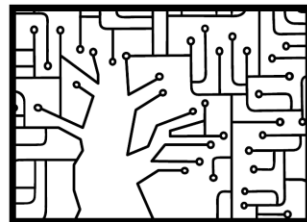


**MEN. WOMEN. BABIES. ELDERLY.  
ARE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow**

# Deep Learning for Computer Vision: Object Detection & Segmentation

Shai Bagon

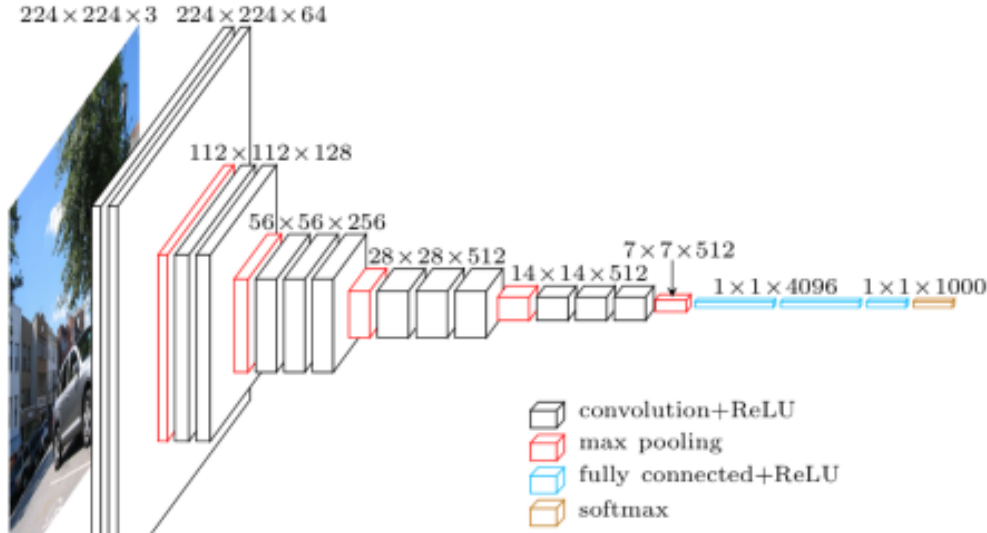


WAIC

# Recap: Deep Learning

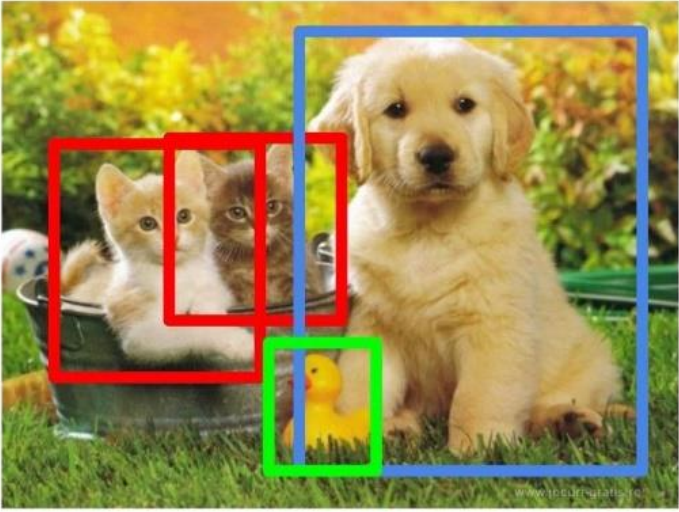
## Key ingredients

- **Data:** labeled examples  $\{(\mathbf{x}_i, y_i)\}_{i=1\dots N}$
- **Model:** a “deep net”  $\hat{y} = f(\mathbf{x}; \boldsymbol{\theta})$
- **Criterion:** how to “match” model and data  $\mathcal{L}(\hat{y}, y)$
- **Optimization:** Stochastic Gradient Descent (SGD)



# Additional Tasks

- Object detection
- Semantic segmentation
- Instance segmentation



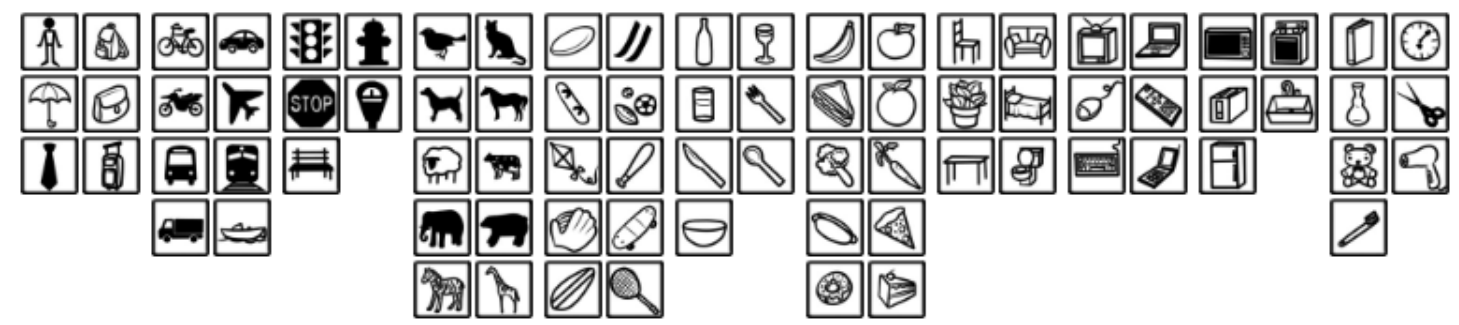
MEN. WOMEN. BABIES. ELDERLY.  
ARE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow

# Additional Tasks: Training Data

MS COCO



- 200K labeled images
- 1.5M instances
- 80 object categories



**MEN. WOMEN. BABIES. ELDERLY.**  
**ARE STILL HELD HOSTAGE BY HAMAS**  
**#BringThemHomeNow**

# Object Detection

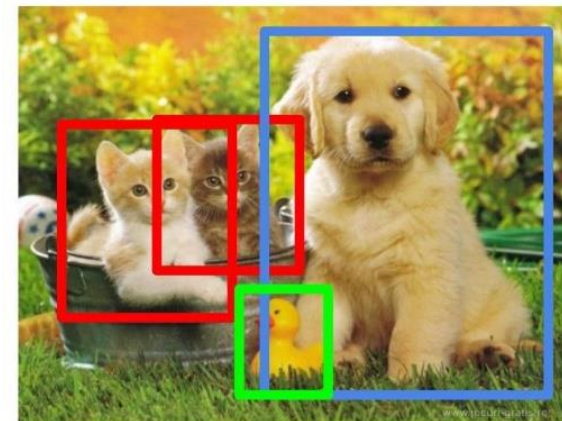
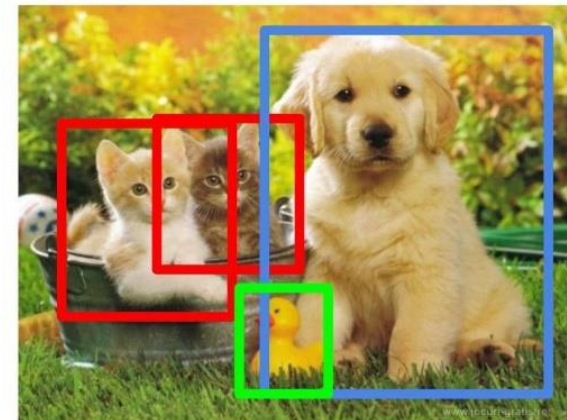


