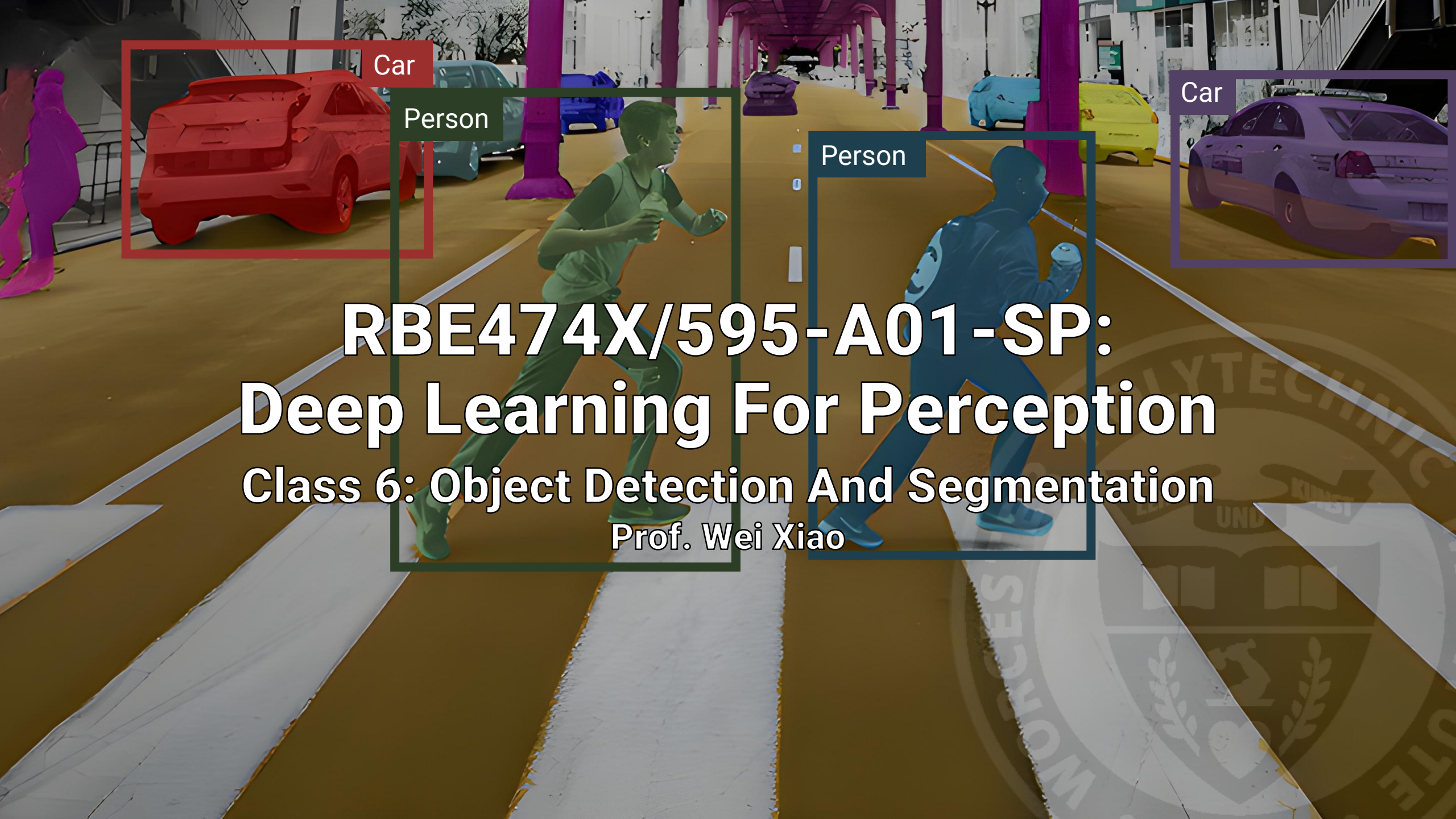


RBE474X/595-A01-SP: Deep Learning For Perception

Class 6: Object Detection And Segmentation

Prof. Wei Xiao

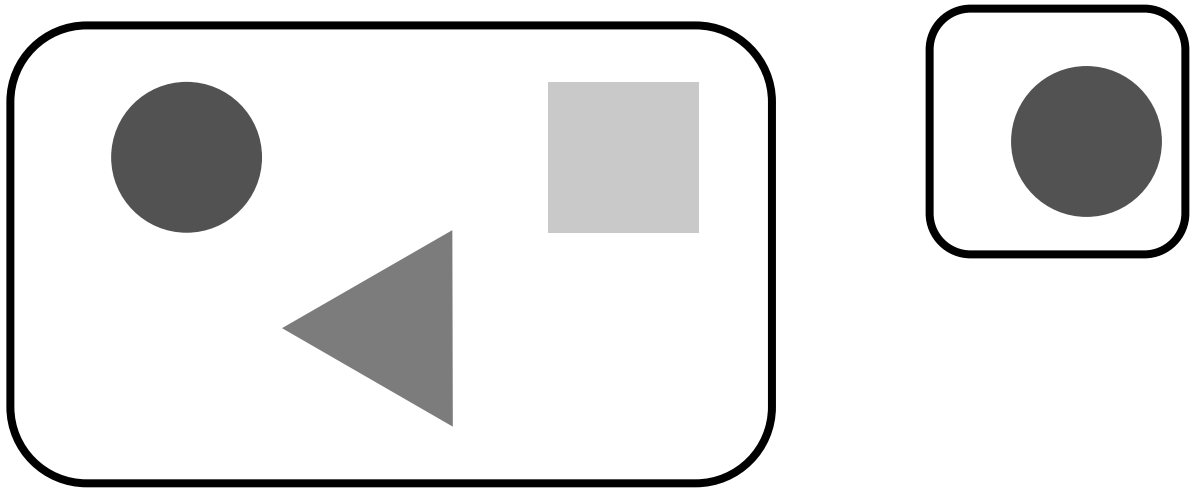
Recall Classification

What Does A Classifier Assume?

Input is crop of **One Object**!

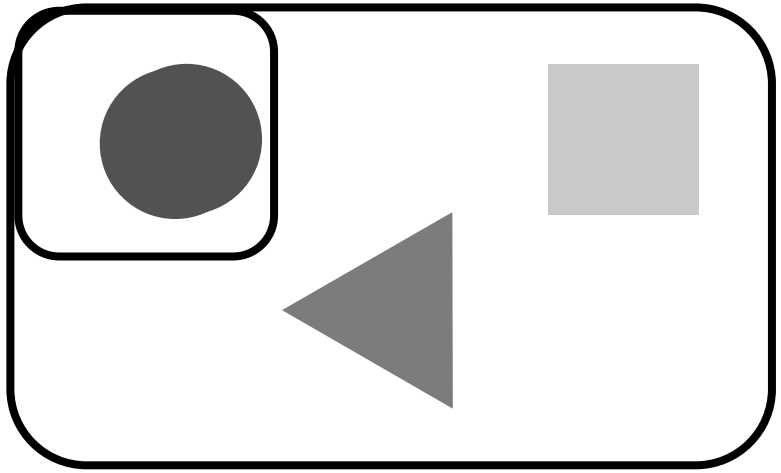
Naïve Way To Detect Objects?

Sliding Window Template Matching



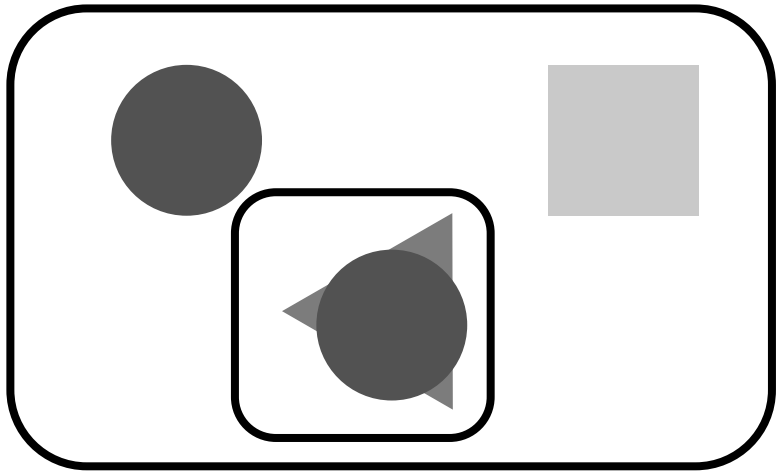
Naïve Way To Detect Objects?

Sliding Window Template Matching

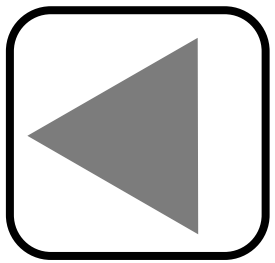
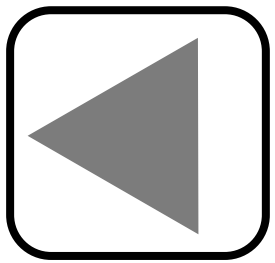


Naïve Way To Detect Objects?

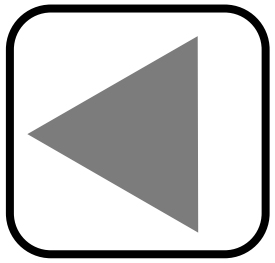
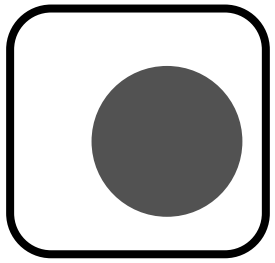
Sliding Window Template Matching



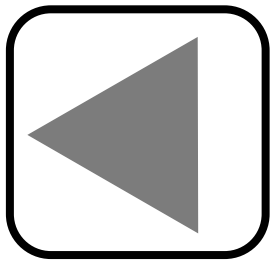
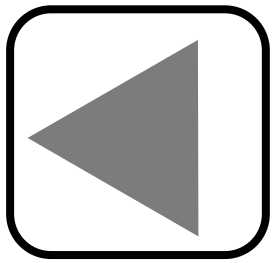
How Do We Match Patches?

$f($  ,  $)$

I_1 I_2

$f($  ,  $)$

Large Value! ⇒ Mis-match or bad match!

$f($  ,  $)$

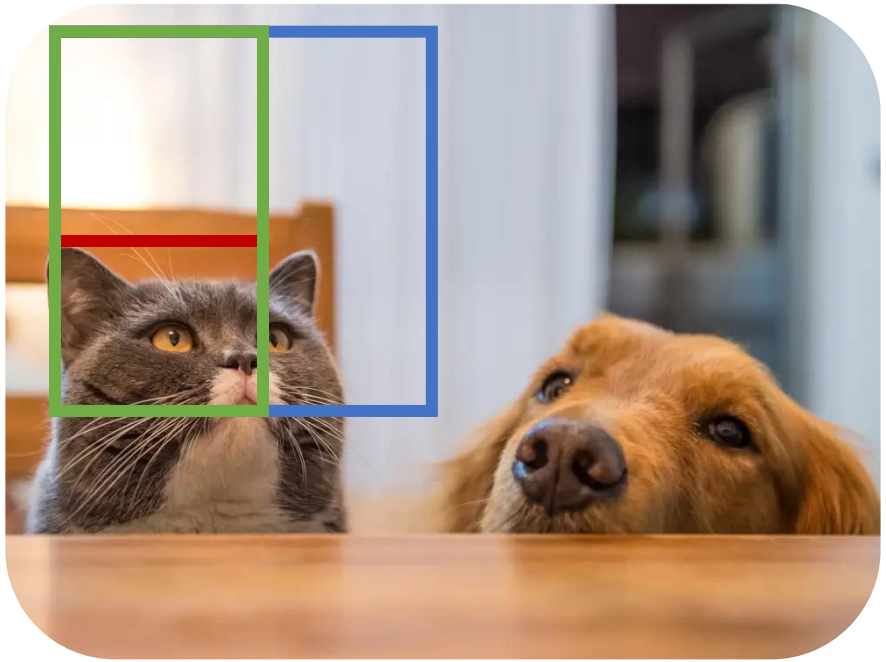
Small Value! ⇒ Amazing match!

Or vice-versa!

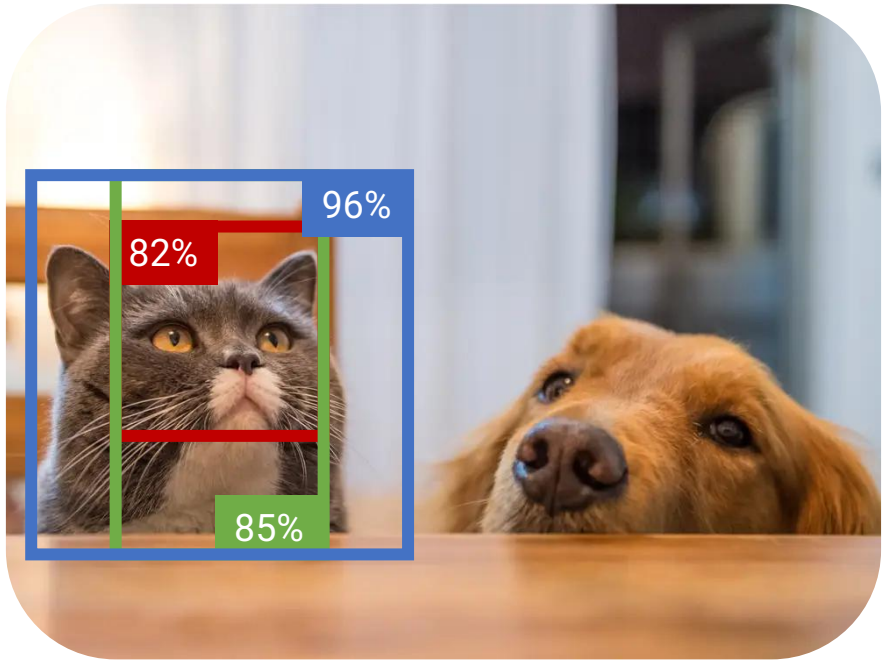
How Do You Choose Scale?

Non-max Suppression

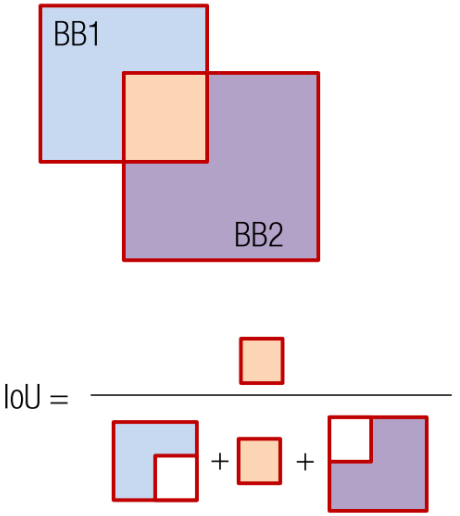
1. Select BB with highest OS
2. Compare IoU with other boxes
3. Remove boxes with IoU > 50%
4. Repeat for next box



Bounding Box (BB)



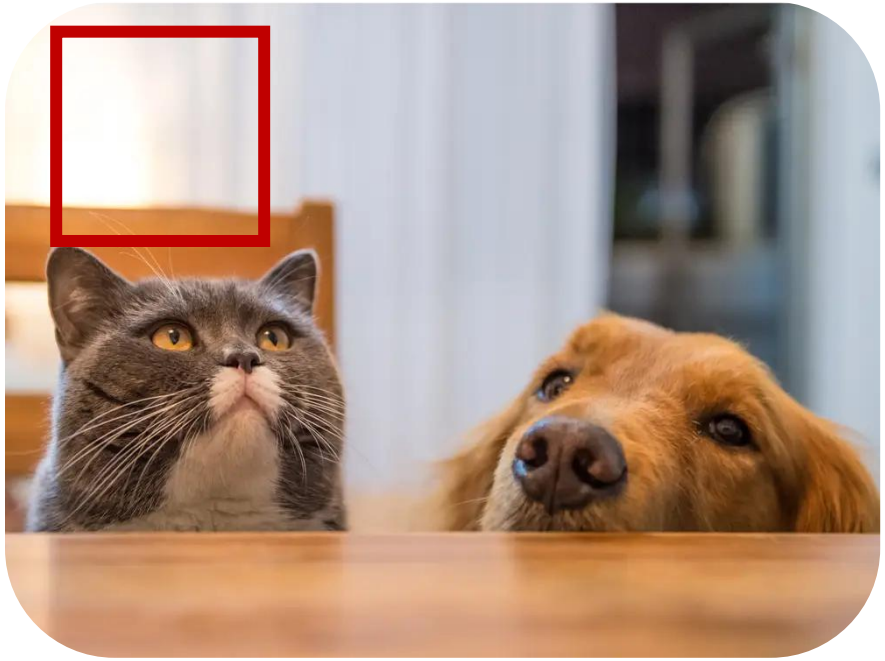
"Objectiveness Score" (OS)



Intersection over Union (IoU)

Era Of Deep Learning

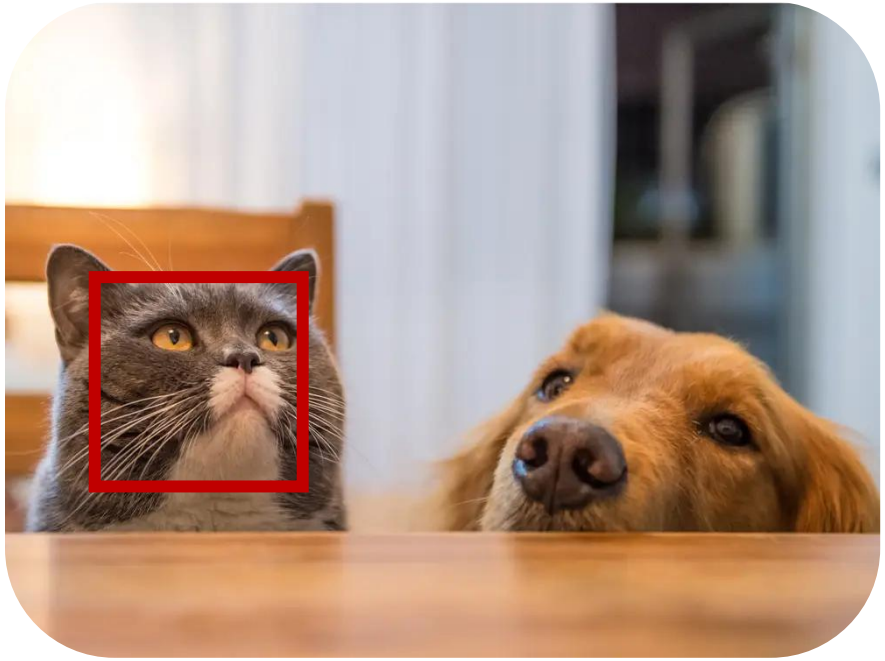
Class Of Region Proposal Networks



Dog? No
Cat? No
Background? Yes

Era Of Deep Learning

Class Of Region Proposal Networks

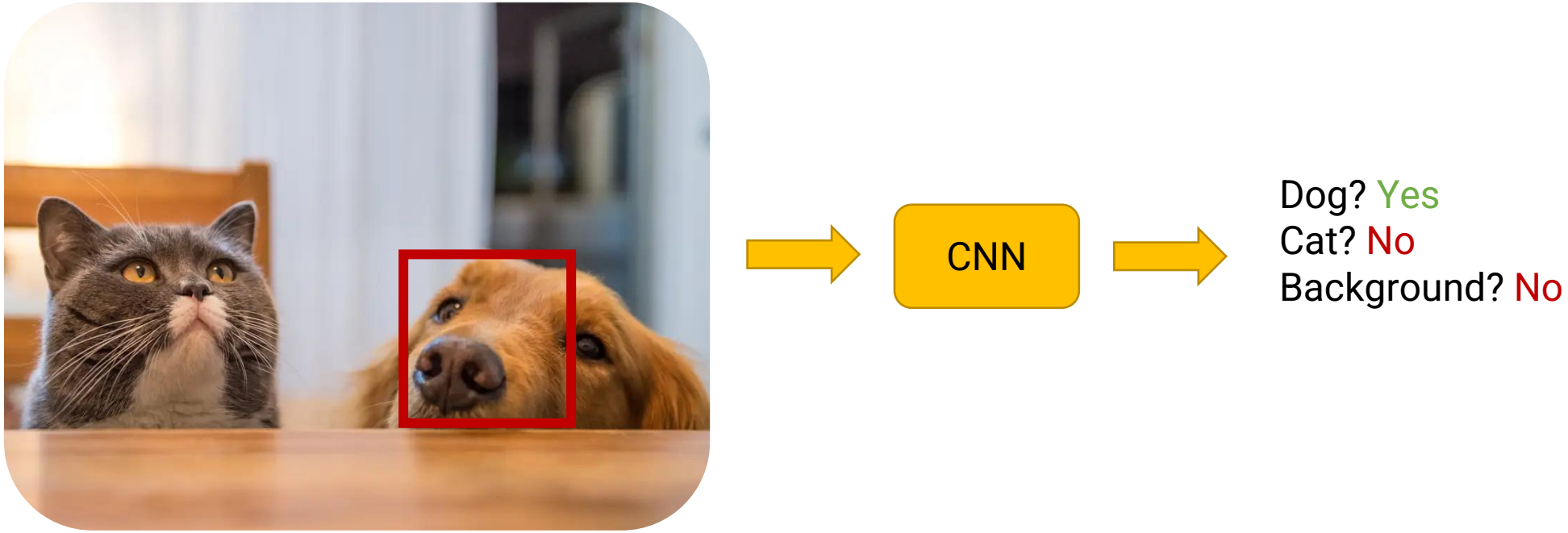


CNN

Dog? No
Cat? Yes
Background? No

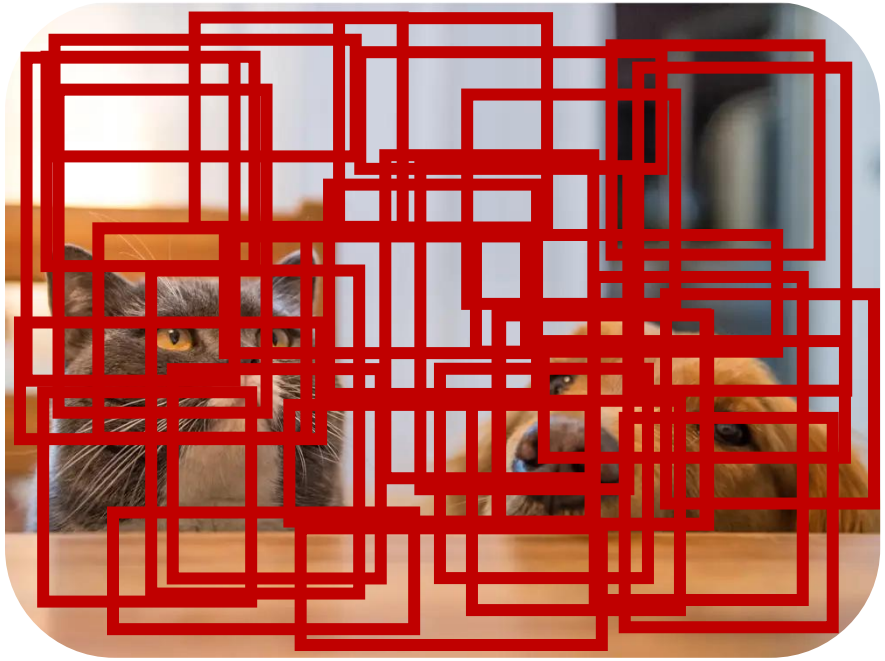
Era Of Deep Learning

Class Of Region Proposal Networks



Era Of Deep Learning

Class Of Region Proposal Networks



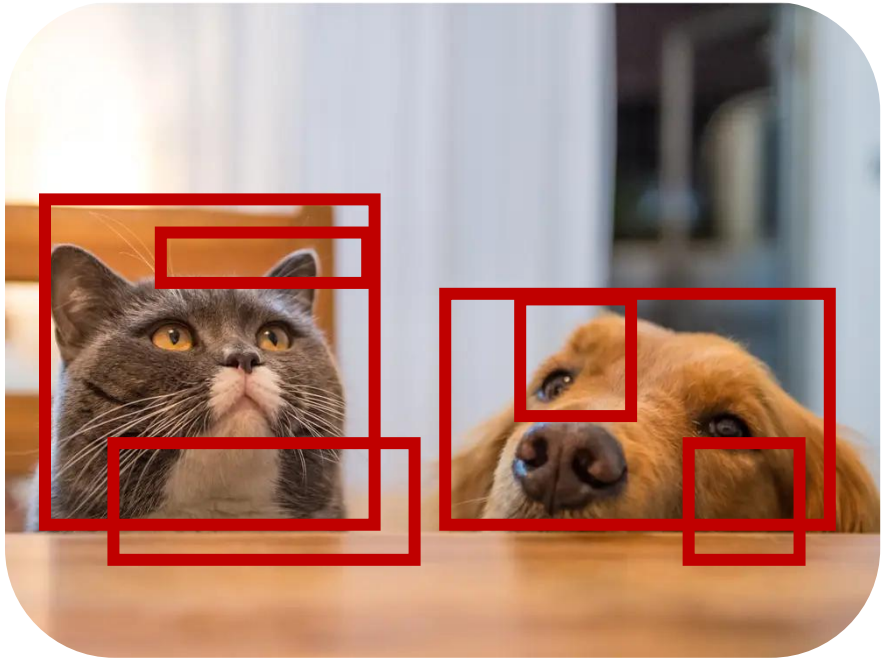
Issues?

Need to apply CNN for a variety of locations, scales, aspect ratios!

So slow ☹

Blobs

- Find blobs (Binary Large Objects) that are likely to contain objects!
- Can use superpixels and other methods!



Input Image | Initial Segmentation | After some iterations | After more iterations

R(Region)-CNN

Ross Girshick

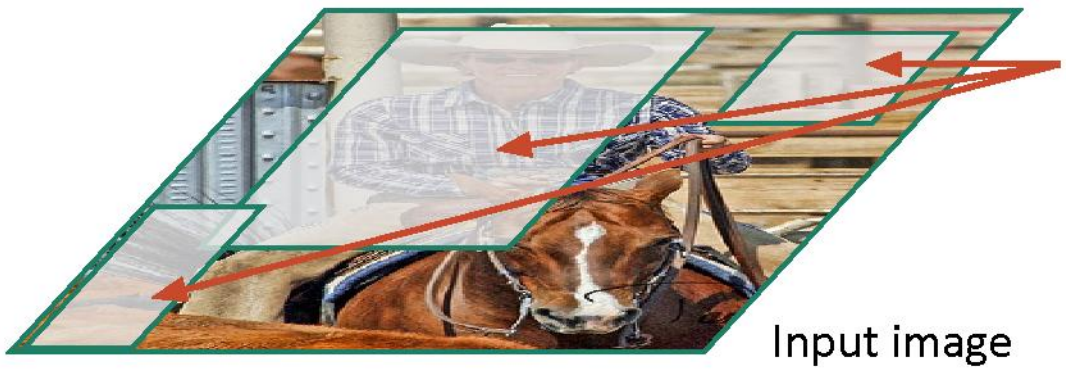
Jitendra Malik

Input image

Slides adapted from Stanford's CS231n

Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

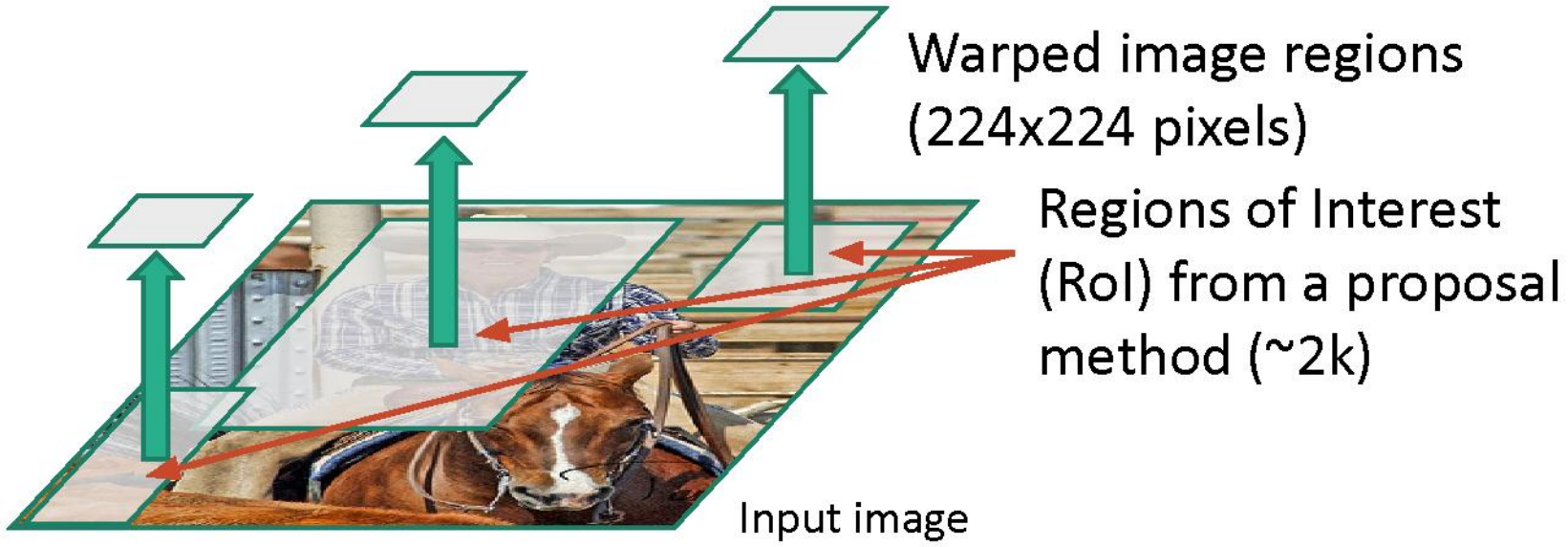
R-CNN



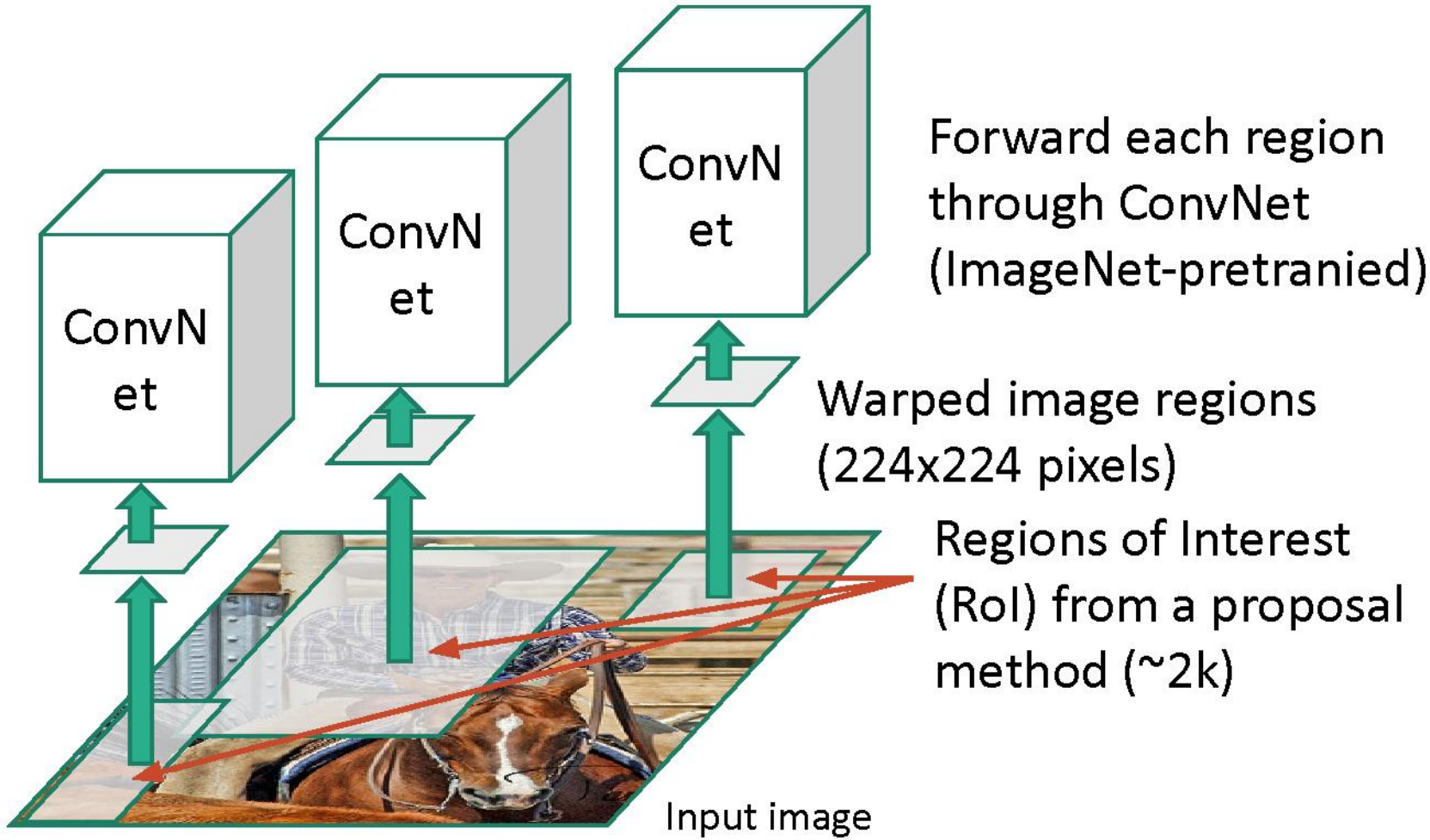
Input image

Regions of Interest (RoI) from a proposal method (~2k)

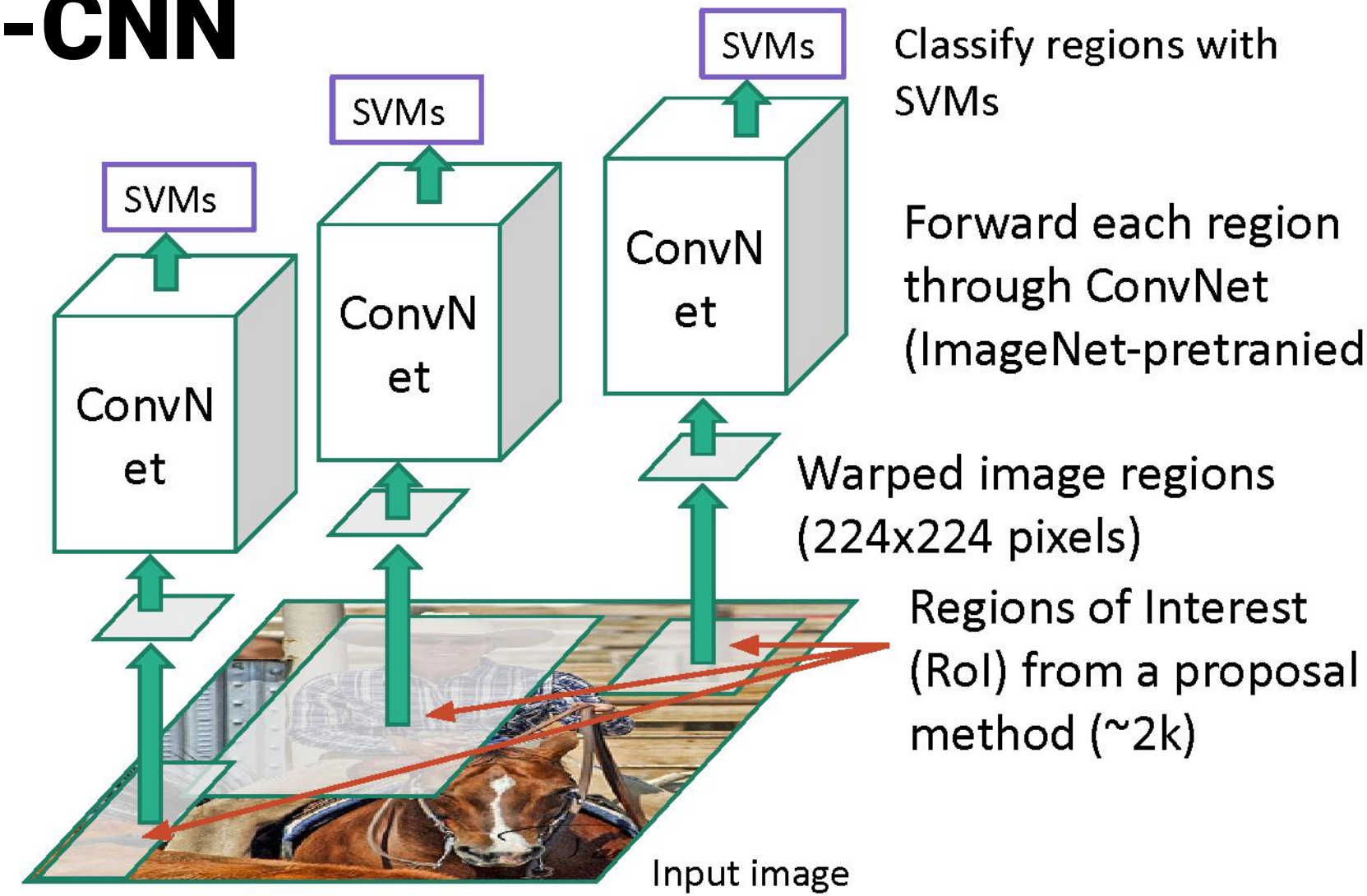
R-CNN


Warped image regions
(224x224 pixels)
Regions of Interest
(RoI) from a proposal
method (~2k)
Input image

