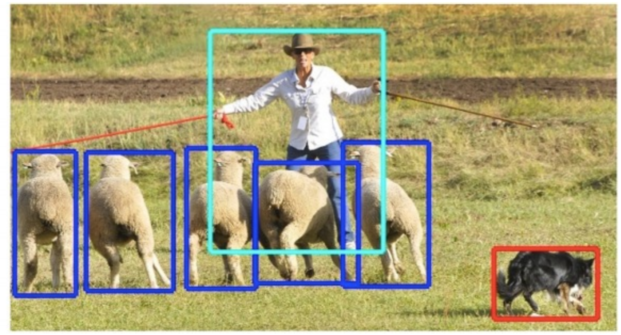
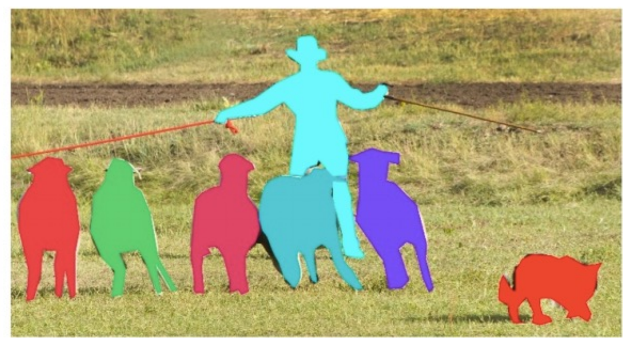
CNNs for dense image labeling

image classification



object detection

semantic segmentation



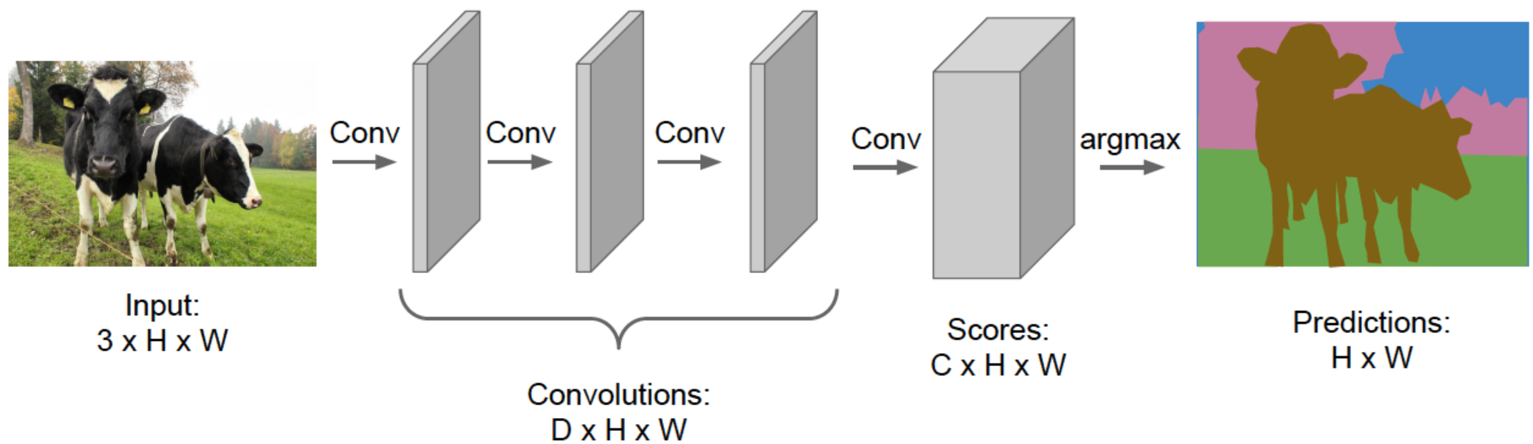
instance segmentation

Outline

- Operations and architectures for dense prediction: U-Net
- Instance segmentation: Mask R-CNN
- Other dense prediction problems

Dense prediction architectures

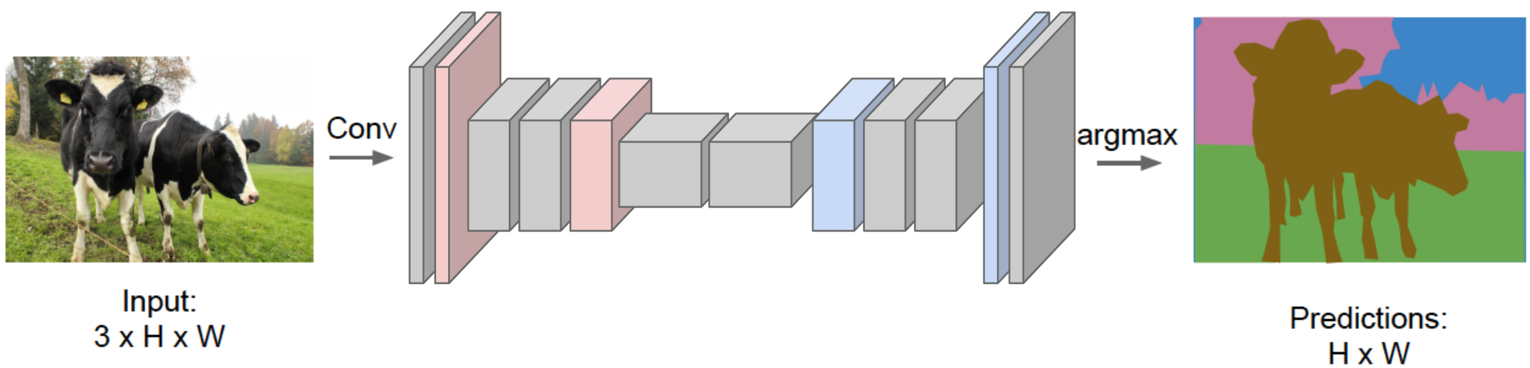
- To make predictions for all pixels at once, we can design a network with only stride-1 convolutions and element-wise operations
- What are the pros and cons of this approach?



Source: [Stanford CS231n]

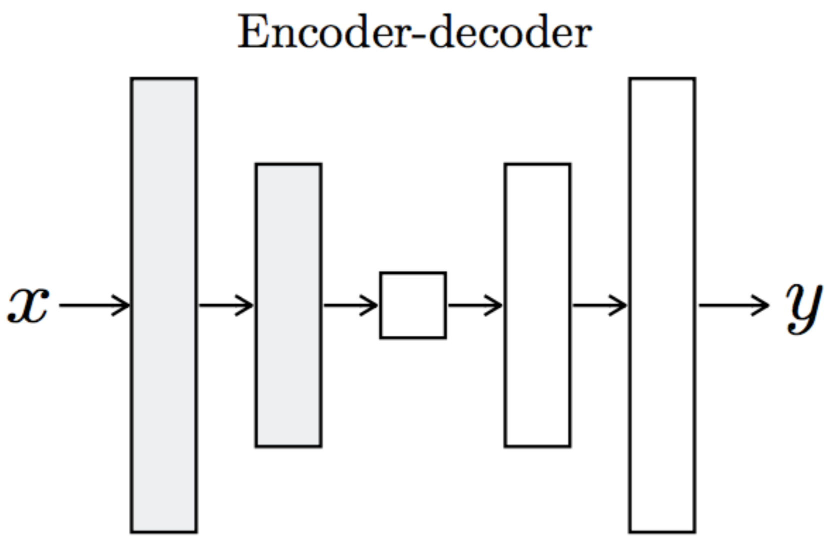
Dense prediction architectures

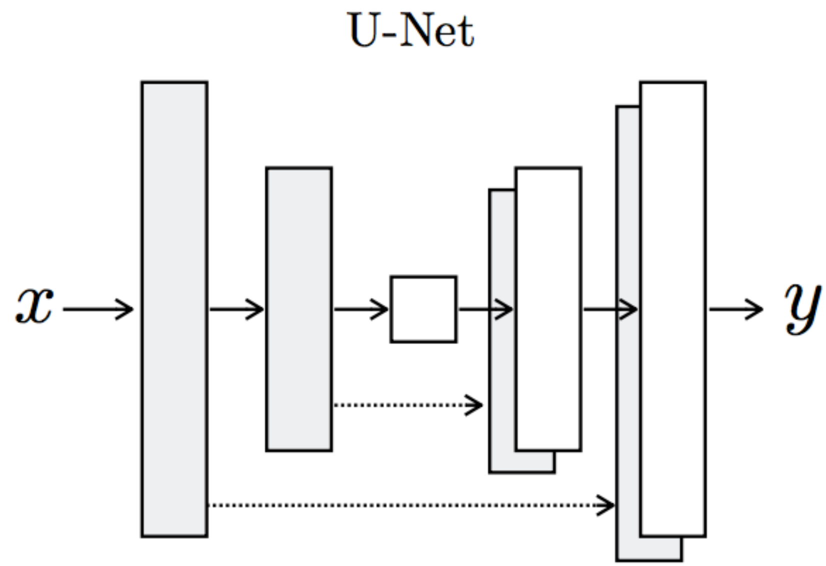
- Practical solution: first downsample, then upsample



Source: [Stanford CS231n]

Dense prediction architectures





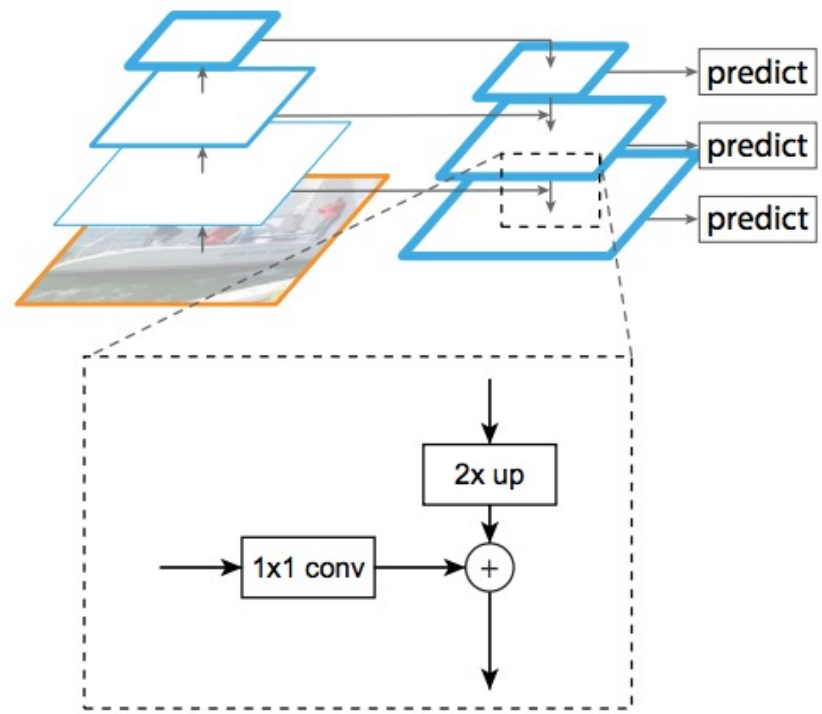
Problem: no way to recover information lost to downsampling

