University of Antwerp

Artificial Neural Networks

[2500WETANN]

José Oramas

Convolutional Neural Networks

[Part 3 – Use Case Discussion]

José Oramas

University of Antwerp

Recap: Supervised Image Recognition Task

Given: an input image *x*

Do: predict a label *y*

(out of a set of class labels)

- **Data**

$$\{x, y\}_i$$

– **Model**

$$\hat{y} \approx f_\theta(x)$$

– **Loss**

$$L(\theta) = \sum_{i=1}^{N} l(f_\theta(x_i), \hat{y}_i)$$

– **Optimization**

$$\theta^* = arg\ min_\theta\ L(\theta)$$

Let’s Consider the Object Detection Task

Given: an input image *x*

How to provide a localization output?

Localized Predictions

[Use Case: Object Detection]

Task-1: Object Detection

Given: an input image *x*

Do: predict a label *y* *(out of a set of class labels)* & *location* *(bounding box)*

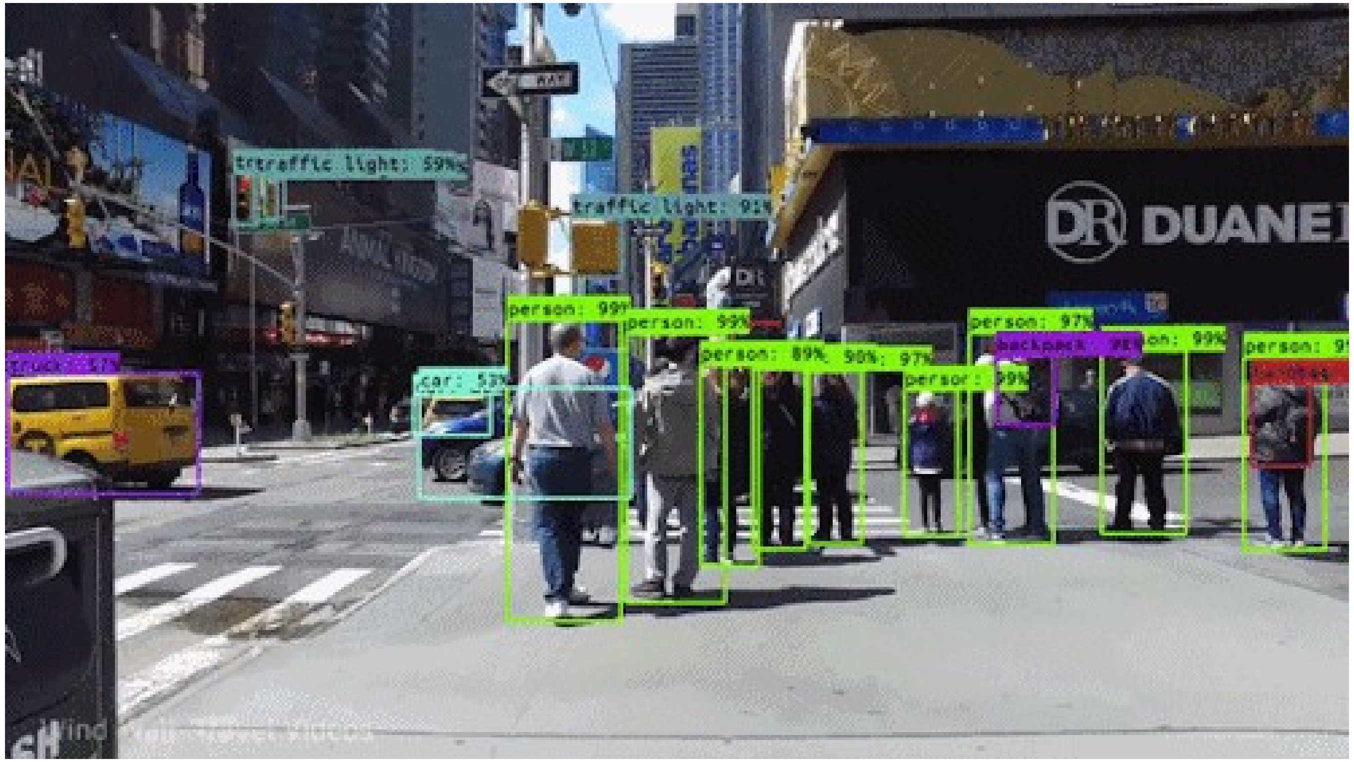
Text
– Role labelling

Audio
– Speech-command detection
– profanity detection

Task-1: Object Detection

Given: an input image *x*

Do: predict a label *y* *(out of a set of class labels)* & *location* *(bounding box)*



Task-1: Object Detection

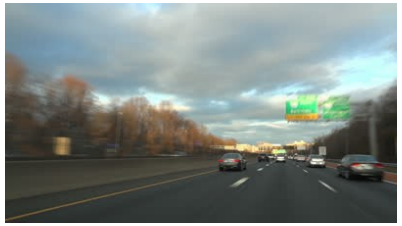
Given: an input image *x*

Do: predict a label *y* *(out of a set of class labels)* & *location* *(bounding box)*

Challenges

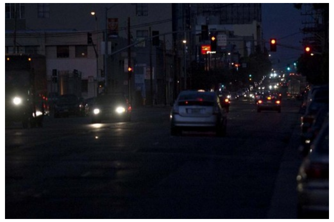
Changes in Viewpoint

Objects at low scale



High Occlusions

Changes in Illumination



University of Antwerp

Task-1: Object Detection

How was it classically done?

- Scan the input image
- Evaluate the scanned regions with a classifier

Sliding Window

Pre-trained Model
(SVM, RF, Boosting, etc.)

Prediction
(class label + location)

Task-1: Object Detection

How was it classically done?

- Scan the input image
- Evaluate the scanned regions with a classifier

Sliding Window

→ **Pre-trained Model** (SVM, RF, Boosting, etc.) → **Prediction** (class label + location)

Task-1: Object Detection

How was it classically done?

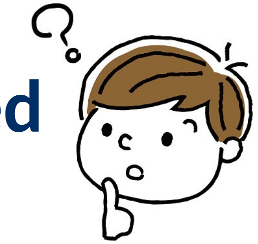
- Scan the input image
- Evaluate the scanned regions with a classifier

Sliding Window

Pre-trained Model
(SVM, RF, Boosting, etc.)

Prediction
(class label + location)

ConvNet pre-trained on ImageNet?



Task-1: Object Detection

Does it addresses the challenges?

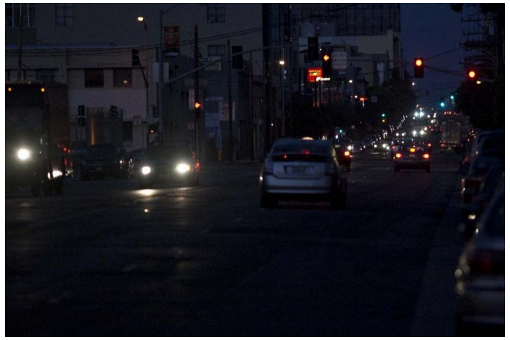
[How? What would be needed?]

High Occlusions

Changes in Viewpoint

Objects at low scale

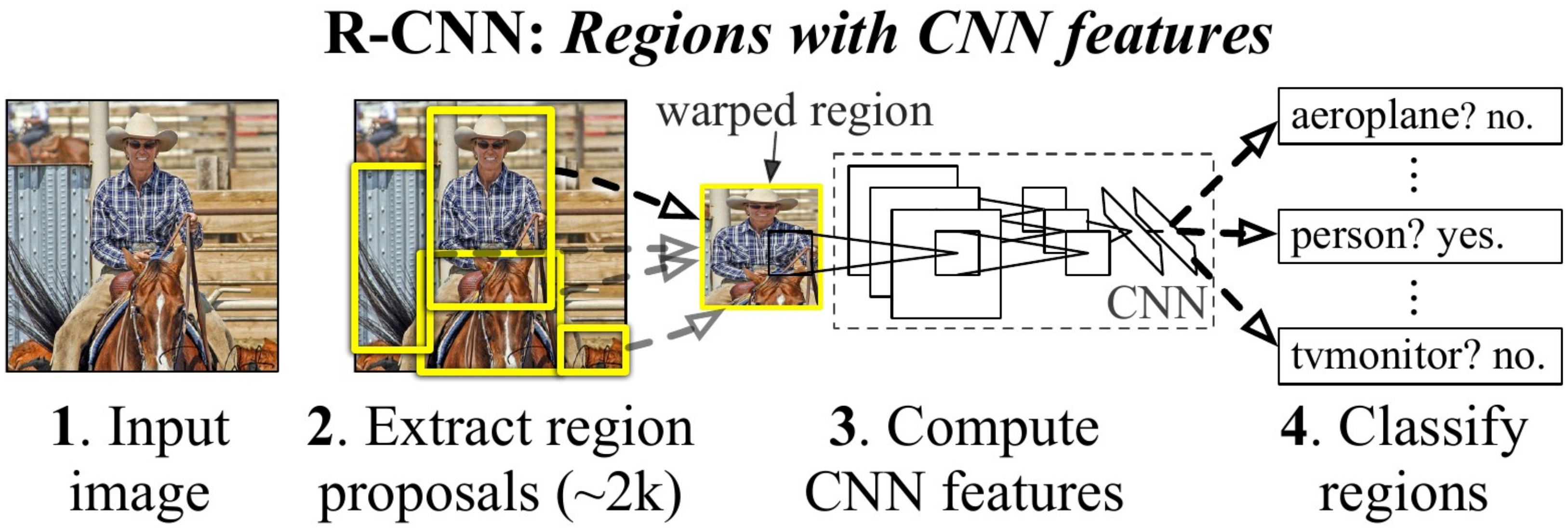
Changes in Illumination



Task-1: Object Detection: R-CNN

How was it done? (at least the first time)

- Generate object proposals
- Evaluate the proposals with a classifier

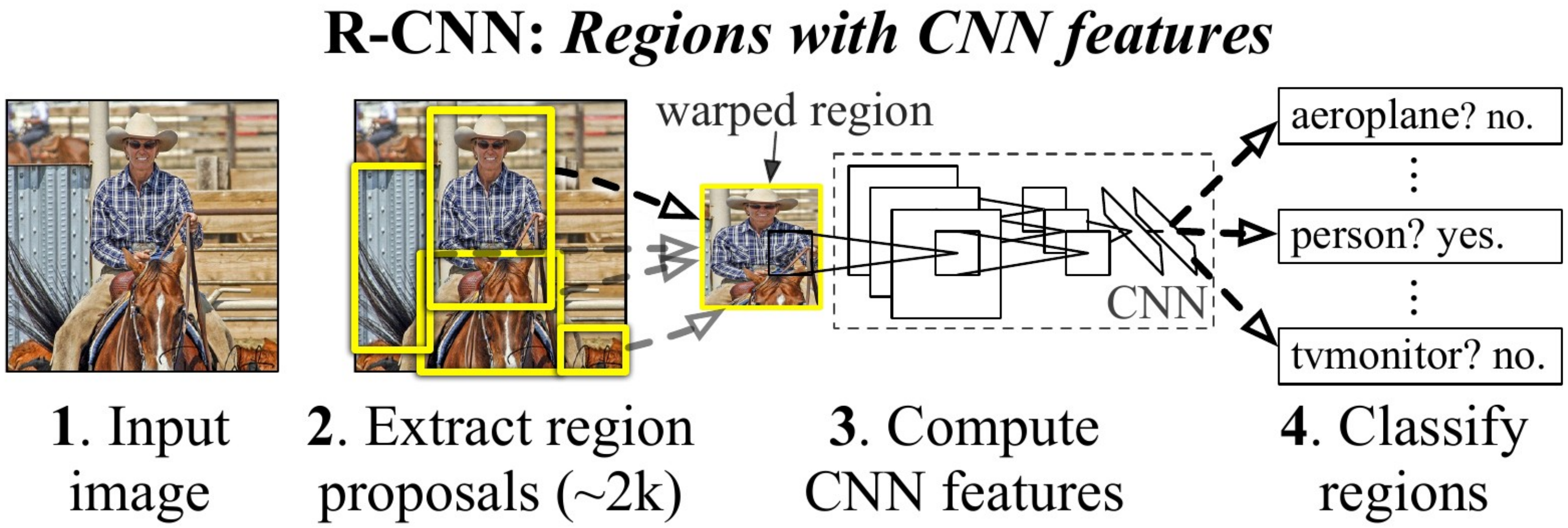


[Girshick et al., 2014]

Task-1: Object Detection: R-CNN

How was it done? (at least the first time)

- Generate object proposals
- Evaluate the proposals with a classifier



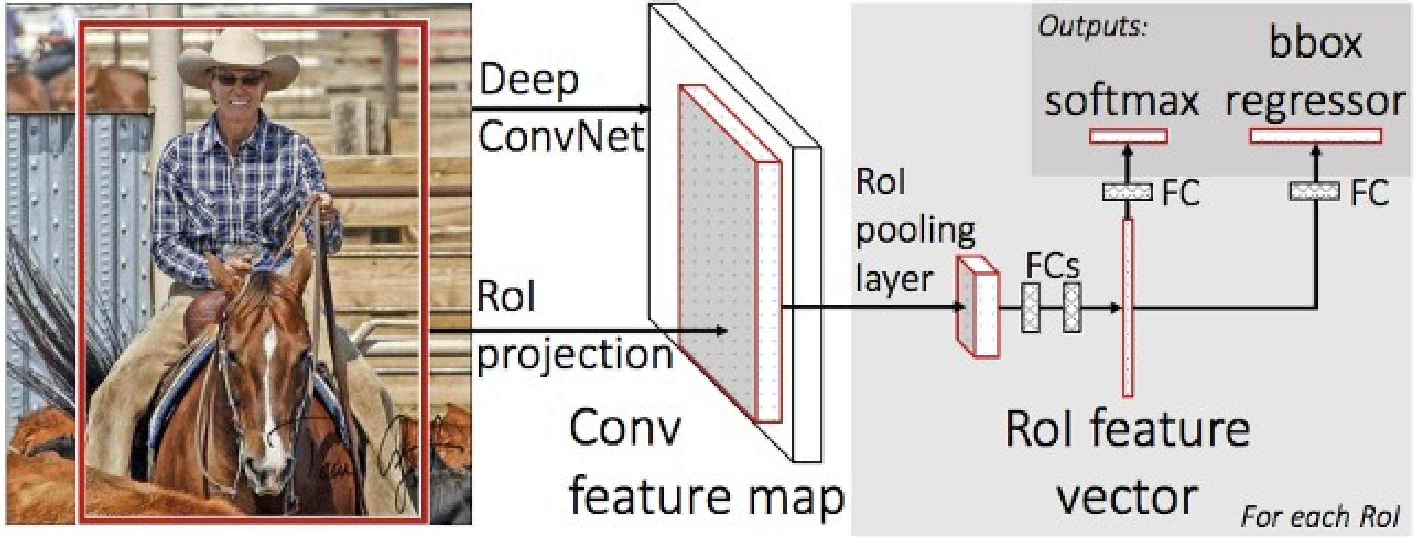
[Girshick et al., 2014]

Q: Does that solves all the problems?

