Lec13 Computer Vision

Yang Shu

School of Data Science and Engineering

yshu@dase.ecnu.edu.cn

[Acknowledgement: Slides are adapted from Deep Learning Course, Mingsheng Long, THU]

Outline

- **Image**
  - **Recognition**, segmentation, detection, stylization
- Video:
  - Recognition, detection
- 3D Vision:
  - Volumetric data

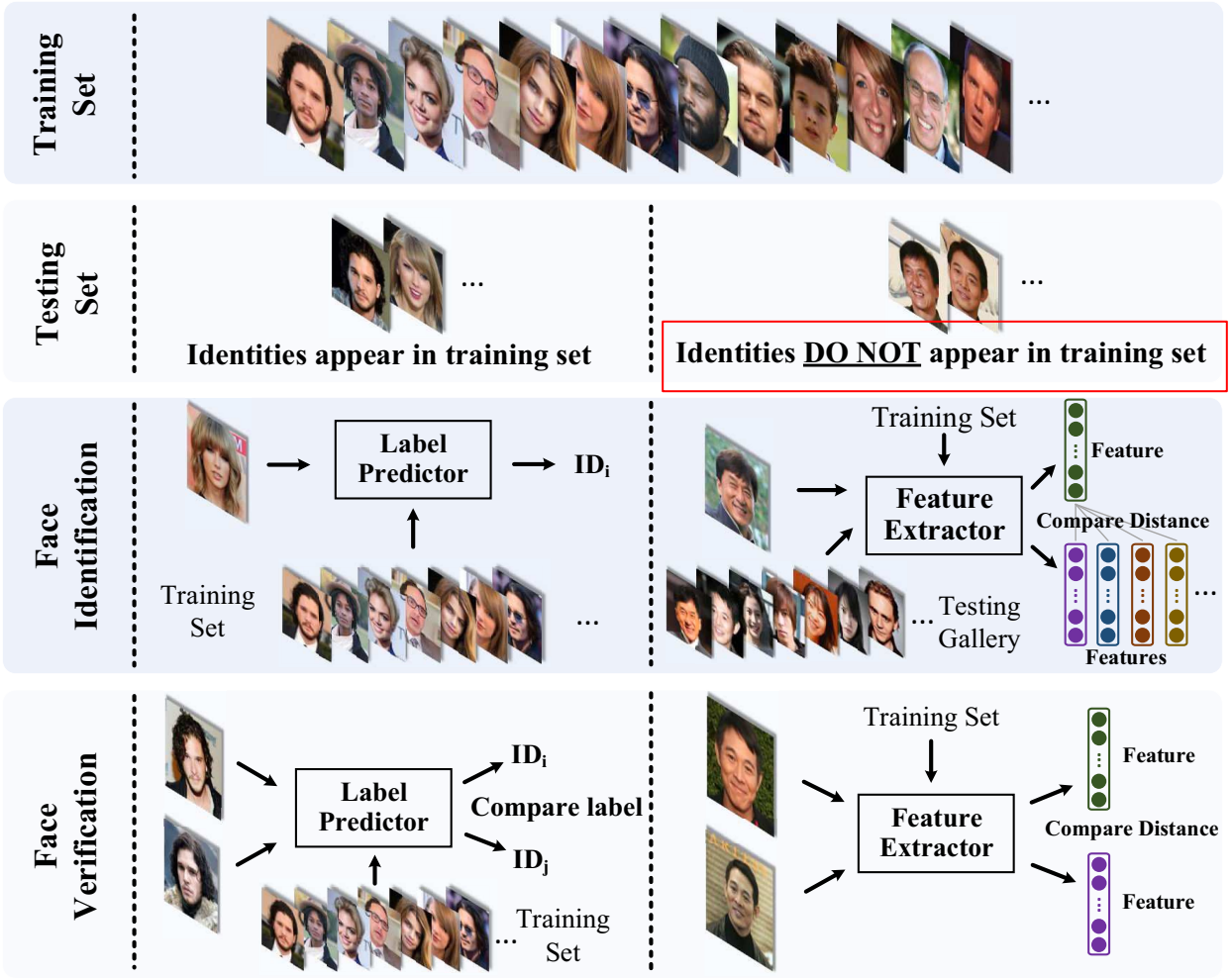
Face Recognition

Closed-set face recognition

Open-set face recognition

人脸识别 (FID)

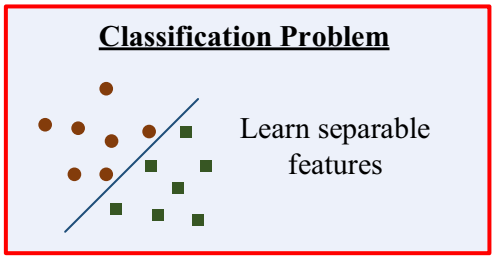
人脸确认 (FV)



Liu, Weiyang, et al. "Sphereface: Deep hypersphere embedding for face recognition." CVPR 2017.

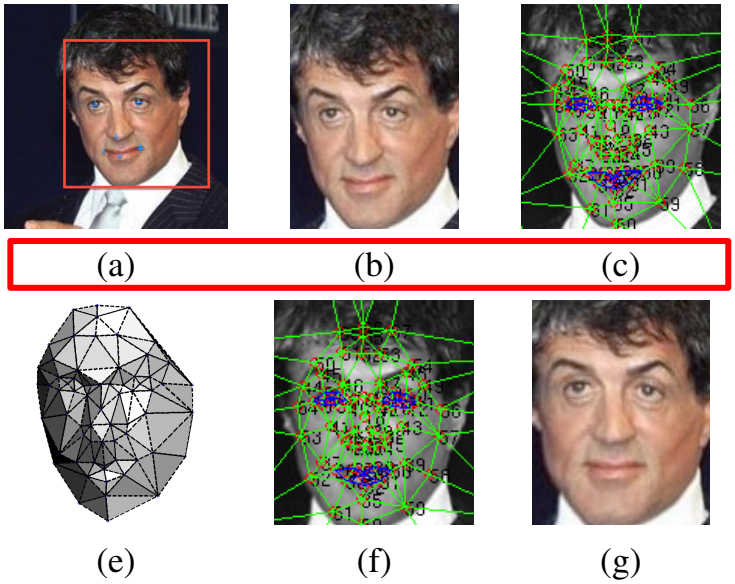
DeepFace

Closed-set face recognition

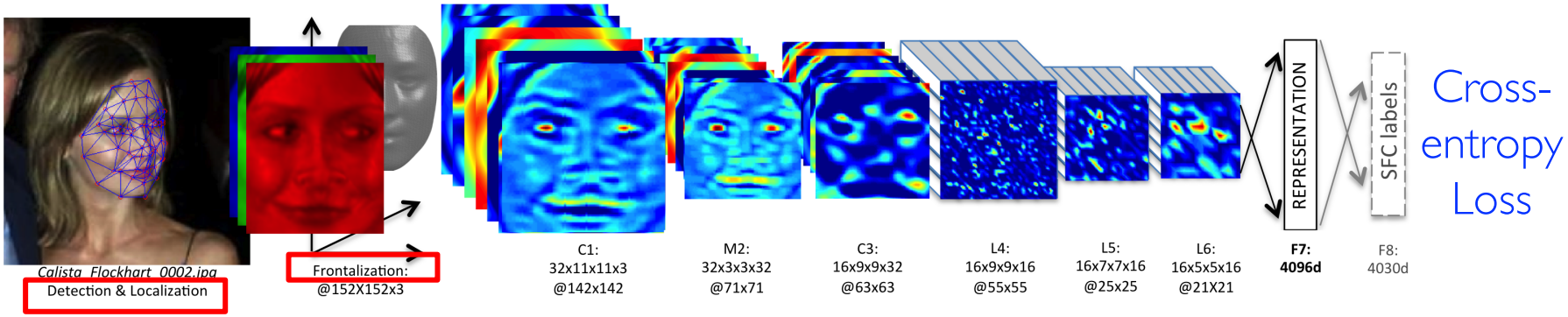


- The first deep method approaching human-level performance
- Two-stage learning:
  - Shallow face alignment
  - CNN feature learning

FR: Who is the person?



Face alignment with shallow method



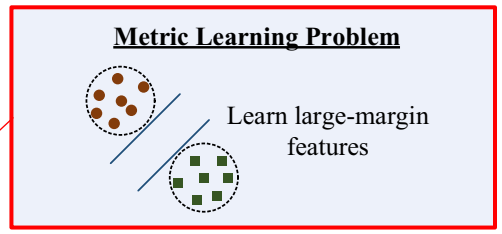
Cross-entropy Loss

Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." CVPR 2014. (Cited 4173)

Face Verification

Open-set face recognition

Large-margin learning:
Generalize to unseen faces.



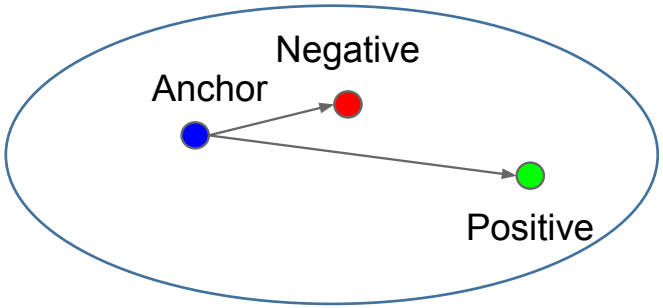
FV:
Who is the same person?

Anchor

Positive

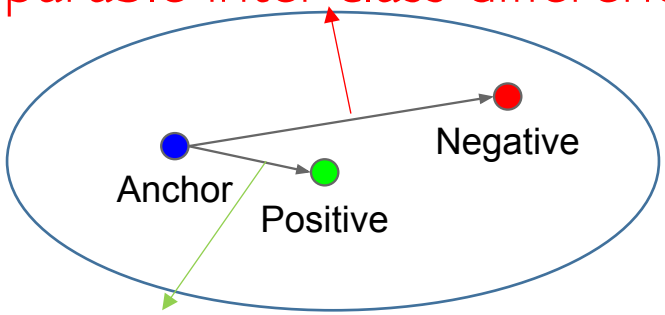
Negative

How to select triplets? $O(n^3)$



LEARNING

Separable inter-class differences



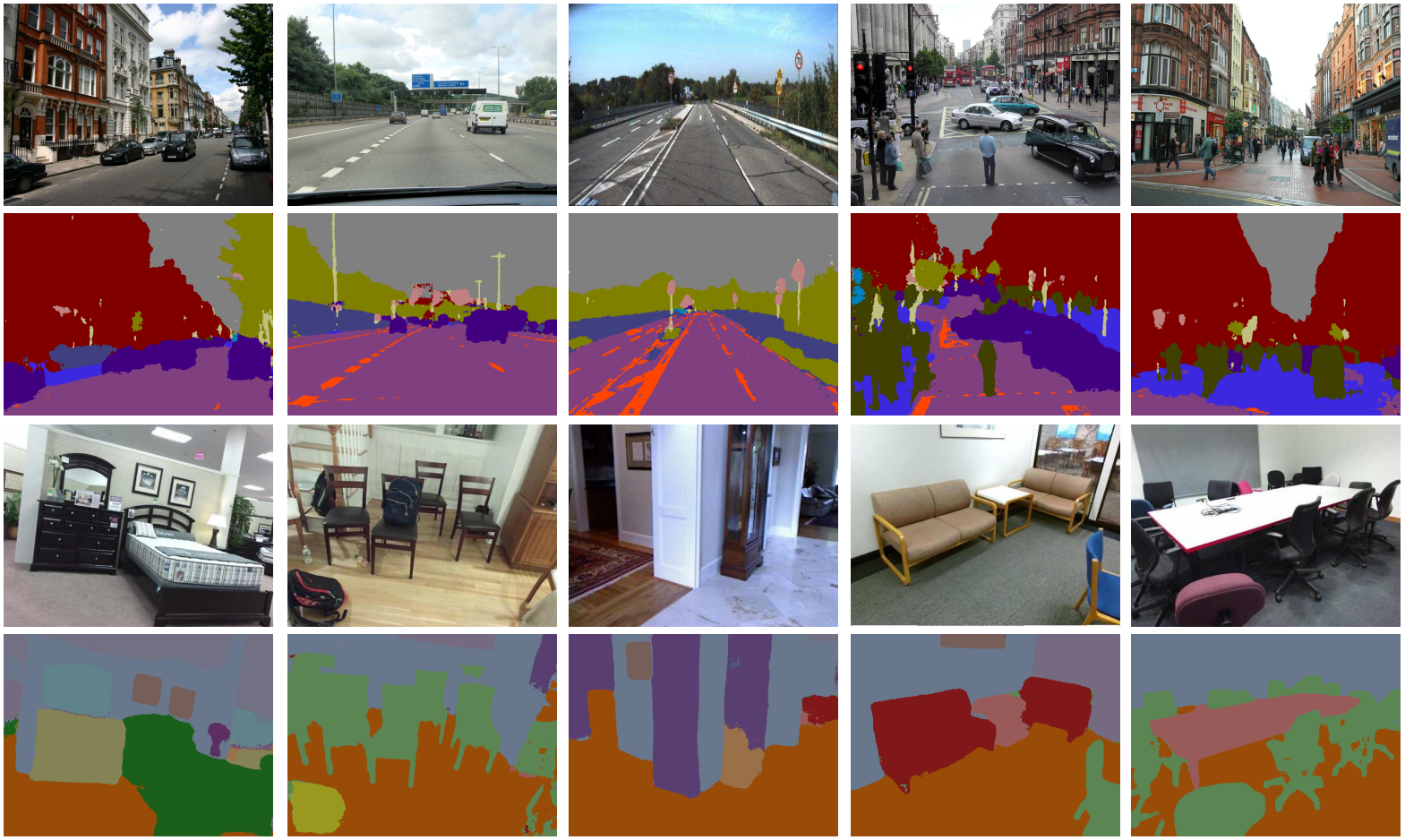
Compact intra-class variation

F. Schroff et al. FaceNet: A Unified Embedding for Face Recognition and Clustering. CVPR 2015.

Outline

- **Image**
  - Recognition, **segmentation**, detection, stylization
- Video:
  - Recognition, detection
- 3D Vision:
  - Volumetric data

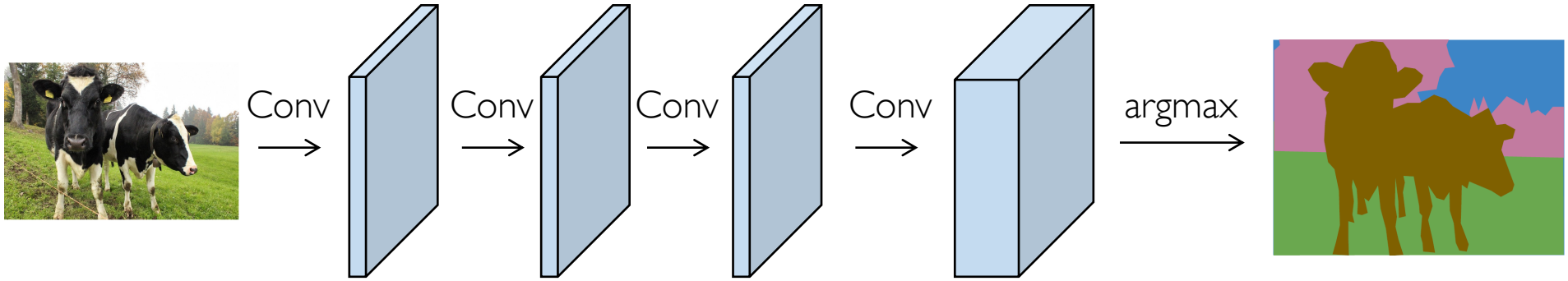
Semantic Segmentation



- Classification at pixel level: low-level vision tasks. (labeling expensive)

Fully Convolutional Network (FCN)

- **Basic idea:** Design a network as a bunch of convolutional layers to make predictions for pixels all at once!
- Convolutional layer keeps spatial information and extracts high-level semantics.