R-CNN

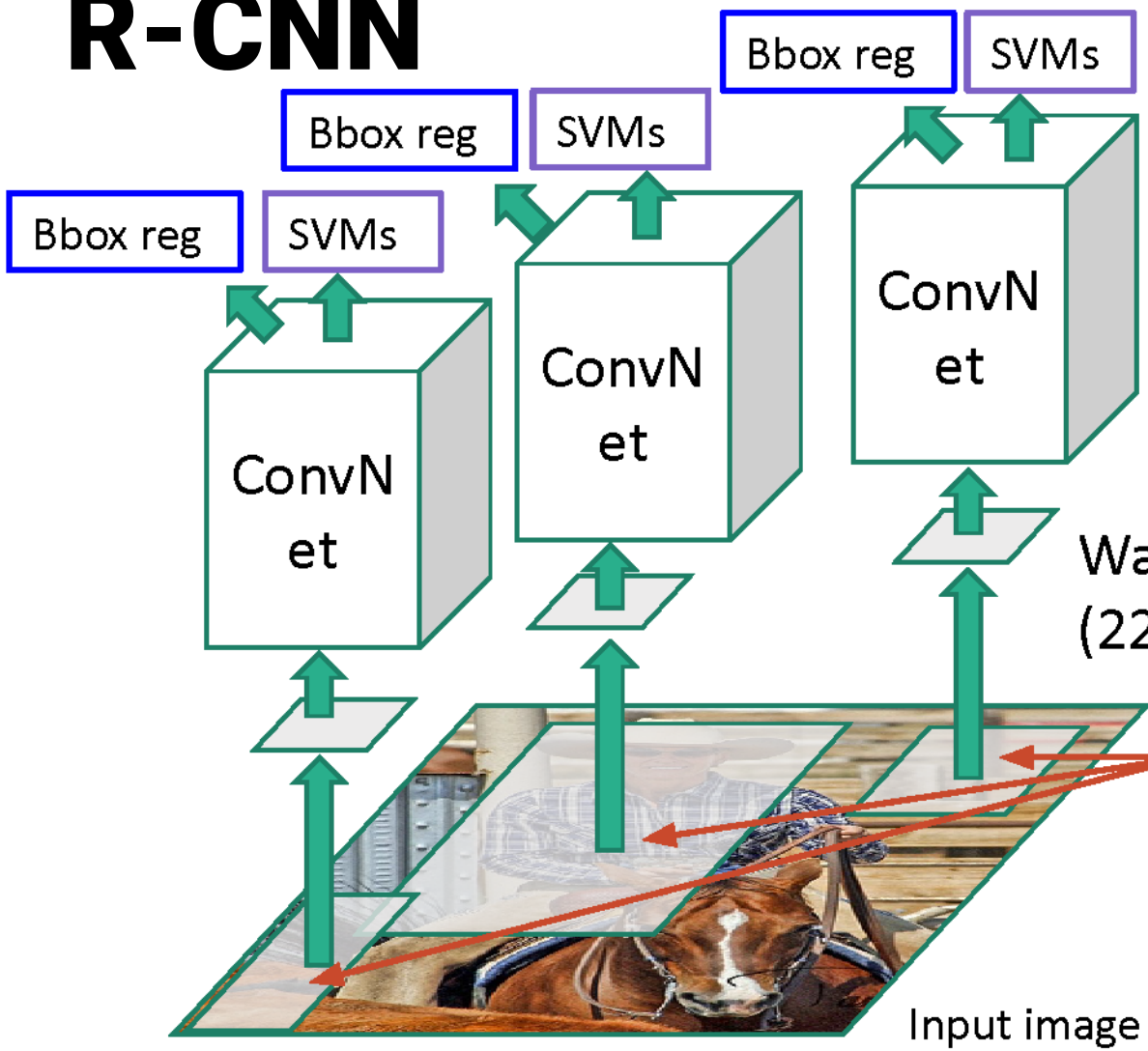

ConvN
et
ConvN
et
ConvN
et
Forward each region
through ConvNet
(ImageNet-pretranied)
Warped image regions
(224x224 pixels)
Regions of Interest
(RoI) from a proposal
method (~2k)
Input image

R-CNN


SVMs
SVMs
SVMs
ConvNet
ConvNet
ConvNet
Classify regions with SVMs
Forward each region through ConvNet (ImageNet-pretranied
Warped image regions (224x224 pixels)
Regions of Interest (RoI) from a proposal method (~2k)
Input image

Predict “corrections” to the RoI: 4 numbers: (dx, dy, dw, dh)

R-CNN



Classify regions with SVMs

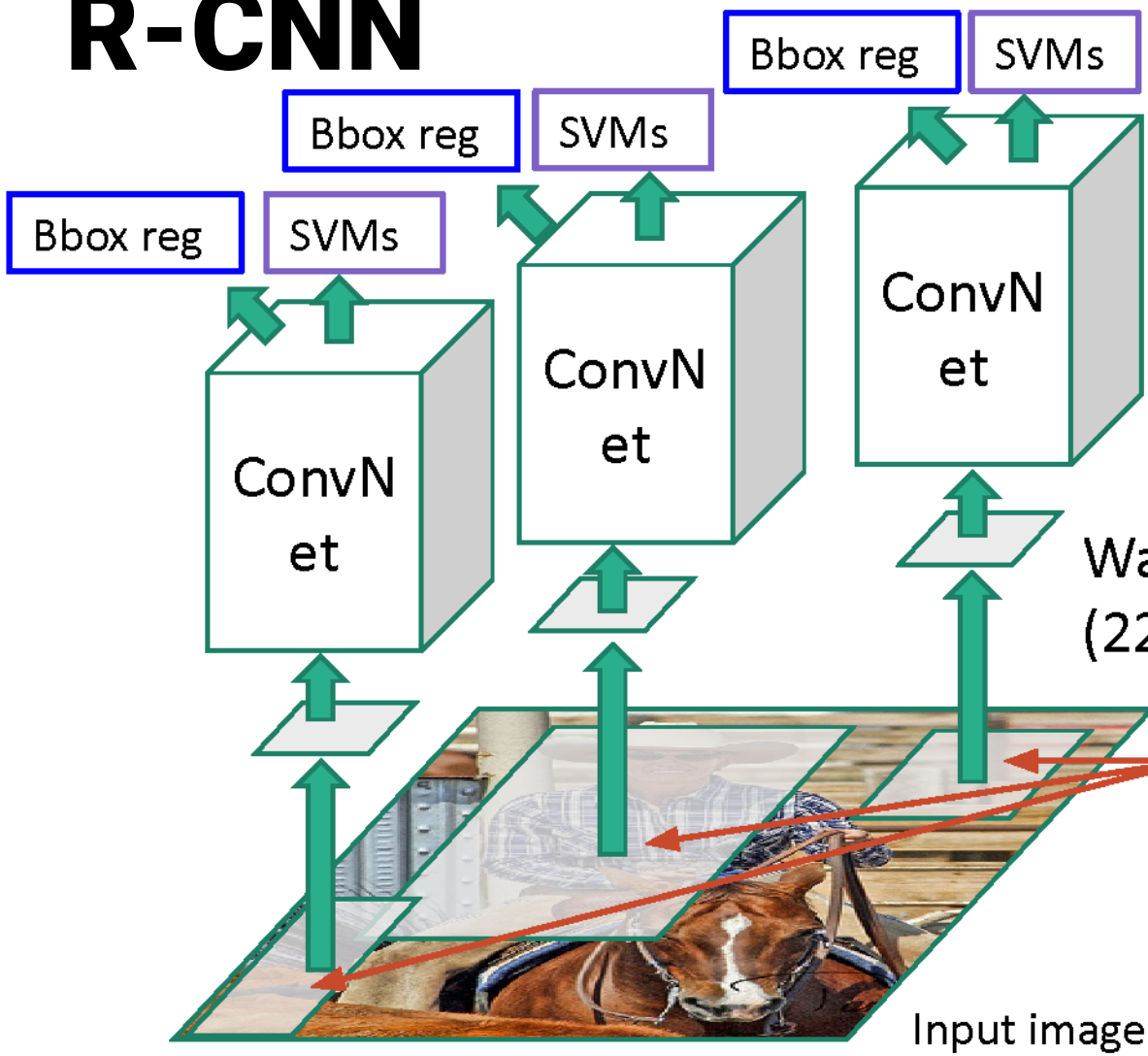
Forward each region through ConvNet (ImageNet-pretranied)

Warped image regions (224x224 pixels)

Regions of Interest (RoI) from a proposal method (~2k)

Predict “corrections” to the RoI: 4 numbers: (dx, dy, dw, dh)

R-CNN



Classify regions with SVMs

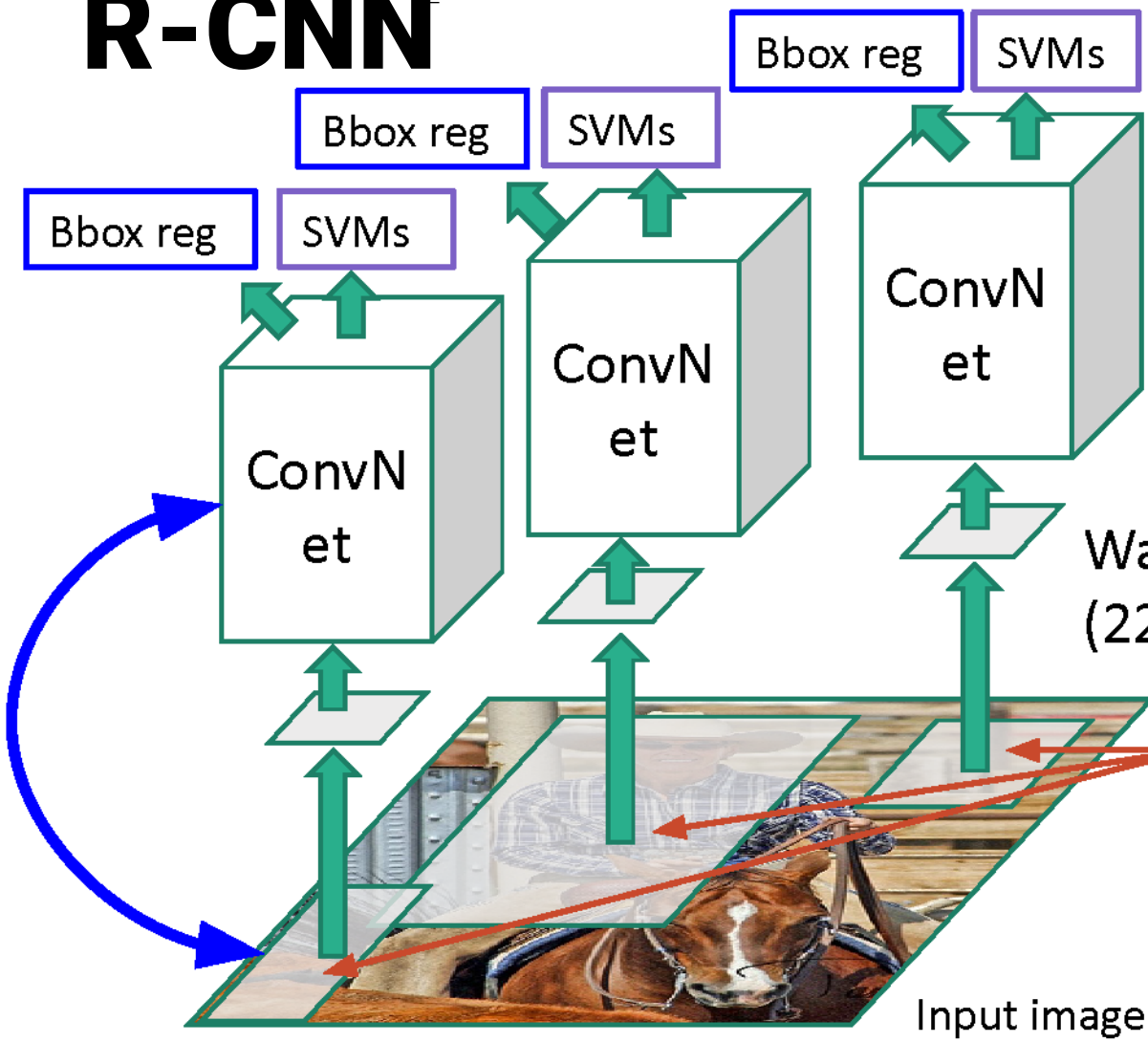
Forward each region through ConvNet

Warped image regions (224x224 pixels)

Regions of Interest (RoI) from a proposal method (~2k)

Problem: Very slow! Need to do ~2k independent forward passes for each image!

R-CNN



Classify regions with SVMs

Forward each region through ConvNet

Warped image regions (224x224 pixels)

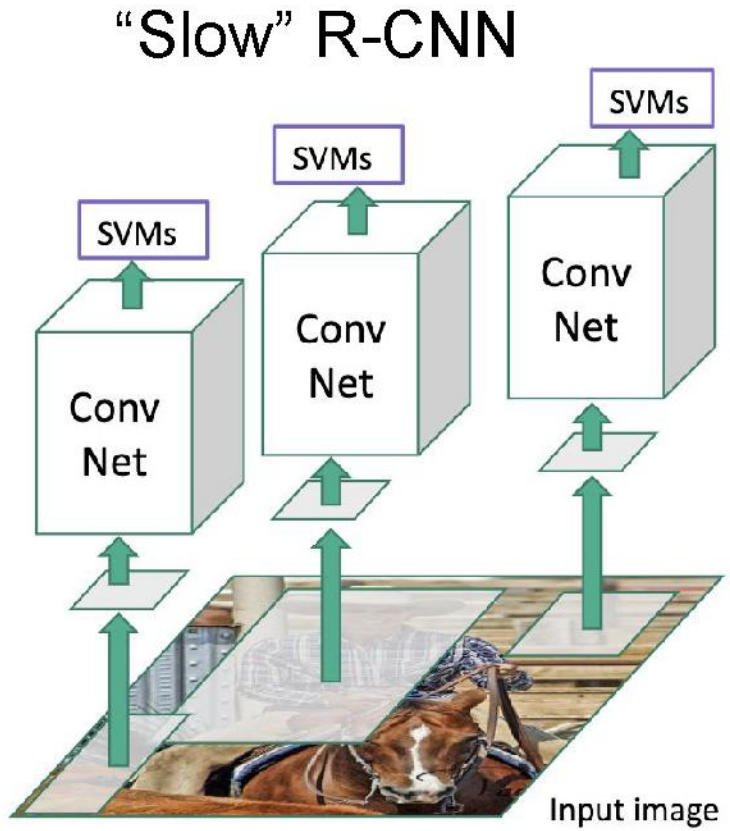
Regions of Interest (RoI) from a proposal method (~2k)

Problem: Very slow! Need to do ~2k independent forward passes for each image!

Idea: Pass the image through convnet before cropping! Crop the conv feature instead!

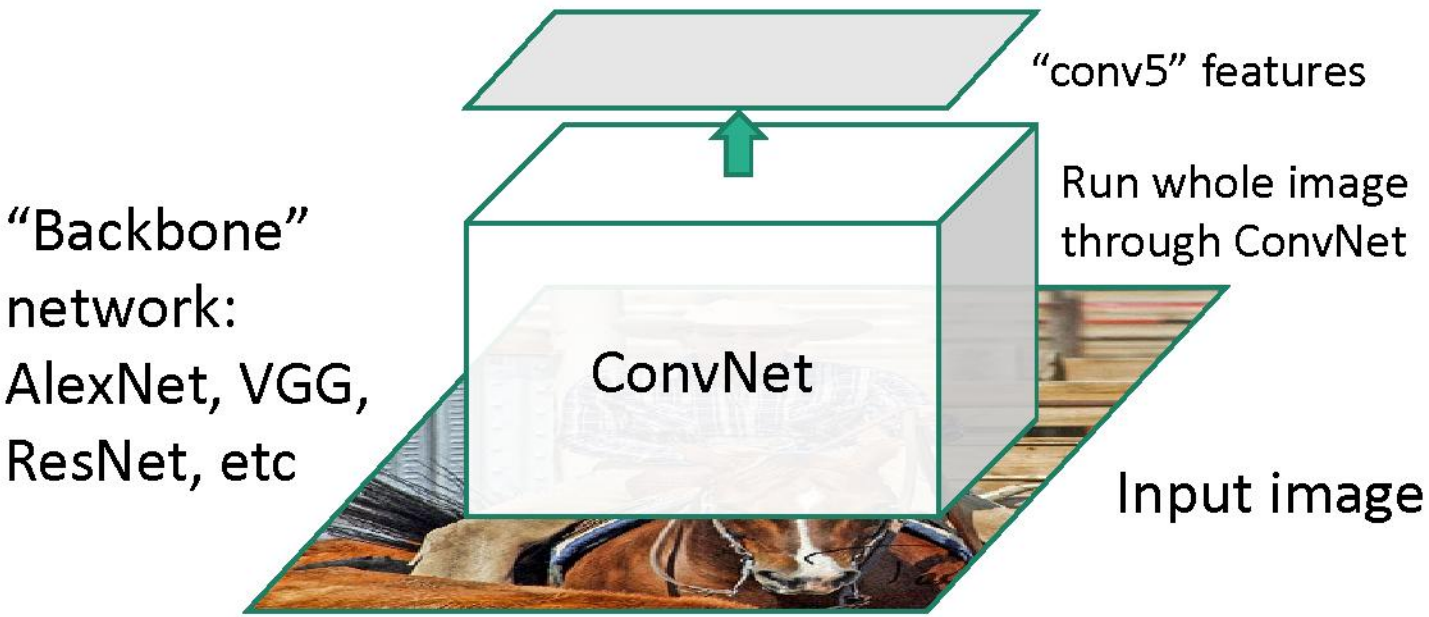
Fast R-CNN

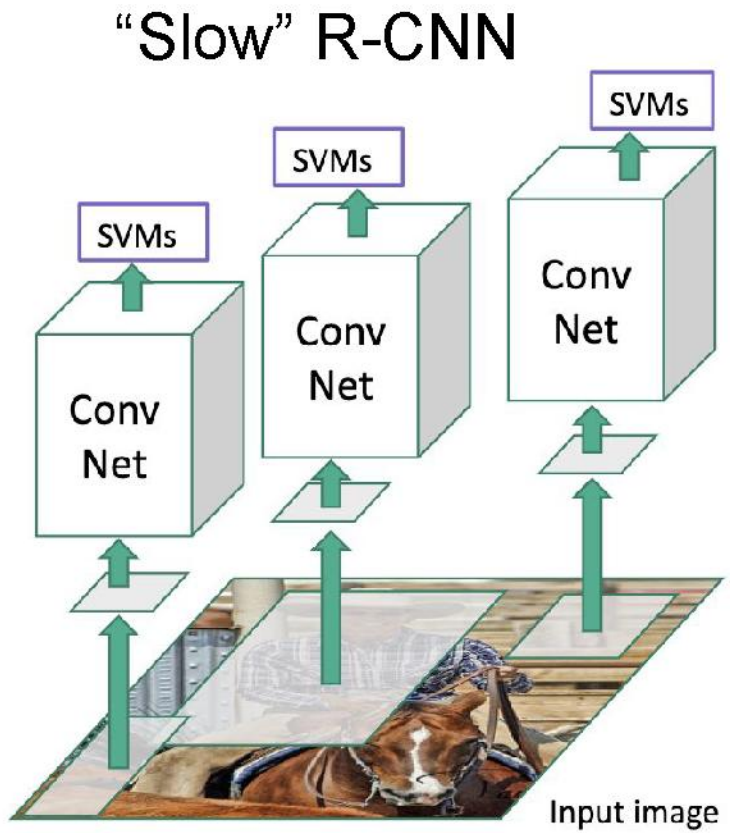
Input image



Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.

Fast R-CNN

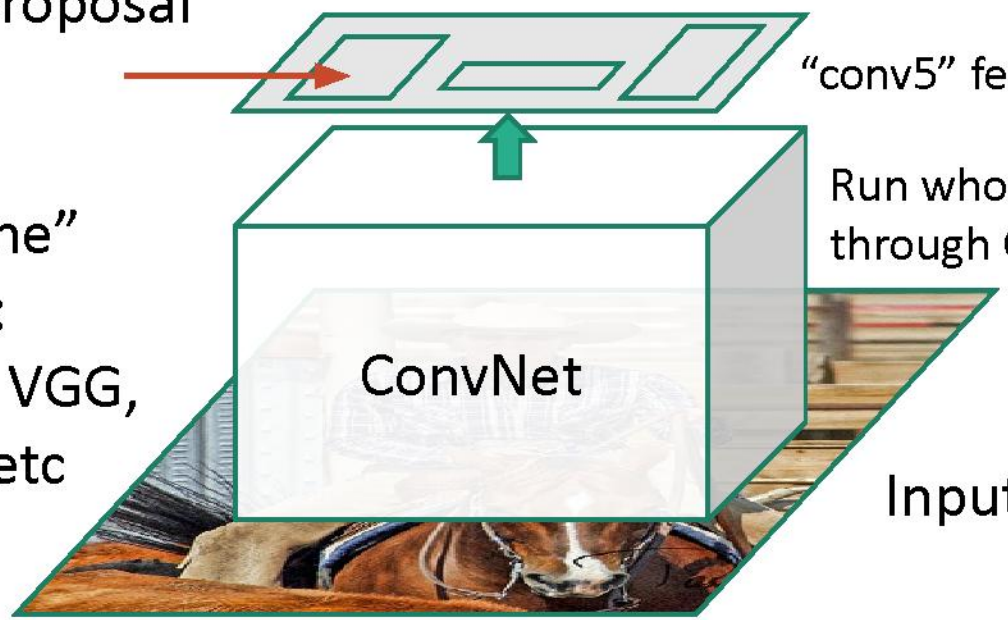

"conv5" features
Run whole image through ConvNet
"Backbone" network: AlexNet, VGG, ResNet, etc
ConvNet
Input image

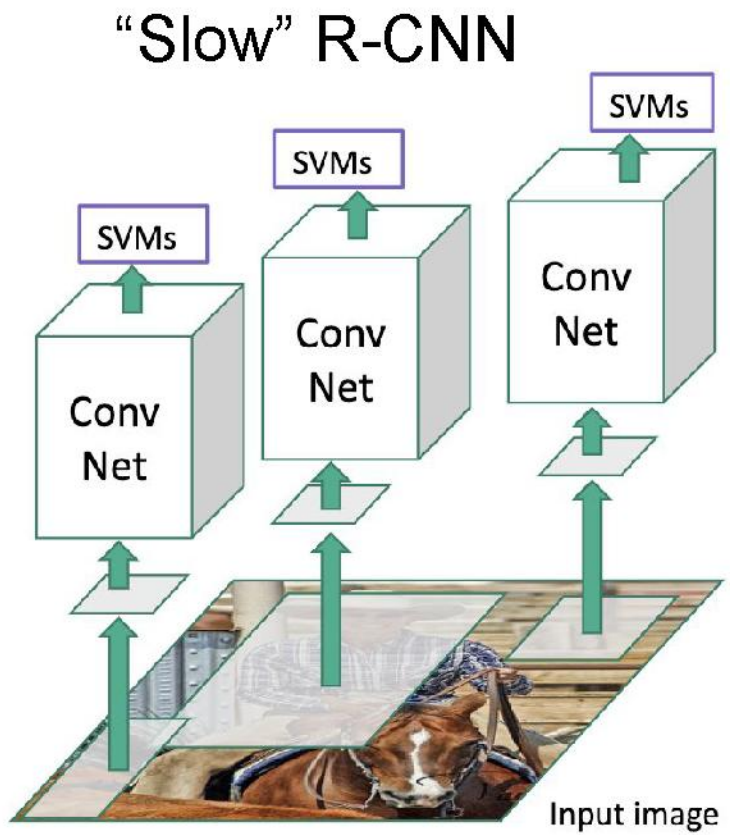

"Slow" R-CNN
SVMs
SVMs
SVMs
Conv Net
Conv Net
Conv Net
Input image

Fast R-CNN

Regions of Interest (RoIs) from a proposal method

"Backbone" network: AlexNet, VGG, ResNet, etc



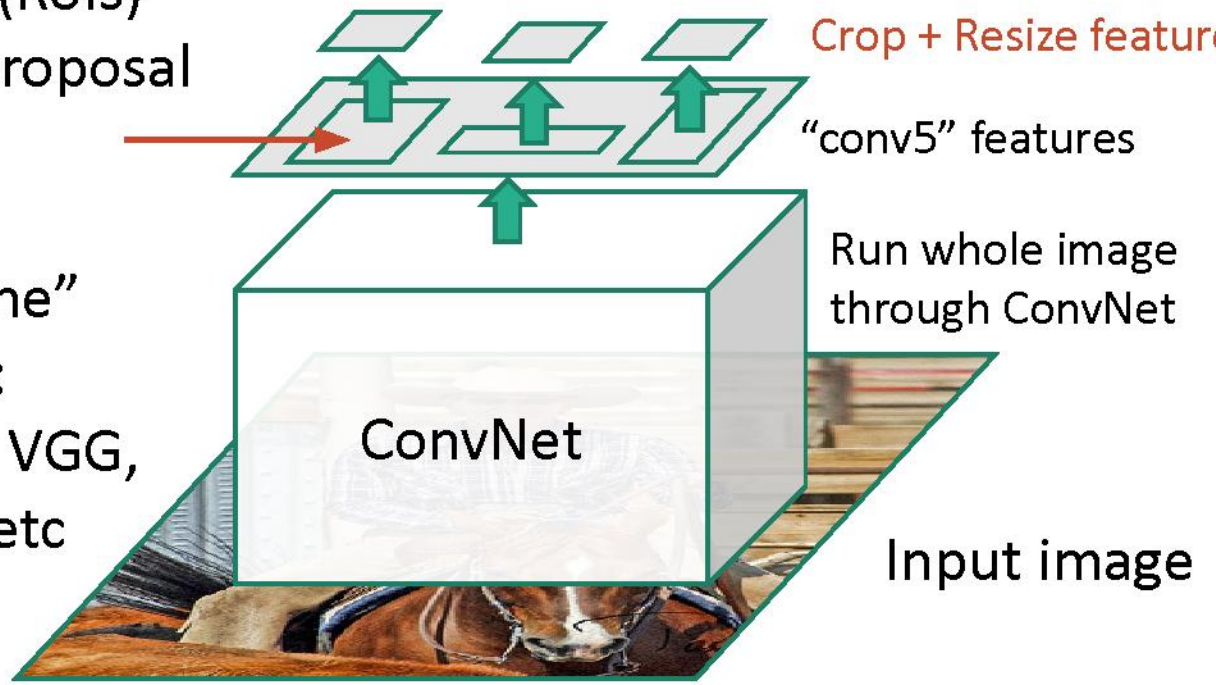


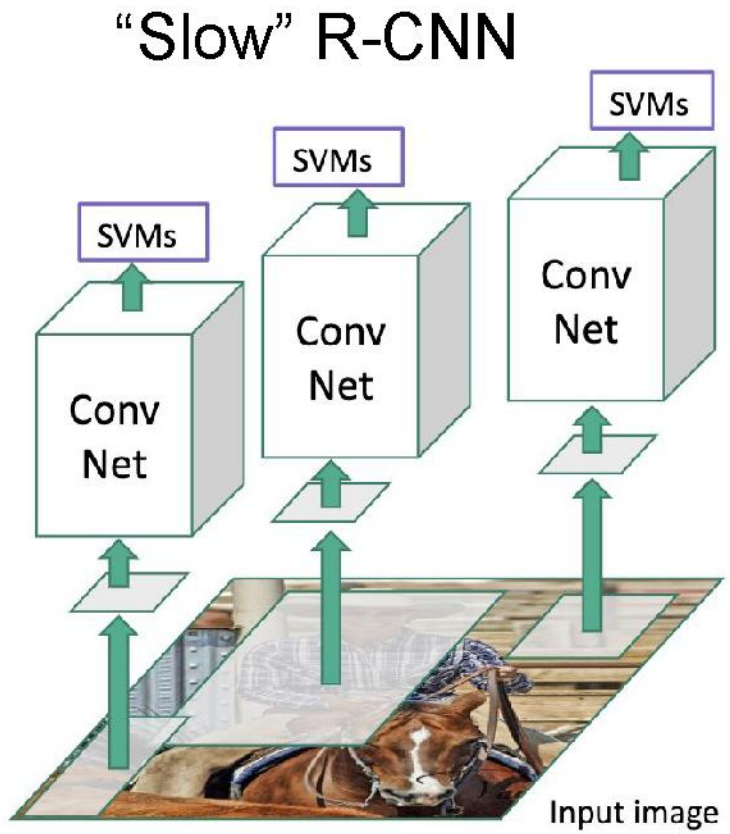
Fast R-CNN

Region Proposals on Features

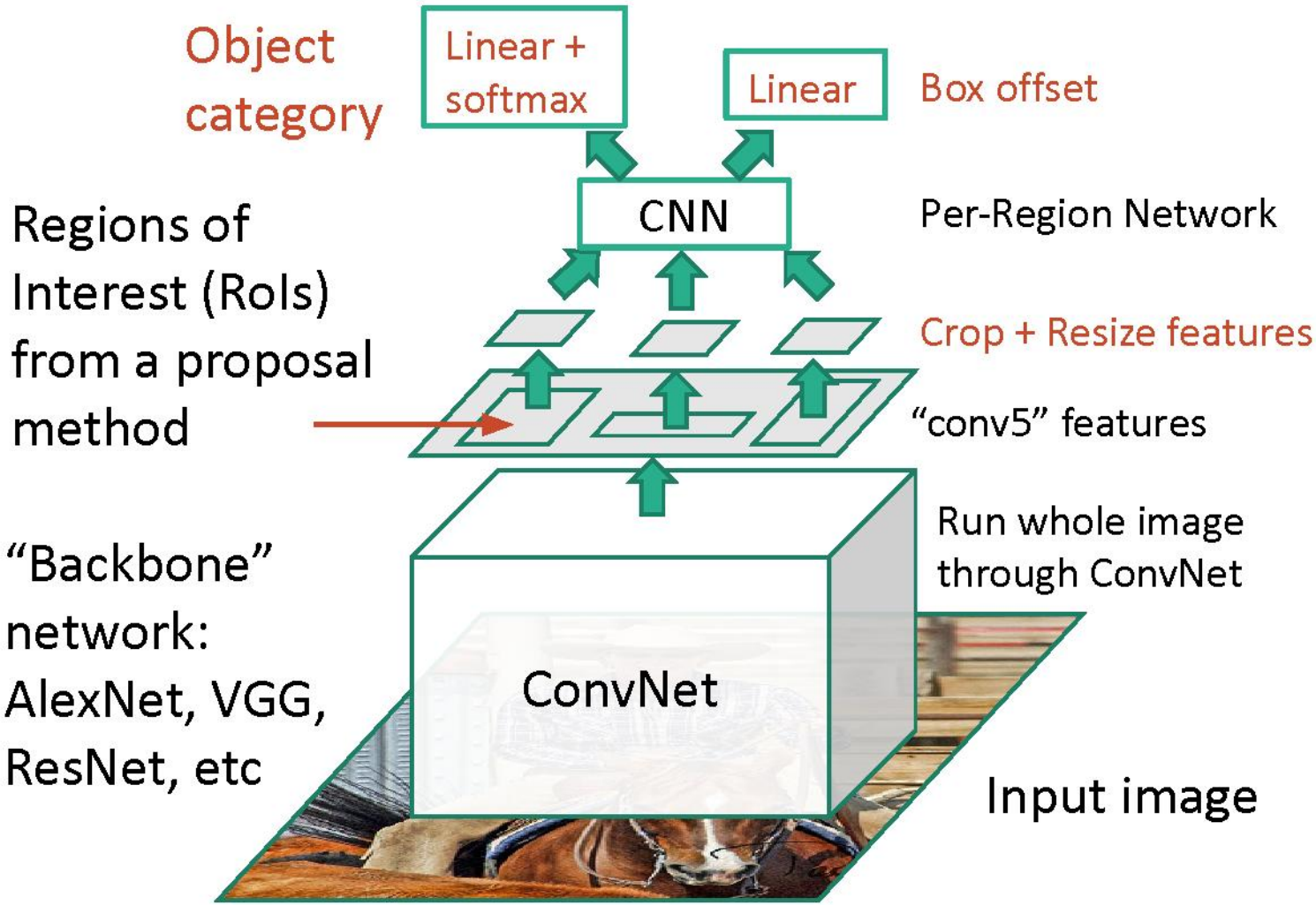
Regions of Interest (RoIs) from a proposal method

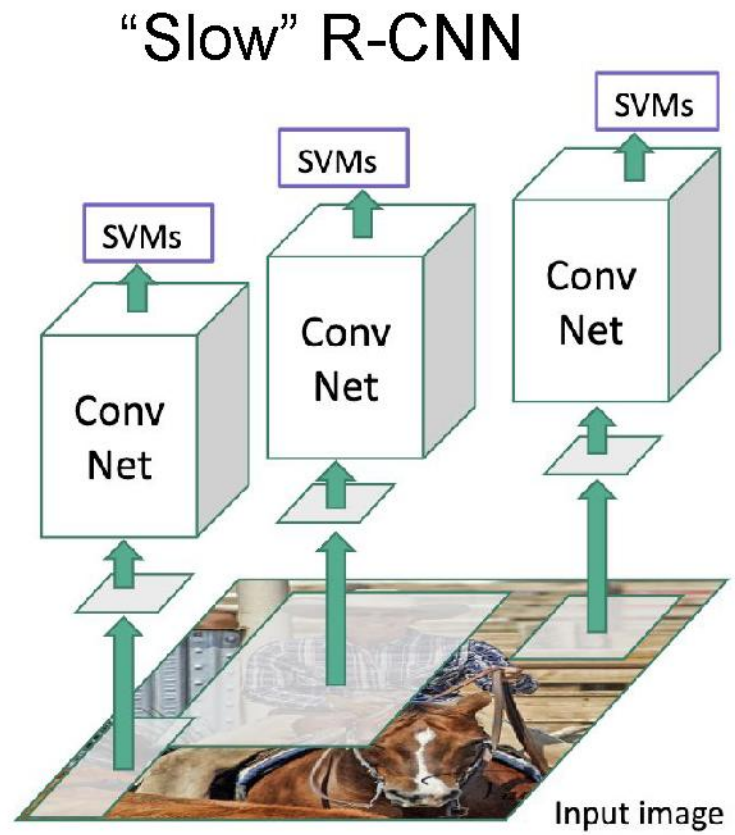
"Backbone" network: AlexNet, VGG, ResNet, etc