Task-1: Object Detection: Fast R-CNN

How was it done?

- Extract the Region Proposals from a Feature Map
- Evaluate the proposal with the FC layers (classifier)

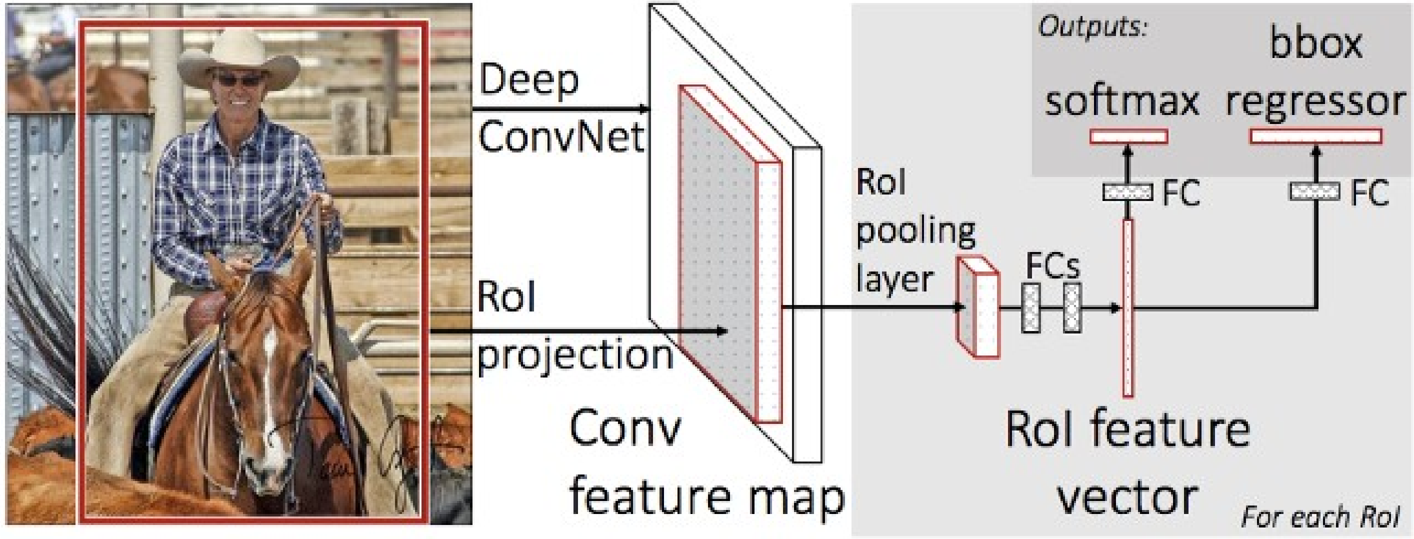


[Girshick et al., 2015]

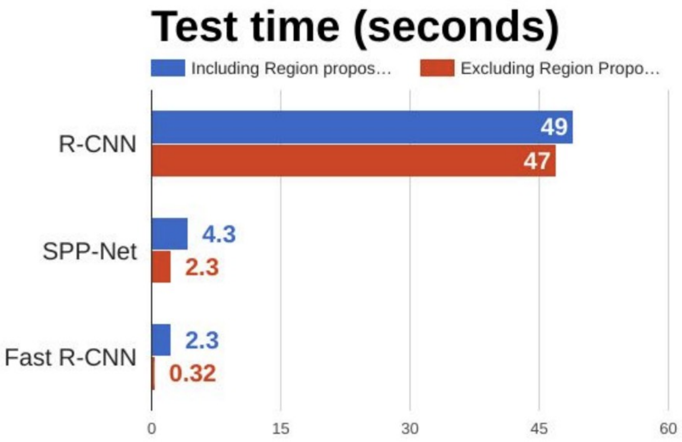
Task-1: Object Detection: Fast R-CNN

How was it done?

- Extract the Region Proposals from a Feature Map
- Evaluate the proposal with the FC layers (classifier)



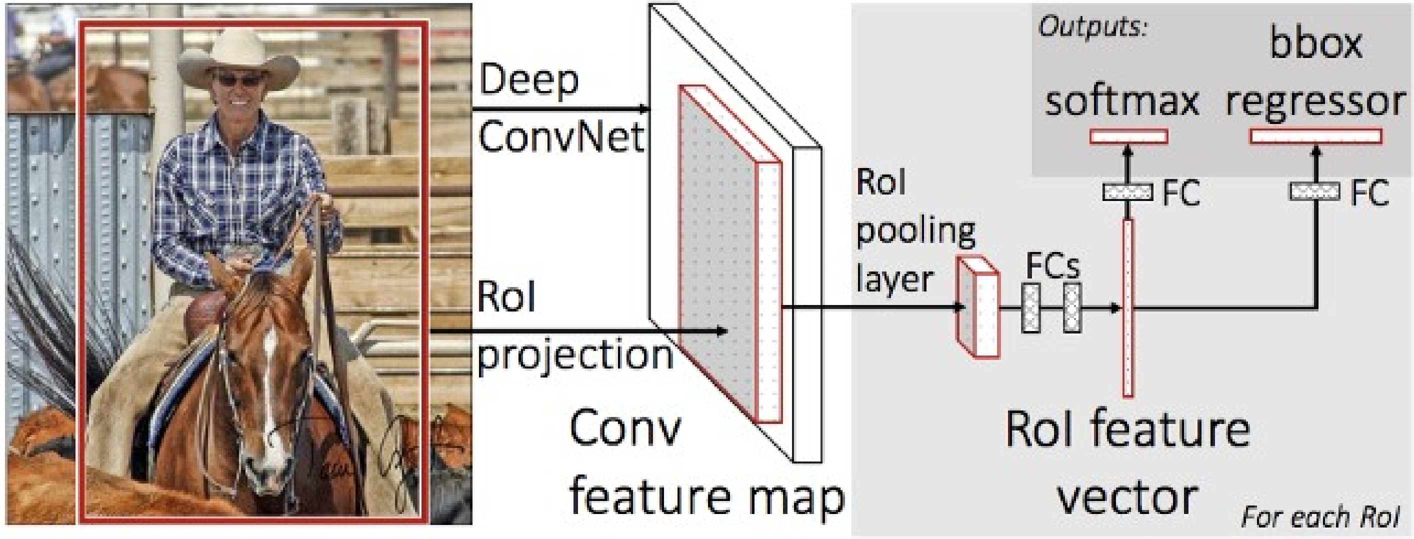
[Girshick et al., 2015]



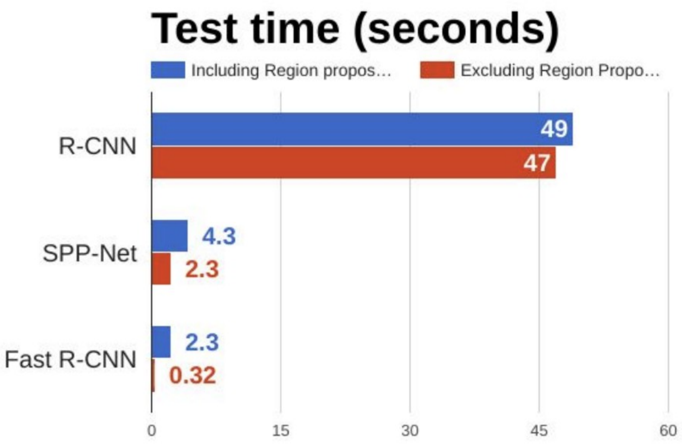
Task-1: Object Detection: Fast R-CNN

How was it done?

- Extract the Region Proposals from a Feature Map
- Evaluate the proposal with the FC layers (classifier)



[Girshick et al., 2015]

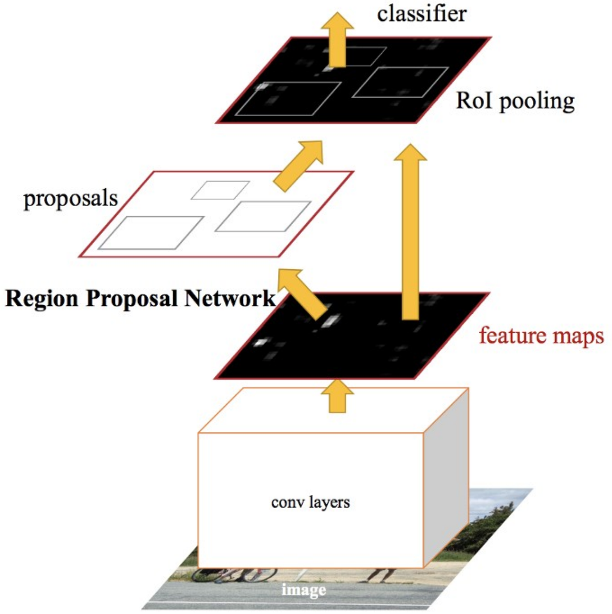


Q: Does that solves all the problems?

Task-1: Object Detection: Faster R-CNN

How was it done?

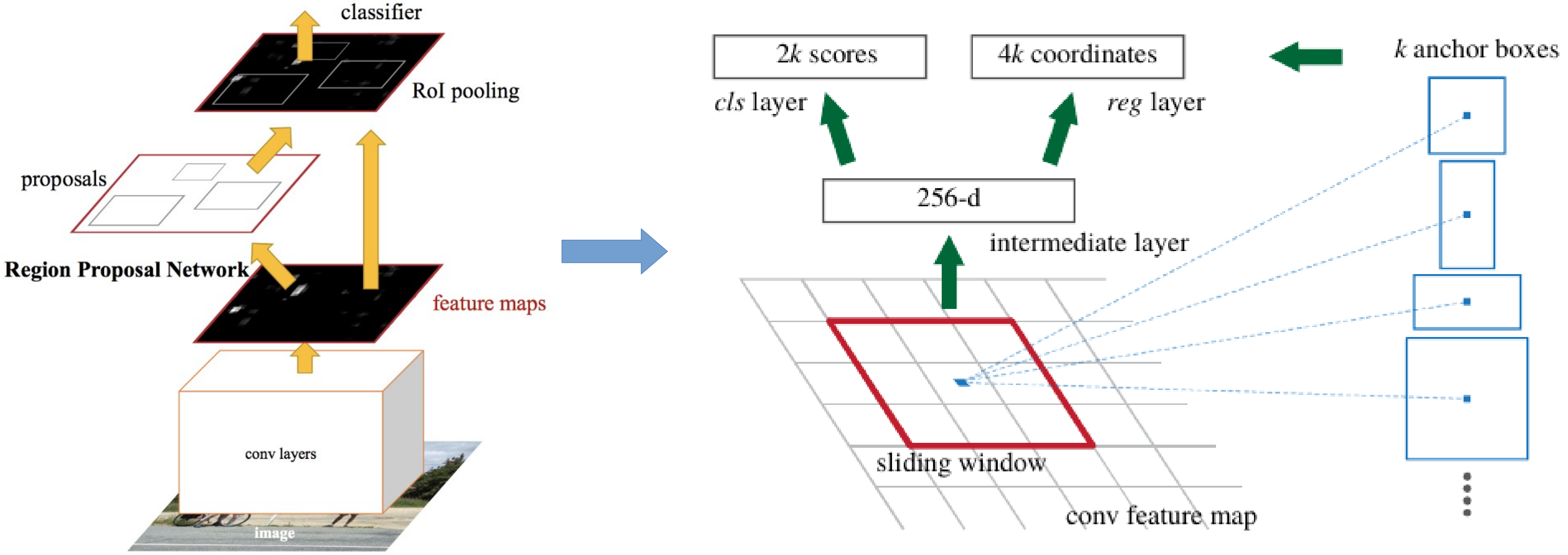
- Use a network (RPN) to detect proposals from the feature map
- Evaluate the proposal with the FC layers



Task-1: Object Detection: Faster R-CNN

How was it done?