Image credit: [medium](#)

# Object Detection - Challenges

- Multiple types of outputs
- Varying number of objects



MEN. WOMEN. BABIES. ELDERLY.  
ARE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow

# Localization

**Classification**



CAT

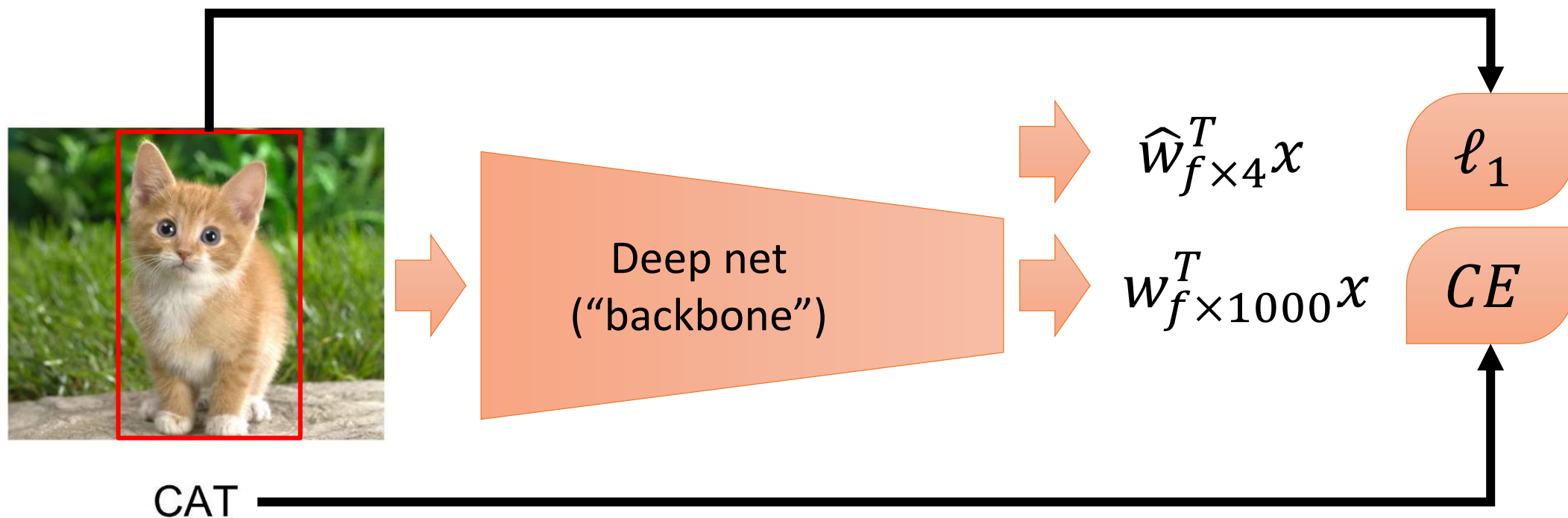
**Classification  
+ Localization**



CAT

MEN. WOMEN. BABIES. ELDERLY.  
ARE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow

