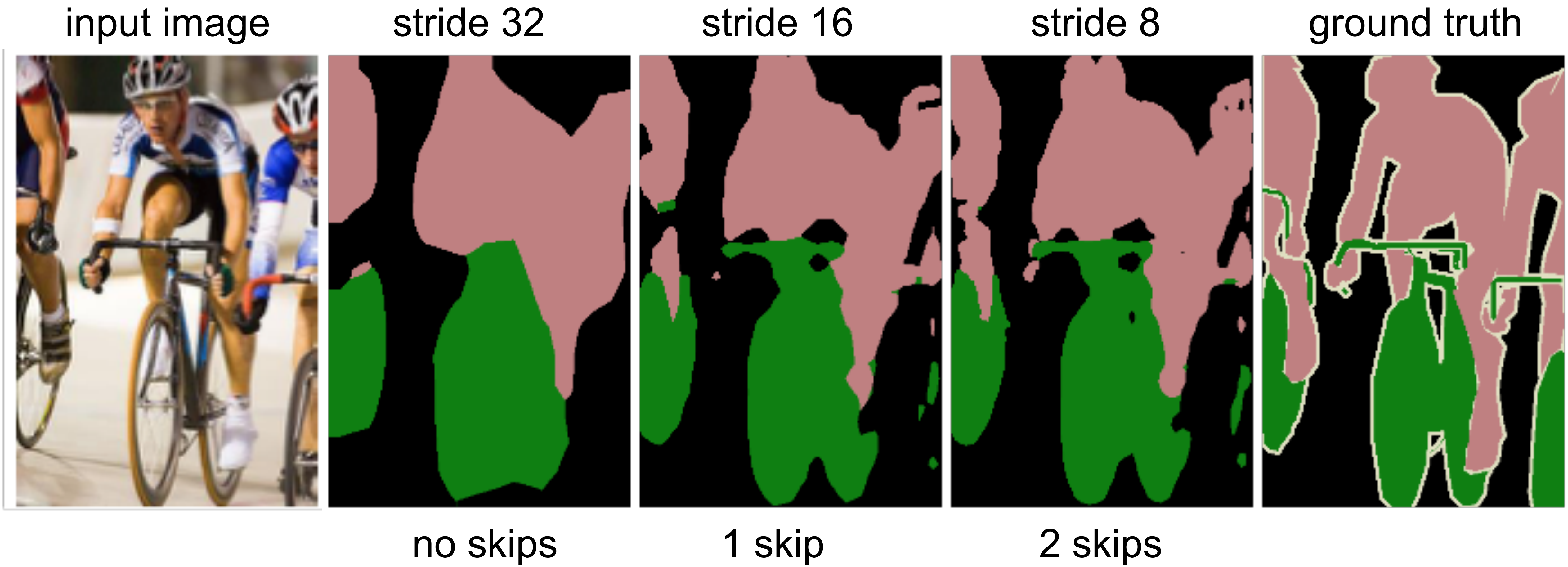
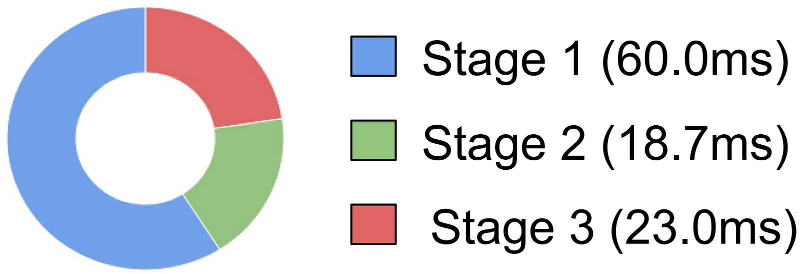


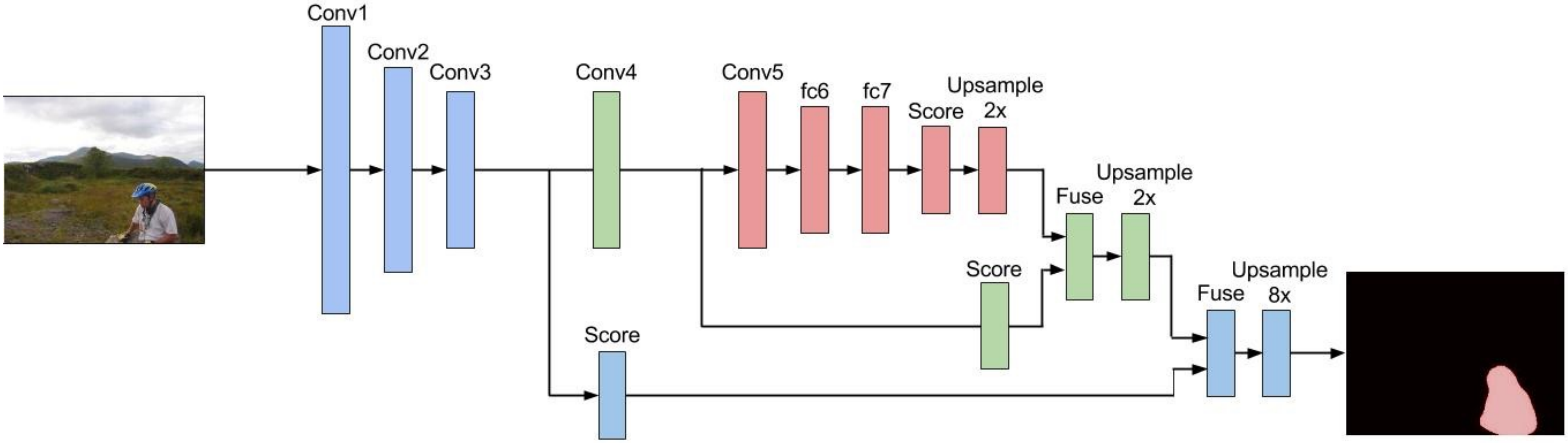
image
conv1
pool1
conv2
pool2
conv3
pool3
conv4
pool4
conv5
pool5
conv6-7
interp + sum
2x conv7
pool4
skip to fuse layers!
interp + sum
4x conv7
2x pool4
pool3
end-to-end, joint learning
of semantics and location
dense output

skip layer refinement

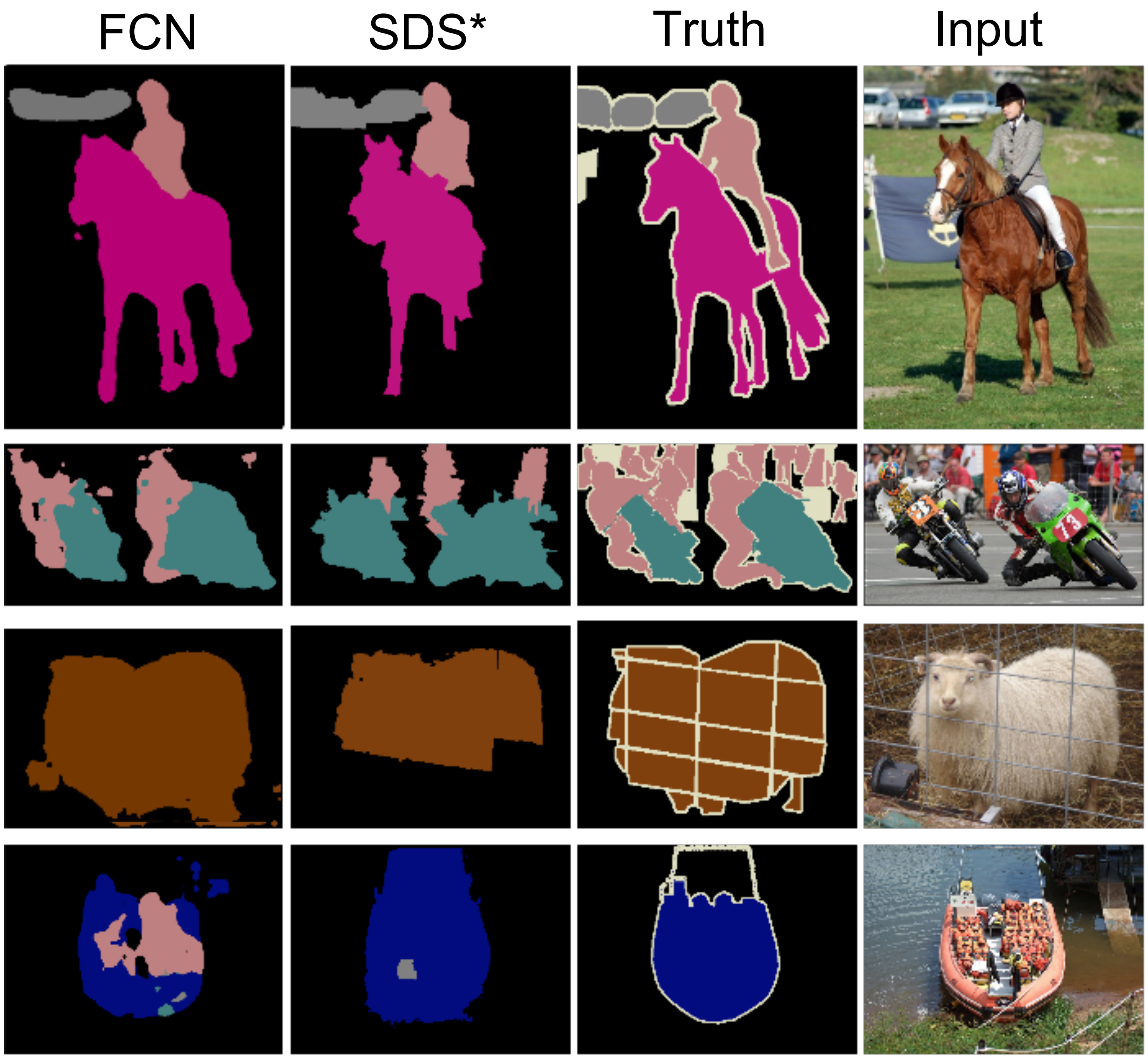


skip FCN computation





A multi-stream network that fuses features/predictions across layers

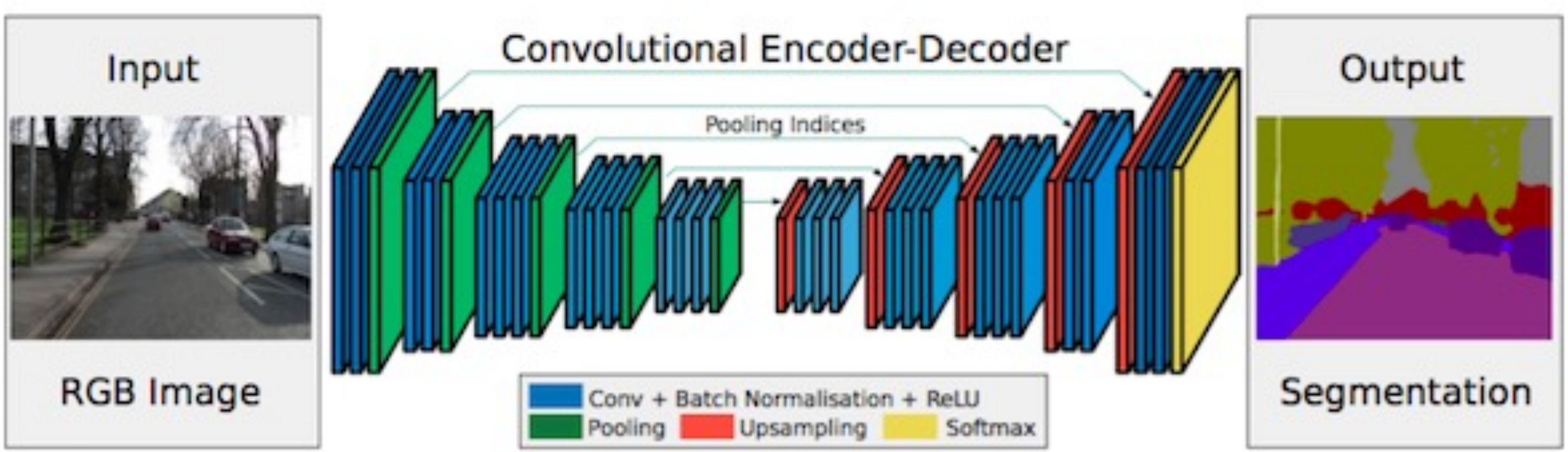


Relative to prior state-of-the-art SDS:

- 30% relative improvement for mean IoU
- 286× faster

*Simultaneous Detection and Segmentation
Hariharan et al. ECCV14

SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

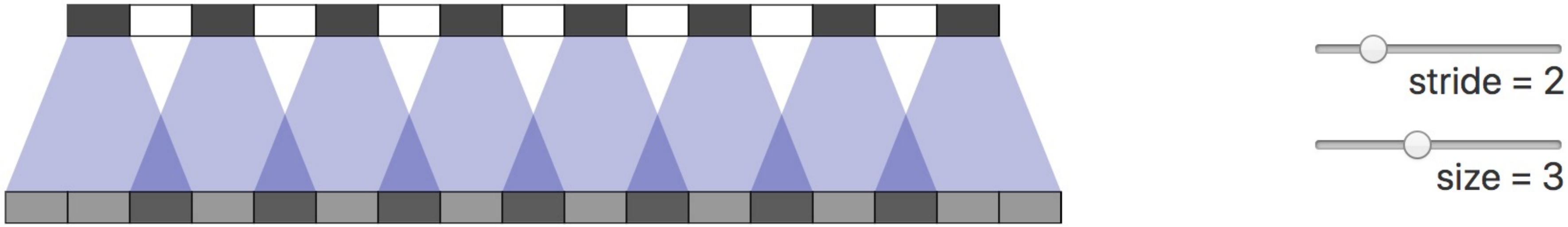


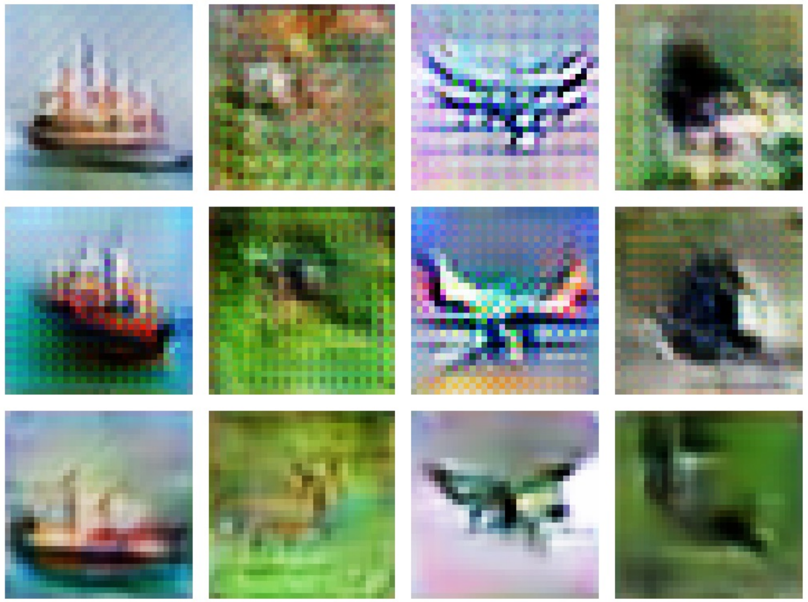
Max pooling indices transferred to decoder to improve output resolution

https://arxiv.org/abs/1511.00561

How to do the Upsampling?