- Use a network (RPN) to detect proposals from the feature map
- Evaluate the proposal with the FC layers

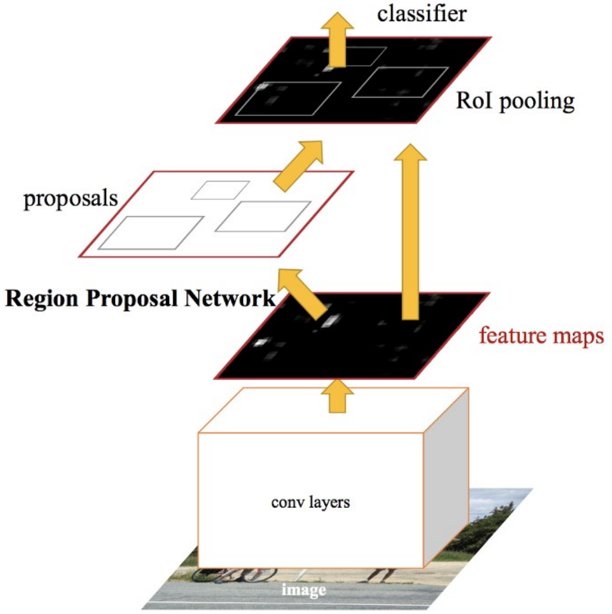


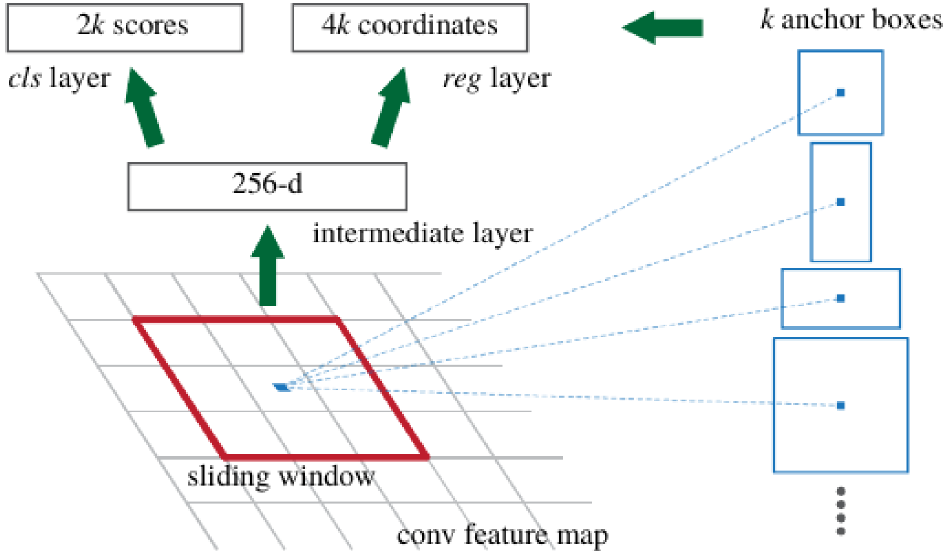
[Ren et al., 2015]

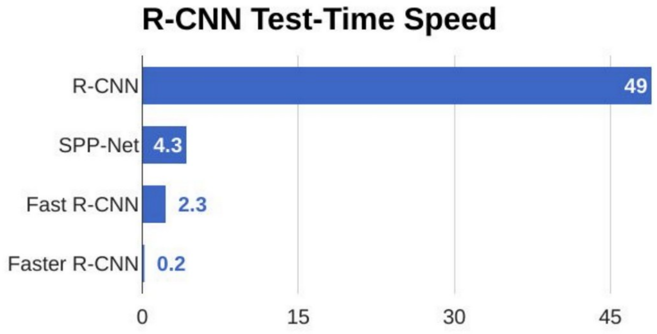
Task-1: Object Detection: Faster R-CNN

How was it done?

- Use a network (RPN) to detect proposals from the feature map
- Evaluate the proposal with the FC layers





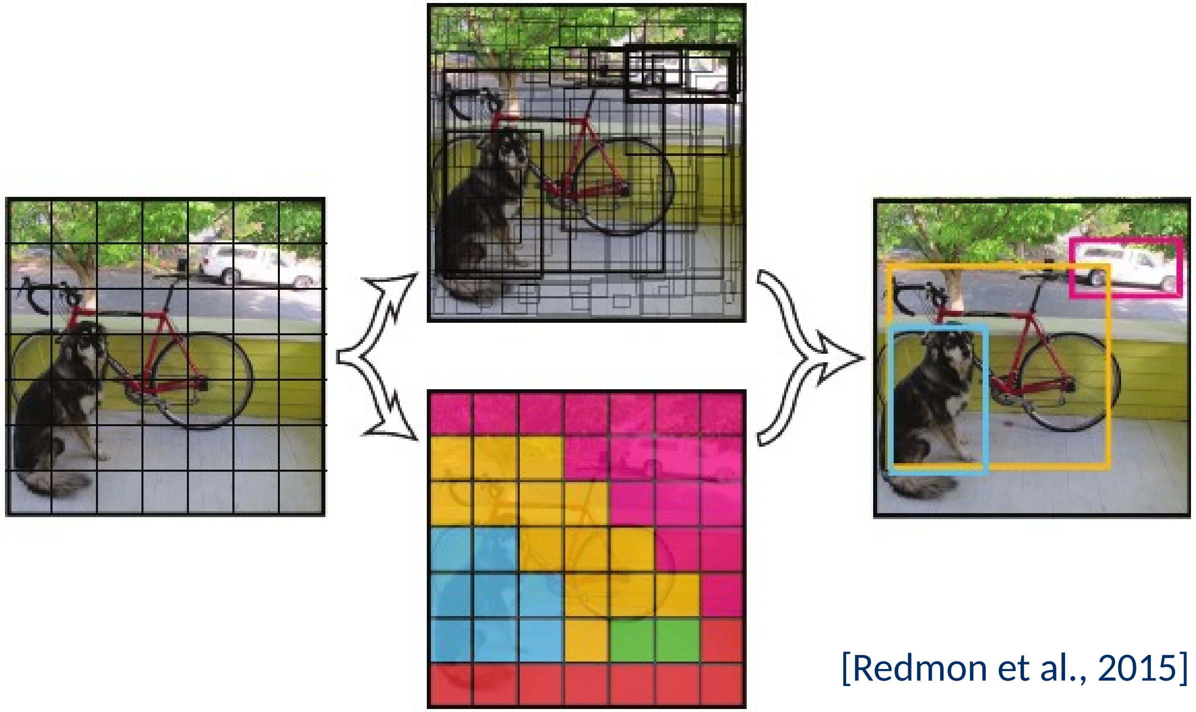


[Ren et al., 2015]

Task-1: Object Detection: YOLO

How was it done?

- Integrating the two stages
  - Divide an image into a grid
  - Predict: bbox with confidence and class probabilities

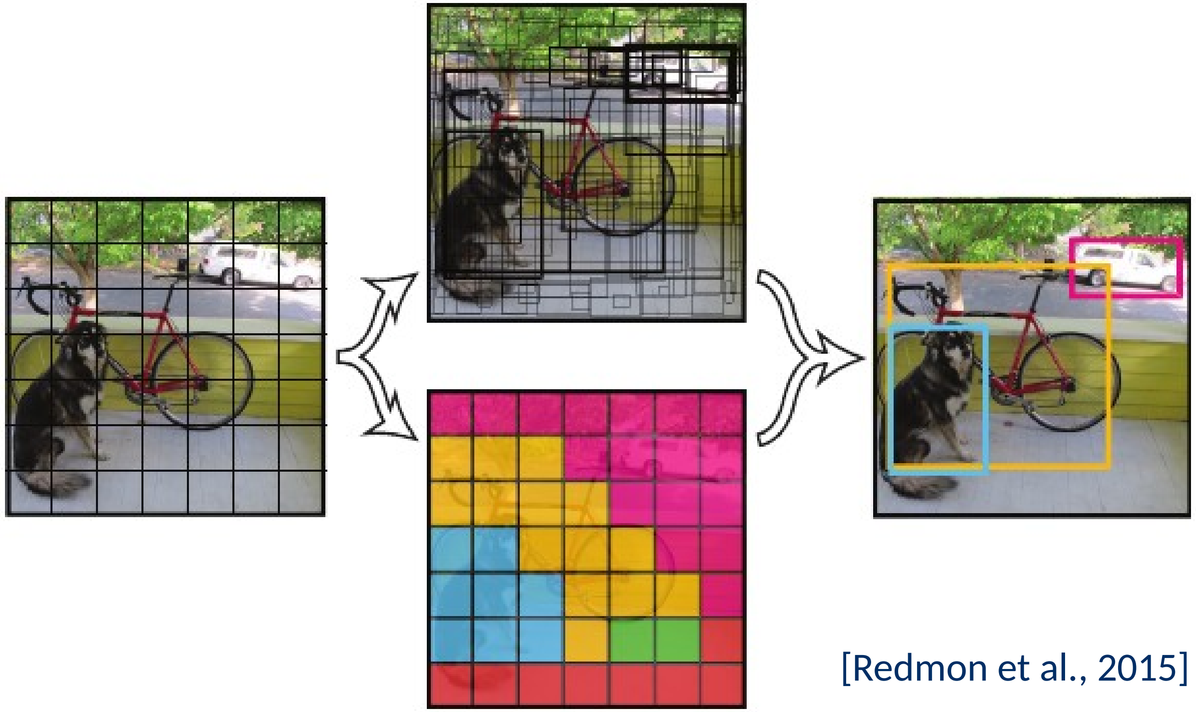


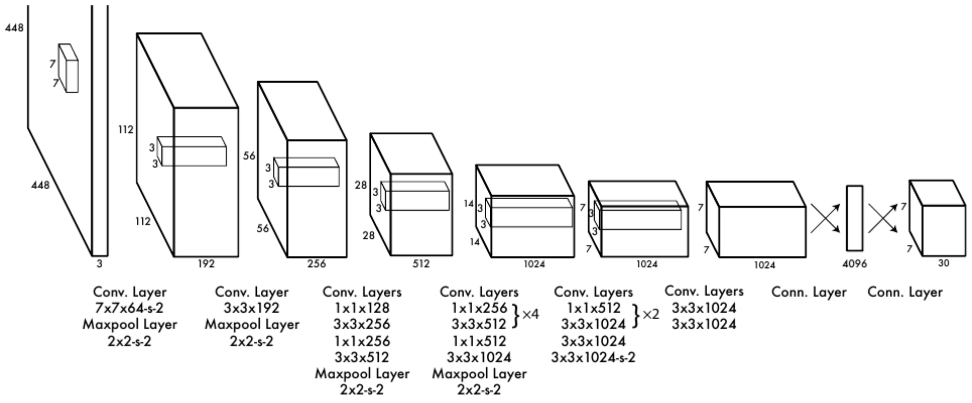
[Redmon et al., 2015]

Task-1: Object Detection: YOLO

How was it done?