Fuse encoder and decoder feature maps at the same resolution

Figure source

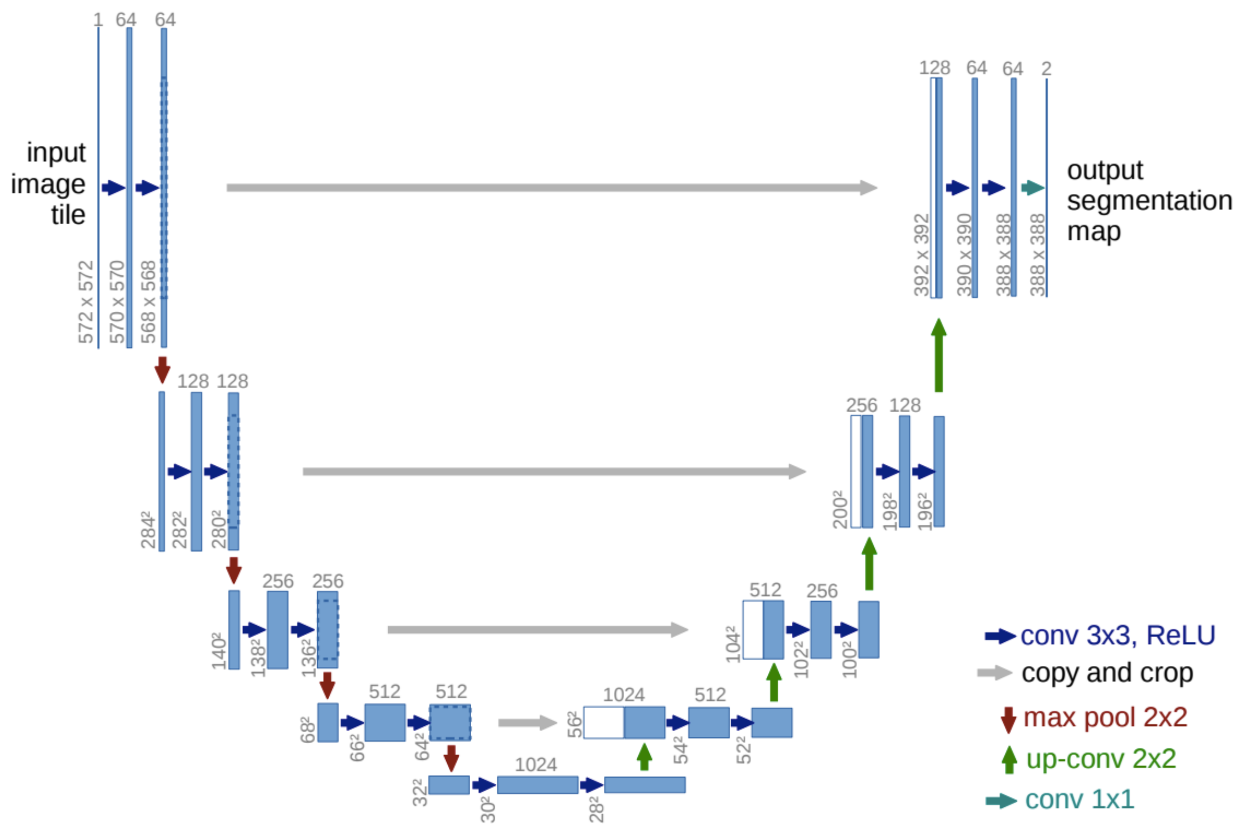
Recall: Feature pyramid networks

- Improve predictive power of lower-level feature maps by adding contextual information from higher-level feature maps
- Predict different sizes of bounding boxes from different levels of the pyramid (but share parameters of predictors)



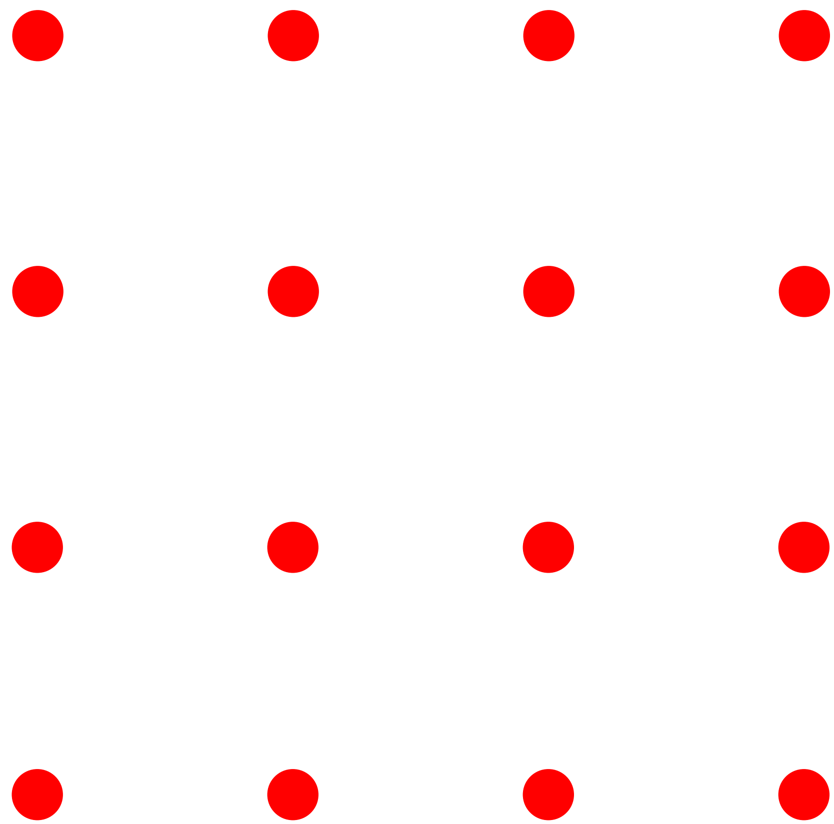
T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, [Feature pyramid networks for object detection](), CVPR 2017

U-Net



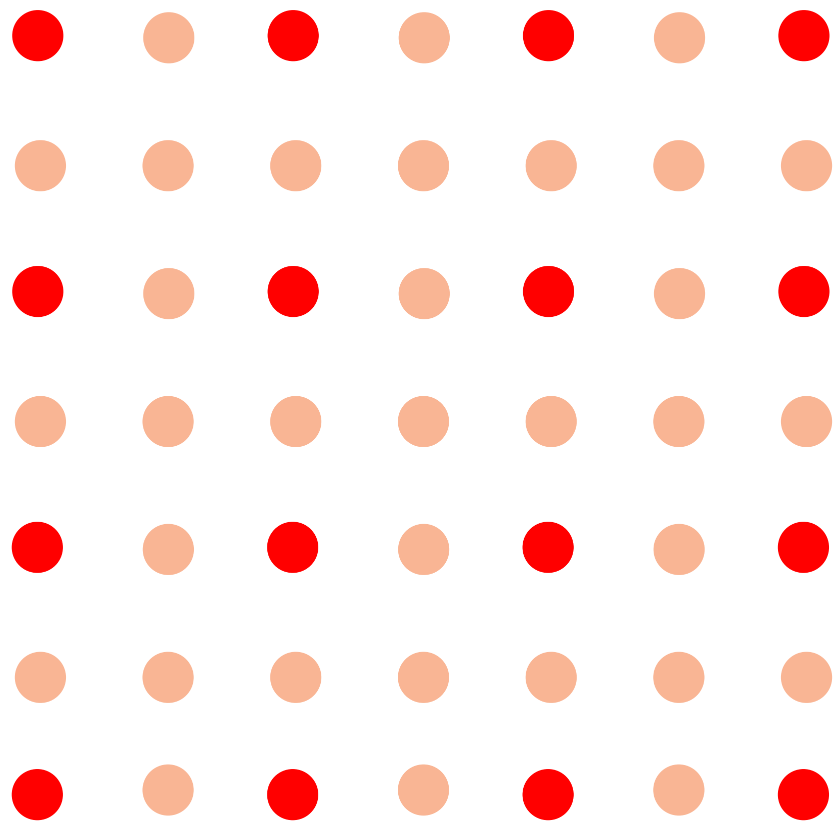
O. Ronneberger, P. Fischer, T. Brox, [U-Net: Convolutional Networks for Biomedical Image Segmentation](), MICCAI 2015

Feature map upsampling



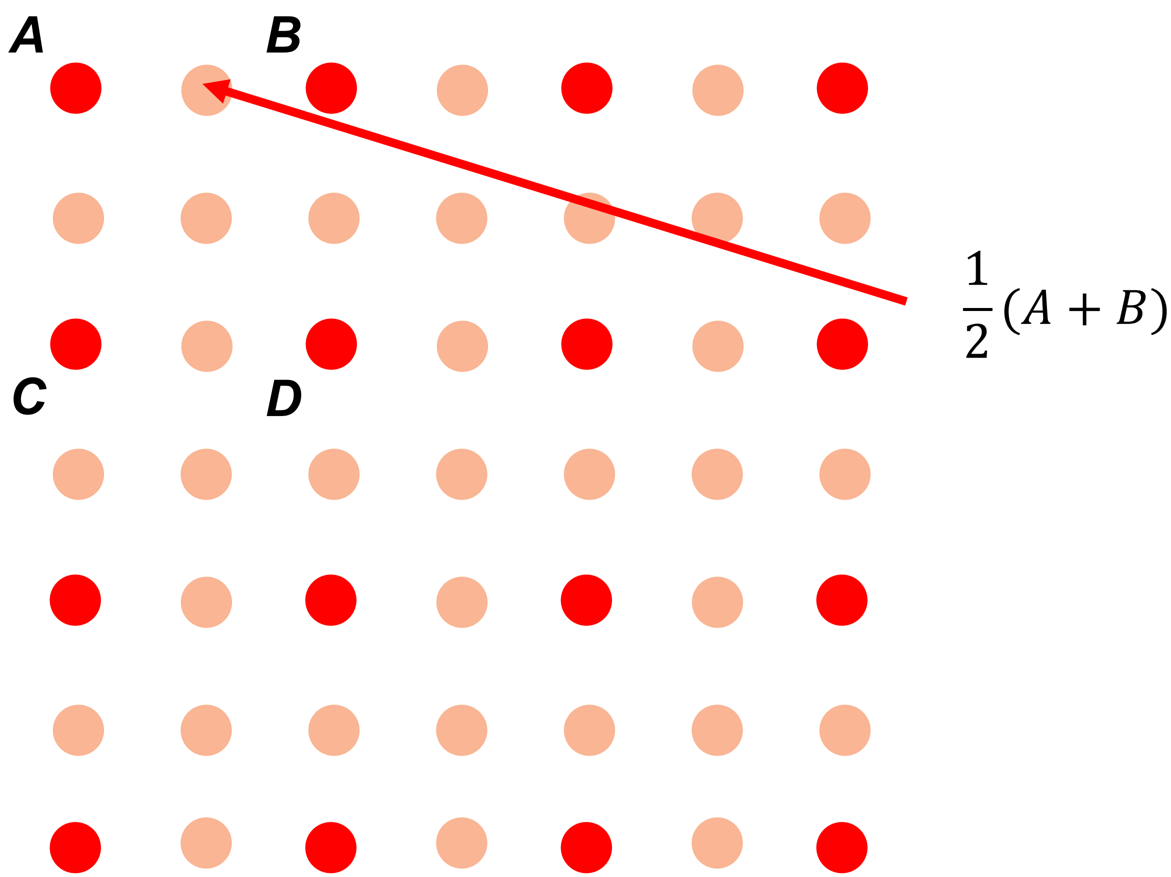
Step 1: increase the resolution of the feature grid