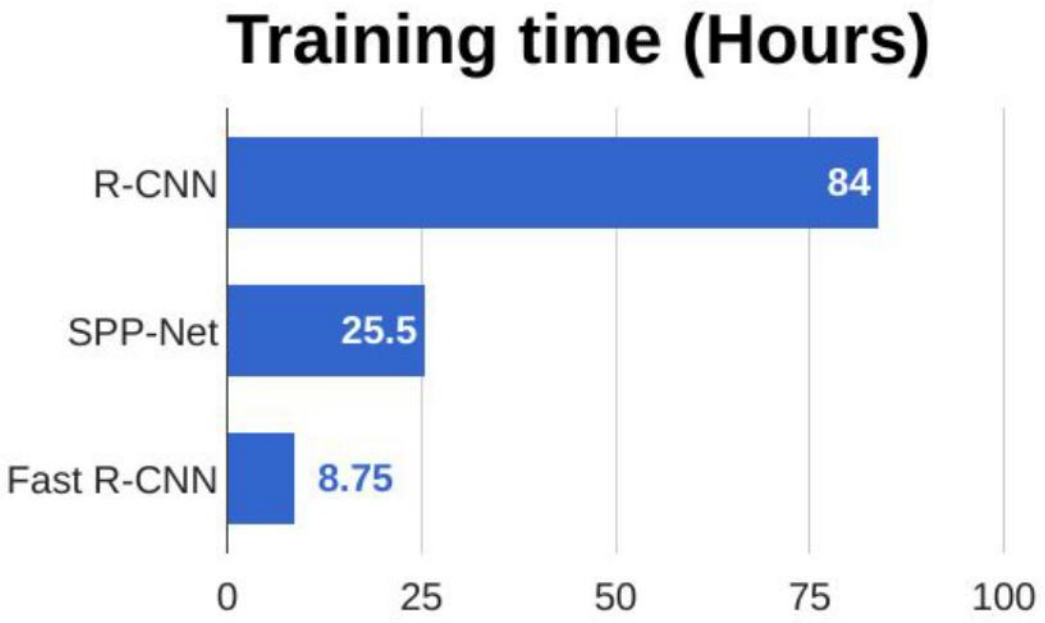


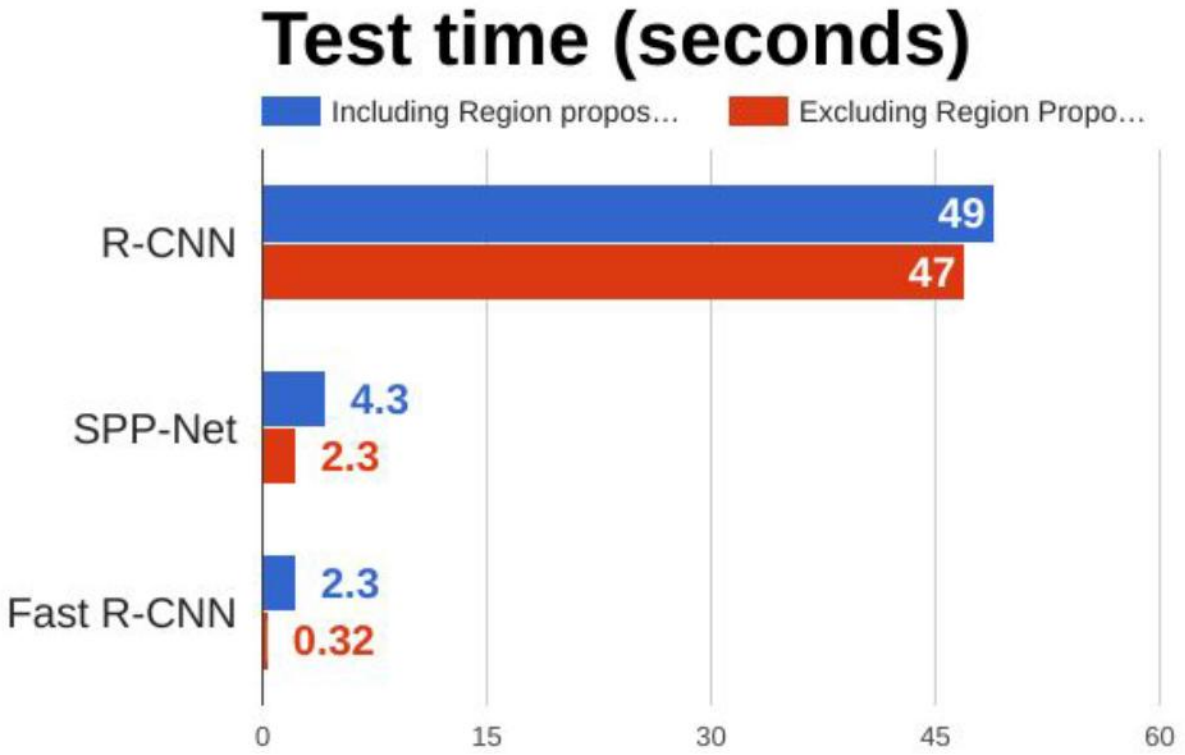
Fast R-CNN


Object category
Linear + softmax
Linear
Box offset
CNN
Per-Region Network
Regions of Interest (RoIs) from a proposal method
Crop + Resize features
"conv5" features
Run whole image through ConvNet
"Backbone" network: AlexNet, VGG, ResNet, etc
ConvNet
Input image

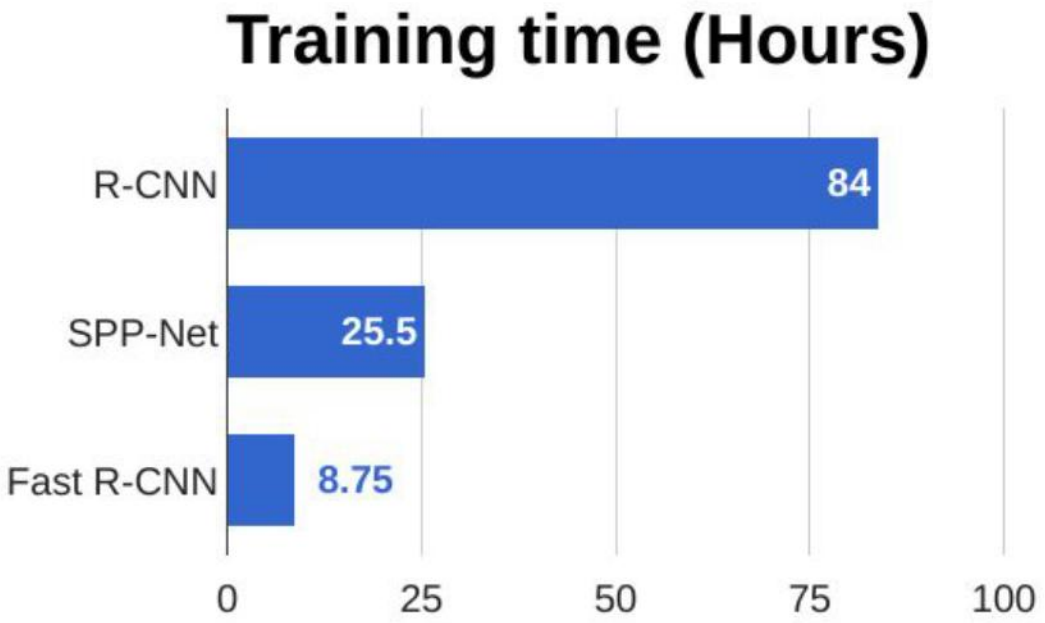

"Slow" R-CNN
SVMs
SVMs
SVMs
Conv Net
Conv Net
Conv Net
Input image

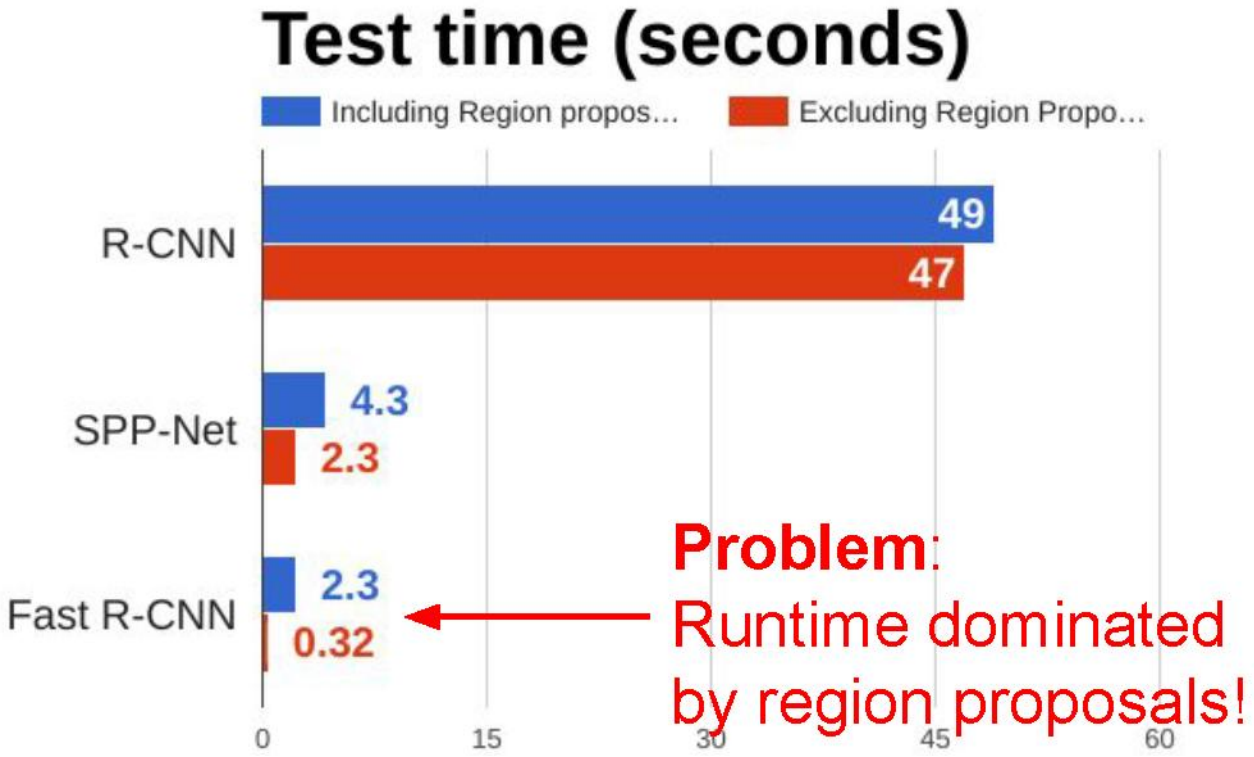
R-CNN Vs Fast R-CNN





R-CNN Vs Fast R-CNN

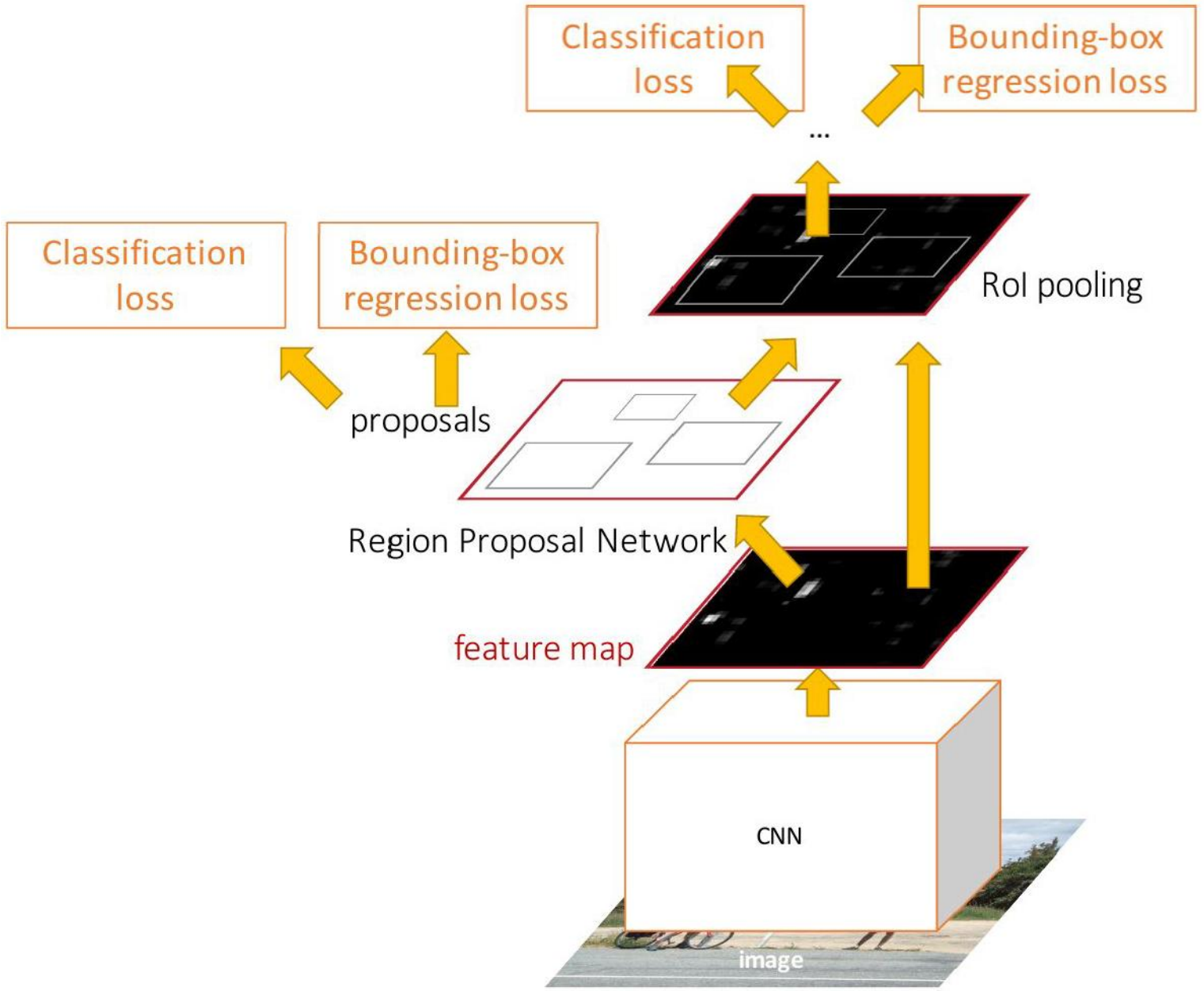




Faster R-CNN

Insert **Region Proposal Network (RPN)** to predict proposals from features

Otherwise same as Fast R-CNN: Crop features for each proposal, classify each one



Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015).

Region Proposal Network

Input Image
(e.g. 3 x 640 x 480)

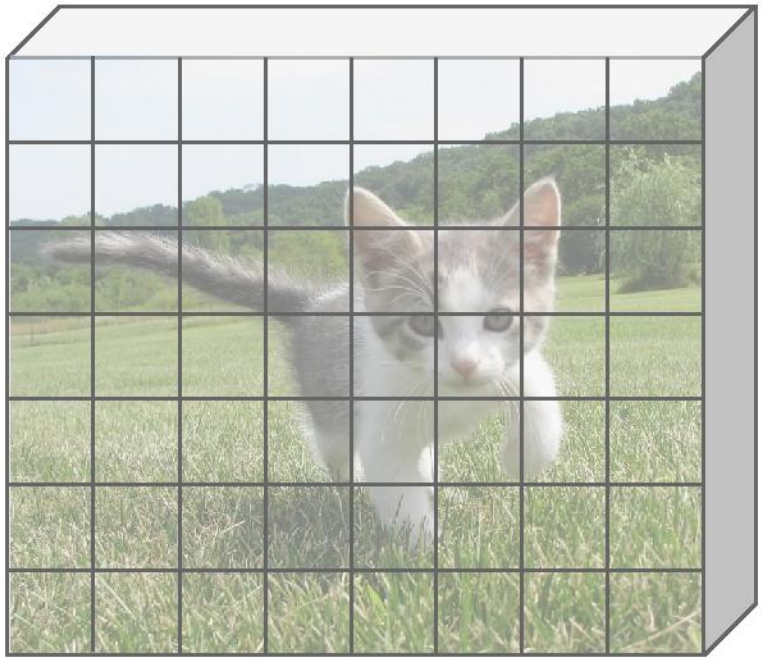


Image features
(e.g. 512 x 20 x 15)

Region Proposal Network

Imagine an **anchor box** of fixed size at each point in the feature map

Input Image
(e.g. 3 x 640 x 480)

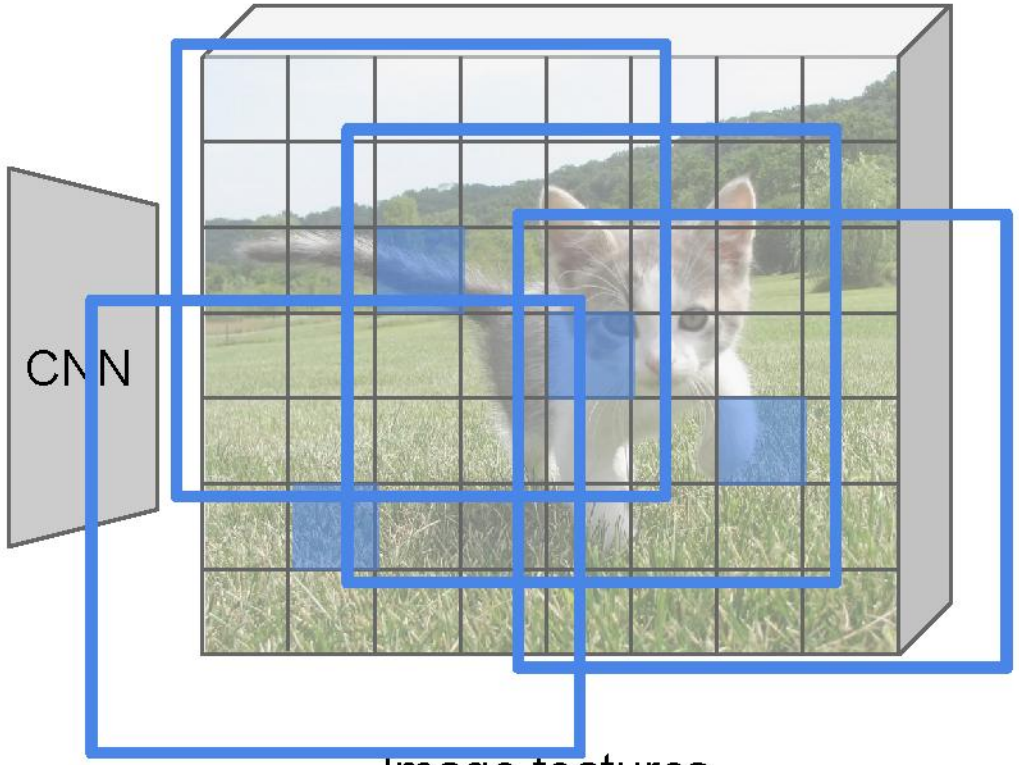
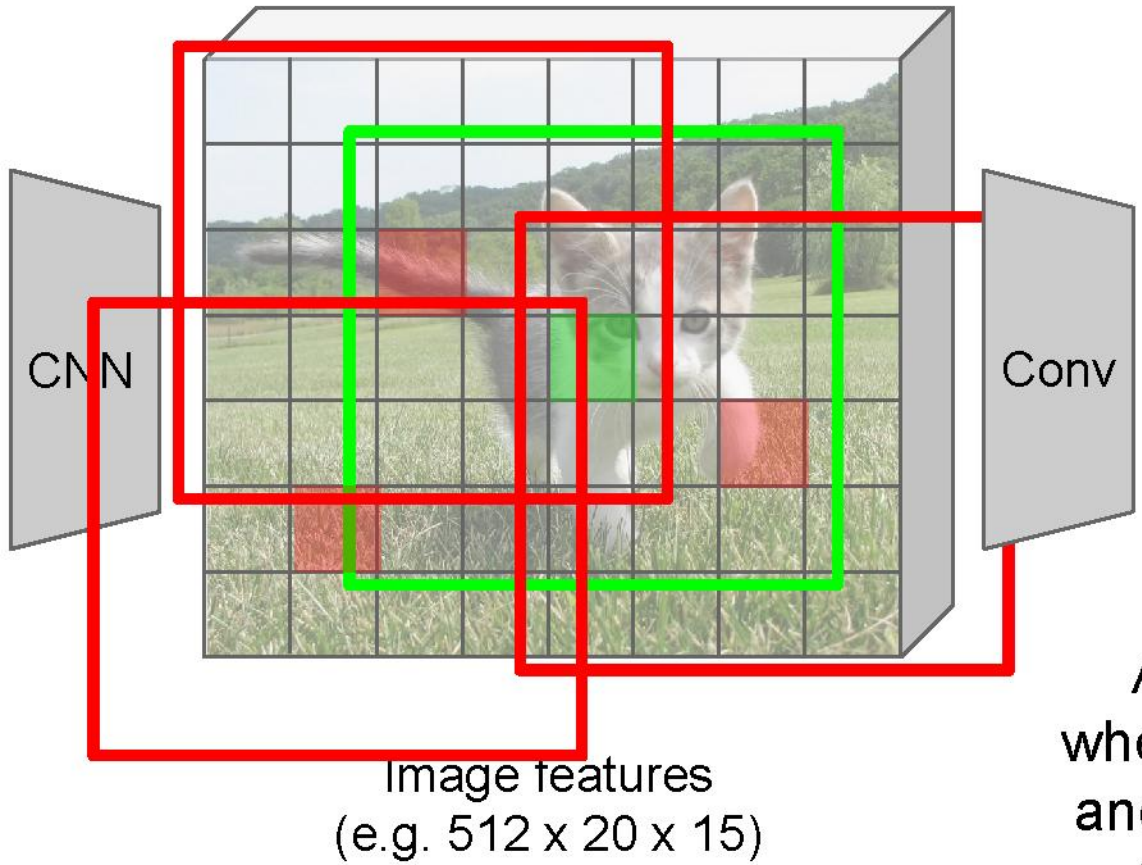


Image features
(e.g. 512 x 20 x 15)

Region Proposal Network

Imagine an **anchor box** of fixed size at each point in the feature map

Input Image
(e.g. 3 x 640 x 480)



At each point, predict whether the corresponding anchor contains an object (binary classification)

Region Proposal Network

Imagine an **anchor box** of fixed size at each point in the feature map

Input Image
(e.g. 3 x 640 x 480)

CNN

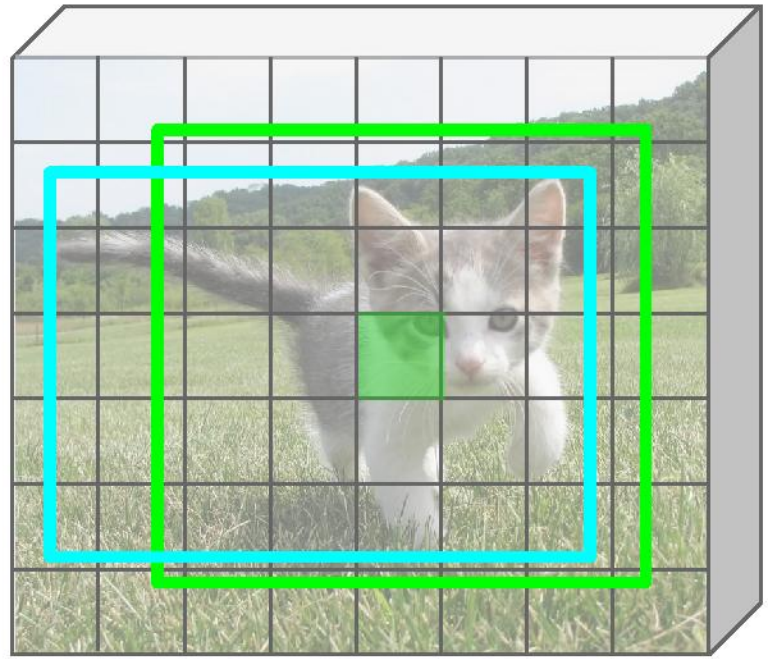


Image features
(e.g. 512 x 20 x 15)

Conv

Anchor is an object?
1 x 20 x 15

Box corrections
4 x 20 x 15

For positive boxes, also predict a corrections from the anchor to the ground-truth box (regress 4 numbers per pixel)

Region Proposal Network

In practice use K different anchor boxes of different size / scale at each point

Input Image
(e.g. 3 x 640 x 480)

CNN

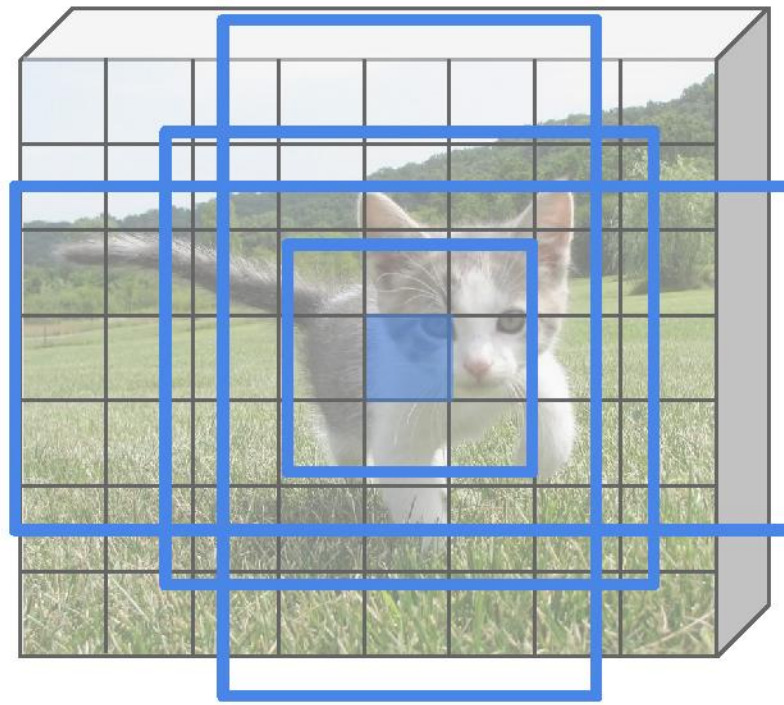


Image features
(e.g. 512 x 20 x 15)

Conv

Anchor is an object?
K x 20 x 15

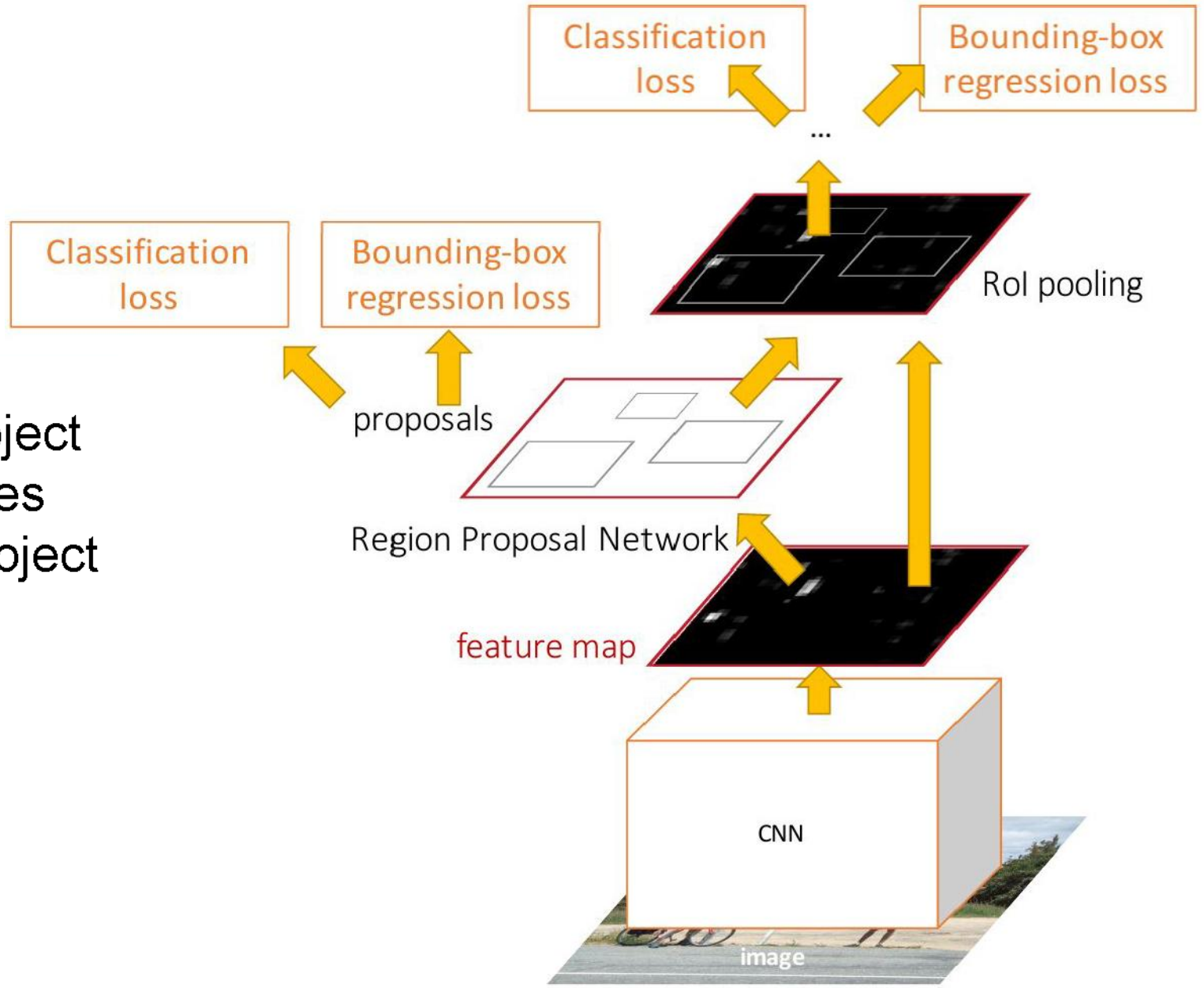
Box transforms
4K x 20 x 15

Sort the K*20*15 boxes by their “objectness” score, take top ~300 as our proposals

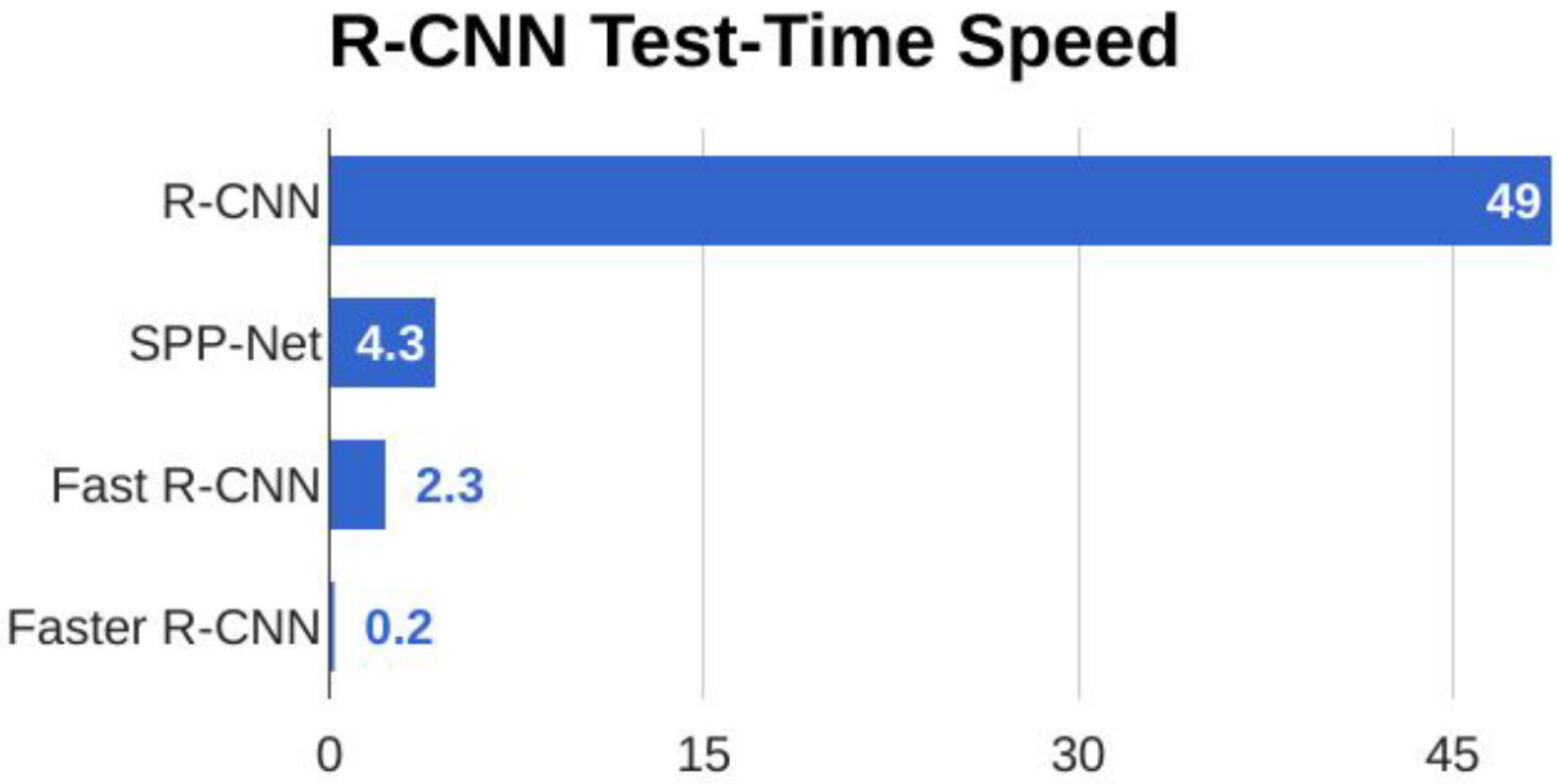
Faster R-CNN

Jointly train with 4 losses:

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



Faster R-CNN

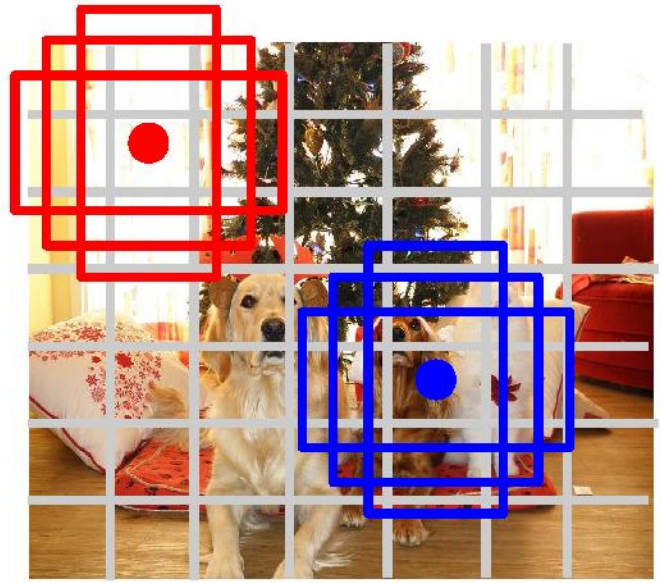

R-CNN Test-Time Speed
R-CNN
SPP-Net
Fast R-CNN
Faster R-CNN
49
4.3
2.3
0.2
0
15
30
45

Single Shot Detectors

Perform Non-max suppression directly

YoLo, SSD, RetinaNet

Input image
3 x H x W



Divide image into grid
7 x 7

Image a set of **base boxes** centered at each grid cell
Here B = 3

Within each grid cell:

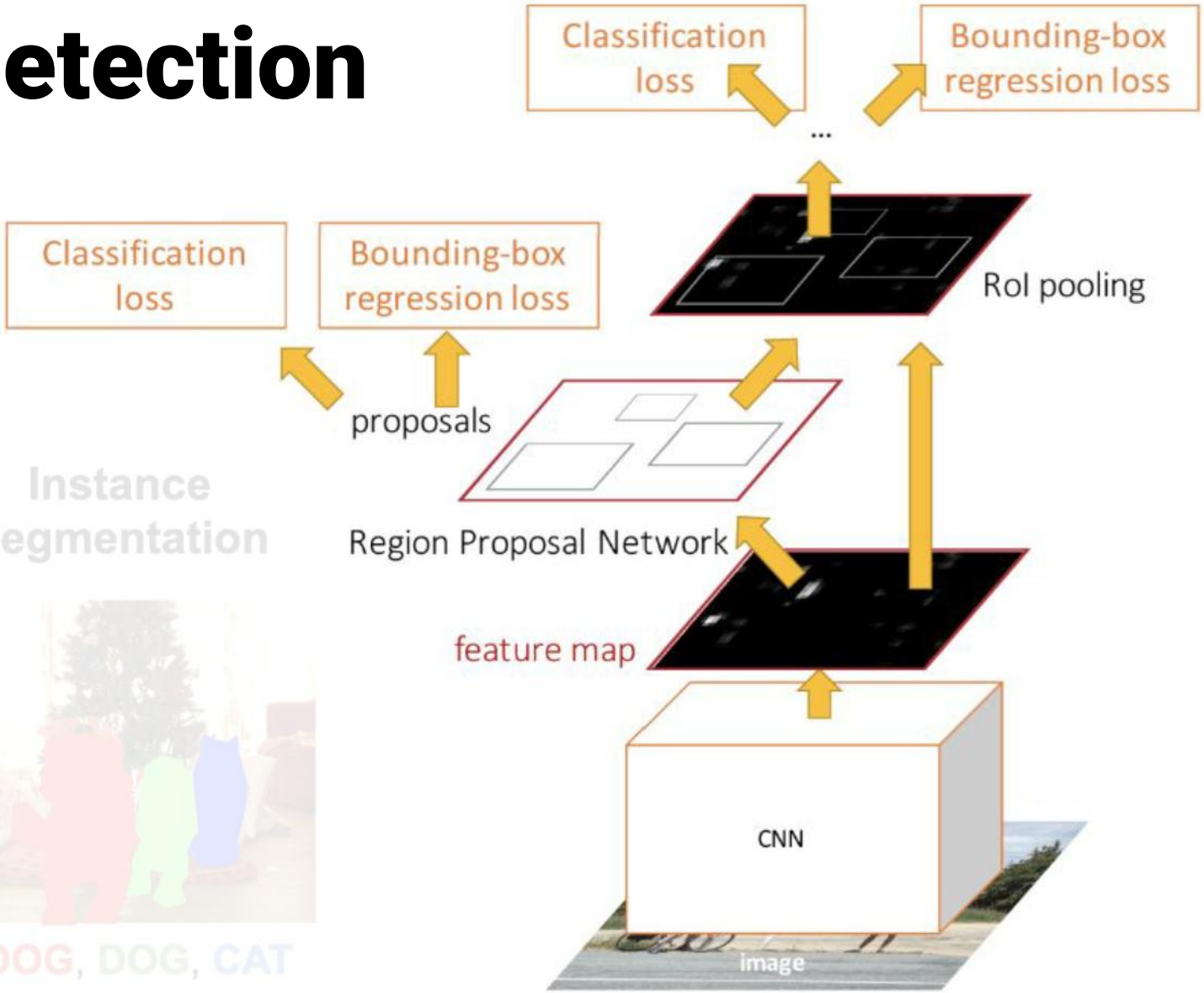
- Regress from each of the B base boxes to a final box with 5 numbers: (dx, dy, dh, dw, confidence)
- Predict scores for each of C classes (including background as a class)
- Looks a lot like RPN, but category-specific!