- Integrating the two stages
  - Divide an image into a grid
  - Predict: bbox with confidence and class probabilities



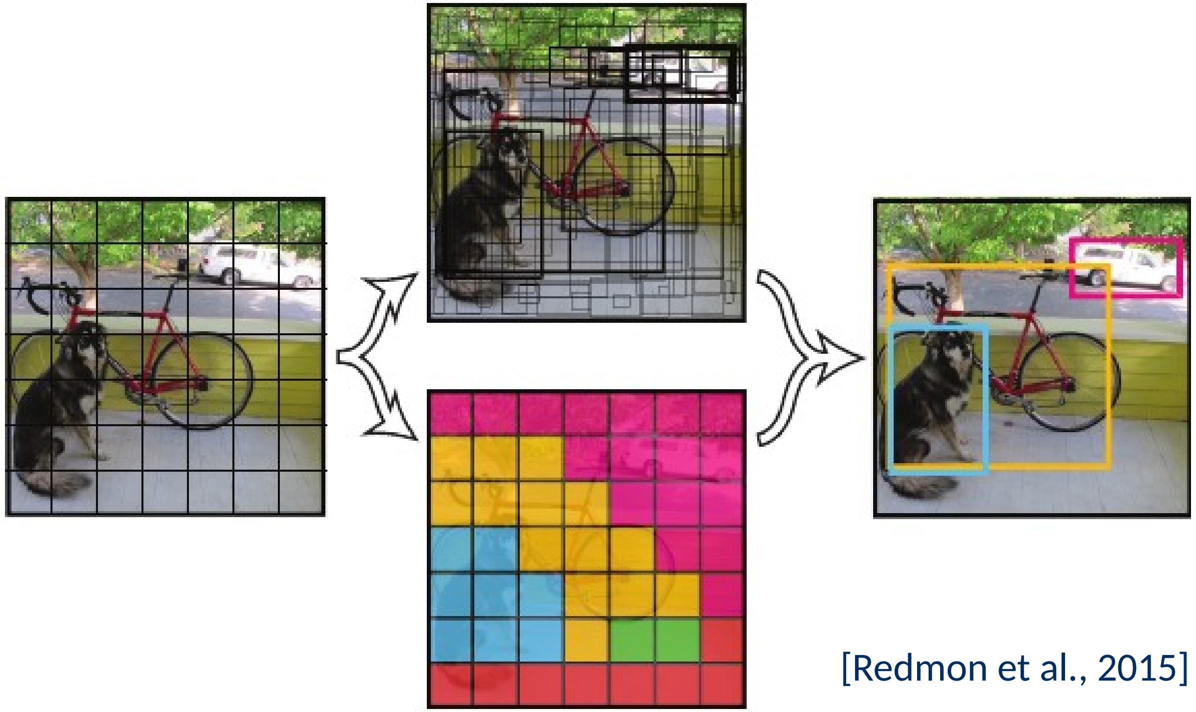


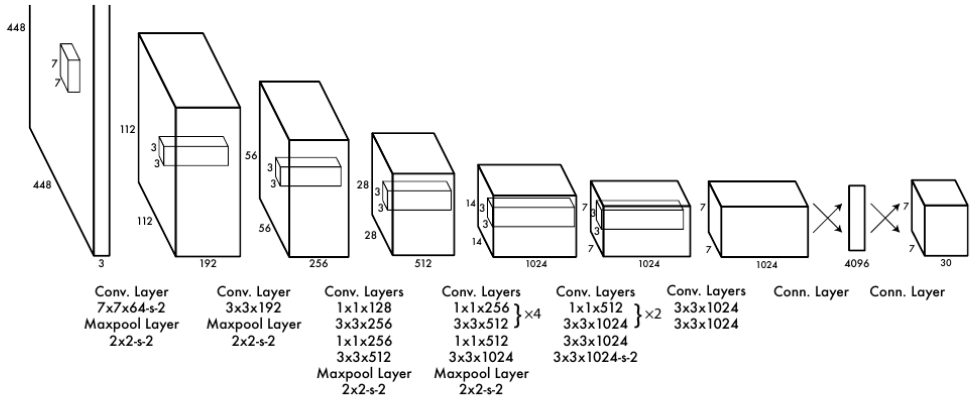
[Redmon et al., 2015]

Task-1: Object Detection: YOLO

How was it done?

- Integrating the two stages
  - Divide an image into a grid
  - Predict: bbox with confidence and class probabilities





[Redmon et al., 2015]

Limitations
– Only 2 boxes per cell
– Only 1 class per cell

University of Antwerp


Object Detection
starring
YOLOv3

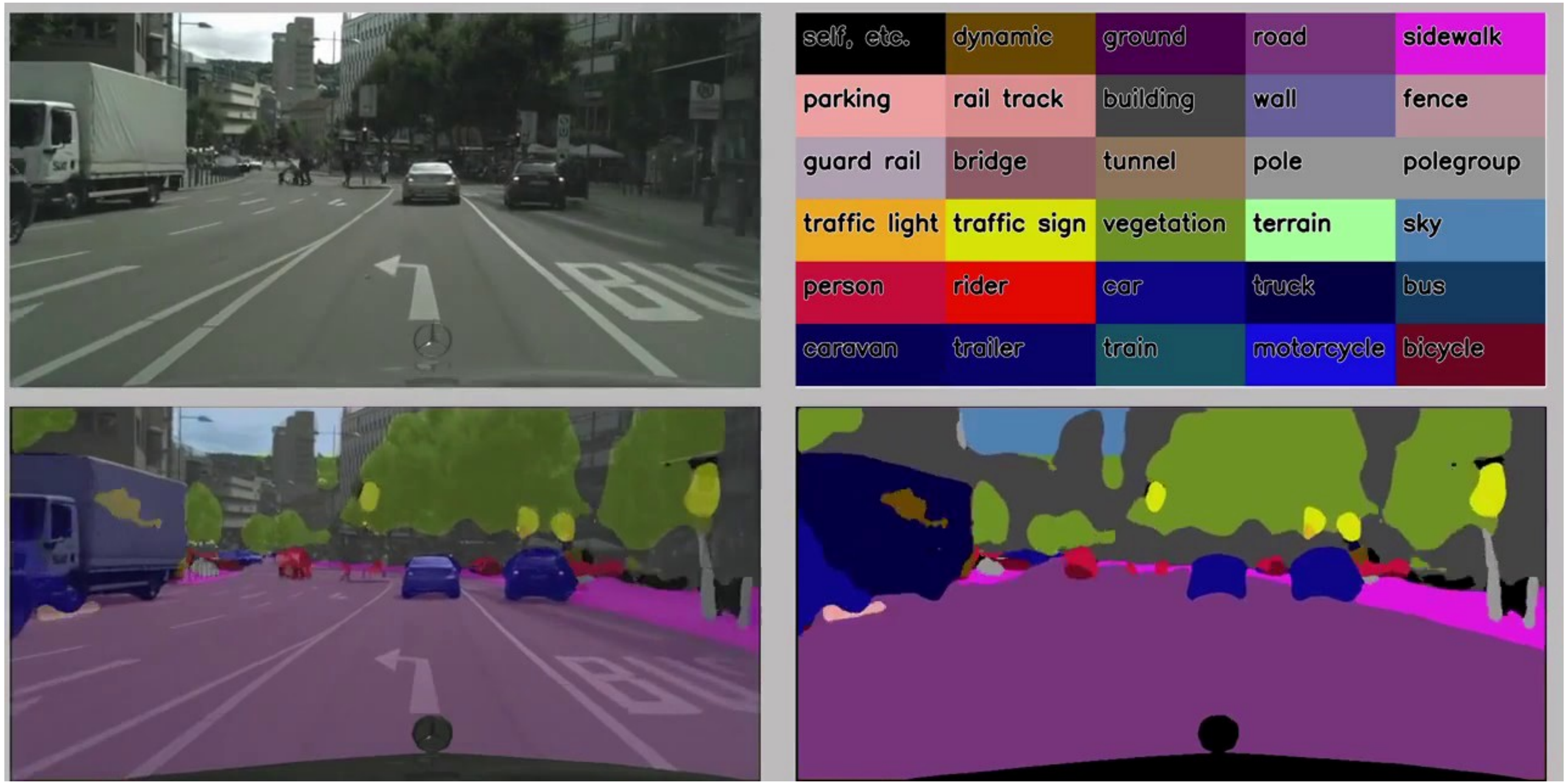
Dense Predictions

[Use Case: Semantic Segmentation]

Task-2: Image Segmentation

Given: an input image *x*

Do: predict a label *y for every pixel in the image*

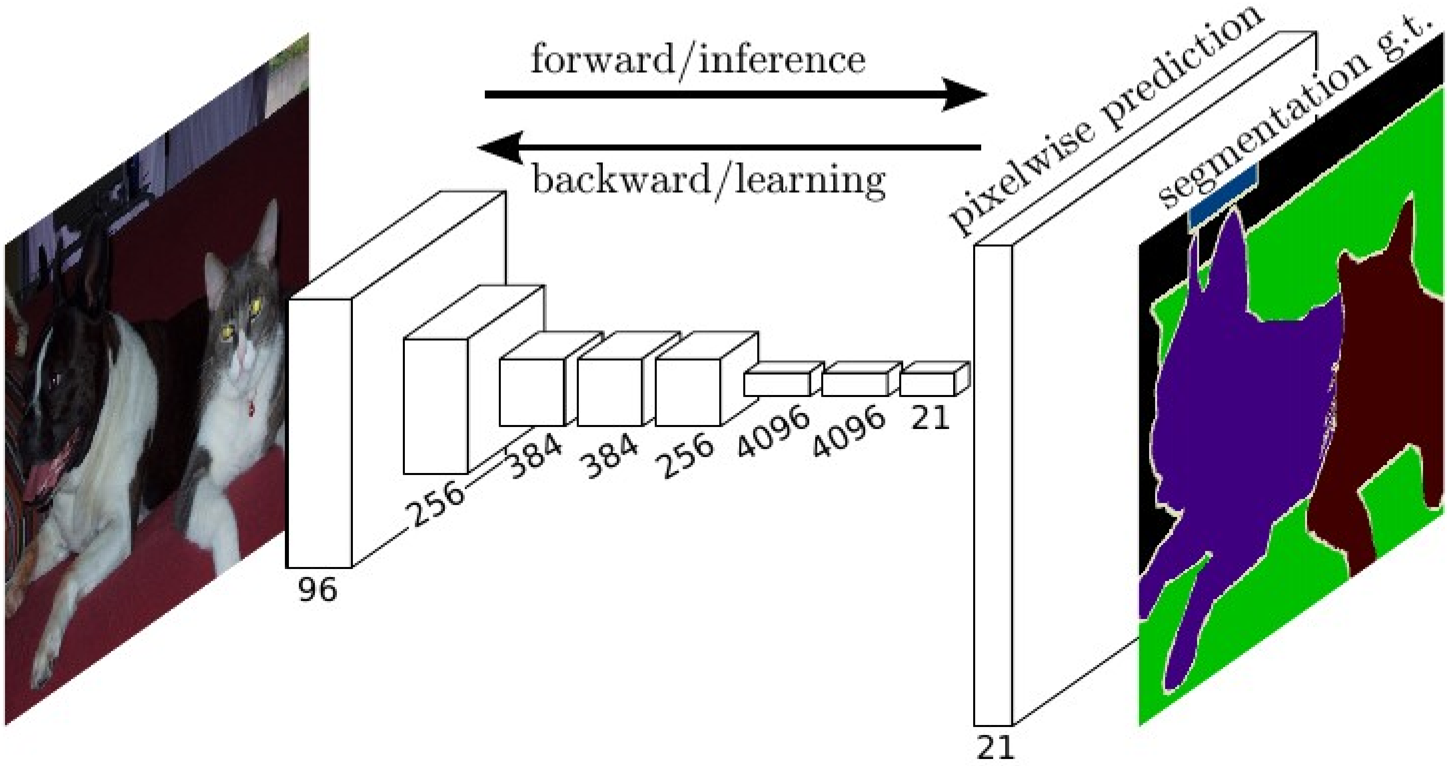


Text
– Part of Speech tagging

Audio
– Source labelling

Task-2: Semantic Segmentation: FCNs

How was it done? (at least the first time)

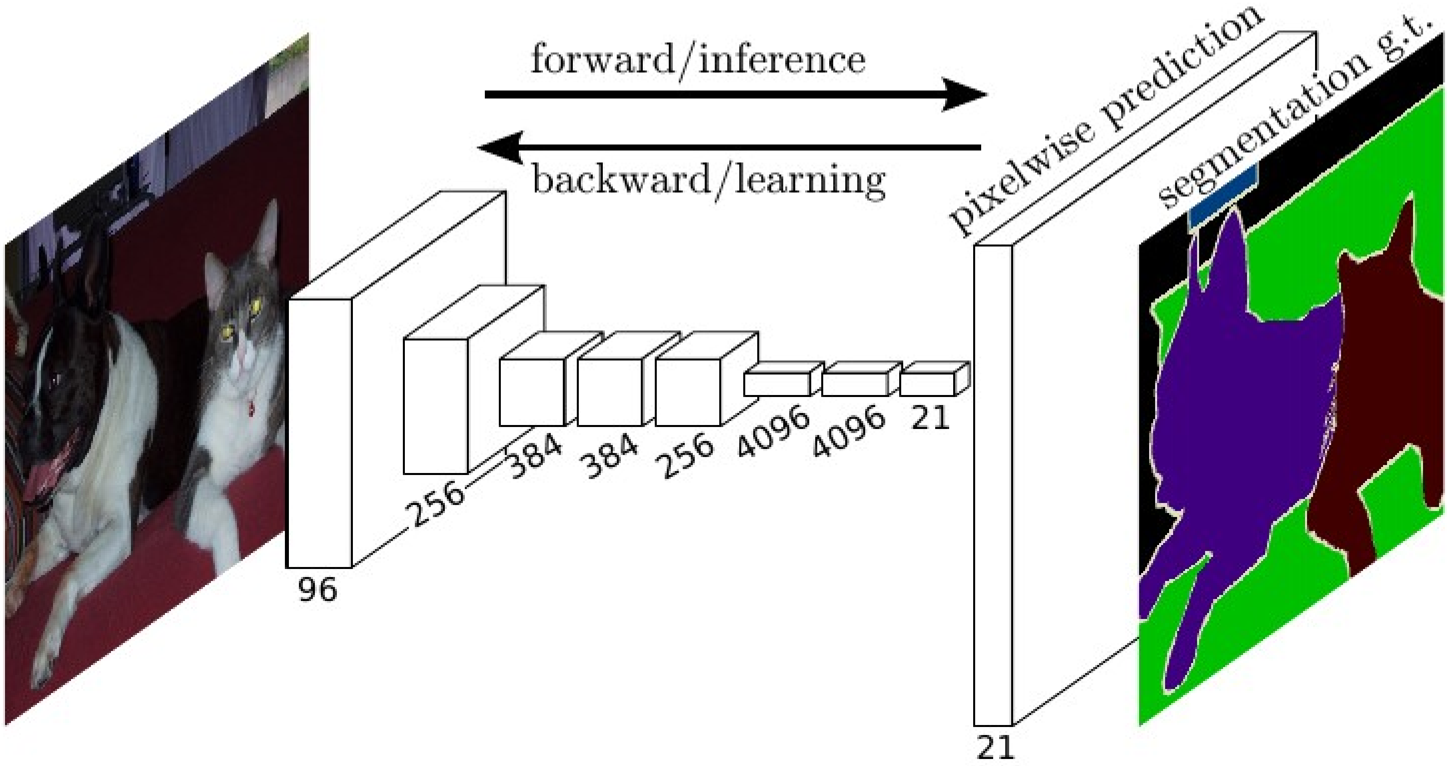


[Long et al., 2015]