Feature map upsampling

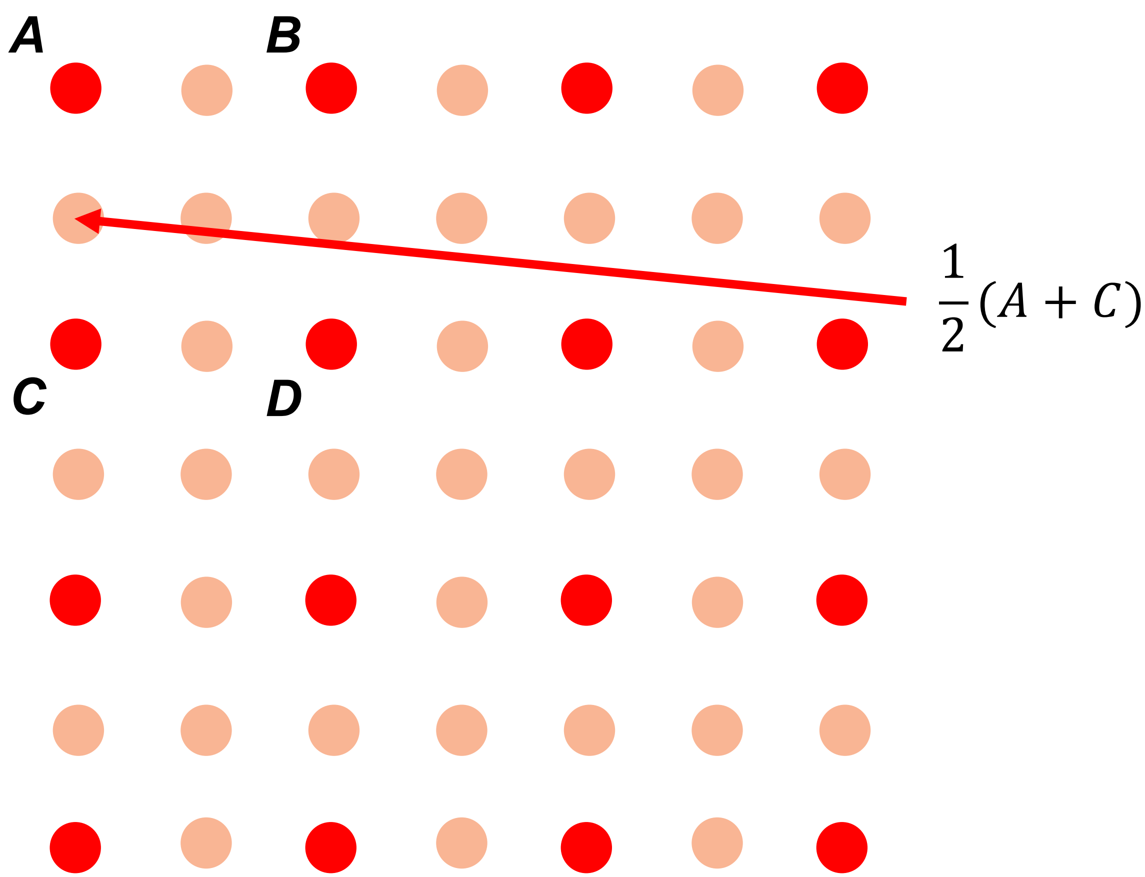


Step 2: *interpolate* to get the missing values

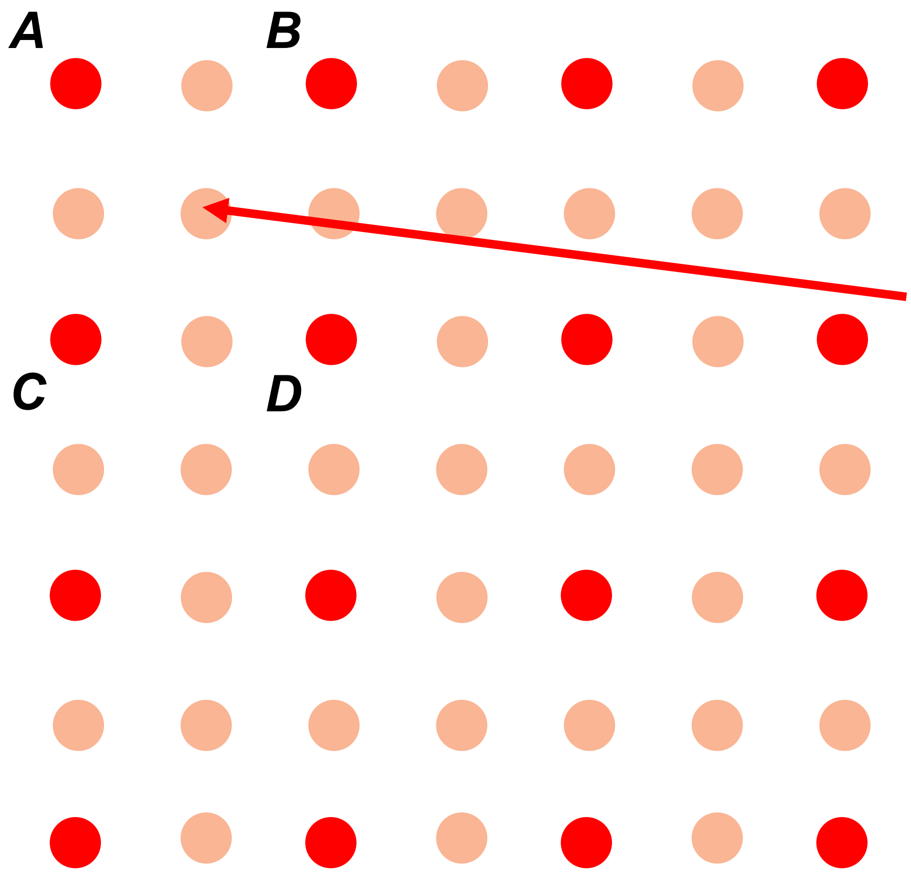
Bilinear interpolation



Bilinear interpolation



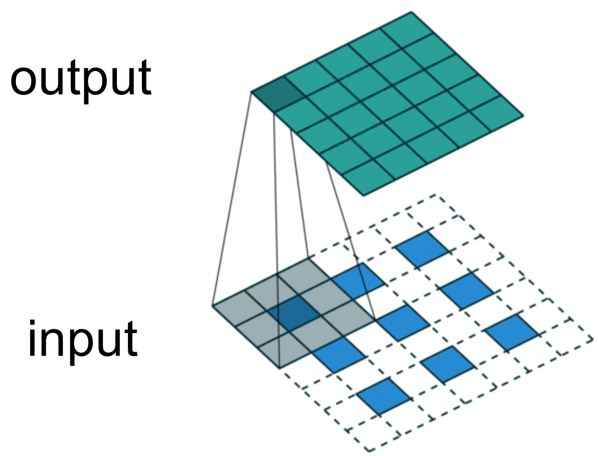
Bilinear interpolation



$$\frac{1}{4}(A + B + C + D)$$

Feature map upsampling

- For 2x upsampling, dilate the input by inserting rows and columns of zeros between adjacent entries, convolve with upsampling filter



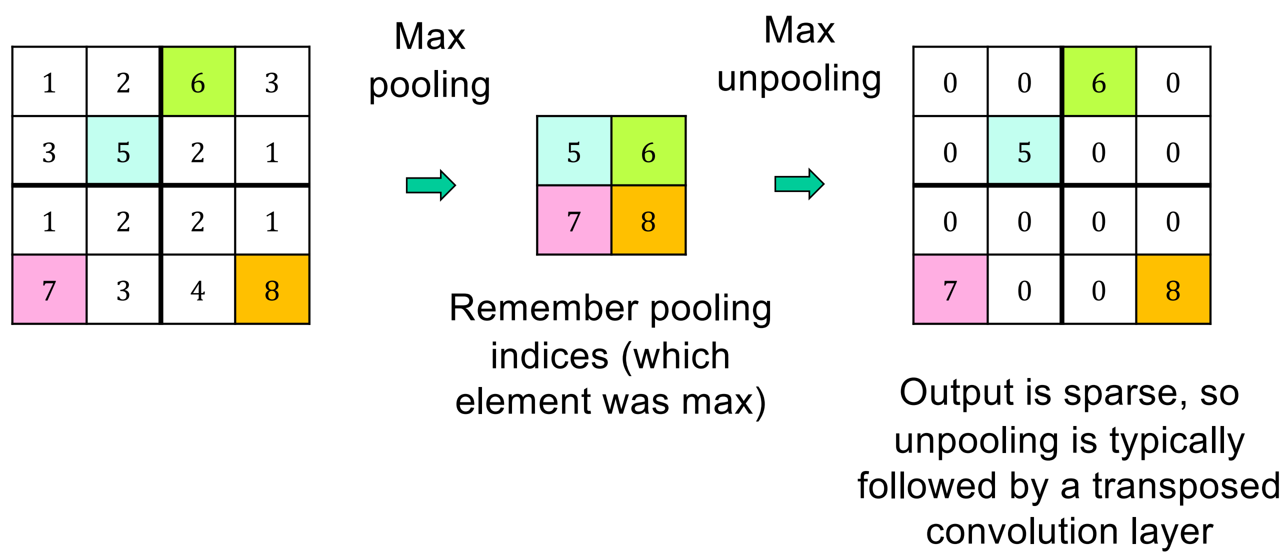
Upsampling filter:

$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
$\frac{1}{2}$	1	$\frac{1}{2}$
$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

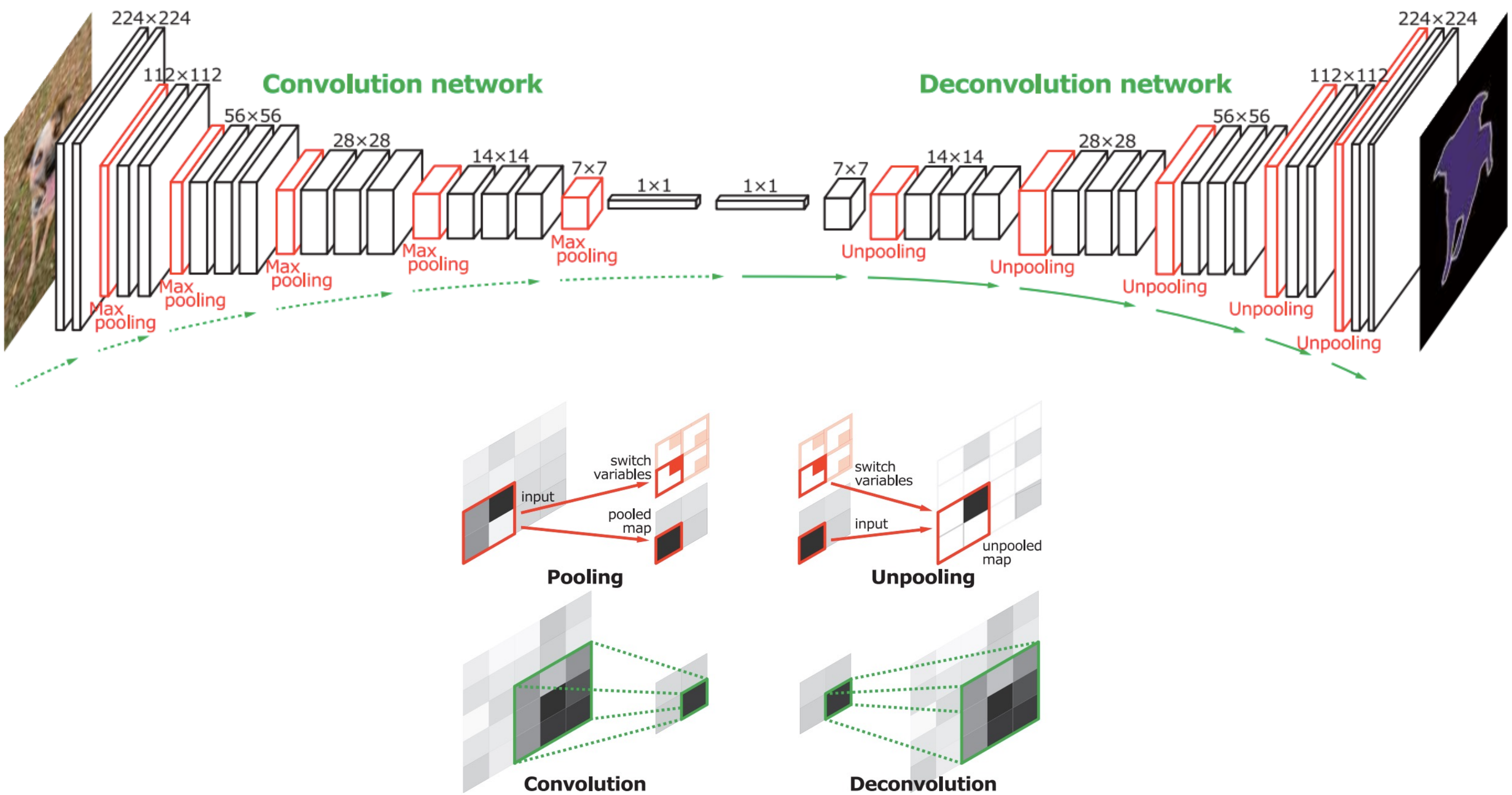
V. Dumoulin and F. Visin, [A guide to convolution arithmetic for deep learning](), arXiv 2018