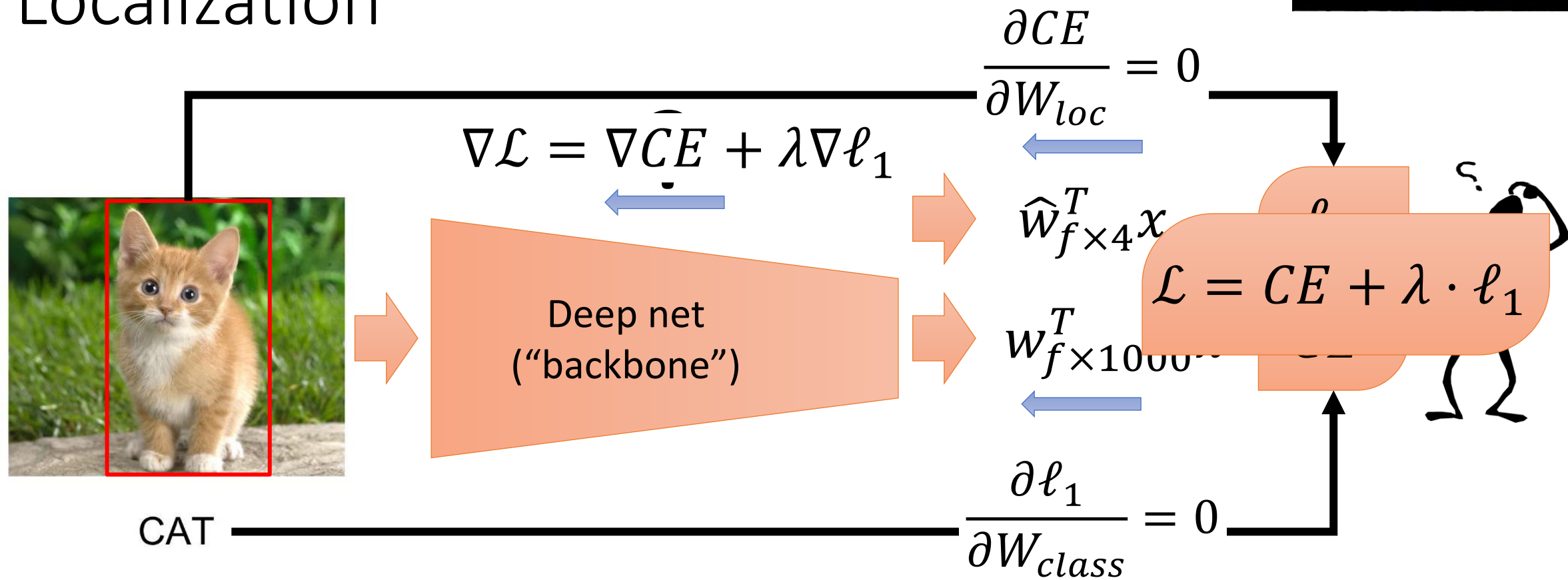
# Localization



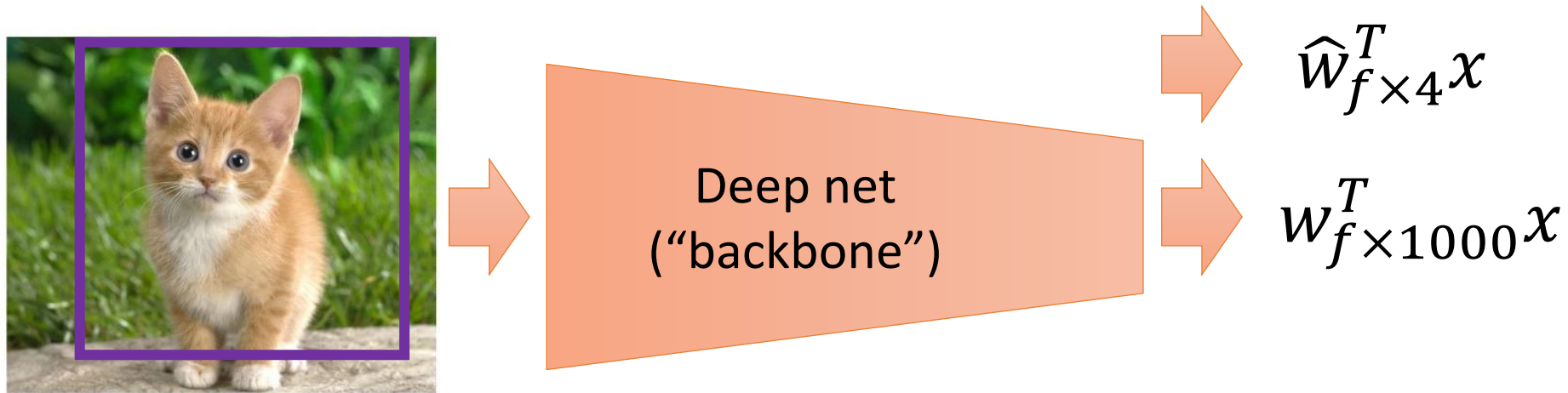


MEN. WOMEN. BABIES. ELDERLY.  
 ARE STILL HELD HOSTAGE BY HAMAS  
 #BringThemHomeNow

# Localization



# From Localization to Detection (v0)



# From Localization to Detection (v0)



Deep net  
("backbone")



$$\widehat{W}_{f \times 4}^T x$$

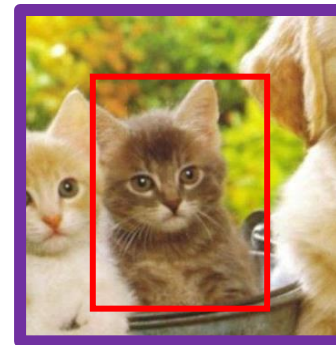


$$W_{f \times 1000}^T x$$

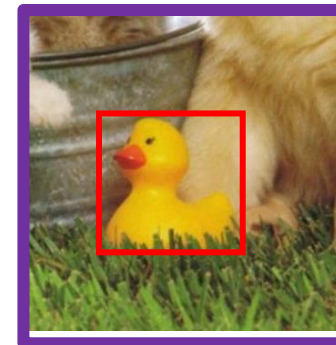
# From Localization to Detection (v0)



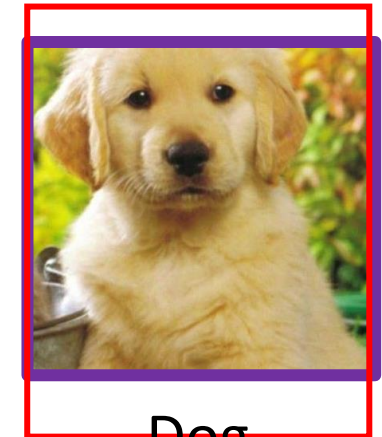
Cat



Cat



Duck



Dog



**“Background”**

# From Localization to Detection (v0)



Deep net  
("backbone")



$$\widehat{W}_{f \times 4}^T x$$



$$W_{f \times 1000}^T x$$

Challenges:

- Multiple types of outputs
- Number of objects

# From Localization to Detection (v0)



Deep net  
("backbone")



$$\widehat{W}_{f \times 4}^T x$$

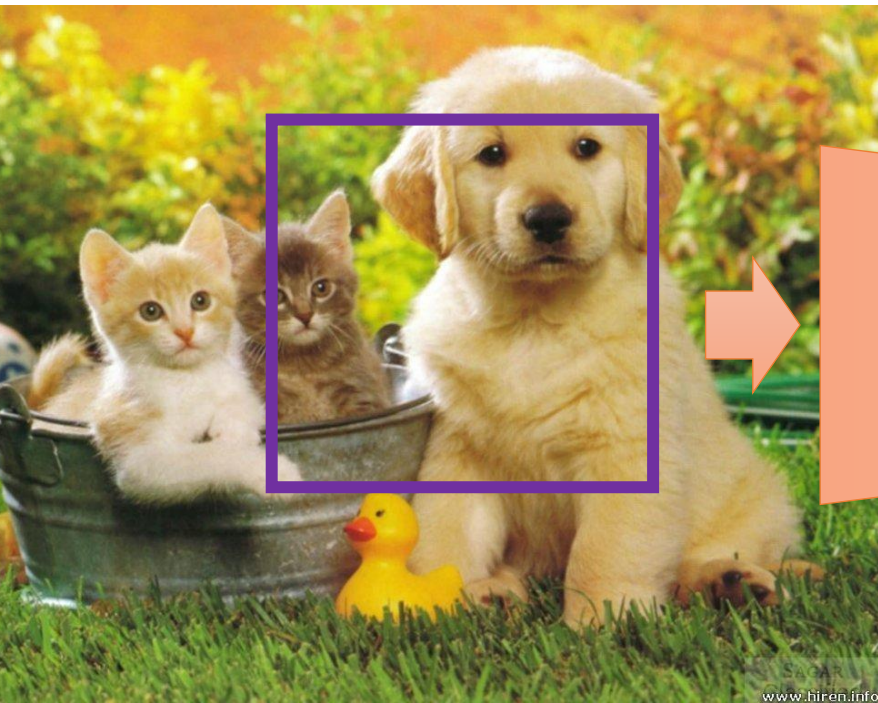


$$W_{f \times 1000}^T x$$

How many "sliding windows" are there?

**There can easily be  $O(1M)$  windows!**

# From Localization to Detection (v1)



Deep net  
("backbone")



$$\widehat{W}_{f \times 1 \times 1 \times 4}^{loc} * \mathcal{X}$$



$$W_{f \times 1 \times 1 \times C}^{class} * \mathcal{X}$$

MEN. WOMEN. BABIES. ELDERLY.  
ARE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow

# Object Detection

Single Shot:  
SSD, YOLO...