Input:
3 x H x W

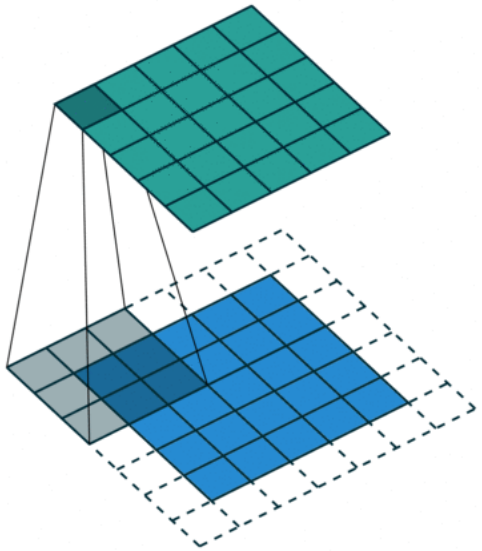
Convolutions:
D x H x W

Scores:
C x H x W

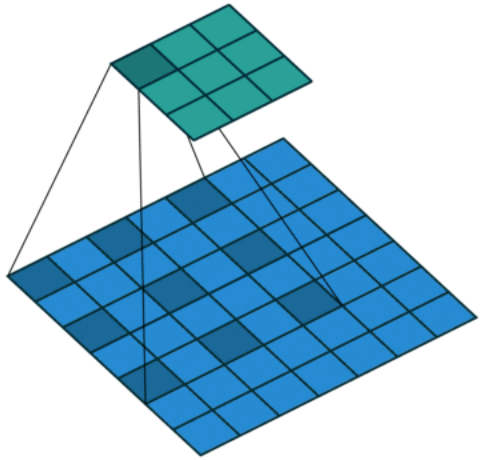
Predictions:
H x W

- Problem: convolutions at the original resolution will be very expensive

Transpose Convolution

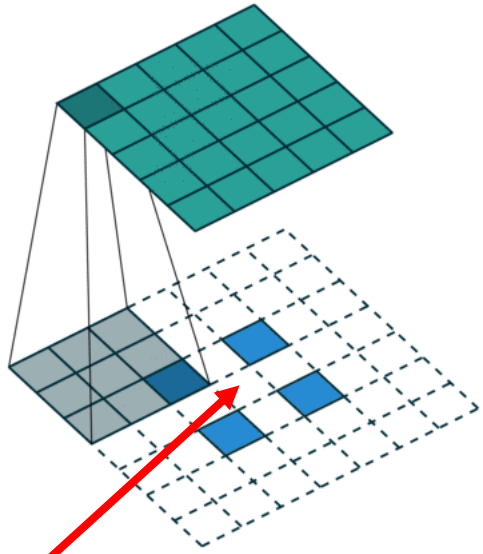


Convolution
Kernel Size: 3x3
Stride: 1
Padding: 1
Dilation: 1



Dilated Convolution
Kernel Size: 3x3
Stride: 1
Padding: 0
Dilation: 2

Original Conv Stride-2

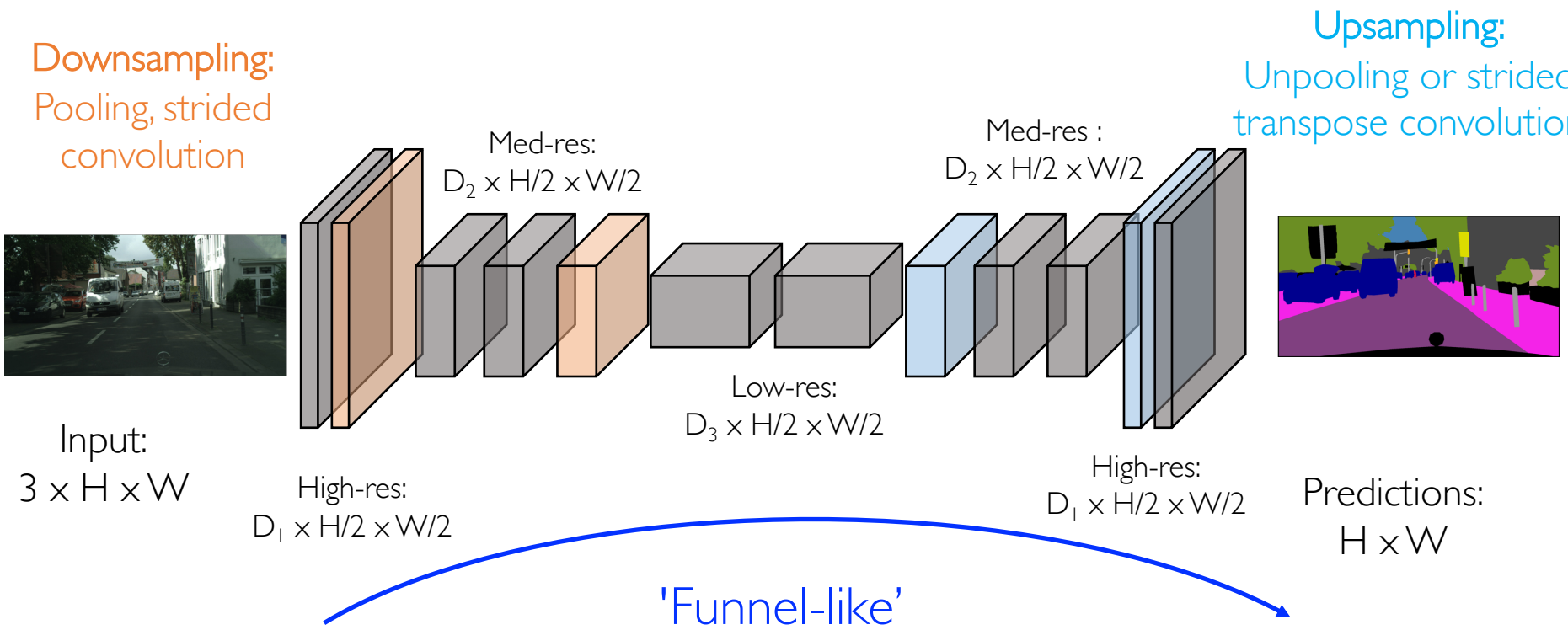


Transpose Convolution
Kernel Size: 3x3
Stride: 1
Padding: 3
Dilation: 1

http://deeplearning.net/software/theano_versions/dev/tutorial/conv_arithmetic.html

Fully Convolutional Network (FCN)

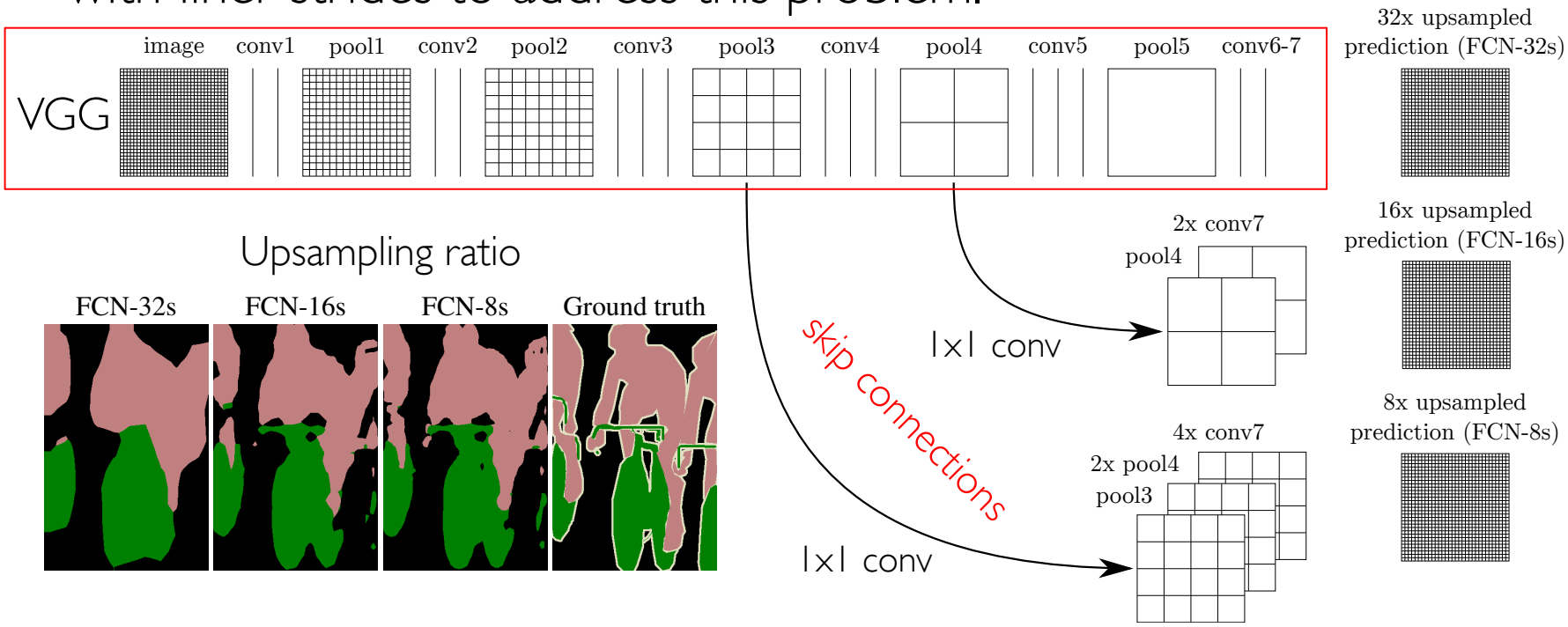
- Design network as a bunch of convolutional layers, with (differentiable) downsampling and upsampling inside the network!



Long, Jonathan et al. "Fully convolutional networks for semantic segmentation." CVPR 2015.

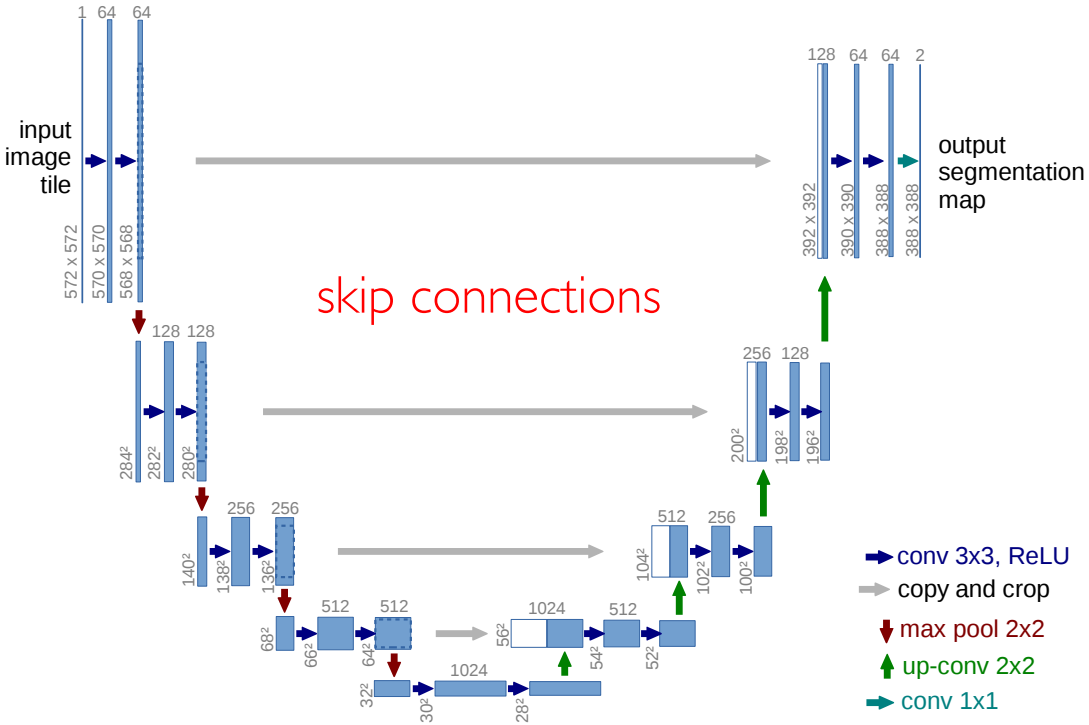
Fully Convolutional Network (FCN)

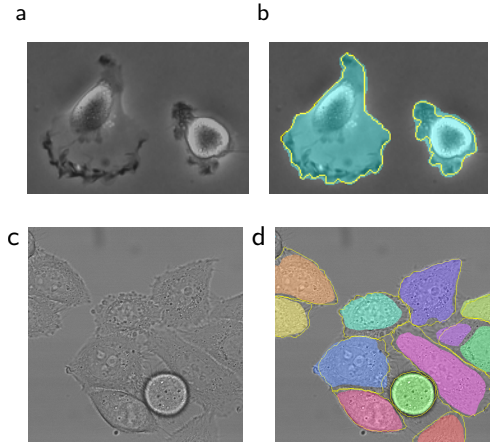
- With downsampling, local details are missing and the prediction is very coarse.
- Add skips that combine the final prediction layer and lower layers with finer strides to address this problem.



Long, Jonathan et al. "Fully convolutional networks for semantic segmentation." CVPR 2015.