Feature map upsampling: Max unpooling

- Can be used when max pooling is used to downsample

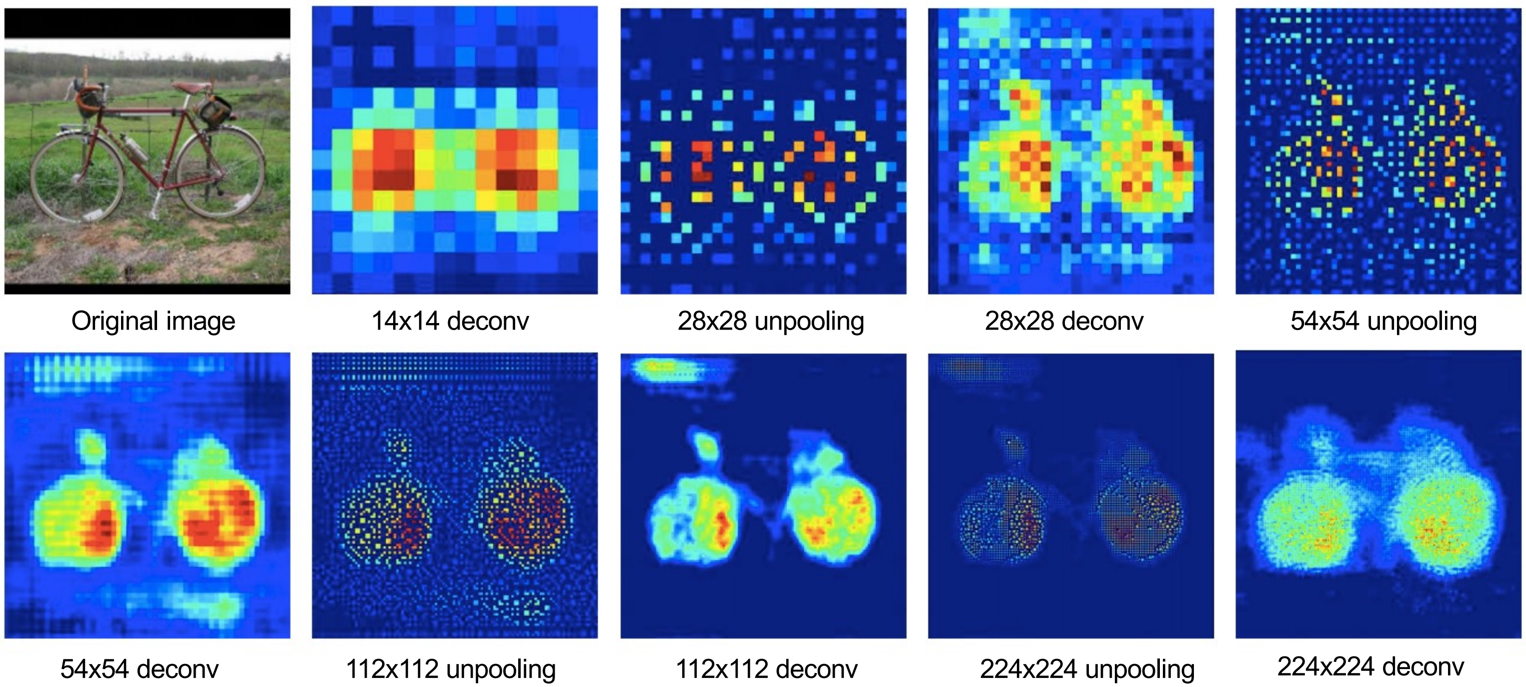


DeconvNet



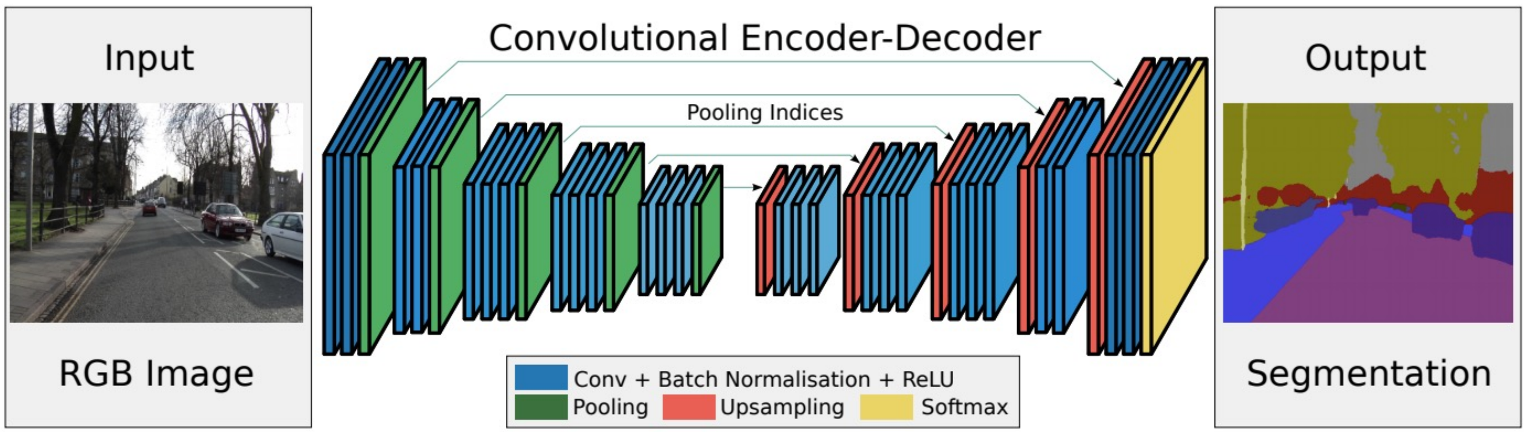
H. Noh, S. Hong, and B. Han, [Learning Deconvolution Network for Semantic Segmentation](), ICCV 2015

DeconvNet



H. Noh, S. Hong, and B. Han, Learning Deconvolution Network for Semantic Segmentation, ICCV 2015

SegNet



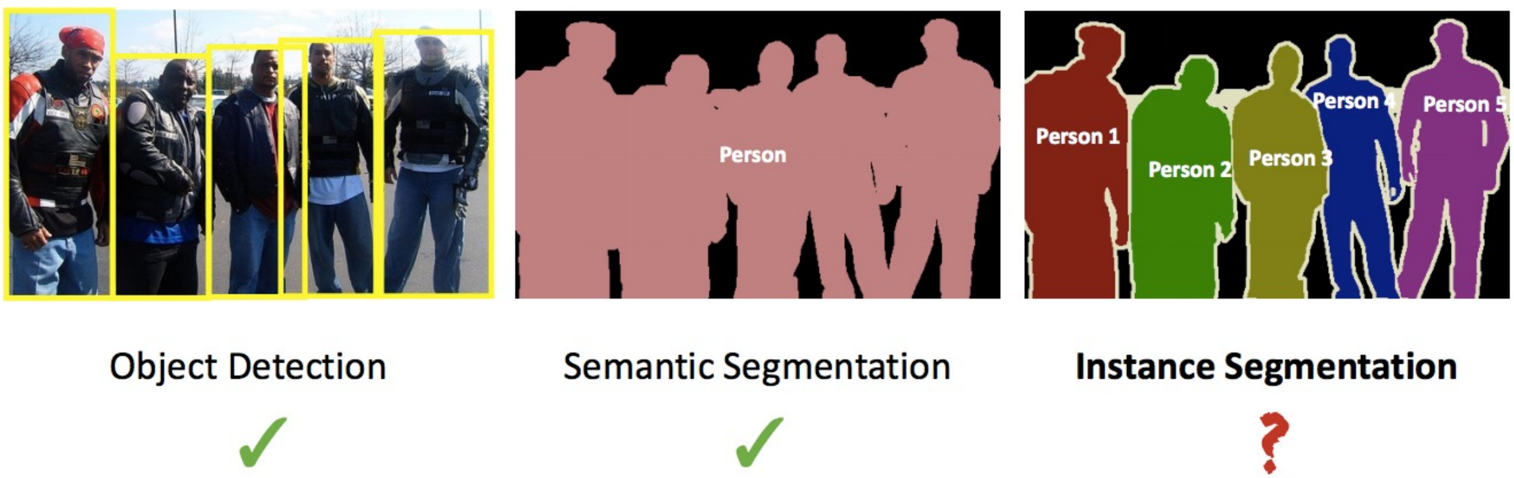
Drop the FC layers,
get better results

V. Badrinarayanan, A. Kendall and R. Cipolla, [SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation](), PAMI 2017

Dense prediction: Outline

- Operations and architectures for dense prediction: U-Net
- Instance segmentation: Mask R-CNN