Fast  
High false rate



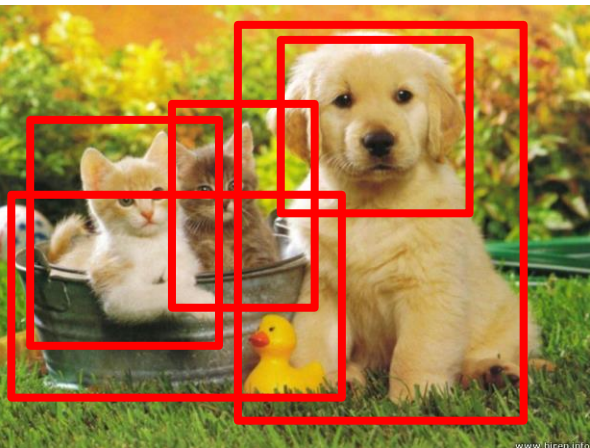
Two Shots:  
R-CNN...

Slower  
More accurate

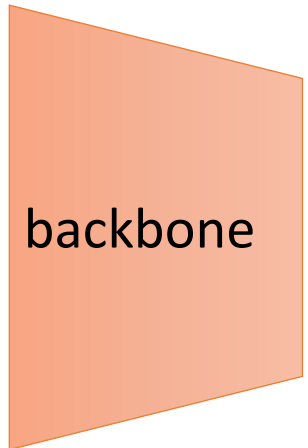


MEN. WOMEN. BABIES. ELDERLY.  
APE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow

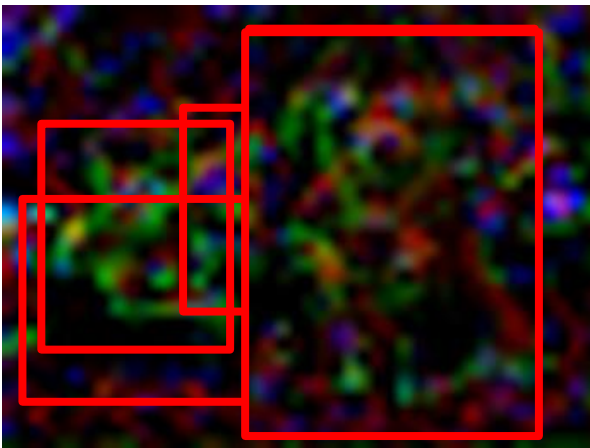
# Faster R-CNN



$3 \times H \times W$



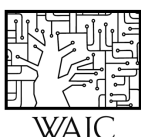
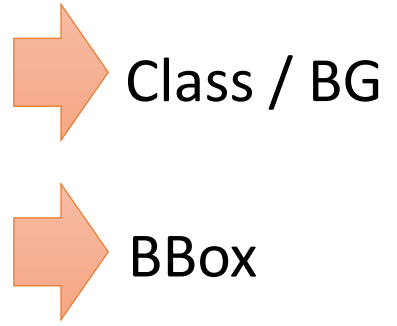
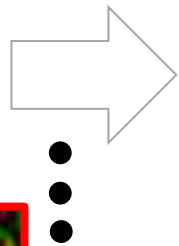
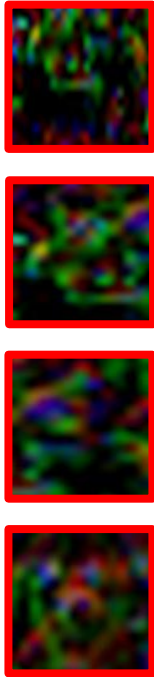
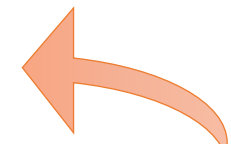
backbone



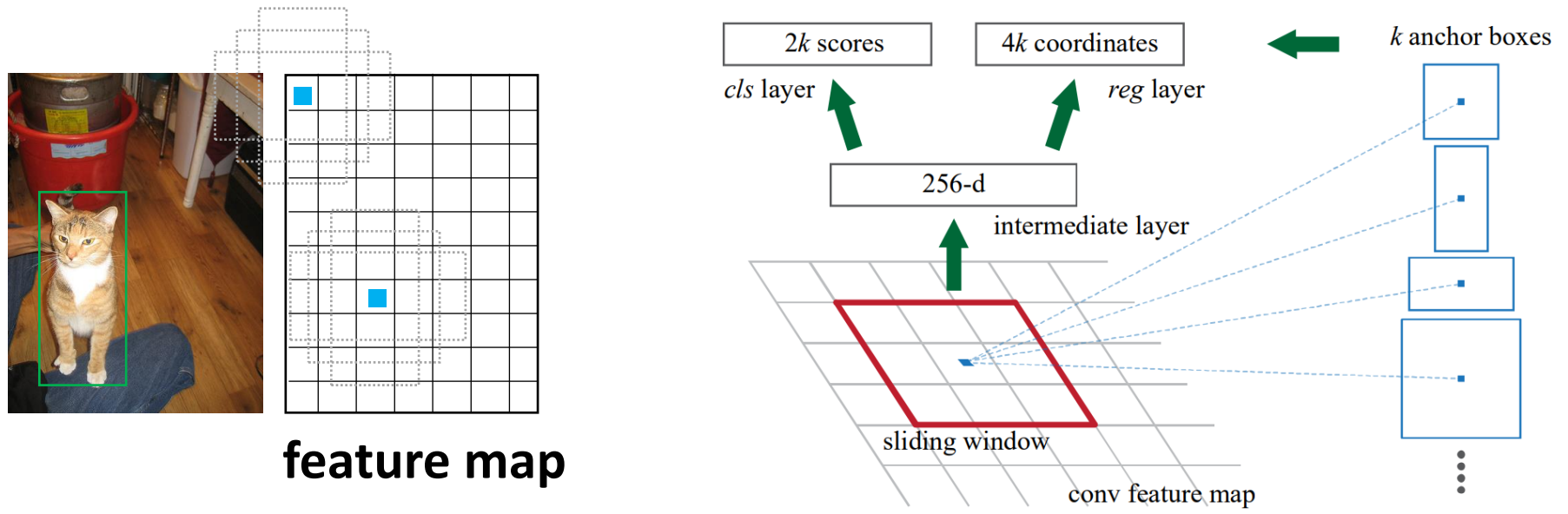
“ROI Pool”  $D \times h \times w$  each proposed region from the **feature map**



