Segmentation

Computer Vision – Lecture 11

Further Reading

- Slides from S Lazebnik
- Slides from J Johnson
- Slides from A Geiger

Granularity

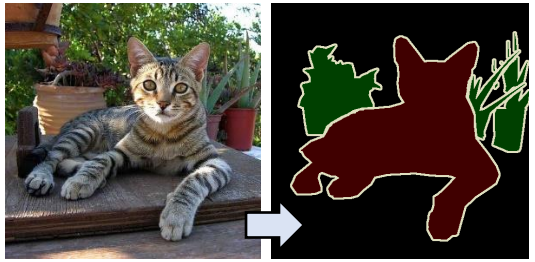
Image classification

Image-scale

Object detection

Localization (Regions)

Segmentation

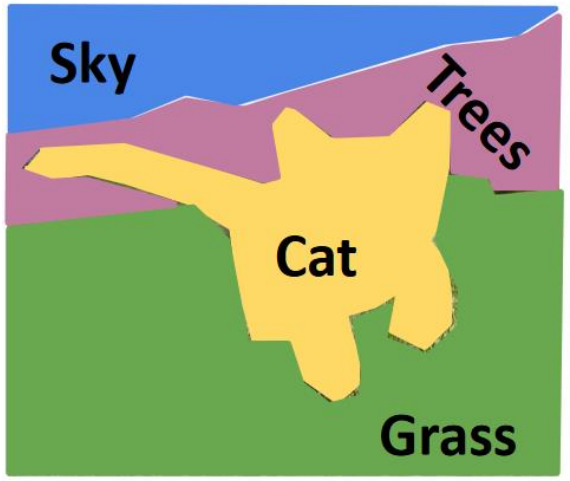


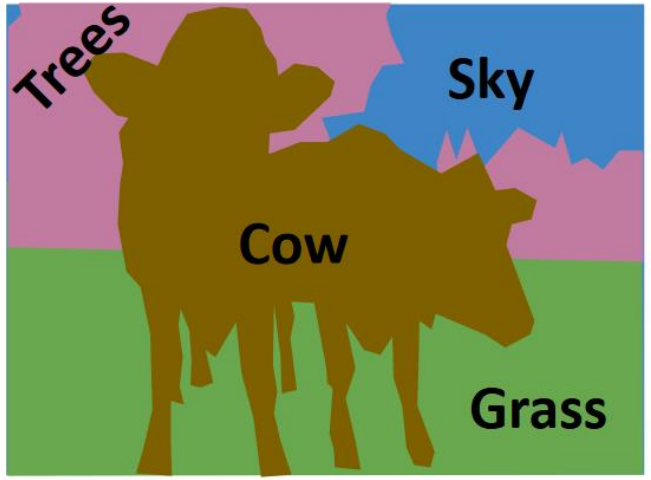
Pixel-level

Spatial Labelling

Semantic Segmentation

- Label each pixel in the image with a category label
- Do not differentiate instances, only care about pixels





Slide: J Johnson

Evaluation

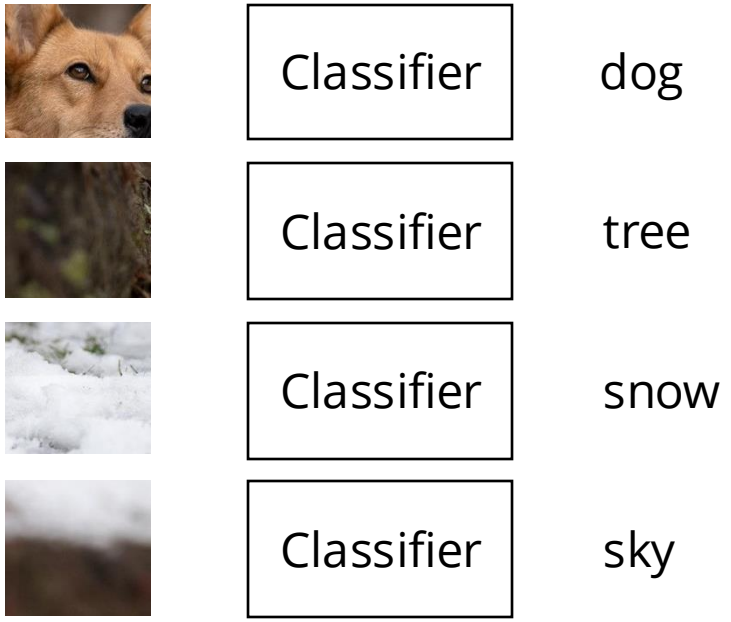
- Per-class IoU.
- Since there are no instances, evaluation is much easier than detection.
- Compare the binary masks of prediction and GT for each class.
- Average over classes.

- Note: IoU is favourable to large objects. (harder to intersect small objects)

Sliding Window

Similar to detection:

- Slide a window over the whole image.
- Classify the centre pixel of the window.



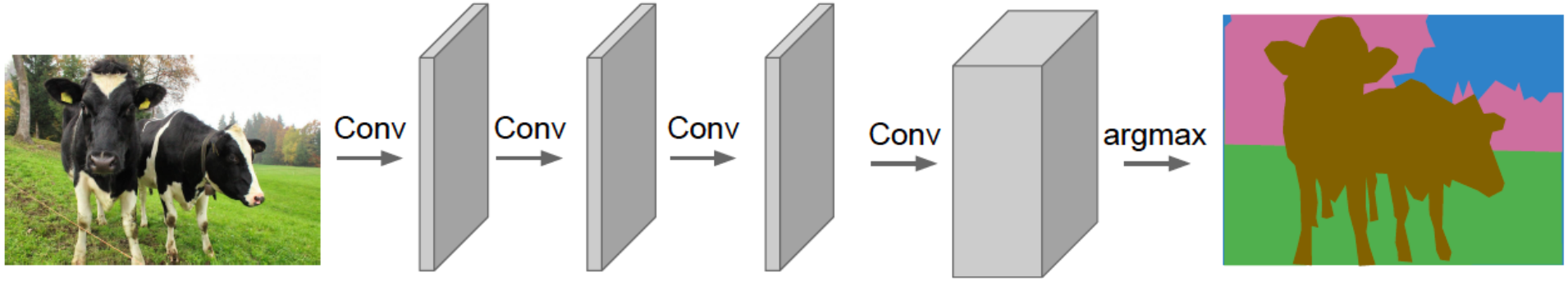
Farabet et al, “Learning Hierarchical Features for Scene Labeling,” TPAMI 2013
Pinheiro and Collobert, “Recurrent Convolutional Neural Networks for Scene Labeling”, ICML 2014

Sliding Window Segmentation

- Very inefficient!
- Very noisy: independent decision for each patch.
- Observation: we are recomputing the same features for overlapping patches.
- How can we share computation?
- Convolutions!

Fully Convolutional Networks

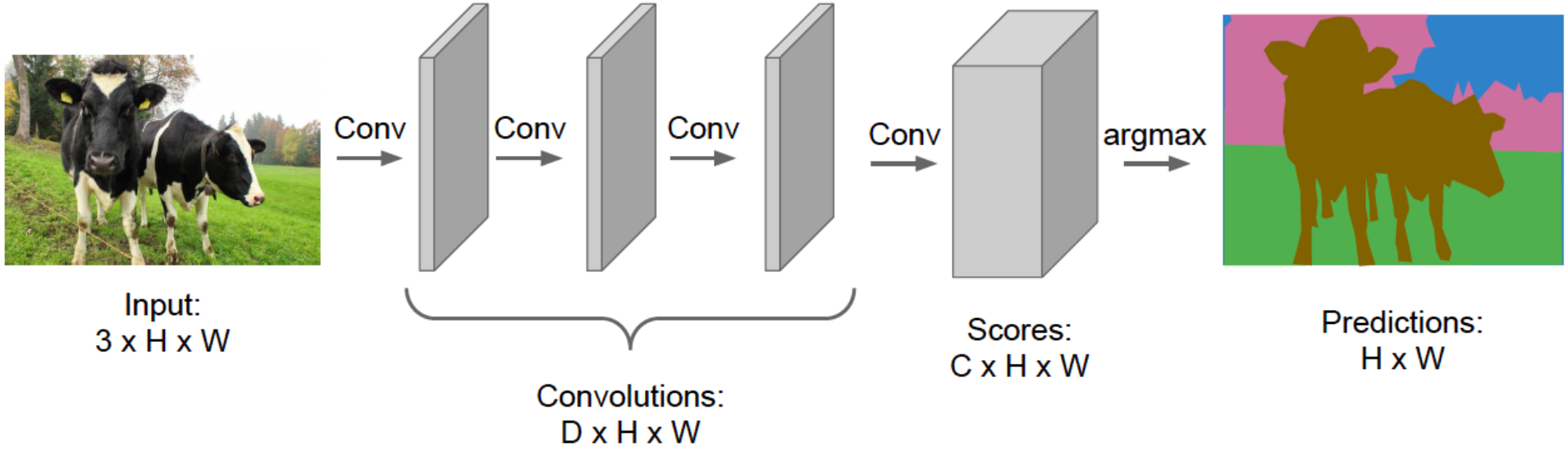
If we do not down-sample (and always pad appropriately) we can design a network with convolutions that has the same input and output size.



Source: Stanford CS231n

Fully Convolutional Networks

If we do not down-sample (and always pad appropriately) we can design a network with convolutions that has the same input and output size.

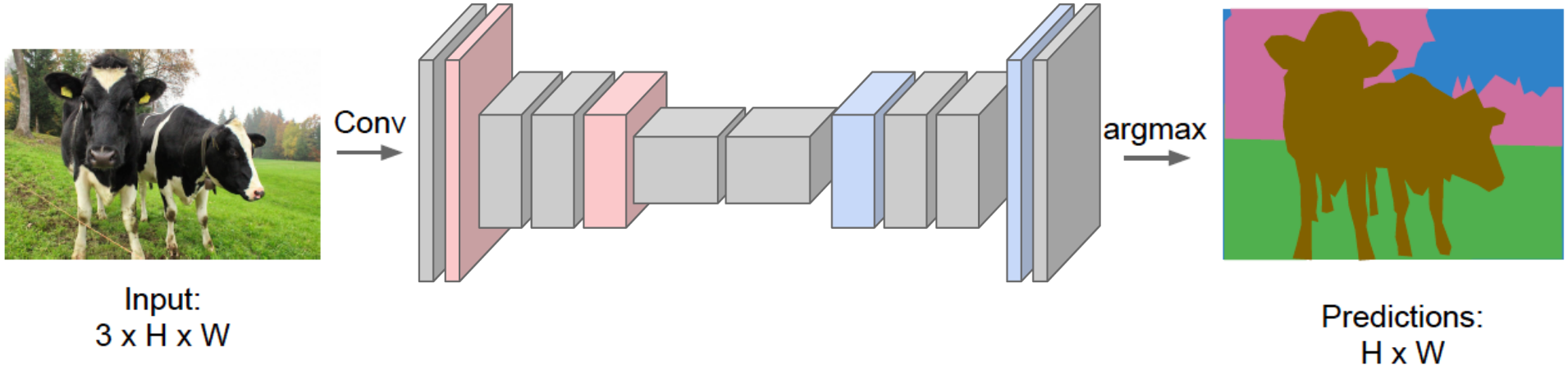


Source: Stanford CS231n

FCN: Fully Convolutional Networks

If we do not down-sample (and always pad appropriately) we can design a network with convolutions that has the same input and output size.