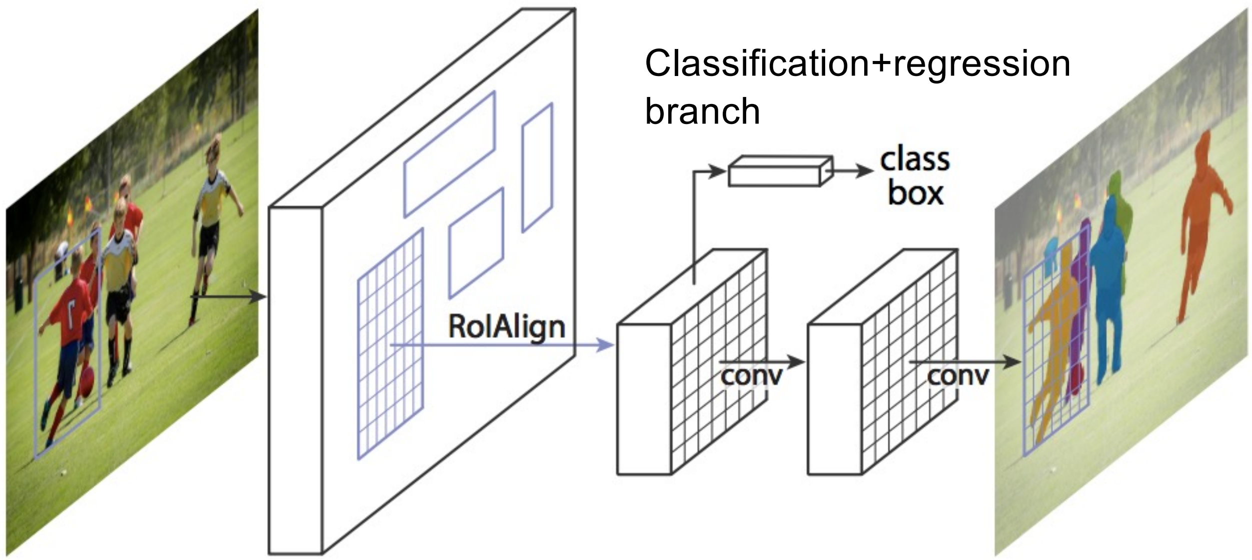
Instance segmentation



Source: Kaiming He

Mask R-CNN

- Mask R-CNN = Faster R-CNN + FCN on RoIs



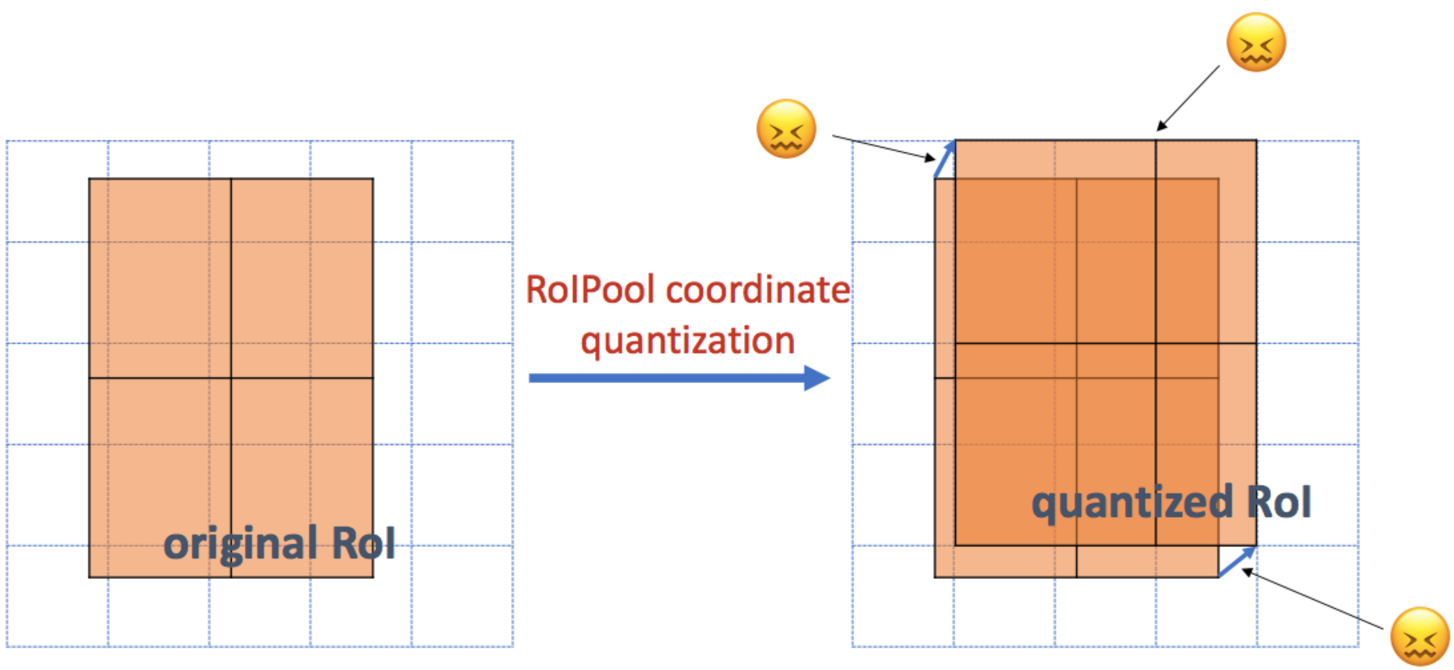
Classification+regression branch

Mask branch: separately predict segmentation for each possible class

K. He, G. Gkioxari, P. Dollar, and R. Girshick, [Mask R-CNN](), ICCV 2017 (Best Paper Award)

RoIAlign vs. RoIPool

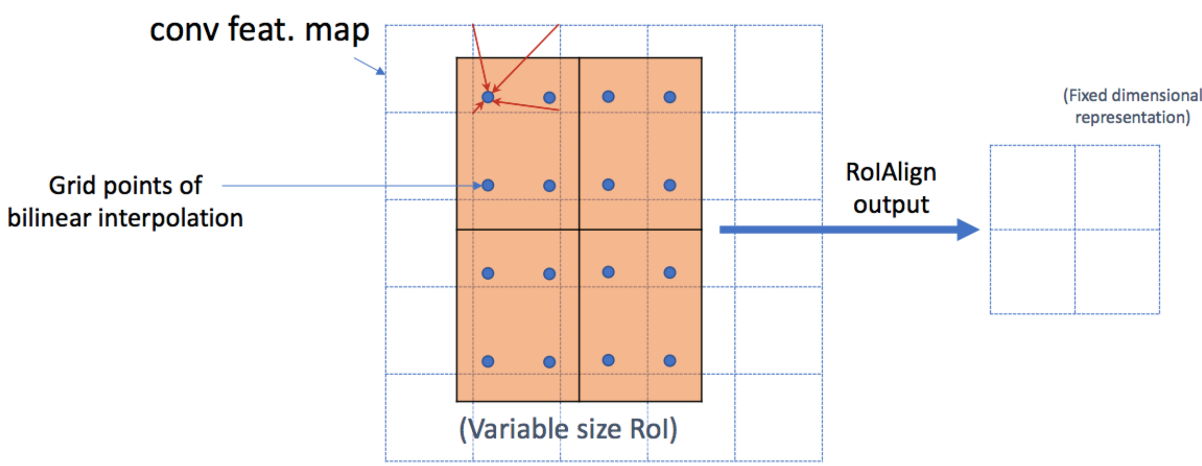
- RoIPool: nearest neighbor quantization



Source: K. He, R. Girshick

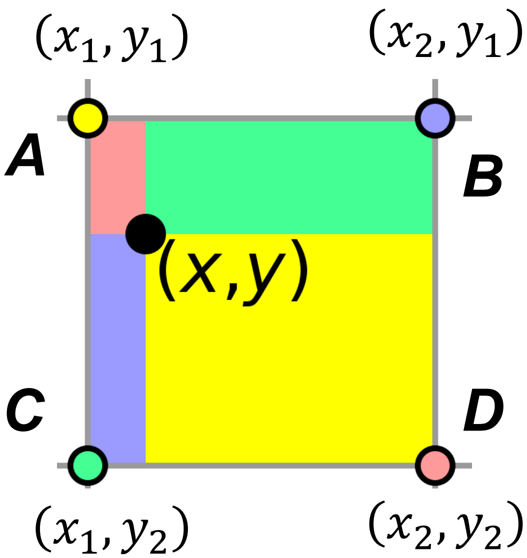
RoIAlign vs. RoIPool

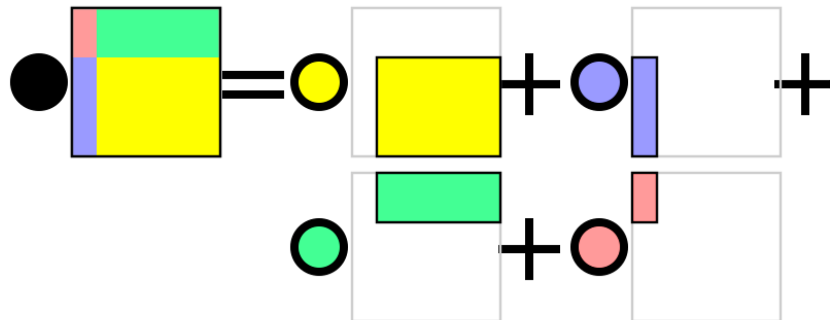
- RoIPool: nearest neighbor quantization
- RoIAlign: bilinear interpolation



Source: K. He, R. Girshick

Bilinear interpolation





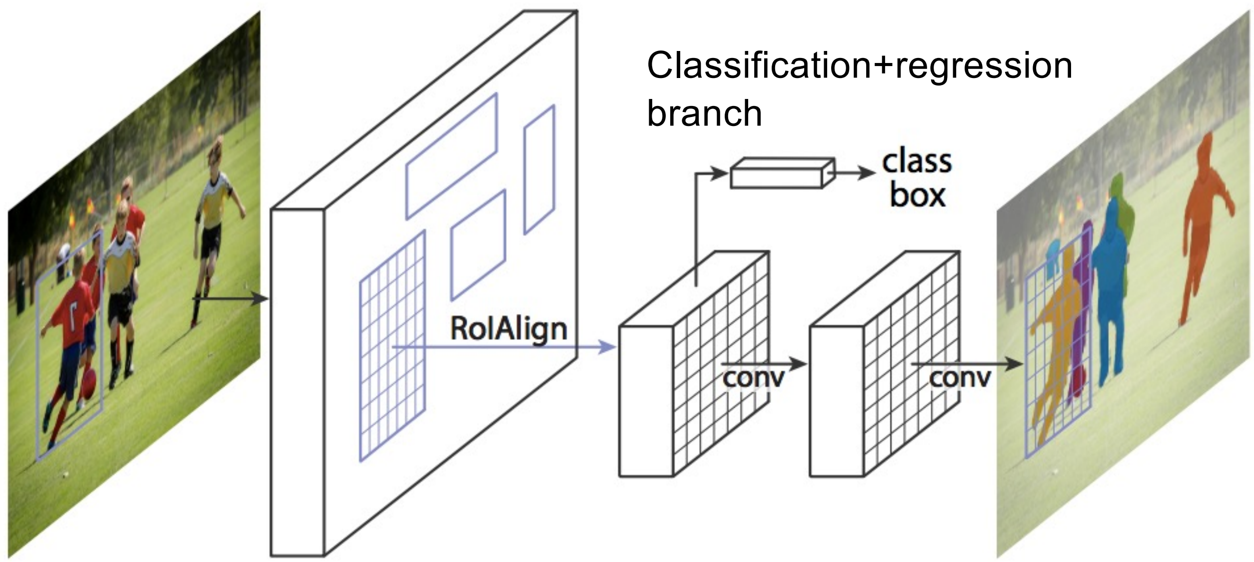
$$f(x,y) = w_{11}A + w_{21}B + w_{12}\,C + w_{22}D$$

$$w_{11} = \frac{(x_2 - x)(y_2 - y)}{(x_2 - x_1)(y_2 - y_1)} \qquad w_{21} = \frac{(x - x_1)(y_2 - y)}{(x_2 - x_1)(y_2 - y_1)}$$

$$w_{12} = \frac{(x_2 - x)(y - y_1)}{(x_2 - x_1)(y_2 - y_1)} \qquad w_{22} = \frac{(x - x_1)(y - y_1)}{(x_2 - x_1)(y_2 - y_1)}$$

http://en.wikipedia.org/wiki/Bilinear_interpolation

Mask R-CNN

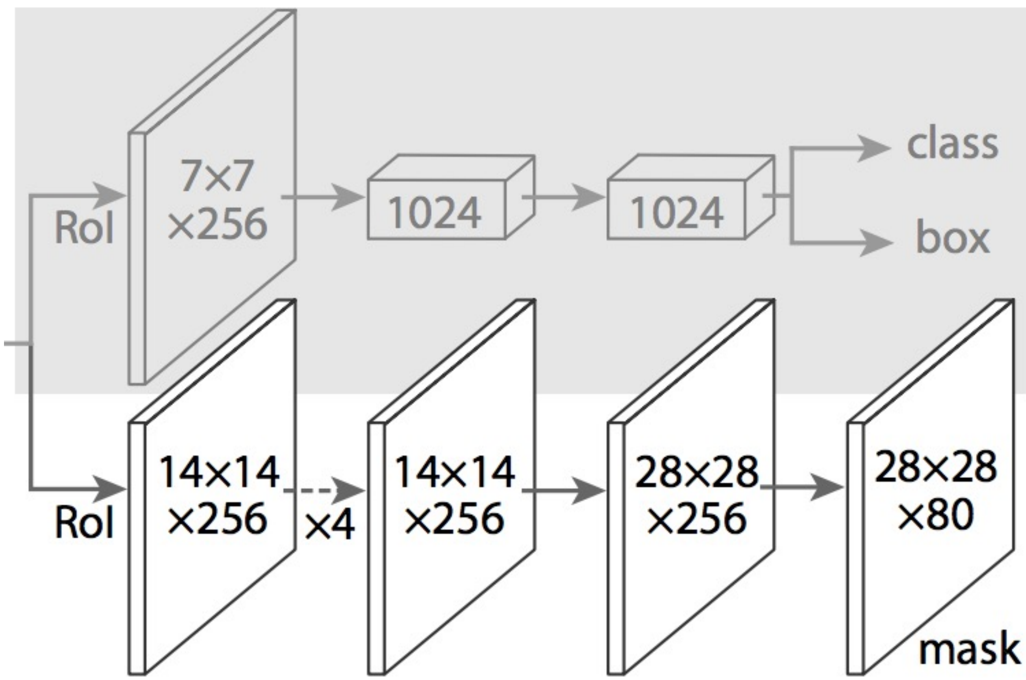


Classification+regression branch

Mask branch: separately predict segmentation for each possible class

Mask R-CNN

- From RoIAlign features, predict class label, bounding box, and segmentation mask