Output:
7 x 7 x (5 * B + C)

Redmon et al, "You Only Look Once: Unified, Real-Time Object Detection", CVPR 2016
Liu et al, "SSD: Single-Shot MultiBox Detector", ECCV 2016
Lin et al, "Focal Loss for Dense Object Detection", ICCV 2017

Object Detection

Faster R-CNN



Object Detection

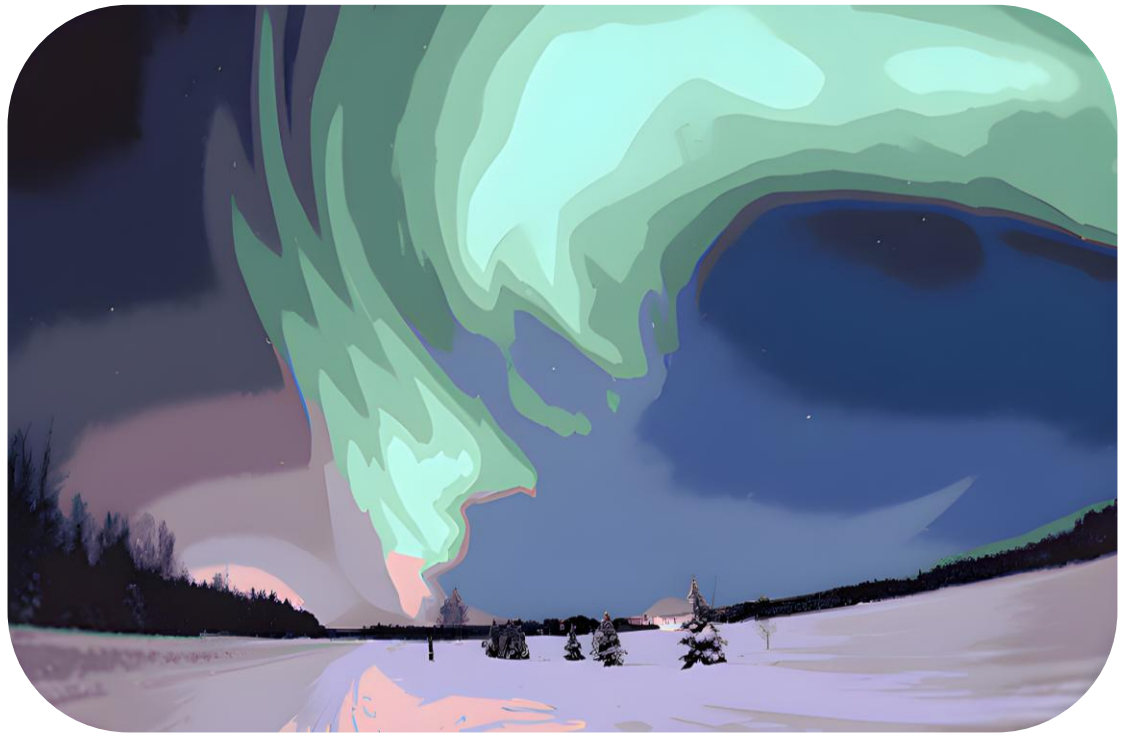
DOG, DOG, CAT

Instance Segmentation

DOG, DOG, CAT

What If We Want To Classify Every Pixel?

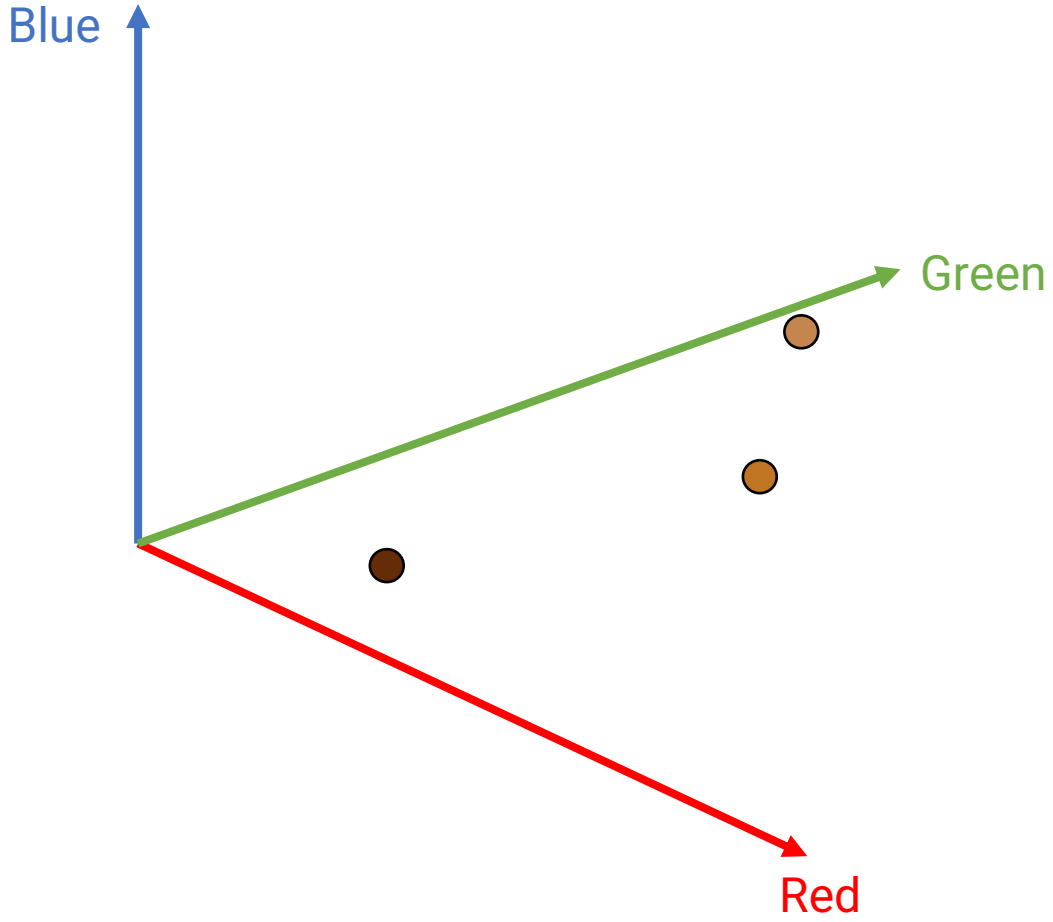
Scene Segmentation



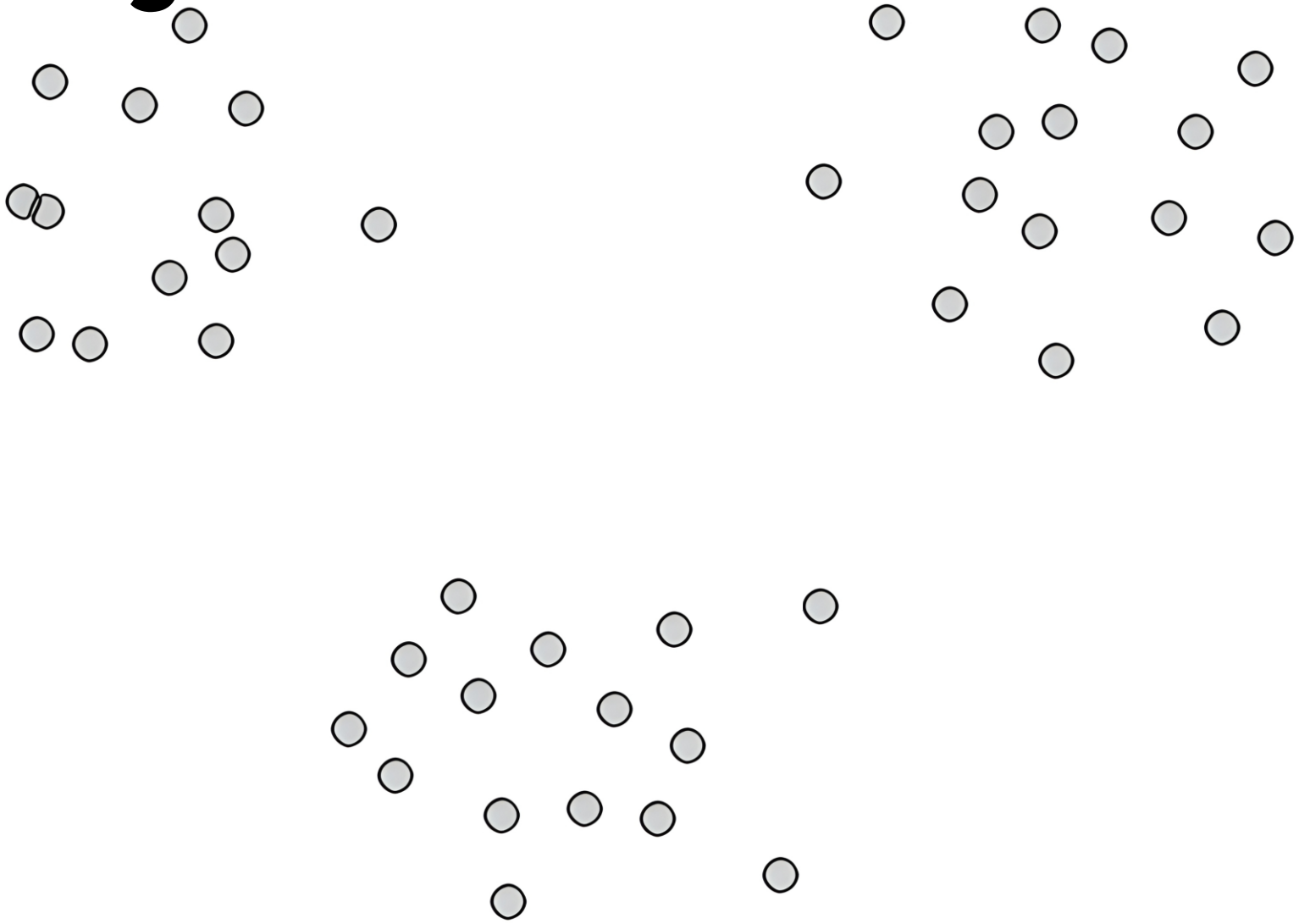
We are trying to group **“similar”** pixels together

Slides adapted from Naitri Rajyaguru

Clustering

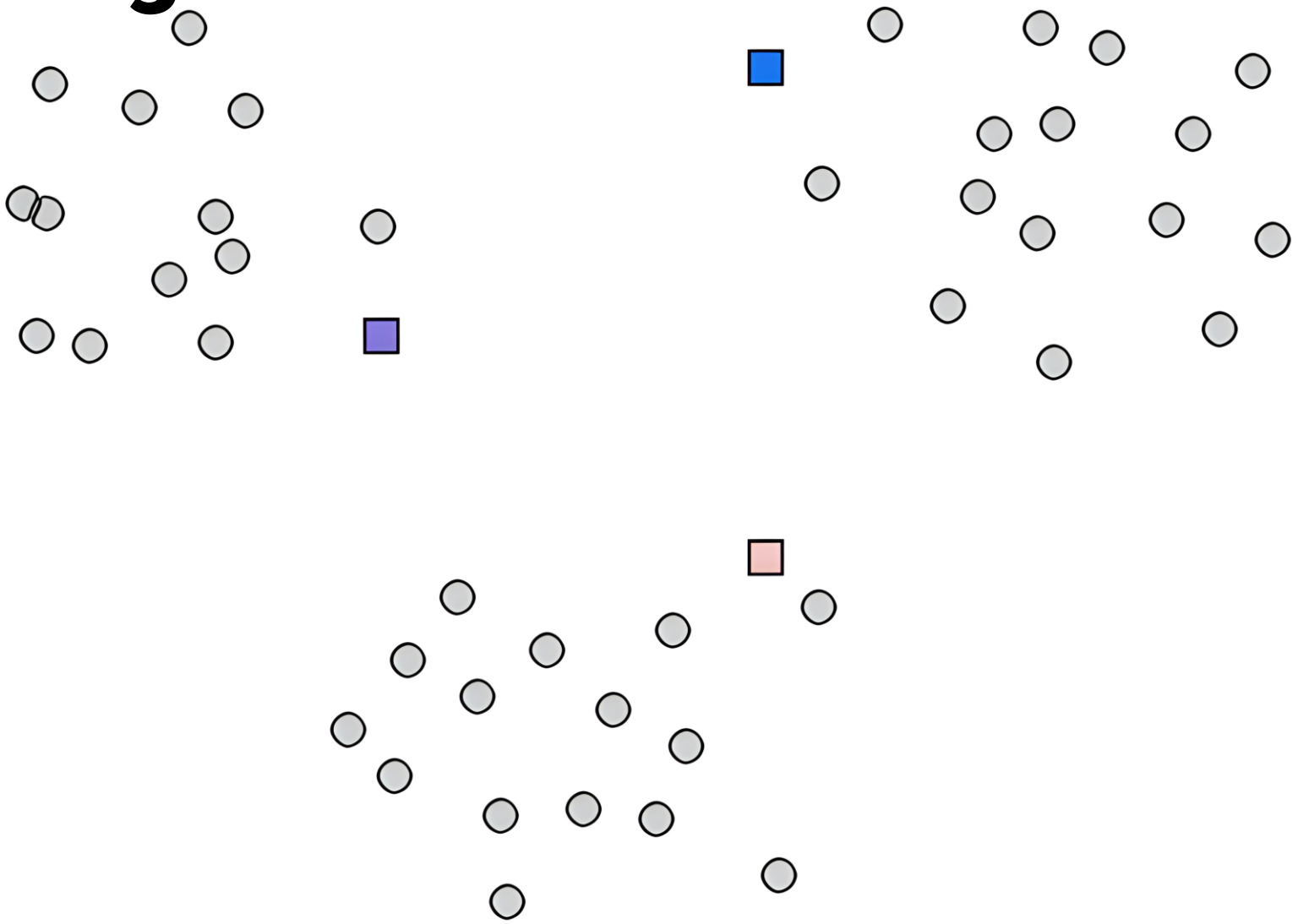


Clustering



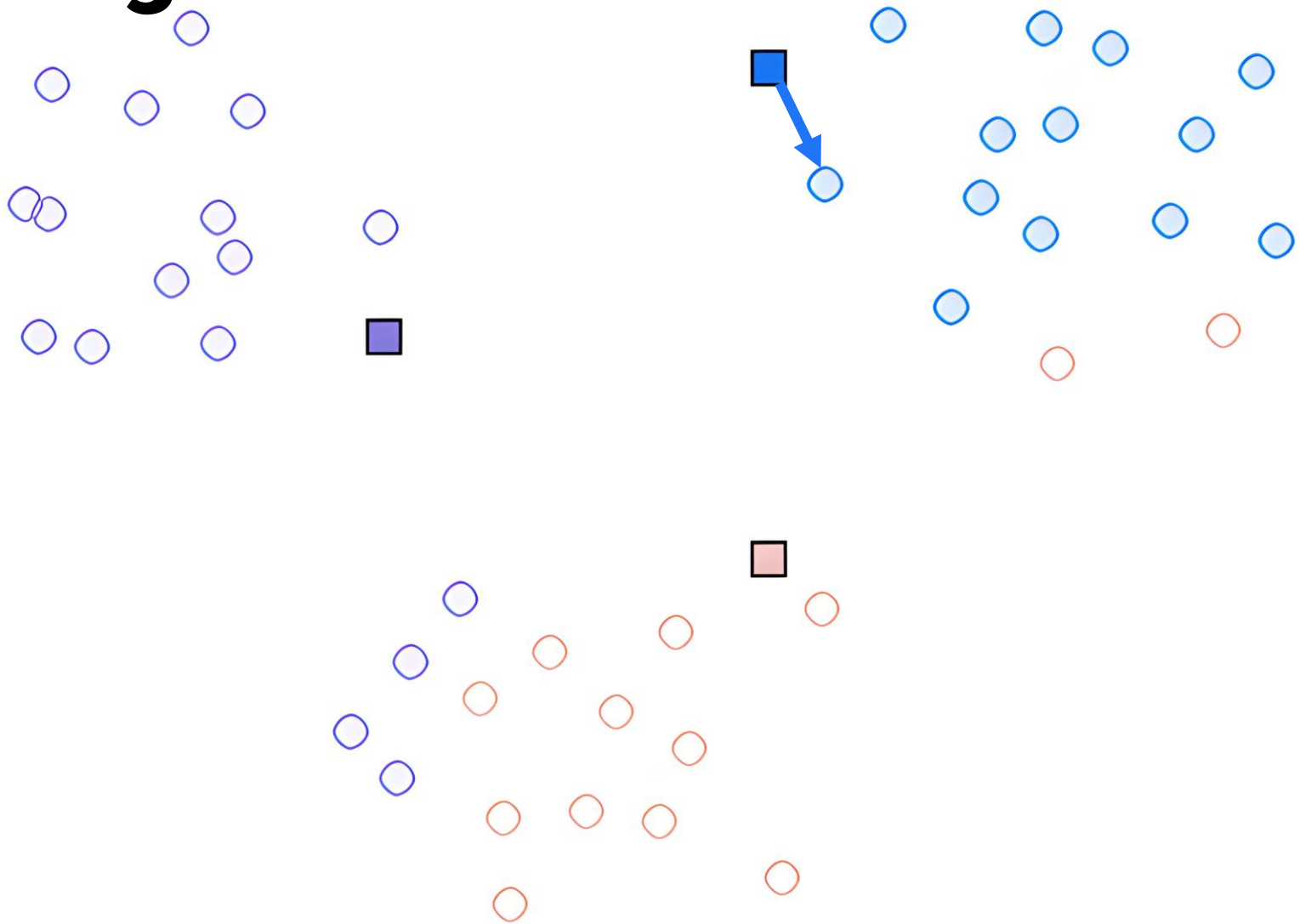
Slide adapted from UPenn's CIS520

Clustering

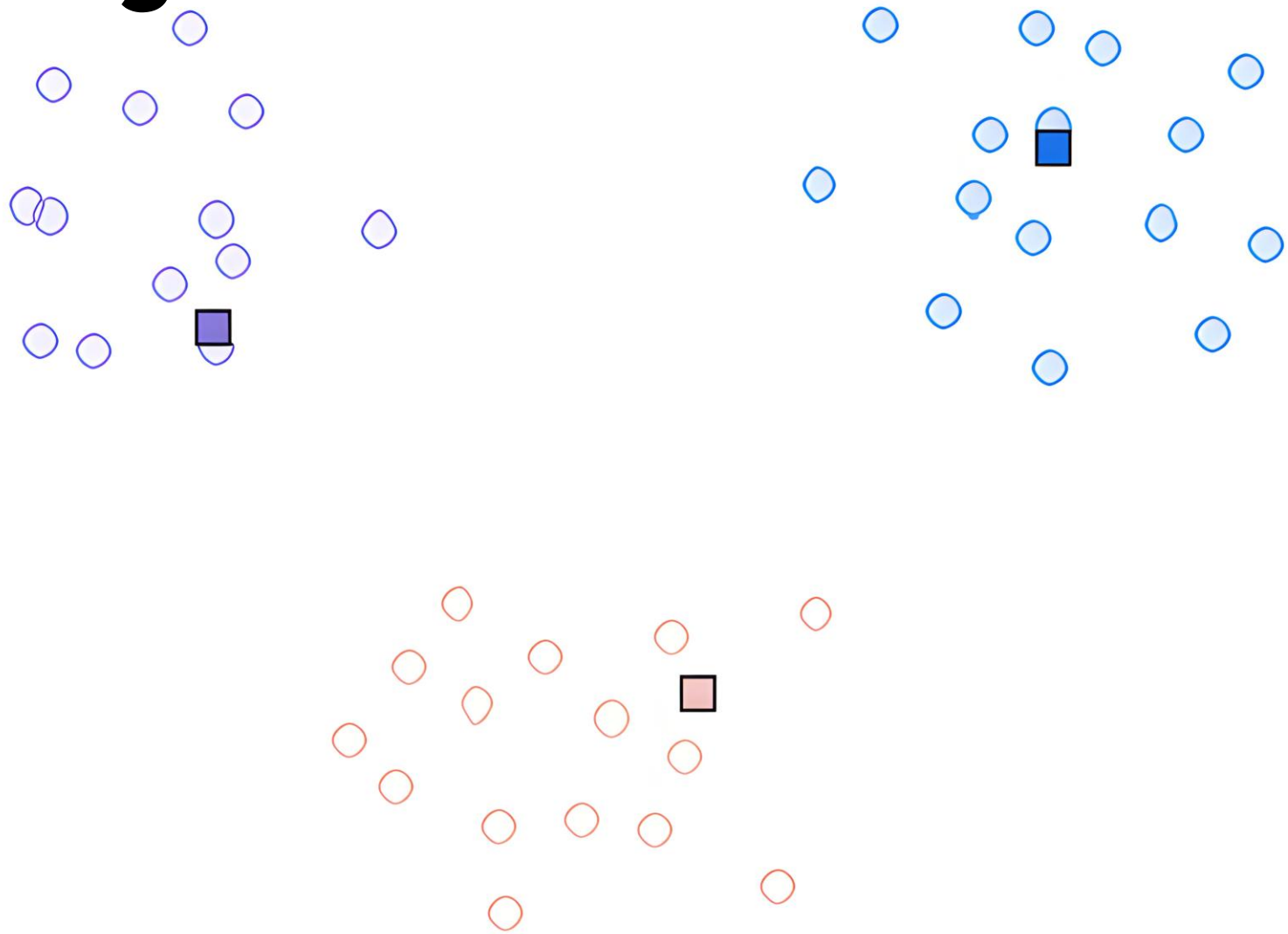


Define distance metric for cluster assignment
Assign clusters within a distance

Clustering

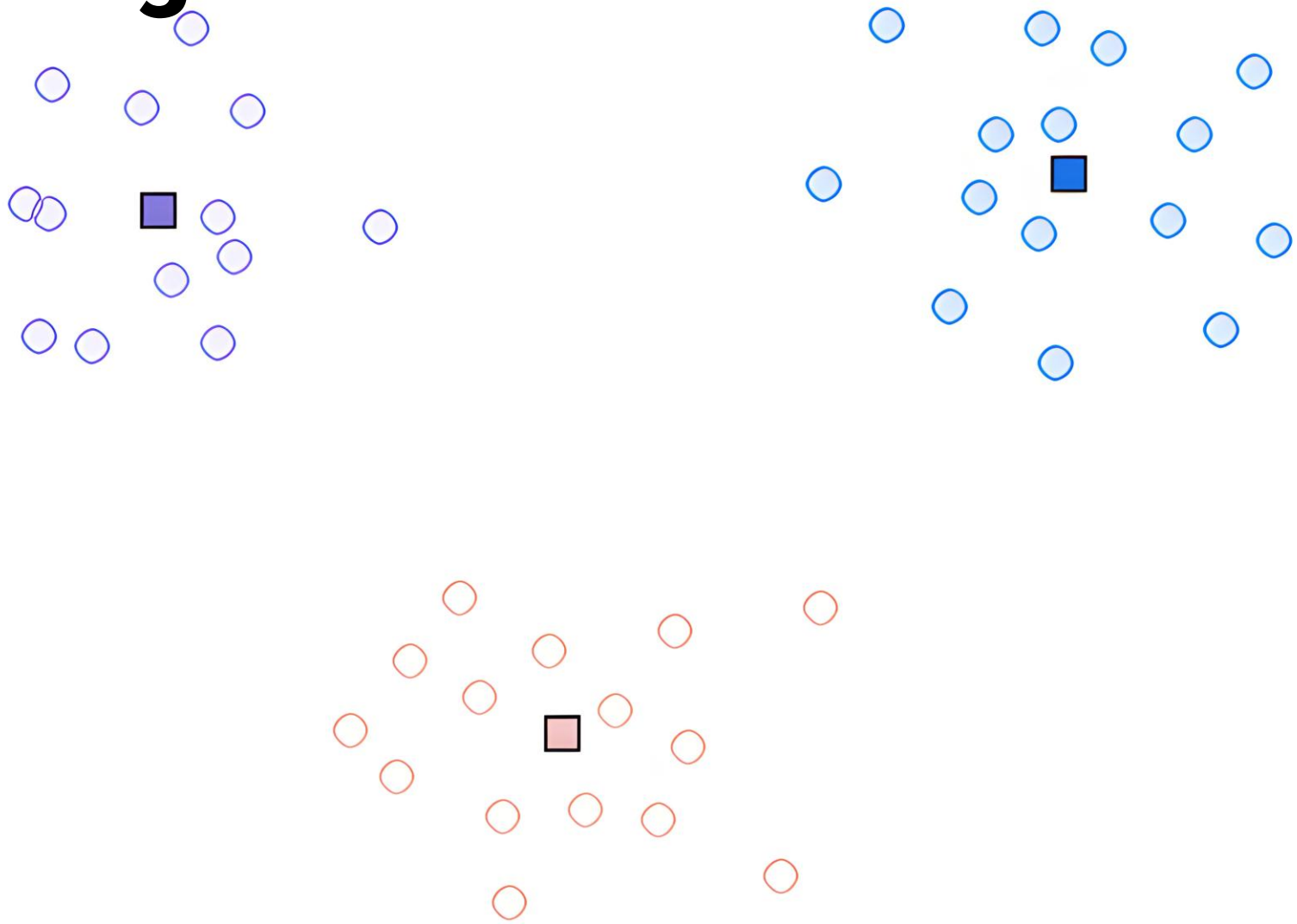


Clustering

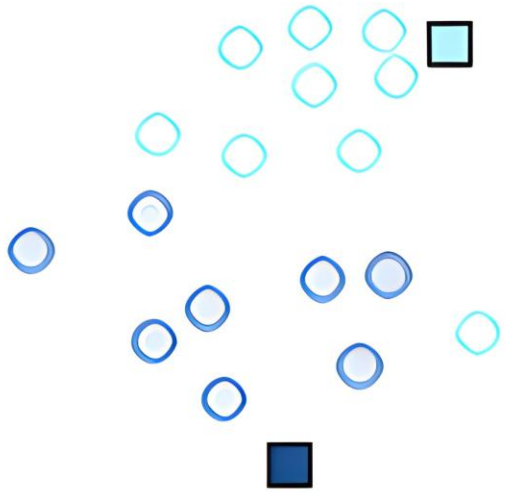


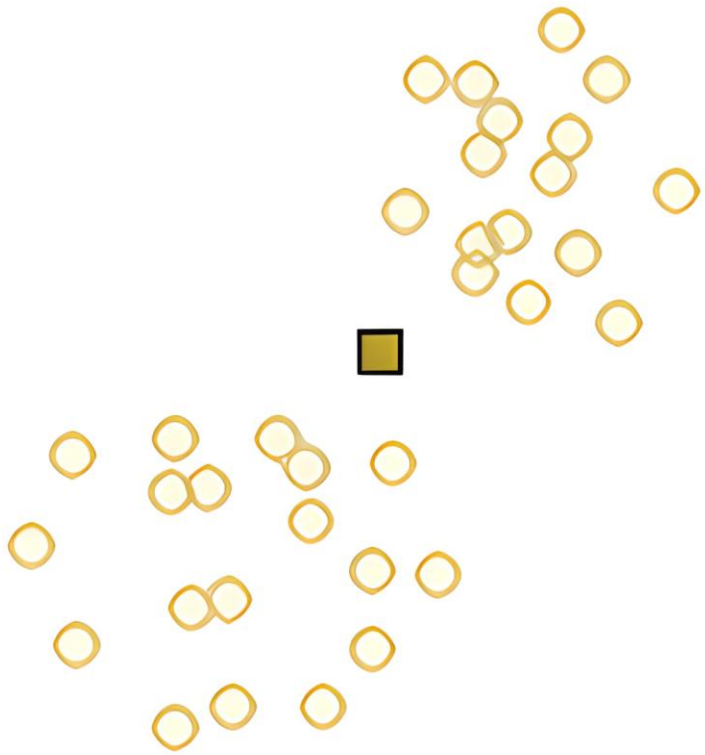
Repeat until convergence
(Cluster centers do not change much)

Clustering

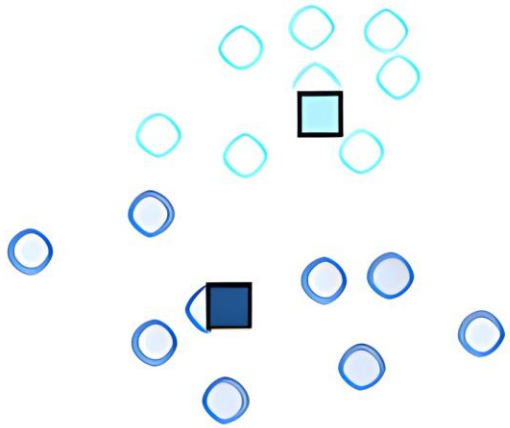


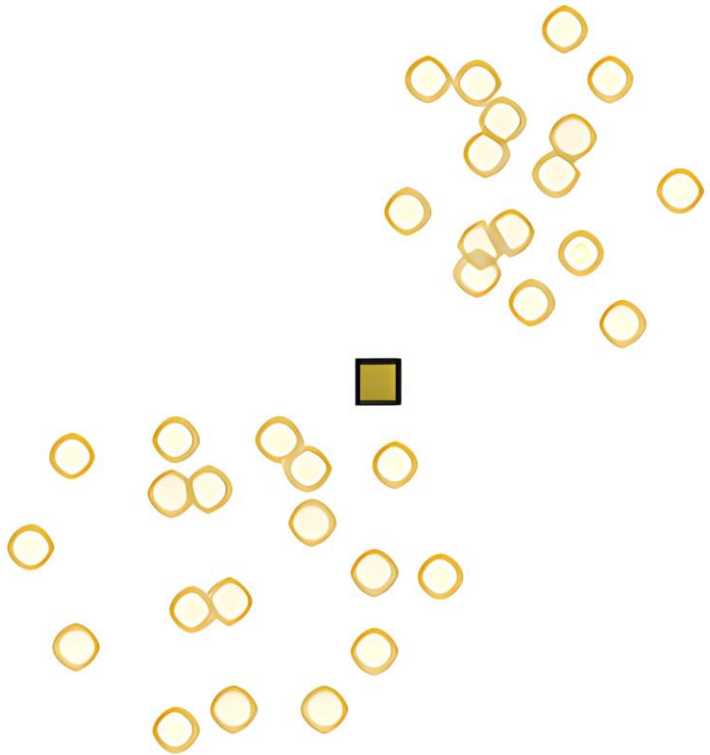
Issues?



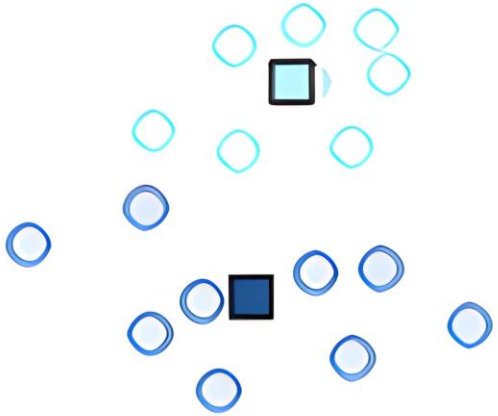


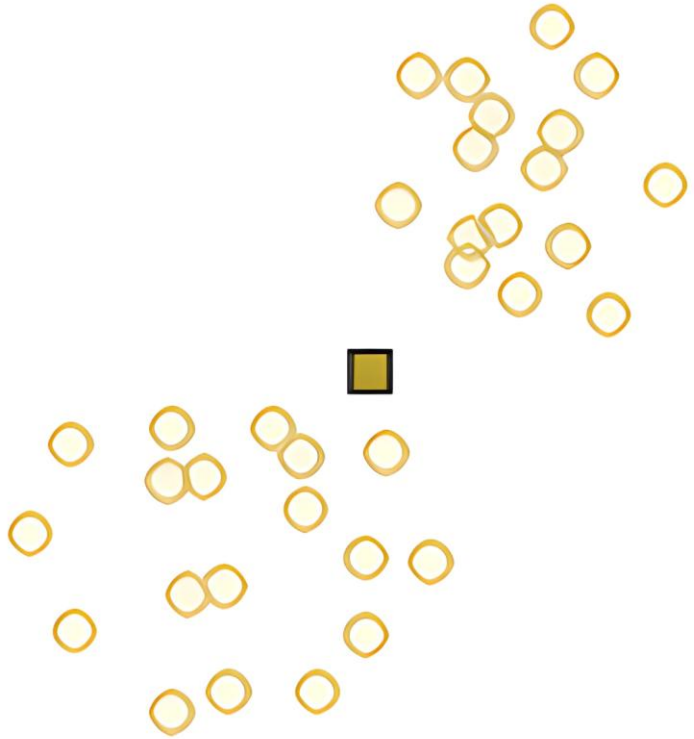
Issues?



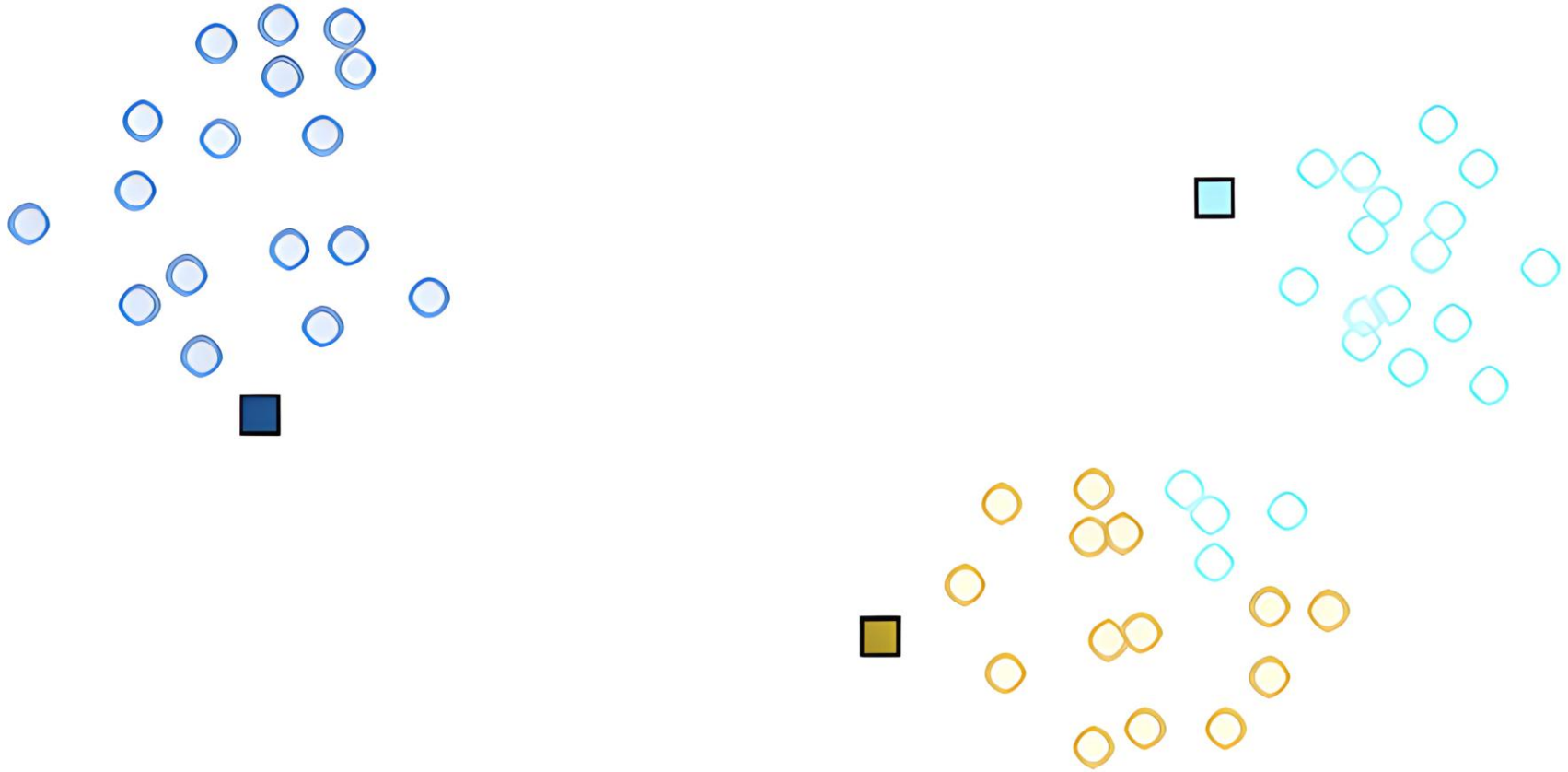


Issues?