Very costly at full resolution. Down-sample, then up-sample!



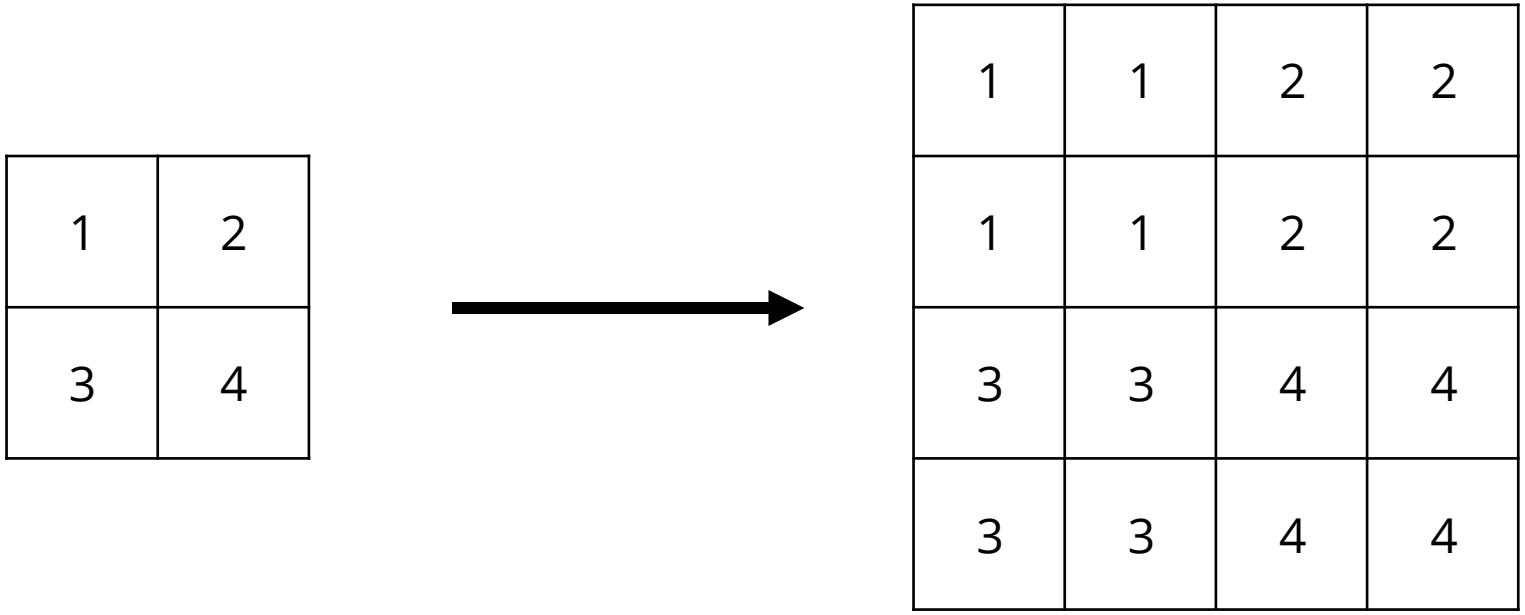
J. Long, E. Shelhamer, and T. Darrell, Fully Convolutional Networks for Semantic Segmentation, CVPR 2015

Source: Stanford CS231n

Upsampling: Unpooling

- “Inverse” max/avg-pooling operation.
- Pooling is not an invertible function: information is lost and cannot be recovered.
- Several options to approximate it.