Also known as Deconvolution
See https://distill.pub/2016/deconv-checkerboard/





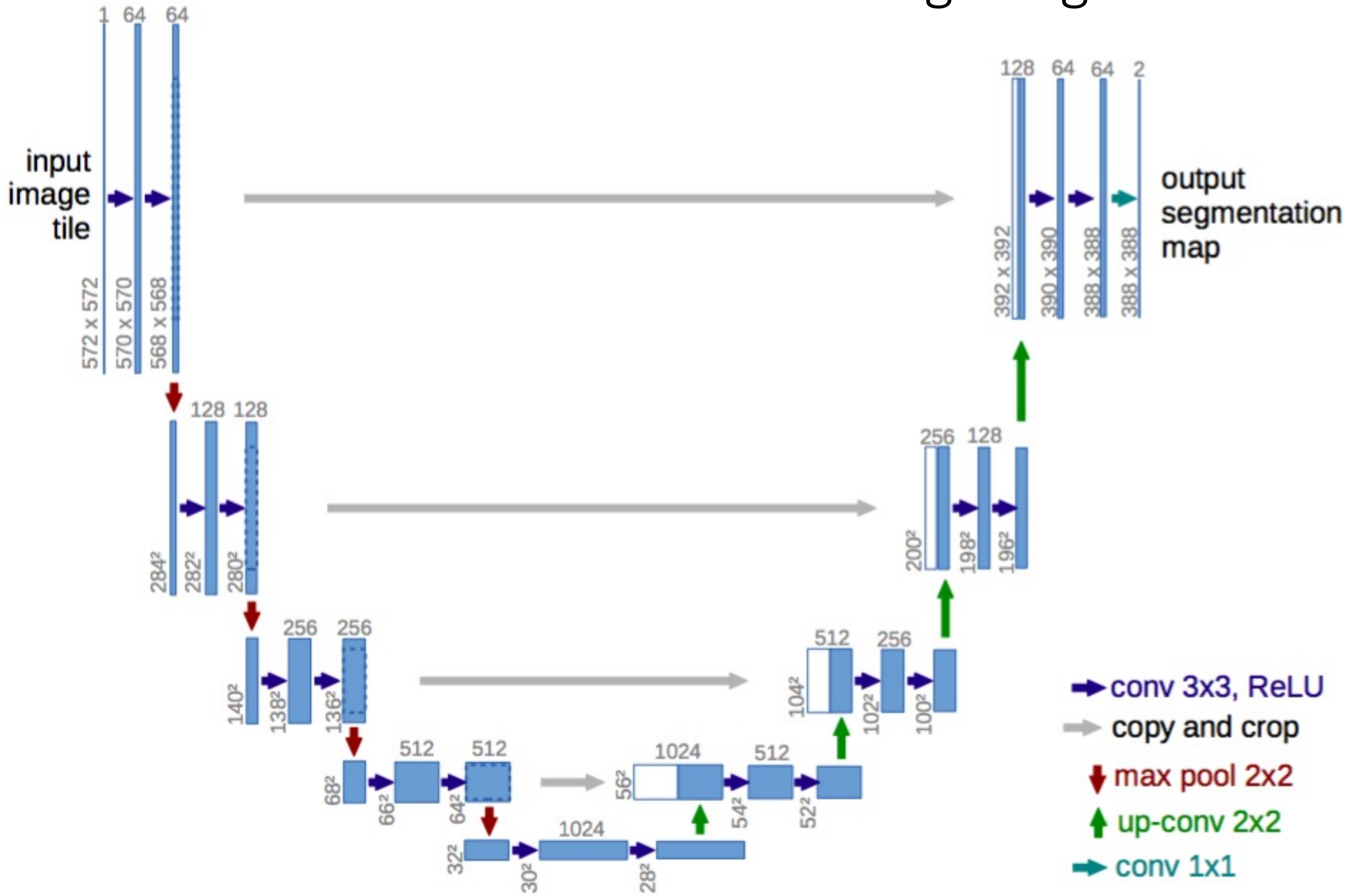
Deconv in last two layers.
Other layers use resize-convolution.
Artifacts of frequency 2 and 4.

Deconv only in last layer.
Other layers use resize-convolution.
Artifacts of frequency 2.

All layers use resize-convolution.
No artifacts.

Avoid artifacts by doing bilinear interpolation

UNet: Convolutional Networks for Biomedical Image Segmentation



https://arxiv.org/abs/1505.04597

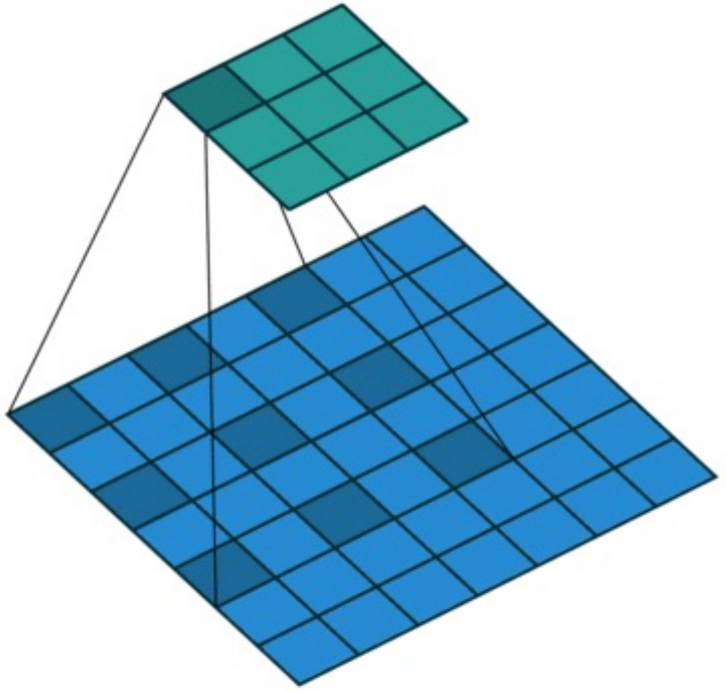
Segmentation of a 512x512 image takes less than a second on a recent GPU

Dilated / Atrous Convolutions

[Multi-Scale Context Aggregation by Dilated Convolutions, Yu and Koltun, 2015]

- No pooling operations
- Constant resolution feature maps
- Integrate increasing spatial context by special kind of dilated convolution

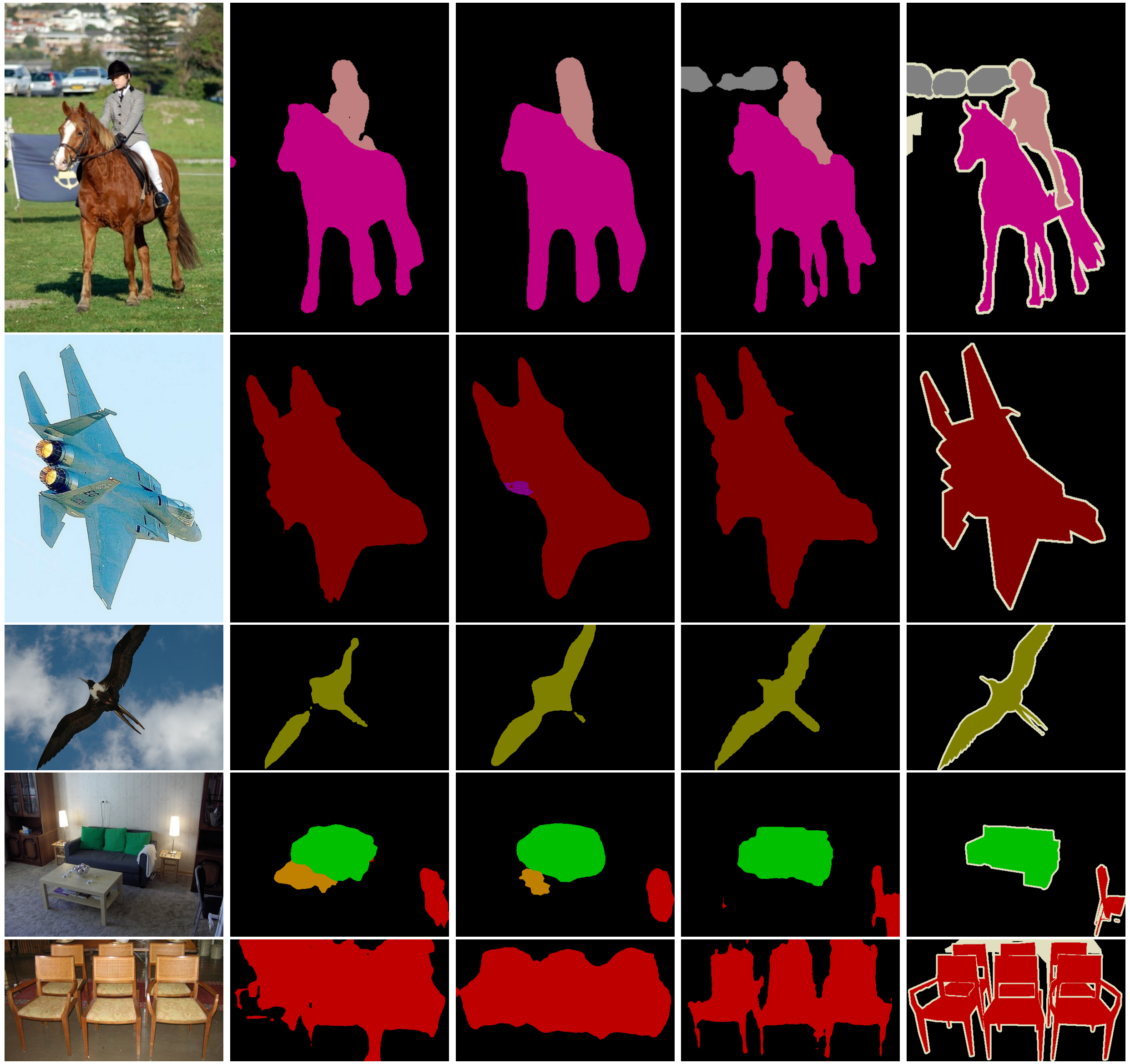
Layer	1	2	3	4	5	6	7	8
Convolution	3×3	3×3	3×3	3×3	3×3	3×3	3×3	1×1
Dilation	1	1	2	4	8	16	1	1
Truncation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Receptive field	3×3	5×5	9×9	17×17	33×33	65×65	67×67	67×67
Output channels								
Basic	C	C	C	C	C	C	C	C
Large	$2C$	$2C$	$4C$	$8C$	$16C$	$32C$	$32C$	C



- Constant 64x64 spatial resolution throughout

Dilated / Atrous Convolutions

[Multi-Scale Context Aggregation by Dilated Convolutions, Yu and Koltun, 2015]



(a) Image (b) FCN-8s (c) DeepLab (d) Our front end (e) Ground truth

Further Resources

http://blog.qure.ai/notes/semantic-segmentation-deep-learning-review

Overview

- Semantic Segmentation
  - Fully Convolutional Nets [Shelhamer et al. 2016] https://arxiv.org/abs/1605.06211
- Panoptic Segmentation
- Image processing with Convnets

Panoptic Segmentation: Task and Approaches

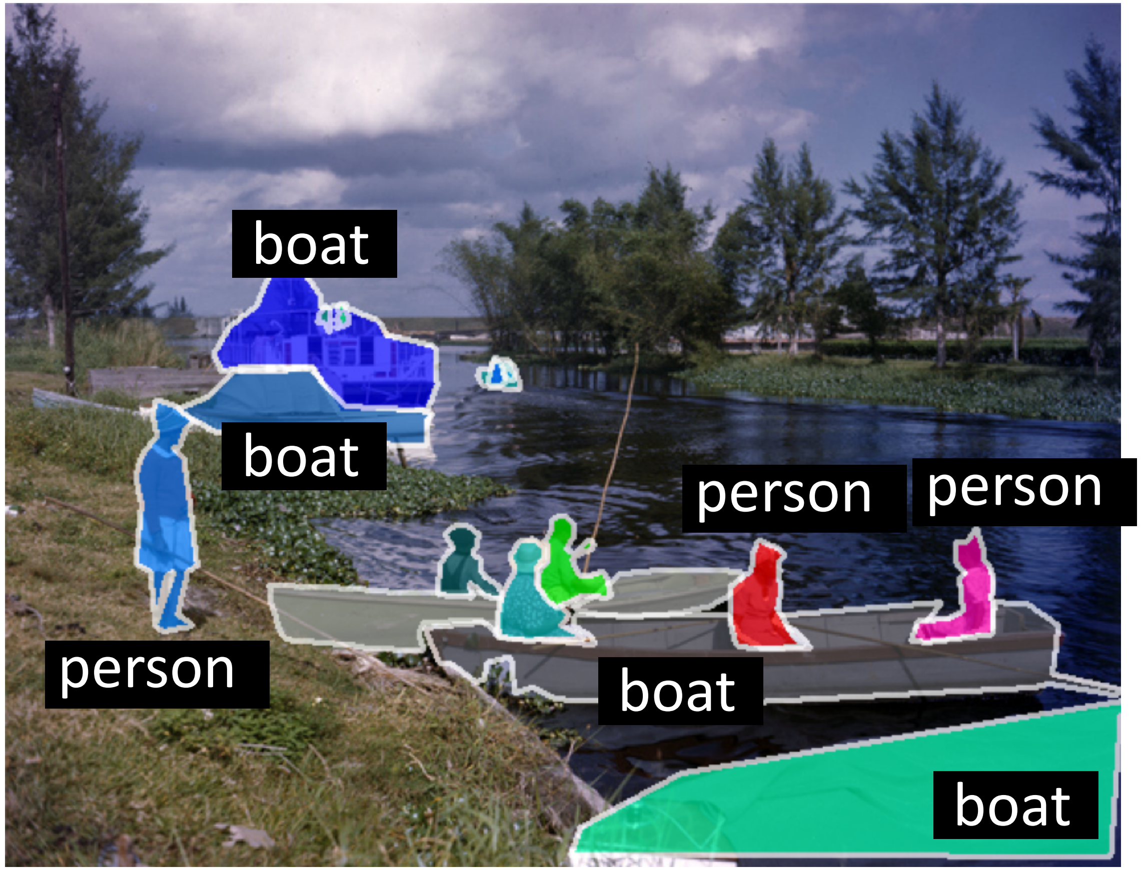
CVPR 2019 Tutorial
Visual Recognition and Beyond

Alexander Kirillov

facebook Artificial Intelligence

CVPR LONG BEACH CALIFORNIA
June 16-20, 2019

Image segmentation tasks last 10 years

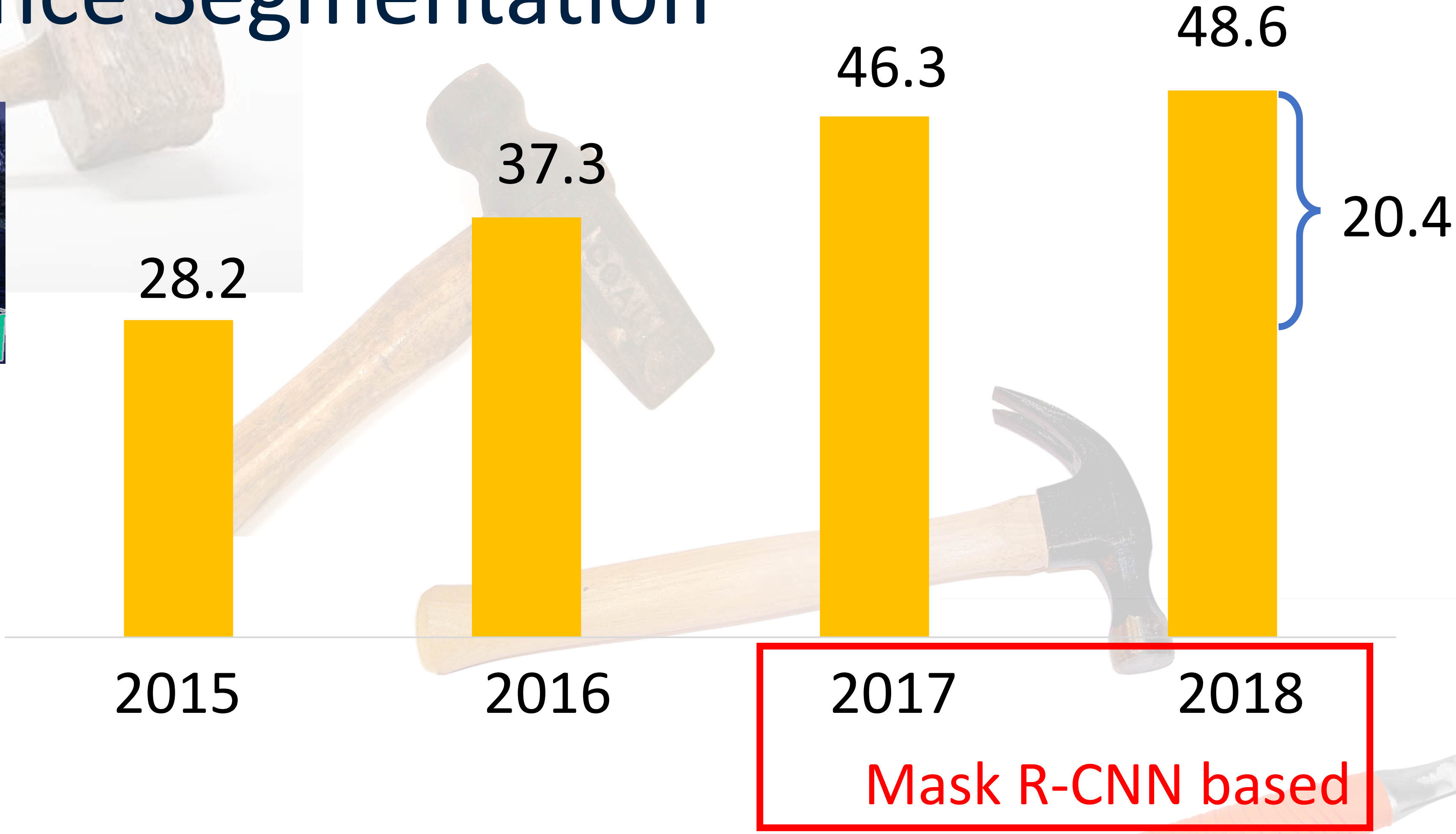


instance segmentation

delineate each object with a mask

[Slide: A. Kirillov]