RPN

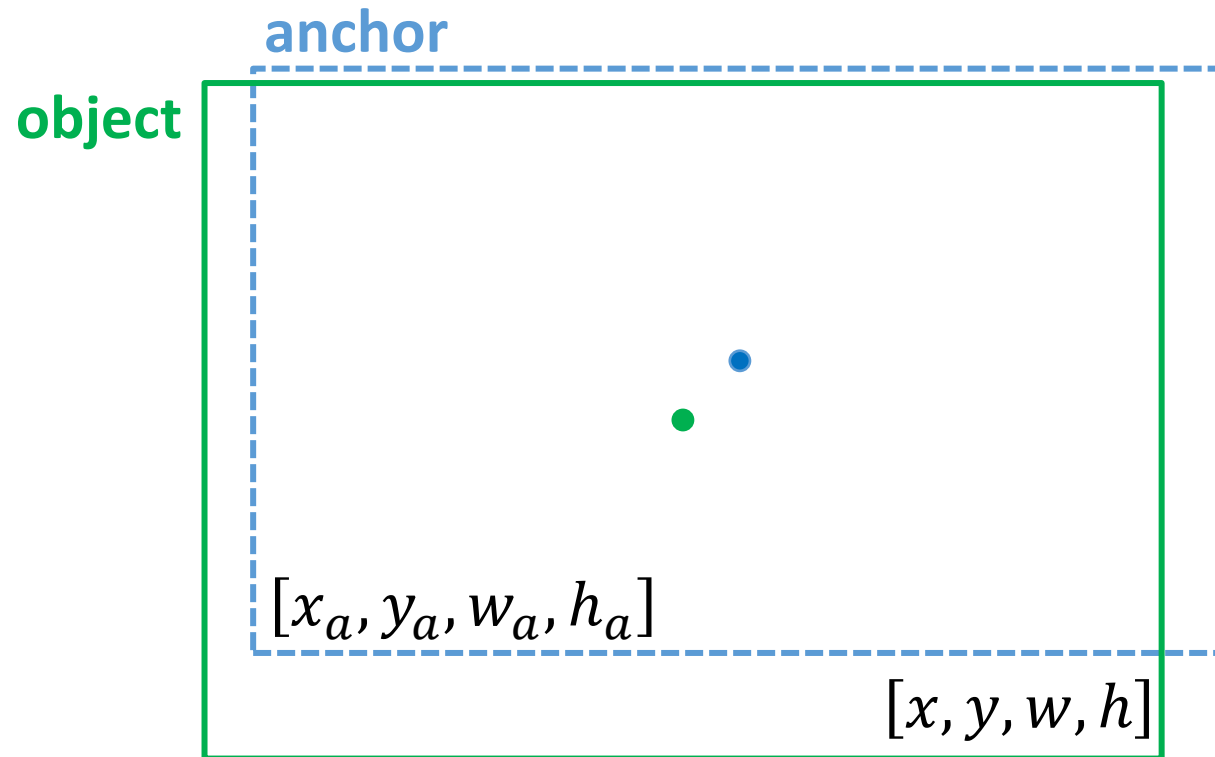


# RPN: Region Proposal Network



# RPN: Region Proposal Network

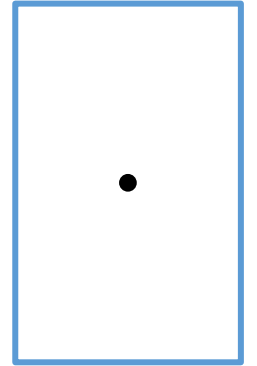
Predicting Bounding Box coordinates from anchors:



# RPN: Region Proposal Network

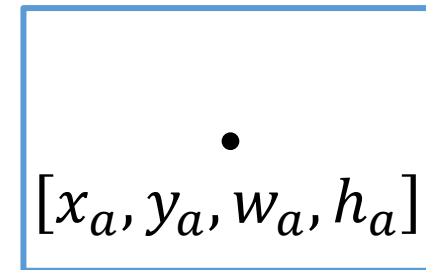
Predicting a bounding box “corrections”  $\mathbf{t} = [t_x, t_y, t_w, t_h]$ :

$$t_x = \frac{x - x_a}{w_a}, \quad t_y = \frac{y - y_a}{h_a}, \quad t_w = \log\left(\frac{w}{w_a}\right), \quad t_h = \log\left(\frac{h}{h_a}\right)$$



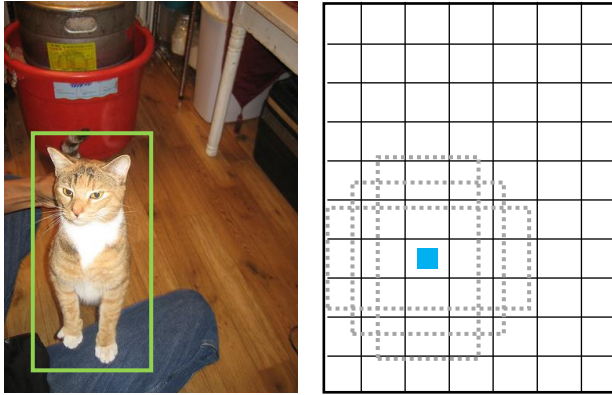
Recovering the actual BBox from  $\mathbf{t} = [t_x, t_y, t_w, t_h]$ :

$$x = w_a \cdot t_x + x_a, \quad y = h_a \cdot t_y + y_a, \\ w = e^{t_w} \cdot w_a, \quad h = e^{t_h} \cdot h_a$$



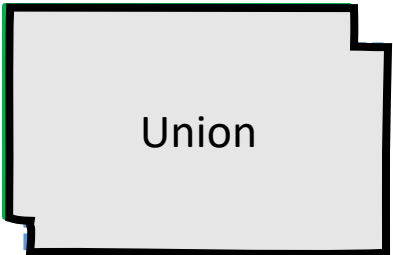
The learned “corrections”  $\mathbf{t}$  are relative to anchor position and scale.