U-net





Result

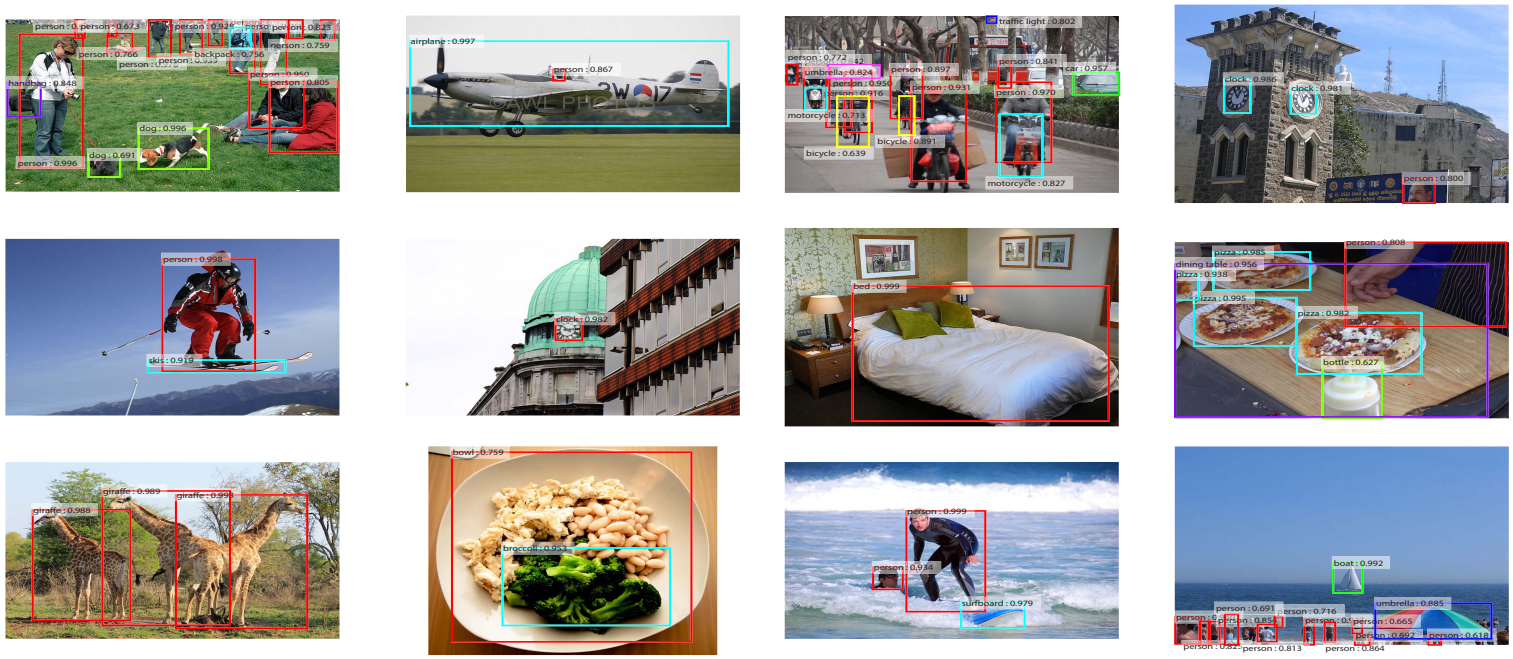
- An important baseline for biomedical image segmentation.
- Similar idea with FCN: down-up sampling + skip-connection.
- Usually no need for using of deep backbone in medical segmentation.

Ronneberger, Olaf et al. "U-net: Convolutional networks for biomedical image segmentation." MICCAI 2015.

Outline

- **Image**
  - Recognition, segmentation, **detection**, stylization
- Video:
  - Recognition, detection
- 3D Vision:
  - Volumetric data

Object Detection

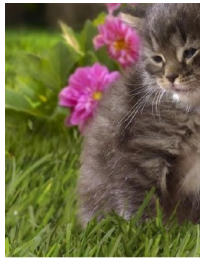


- Objective Detection: Classification + Localization
- Basic idea:
  - Find box that may contain objects (region proposals)
  - Classify the object (classification) + Refine the box (regression)

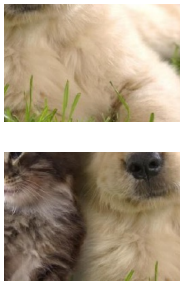
Region Proposals

There are $(256 \times 256)^2$ boxes in a 256x256 images

But not all of the boxes are useful.



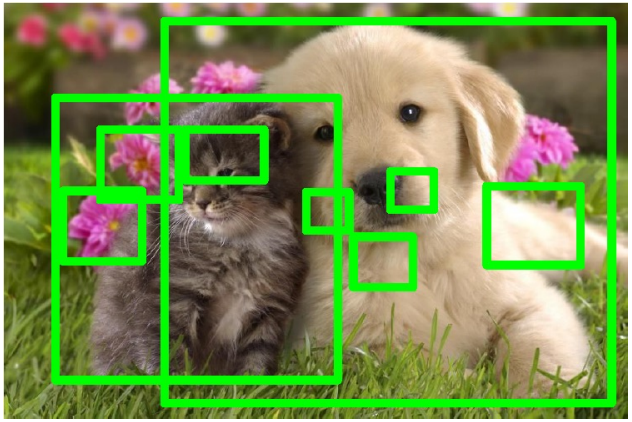
It's OK



- Key problem: Find "blobby" image regions that are likely to contain objects
- Selective Search gives ~1000 region proposals in a few seconds on CPU

Main idea: Compute similarity between adjacent regions and combine similar regions.

Shallow Method



Uijlings et al, "Selective Search for Object Recognition", IJCV 2013

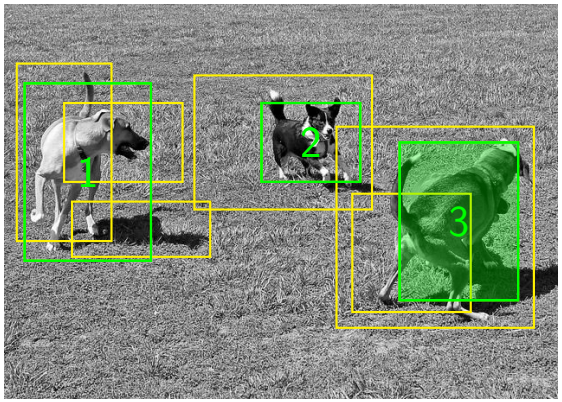
Non-Maximum Suppression (NMS)

- Given regions $r_1, r_2, \ldots$ of each class, ranked by order of confidence.
- Reject each region r_i if it has higher intersection-over-union (IoU):

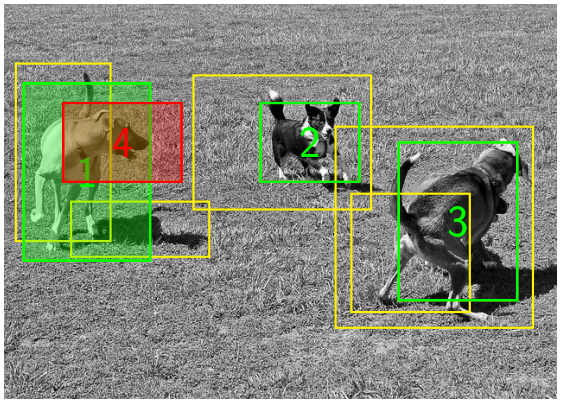
$$\text{IOU}(r_i, r_j) > \tau$$

with some higher scoring region $r_j, j < i$ that has not been rejected

Original

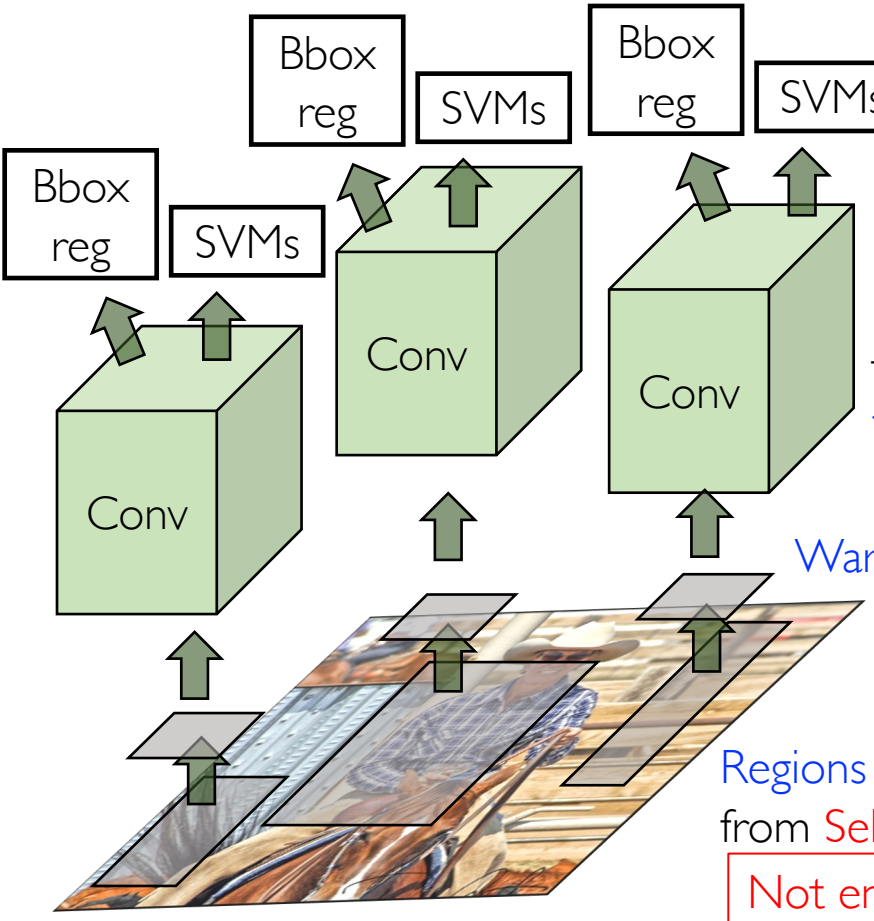


Accept 1, 2, 3



Reject 4

R-CNN



4 Linear Regression for bounding box offsets

4 Classify regions with SVMs (including background as a class)

Not end-to-end

3 Slow

Forward each region through ConvNet to a feature vector

2 Warped image regions

1 Regions of Interest (RoI) from Selective Search (2k+)

Not end-to-end | Slow

Bounding-box regressors

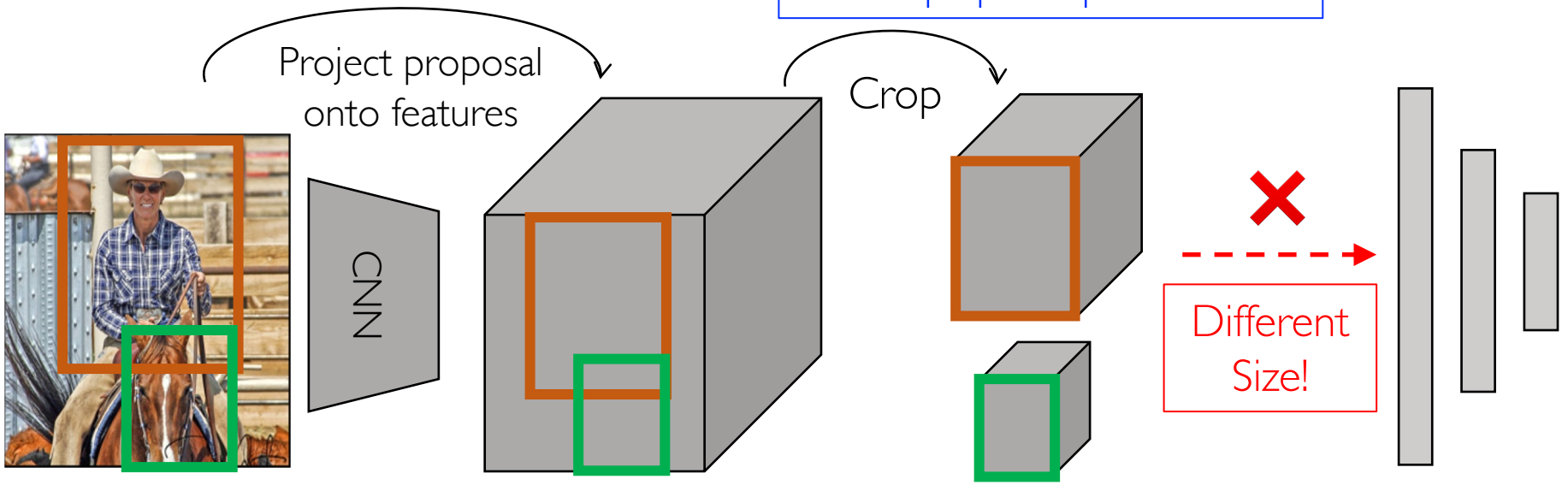
$$(\underbrace{\Delta x, \Delta y}_{\text{Translation}}, \underbrace{S_w, S_h}_{\text{Scaling}})$$

- Training is slow (84h), takes a lot of disk space
- Inference (detection) is slow
- Not end-to-end

Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." CVPR 2014.

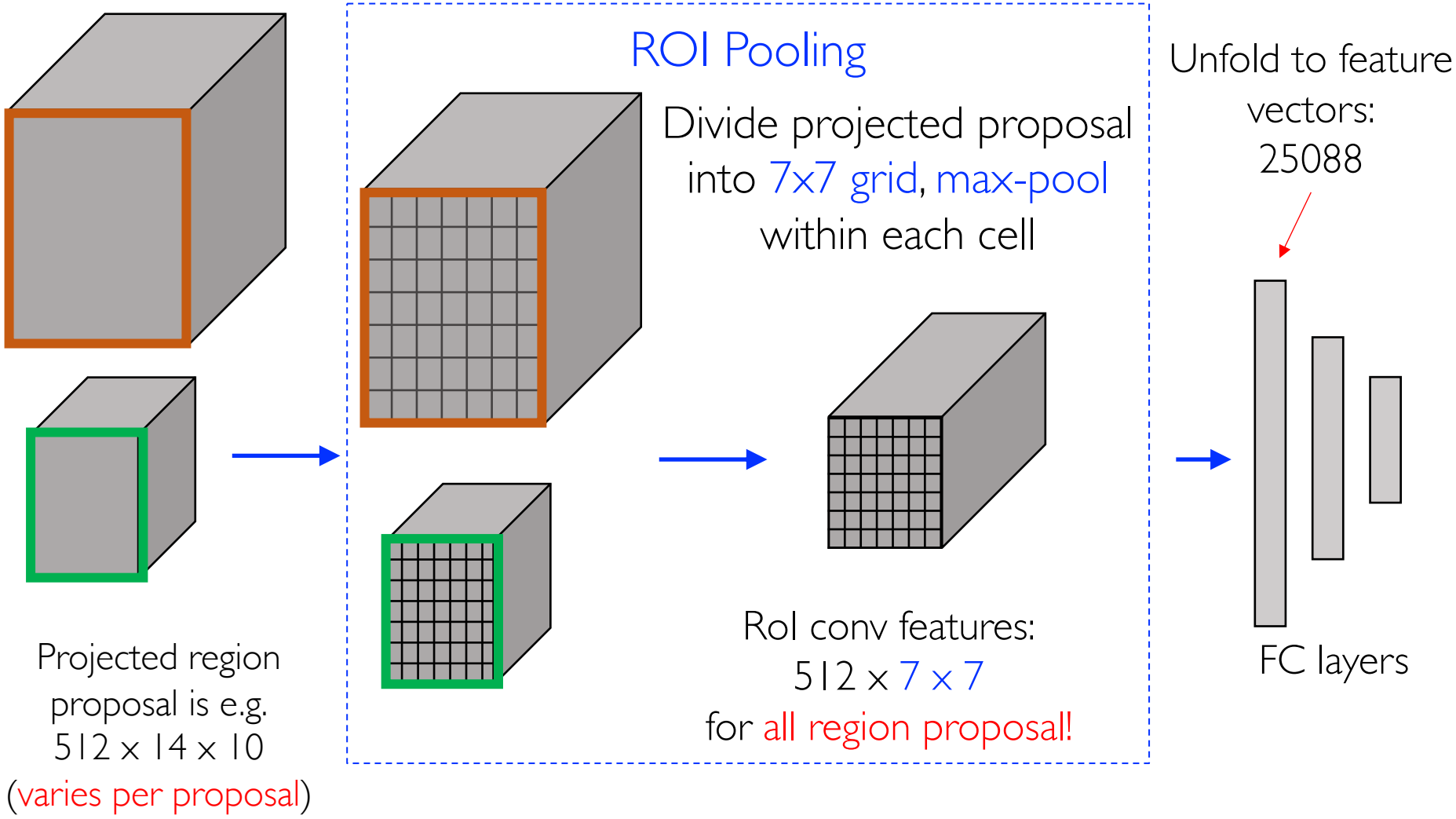
Fast R-CNN

- R-CNN needs to forward through CNN ~1000 times for a image.
- Can we forward only once?