Instance Segmentation



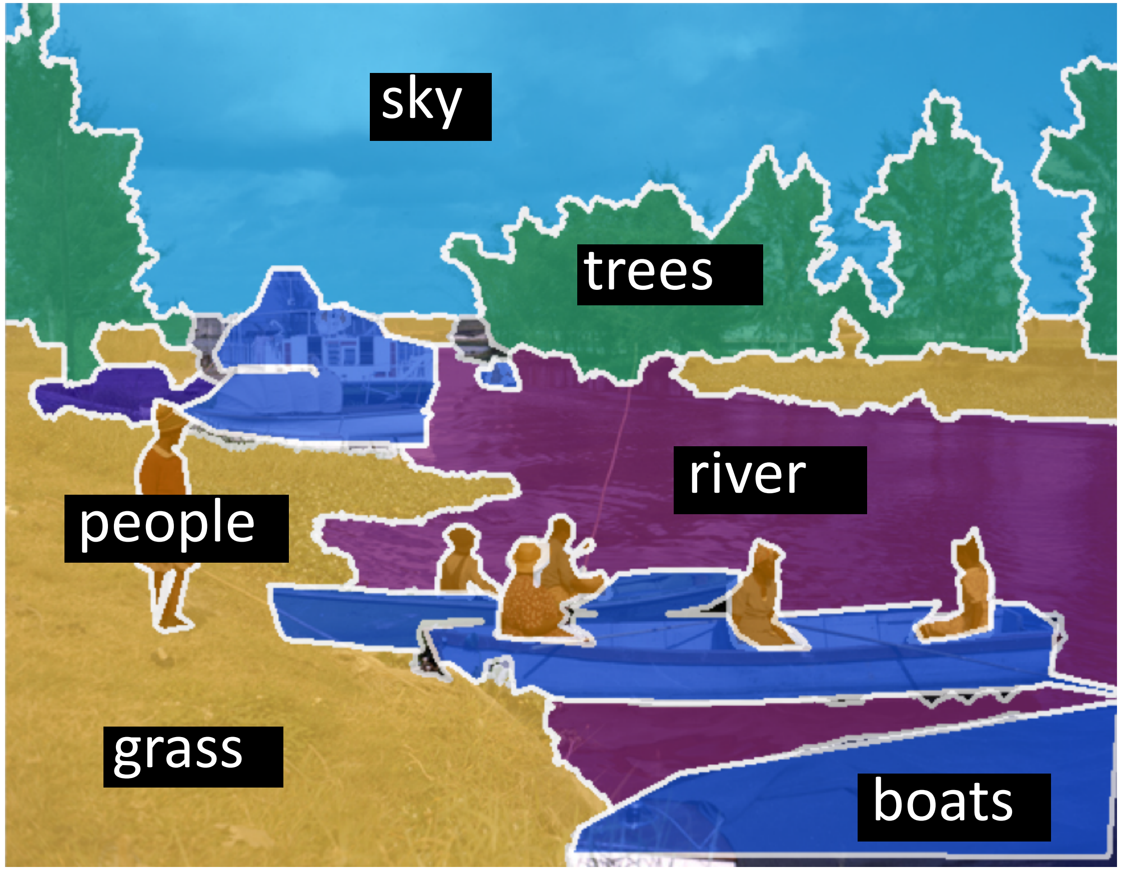
COCO-challenge winner instance segmentation AP (%)

Hammers credits:
Ross Girshick

[Slide: A. Kirillov]

Image segmentation tasks last 10 years

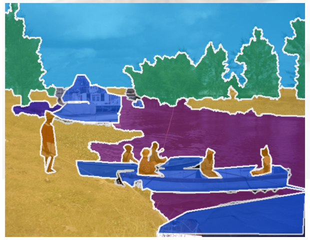
assign semantic
label to each pixel

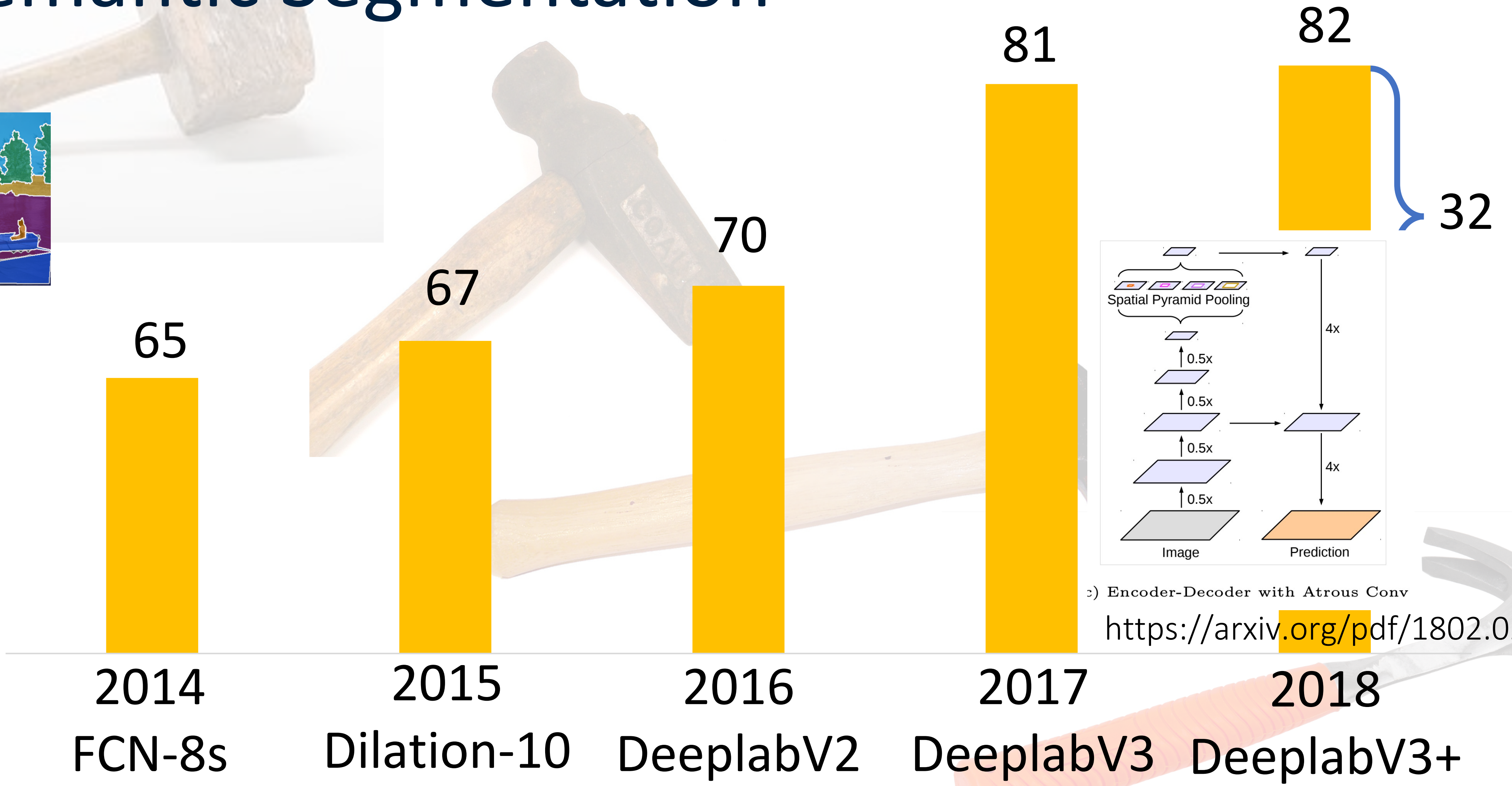


semantic segmentation

[Slide: A. Kirillov]

Semantic Segmentation





c) Encoder-Decoder with Atrous Conv

https://arxiv.org/pdf/1802.02611.pdf

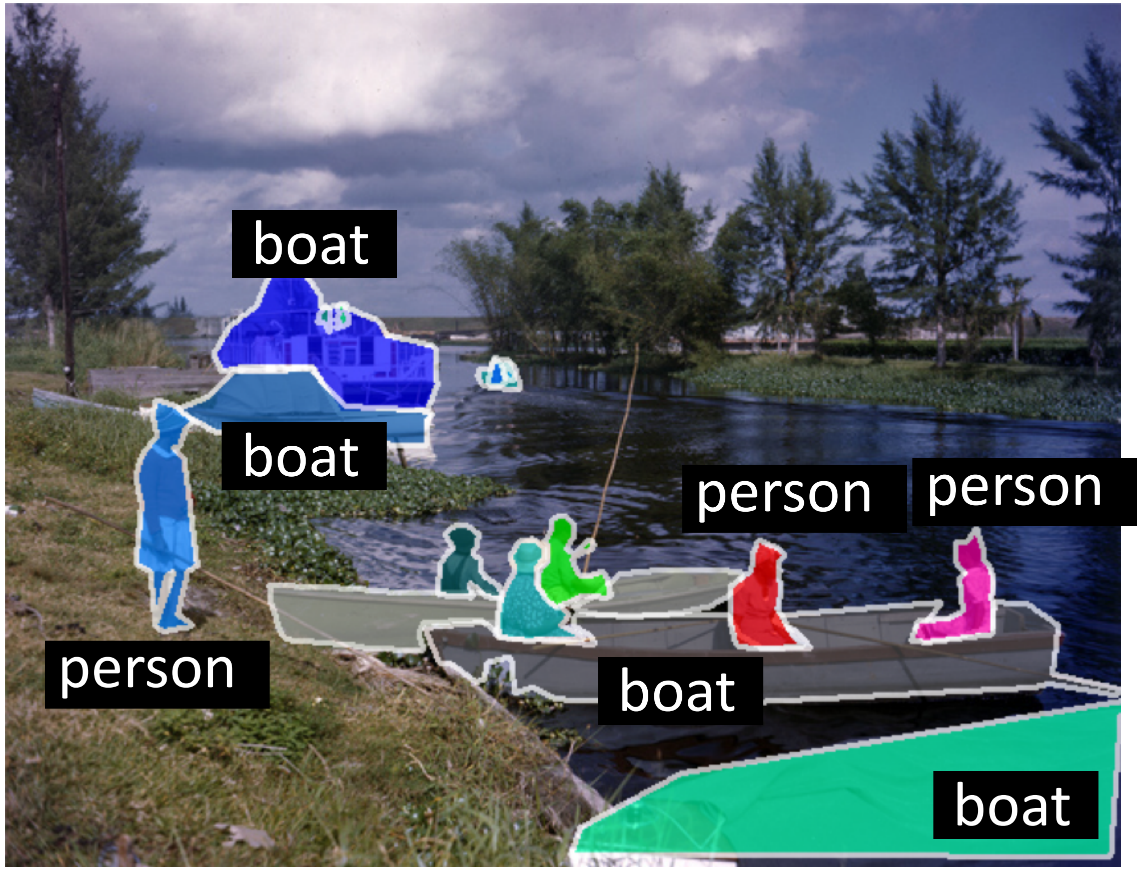
Cityscapes semantic segmentation IoU (%)

Hammers credits:
Ross Girshick

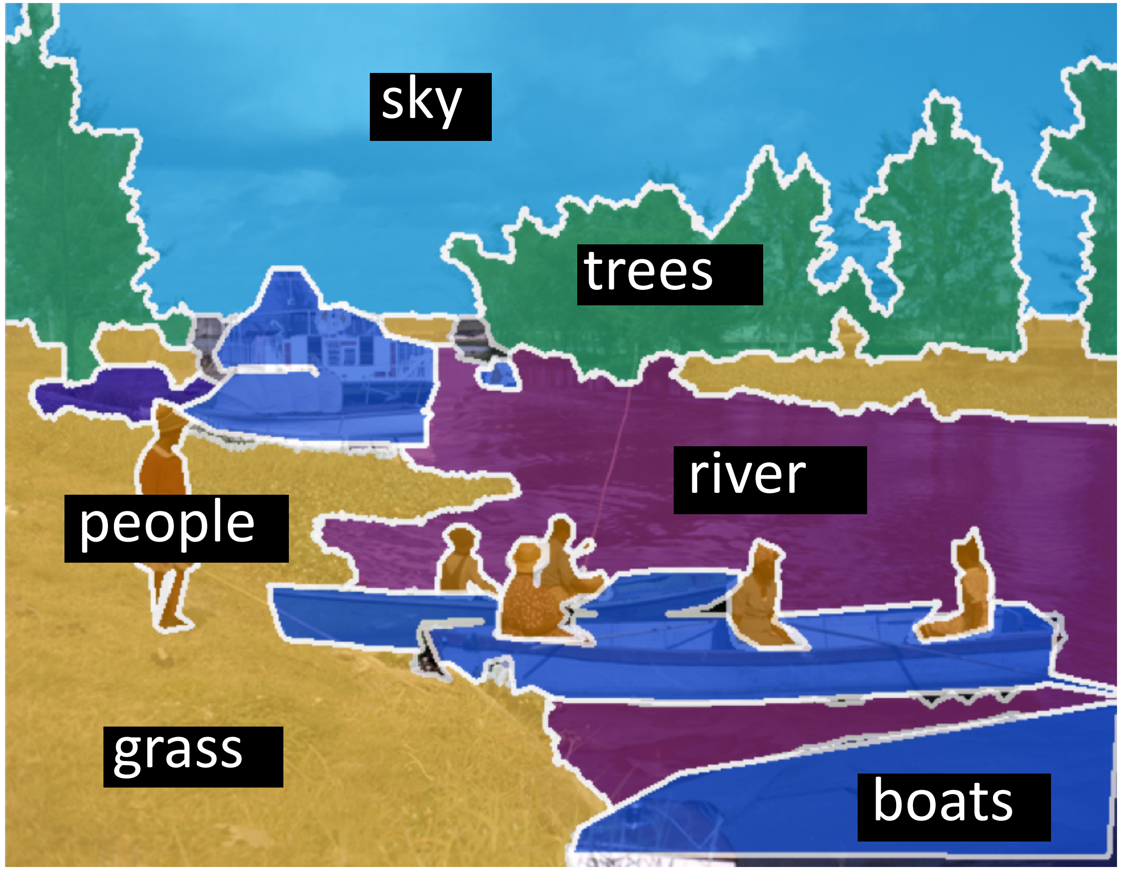
Cityscapes leaderboard performance

[Slide: A. Kirillov]