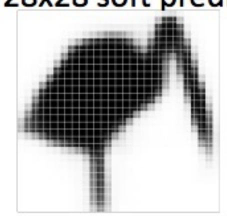
Classification/regression head from an established object detector (e.g., FPN)

Separately predict binary mask for each class with per-pixel sigmoids, use average binary cross-entropy loss

Mask R-CNN

Validation image with box detection shown in red

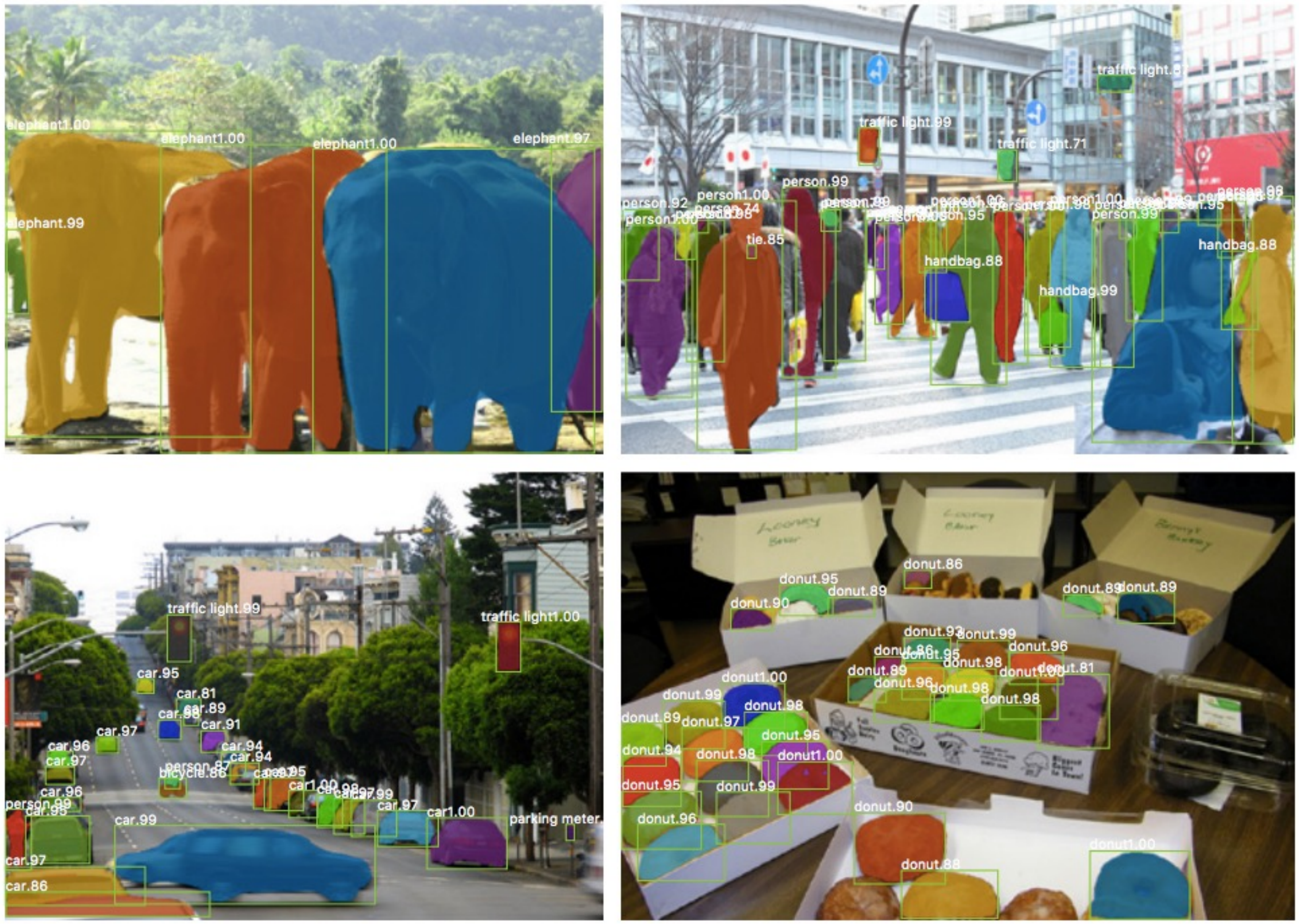
28x28 soft prediction



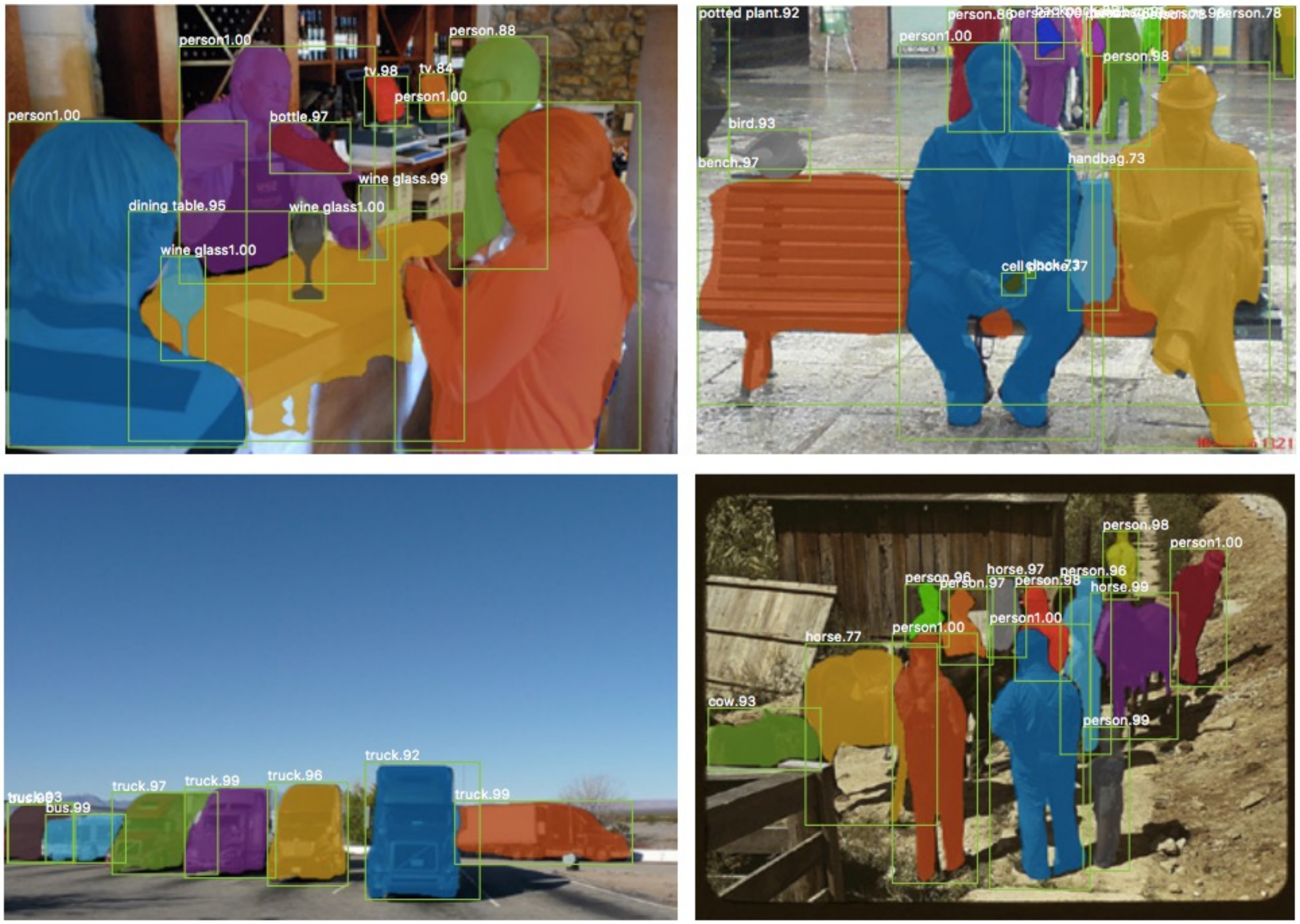
Resized Soft prediction

Final mask

Example results



Example results



Instance segmentation results on COCO

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [10]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [26] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [26] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-	-	-	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**	**16.9**	**39.9**	**53.5**

AP at different IoU thresholds

AP for different size instances

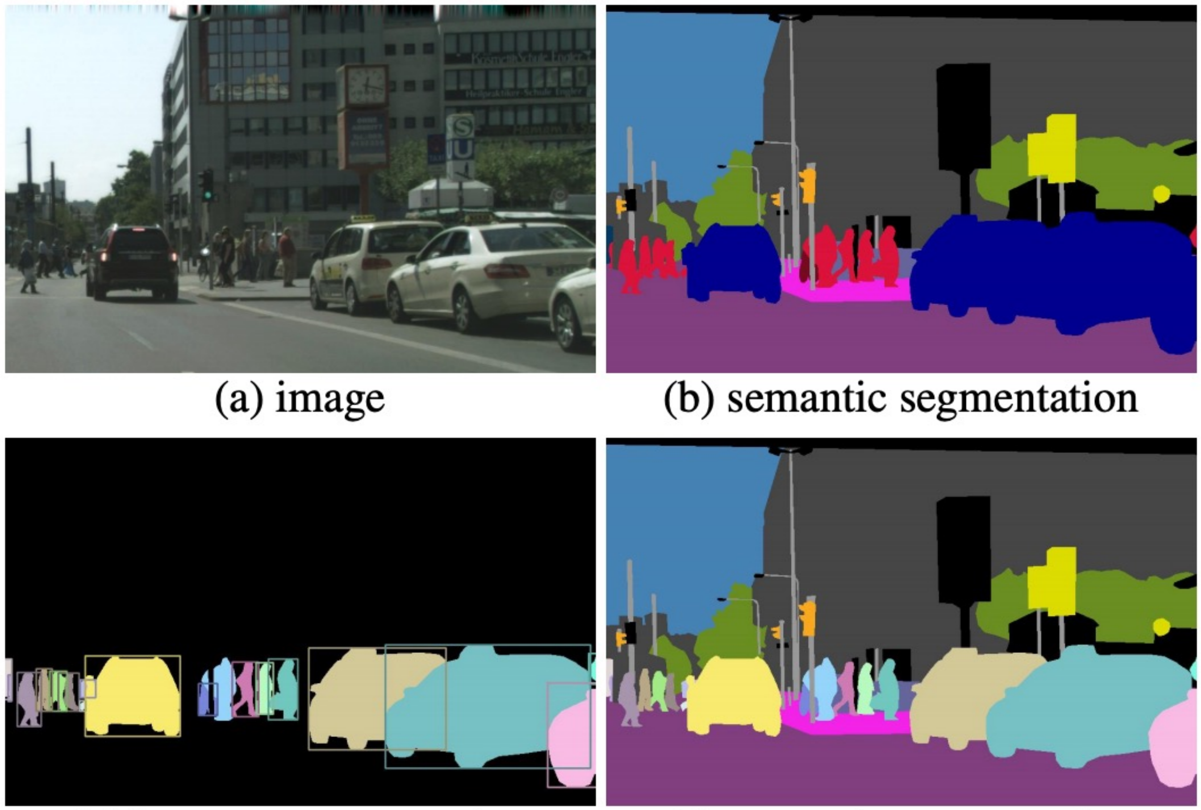
Keypoint prediction

- Given K keypoints, train model to predict K $m \times m$ *one-hot* maps with cross-entropy losses over m^2 outputs