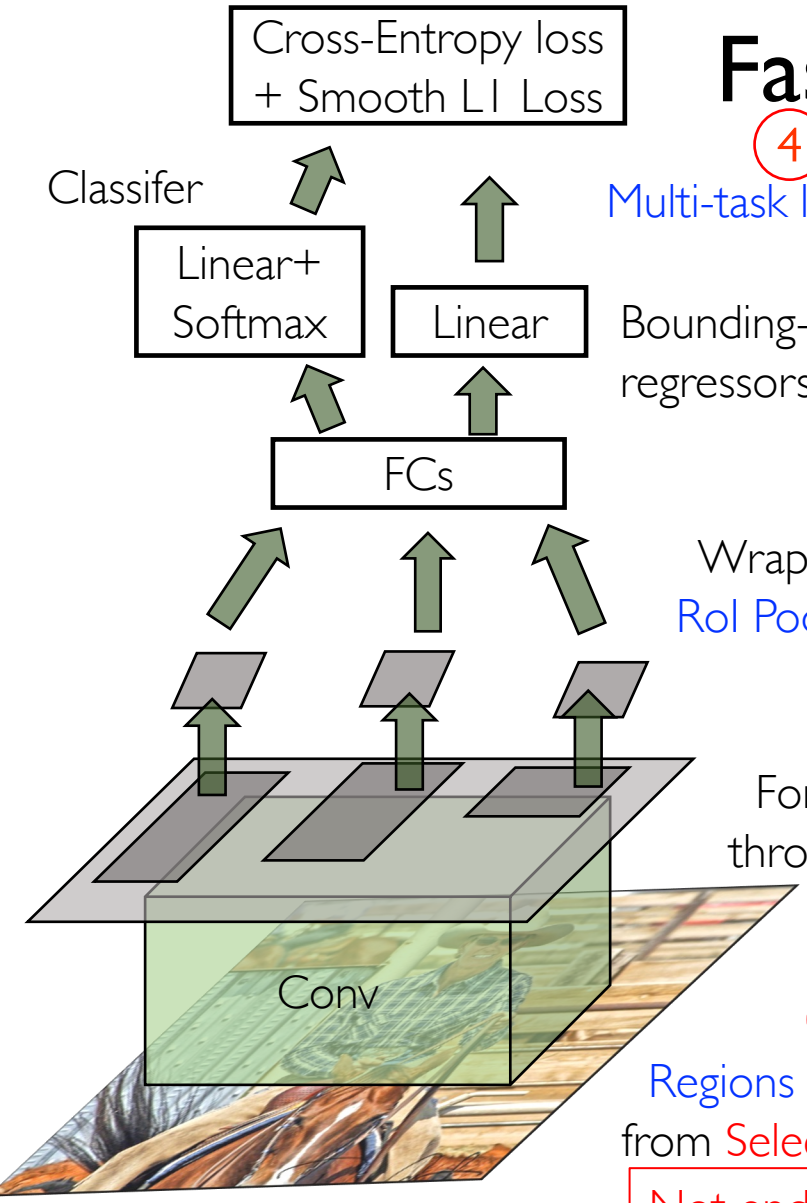
Girshick, "Fast R-CNN", ICCV 2015.

ROI Pooling



Girshick, "Fast R-CNN", ICCV 2015.

Fast R-CNN



4 Multi-task learning ✅

3 Wrap features and RoI Pooling ✅

2 Forward image through ConvNet ✅

1 Regions of Interest (RoI) from Selective Search (2k+)

Not end-to-end | Slow

Better loss for Bounding-box regressors:

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases}$$

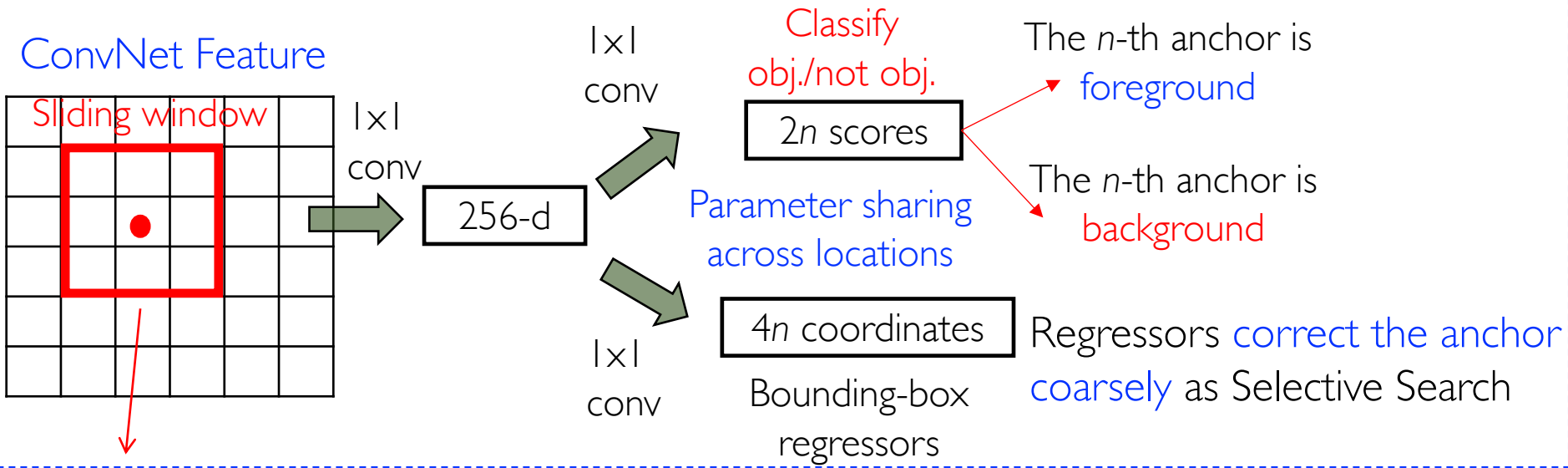
End-to-end trainable!
10x faster than R-CNN

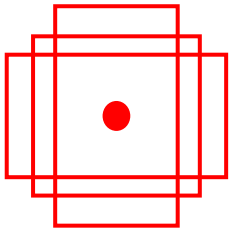
- Training and inference is faster.
- Selective Search is still relatively slow and not end-to-end

Could we use NN to find RoI instead?

Girshick, "Fast R-CNN", **ICCV** 2015.

Region Proposal Network (RPN)





Anchors
(*n* bounding boxes, *n*=9 usually)

- Slide a small window on the feature map.
- Build a small network for:
  - Classifying object or not-object (fg or bg)
  - Regressing bounding-box localizations
- Use *n* anchor boxes with different sizes at each location

Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." **NIPS** 2015.

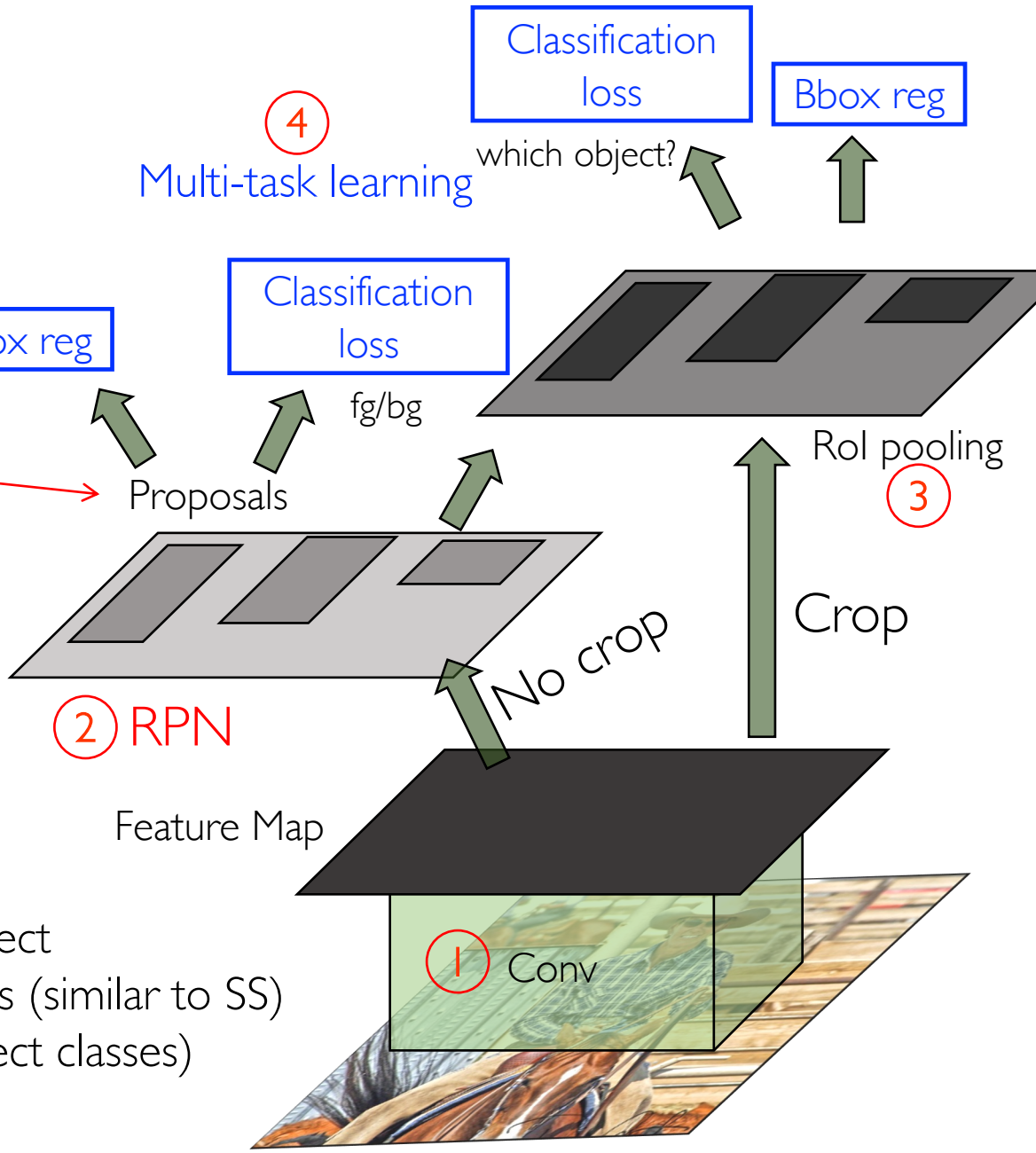
Faster R-CNN

Make CNN do proposals!

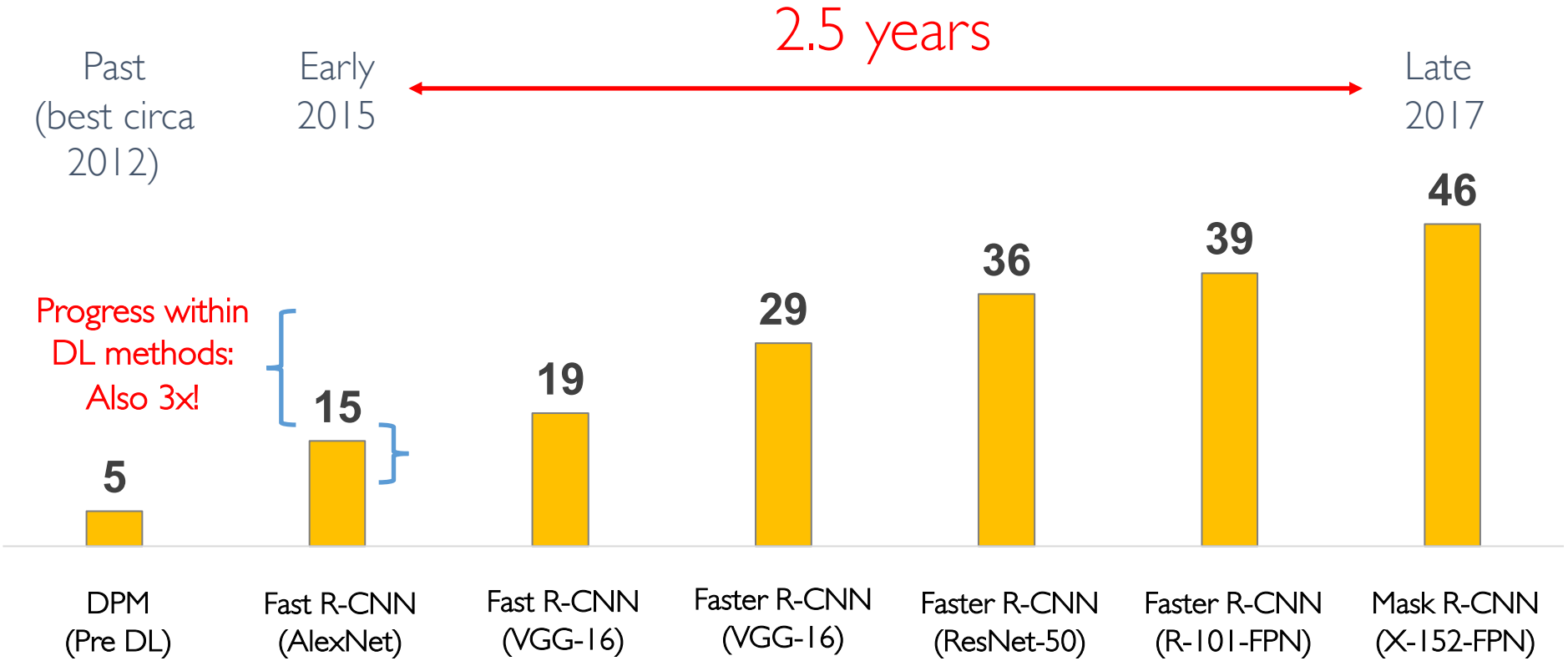
9 anchors for each position in feature maps

Insert Region Proposal Network (RPN) to predict proposals from features
Jointly train with 4 losses:

1. RPN classify object / not object
2. RPN regress box coordinates (similar to SS)
3. Final classification score (object classes)
4. Final box coordinates



COCO Object Detection (mAP)



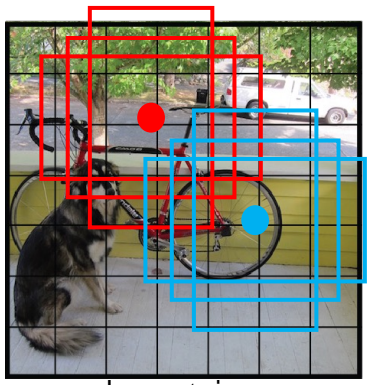
http://cocodataset.org/

Liu et al. Deep Learning for Generic Object Detection: A Survey. IJCV 2020.