# Object Detection



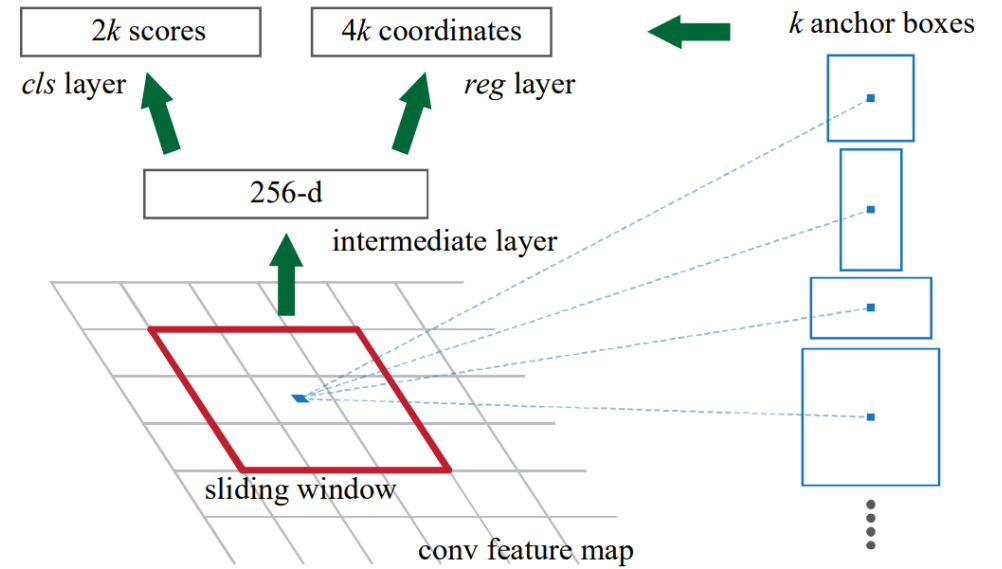
Training

object



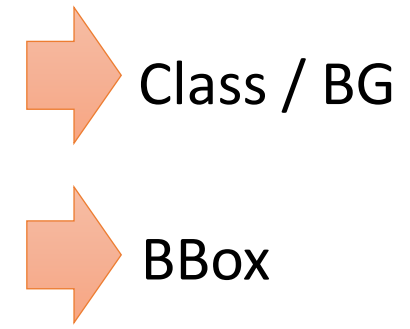
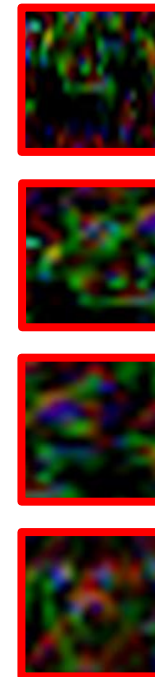
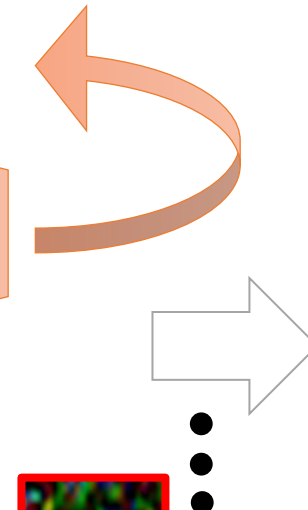
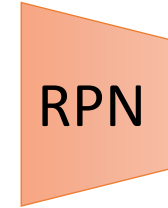
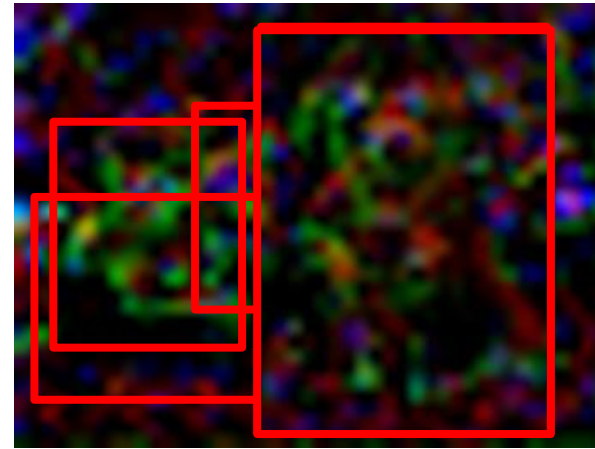
anchor

$$\text{IoU} = \frac{\text{Intersection}}{\text{Union}}$$



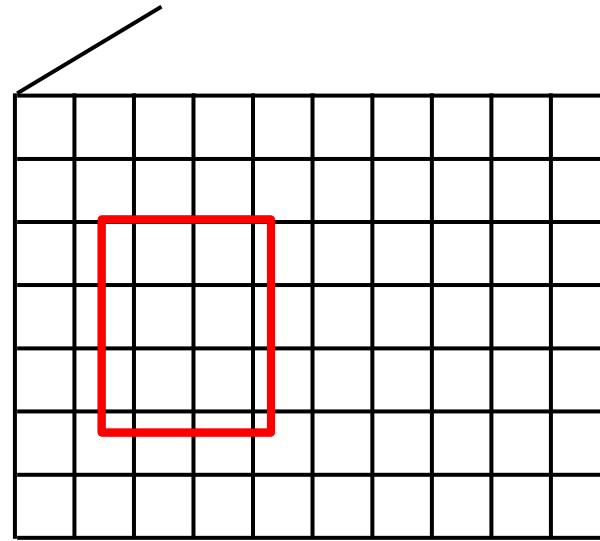
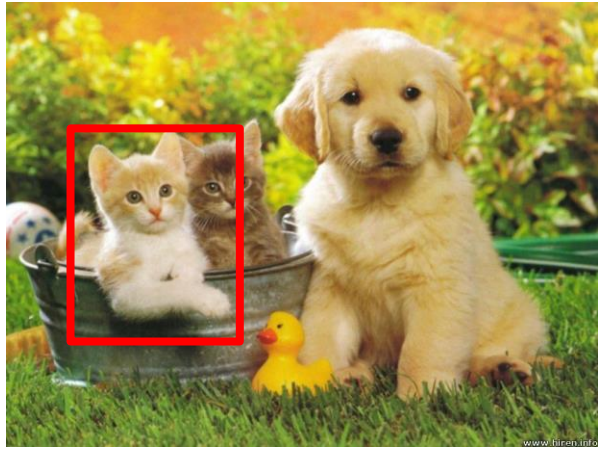