Image segmentation tasks last 10 years



instance segmentation

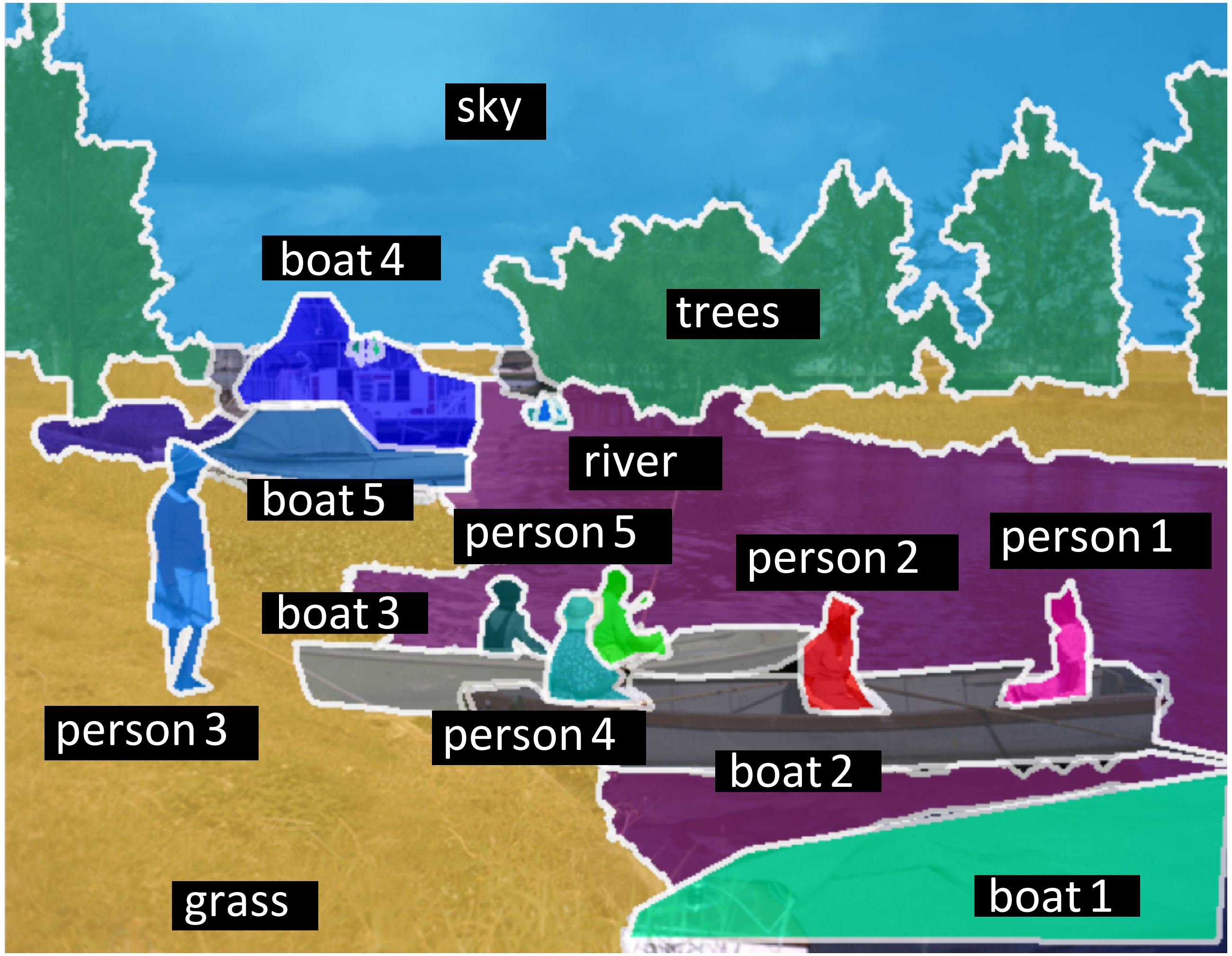


semantic segmentation

real-world application likely requires both: things + stuff

[Slide: A. Kirillov]

Panoptic segmentation



assign semantic labels to pixels
+ segment each instance
separately

[Slide: A. Kirillov]

Available panoptic segmentation datasets

COCO (2014) + COCO-stuff (2017)
COCO-panoptic challenges:
ECCV`18, **ICCV`19**

Mapillary Vistas (2017)
Vistas-panoptic challenges:
ECCV`18, **ICCV`19**

Cityscapes (2015)
panoptic test set
leaderboard (2019)

ADE20k (2016)
>22k images, 150 categorie

[Slide: A. Kirillov]

Panoptic quality (PQ) measure

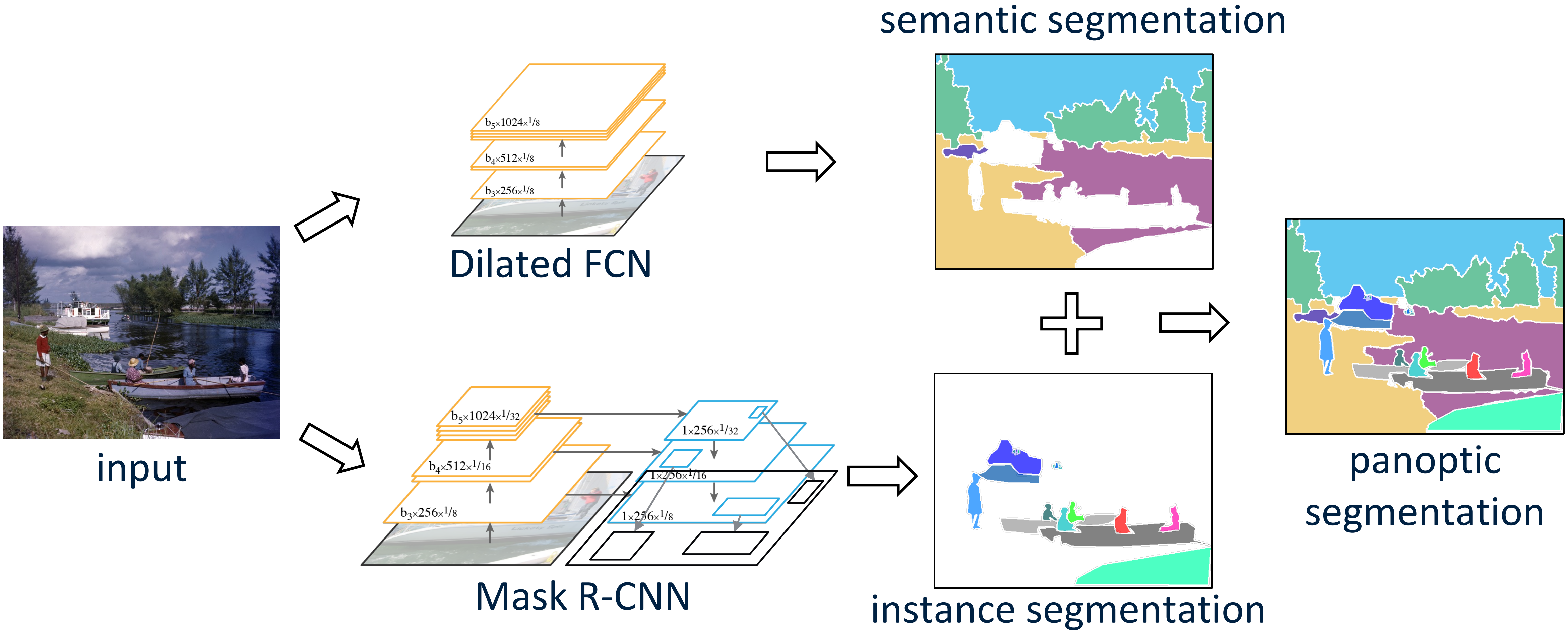
$$\mathrm{PQ} = \frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} = \underbrace{\frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP|}}_{\text{Segmentation Quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{Recognition Quality (RQ)}}$$

- symmetric
- unified for categories with and without instance-level annotation (**analysis**)

evaluation code: https://github.com/cocodataset/panopticapi

[Slide: A. Kirillov]

Panoptic segmentation: naïve approach

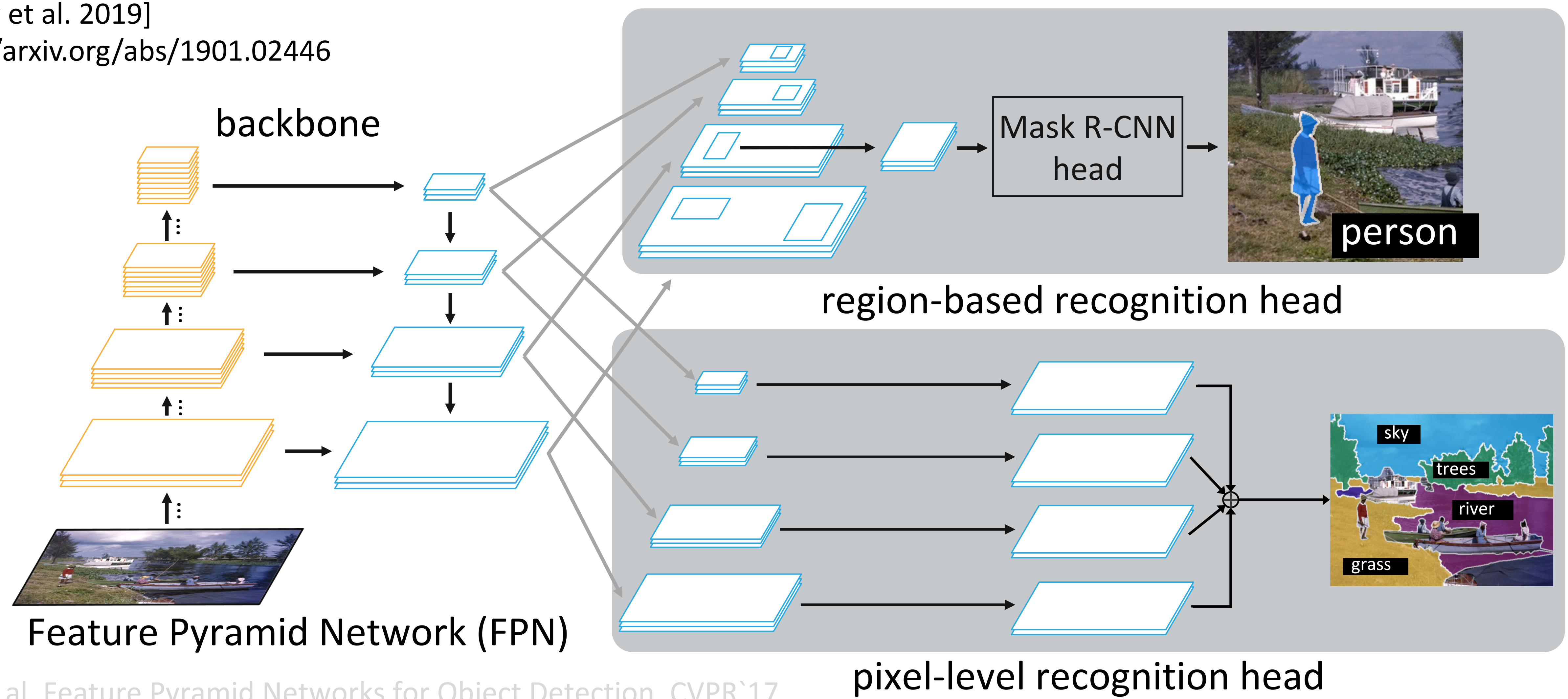


resolve overlaps between different
instances and stuff classes

[Slide: A. Kirillov]

Panoptic FPN: unified framework

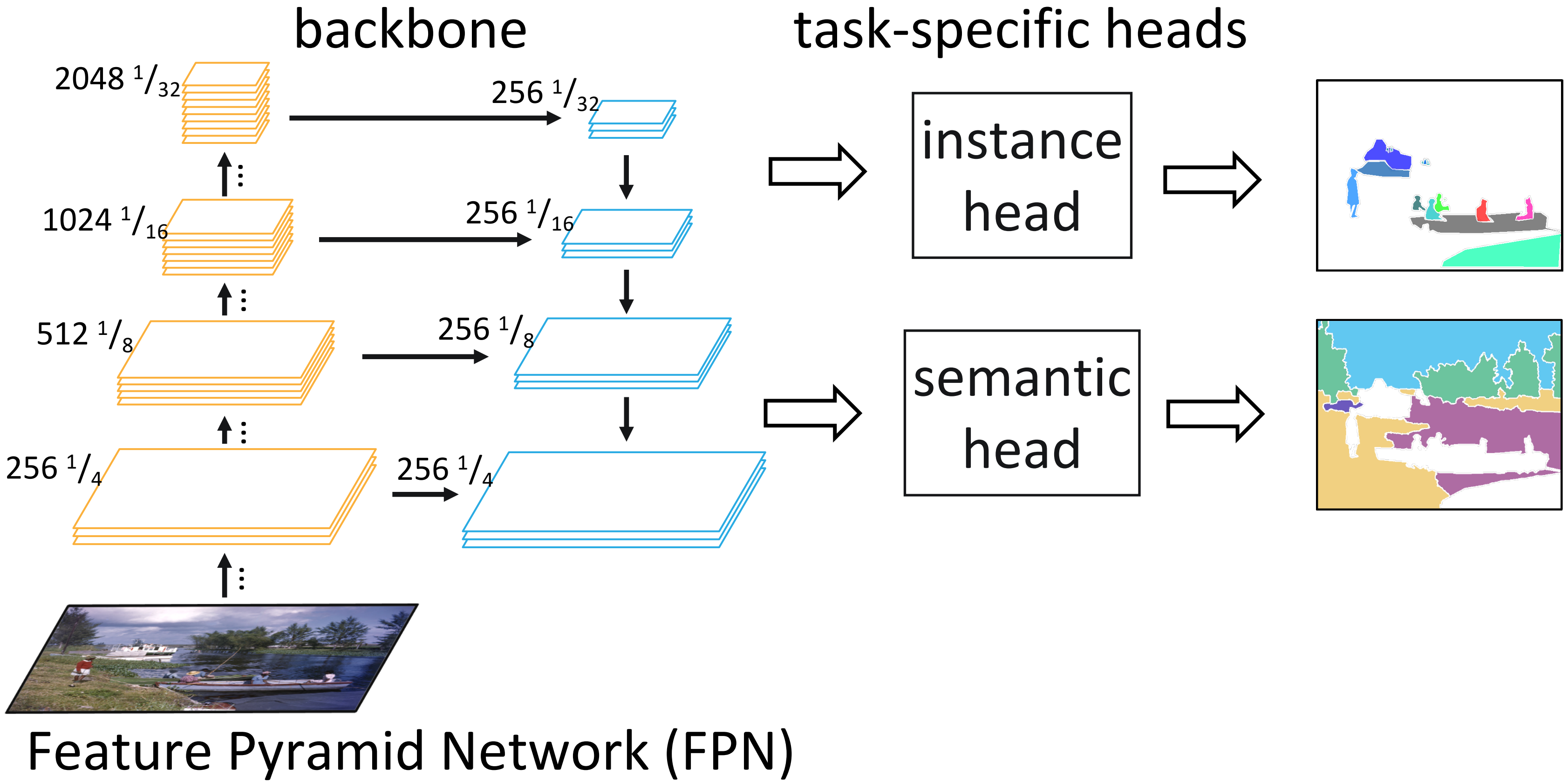
[Kirillov et al. 2019]
https://arxiv.org/abs/1901.02446



Lin et al. Feature Pyramid Networks for Object Detection, CVPR`17
He et al. Mask R-CNN, ICCV`17
Kirillov et al. Panoptic Feature Pyramid Networks, CVPR`19

[Slide: A. Kirillov]