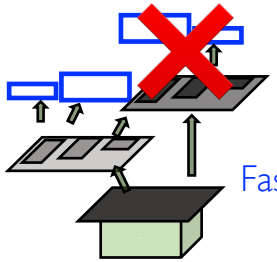
You Only Look Once (YOLO)

Class-specified RPN

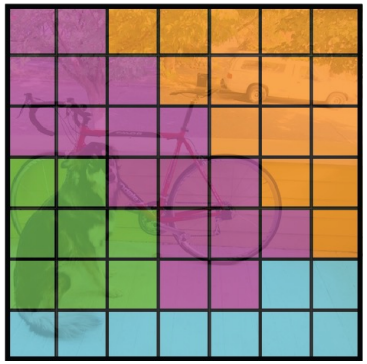
Divide image into grid 7 x 7.
Set a group of **anchors** centered at each grid cell



Input image
3 x H x W

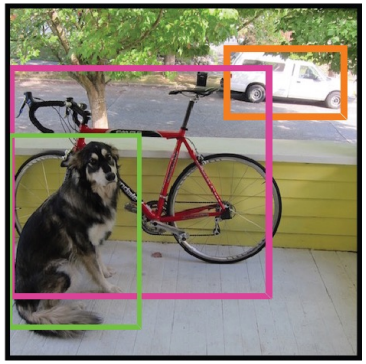


Faster R-CNN



Grid-wise classification

Regress from each of the *B* anchors to a final box with 5 numbers:
(dx, dy, dh, dw, box confidence)



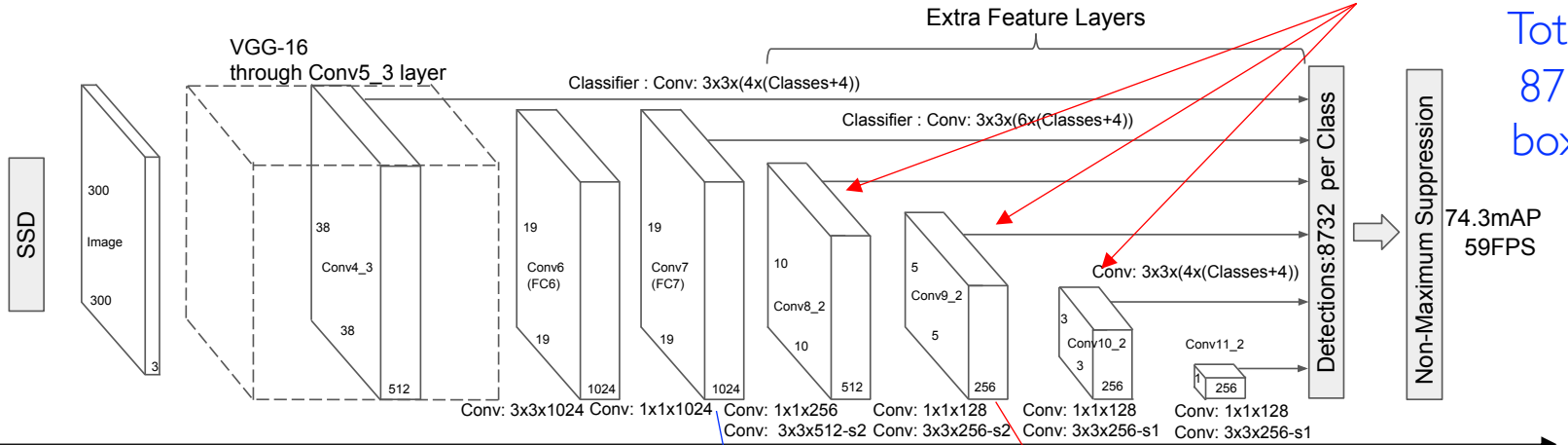
x = Final score

Predict scores for each of *C* classes (including background as a class)

Redmon et al, "You Only Look Once: Unified, Real-Time Object Detection", **CVPR** 2016

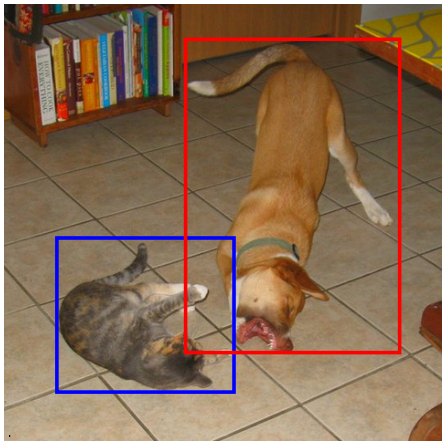
Single-Shot Detector (SSD)

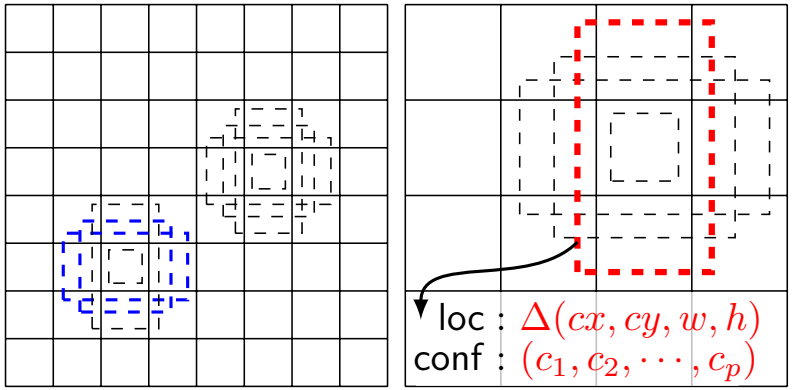
Class-specified RPN on features from multiple conv layers.



Totally 8732 boxes

Receptive fields grows, anchor size grows



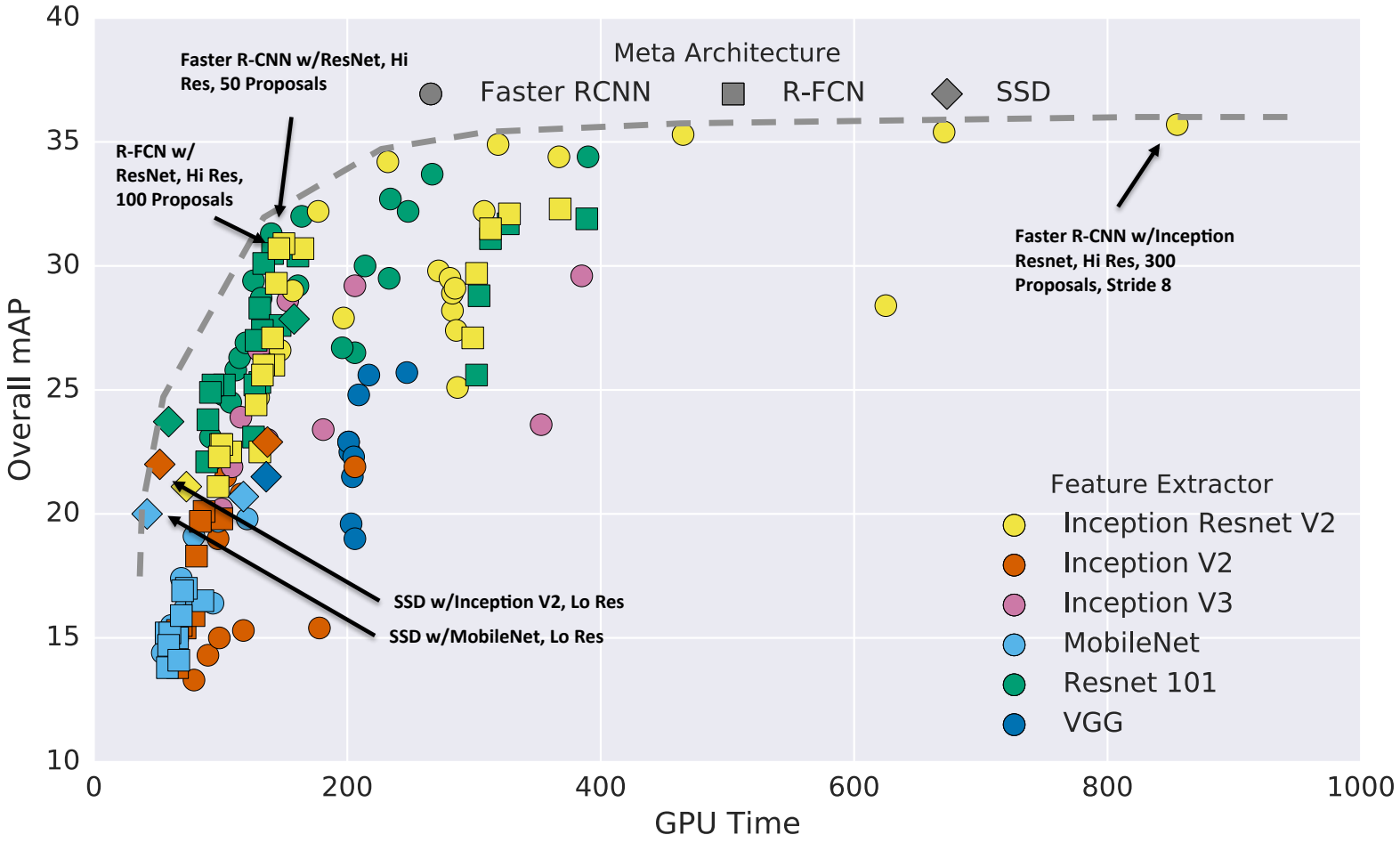


Smaller objects detected on lower layer

Larger objects detected on higher layer

Liu et al, "SSD: Single-Shot MultiBox Detector", **ECCV** 2016

Benchmarking

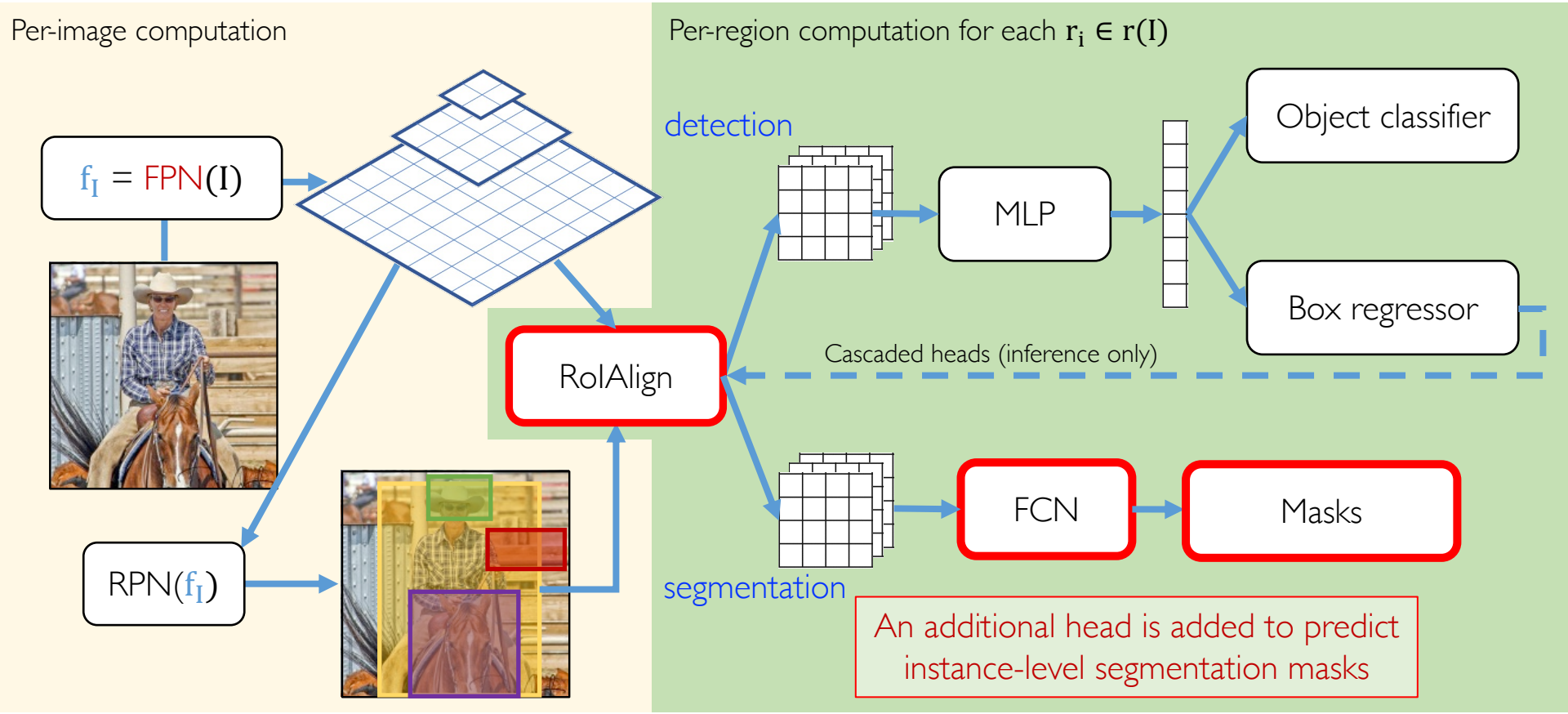


Huang et al. Speed/accuracy trade-offs for modern convolutional object detectors. CVPR 2017.