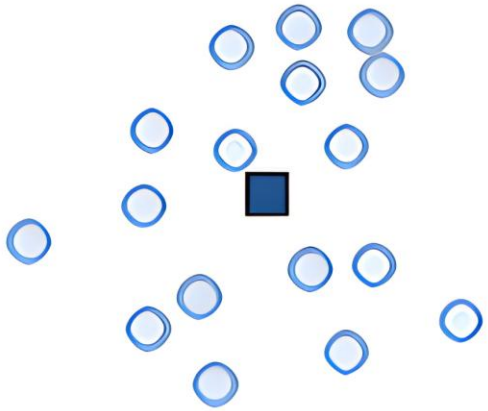


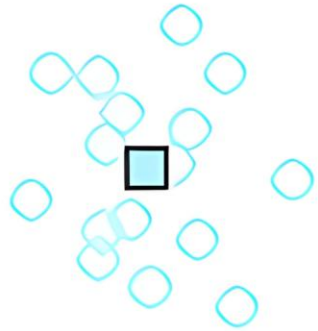


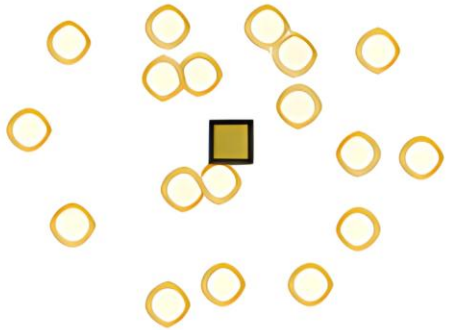
Let’s Change Initialization



Let’s Change Initialization

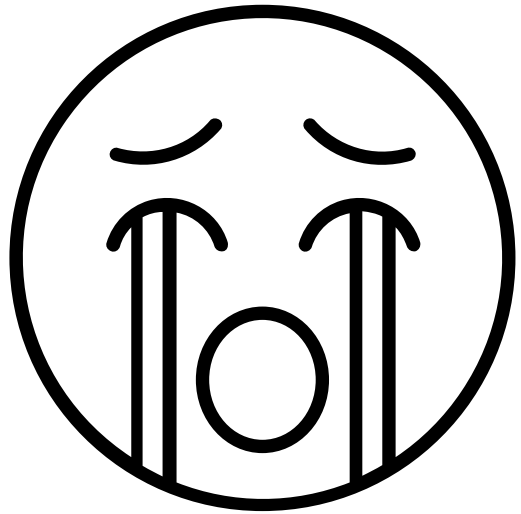




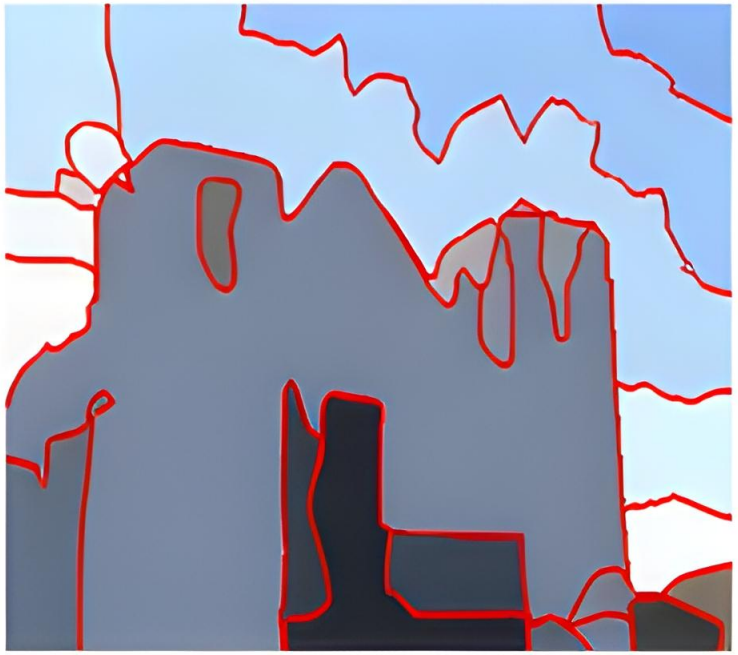


Issues?

- Needs a good initialization
  - Fixed by prior
- Needs a good distance metric
  - Domain knowledge
- Hard clusters
  - Fixed by GMM (Gaussian Mixture Model)
- No higher order semantics
  - Can be fixed by clustering in higher space
    - This has other issues



K-Means Color Clustering



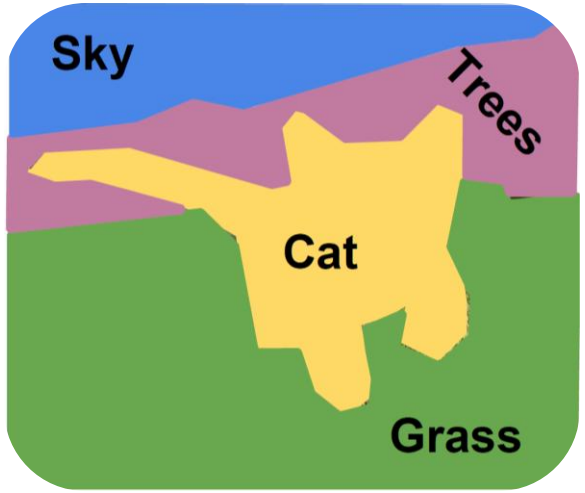
Group of **“similar” pixels** are called **Superpixels**

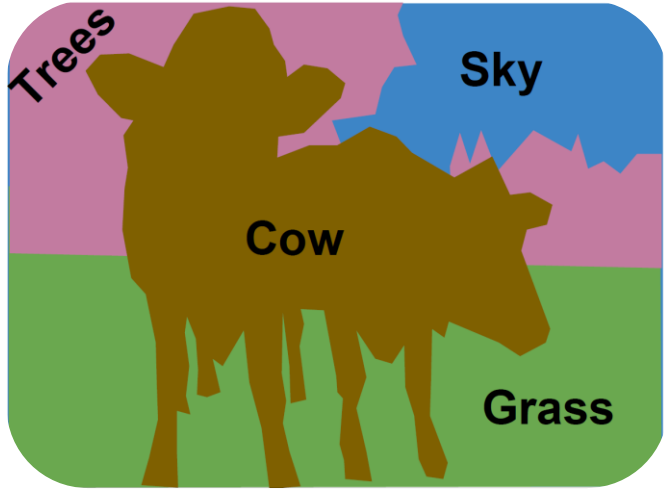
Super-resolution using DeepAI

Reddy, S., M. S. U. Faruq, and K. R. Lakshmi. "Image segmentation by using linear spectral clustering." J. Telecommun. Syst. Manage 5.3 (2016): 1-5.

Semantic Segmentation

- Label each pixel in the image with a category label
- Doesn't differentiate instances

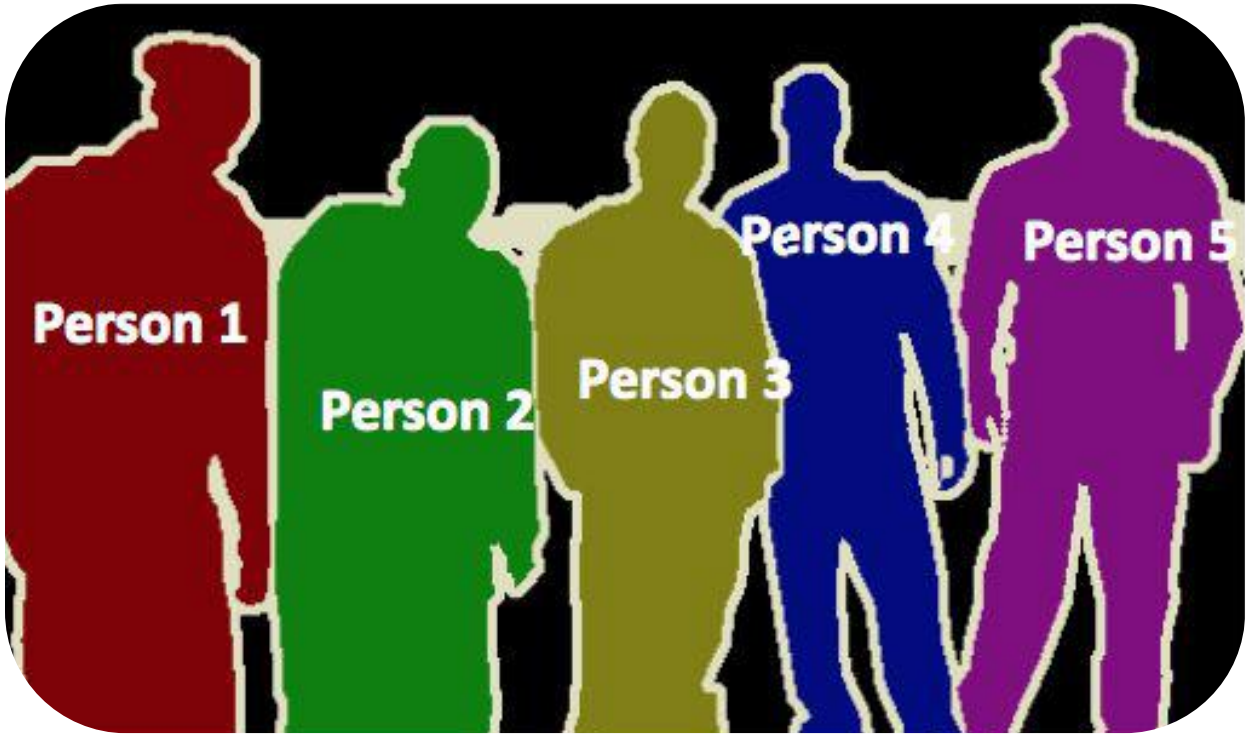




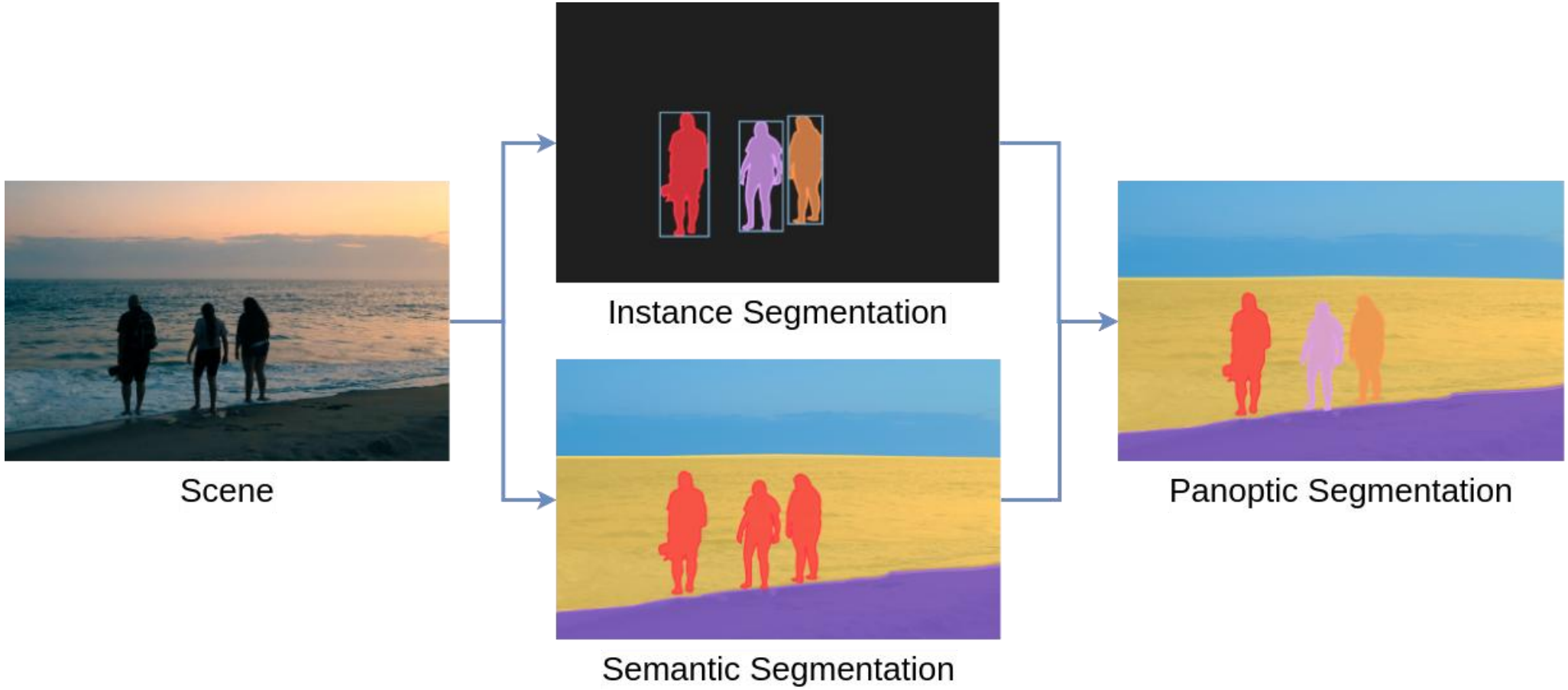
Slides adapted from Stanford's CS231n

Instance Segmentation

- Label each foreground pixel with object and instance
- Object detection + semantic segmentation



Types of Segmentation



- **Semantic segmentation:** treats multiple objects within a single category as one entity
- **Instance segmentation:** identifies individual objects within these categories

Slide adapted from Naitri Rajyaguru

Panoptic Segmentation

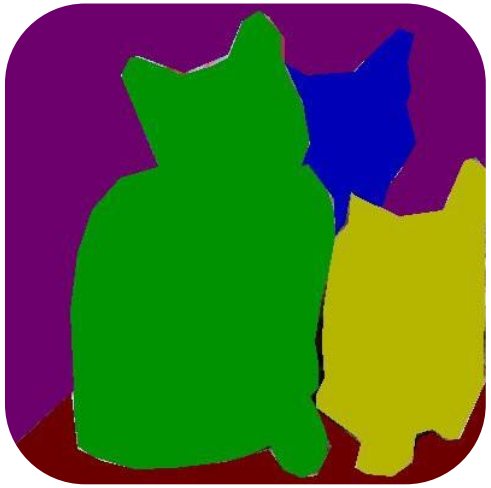
- Classify all the pixels in the image as belonging to a class label
- Identify what instance of that class they belong to

Semantic Segmentation

+

Instance Segmentation

=



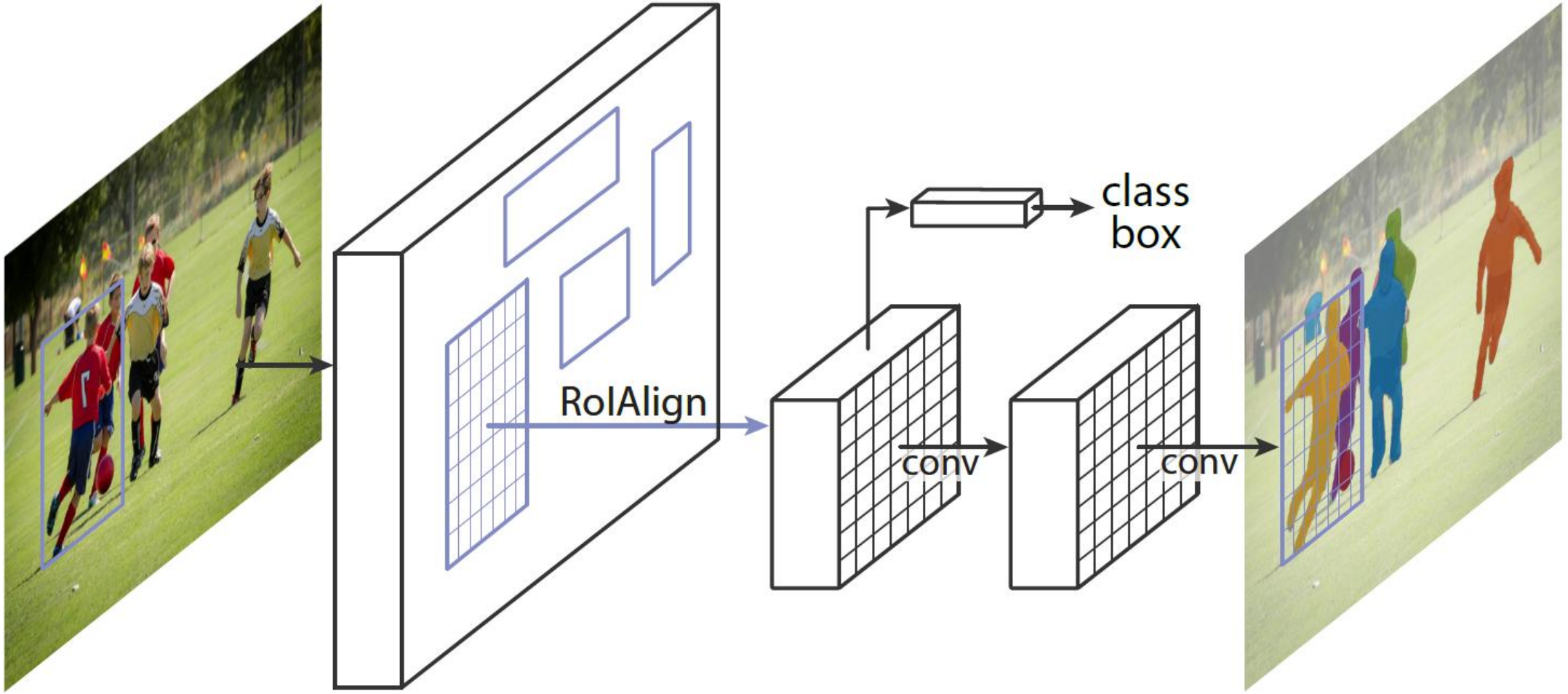
Panoptic Segmentation

Semantic Segmentation

How would you do this classically?

Classify each cluster using HoG ((Histogram of Oriented Gradients)**) + SVM or something**

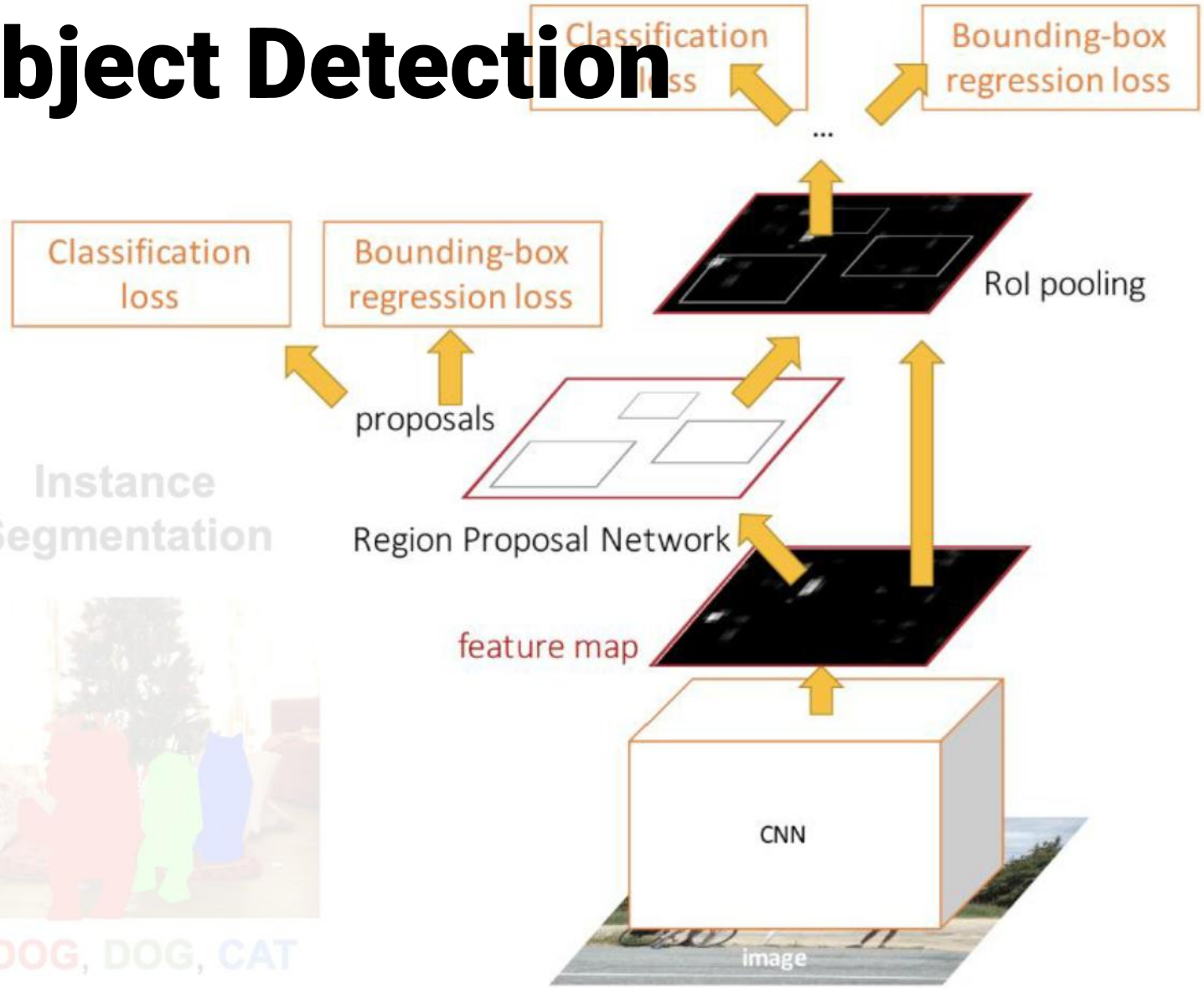
Mask R-CNN
Instance Segmentation



He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

Recall Object Detection

Faster R-CNN



Object Detection

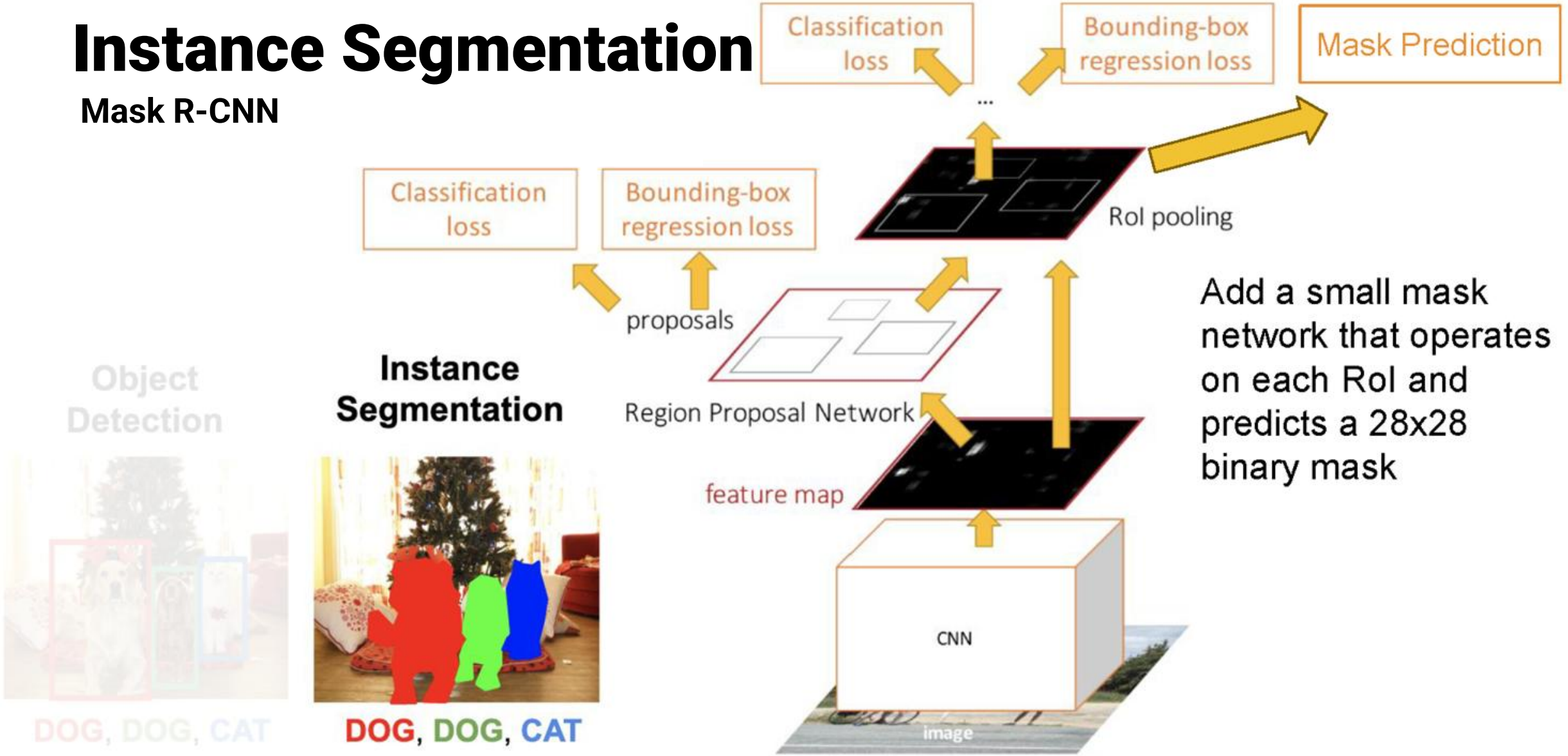
DOG, DOG, CAT

Instance Segmentation

DOG, DOG, CAT

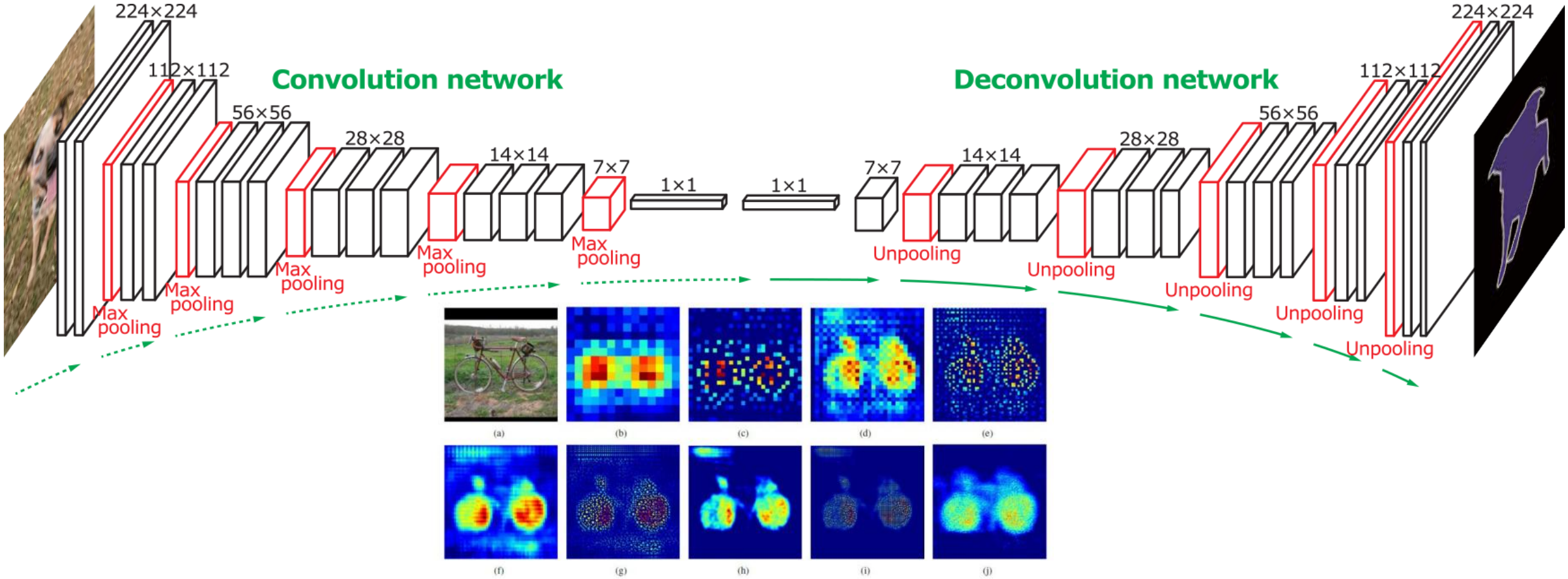
Instance Segmentation

Mask R-CNN



He et al, "Mask R-CNN", ICCV 2017

Encoder-Decoder Network

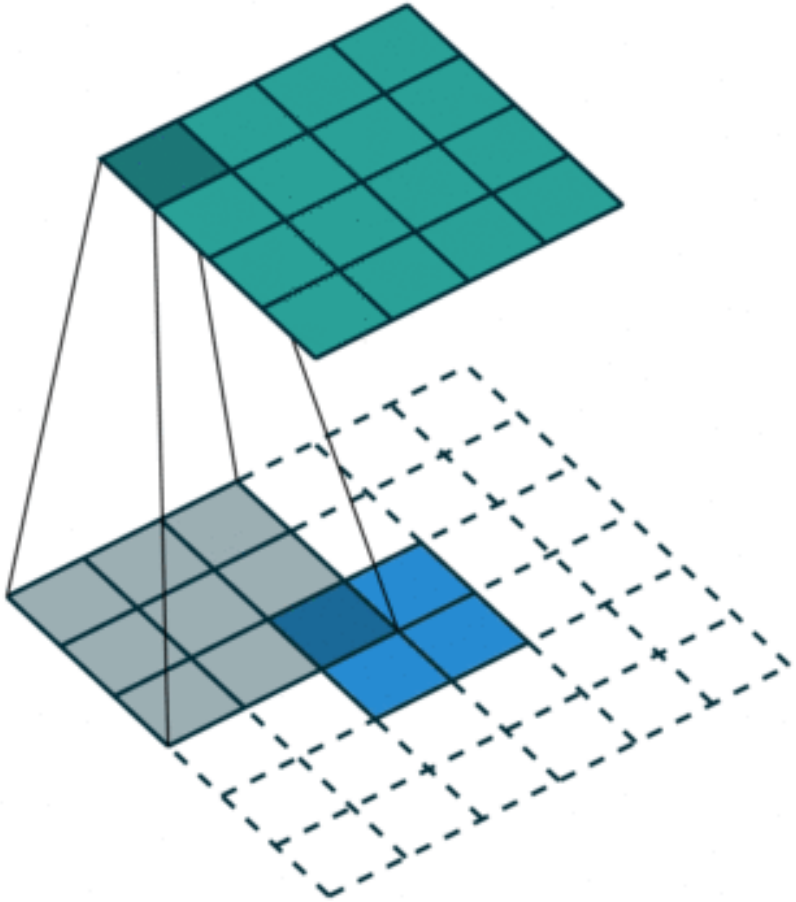


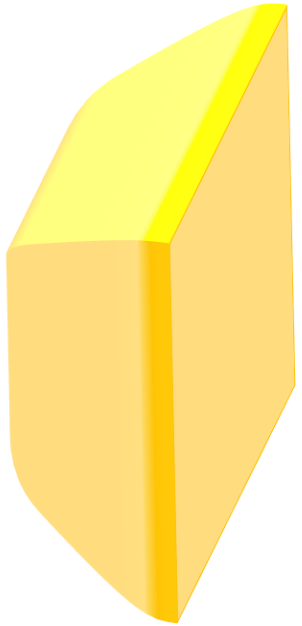
Slide adapted from UMN's CSCI5561

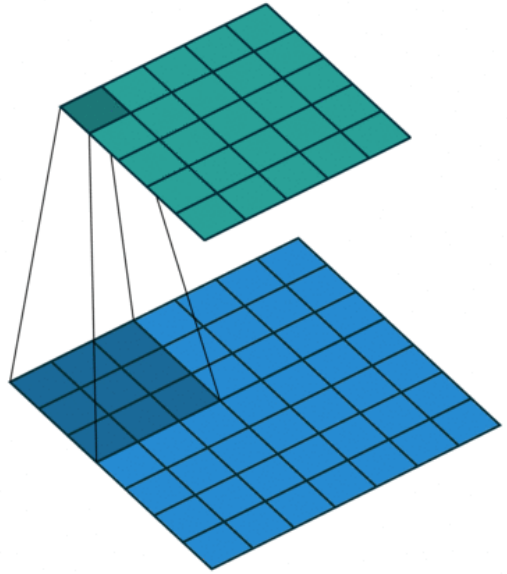
Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning deconvolution network for semantic segmentation." Proceedings of the IEEE international conference on computer vision. 2015.

Recall Transposed Convolution

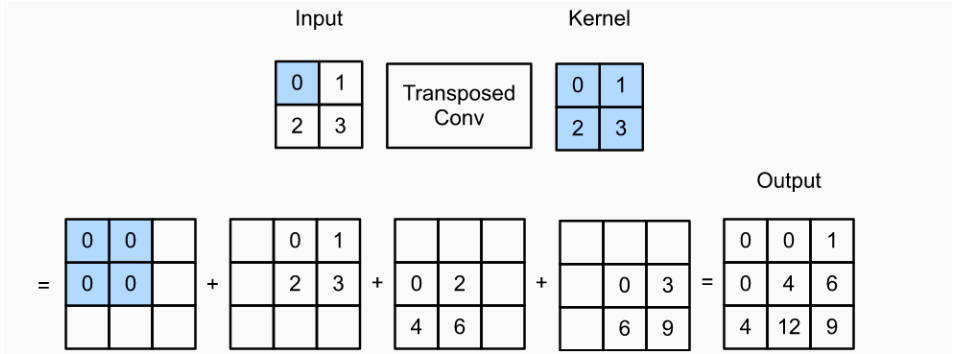
AKA Deconvolution





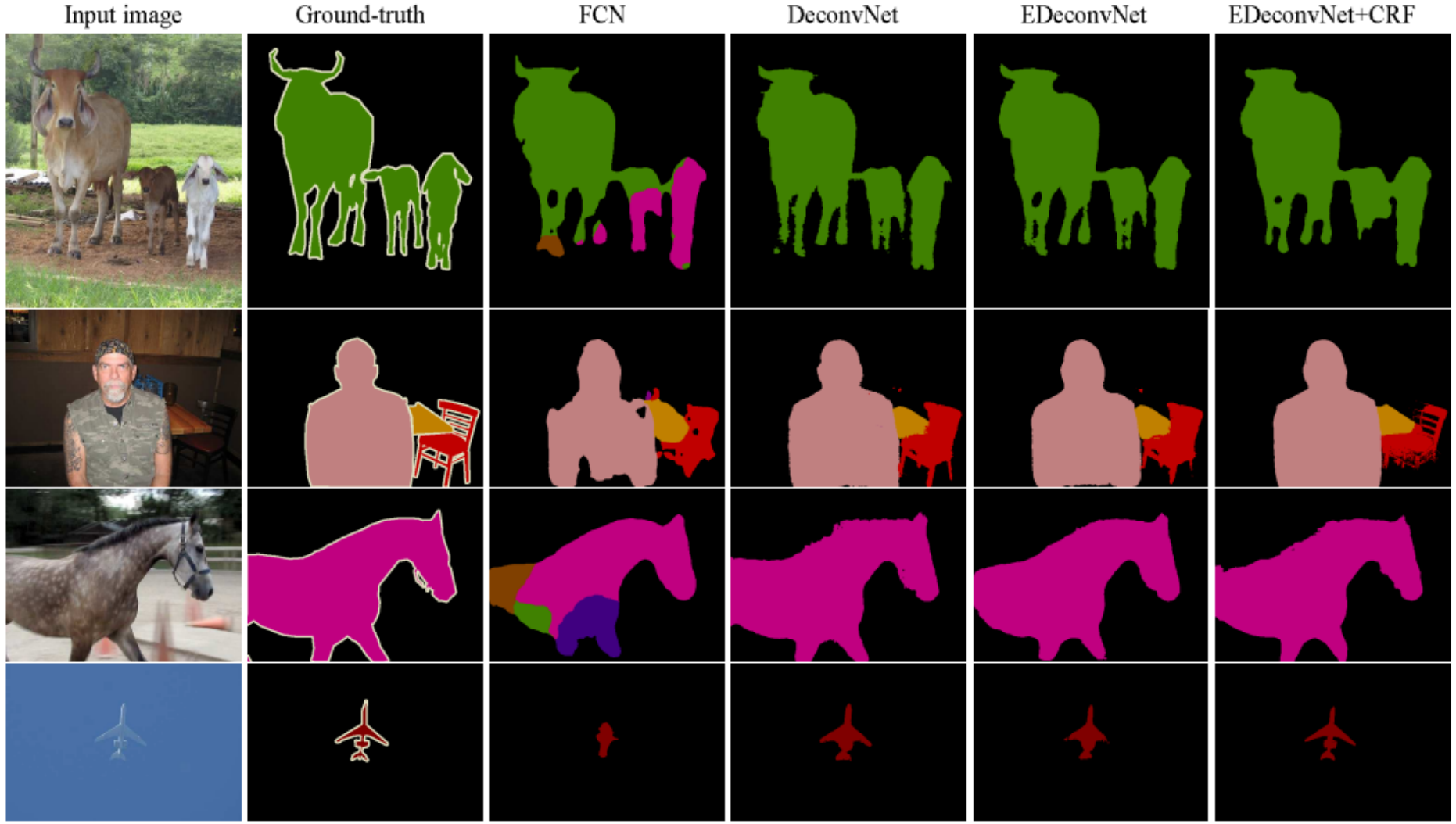


Padded Decovolution



https://github.com/vdumoulin/conv_arithmetic

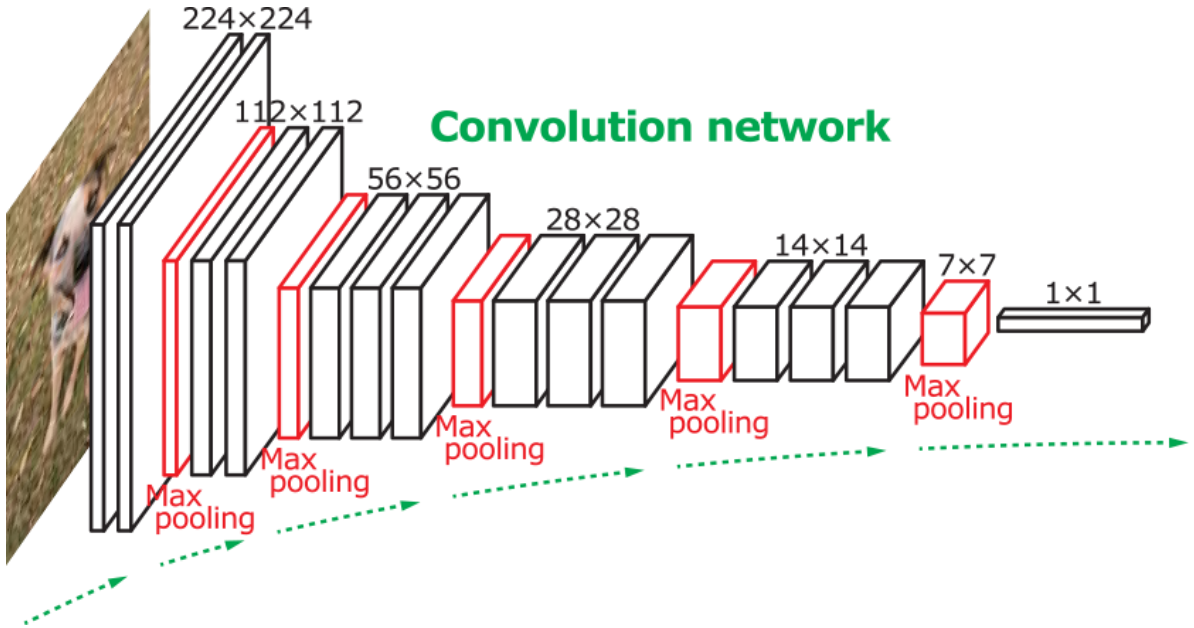
Encoder-Decoder Network



Slide adapted from UMN's CSCI5561

Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning deconvolution network for semantic segmentation." Proceedings of the IEEE international conference on computer vision. 2015.