Outline

- Operations and architectures for dense prediction: U-Net
- Instance segmentation: Mask R-CNN
- Other dense prediction problems

Panoptic segmentation

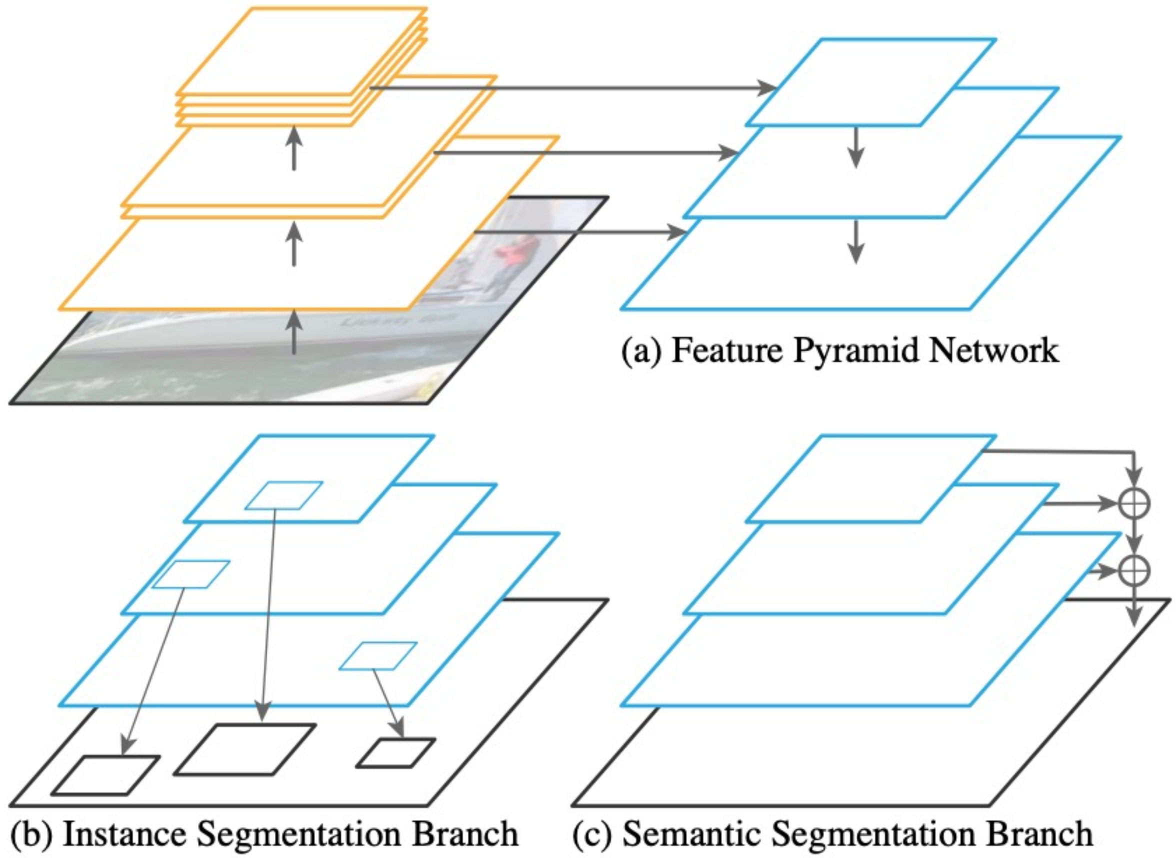


(a) image (b) semantic segmentation

(c) instance segmentation (d) panoptic segmentation

A. Kirillov et al. [Panoptic segmentation](). CVPR 2019

Panoptic feature pyramid networks



A. Kirillov et al. [Panoptic feature pyramid networks](). CVPR 2019

Panoptic feature pyramid networks

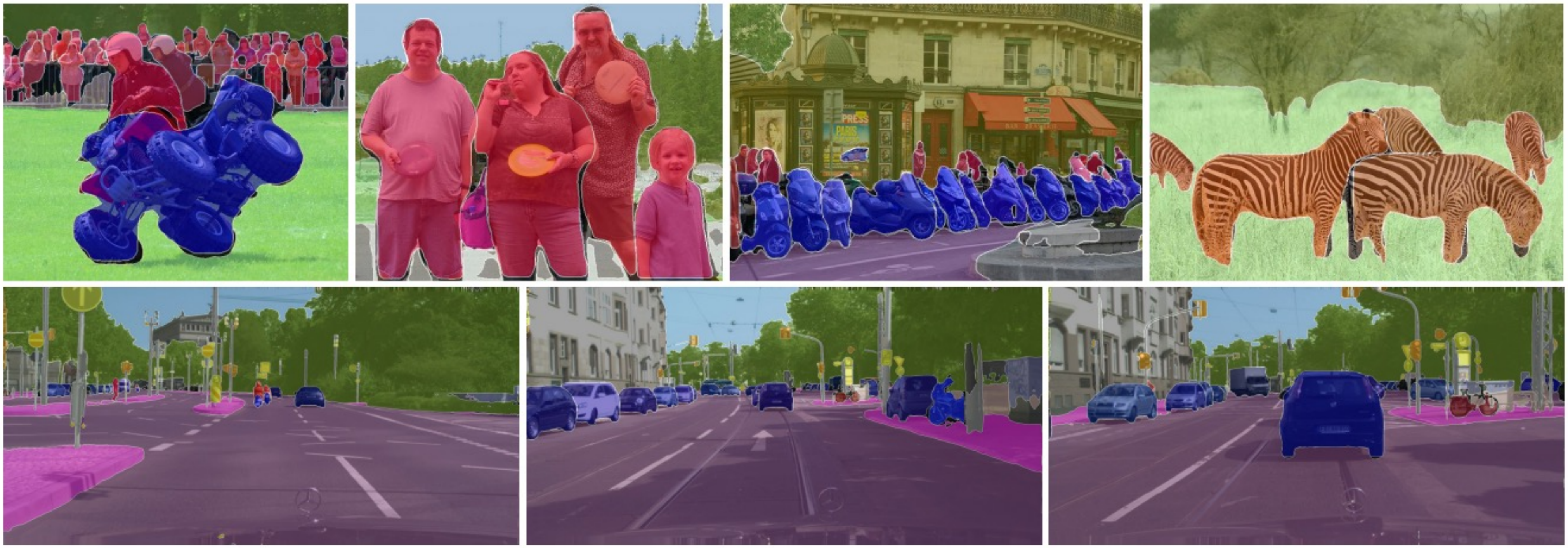


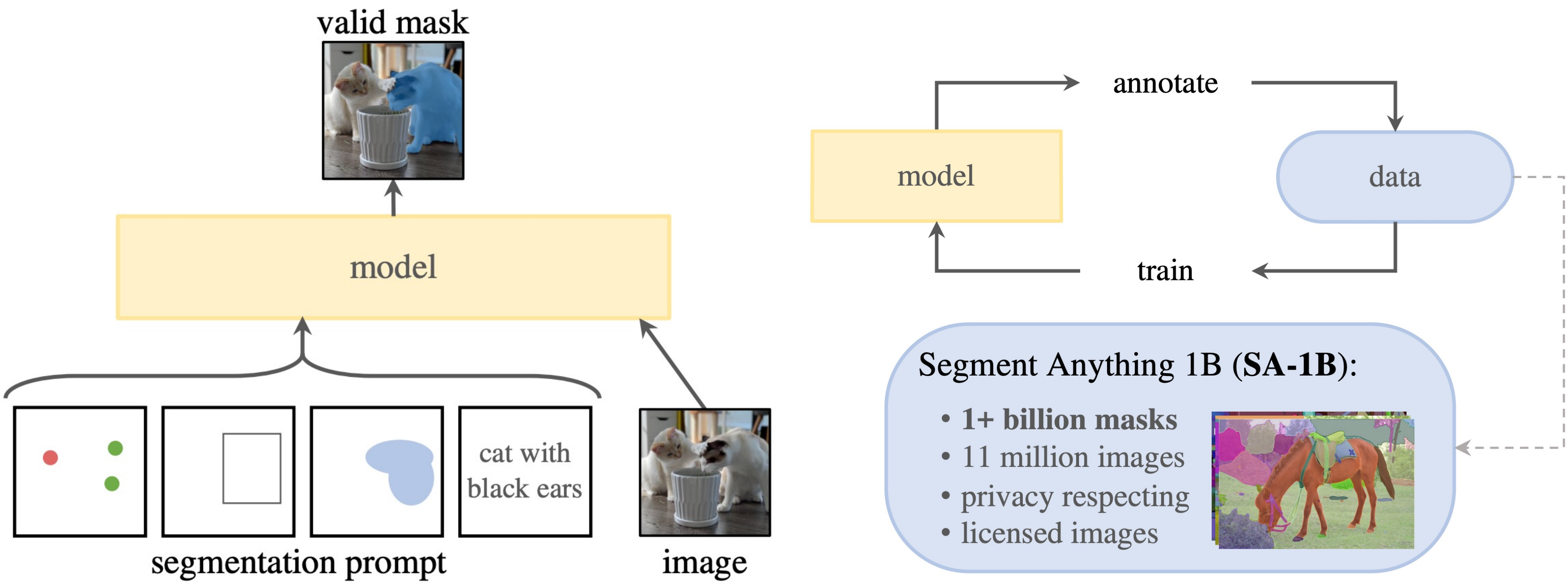
Figure 2: Panoptic FPN results on COCO (top) and Cityscapes (bottom) using a single ResNet-101-FPN network.

A. Kirillov et al. [Panoptic feature pyramid networks](). CVPR 2019

Amodal instance segmentation

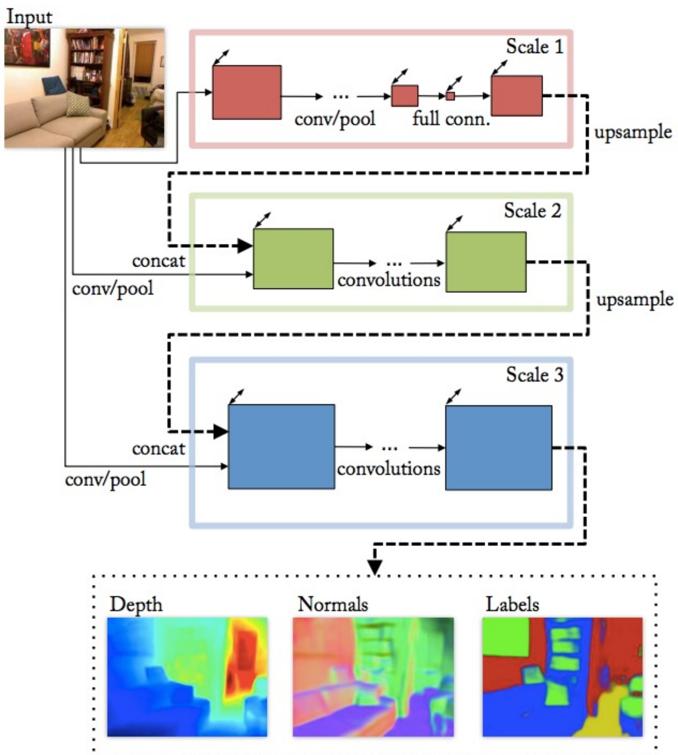
K. Li and J. Malik. [Amodal instance segmentation.](#) ECCV 2016