Instance Segmentation

He, Kaiming, et al. "Mask r-cnn." **ICCV** 2017. (Best Paper)

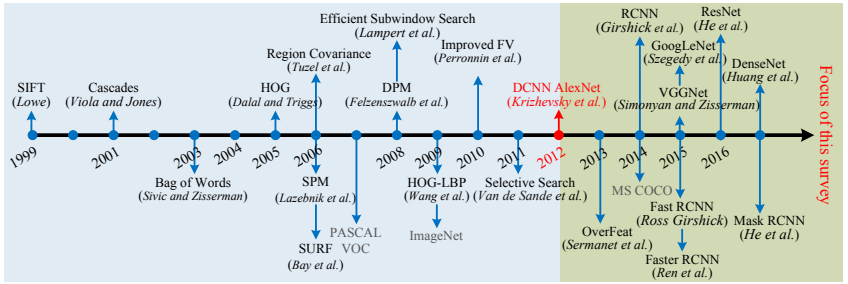
Mask R-CNN

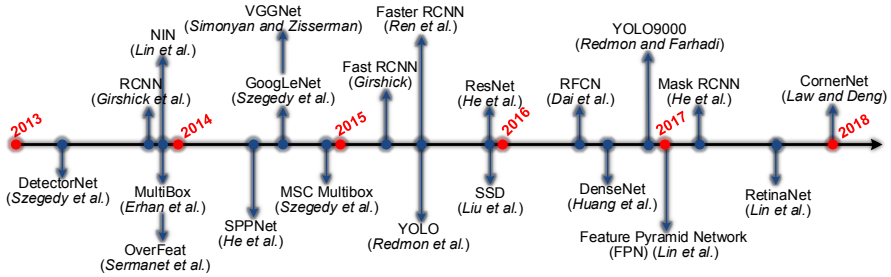


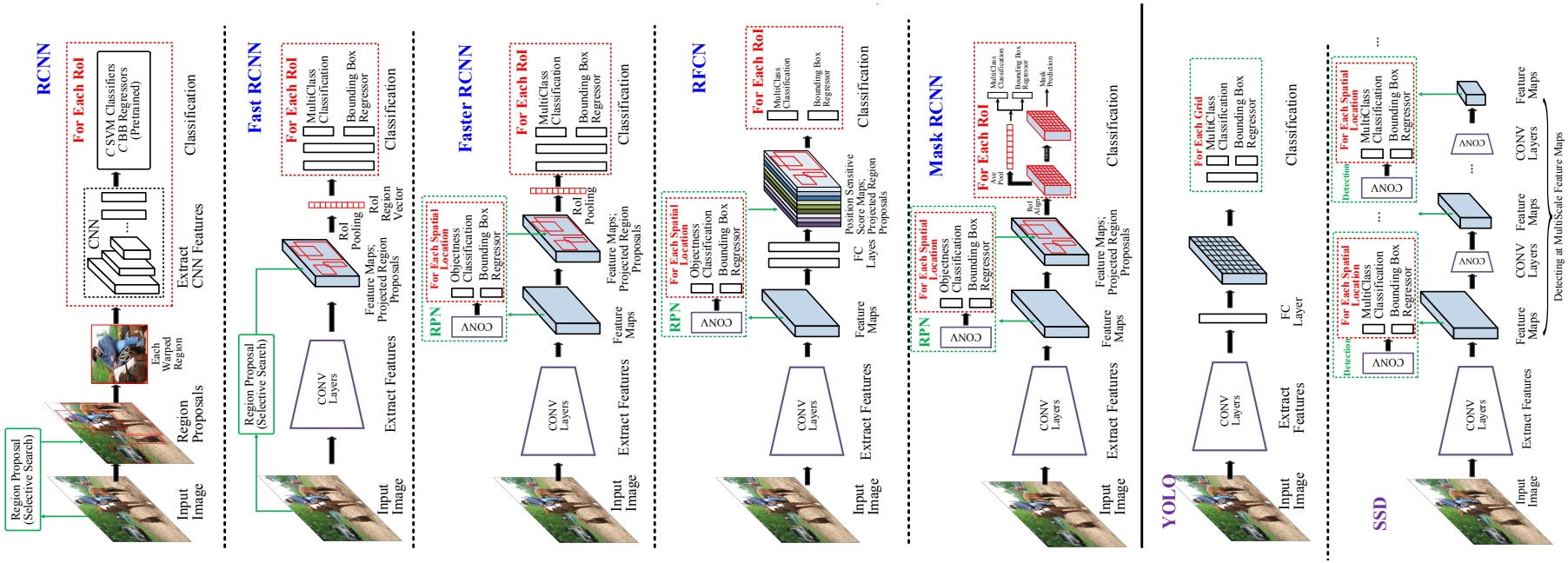
- A combination of semantic segmentation and object detection.
- Enhance detection result by multi-task learning.

He, Kaiming, et al. "Mask r-cnn." **ICCV** 2017. (Best Paper)

Object Detection in 20 Years







Liu et al. Deep Learning for Generic Object Detection: A Survey. IJCV 2020.

Outline

- **Image**
  - Recognition, segmentation, detection, **stylization**
- Video:
  - Recognition, detection
- 3D Vision:
  - Volumetric data

Neural Style Transfer

Content Image

This image is licensed under CC-BY 3.0

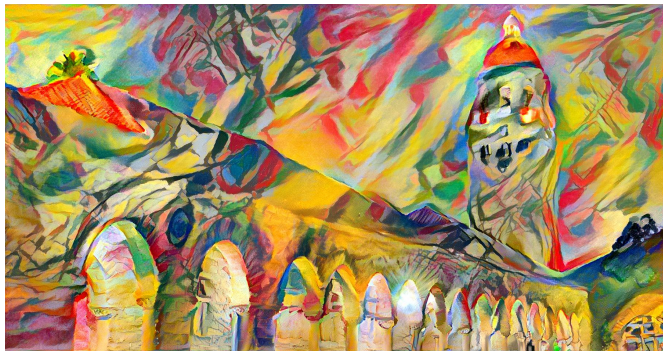
+

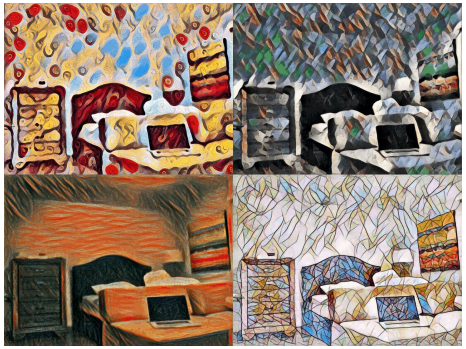
Style Image

Starry Night by Van Gogh

=

Style Transfer

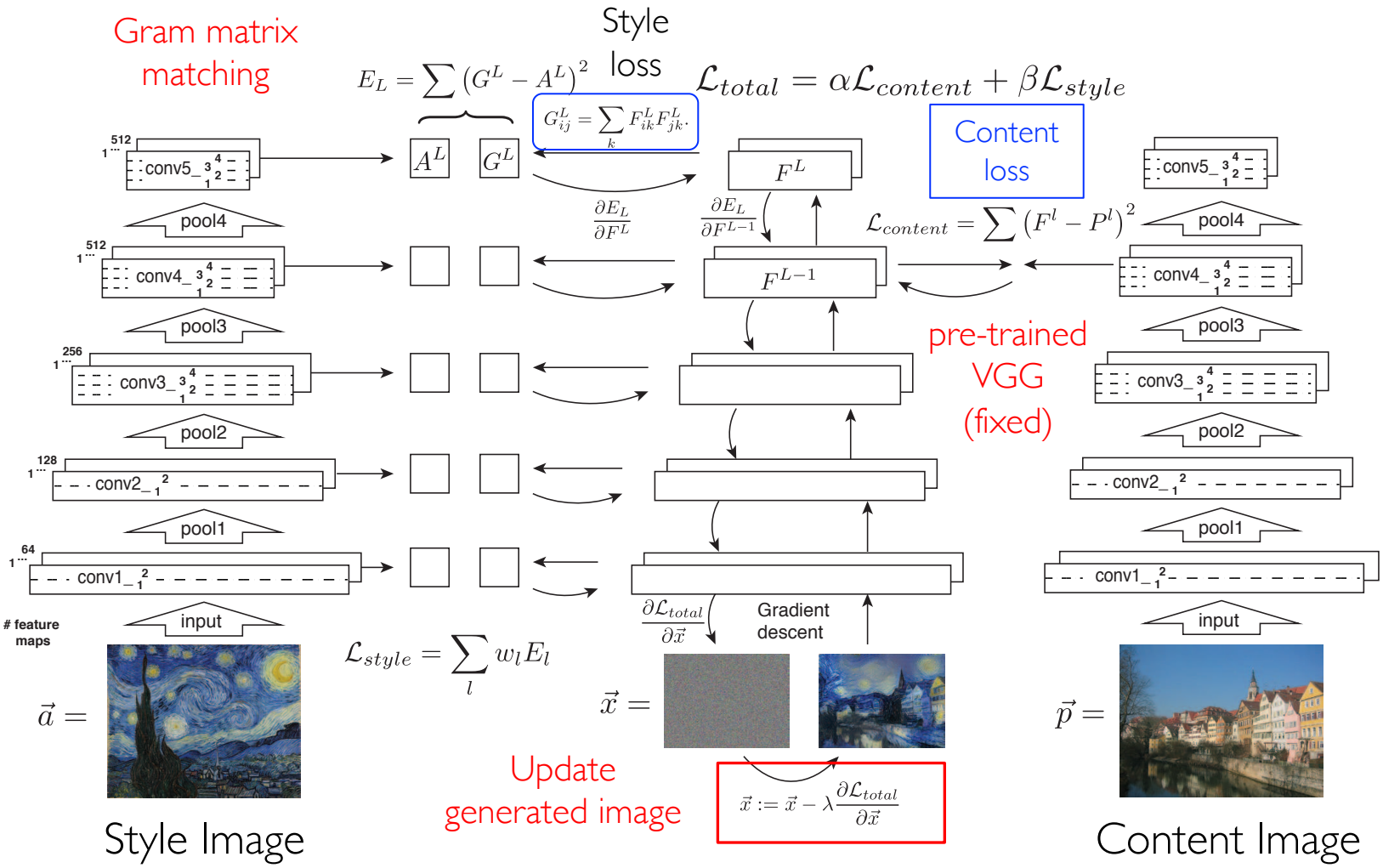




https://github.com/jcjohnson/fast-neural-style

L A. Gatys et al. Image Style Transfer Using Convolutional Neural Networks. CVPR 2016.

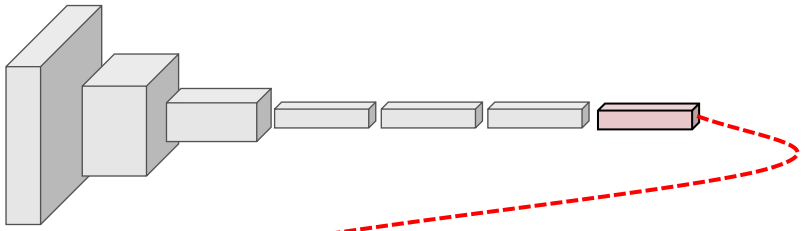
Neural Style Transfer



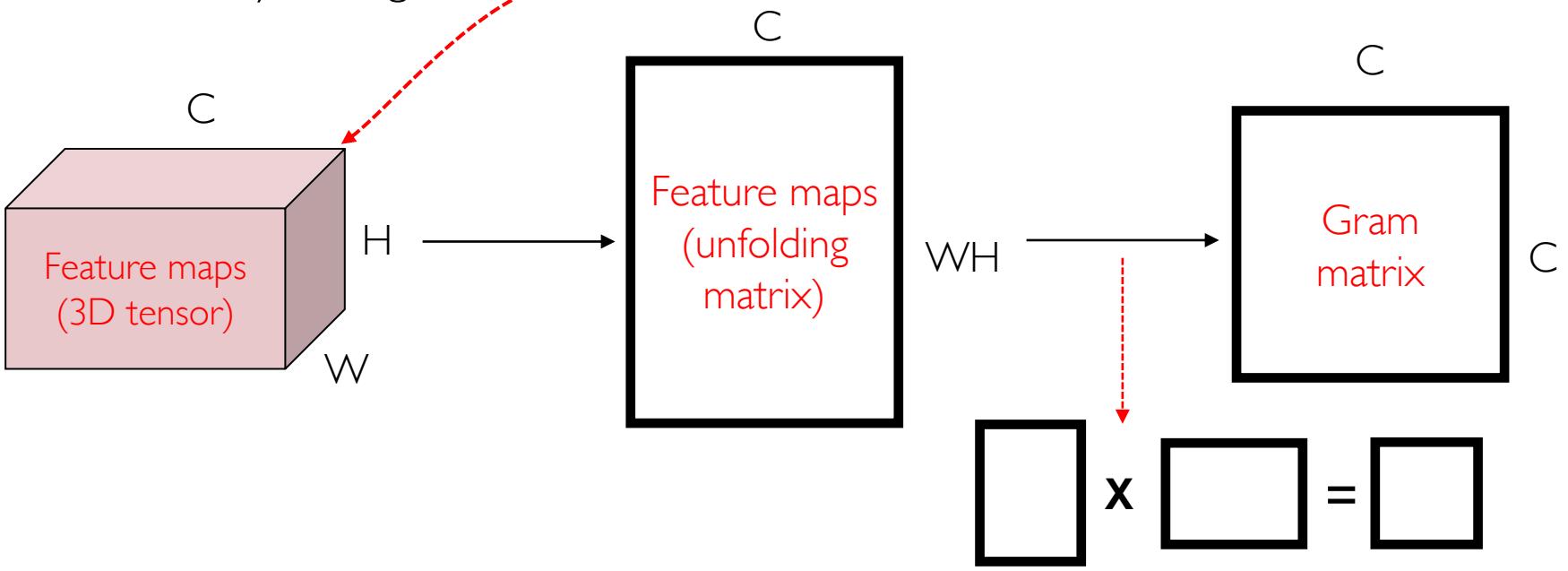
L A. Gatys et al. Image Style Transfer Using Convolutional Neural Networks. **CVPR** 2016.

Neural Style Transfer

Style Image



Gram matrix computation

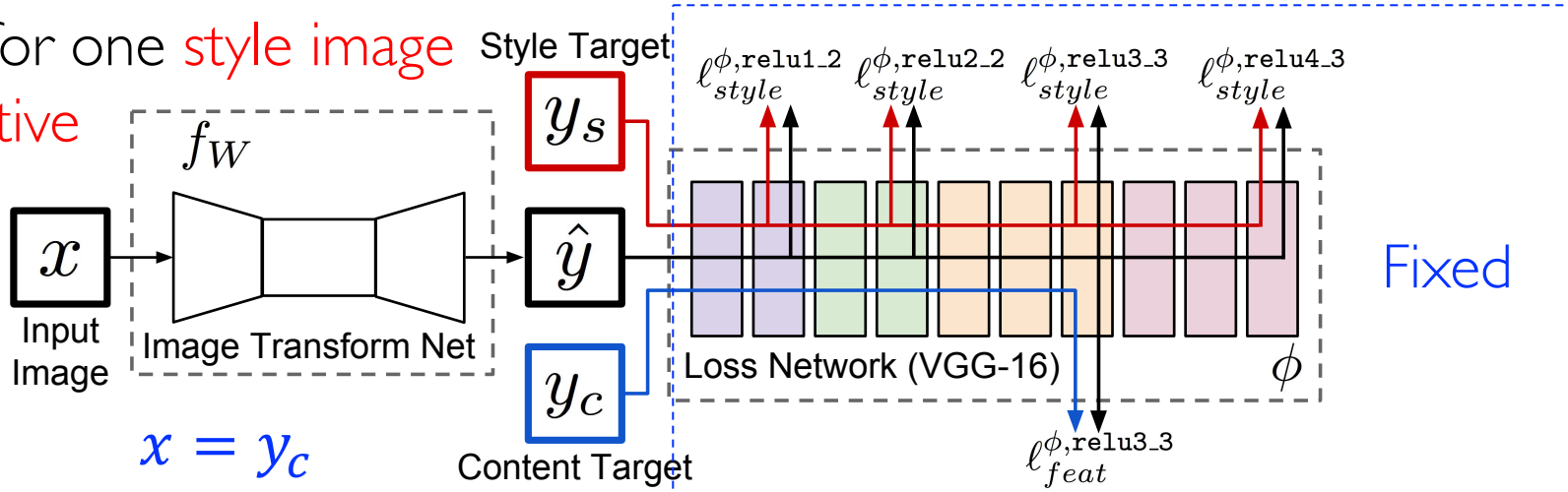


L A. Gatys et al. Image Style Transfer Using Convolutional Neural Networks. CVPR 2016.