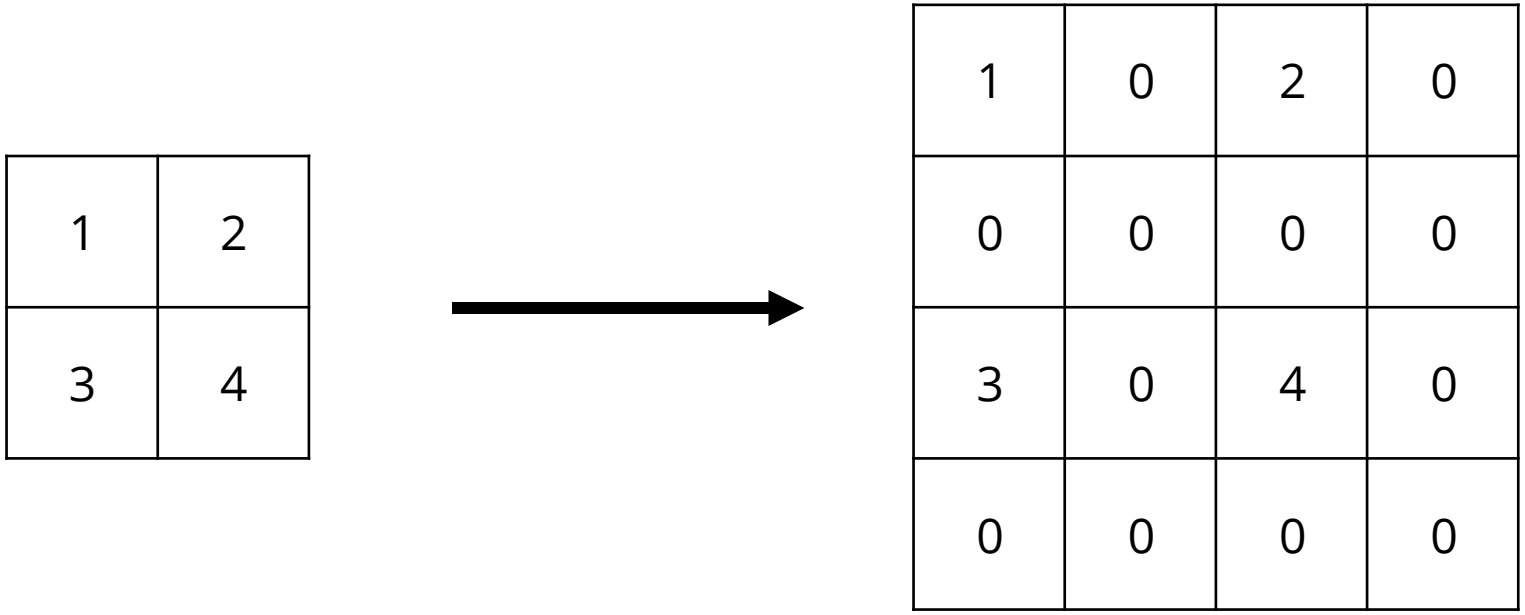
Unpooling: Nearest Neighbour



Input
C x H x W

Output
C x 2H x 2W

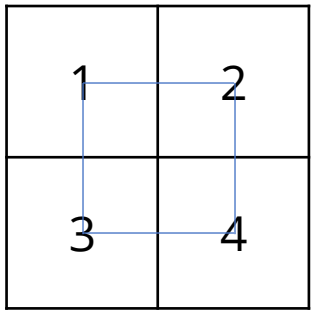
Unpooling: Bed of Nails



Input
C x H x W

Output
C x 2H x 2W

Unpooling: Bilinear Interpolation

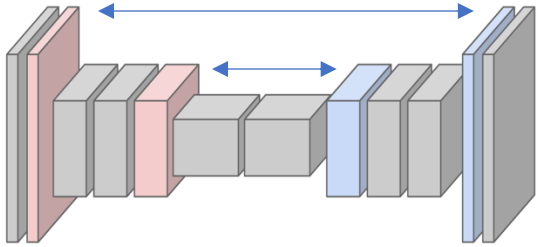


1	1.25	1.75	2
1.50	1.75	2.25	2.50
2.50	2.75	3.25	3.50
3	3.25	3.75	5

Input
C x H x W

Output
C x 2H x 2W

Max-Unpooling



While down-sampling:
remember locations

1	2	**6**	3
3	**5**	2	1
4	5	6	2
7	1	4	**8**

→

5	6
7	8

→ Other layers →

1	2
3	4

→

While up-sampling: use
remembered locations

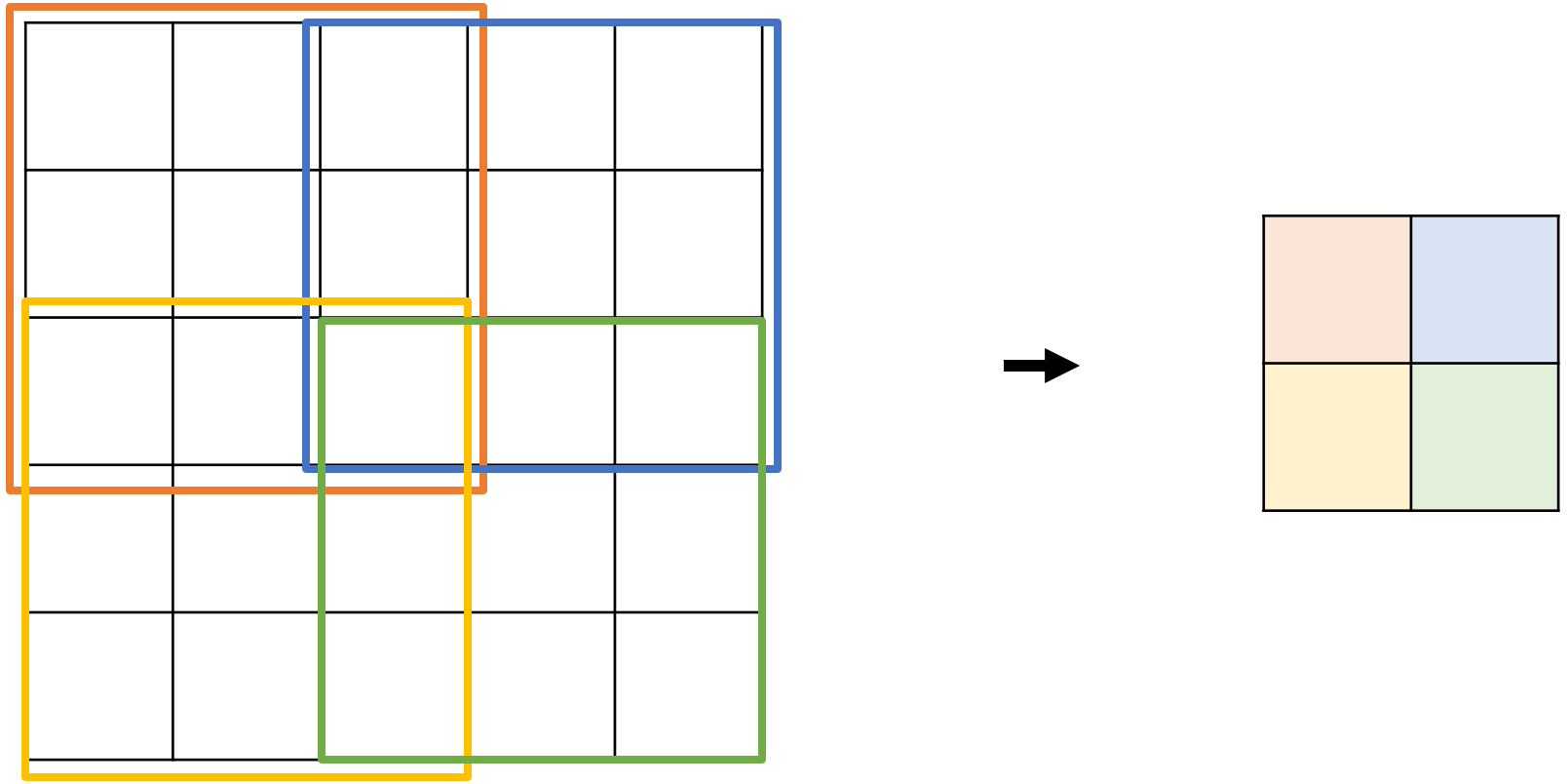
0	0	2	0
0	1	0	0
0	0	0	0
3	0	0	4

Noh et al, “Learning Deconvolution Network for Semantic Segmentation”, ICCV 2015

Slide: J Johnson

Convolution with Stride

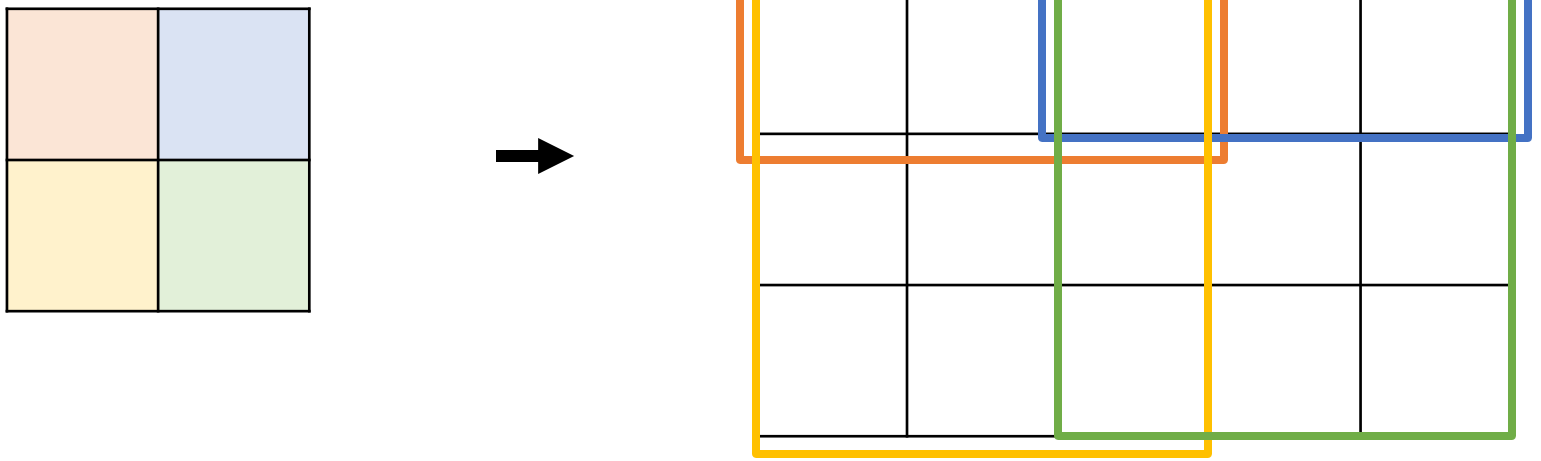
3x3 Convolution with stride 2



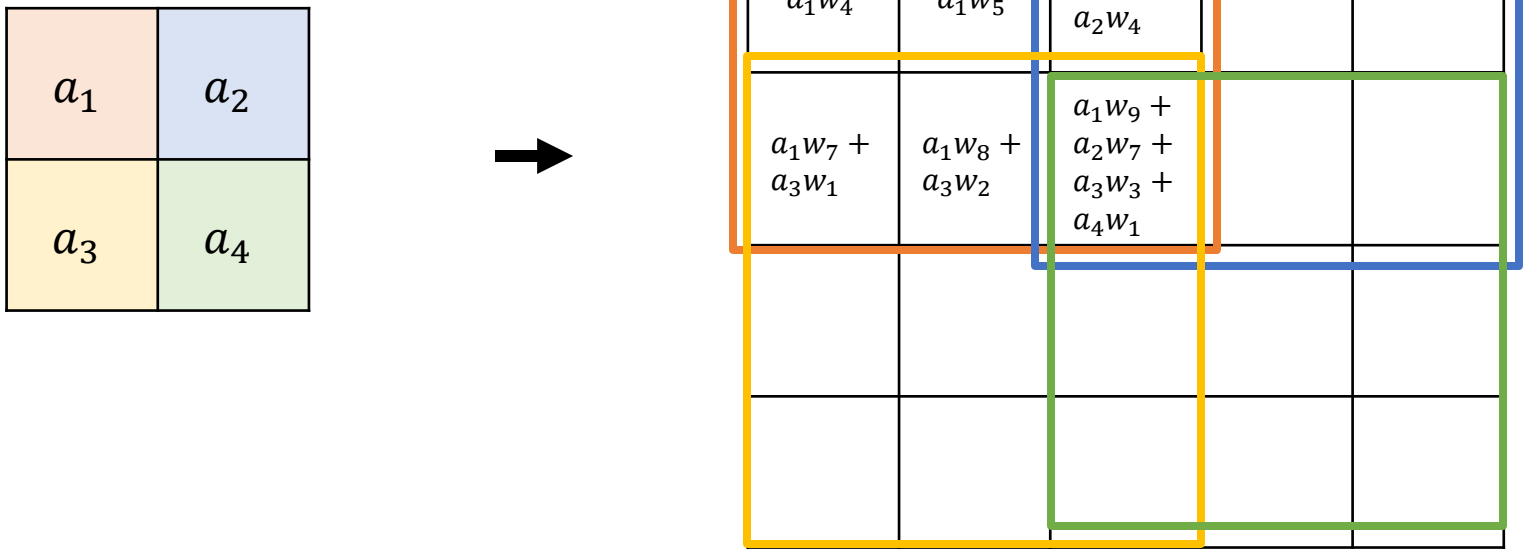
Transposed Convolution

3x3 convolution transpose, stride 2

- 9 weights
- Scale by input value
- Sum in overlapping regions



Transposed Convolution



Transposed Convolution

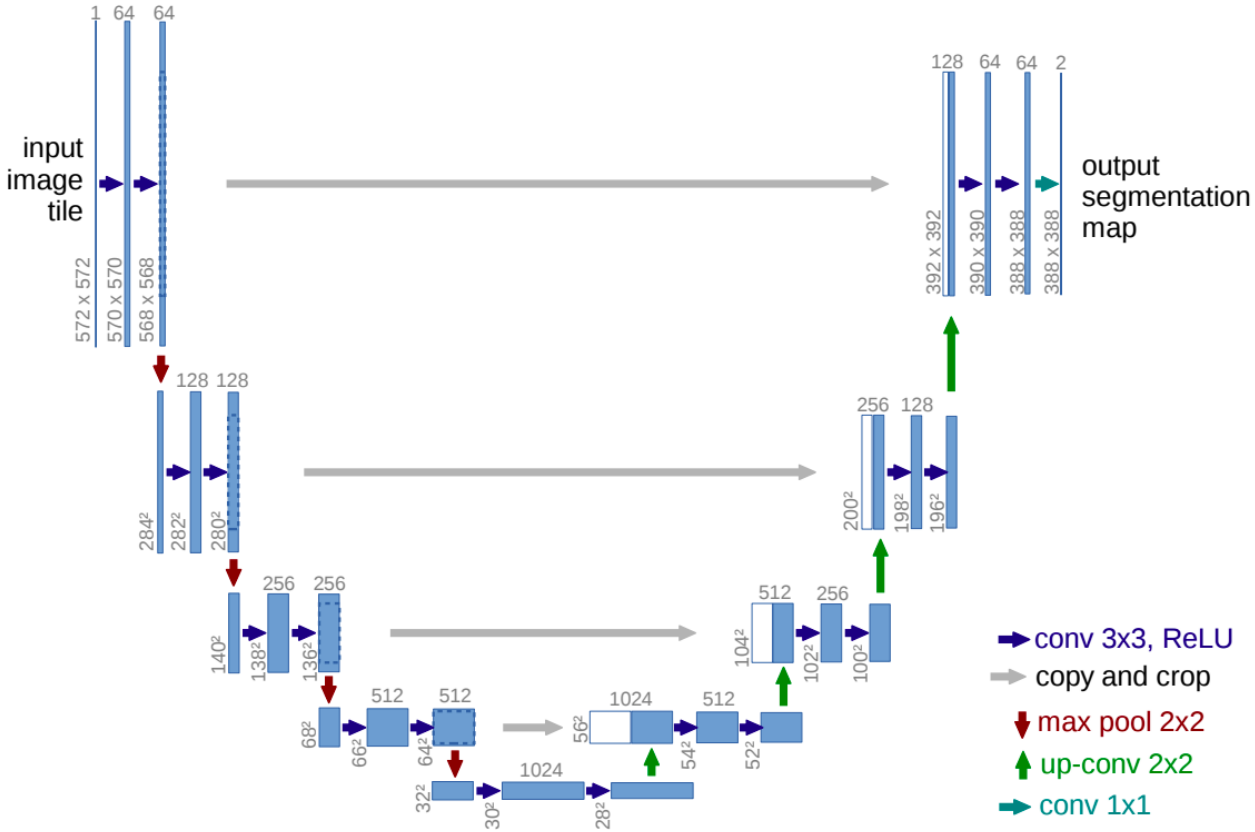
- Learnable upsampling.
- Outputs are the sum of 1-4 values: often grid pattern in output. Add normal convolution to learn smoothing.

Other names (can be confusing)

- Deconvolution
- Upconvolution
- Fractionally strided convolution
- Backward strided convolution

U-Net

- Skip-connections.
- Concatenate upsampled higher-level feature maps with higher-res, lower-level feature maps.
- Low-level feature maps: details.
- High-level feature maps: semantics.