Receptive Field Vs Resolution



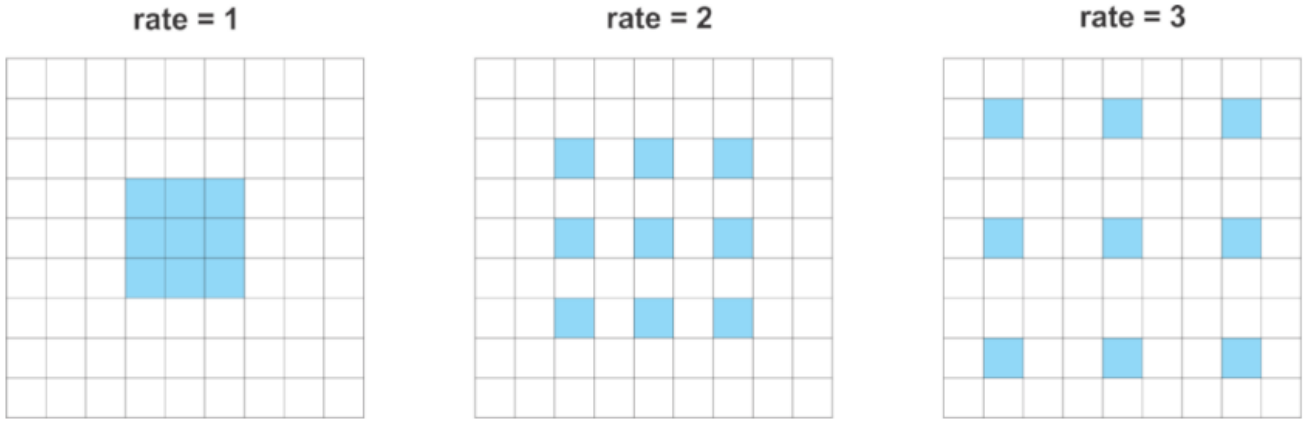
Larger receptive field ⇒ Bigger spatial context
⇒ Reduce resolution

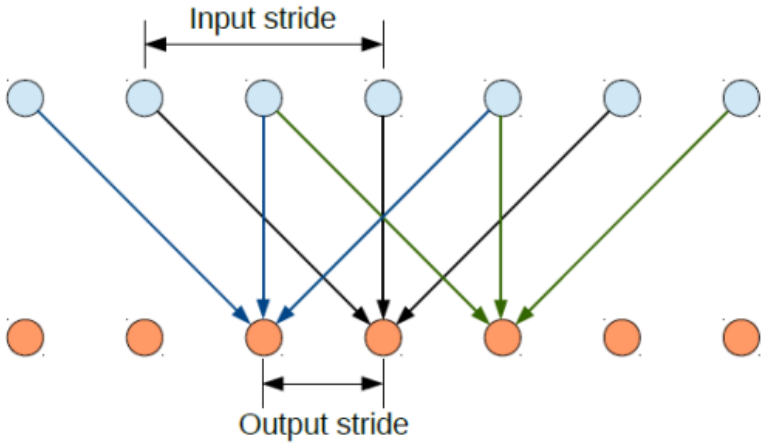
Higher output resolution ⇒ Reducing receptive field

We want higher resolution and larger receptive field :/

Receptive field refers to the region of the input image that a particular neuron in a convolutional layer is “looking at” or taking into account when making its predictions or feature extractions

New Trick: Dilated Convolution

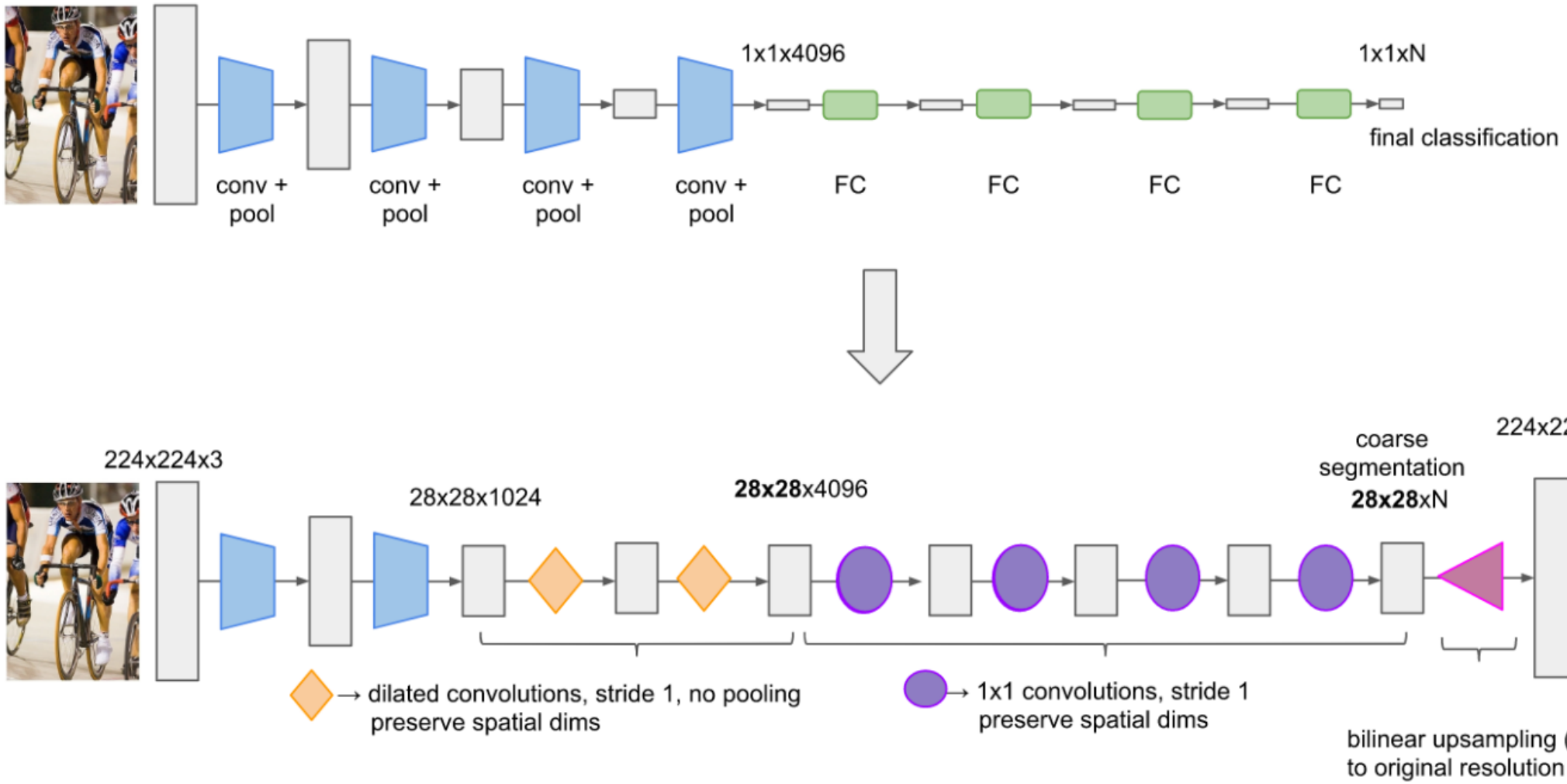




Enlarge receptive field without reducing resolution ☺

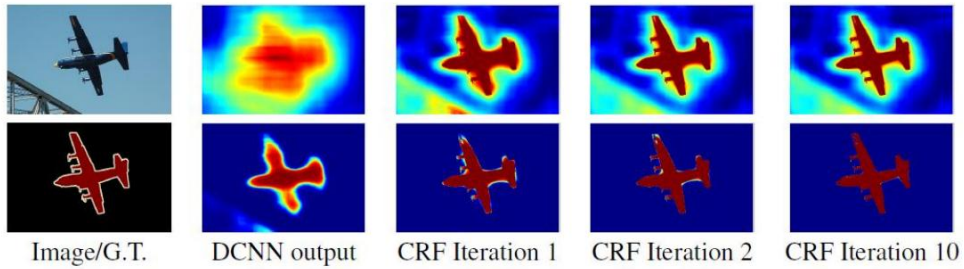
Chen, Liang-Chieh, et al. "Semantic image segmentation with deep convolutional nets and fully connected crfs." ICLR 2015.

DeepLab

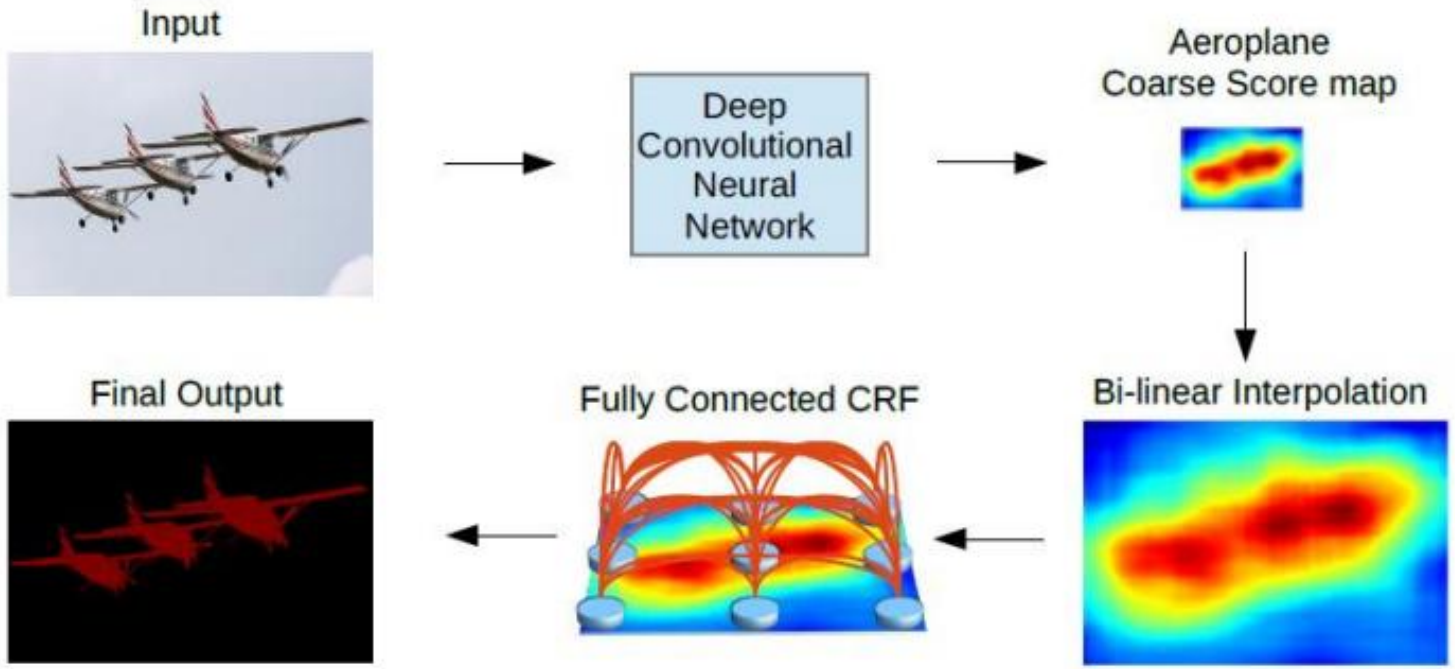


Chen, Liang-Chieh, et al. "Semantic image segmentation with deep convolutional nets and fully connected crfs." ICLR 2015.

DeepLab

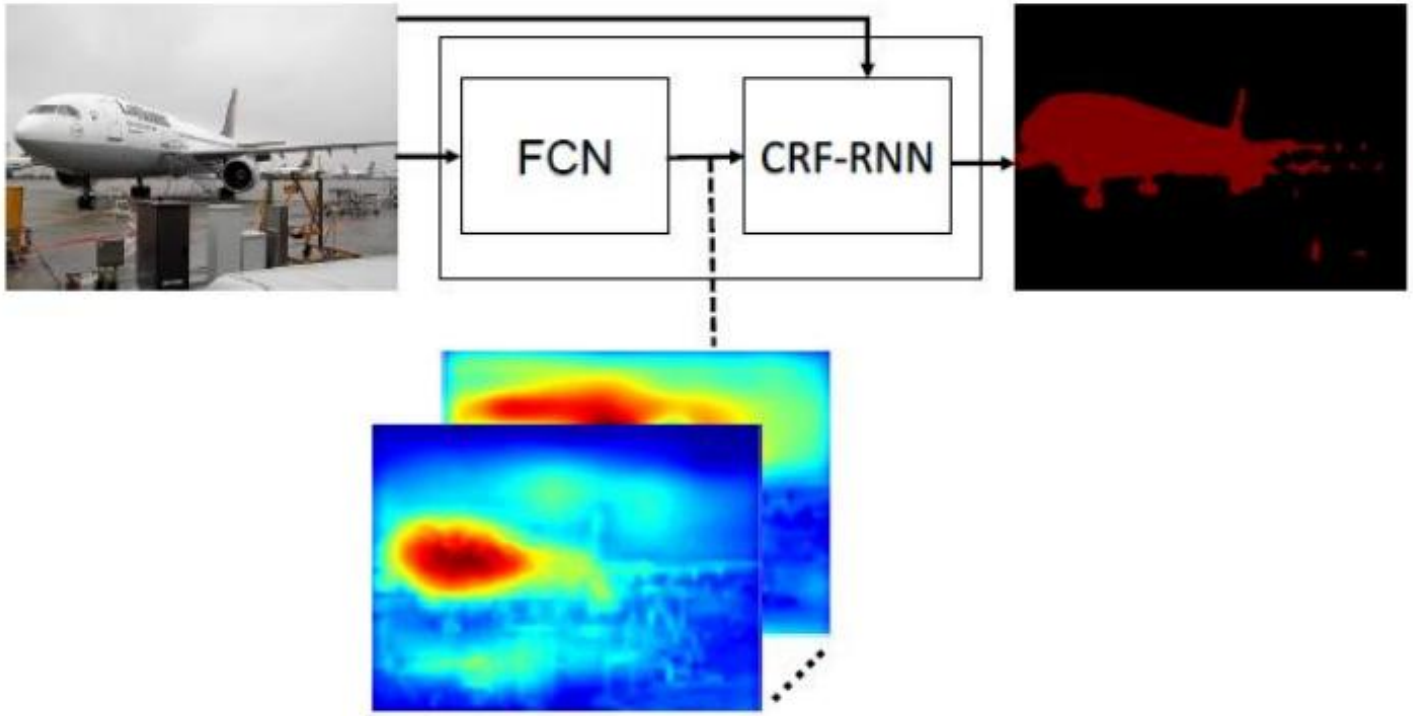


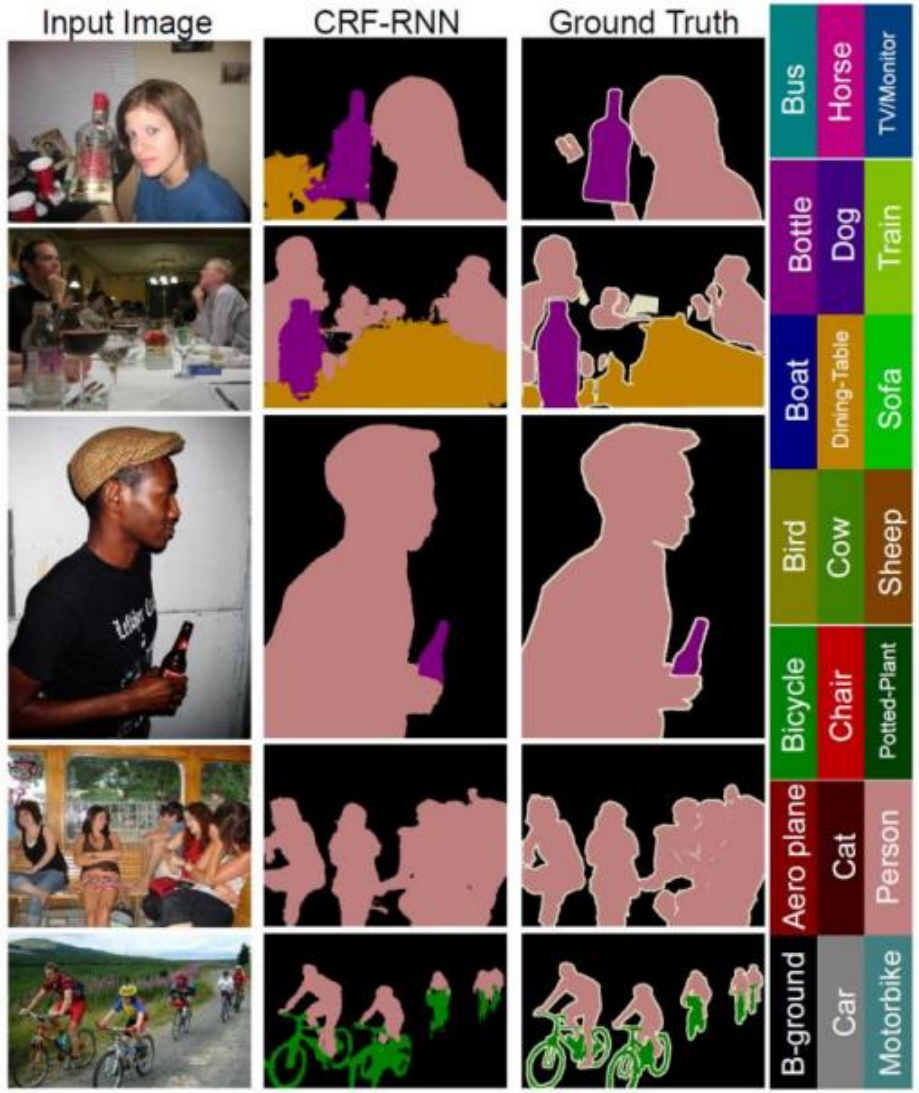
Post-process using CRFs (Conditional Random Field)



Chen, Liang-Chieh, et al. "Semantic image segmentation with deep convolutional nets and fully connected crfs." ICLR 2015.

CRFs as RNN

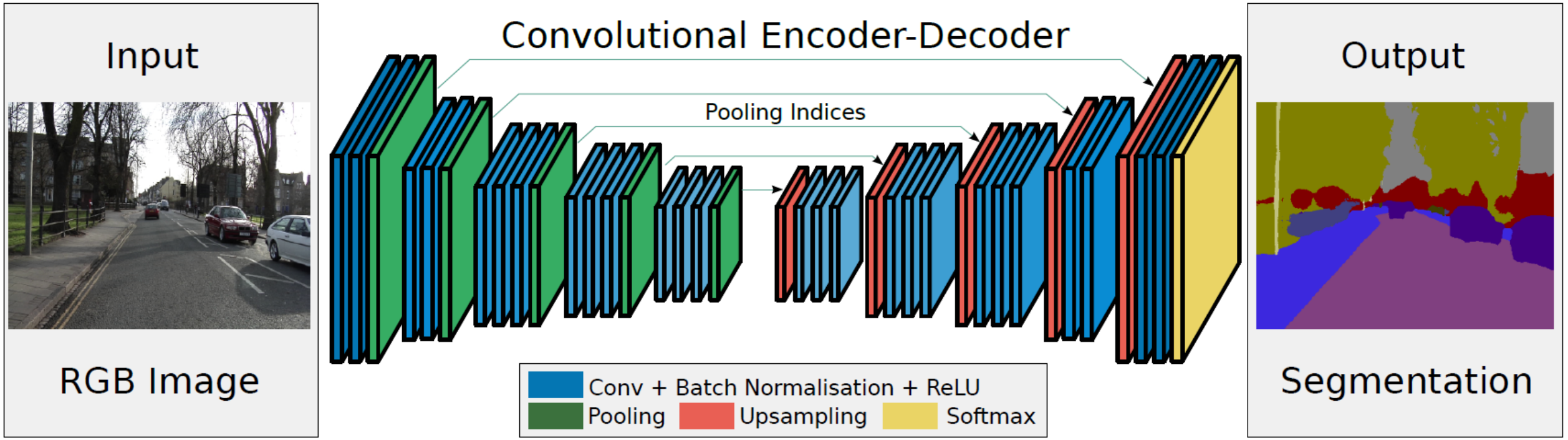




Zheng, Shuai, et al. "Conditional random fields as recurrent neural networks." *Proceedings of the IEEE international conference on computer vision*. 2015.

SegNet

Roberto Cipolla

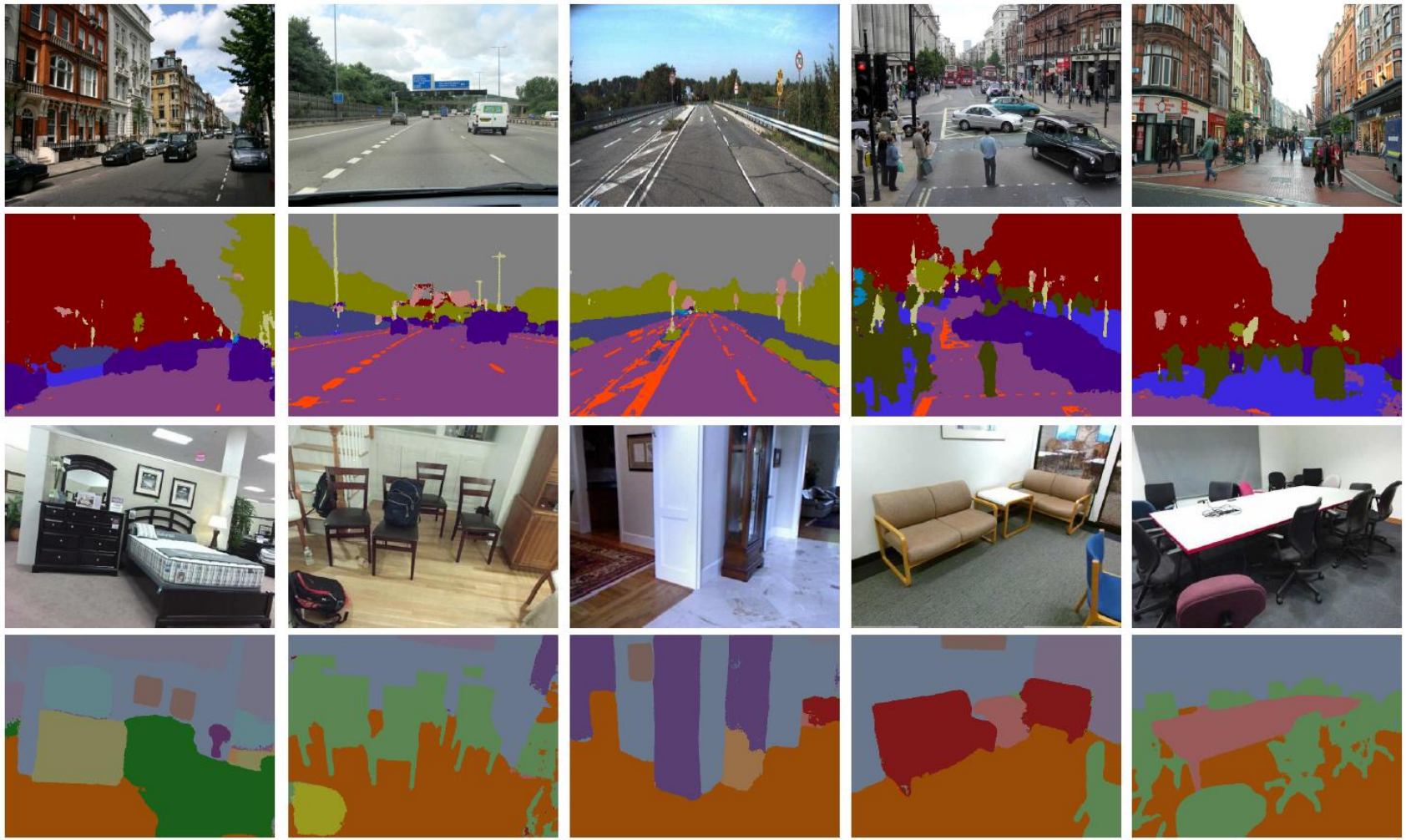


Cross-Entropy Loss

Learn about Transposed Convolution CSCI5561 Lec34 Semantic Segmentation (umn.edu)

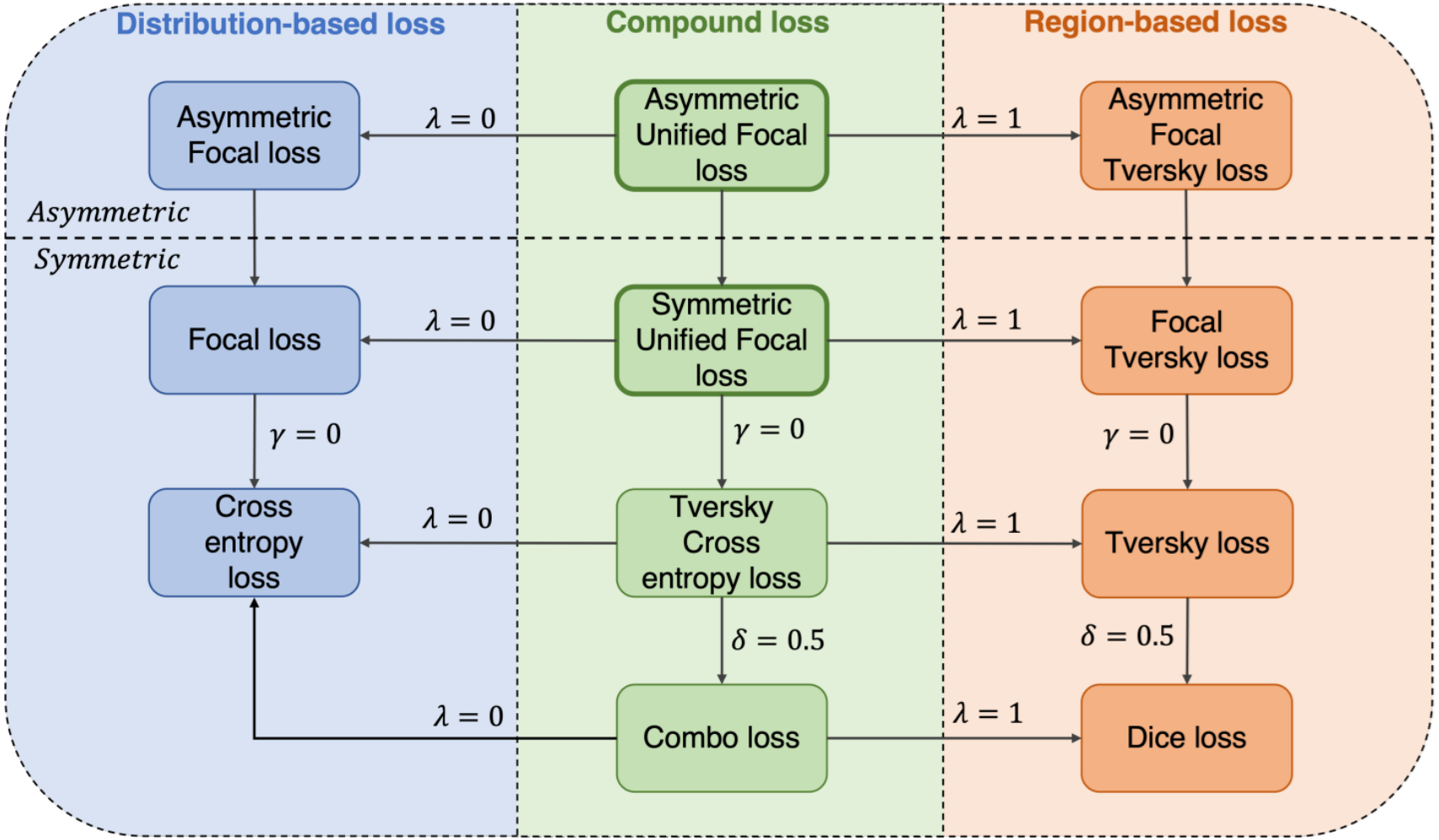
Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.

SegNet



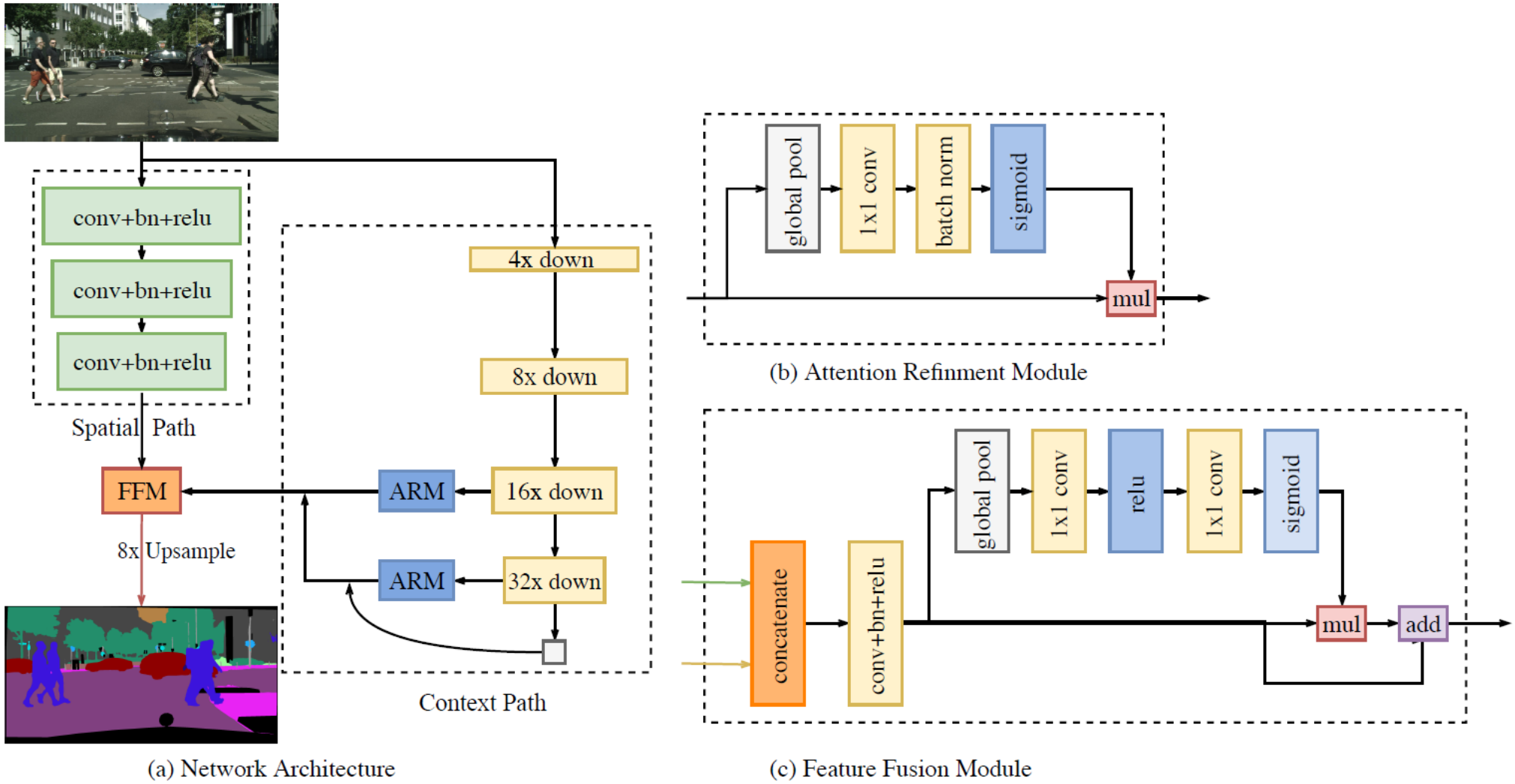
Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.

Losses



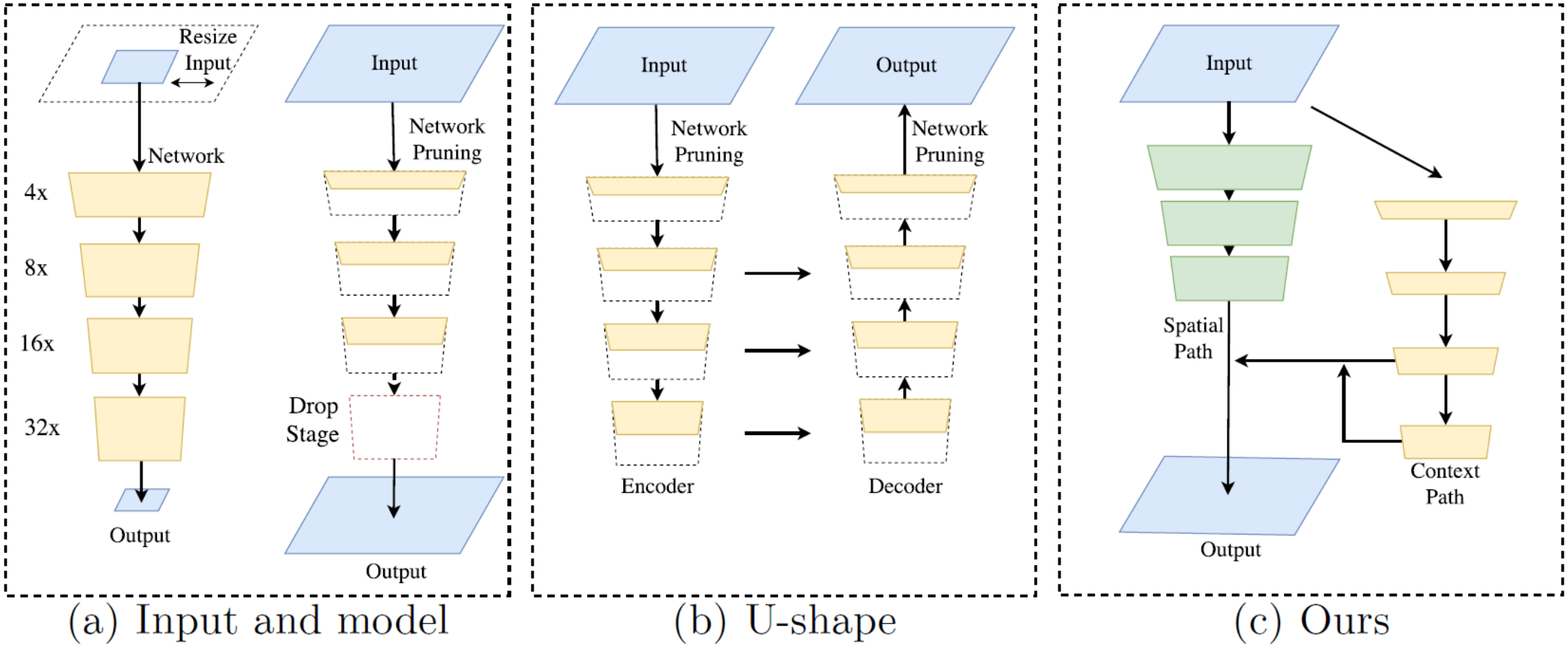
Yeung, Michael, et al. "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation." Computerized Medical Imaging and Graphics 95 (2022): 102026.