Panoptic FPN



$$\mathcal{L}_{\text{bbox}} + \mathcal{L}_{\text{classification}} + \mathcal{L}_{\text{mask}} +$$

$$\mathcal{L}_{\text{semantic}}$$

[Slide: A. Kirillov]


LAKE

Beyond Object Classification with Convolutional Networks

David Eigen (NYU -> Clarifai)
Rob Fergus (Facebook / NYU)

Motivation

Input Image

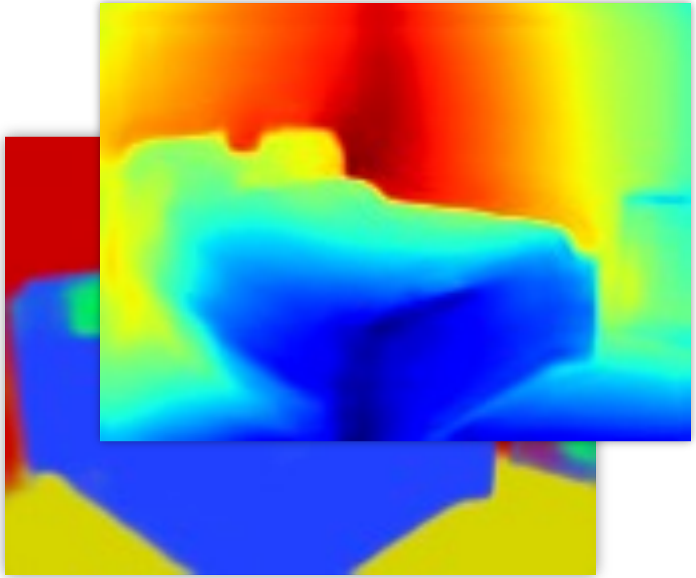
Semantic Map

- Understand input scene
  - Semantic
  - Geometric

Motivation

Input Image

Depth



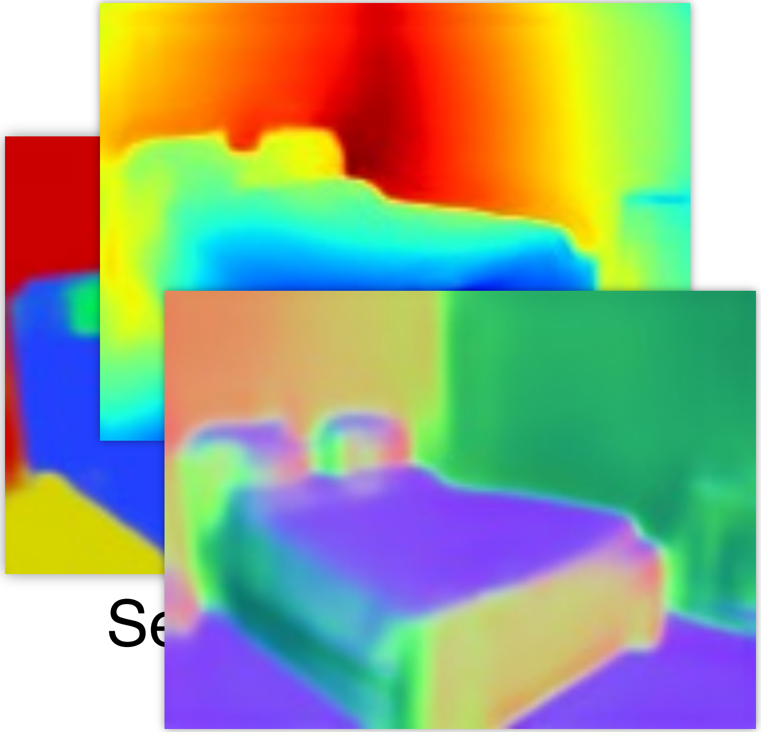
Semantic Map

- Understand input scene
  - Semantic
  - Geometric

Motivation

Input Image

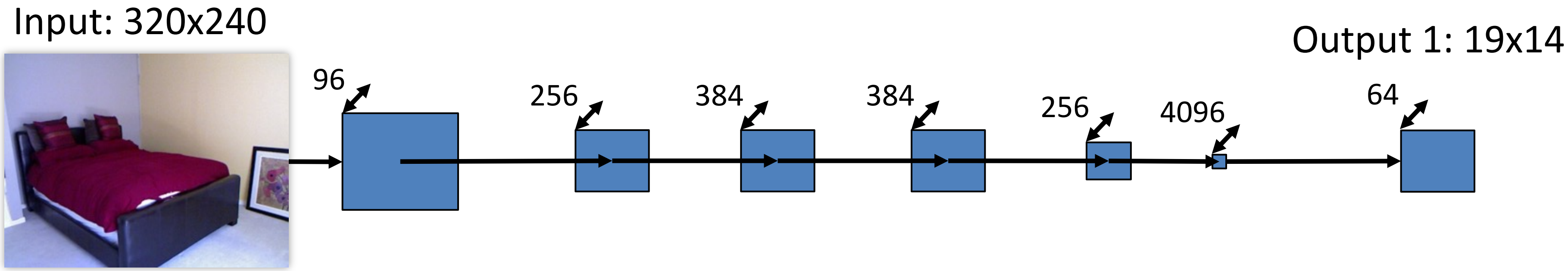
Depth



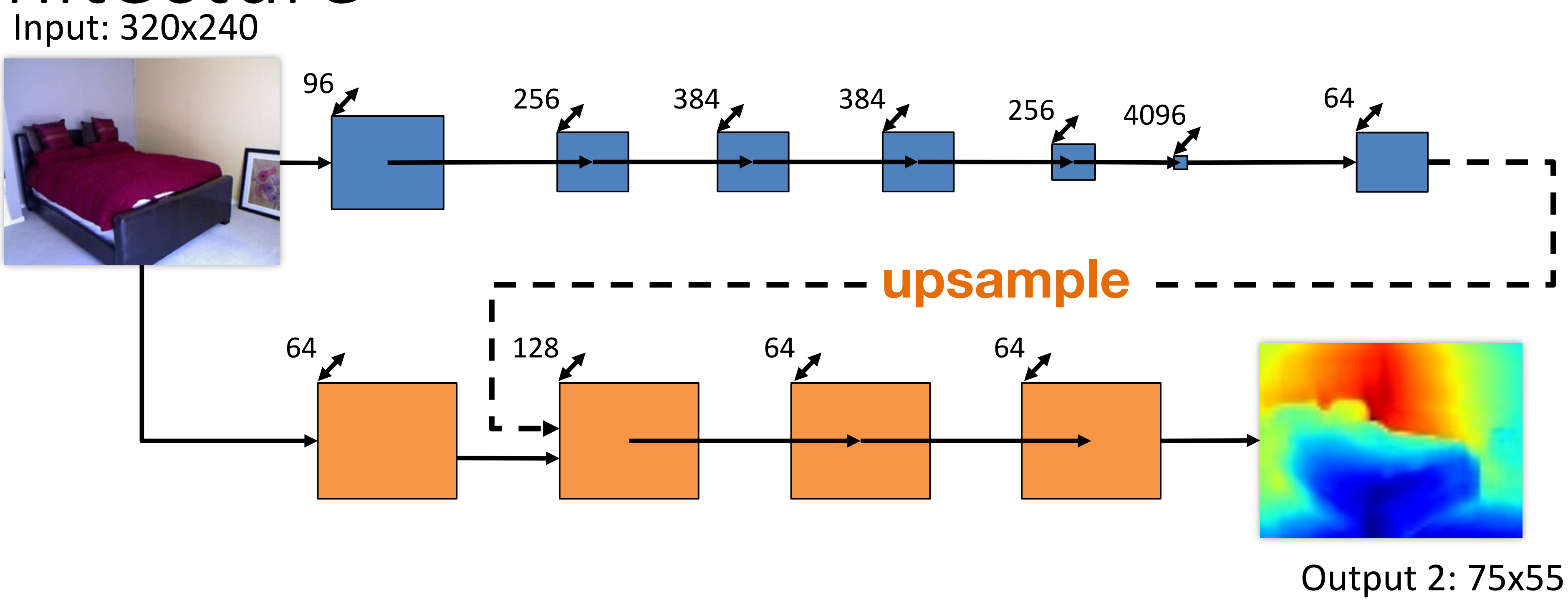