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

Dense Prediction Architectures

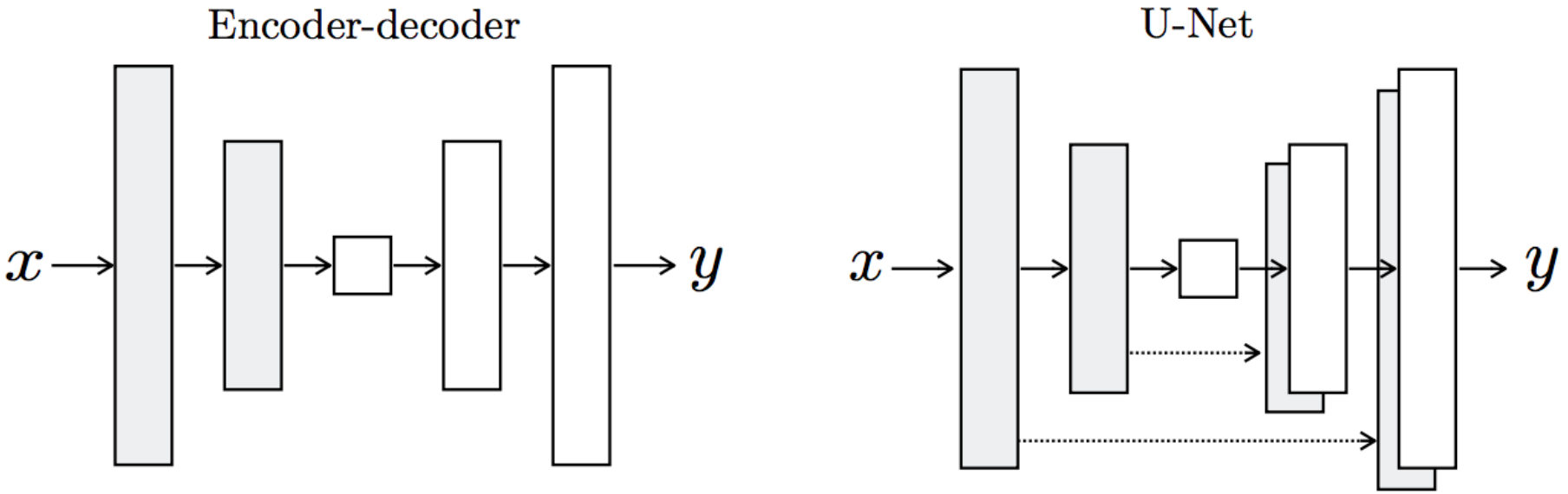
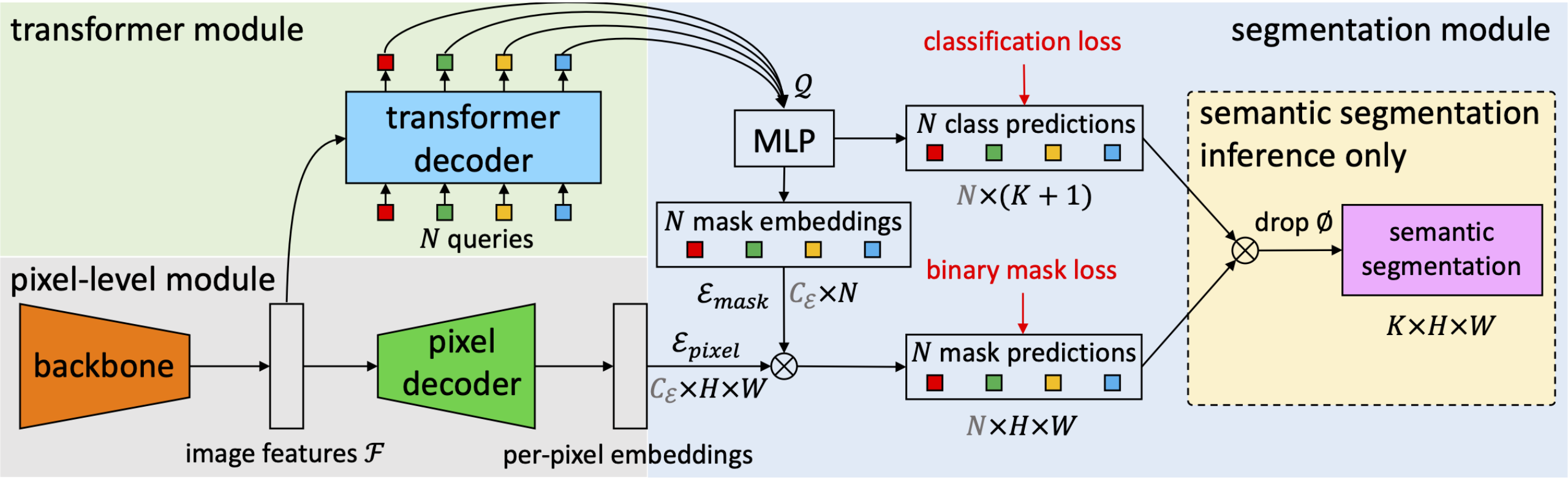


Figure source

Transformer Architectures



Cheng, Schwing, Kirillov, "MaskFormer: Per-Pixel Classification is Not All You Need for Semantic Segmentation", NeurIPS 2021

Interactive Segmentation

- User specifies what to segment.
- Input:
  - Seed points
  - Scribbles
  - Bounding box
  - Text
  - …
- Model segments corresponding object(s).

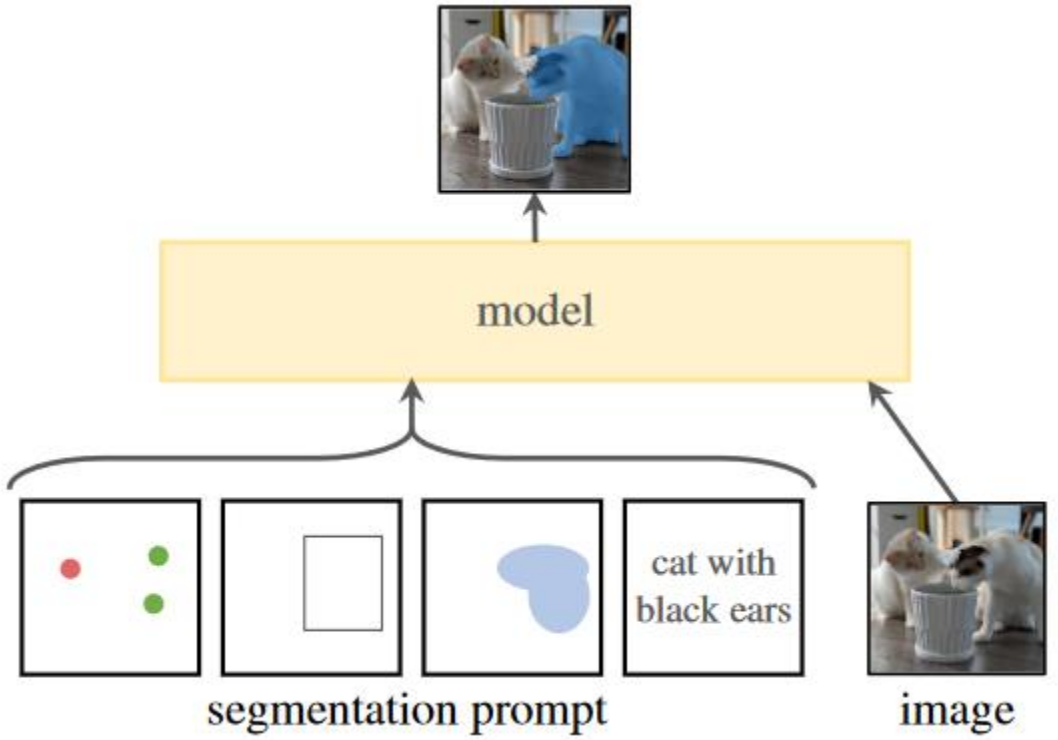


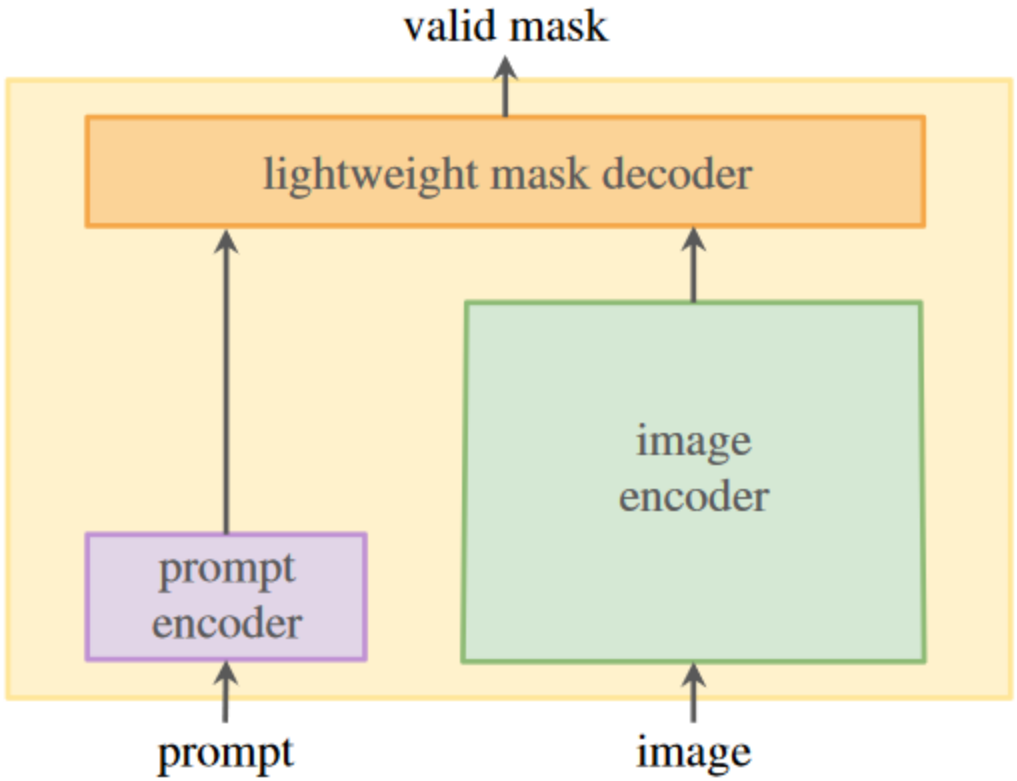
Image source

SAM: Segment Anything Model

Kirilliov et al., 2023

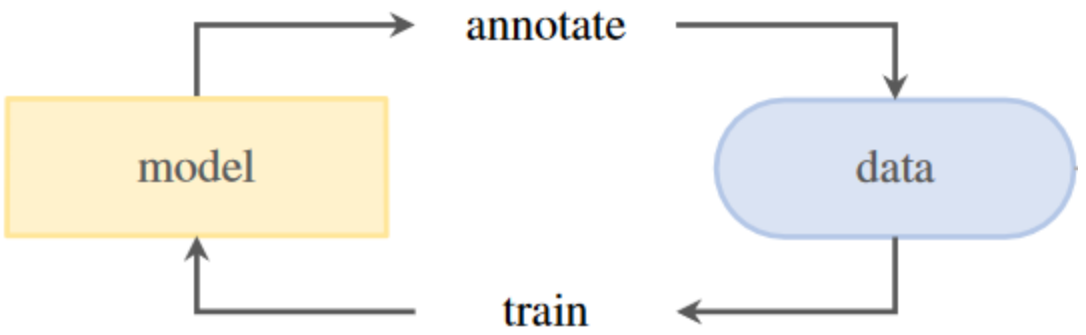
- Trained a model on various input modalities.