Task-2: Semantic Segmentation: FCNs

How was it done? (at least the first time)

- Formulate FC-layers as Conv-layers → deep-filter / FCN
- Upsampling: backwards strided convolution

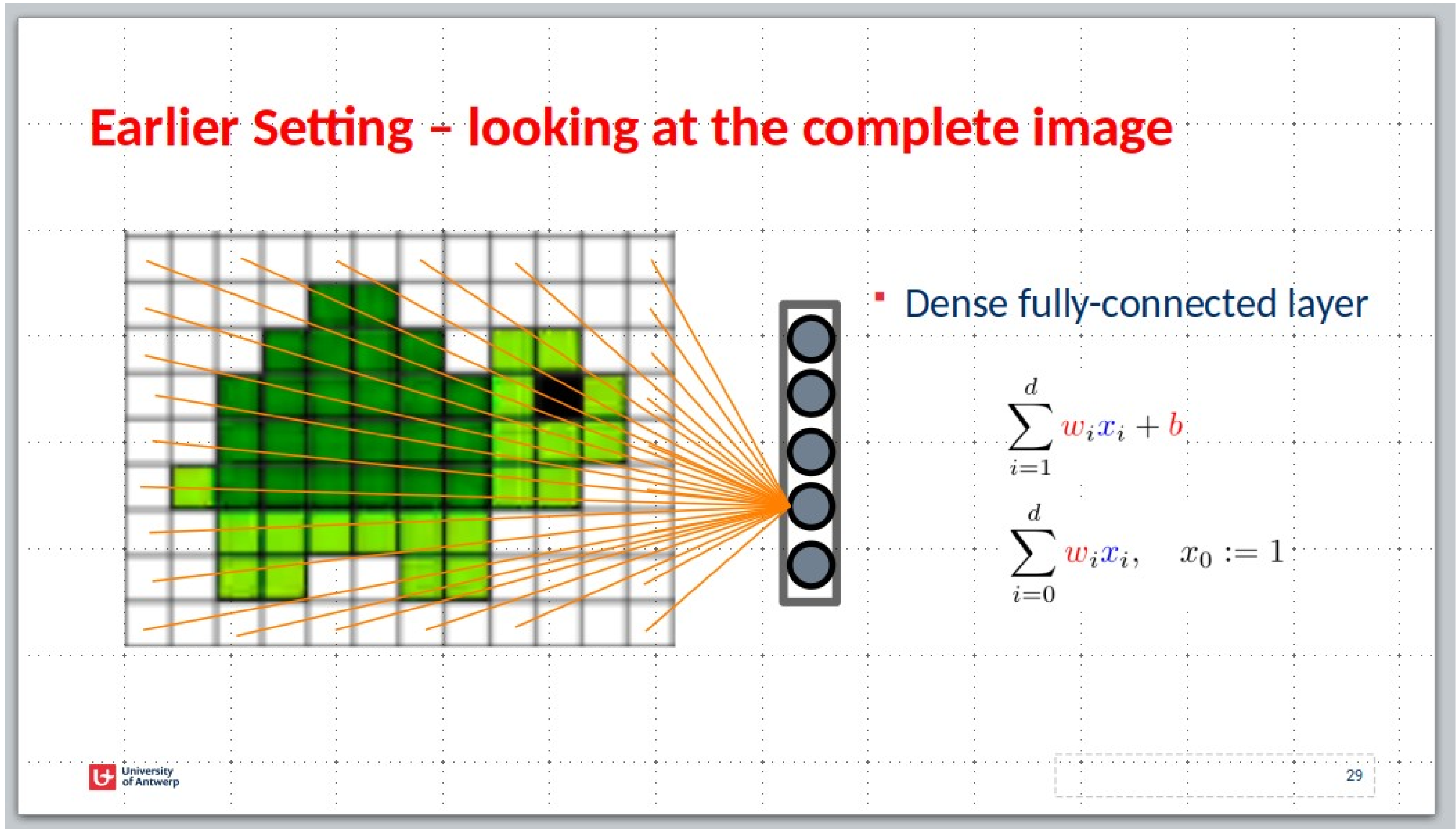


[Long et al., 2015]

Task-2: Semantic Segmentation: FCNs

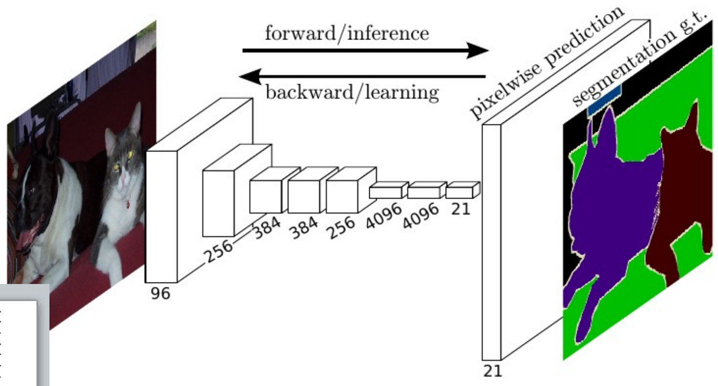
How was it done? (at least the first time)

- Formulate FC-layers as Conv-layers → deep-filter / FCN



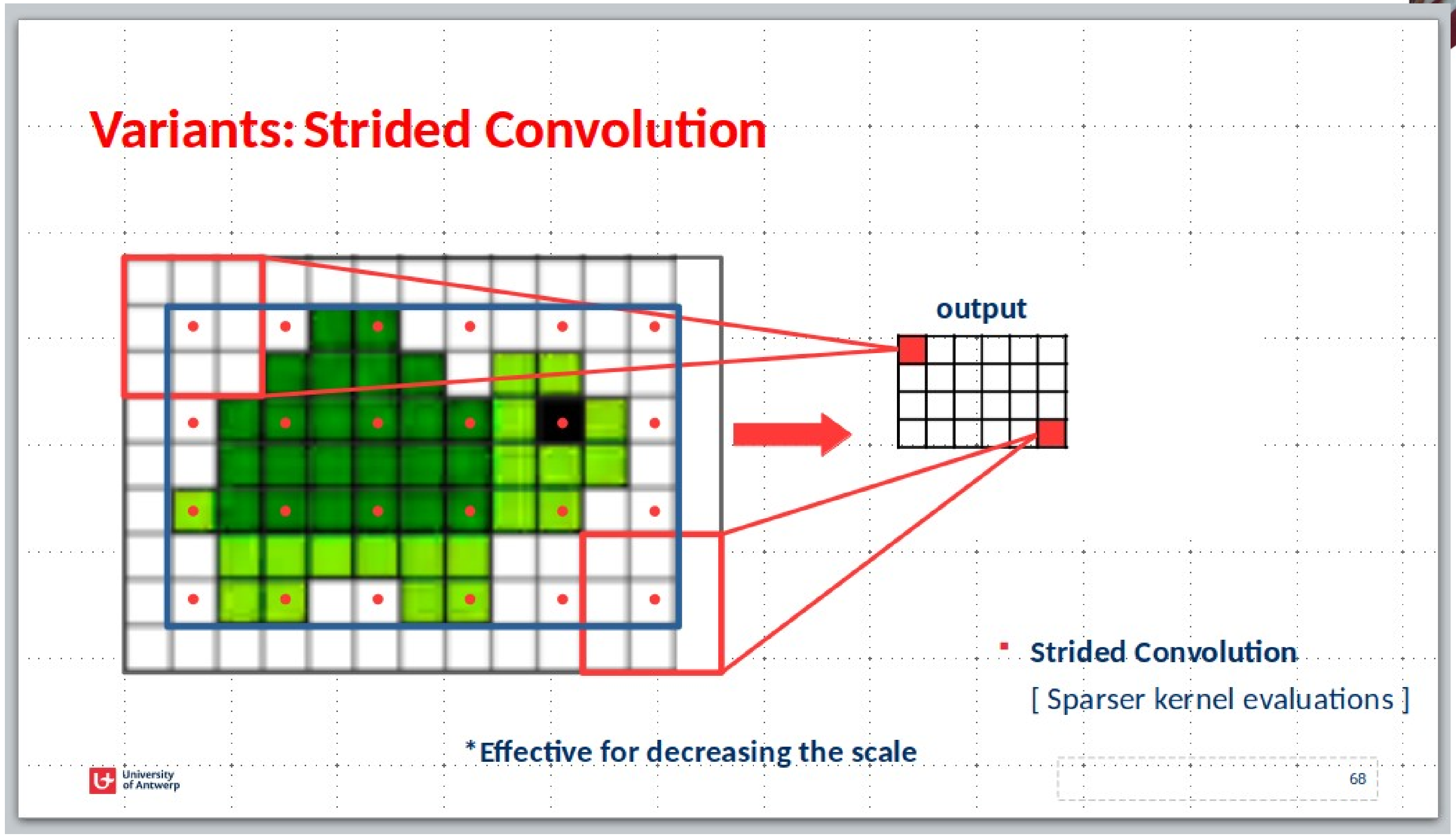
From lecture-2

Task-2: Semantic Segmentation: FCNs



How was it done? (at least the first time)

- Upsampling: backwards strided convolution

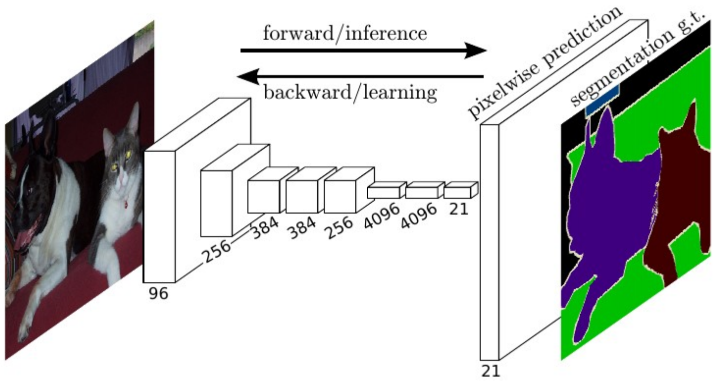


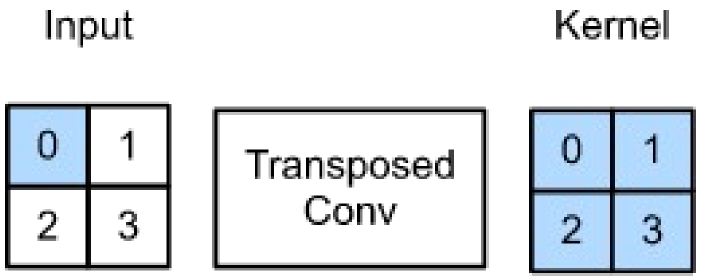
[Long et al., 2015]

Task-2: Semantic Segmentation: FCNs

How was it done? (at least the first time)