BiSeNet

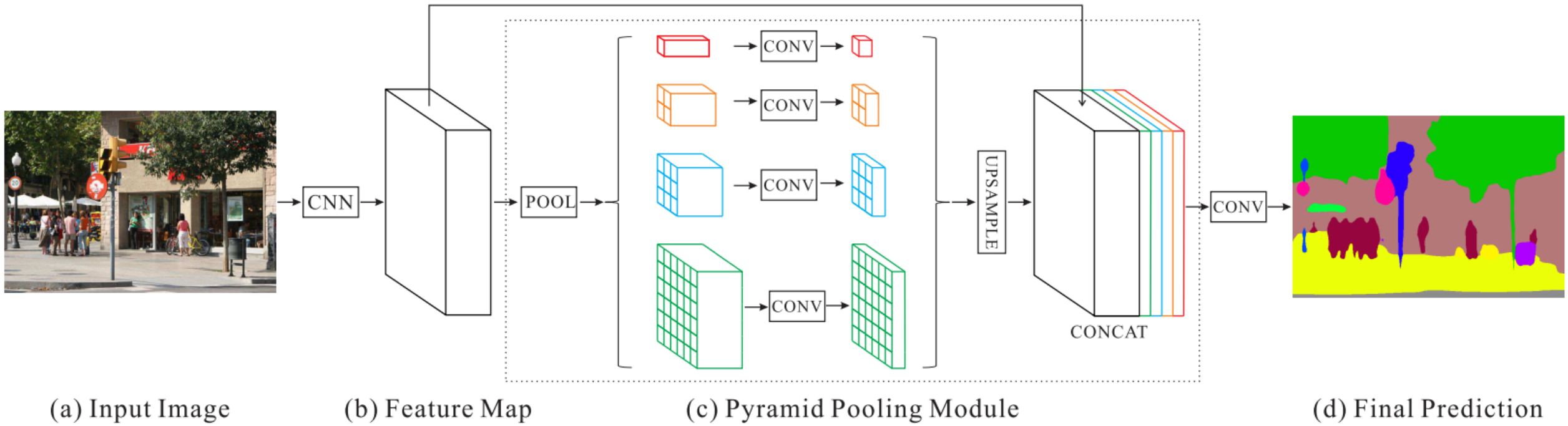


Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 325-341).

How to Speedup Networks?

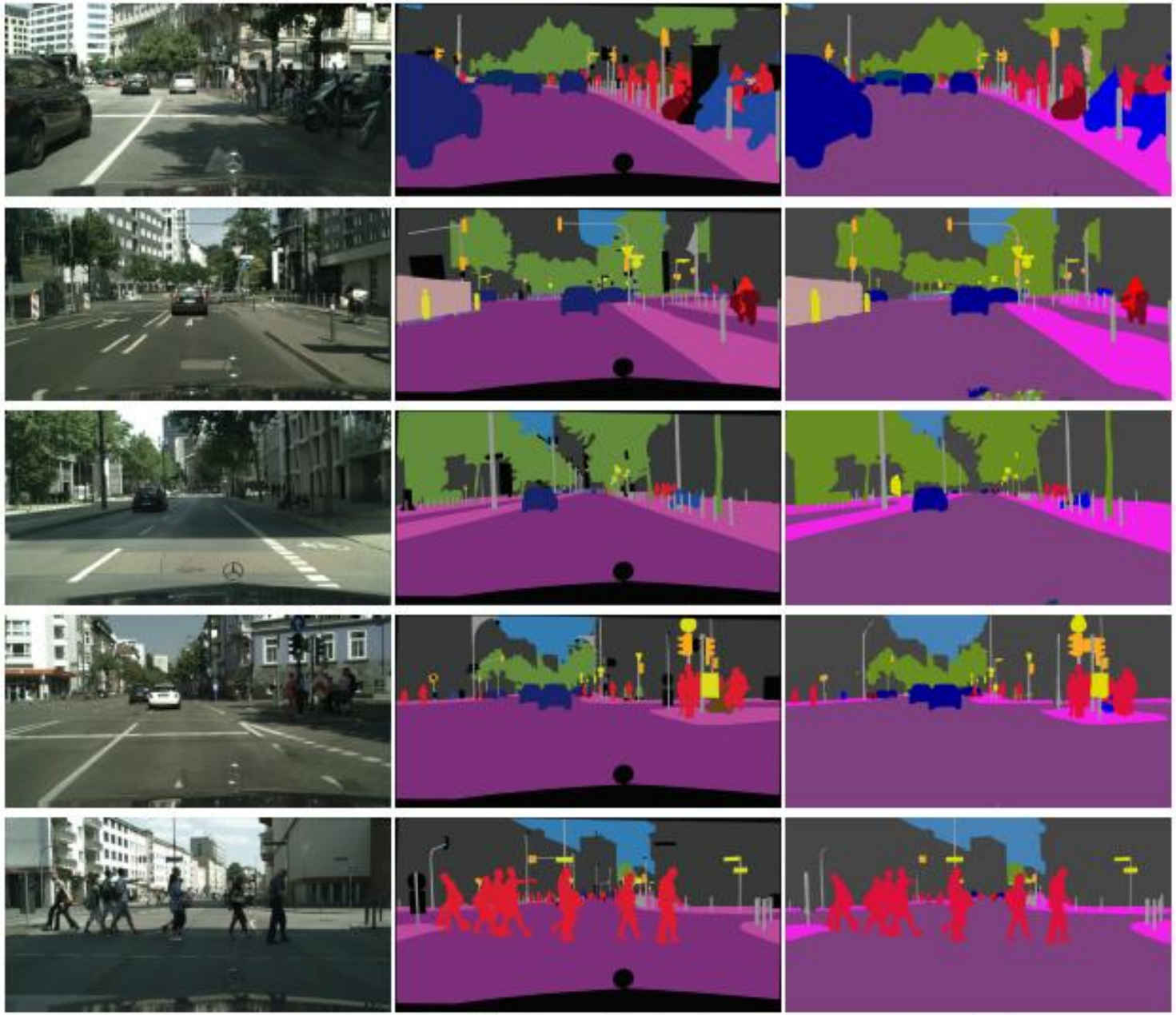


PSPNet



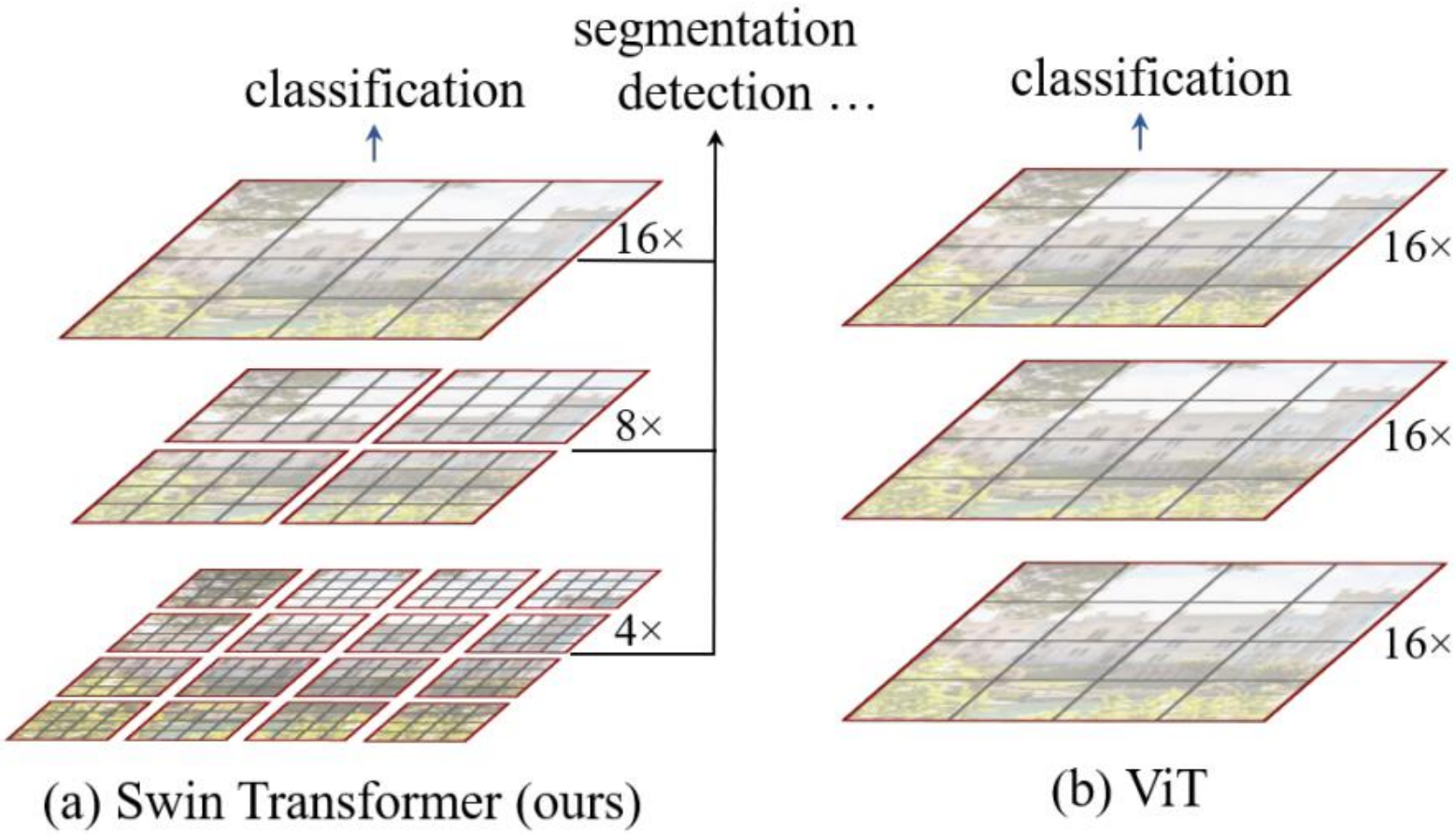
Zhao, Hengshuang, et al. "Pyramid scene parsing network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

PSPNet



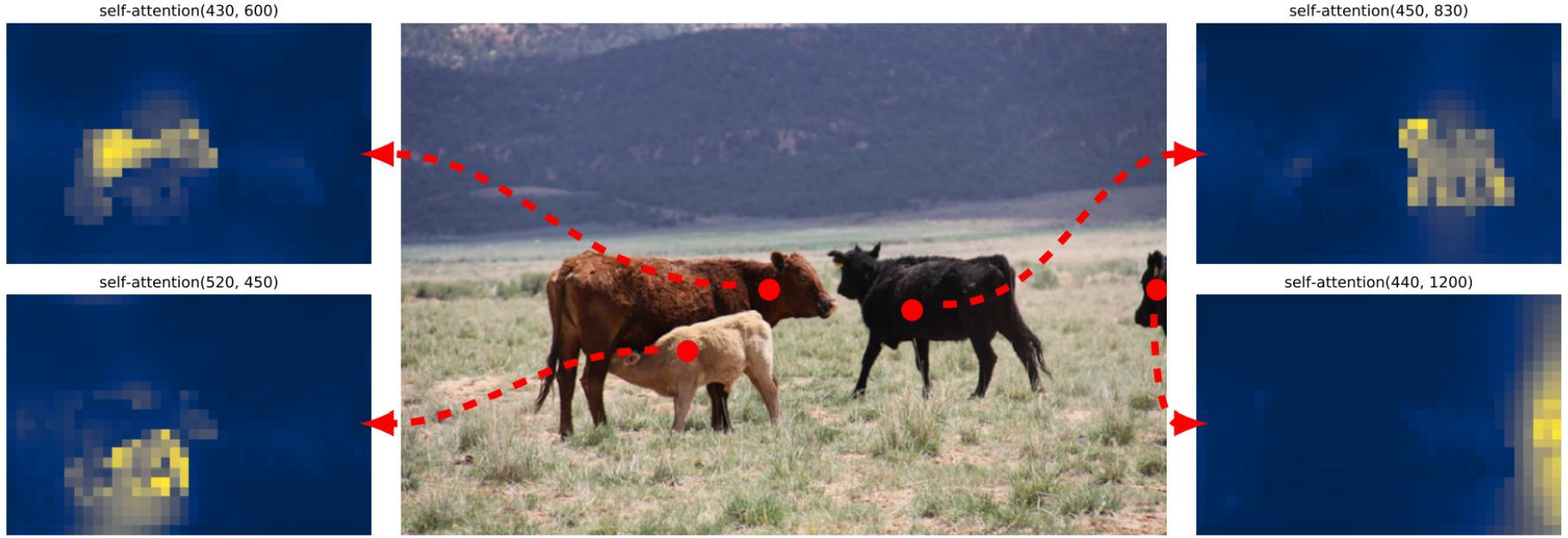
(a) Image (b) Ground Truth (c) PSPNet

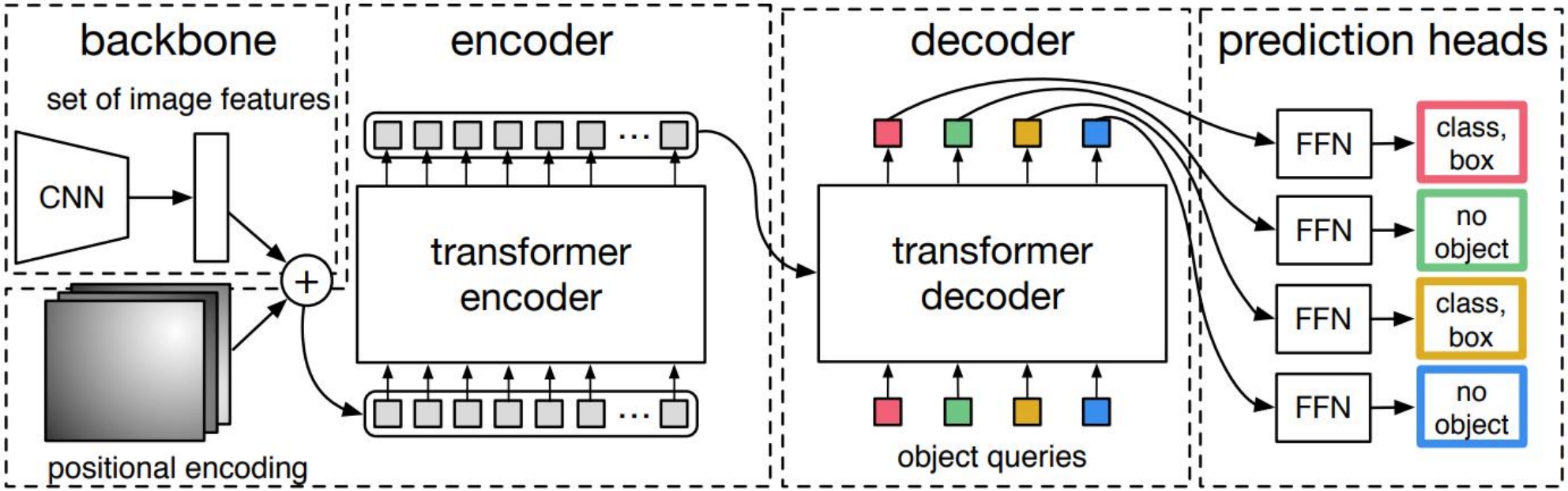
Today, You Use Transformers!

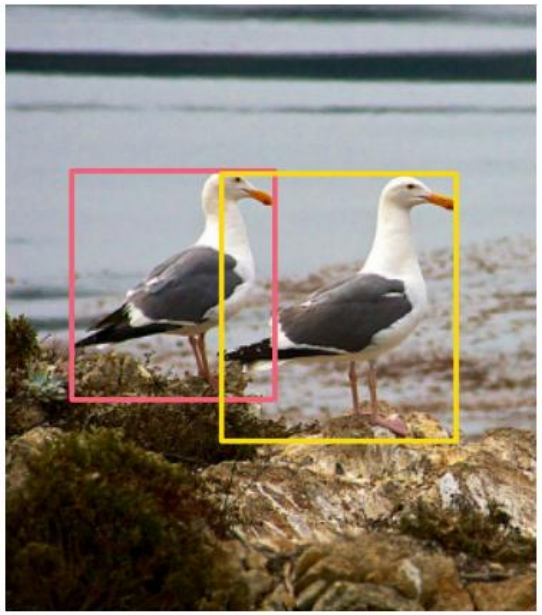


Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

DETR

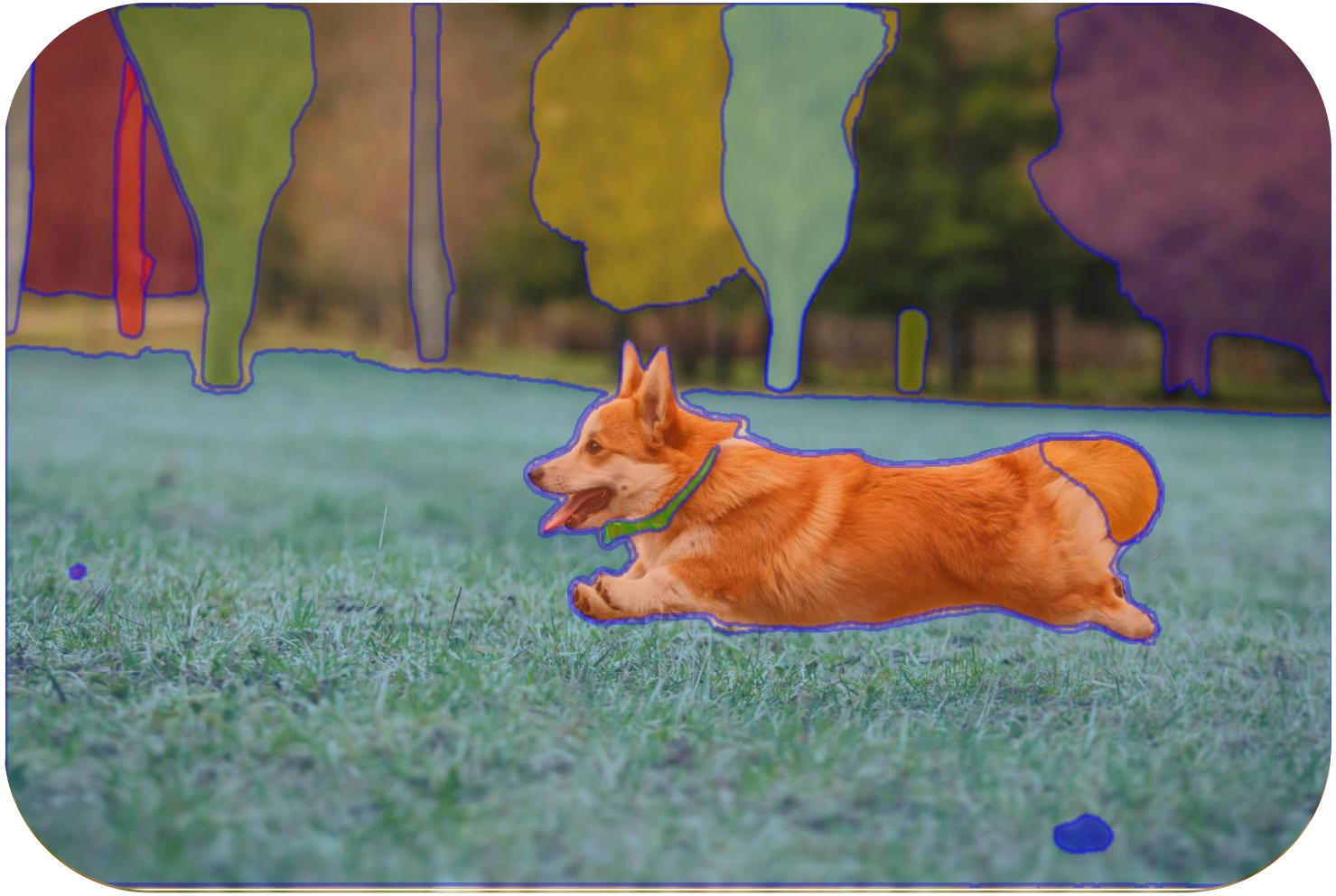






Carion, Nicolas, et al. "End-to-end object detection with transformers." European conference on computer vision. Springer, Cham, 2020.

SAM
Segment Anything Model



Segment Anything | Meta AI (segment-anything.com)

Kirillov, Alexander, et al. "Segment anything." arXiv preprint arXiv:2304.02643 (2023).

Segmentation Via Recognition

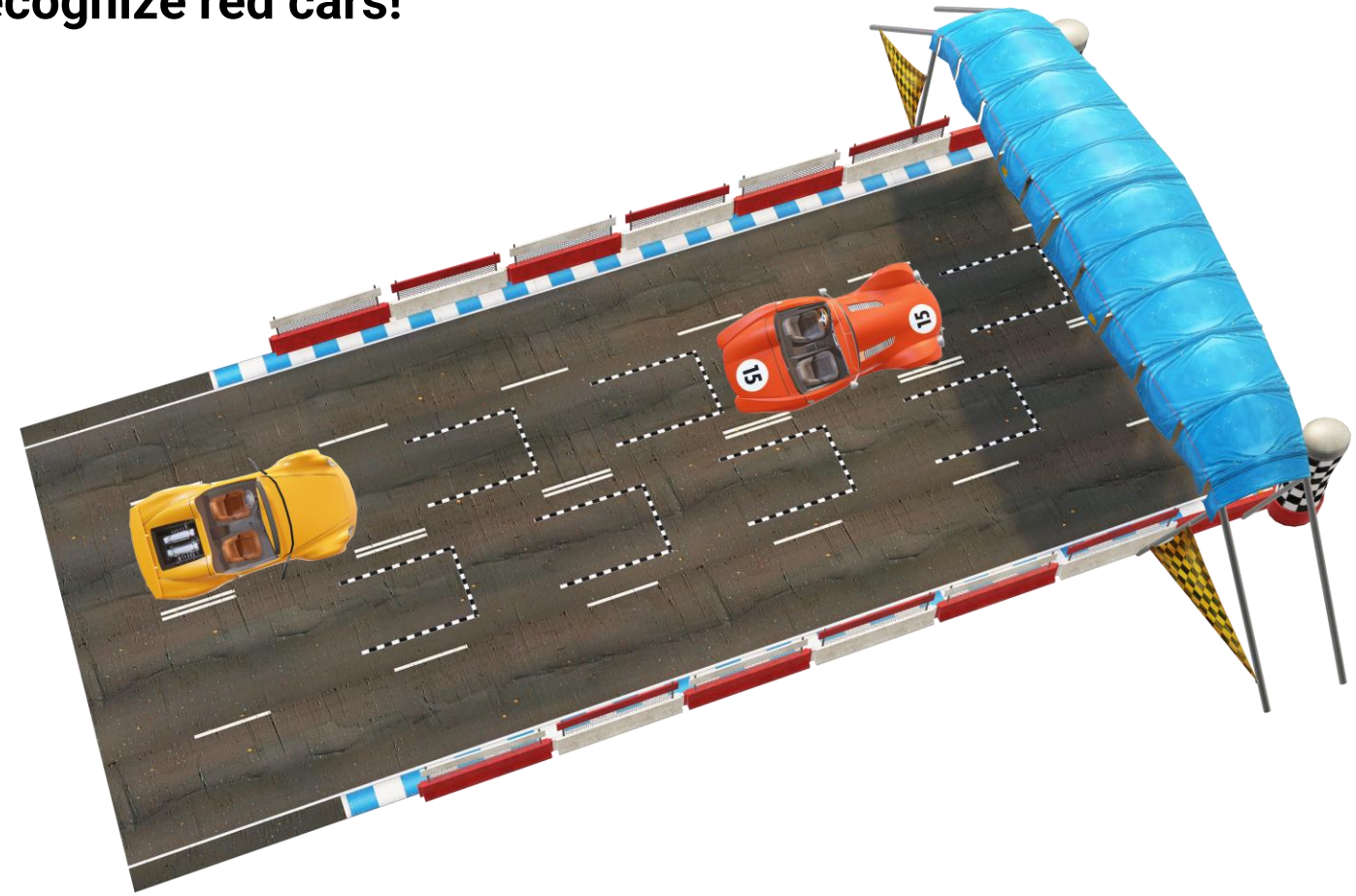
Why this might **NOT** work?
Well, cars don't have wings! So your dataset does not have upside down cars!

You cannot learn everything that can hit you!

How do you deal with this?
Segment using **Motion**!

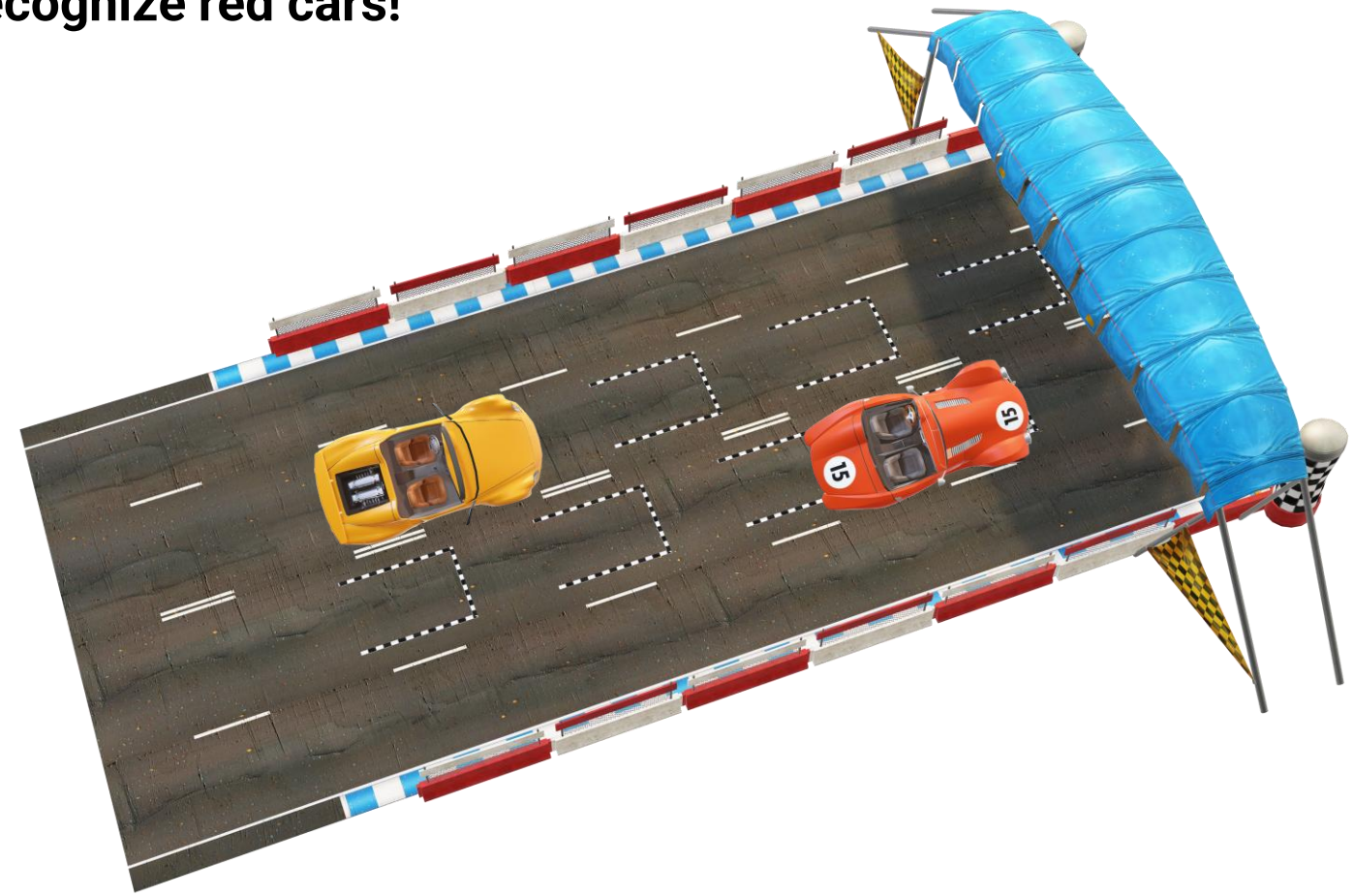
How To Segment?

Can’t recognize red cars!



How To Segment?

Can’t recognize red cars!



What cue can I use? Remember, scene and object(s) are moving!

Optical Flow

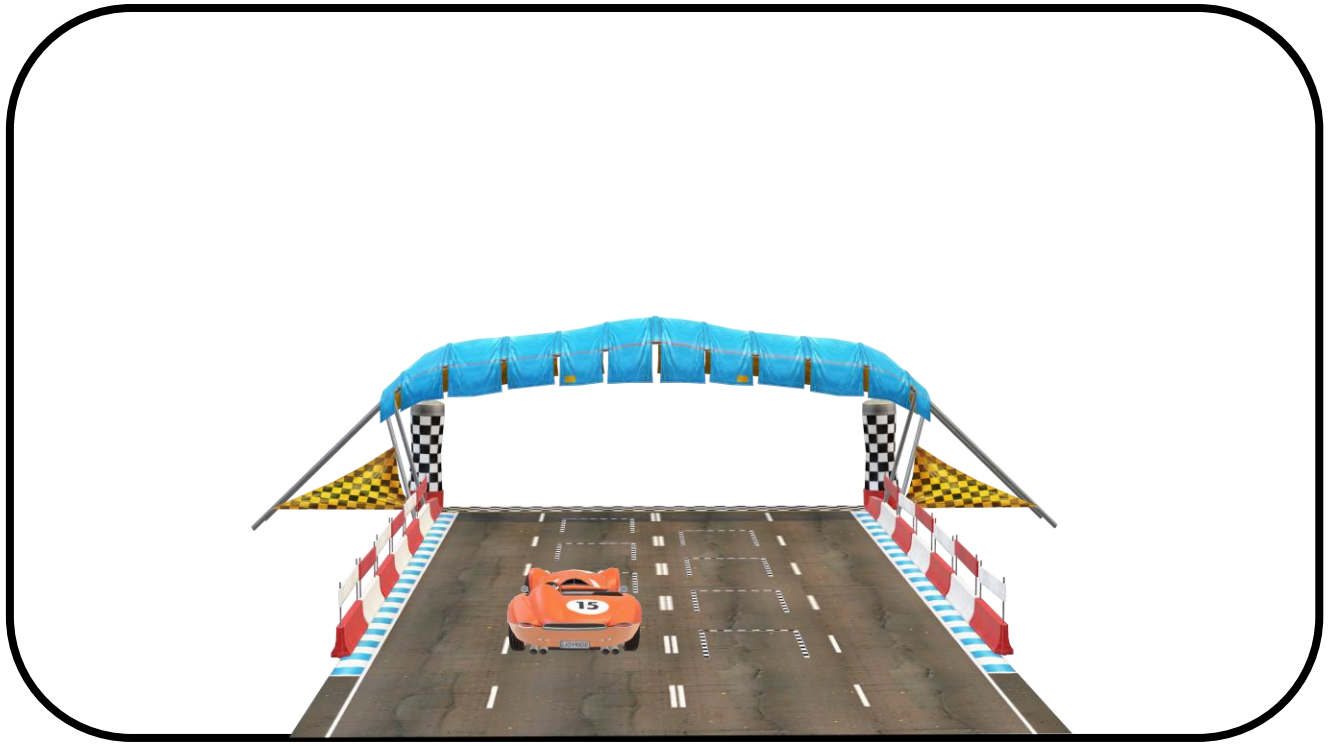


Image Plane

Optical Flow

Image Plane

Optical Flow

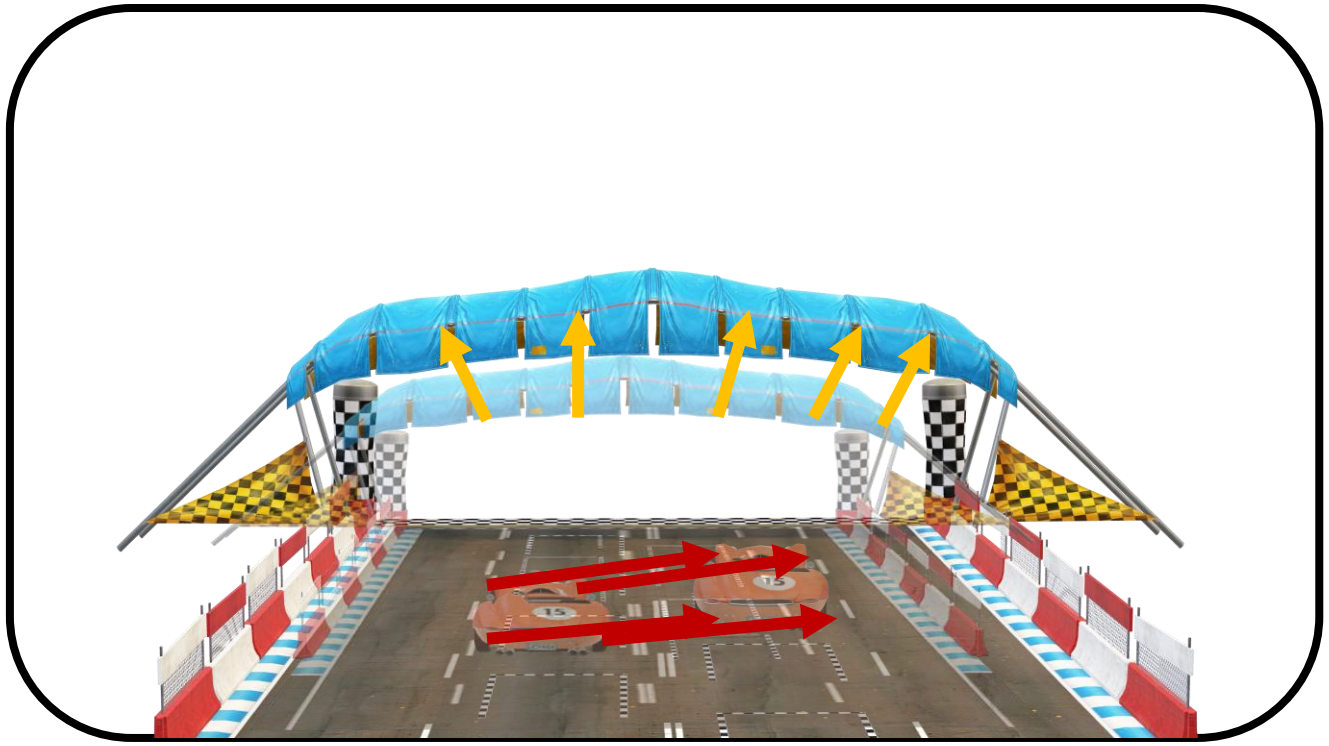


Image Plane

Optical Flow

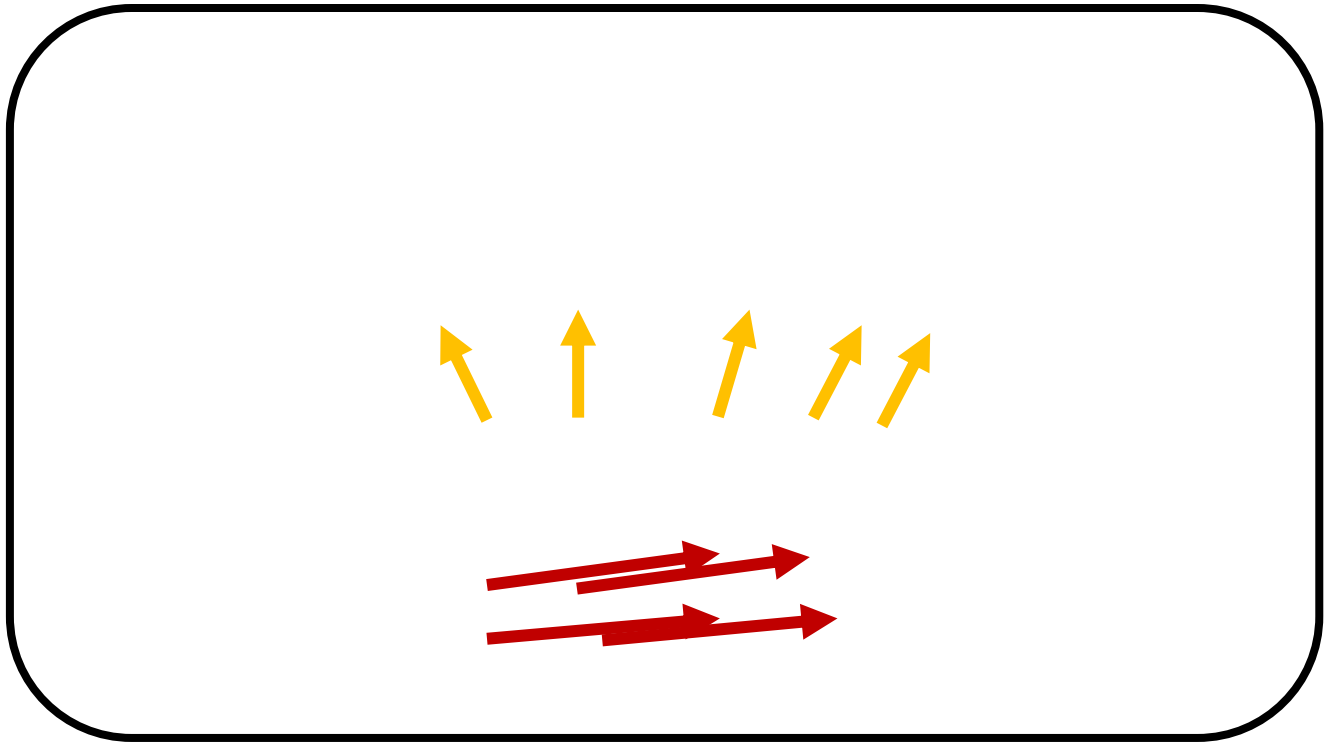


Image Plane

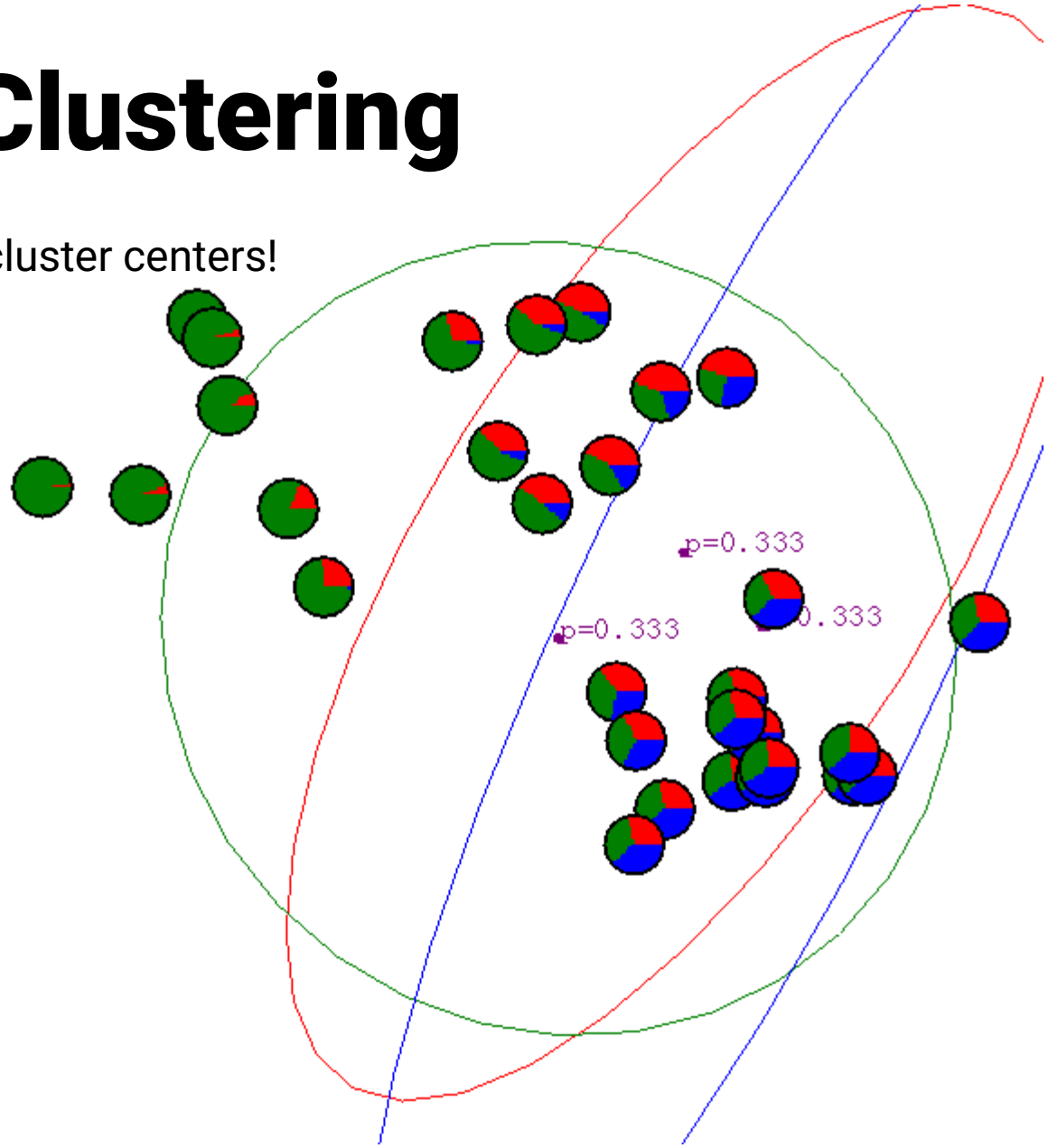
Can cluster flow direction and magnitude!