- Large scale supervision
  - 1B masks
  - 11M images



Training with Humans in the Loop

1. Annotate data.
2. Train model.
3. Label more data with the model.
4. Humans fix, improve labels.
5. Goto 2.



Large-Scale Annotations

Image source

Interactive Segmentation

- User centric.
- SAM: no class labels, just binary segmentation.

Other types:

- Foreground/background segmentation
- Referring expressions segmentation (“The man with the blue hat”)
- Saliency segmentation (what stands out in the image)

Things and Stuff

Thing: An object with a specific size and shape.

Stuff: Material defined by a homogeneous or repetitive pattern of fine-scale properties, but has no specific or distinctive spatial extent or shape

- Object detection: things (instances)
- Semantic segmentation: things and stuff (but no instances)

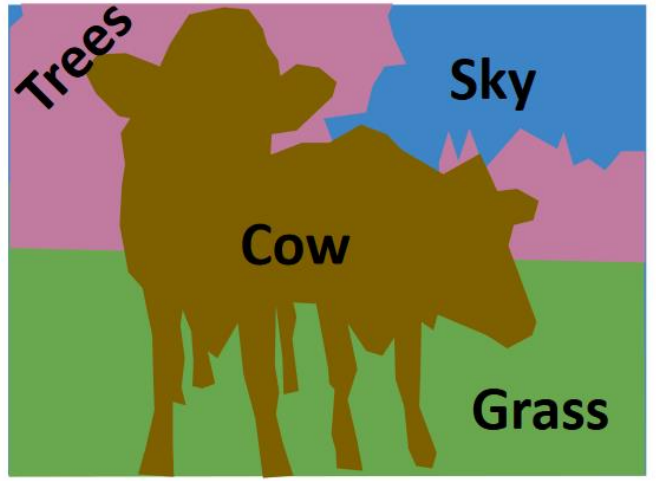
Things:
- Dog
- Tree
- Lantern
- …

Stuff:
- Sky
- Snow
- …

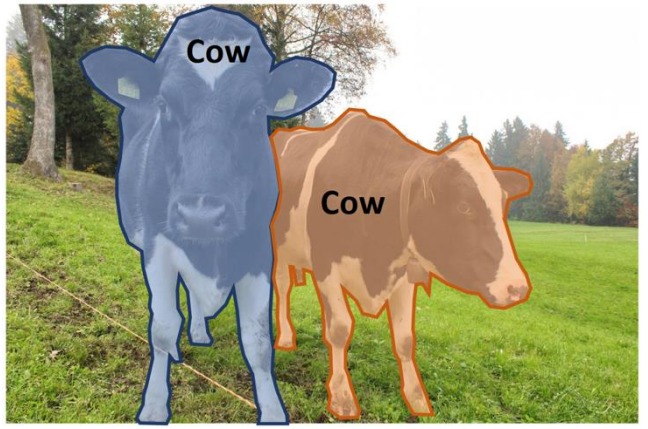
Instance Segmentation

- Semantic segmentation does not separate objects of the same class.
- Object detection finds individual objects (instances).
- Instance segmentation: segmentation at instance-level.

Semantic segmentation



Instance segmentation



Instance Segmentation

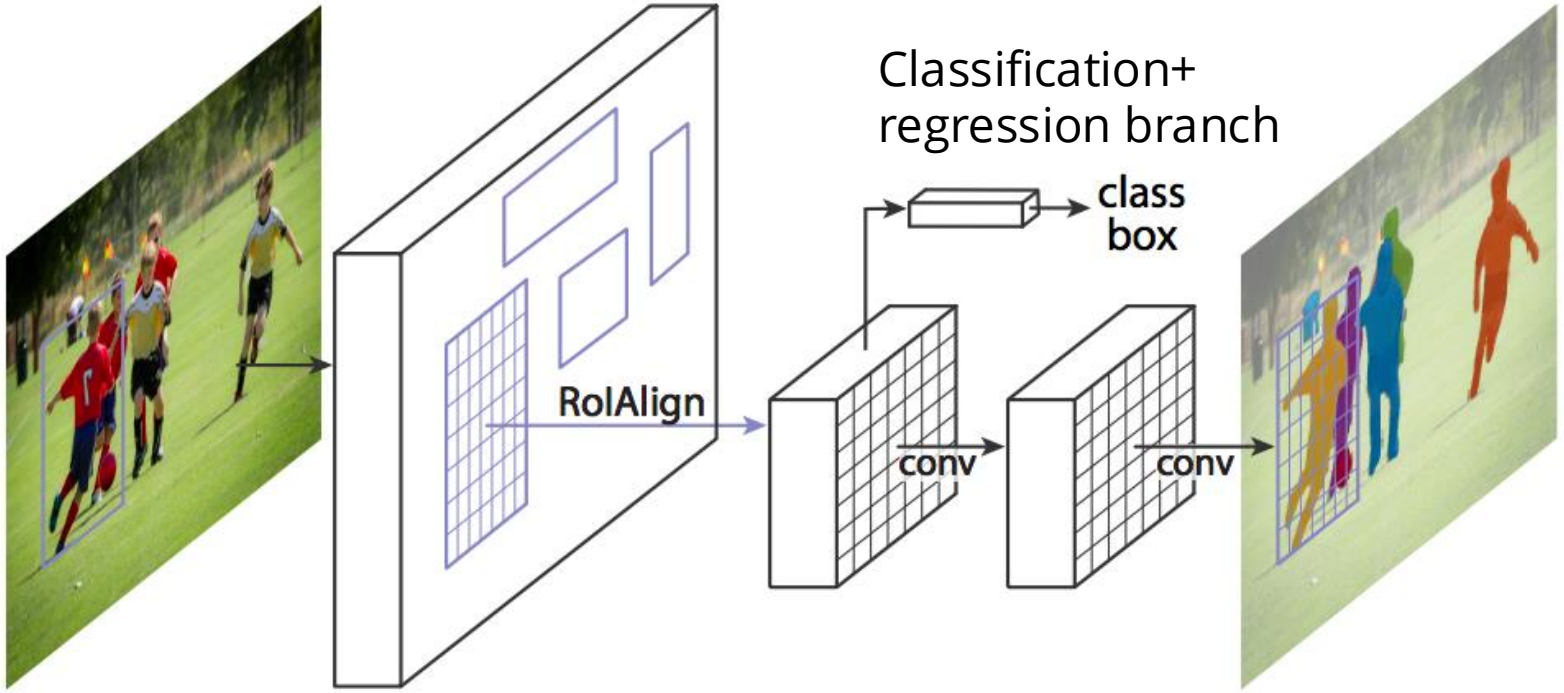
- Detect all objects in the image, and identify the pixels that belong to each object (only things, not stuff)

- Intuitive approach:
  - Detect objects
  - predict a segmentation mask for each object

- Practice: add another branch to your detector that predicts a mask for each box.