- Upsampling: backwards strided convolution



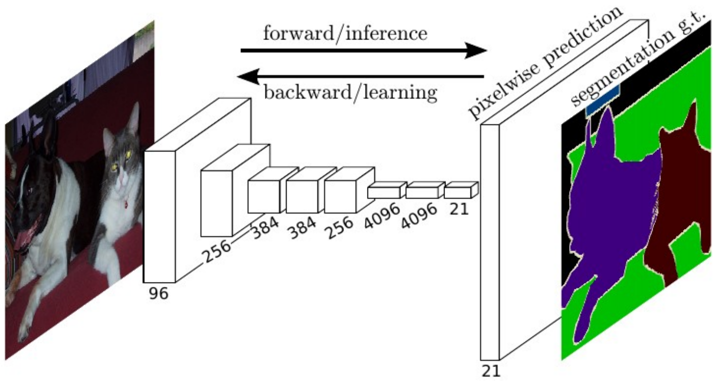


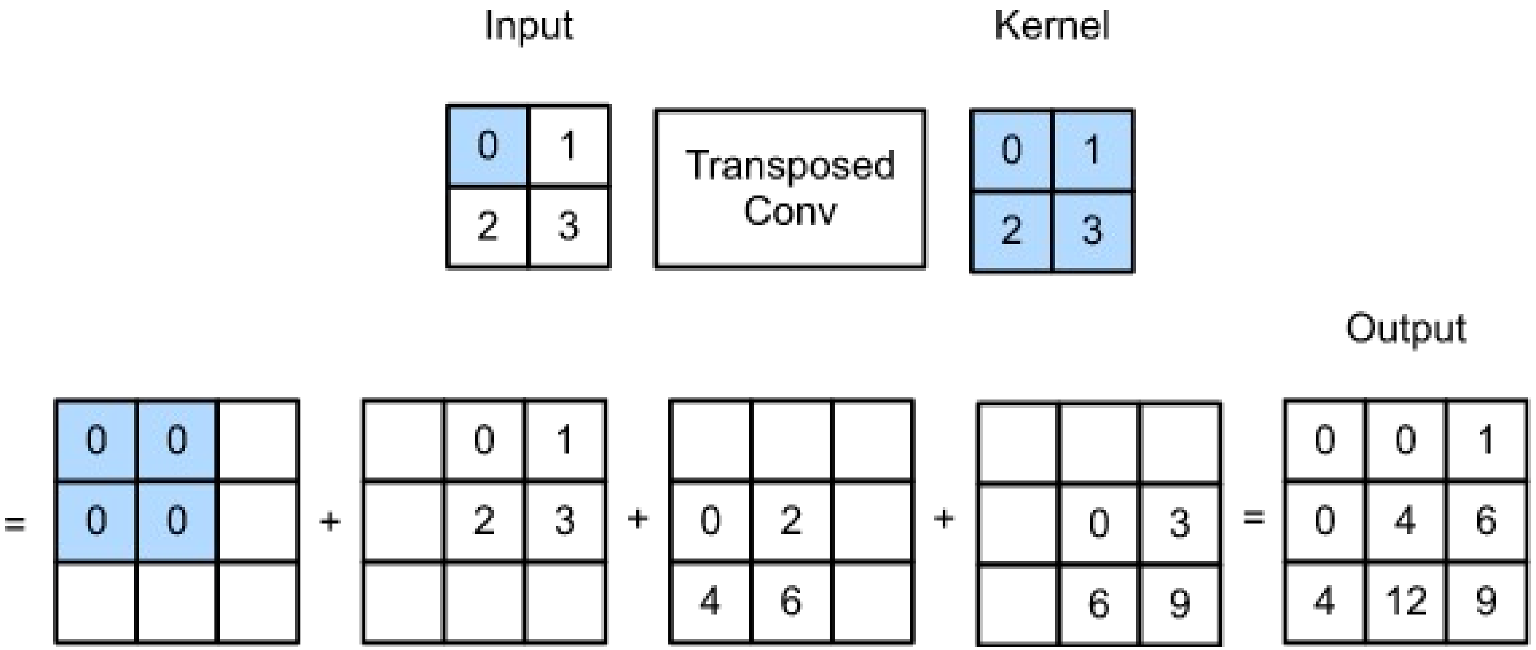
[D2L Chp 13.10]

Task-2: Semantic Segmentation: FCNs

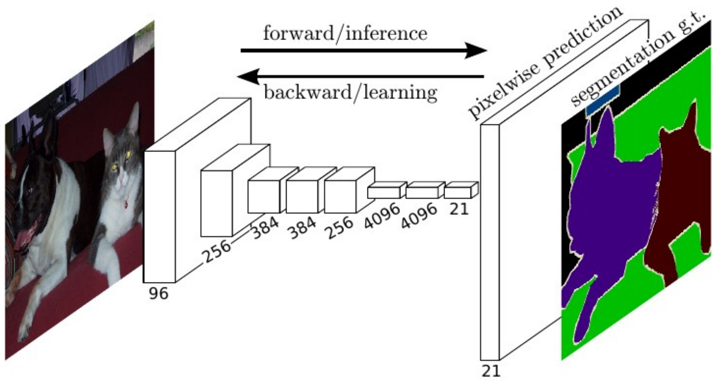
How was it done? (at least the first time)

- Upsampling: backwards strided convolution



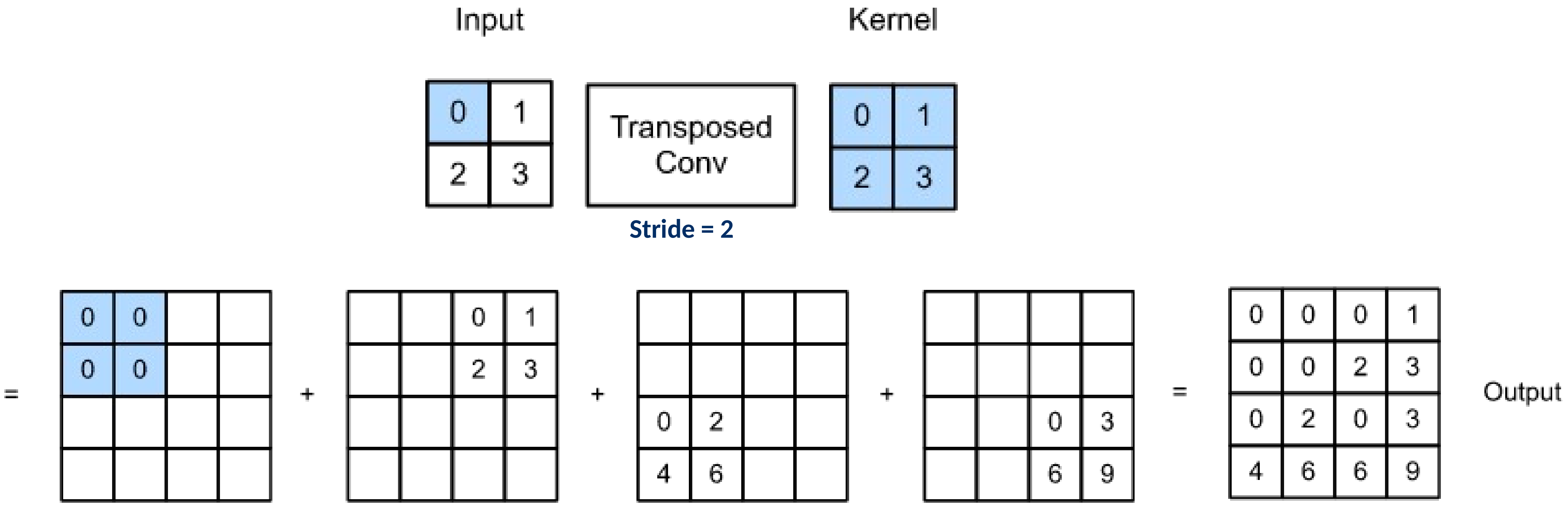


Task-2: Semantic Segmentation: FCNs



How was it done? (at least the first time)

- Upsampling: backwards strided convolution

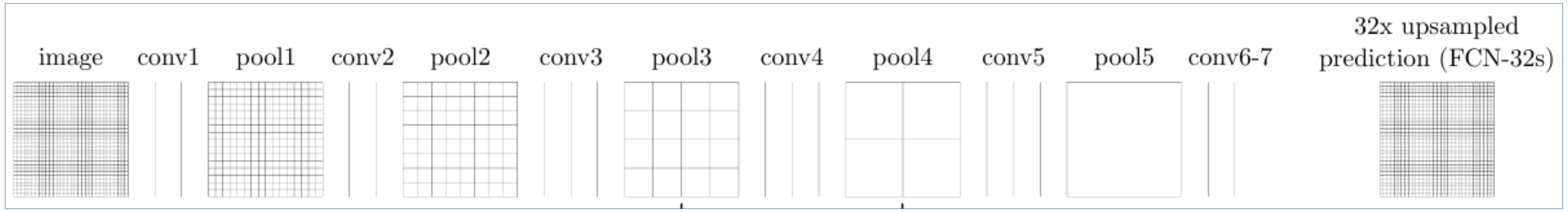


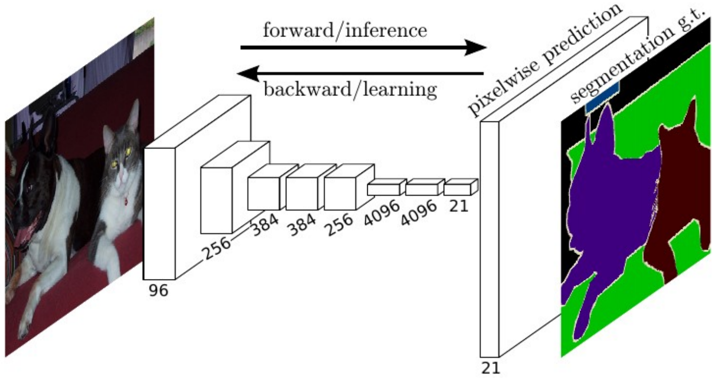
[D2L Chp 13.10]

Task-2: Semantic Segmentation: FCNs

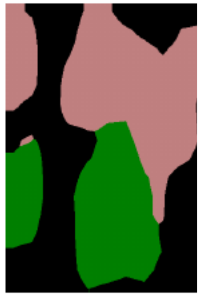
How was it done? (at least the first time)

– Further Improvements





FCN-32s



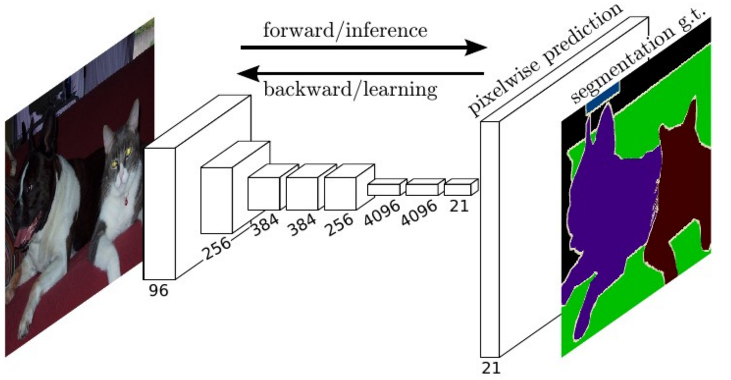
Ground truth

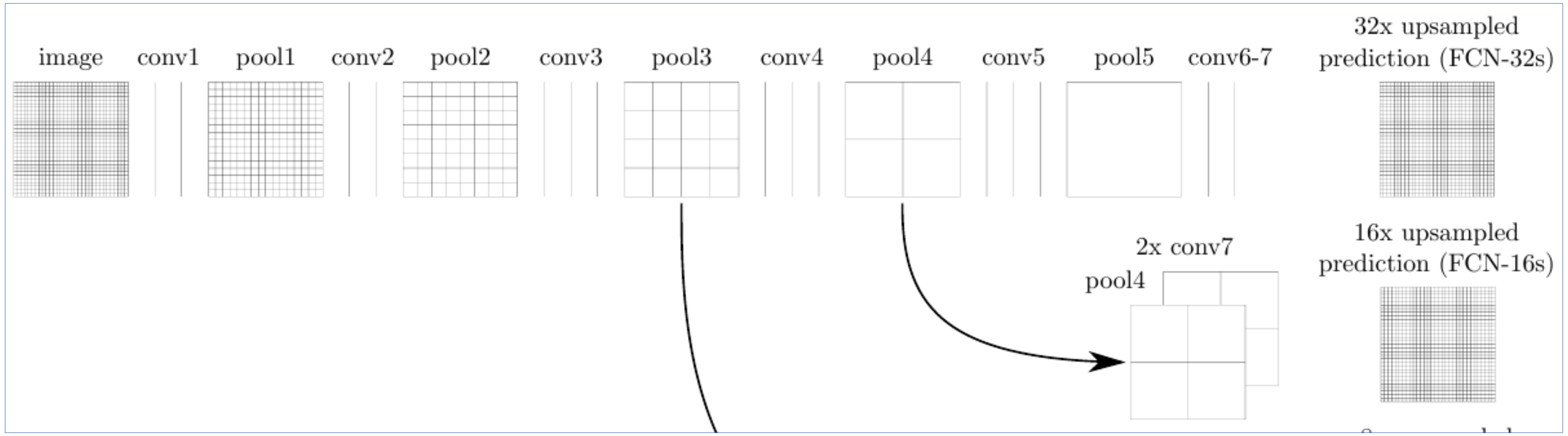
[Long et al., 2015]

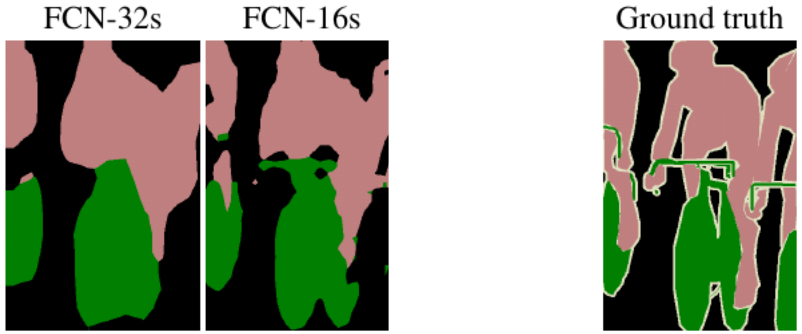
Task-2: Semantic Segmentation: FCNs

How was it done? (at least the first time)