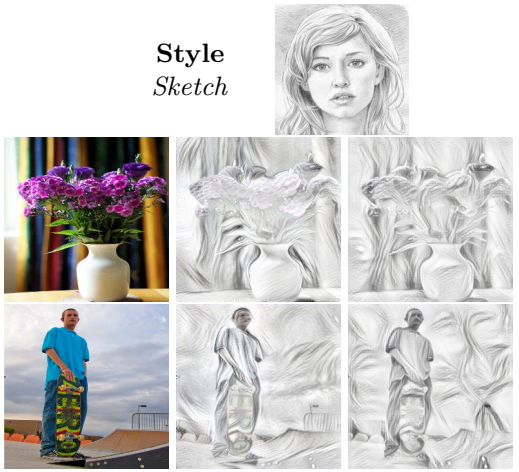
Real-Time Neural Style Transfer

One net for one style image

Not adaptive



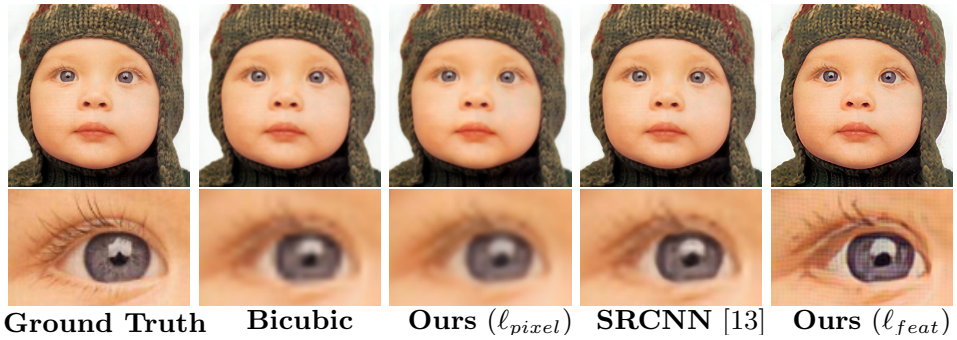
- Train an image transformation network to transform input images into output images instead of directly updating the image.
  - Content loss $\ell^{\phi,j}_{feat}(\hat{y}, y) = \frac{1}{C_j H_j W_j} \|\phi_j(\hat{y}) - \phi_j(y)\|_2^2$
  - Style loss $\ell^{\phi,j}_{style}(\hat{y}, y) = \|G^\phi_j(\hat{y}) - G^\phi_j(y)\|_F^2.$
  - Pixel-level loss $\ell_{pixel}(\hat{y}, y) = \|\hat{y} - y\|_2^2 / CHW.$



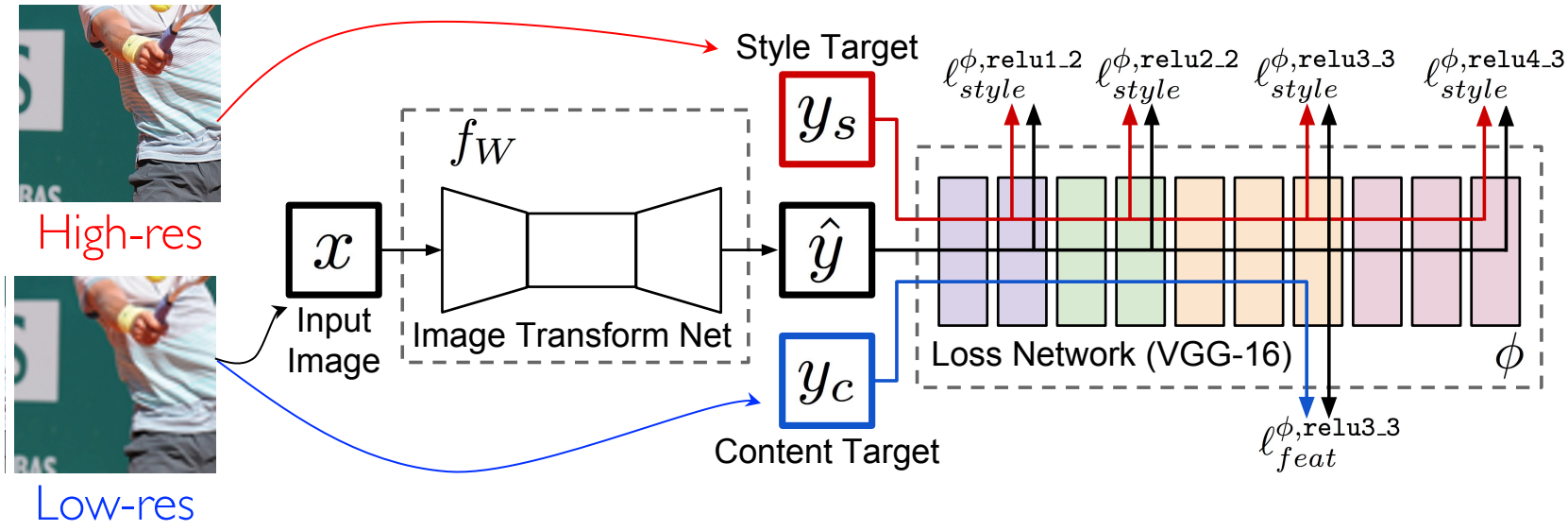
Johnson, Justin, et al. "Perceptual losses for real-time style transfer and super-resolution." ECCV, 2016.

Super-Resolution As Style Transfer

- This framework can also be used for image resolution.



Paired training data



Johnson, Justin, et al. "Perceptual losses for real-time style transfer and super-resolution." ECCV, 2016.

Outline

- Image
  - Recognition, segmentation, detection, stylization
- **Video:**
  - **Recognition, detection**
- 3D Vision:
  - Volumetric data

Video Recognition

How is it different from image recognition?

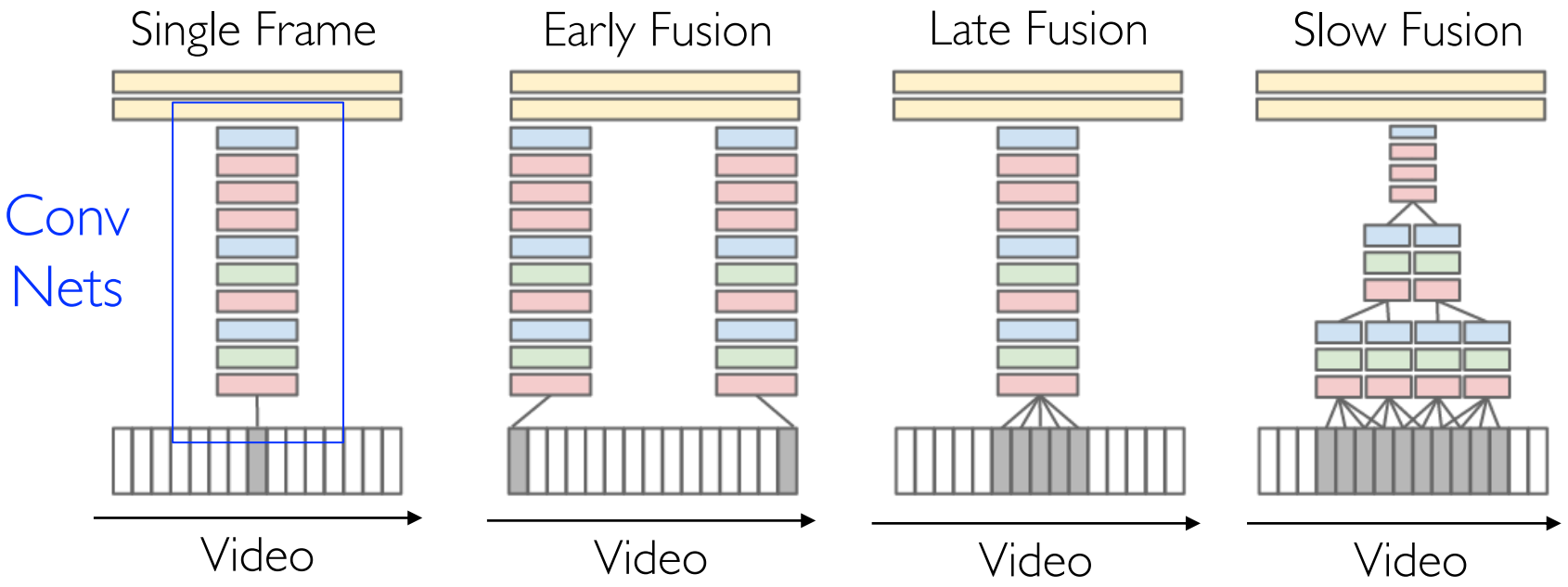
Dog

Motion

Basketball

Temporal Fusion

- How to aggregate visual information across temporal domain?
- A first attempt:



- Weak understanding on motions! (How about self-attention?)

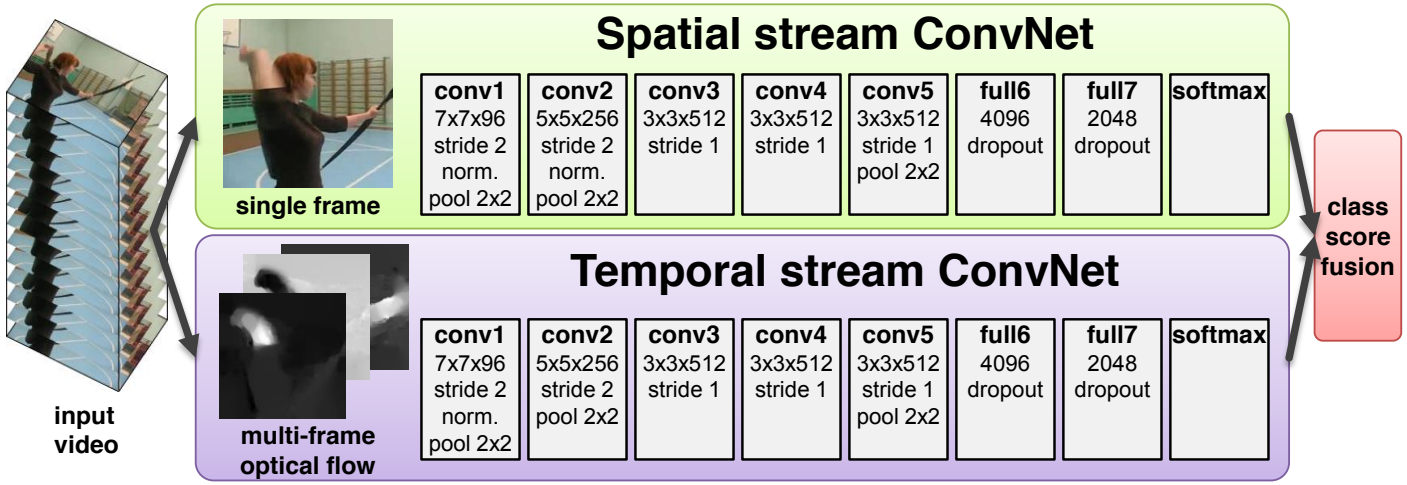
A. Kaparthy et al. Large-scale Video Classification with Convolutional Neural Networks. **CVPR** 2014.

Two-Stream Convolutional Networks

Model ensemble: spatial network + temporal network

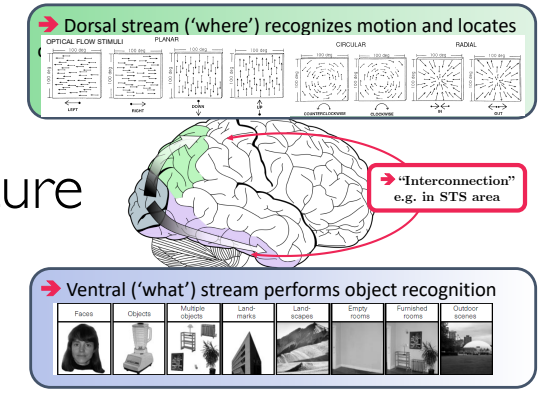
RGB

Optical Flow



Motivation:

Separate visual pathways in nature

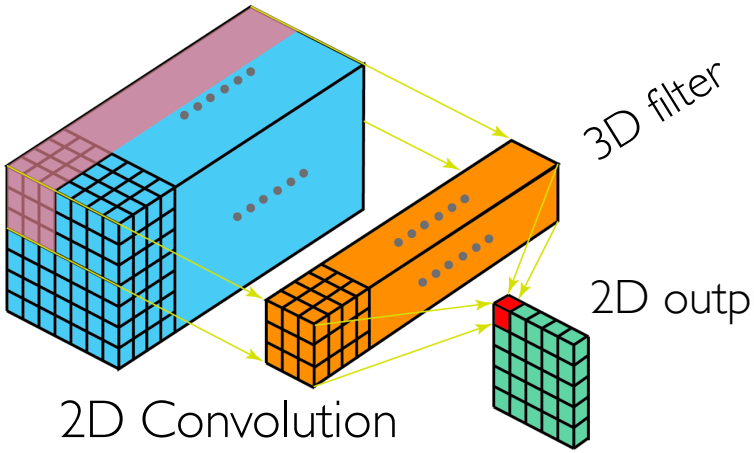


https://en.wikipedia.org/wiki/Optical_flow

FlowNet: Learning Optical Flow With Convolutional Networks. CVPR 2015.

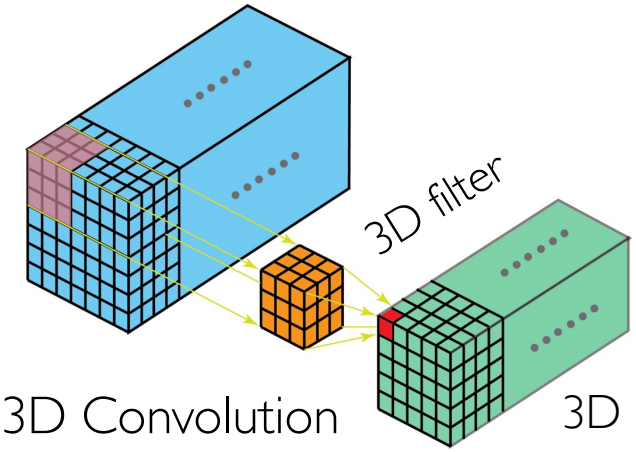
K. Simonyan et al. Two-Stream Convolutional Networks for Action Recognition in Videos. NIPS 2014.