Mask R-CNN

Faster R-CNN + FCN on RoIs

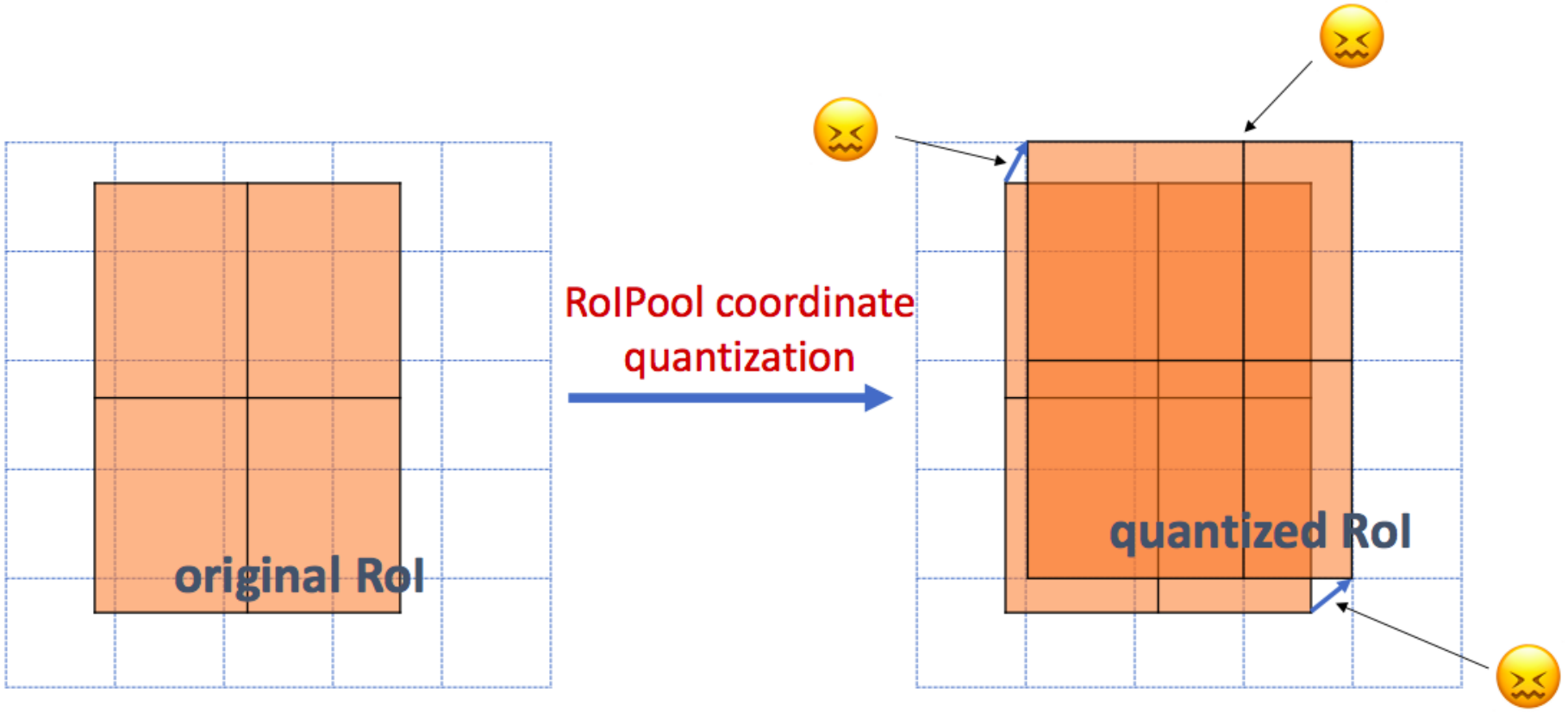


Mask branch: separately predict segmentation for each possible class

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

RoIPool

Nearest neighbor quantization results in small errors

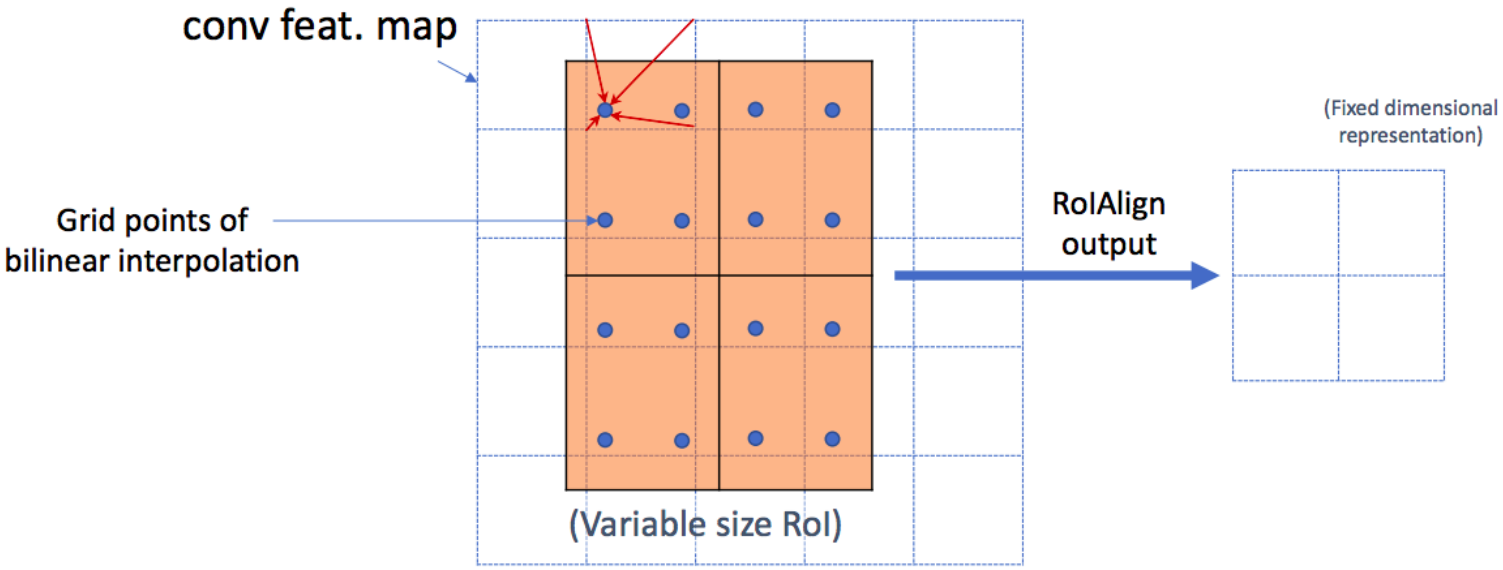


Slide: R Girshick

RoIAlign vs. RoIPool

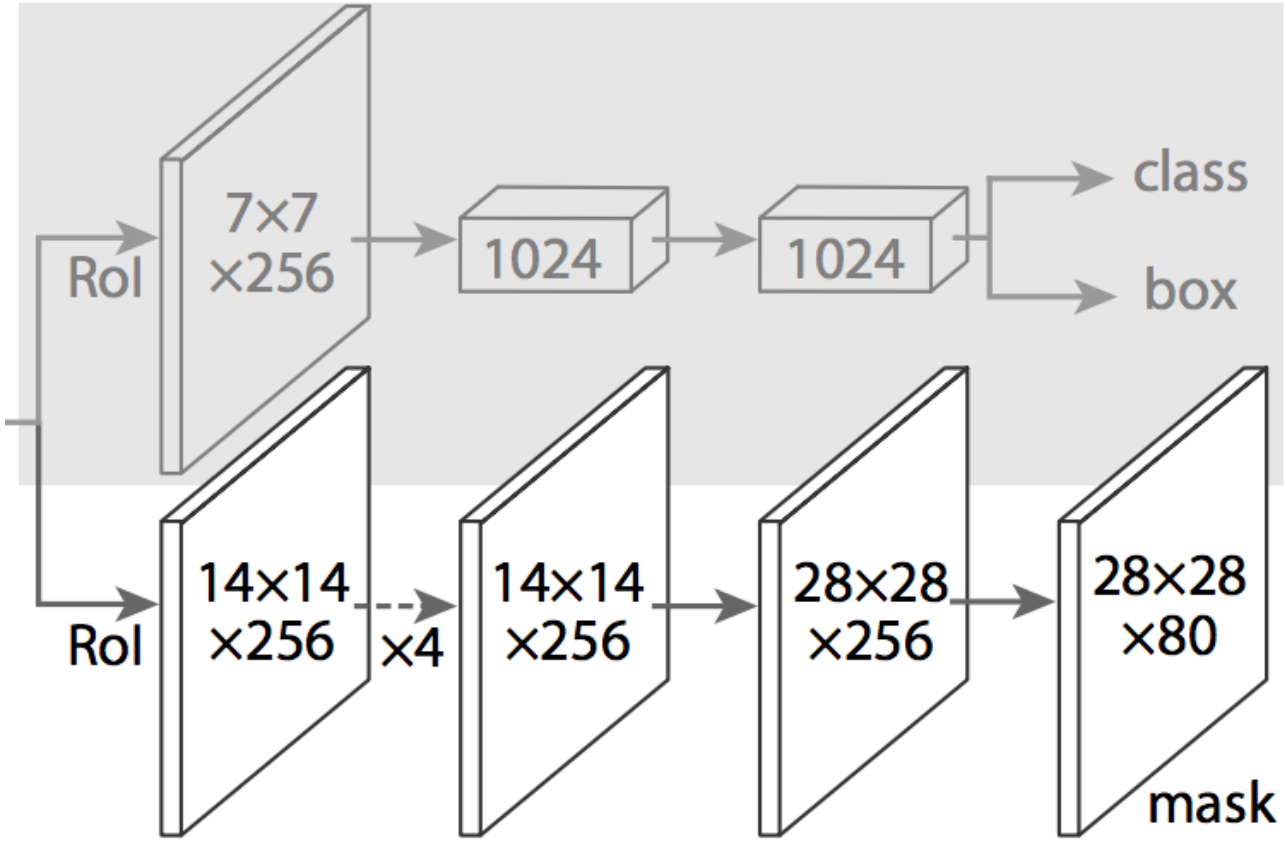
RoIPool: nearest neighbor quantization

RoIAlign: bilinear interpolation



Mask R-CNN

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



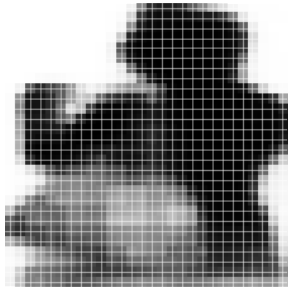
Classification head

Regression head

Separately predict binary mask (28x28) for each class with per-pixel sigmoids, use average binary cross-entropy loss (80 classes)