Normals

- Understand input scene
    - Semantic
    - Geometric

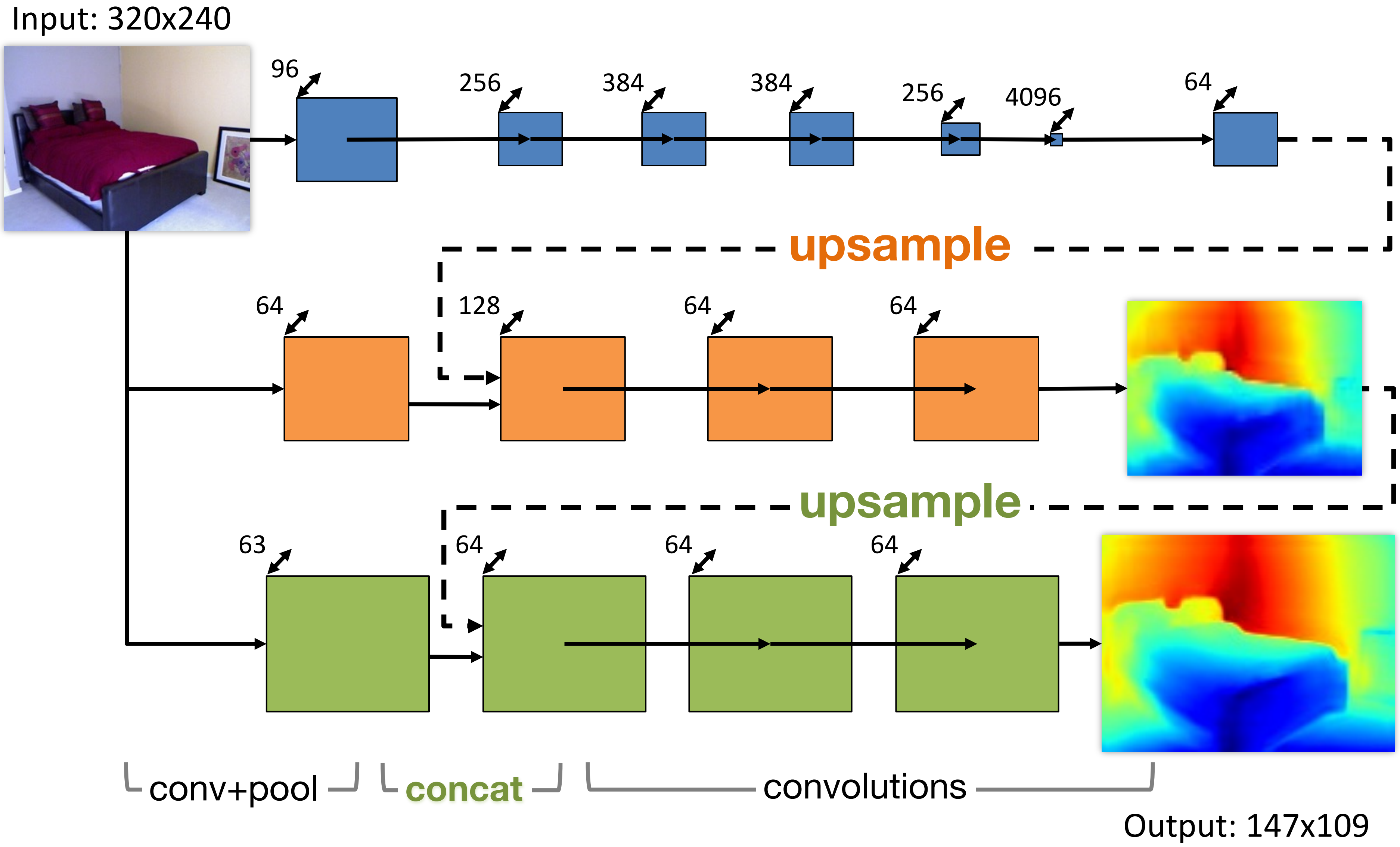
Architecture



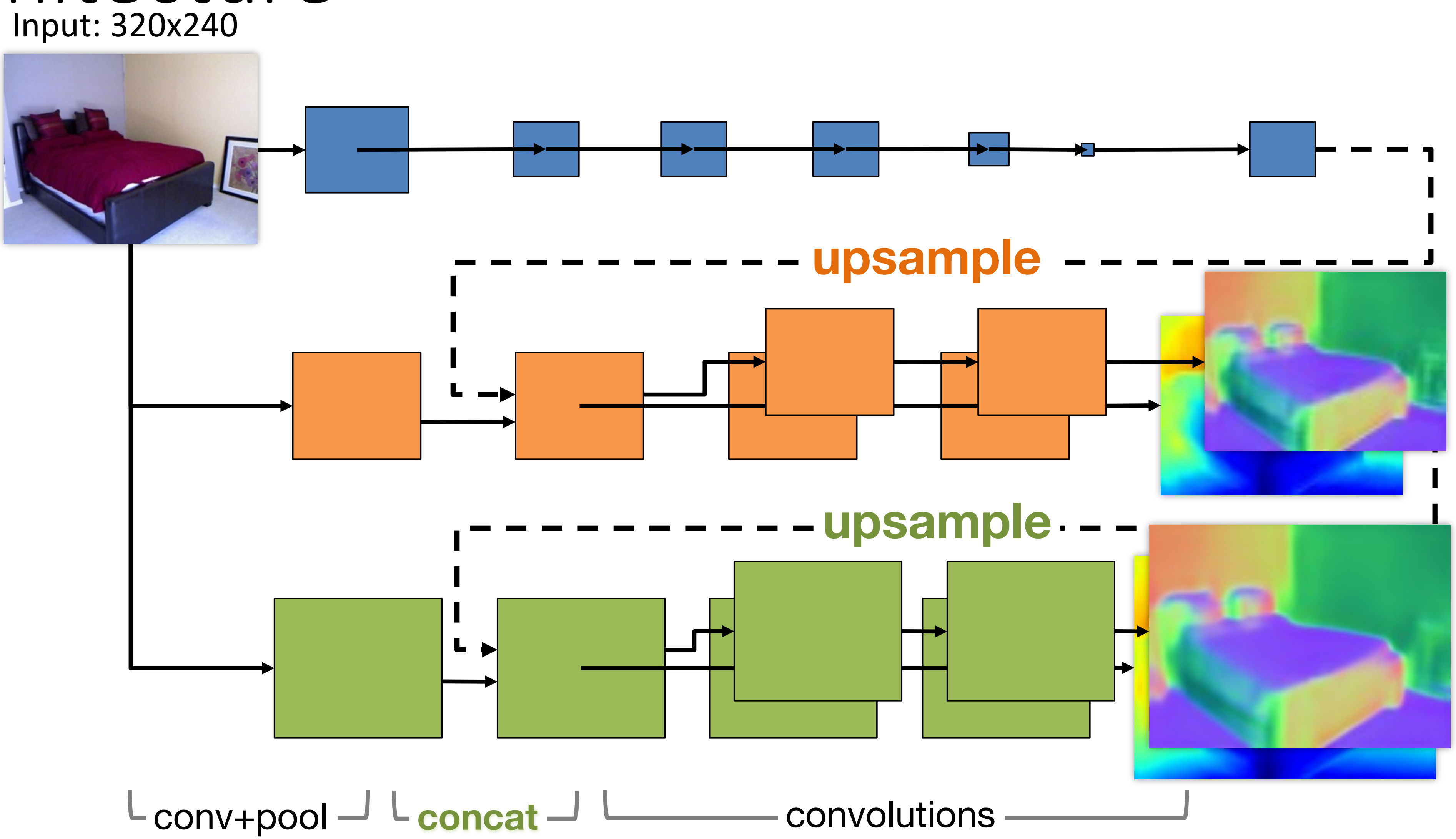
Architecture



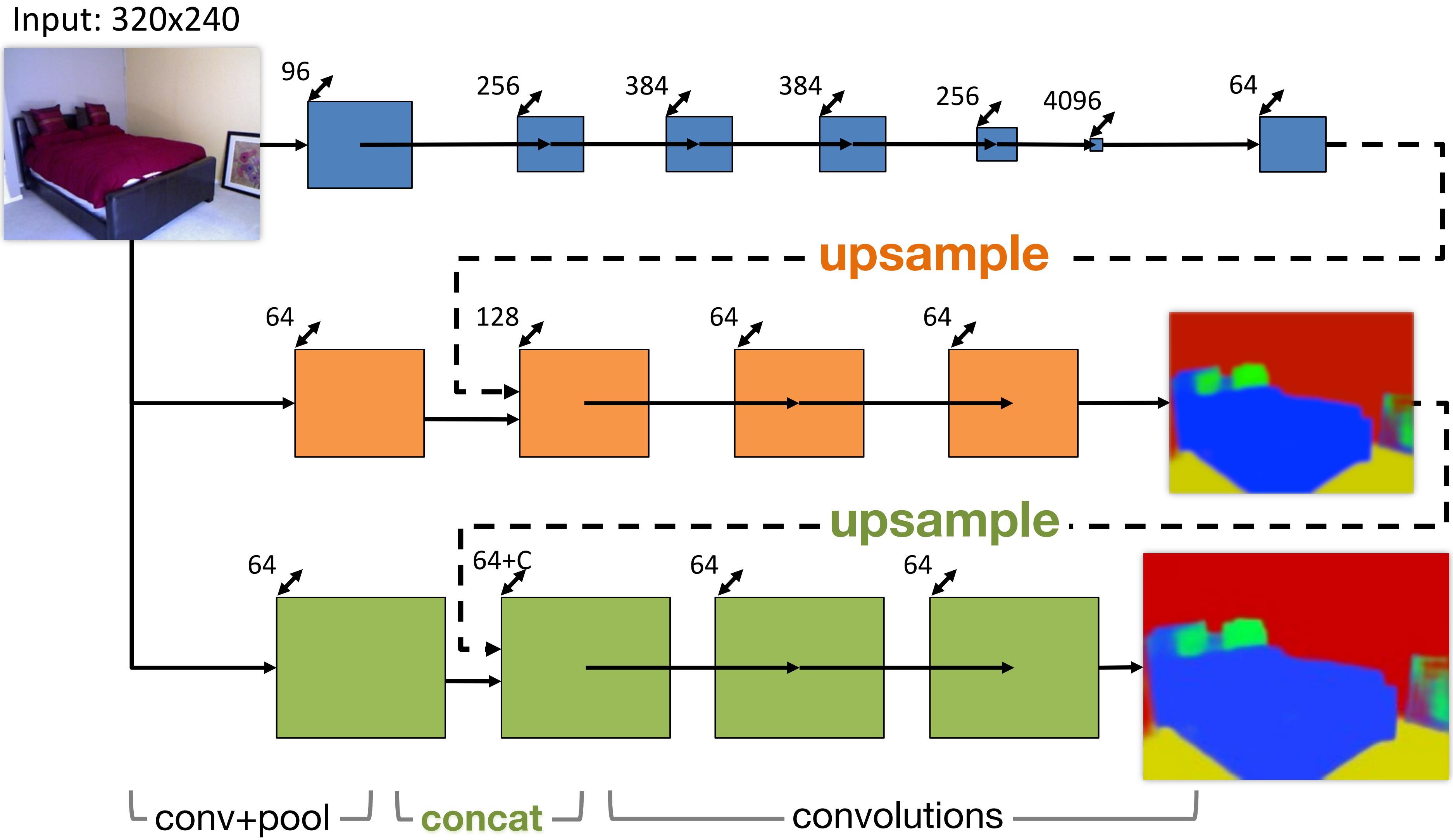
Architecture



Architecture



Architecture



Losses

Depth:

$$d = D - D^*$$

D = log predicted depth, D* = log true depth

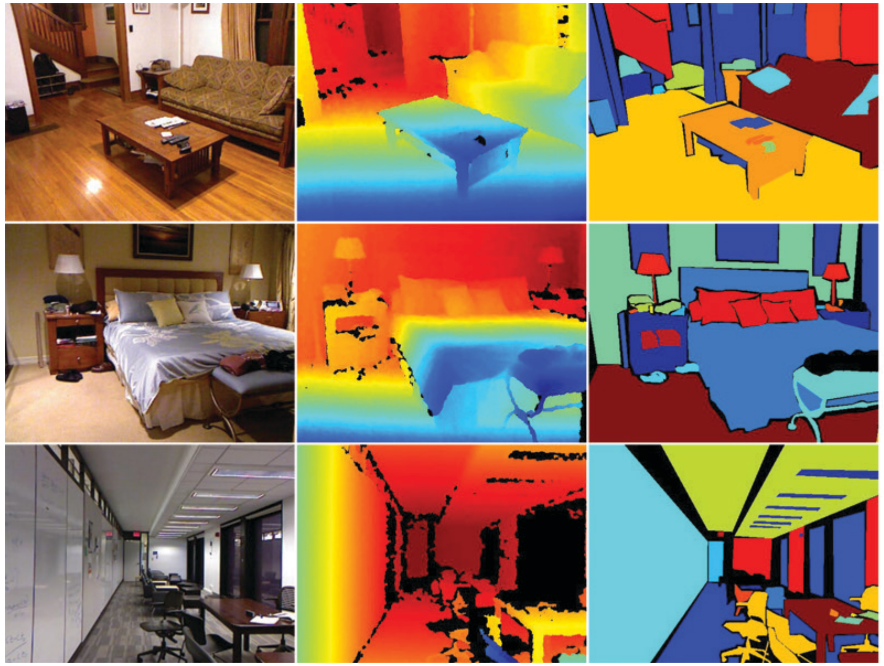
$$L_{depth}(D, D^*) = \frac{1}{n}\sum_i d_i^2 - \frac{1}{2n^2}\left(\sum_i d_i\right)^2 + \frac{1}{n}\sum_i [(\nabla_x d_i)^2 + (\nabla_y d_i)^2]$$

Norma˙

Labels

Evaluation

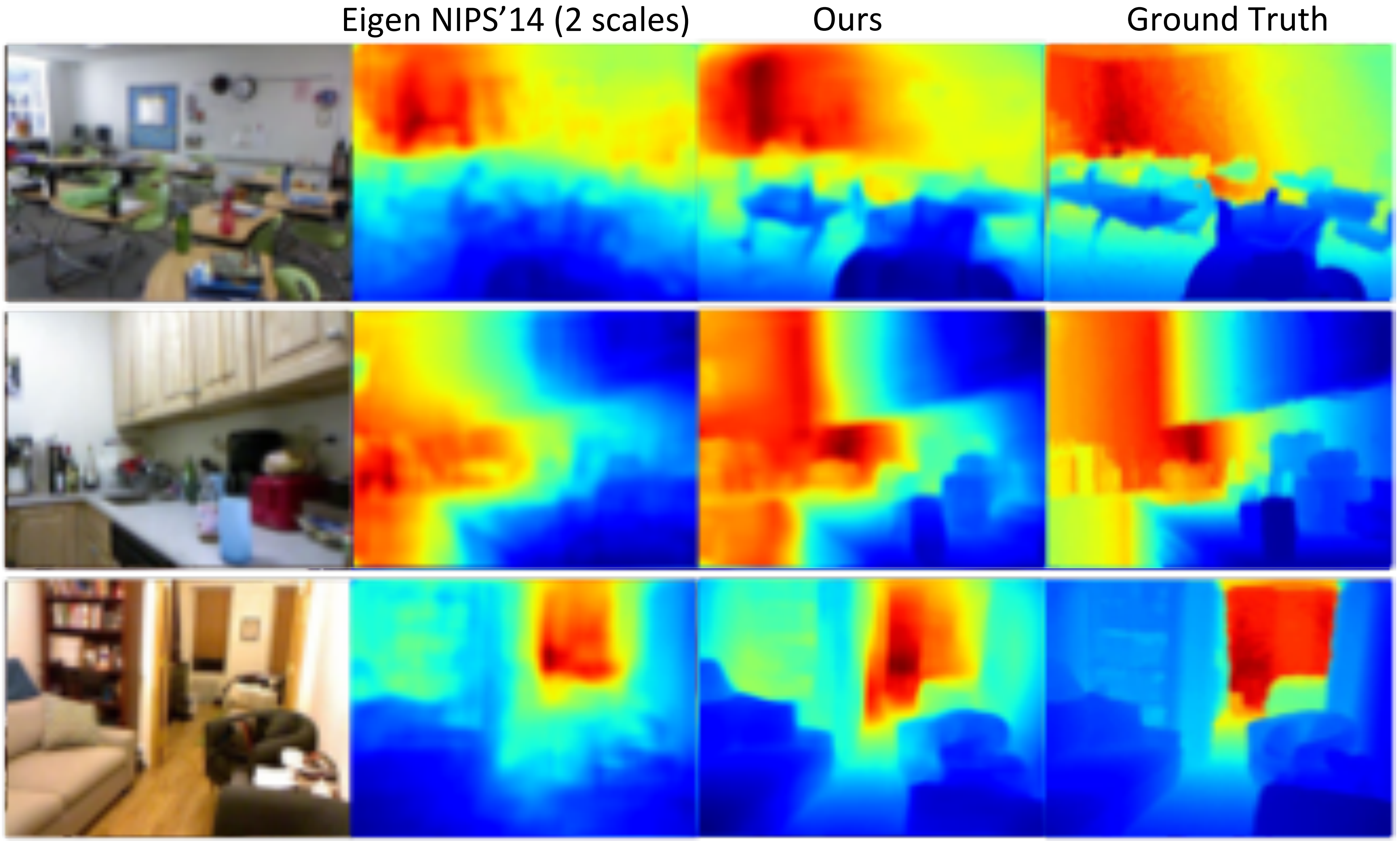
- NYU Depth dataset
  - RGB, Depth
    and per-pixel labels
  - Indoor scenes