– Further Improvements





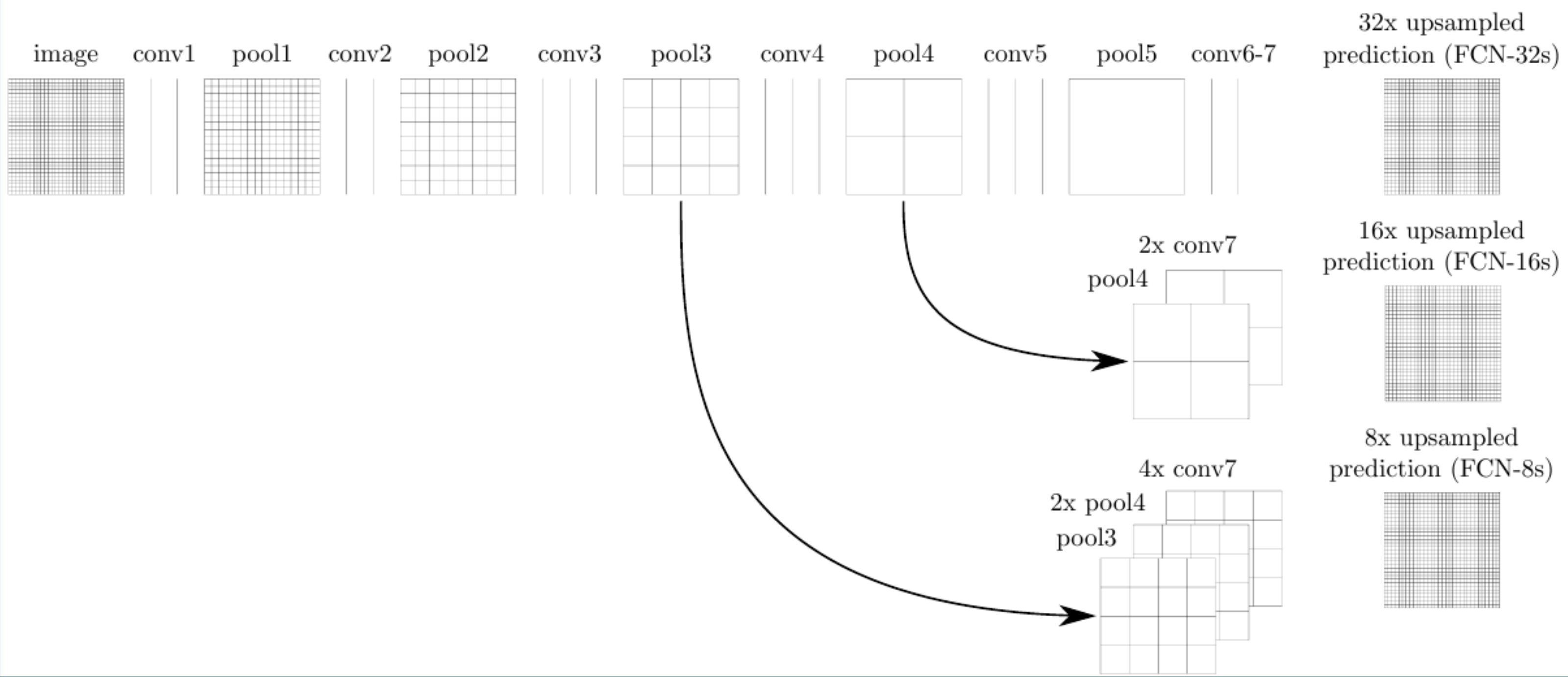


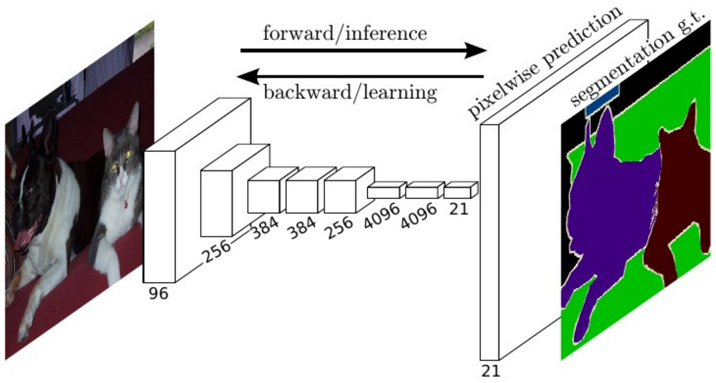
[Long et al., 2015]

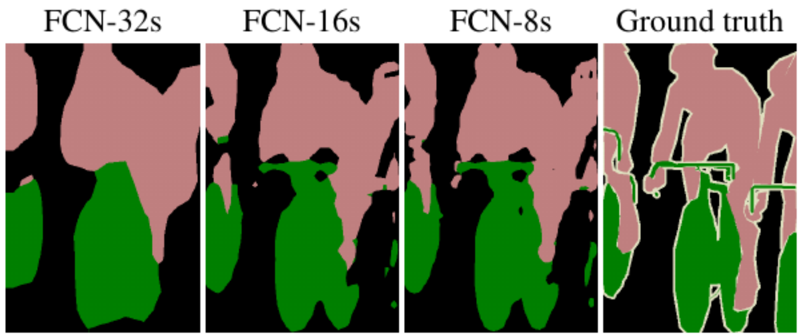
Task-2: Semantic Segmentation: FCNs

How was it done? (at least the first time)

- Further Improvements







[Long et al., 2015]

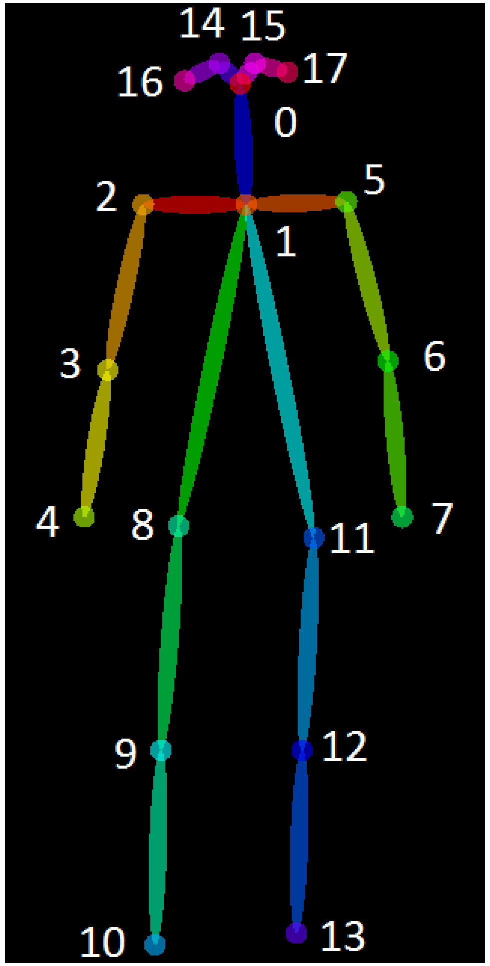
Dense-Sparse Predictions

Task-3: Just a Simple Cue

Can you guess what we will be talking about?

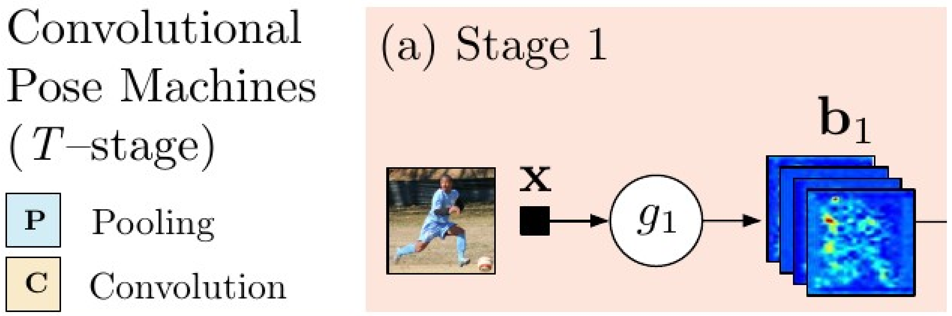
Task-3: Just a Simple Cue

Can you guess what we will be talking about?



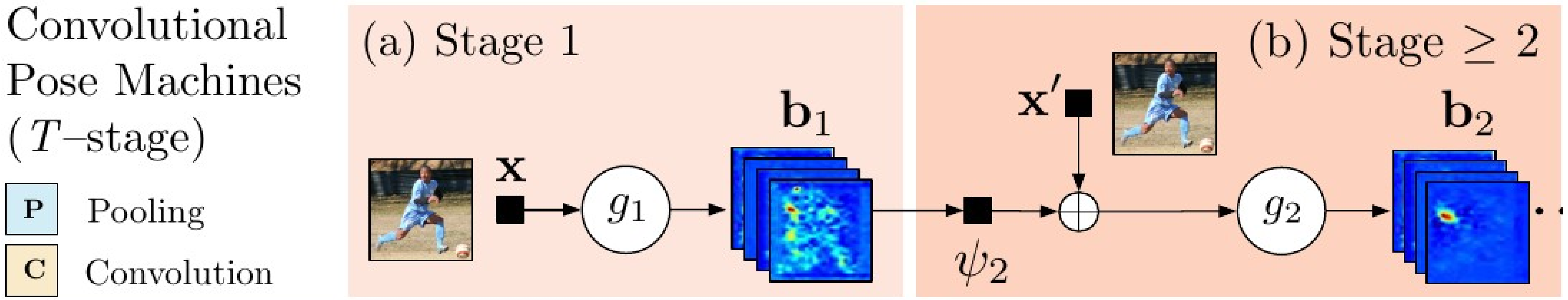
Task-3: Human Pose Estimation

Convolutional Pose Machines [Wei et al., 2016]



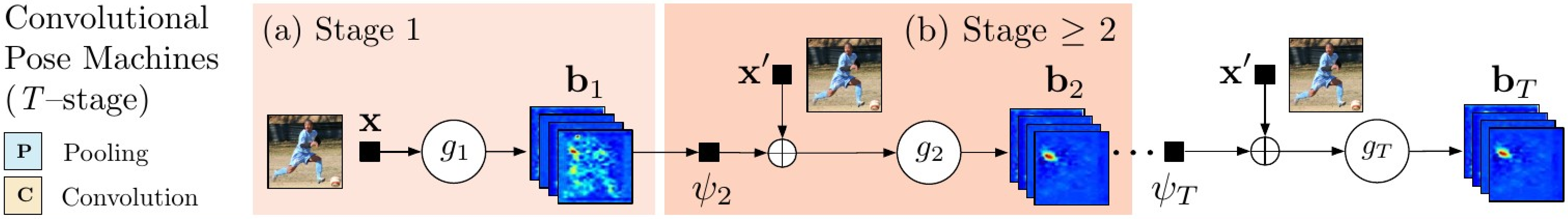
Task-3: Human Pose Estimation

Convolutional Pose Machines [Wei et al., 2016]



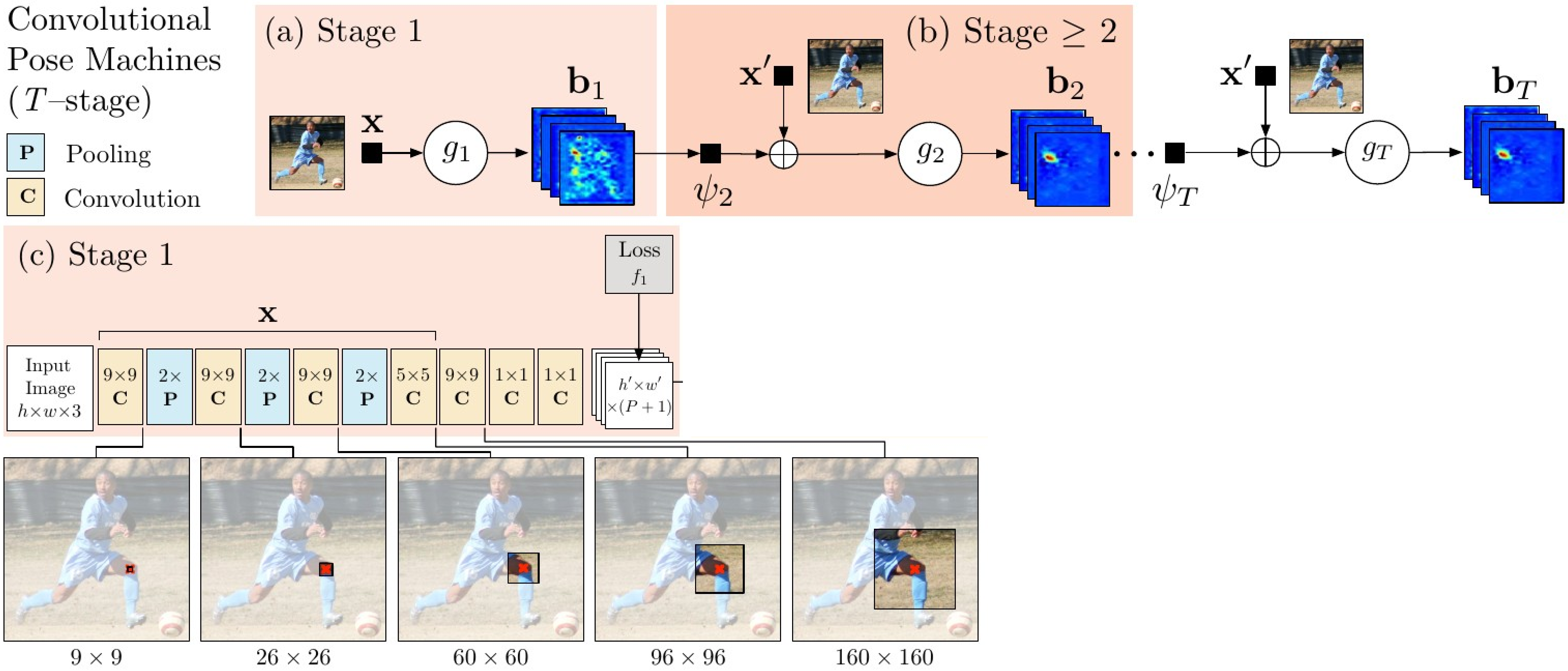
Task-3: Human Pose Estimation

Convolutional Pose Machines [Wei et al., 2016]