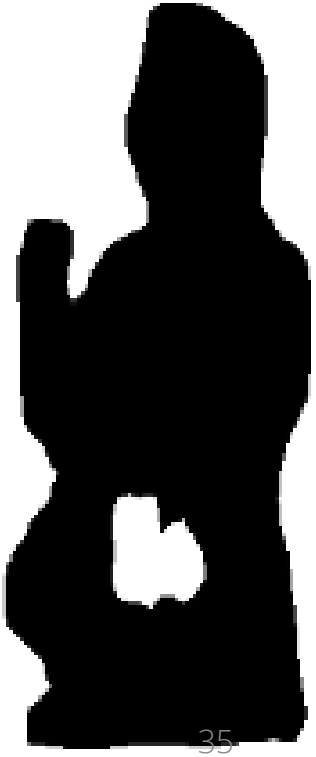
Mask R-CNN Prediction

28x28 soft prediction



Resized soft prediction

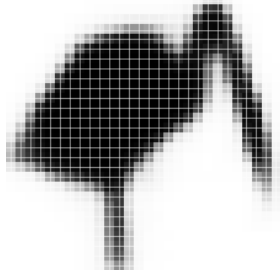
Final mask



Slide: R Girshick

Mask R-CNN Prediction

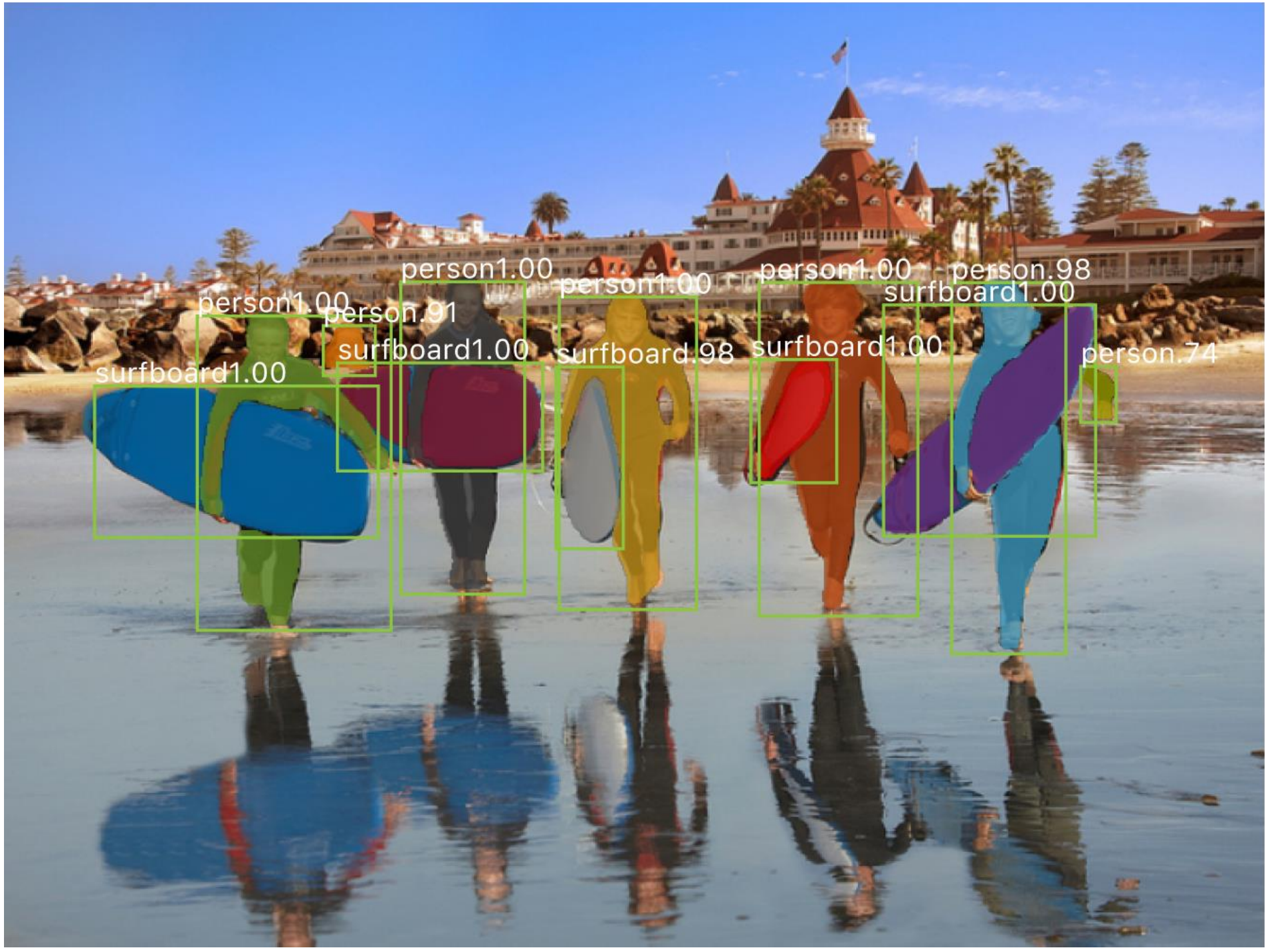
28x28 soft prediction



Resized Soft prediction

Final mask

Slide: R Girshick



Slide: R Girshick

Slide: R Girshick

Panoptic Segmentation

- Combine semantic segmentation for stuff,
- with instance segmentation for things.

- Can be solved separately.
- More efficient: share some computation between tasks.

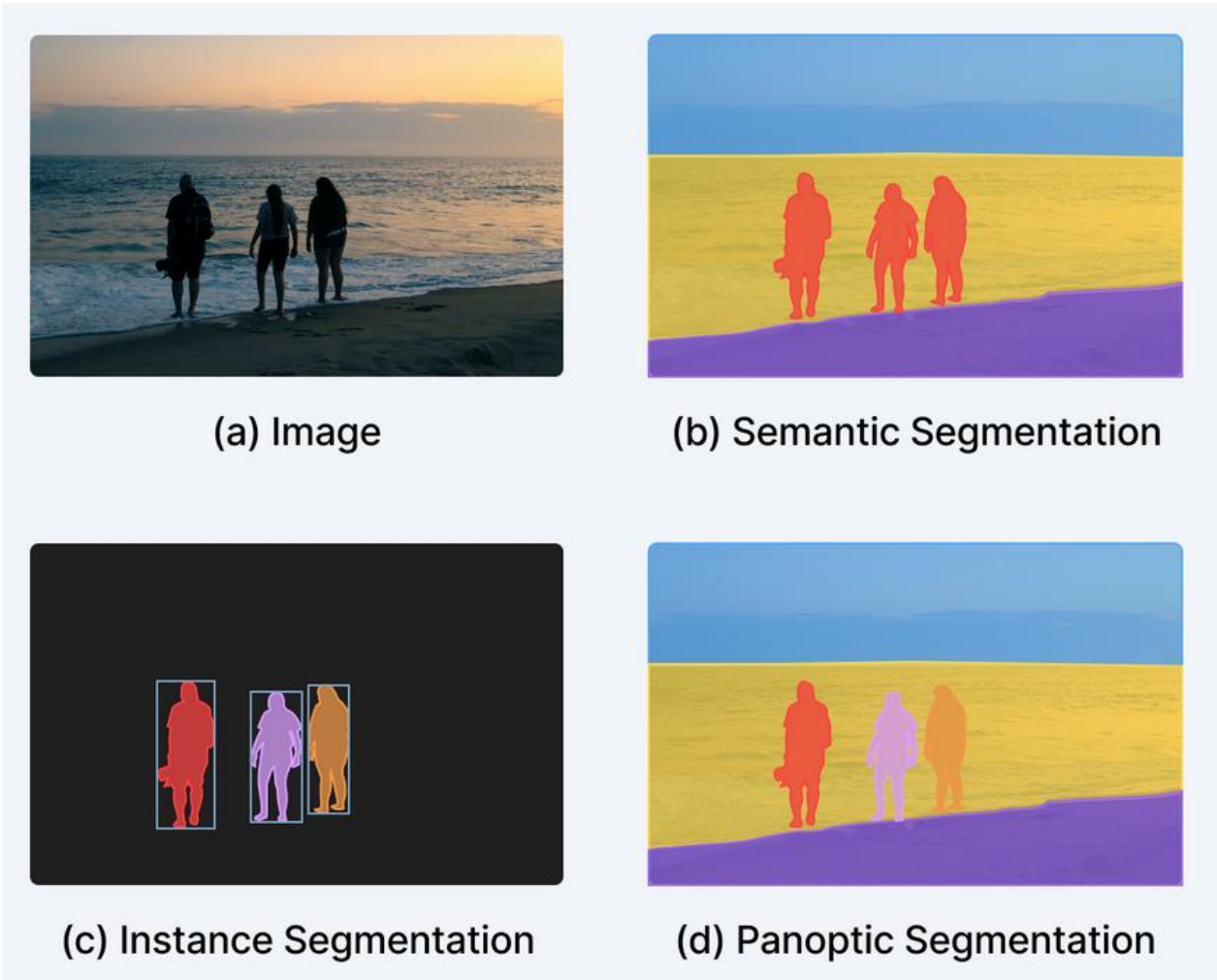
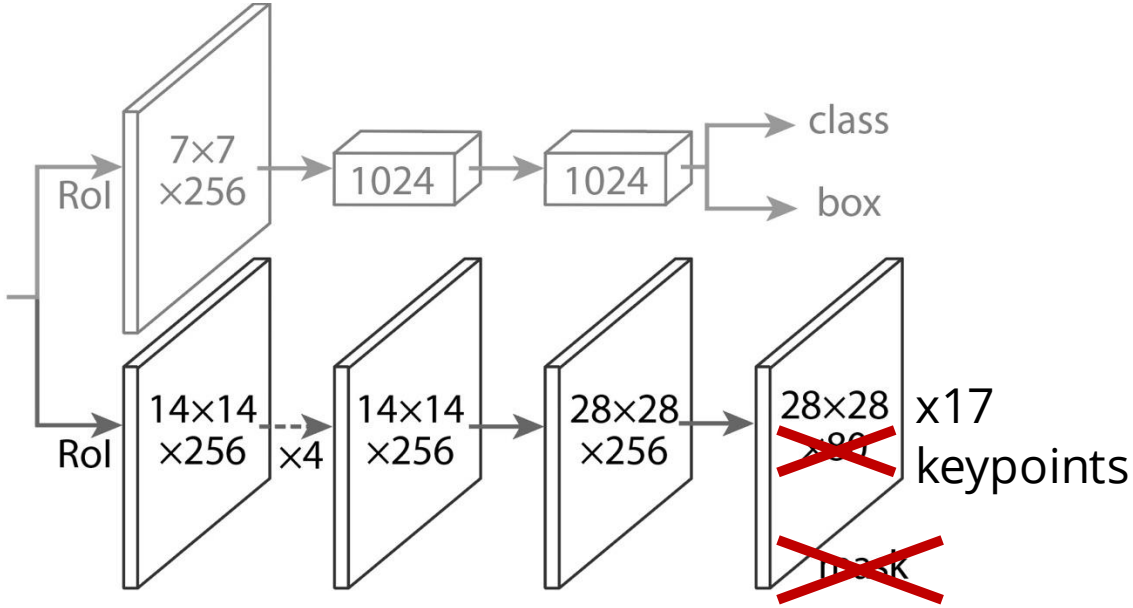


Image source

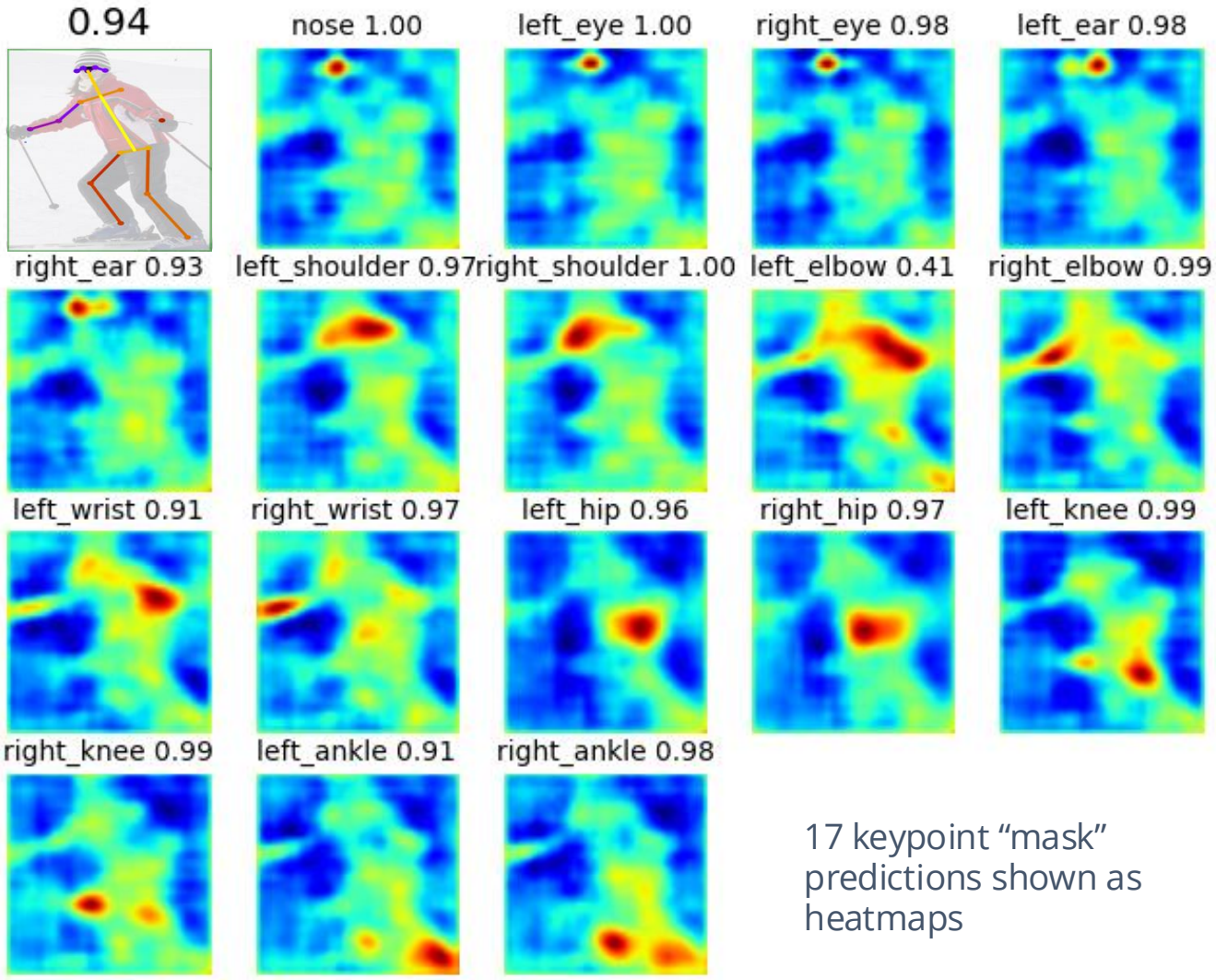
Keypoints

Instead of predicting masks, we can predict other things such as keypoints.