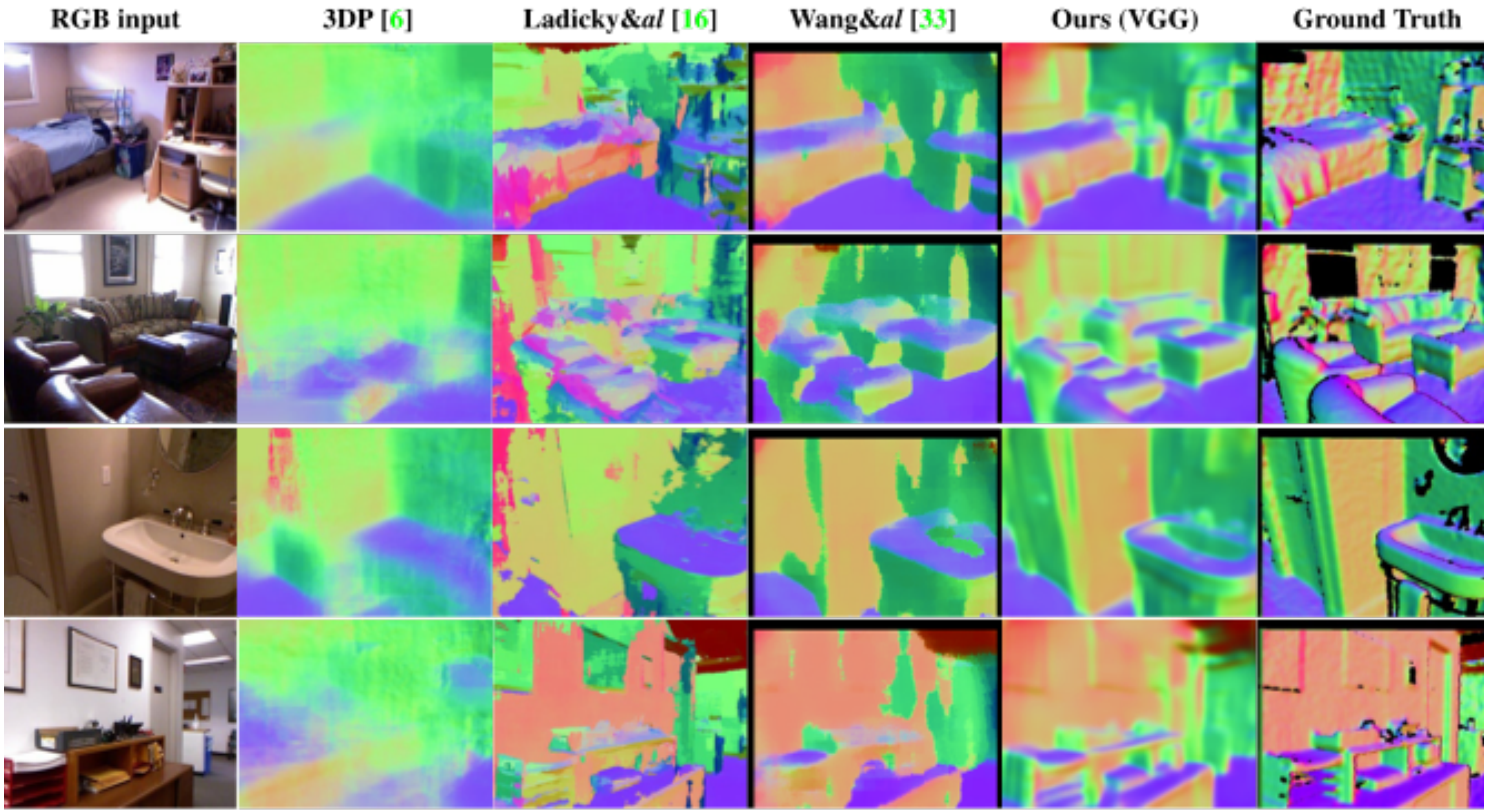
- Supervised training of models

- Compare to range of other methods
  - Also on SIFTFlow and PASCAL VOC’11

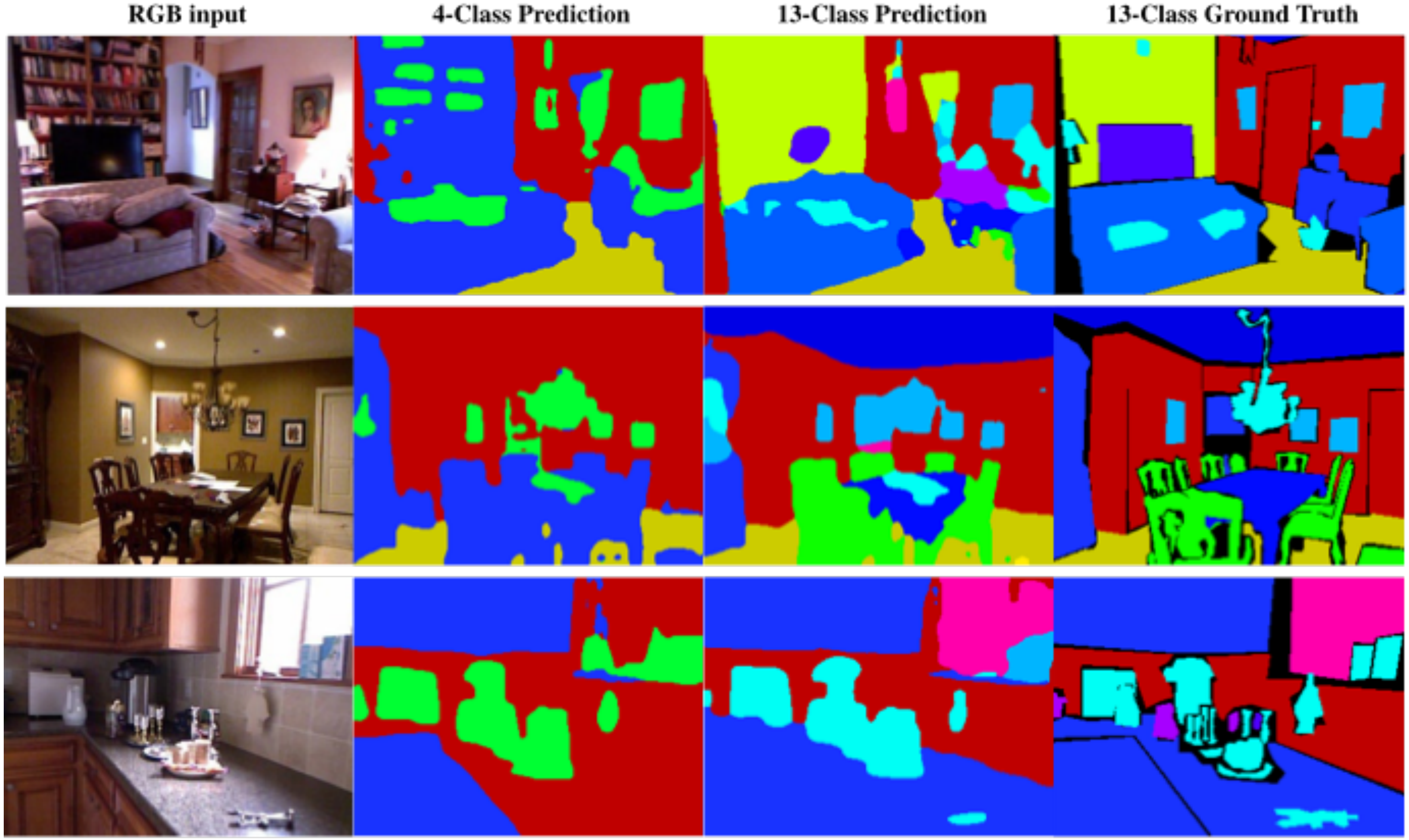
Depths Comparison



Surface Normals



Semantic Labels: NYUD



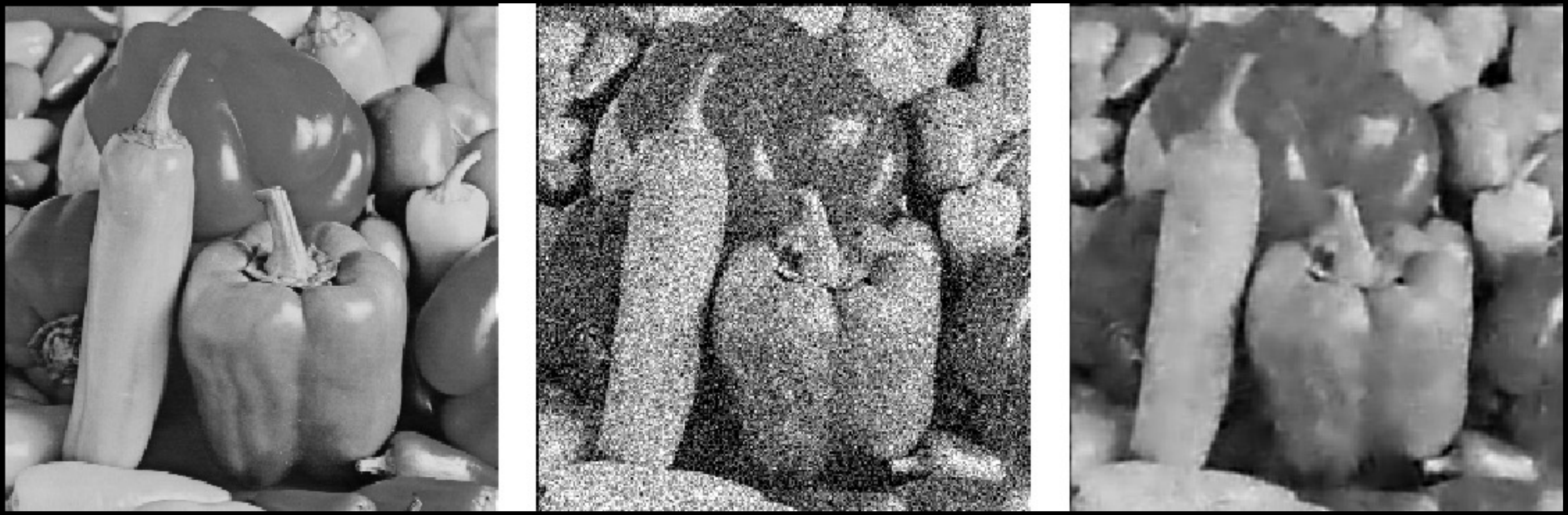
Overview

- Semantic Segmentation
  - Fully Convolutional Nets [Shelhamer et al. 2016] https://arxiv.org/abs/1605.06211
- Panoptic Segmentation
- Image processing with Convnets

Denoising with ConvNets

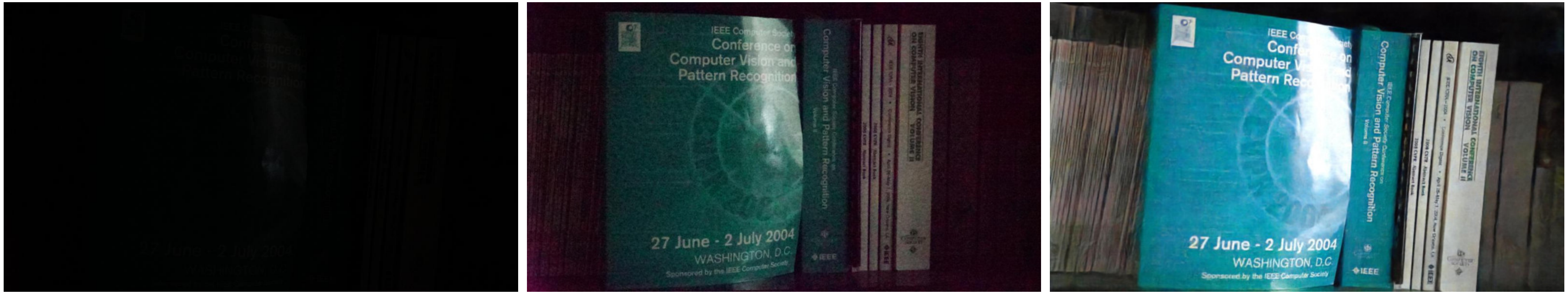
- Burger et al. “Can plain NNs compete with BM3D?” CVPR 2012
- Deep Learning for Image Denoising: a survey, Tian et al. https://arxiv.org/abs/1810.05052, 2018

Original Noised Denoised

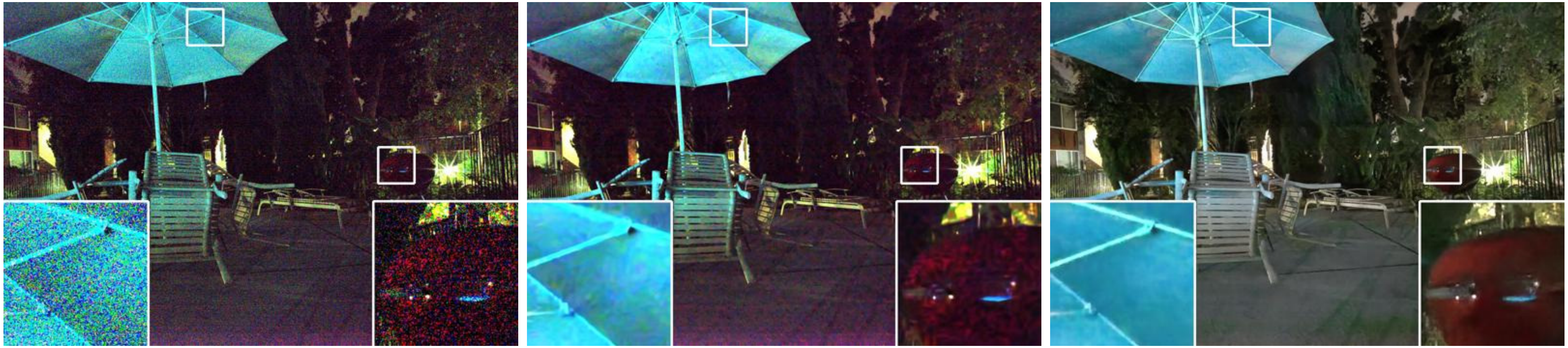


Learning to See in the Dark

[Chen et al., arXiv 1805.01934]



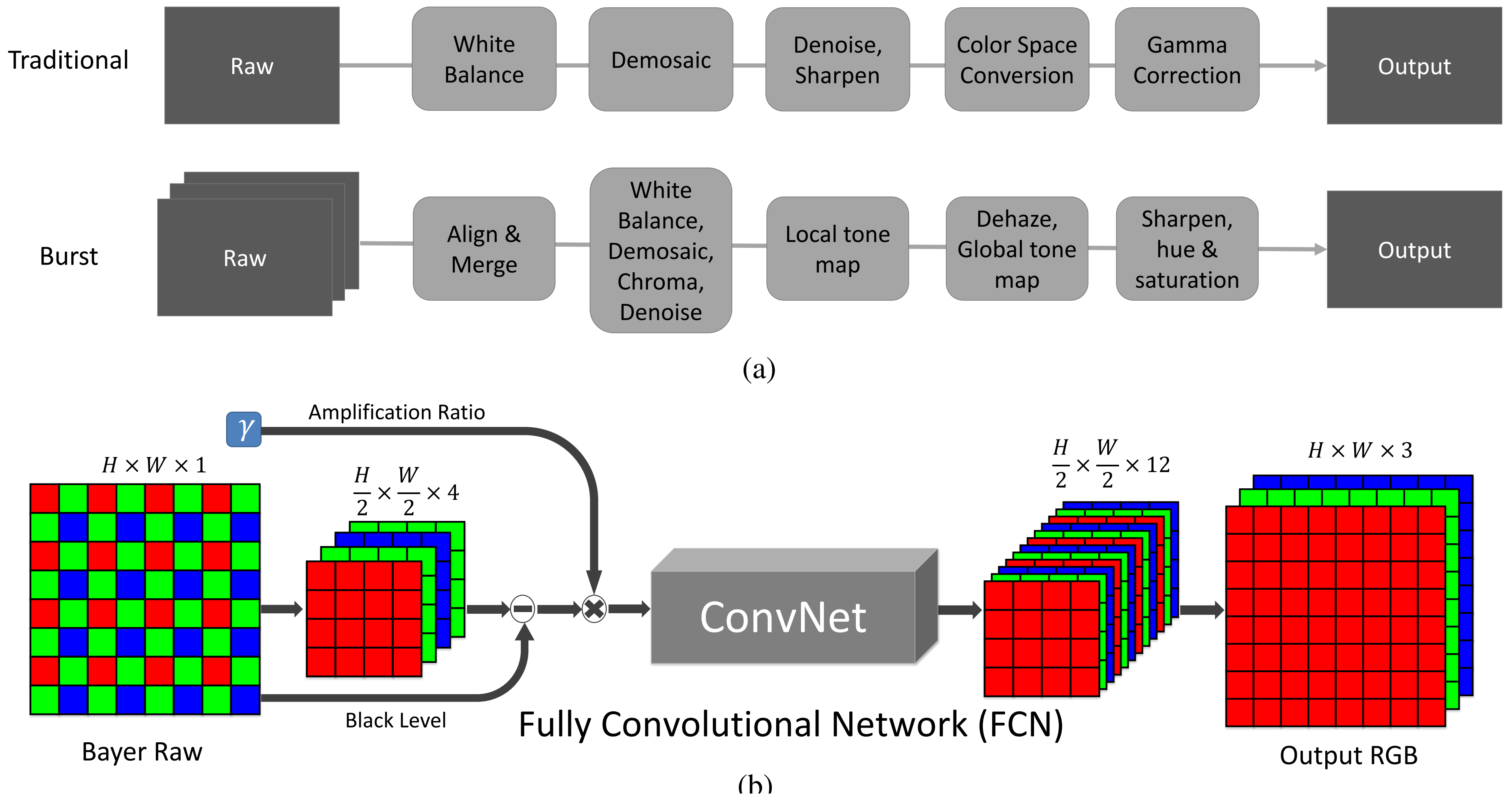
(a) Camera output with ISO 8,000 (b) Camera output with ISO 409,600 (c) Our result from the raw data of (a)



(a) Traditional pipeline (b) ... followed by BM3D denoising (c) Our result

Learning to See in the Dark

[Chen et al., arXiv 1805.01934]



Deblurring with Convnets

- Deep Image Deblurring: A Survey, https://arxiv.org/pdf/2201.10700.pdf, 2022.

(a) Camera shake blur (b) Out-of-focus blur

(c) Moving object blur (d) Mixed blur

Deblurring with Convnets

- Blind deconvolution
  - Learning to Deblur, Schuler et al., arXiv 1406.7444, 2014

Blurry image with ground truth kernel | Result of [Zho+13] PSNR 23.17 | Deblurring result w. noise *agnostic* training PSNR 23.29 | Deblurring result w. noise *specific* training **PSNR 23.41**

Architectures for Deblurring

https://arxiv.org/pdf/2201.10700.pdf

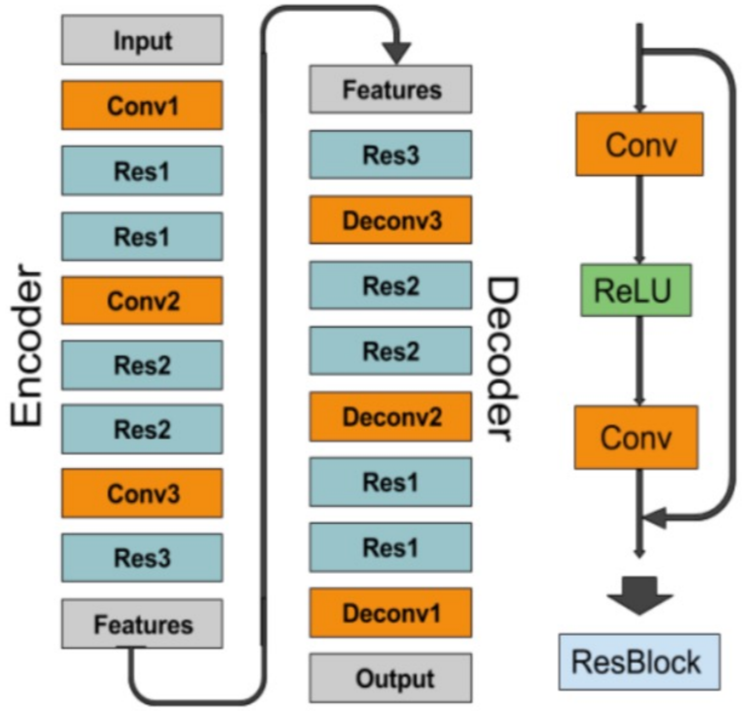


Fig. 3 Deep single image deblurring network based on the Deep Auto-Encoder (DAE) architecture [91].

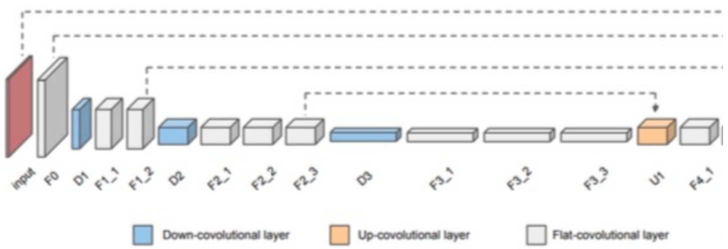


Fig. 4 Deep video deblurring network ba
architecture [122].