Recap: 3D Convolution



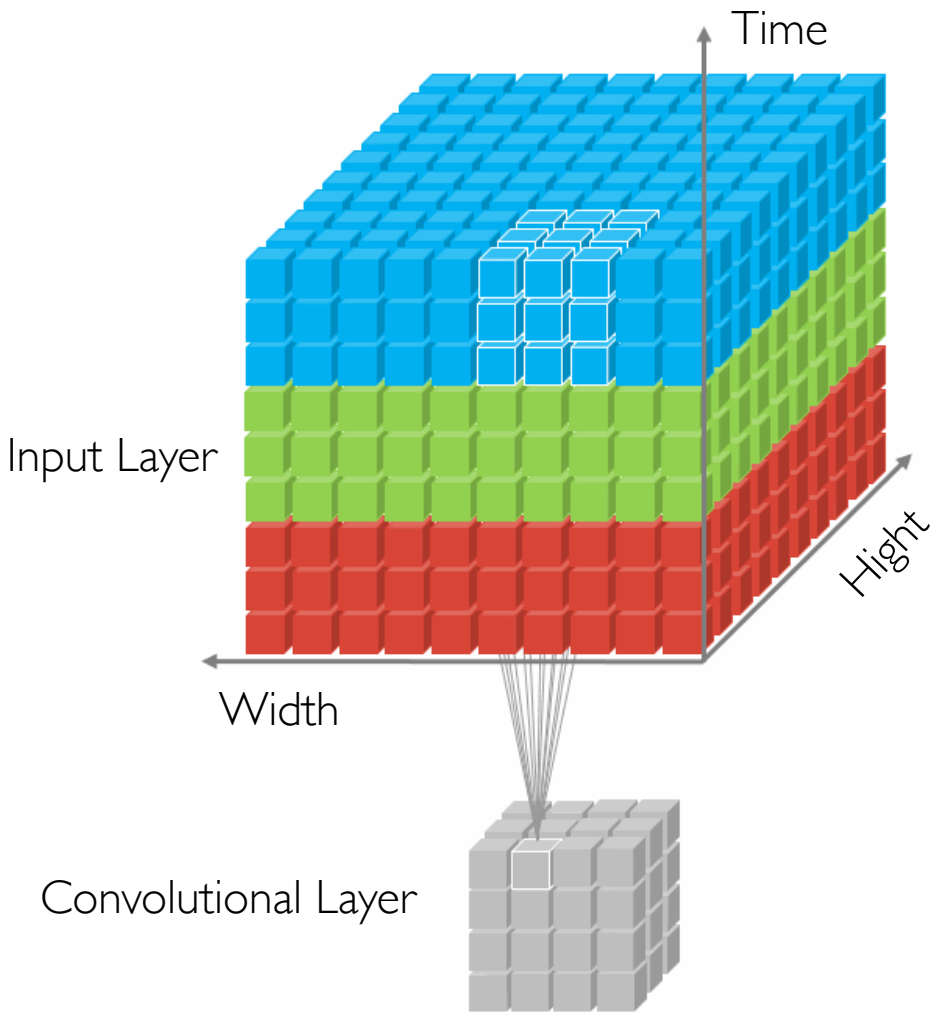
2D Convolution

Sliding over W × H



3D Convolution

Sliding over W × H × T

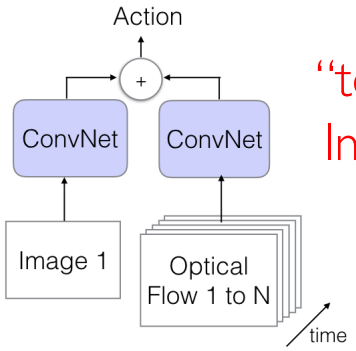


Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh: Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? **CVPR** 2018: 6546-6555

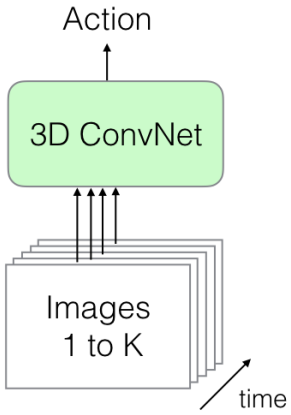
Inflated 3D Convolution (I3D)

RGB

Optical Flow

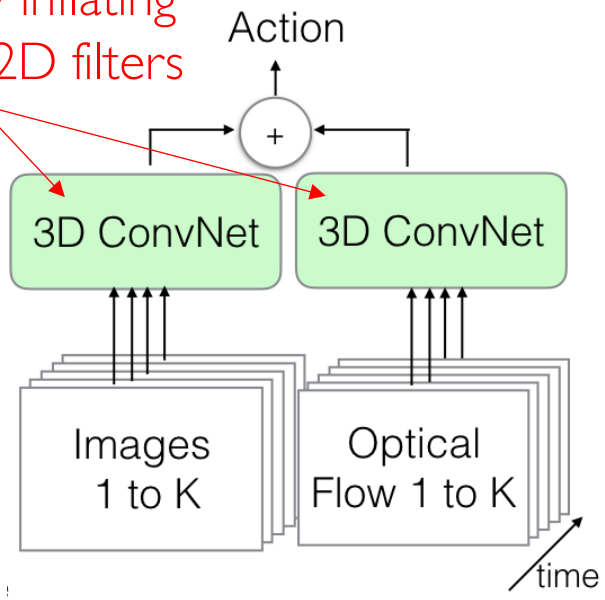


Two-stream



3D-Conv

Initialized by
"temporally inflating"
ImageNet 2D filters



Two-stream 3D
ConvNets

better results for Video

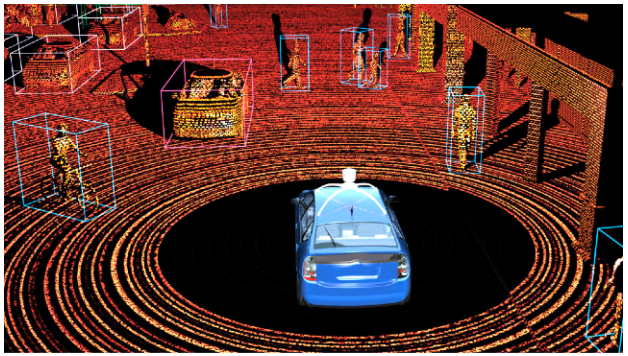
Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. **CVPR** 2017.

Outline

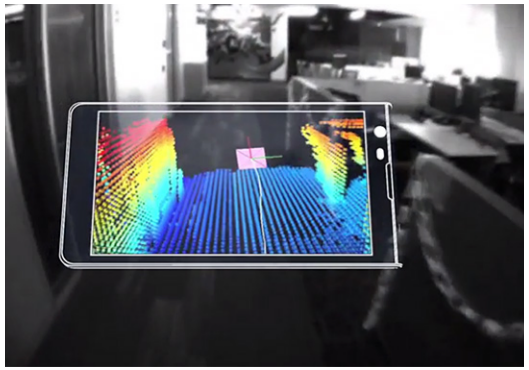
- Image
  - Recognition, segmentation, detection, stylization
- Video:
  - Recognition, detection
- **3D Vision:**
  - **Volumetric data**

3D Vision

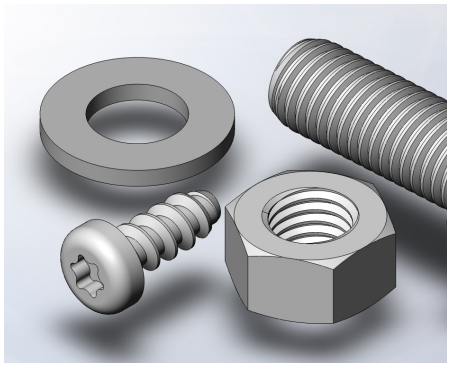
- Why we need to analysis 3D data?



Robot Perception



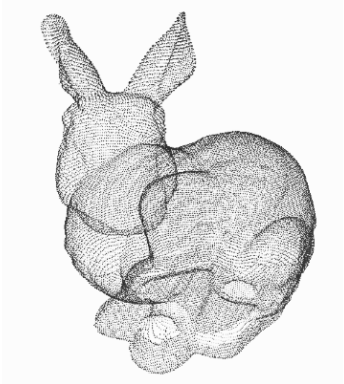
Augmented Reality



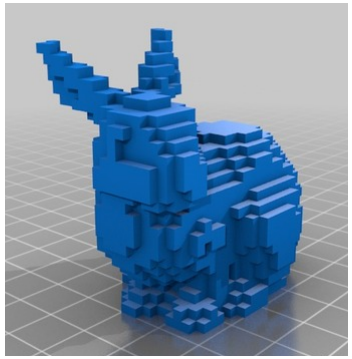
Shape Design

- Different forms of 3D data:

3D surface description by a set of points



Point Cloud

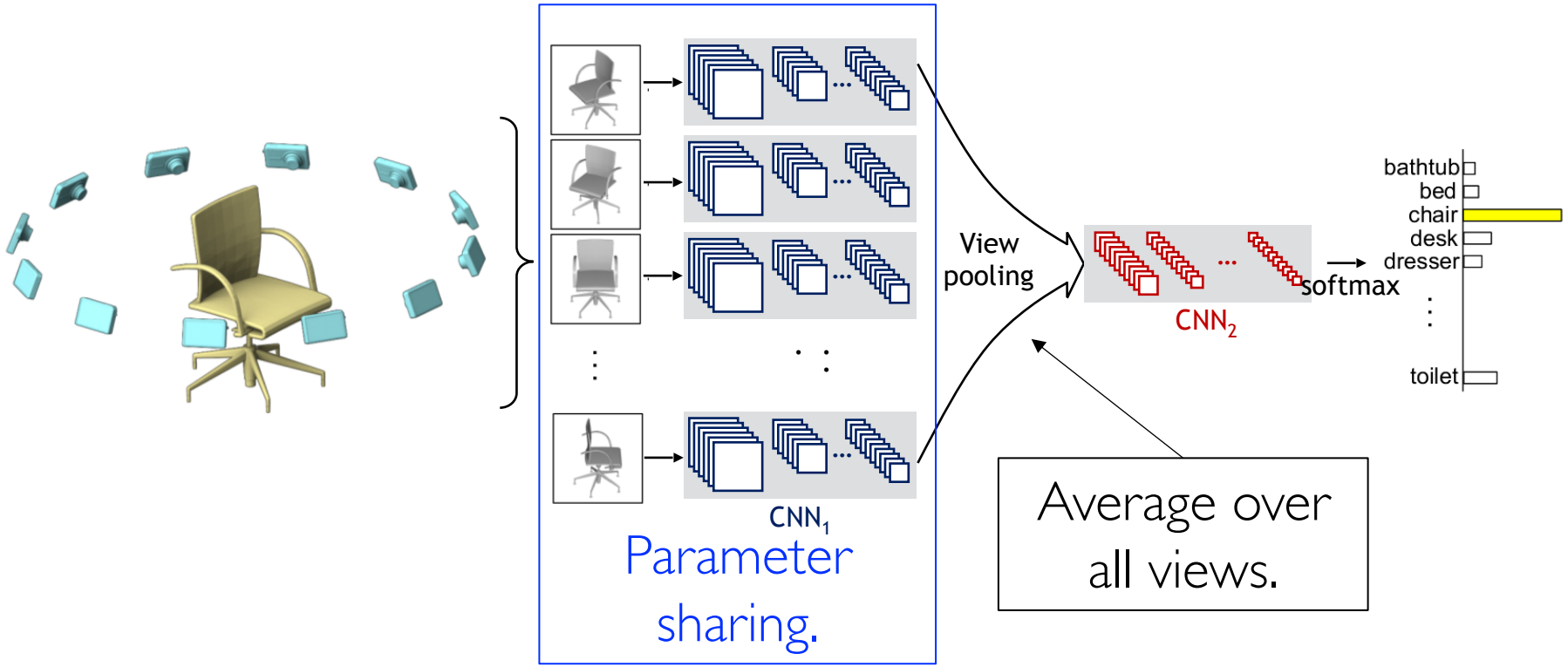


Volumetric

Projected View

Multiview Data

- How to extract or learn features from Multiview 3D data?
  - Aggregate features of every views.
  - With convolution networks.

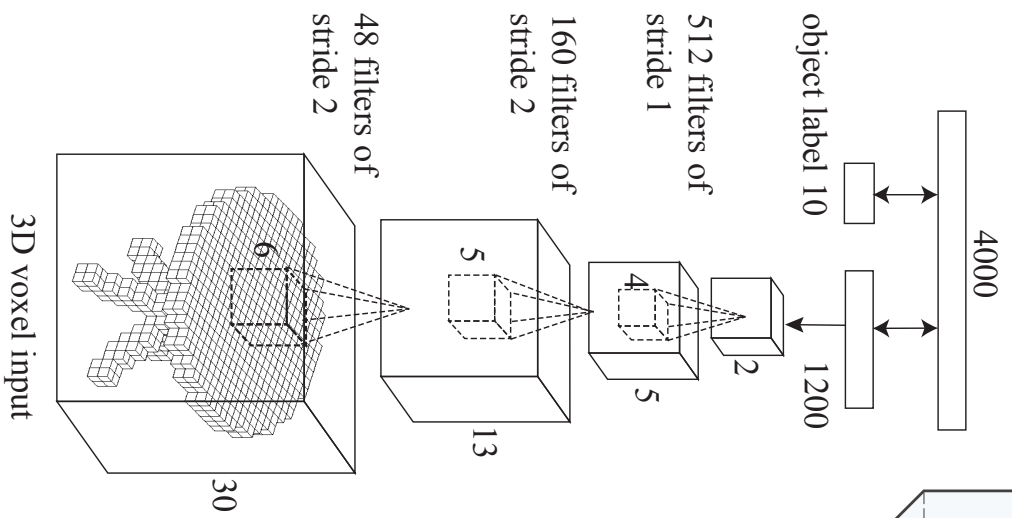


Su, Hang, et al. Multi-View Convolutional Neural Networks for 3D Shape Recognition. **ICCV** 2015.

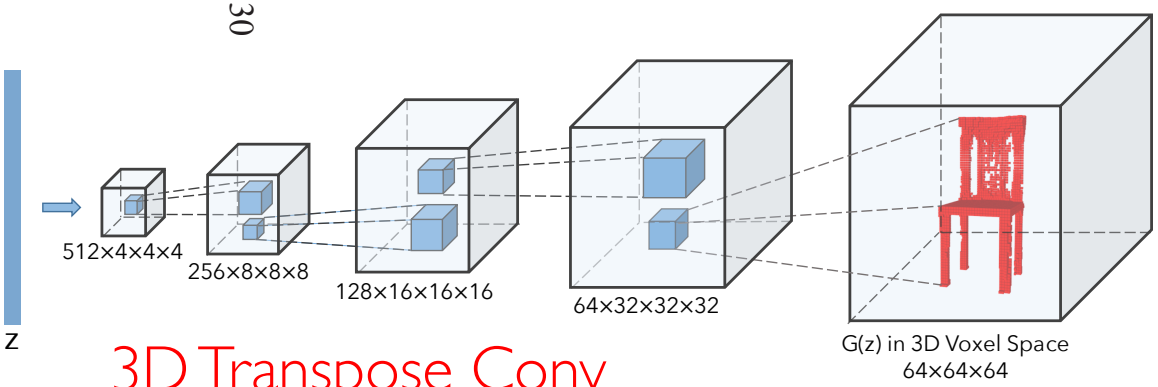
Volumetric Data

- How to extract features from or reconstruct Volumetric 3D data?
  - Using 3D conv!

Feature Encoder:



Reconstruction Decoder:



Wu, Zhirong, et al. 3D ShapeNets: A Deep Representation for Volumetric Shapes. **CVPR** 2015.