Keypoints here: object-specific landmarks, e.g. joints.



Human Pose



17 keypoint “mask” predictions shown as heatmaps

Keypoint Classification Loss

- Turn the GT location into a class.
- As many classes as pixels in the heatmap.
- Train with softmax cross-entropy loss.
- Output resolution limited to heatmap resolution (28x28 for Mask R-CNN)

Differentiable Keypoint Regression

- We can use softmax to formulate the keypoint regression task as a heatmap prediction problem.

- Compute softmax over heatmap: $H = \mathrm{softmax}(h)$

$$(p_x, p_y)^T = \sum_{u,v} \binom{u}{v} H(u, v)$$

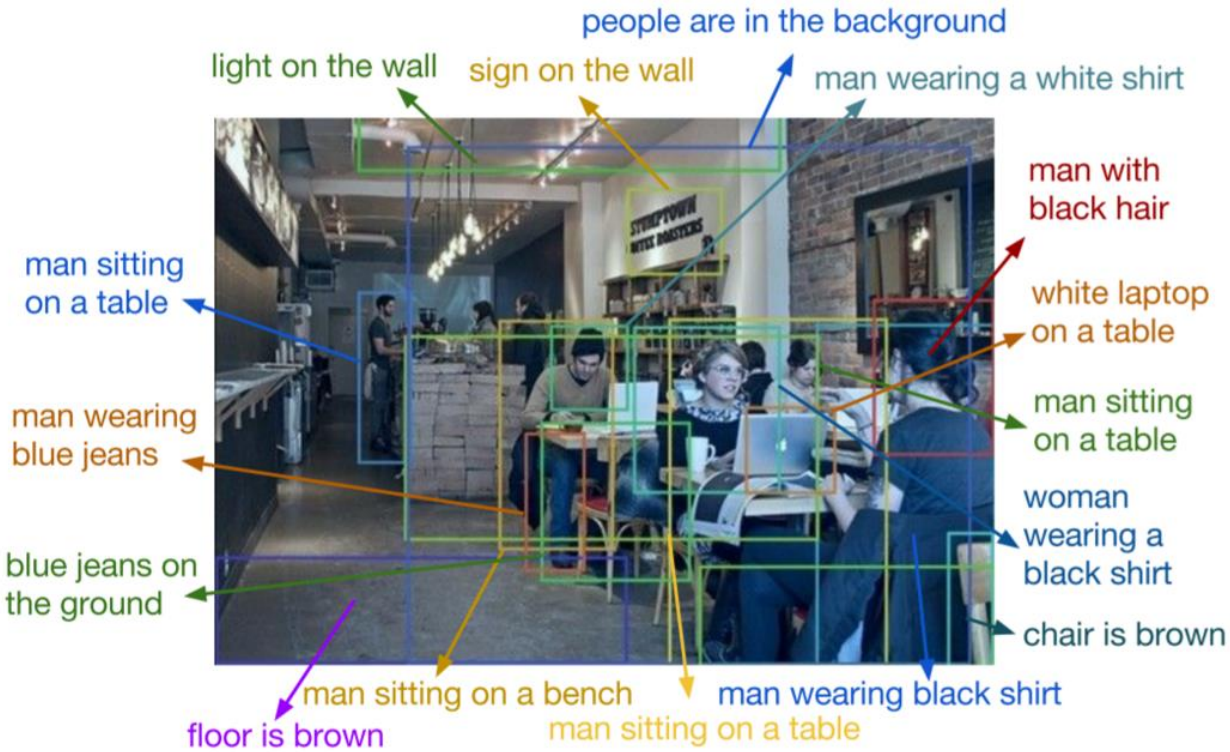
- Output location is the weighted sum of pixel locations.
- Use temperature to make it sharper.

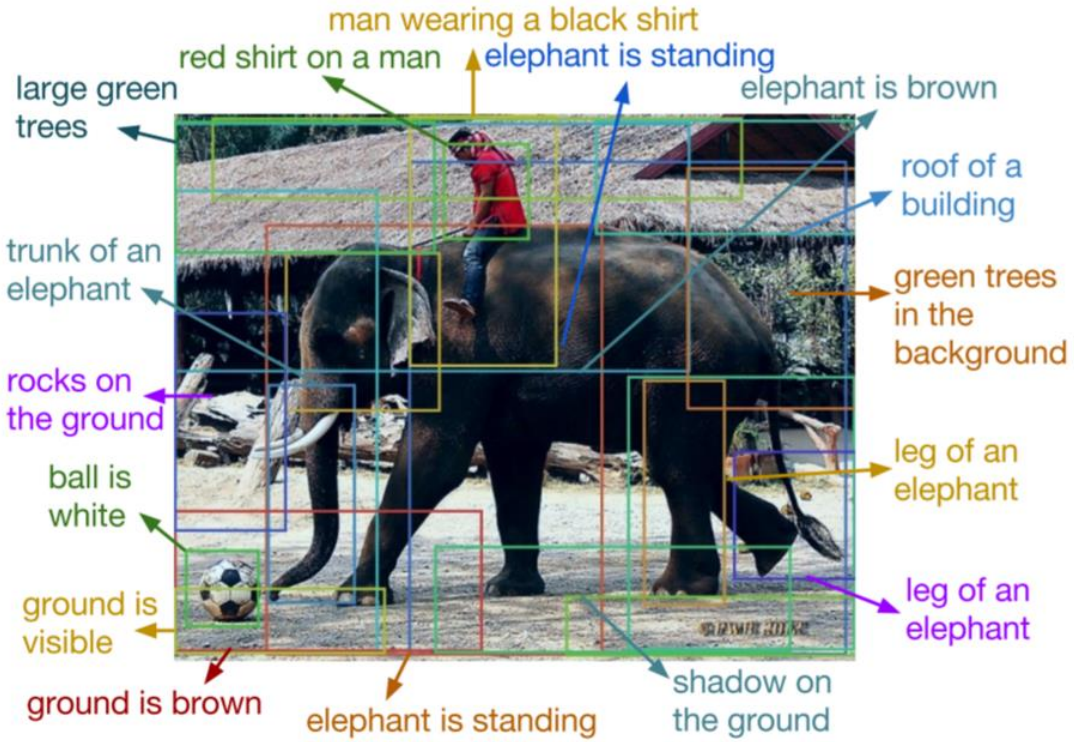
Combining Tasks

Heads can be combined to solve multiple tasks simultaneously with minimal overhead.

Dense Captioning

Add a text-output head: predict captions.

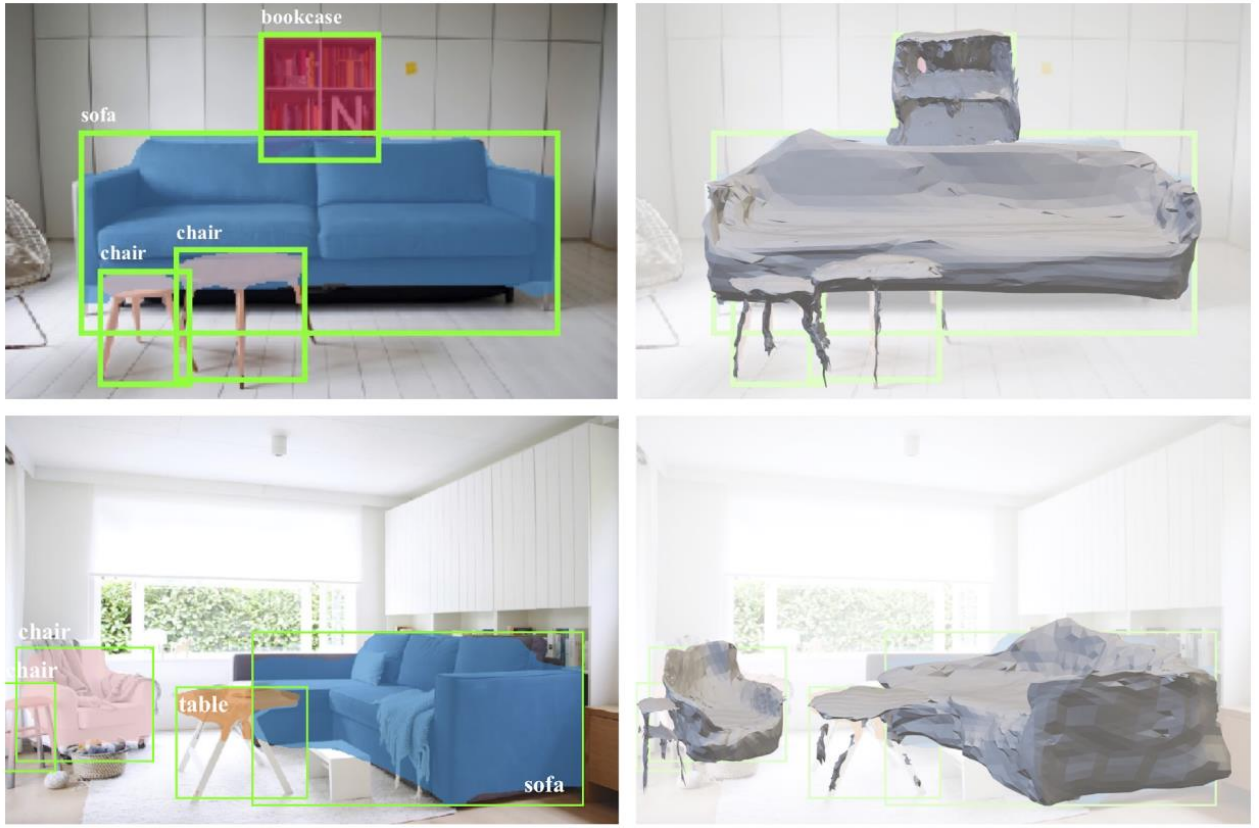




Johnson, Karpathy, and Fei-Fei, “DenseCap: Fully Convolutional Localization Networks for Dense Captioning”, CVPR 2016

3D Shape

Add a mesh prediction head.



Gkioxari, Malik, and Johnson, “Mesh R-CNN”, ICCV 2019