Segmenting multiple motions: **Multi-motion Segmentation (MMS)** Segmenting motions: **Motion Segmentation (MS)**

K-Means Clustering

Initialization

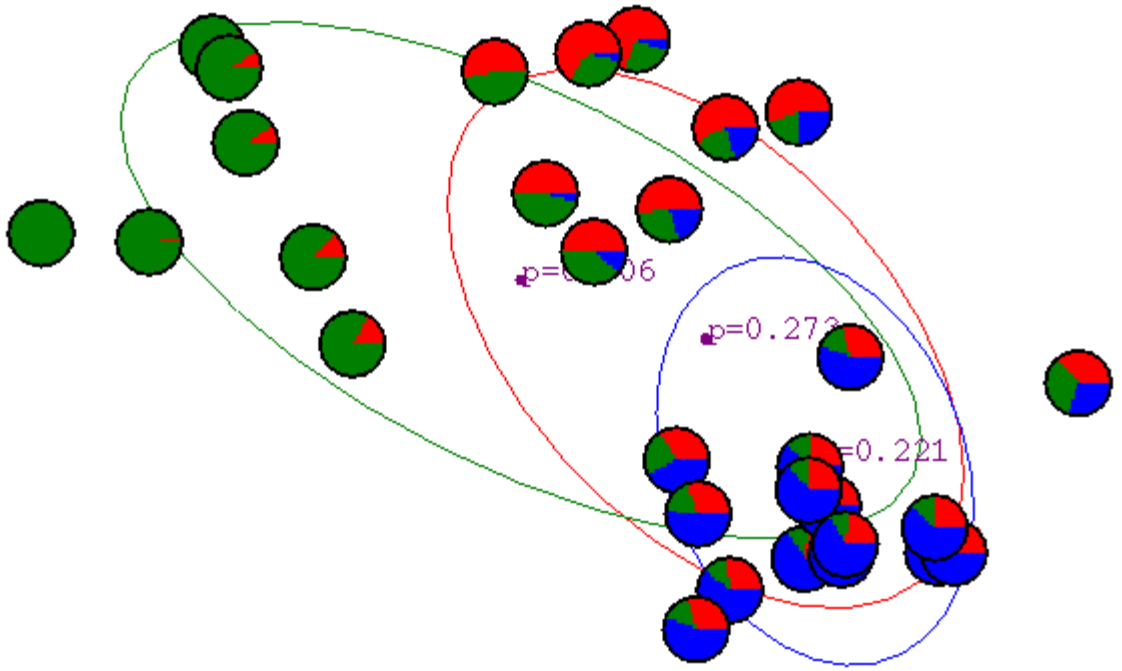
Randomly choose K and cluster centers!



K-Means Clustering

Iteration 1

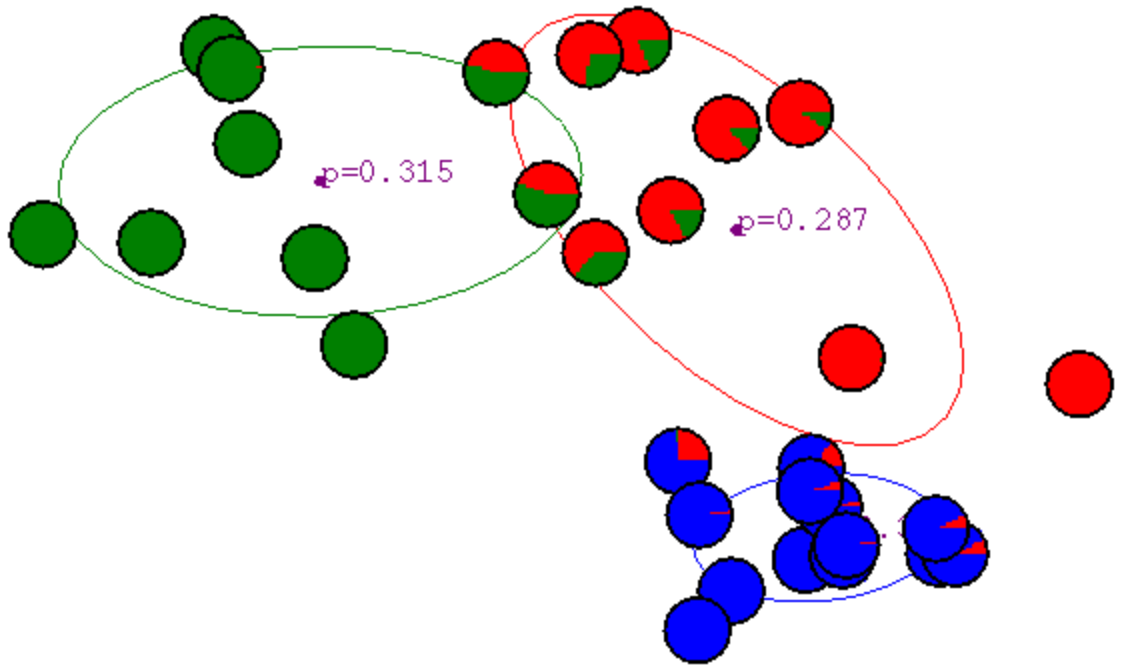
Compute nearest cluster
Recompute means



K-Means Clustering

Iteration 5

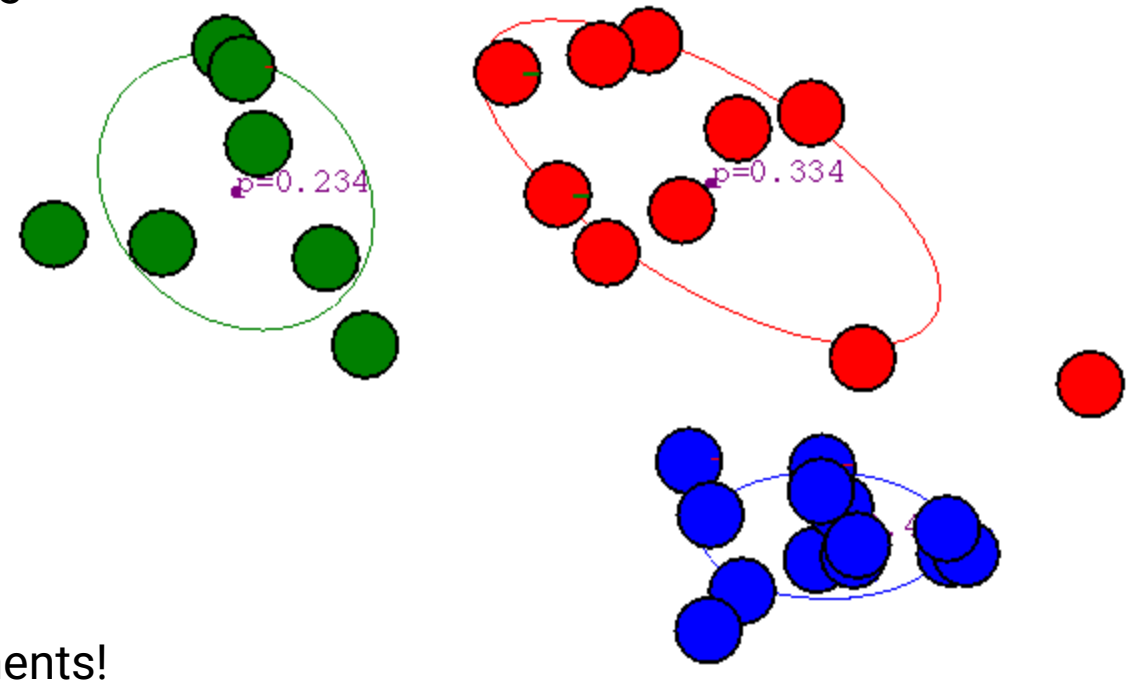
Compute nearest cluster
Recompute means



K-Means Clustering

Iteration N

Repeat until minimal change to cluster means



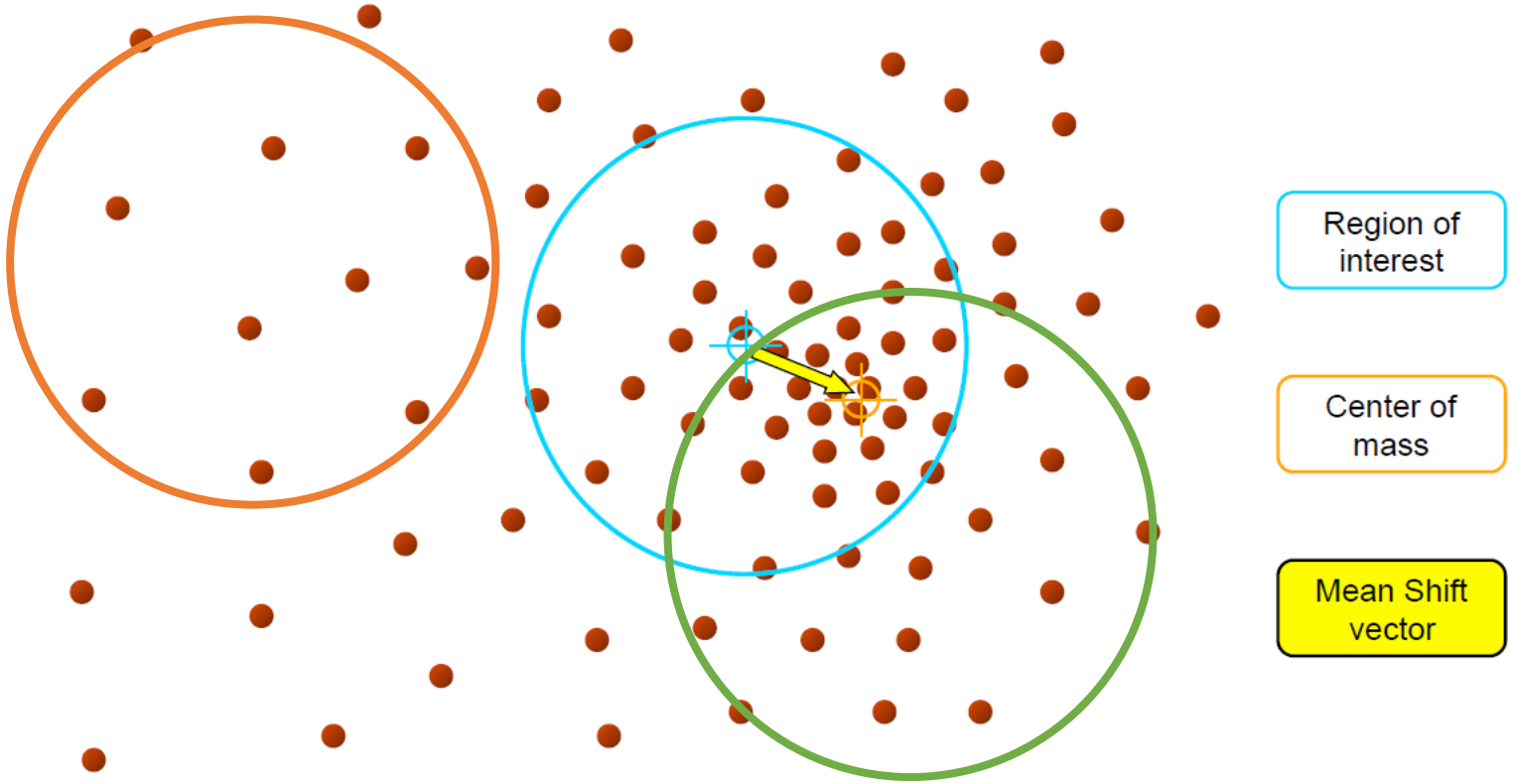
Problems?

- Need to know K
- Clusters are hard assignments!
  - Sometimes you are just not sure enough
  - Soft assignments made through GMM
- Cannot parallelize
- Tuning K is non-intuitive

Mean Shift

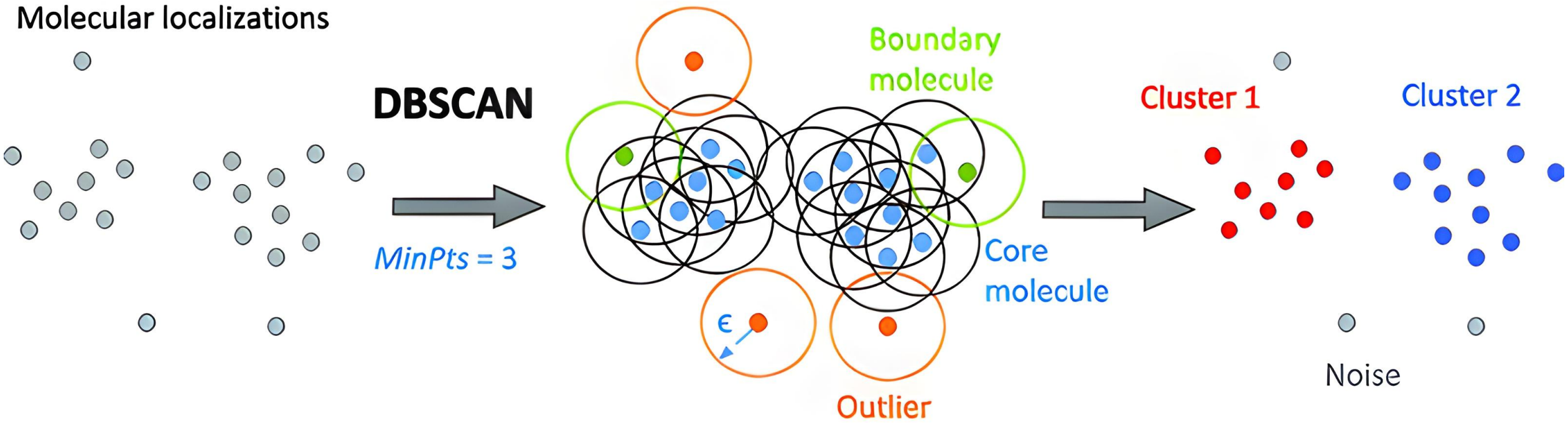
Better than tuning K?

Tune Bandwidth/Density parameters

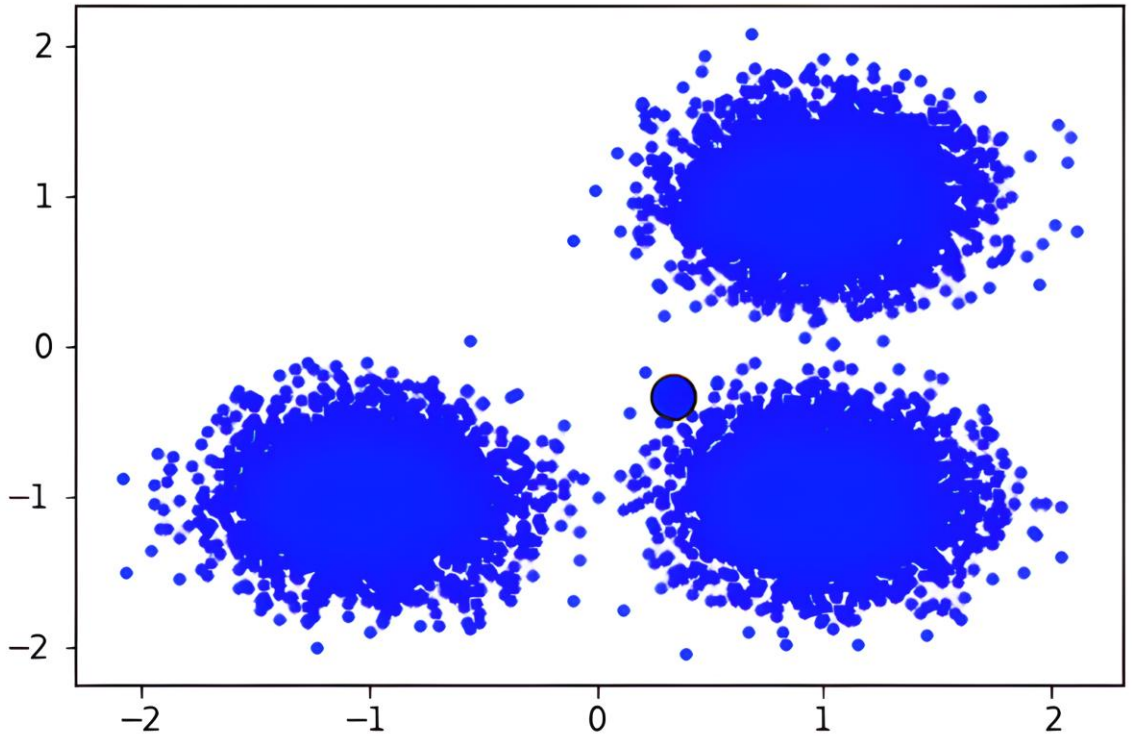


More Robust: DBSCAN

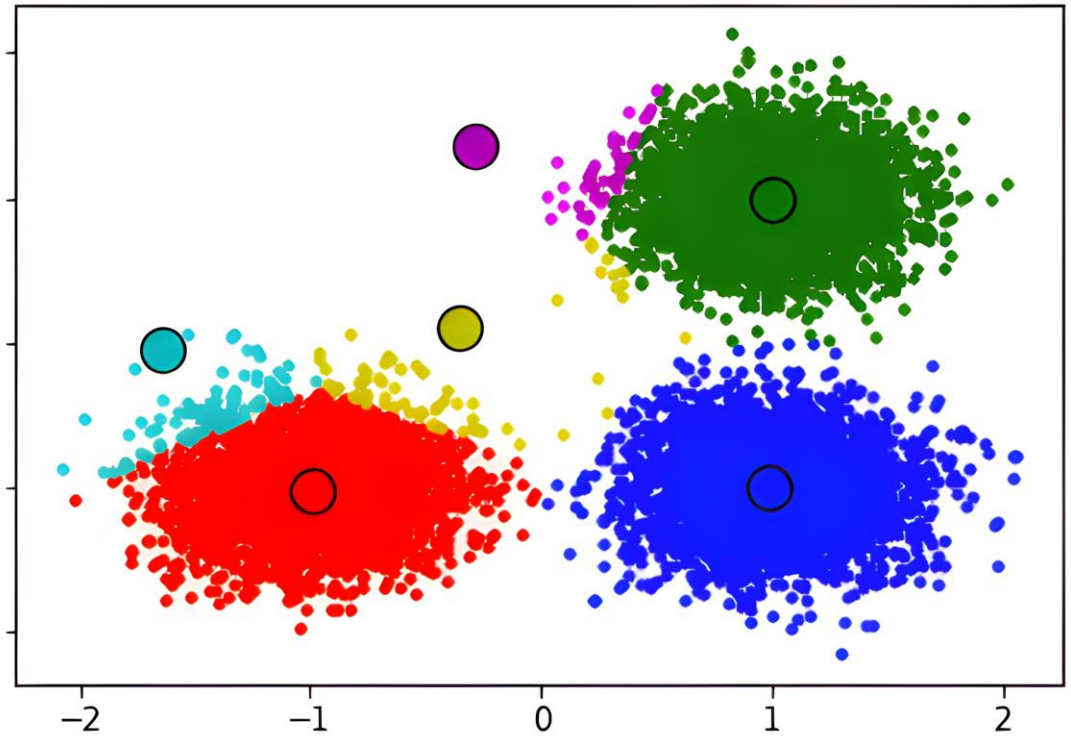
Tune min_pts and ϵ



Split And Merge



Now can start **splitting** clusters!
Undersegment first then fix

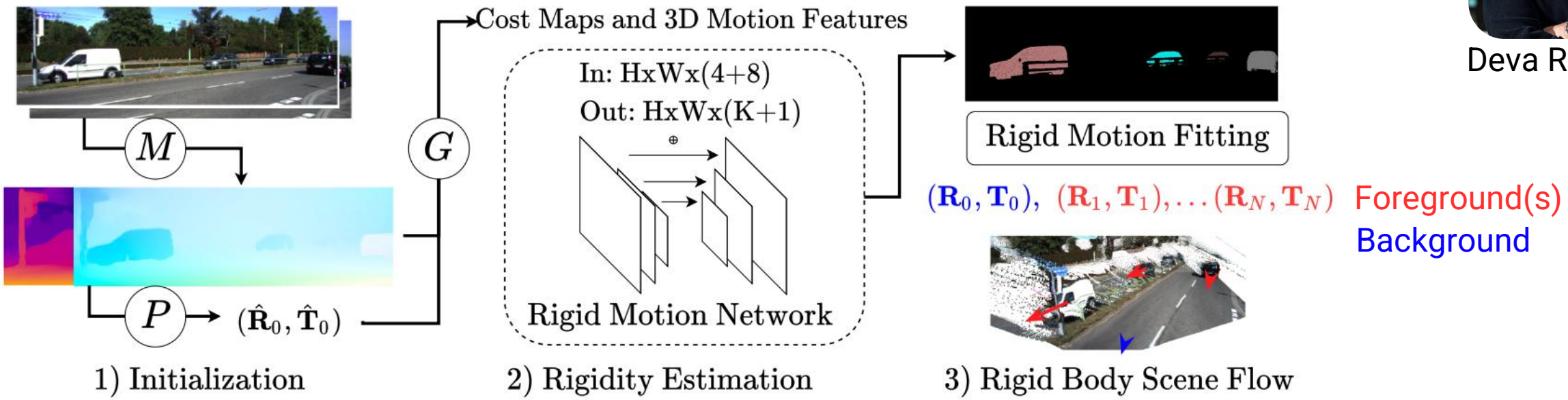


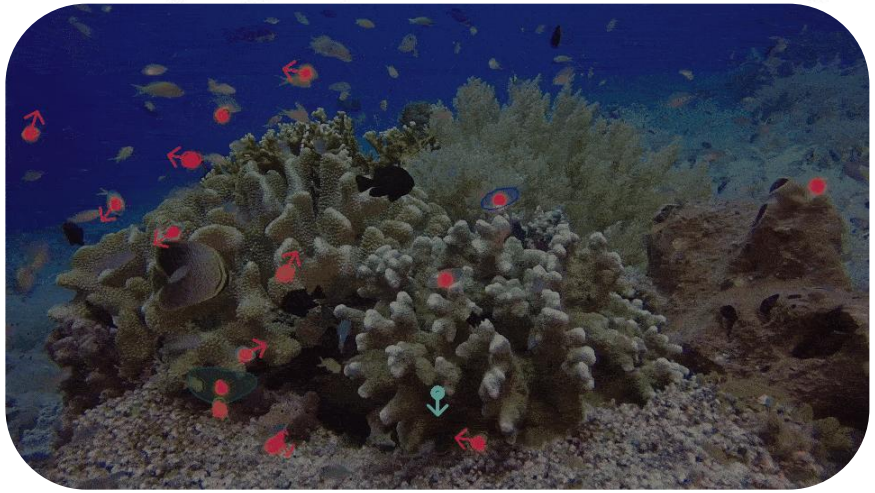
Now can start **merging** clusters!
Oversegment first then fix

Understanding Mean Shift Clustering and Implementation with Python | by Yufeng | Towards Data Science

Era Of Deep Learning

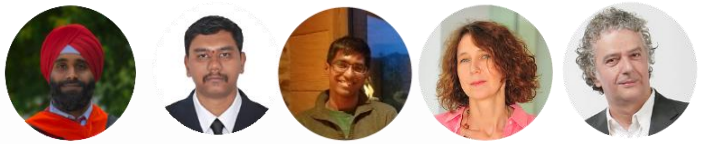
Deva Ramanan





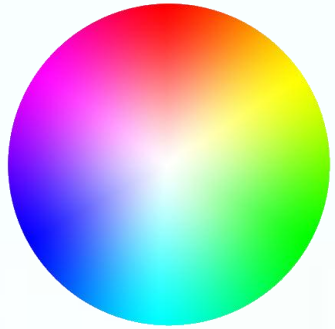
Yang, Gengshan, and Deva Ramanan. "Learning to segment rigid motions from two frames." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

Motion flow vector predicted at every 3D location



IROS 2021

NudgeSeg: Zero-Shot Object Segmentation by Repeated Physical Interaction



Flow-based Clustering

- Distance: $\|X - Y\|_2 < \tau_d \rightarrow$ DBSCAN
- Magnitude of Optical Flow: $\left|\left|M_X - M_Y\right|\right|_2 < \tau_M$
- Angle of Optical Flow: $\min\left(\left|A_x - A_y\right|,\ 2\pi - \left|A_x - A_y\right|\right) < \tau_M$

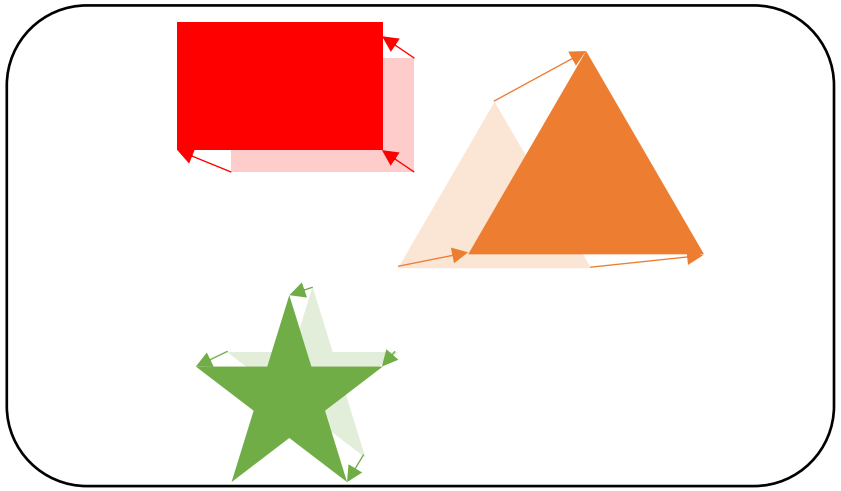
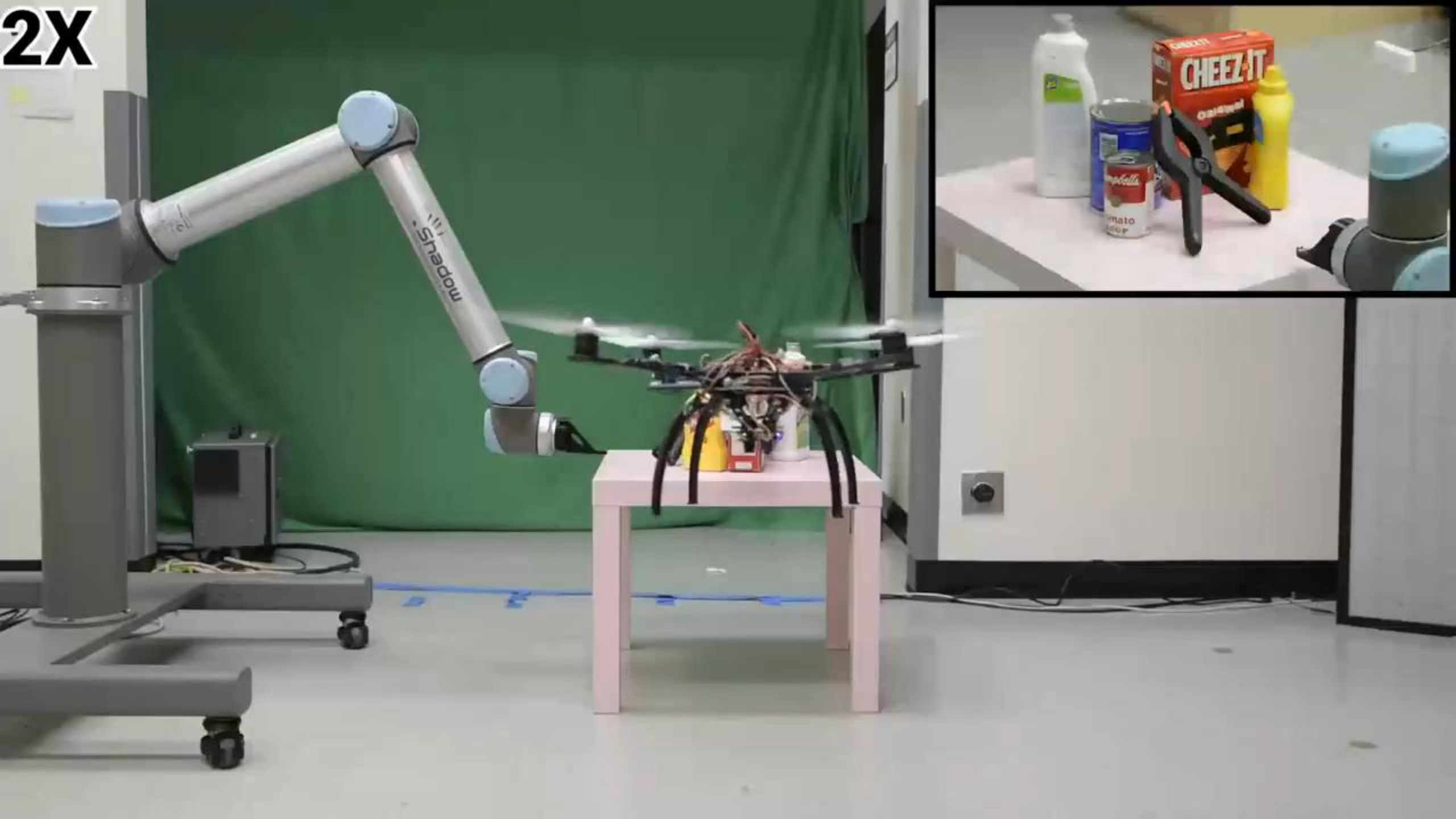


Image Plane


2X
CHEEZ-IT

What Is The Implicit Assumption?

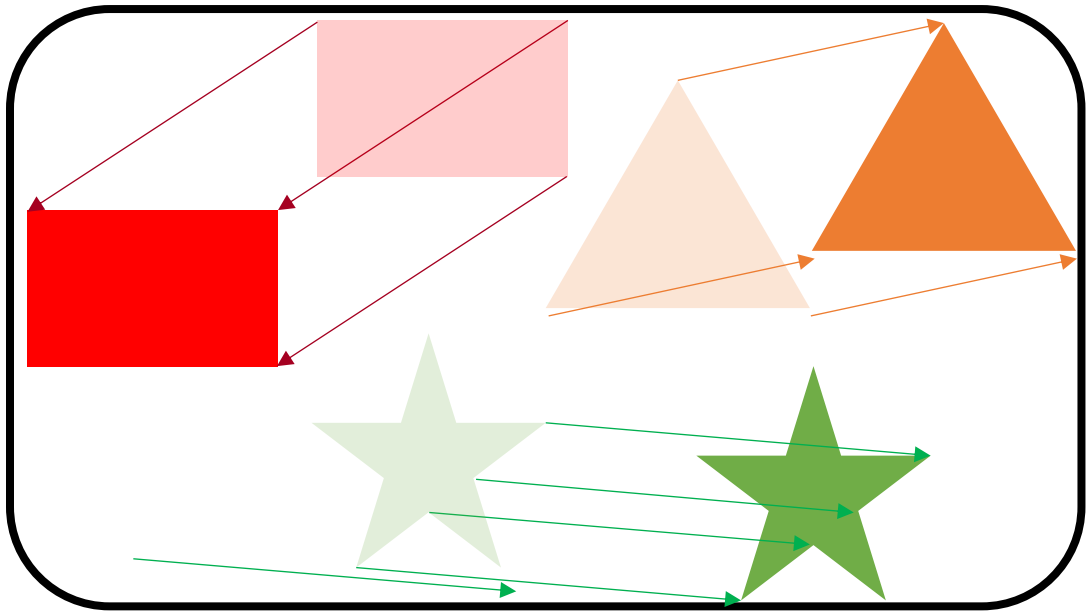


Image Plane

Each object has **one dominant flow** direction

But, objects have **one dominant motion field**

Motion Is Powerful!

Find the gap!

Can You Do It Now?

What Is The “Mostly” The Implicit Assumption?

Objects are **Rigid**

Very Cool Research Problem!

But, objects are **Deformable**

Next Class!

Learned Depth: Monocular + Stereo