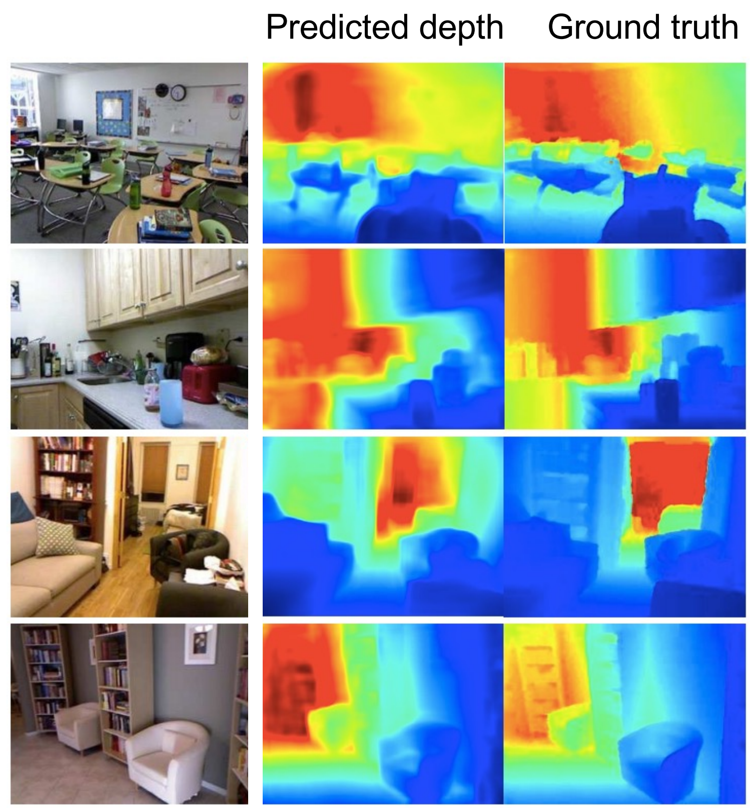
Promptable segmentation



A. Kirilllov et al. [Segment anything](). ICCV 2023

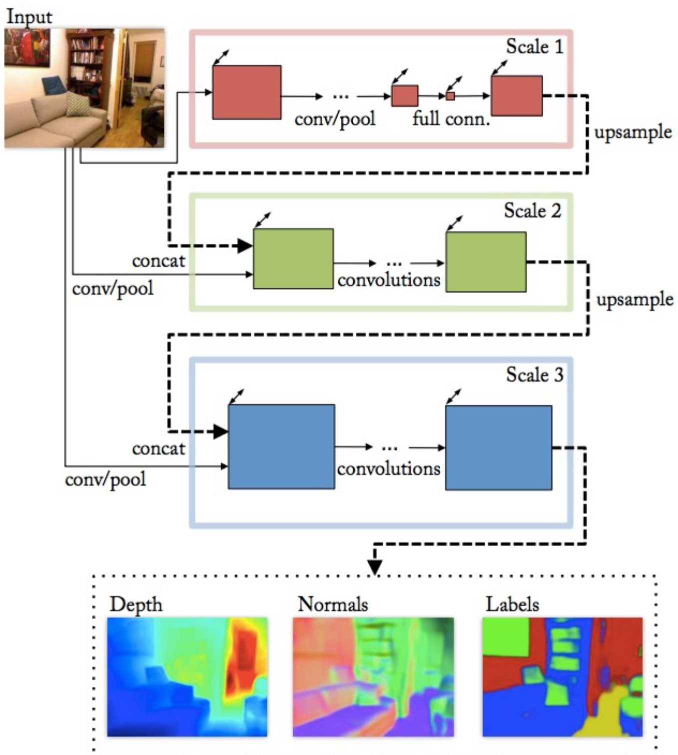
Depth and normal estimation





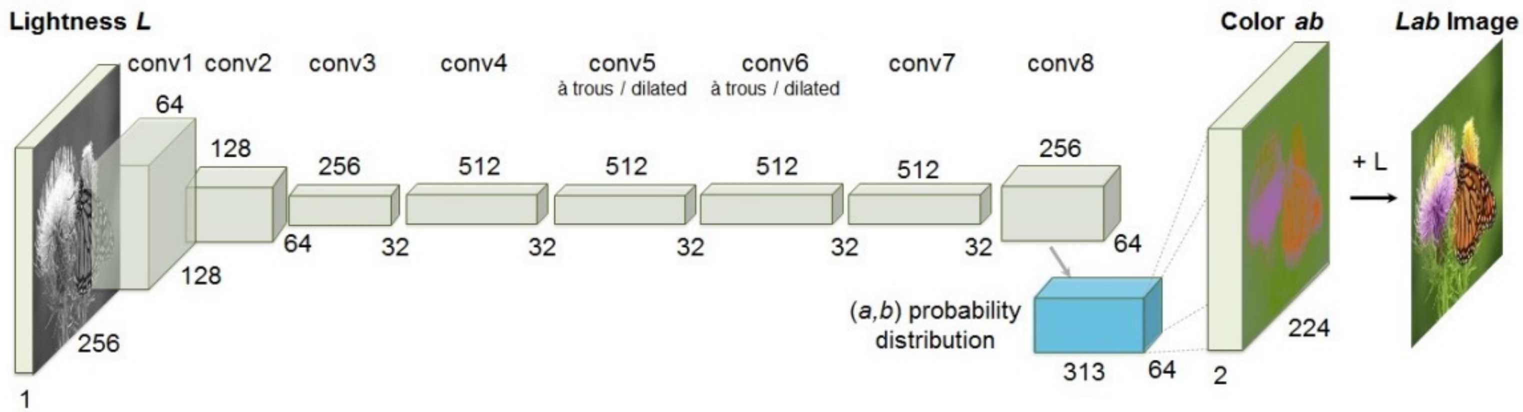
D. Eigen and R. Fergus, [Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture](), ICCV 2015

Depth and normal estimation



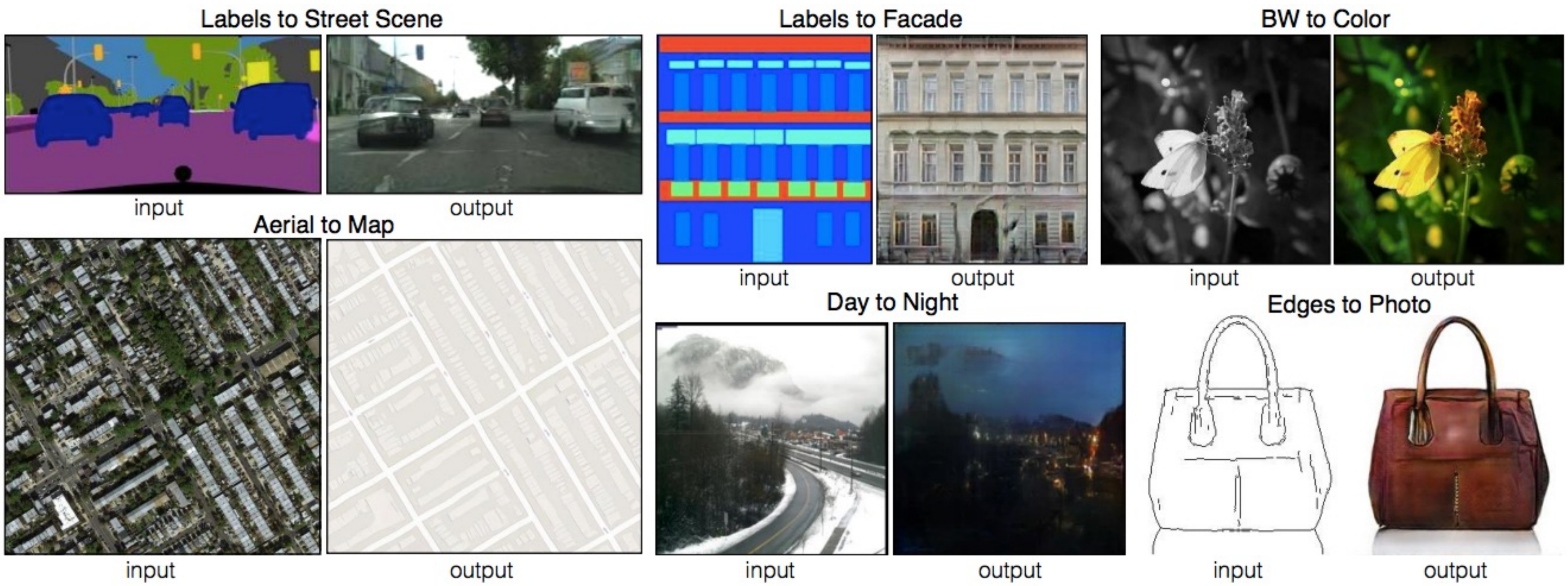
D. Eigen and R. Fergus, [Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture](), ICCV 2015

Colorization



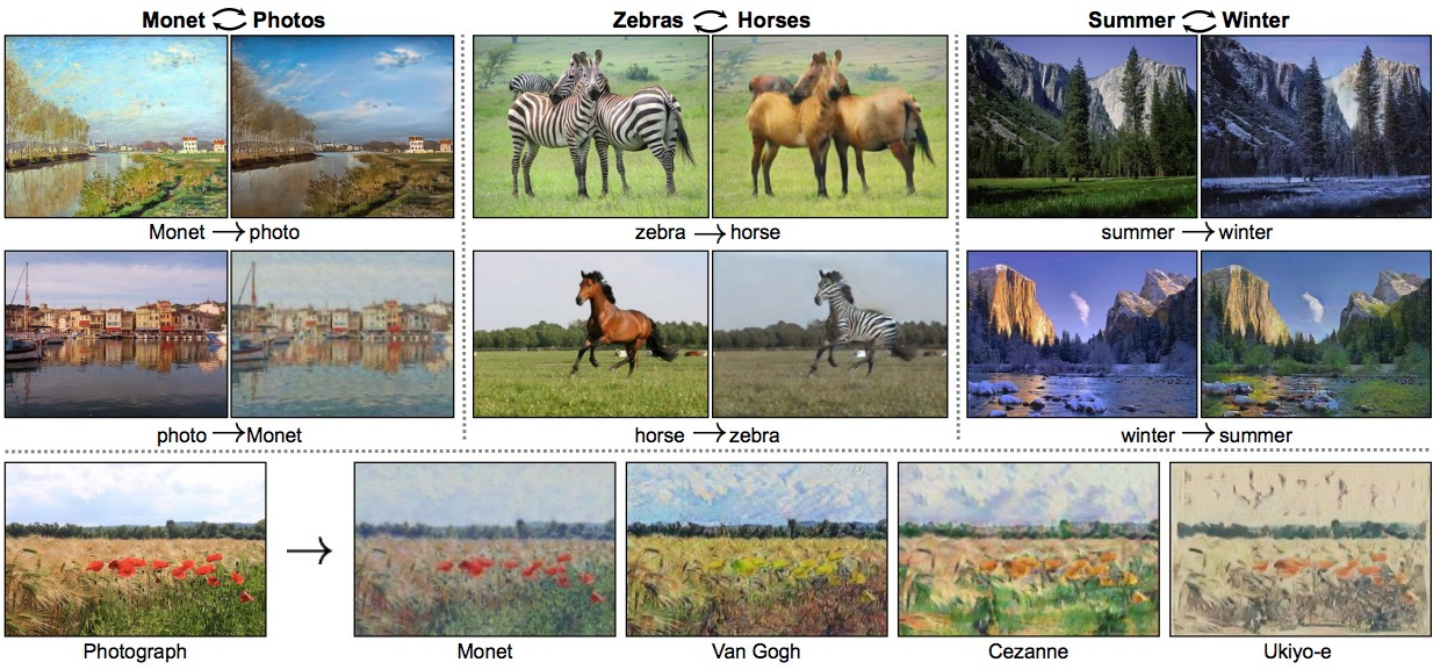
R. Zhang, P. Isola, and A. Efros, [Colorful Image Colorization](), ECCV 2016

Image-to-image translation (paired)



P. Isola, J.-Y. Zhu, T. Zhou, A. Efros, [Image-to-Image Translation with Conditional Adversarial Networks](), CVPR 2017

Image-to-image translation (unpaired)



J.-Y. Zhu, T. Park, P. Isola, A. Efros, [Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks](), ICCV 2017

Image generation

Encoder-decoder