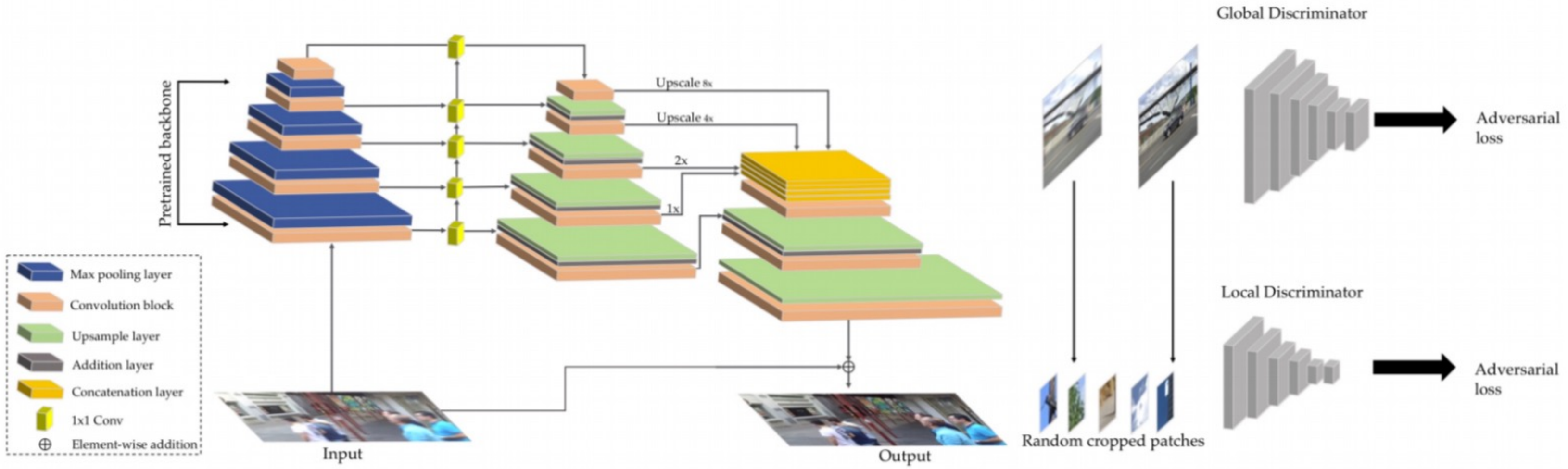


Fig. 7 Deep single image deblurring network based on the GAN architecture [60].

Fig. 12 Evaluation results of the state-of-the-art deblurring methods on the GoPro dataset [86]. From left to right: blurry images, results of Nah *et al.*[86], Tao *et al.*[131], DBGAN [154] and DeblurGAN-v2 [60]. [86] and [131] are two multi-scale based image deblurring networks. [154] and [60] are two GAN based image deblurring networks.

https://arxiv.org/pdf/2201.10700.pdf

Inpainting with Convnets

- Image Denoising and Inpainting with Deep Neural Networks, Xie et al. NIPS 2012.
- Mask-specific inpainting with deep neural networks, Köhler et al., Pattern Recognition 2014

Original Schmid CVPR'10 Köhler et al. '14

Removing Local Corruption

- Restoring An Image Taken Through a Window Covered with Dirt or Rain, Eigen et al., ICCV 2013.

Removing Local Corruption

Restoring An Image Taken Through a Window Covered with Dirt or Rain

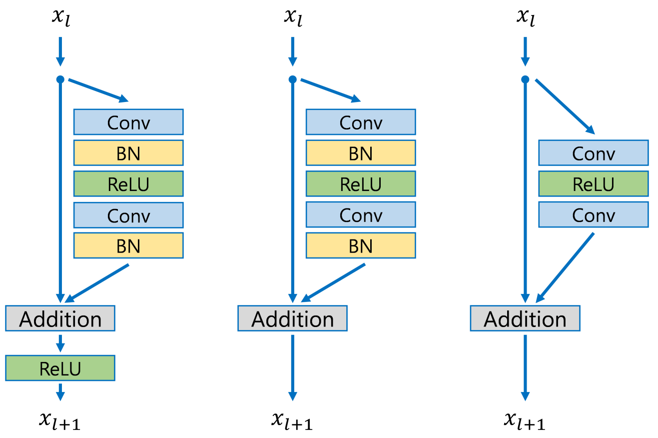
Rain Sequence

Each frame processed independently

David Eigen, Dilip Krishnan and Rob Fergus

ICCV 2013

Enhanced Deep Residual Networks for Single Image Super-Resolution, Bee Lim Sanghyun Son Heewon Kim Seungjun Nah Kyoung Mu Le, CVPR 2017 workshop



(a) Original (b) SRResNet (c) Proposed

Figure 2: Comparison of residual blocks in original ResNet, SRResNet, and ours.

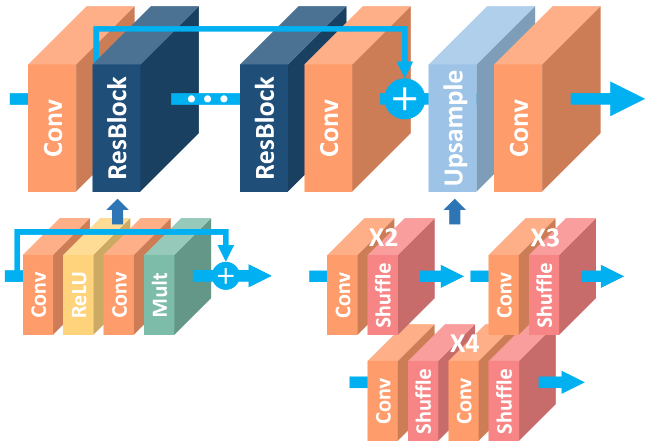
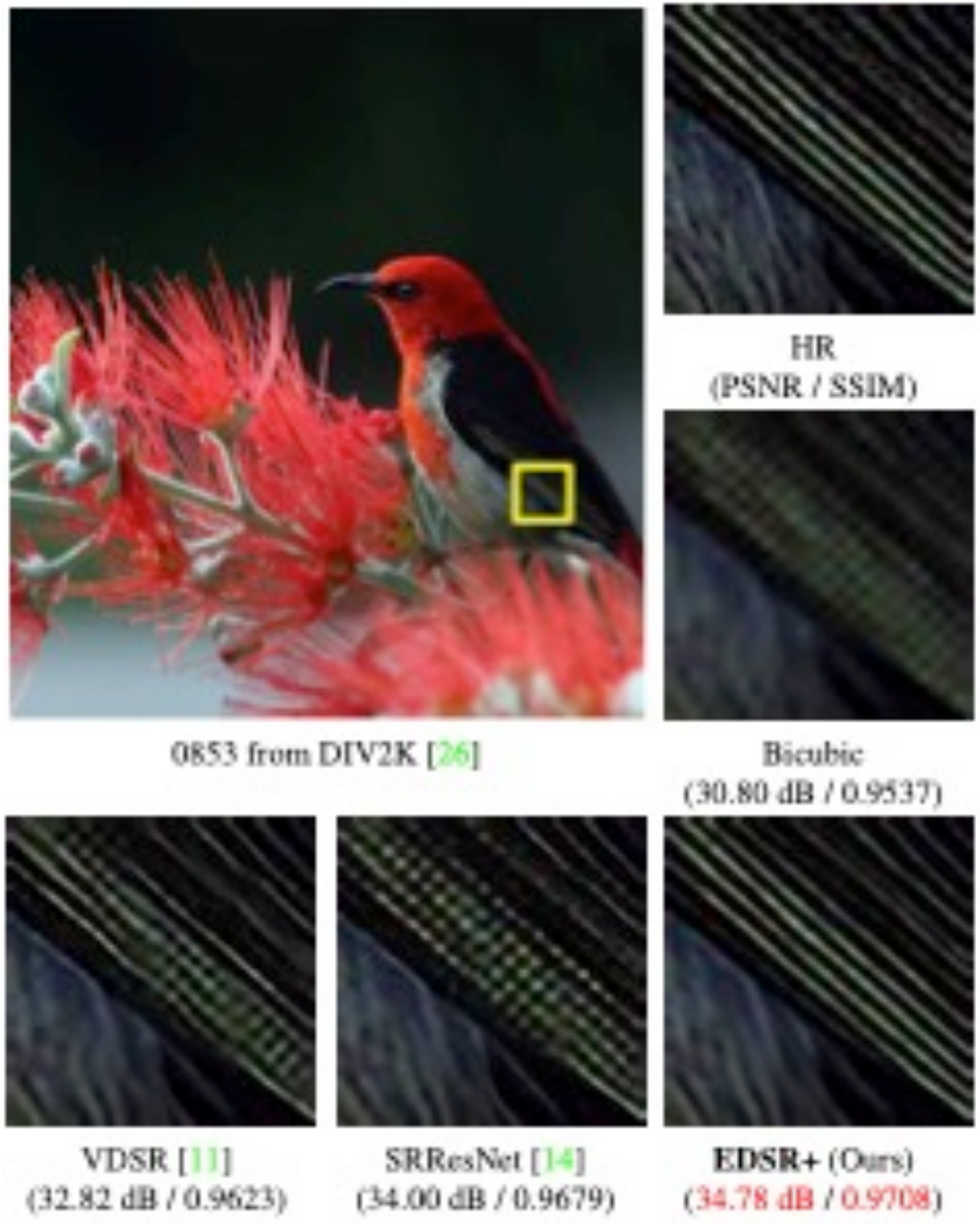


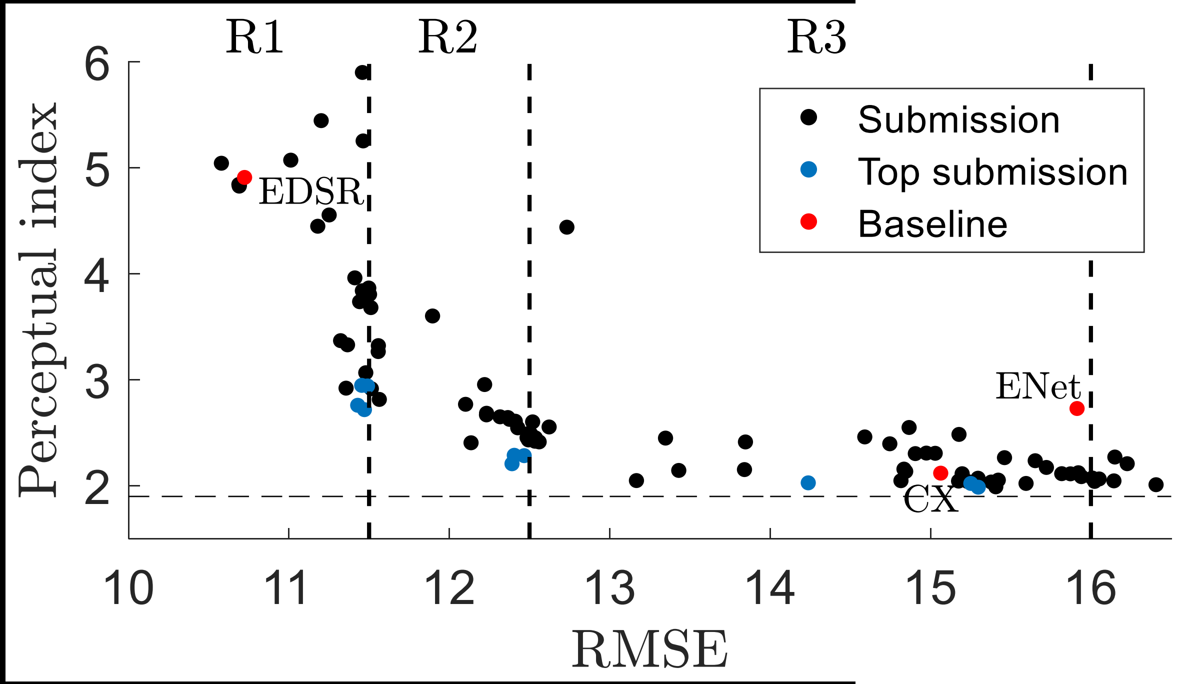
Figure 3: The architecture of the proposed single-scale SR network (EDSR).

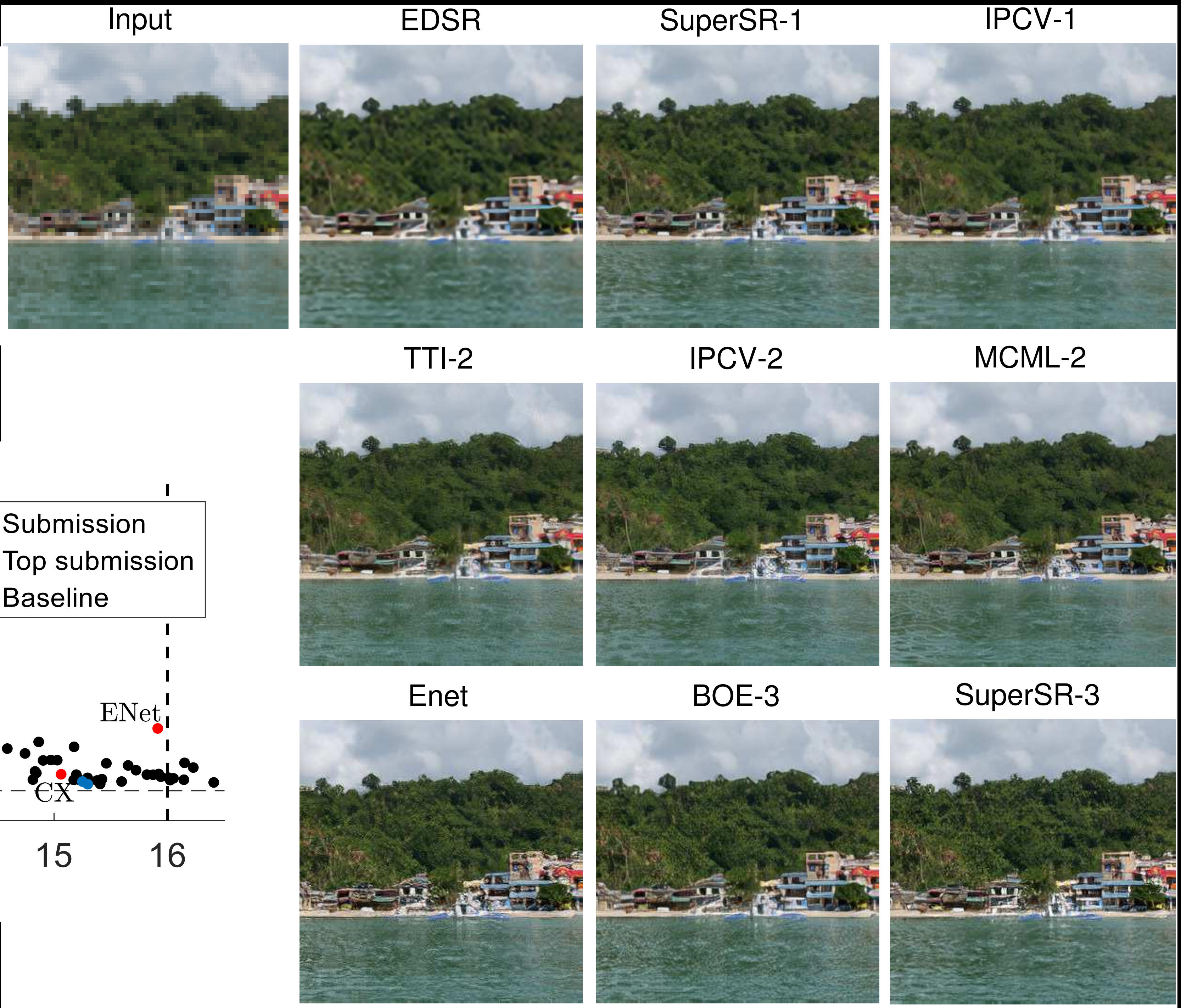


The 2018 PIRM Challenge on Perceptual Image Super-resolution

Yochai Blau[1*], Roey Mechrez[1*], Radu Timofte[2], Tomer Michaeli[1], and Lihi Zelnik-Manor[1]

[1] Technion–Israel Institute of Technology, Haifa, Israel
[2] ETH Zurich, Switzerland
{yochai,roey}@campus.technion.ac.il





Project Abstracts

Project Abstracts are due next week (Thursday 13th @ midnight)

Each project group should email me (fergus@cs.nyu.edu) with an abstract paragraph (100-150 words max; text format) giving:

- 1. Couple of sentences (2-3) describing your intended project.
- 2. Which datasets you will use.
- 3. Any directly related existing papers that you might build off of.

I will review them and let you know if I can concerns about the feasibility of the project, given time & compute constraints.