# ROI Pooling



Input: varying size/aspect ratio "proposals"  
Output: Fixed size feature crop

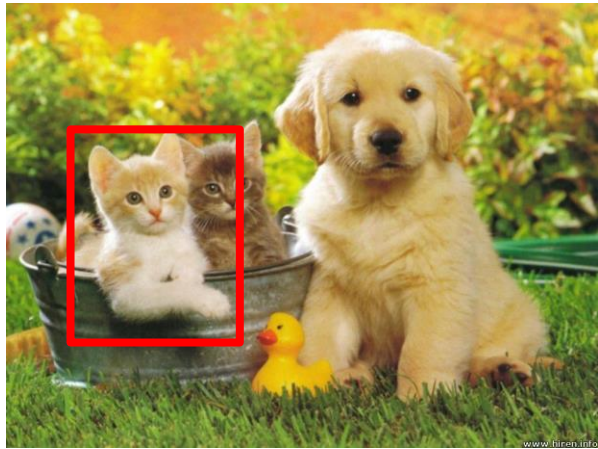
# ROI Pooling



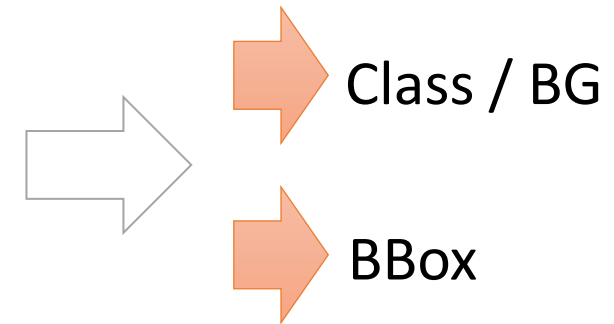
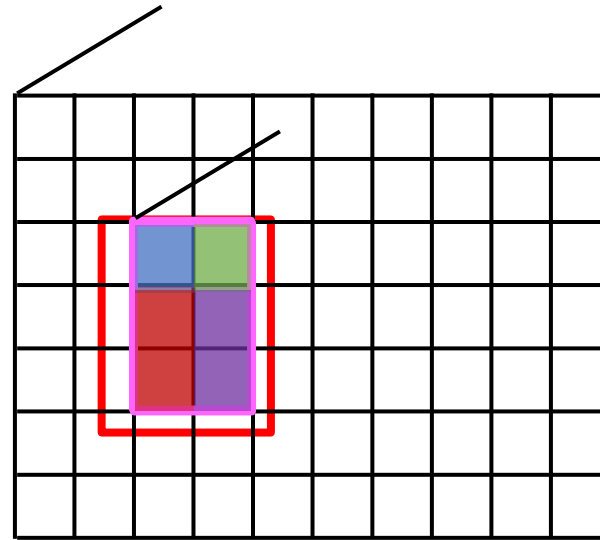
Input: varying size/aspect ratio “proposals”

Output: Fixed size feature crop

# ROI Pooling



backbone



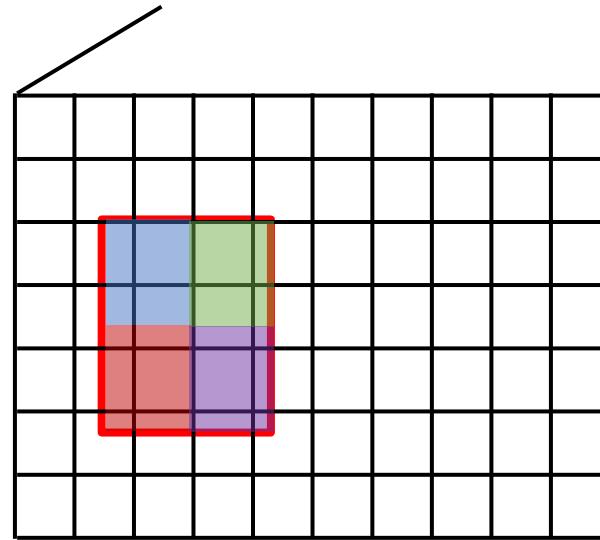
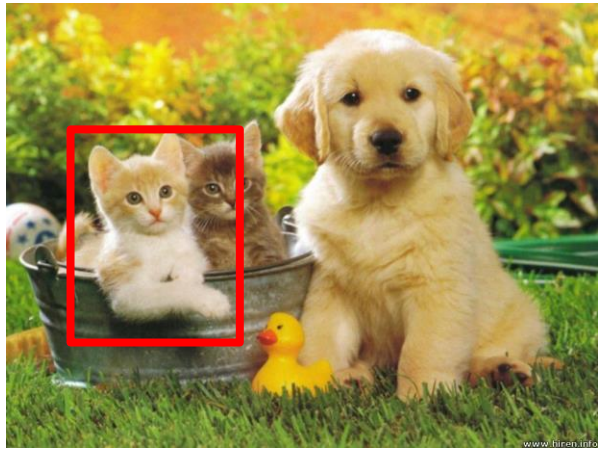
Input: varying size/aspect ratio “proposals”

Output: Fixed size feature crop

ROI Pool – “snap” to closest grid



# ROI Align



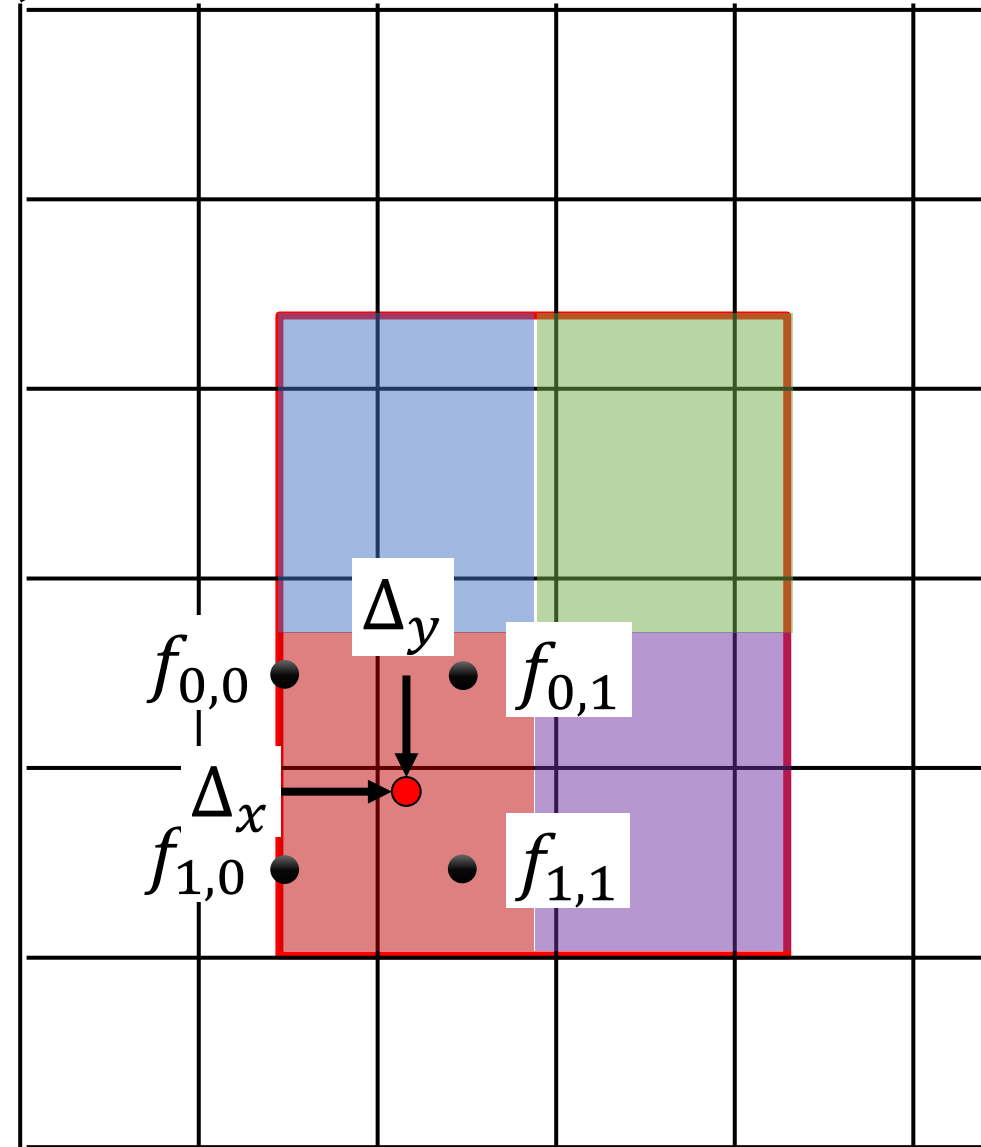
Input: varying size/aspect ratio “proposals”