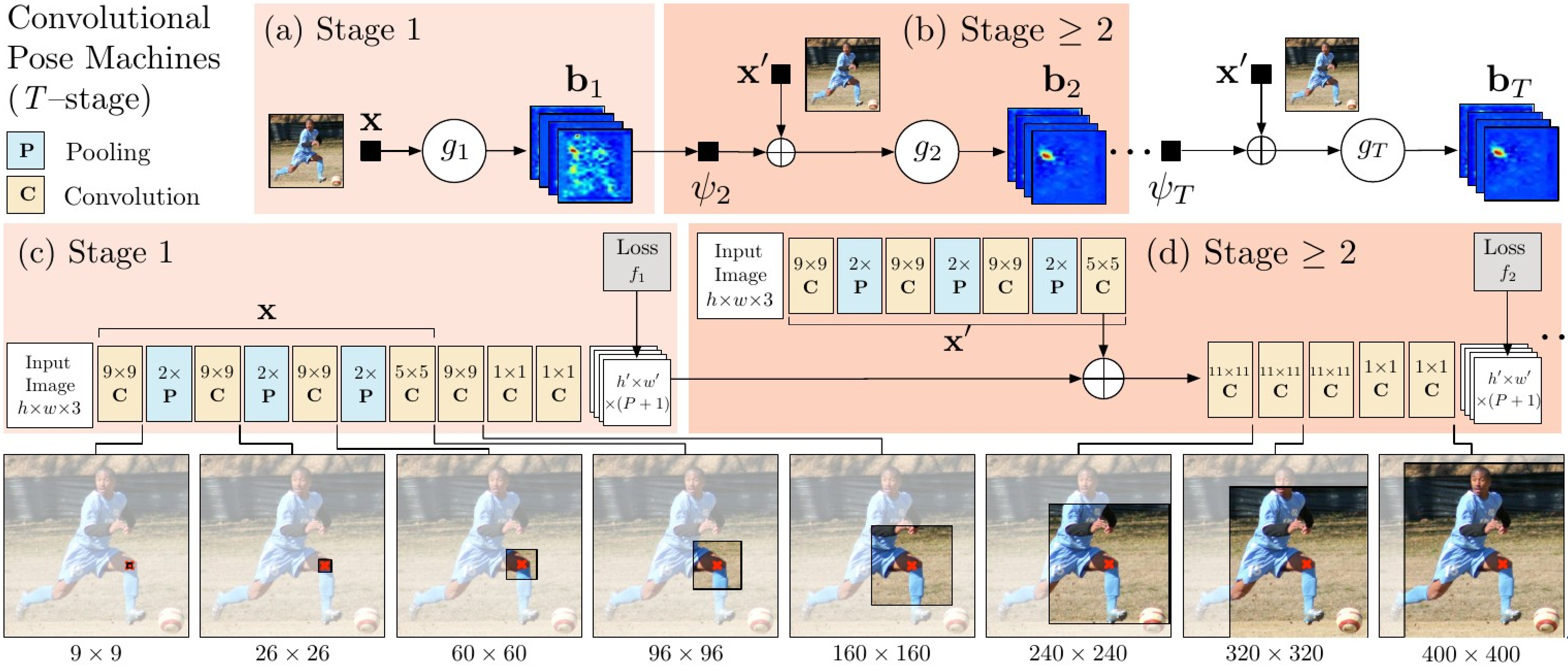
Task-3: Human Pose Estimation

Convolutional Pose Machines [Wei et al., 2016]



Task-3: Human Pose Estimation

Convolutional Pose Machines [Wei et al., 2016]

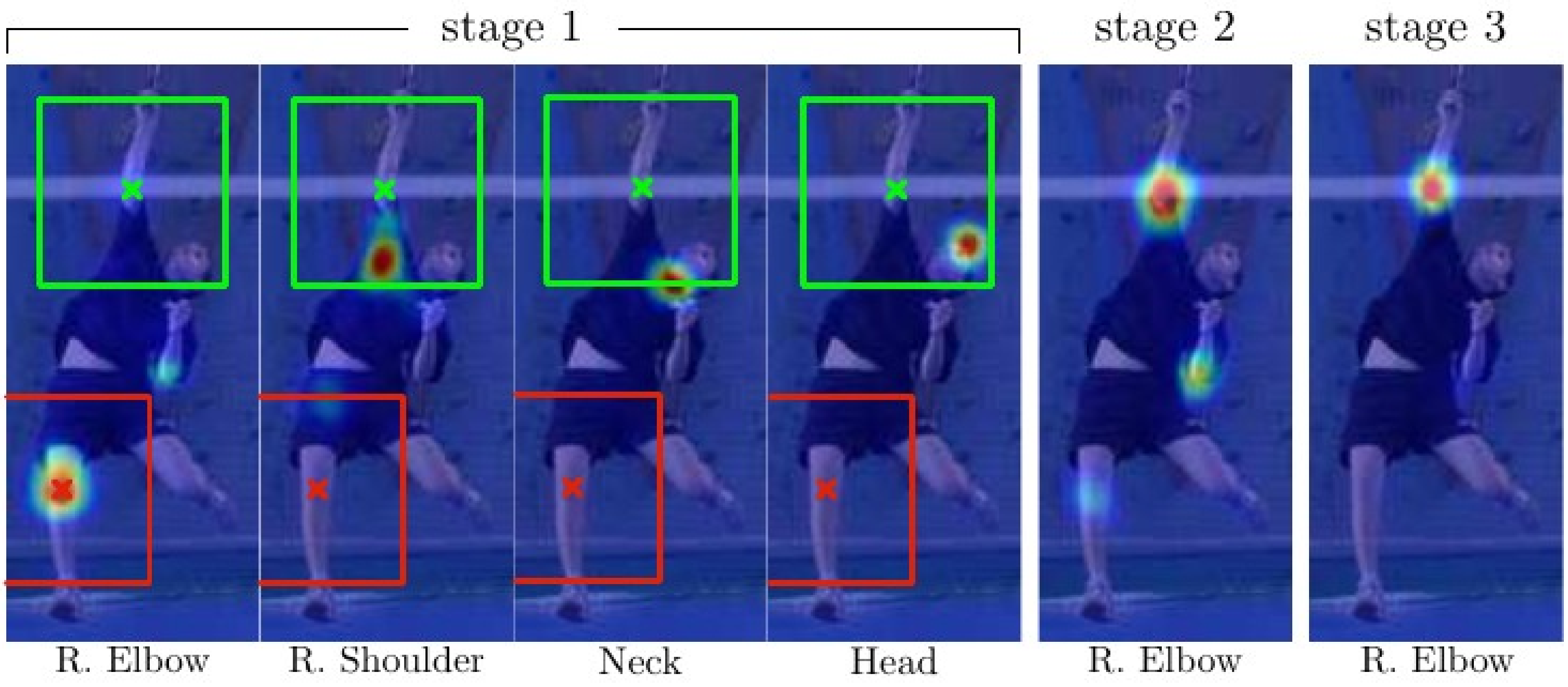


Task-3: Human Pose Estimation

Convolutional Pose Machines [Wei et al., 2016]

- Spatial Context induced by depth.

Elbow Prediction
– Correct
– Incorrect



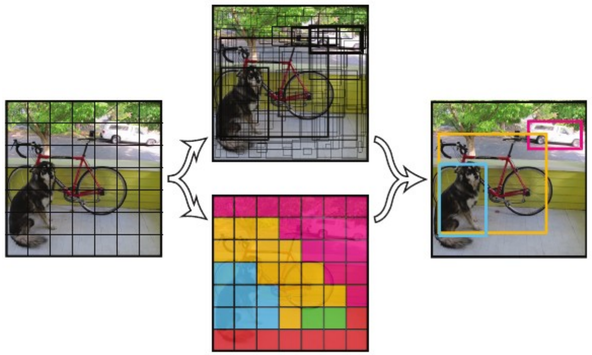
Summarizing

[Different Tasks, Same Components]

Summarizing

- **Convolutions go beyond simple classification**

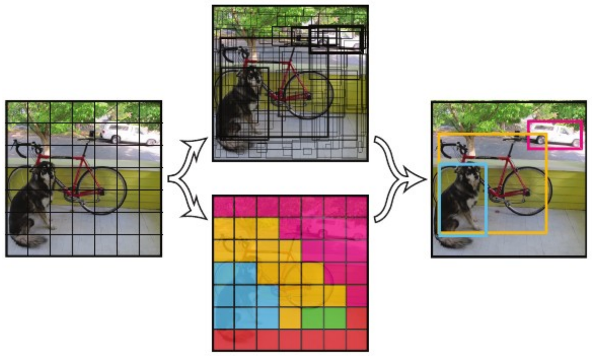
  localization | dense prediction



Summarizing

- **Convolutions go beyond simple classification**

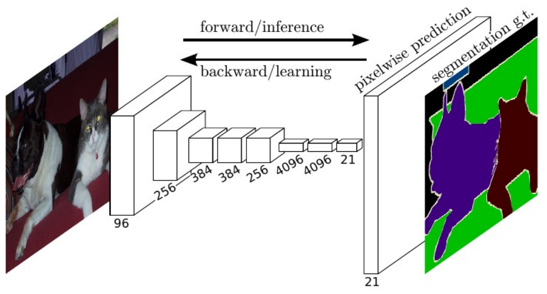
  localization | dense prediction



- **Additional use of convolutions**

  Transpose → Useful for upscaling operations

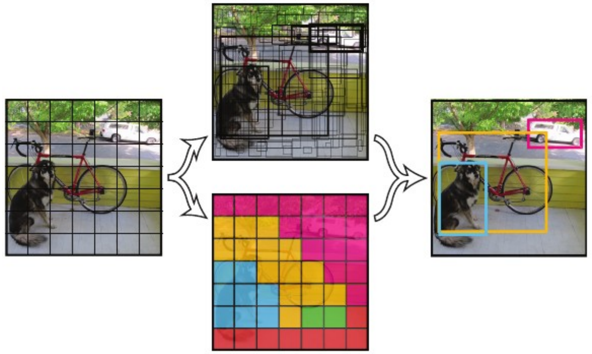
  FC Layers as Convolutions → useful for resolution invariance



Summarizing

- **Convolutions go beyond simple classification**

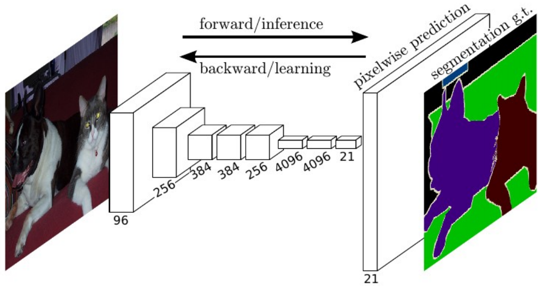
  localization | dense prediction



- **Additional use of convolutions**

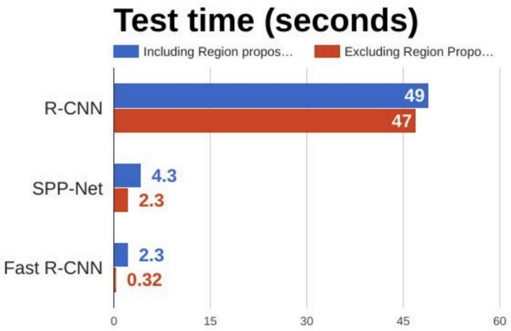
  Transpose → Useful for upscaling operations

  FC Layers as Convolutions → useful for resolution invariance



- **Suitable designs → better performance**

  time invested at design-time eventually pays off



References

- Dive into Deep Learning (D2L)
  - Chapter 13.10: **Transposed Convolutions** – https://d2l.ai/chapter_computer-vision/transposed-conv.html
  - Chapter 13.11: **Fully Convolutional Layers** – https://d2l.ai/chapter_computer-vision/fcn.html
- R. Girshick, J. Donahue, T. Darrell, J. Malik, **Rich feature hierarchies for accurate object detection and semantic segmentation**, CVPR 2014
- R. Girshick, **Fast R-CNN**. ICCV 2015
- S. Ren, K. He, R. Girshick J. Sun, **Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks**, NeurIPS 2015
- J. Redmon, S. Divvala, R. Girshick, A. Farhadi, **You Only Look Once: Unified, Real-Time Object Detection**, NeurIPS 2015
- J. Long, E. Shelhamer, T. Darrell, **Fully Convolutional Networks for Semantic Segmentation**, CVPR 2015
- S. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, **Convolutional Pose Machines**, CVPR 2016.

University of Antwerp

Artificial Neural Networks

[2500WETANN]

José Oramas