Output: Fixed size feature crop

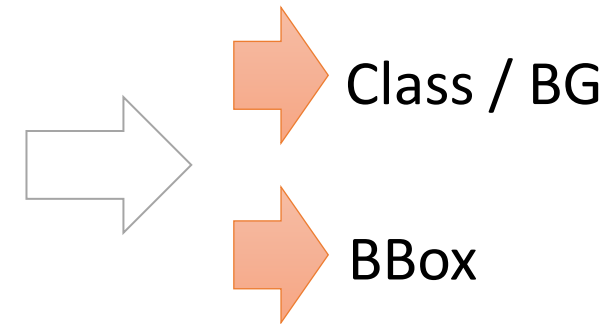
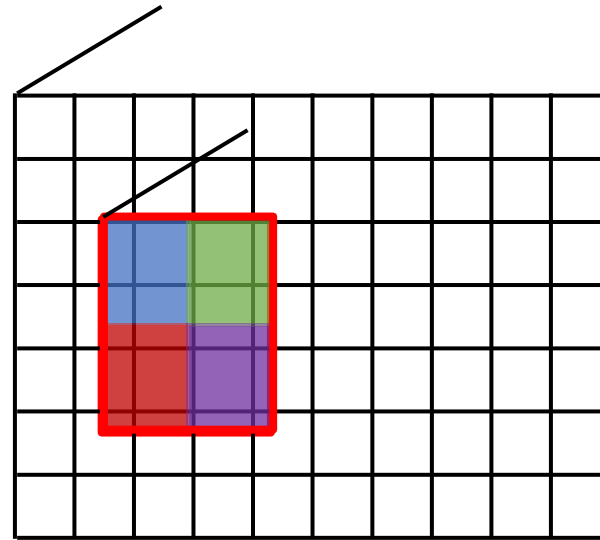
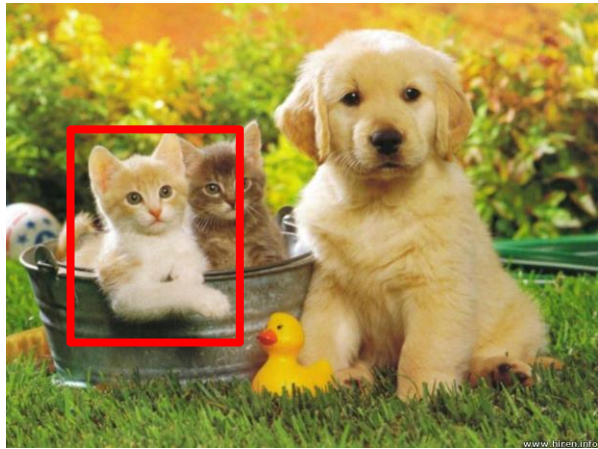
ROI Align – interp. features

# ROI Align

$$\begin{aligned} f_{xy} &= \sum_{ij} (1 - |x - x_i|)(1 - |y - y_i|) f_{ij} \\ &= (1 - \Delta_x)(1 - \Delta_y) f_{0,0} \\ &\quad + \Delta_x(1 - \Delta_y) f_{0,1} \\ &\quad + (1 - \Delta_x)\Delta_y f_{1,0} \\ &\quad + \Delta_x\Delta_y f_{1,1} \end{aligned}$$



# ROI Align



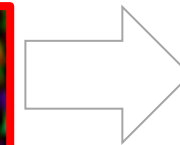
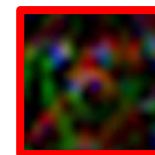
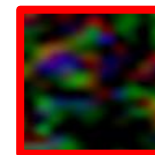
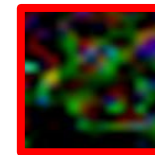
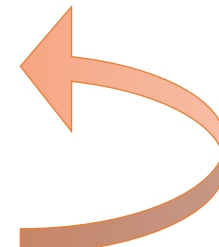
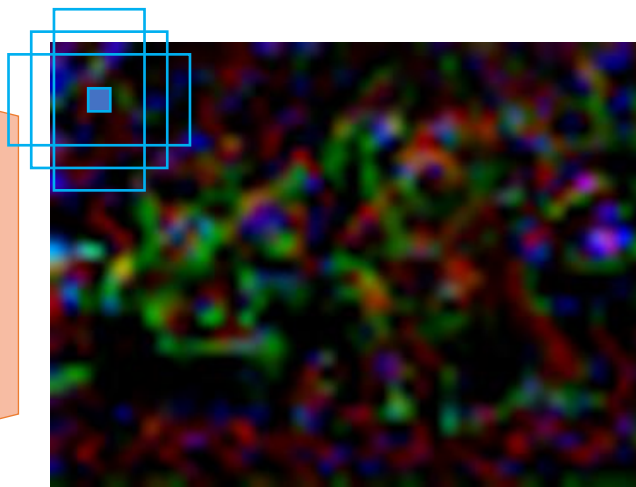
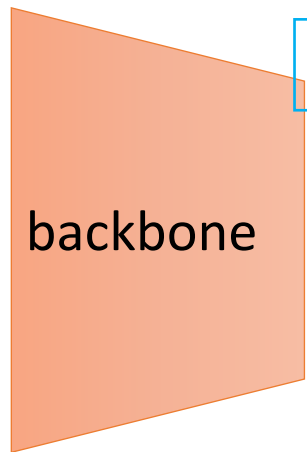
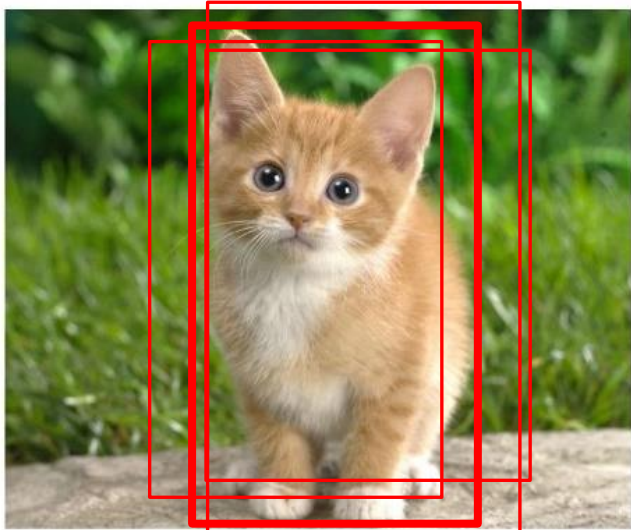
Input: varying size/aspect ratio “proposals”

Output: Fixed size feature crop

ROI Align – interp. features

MEN. WOMEN. BABIES. ELDERLY.  
ARE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow

# Inference



Class / BG

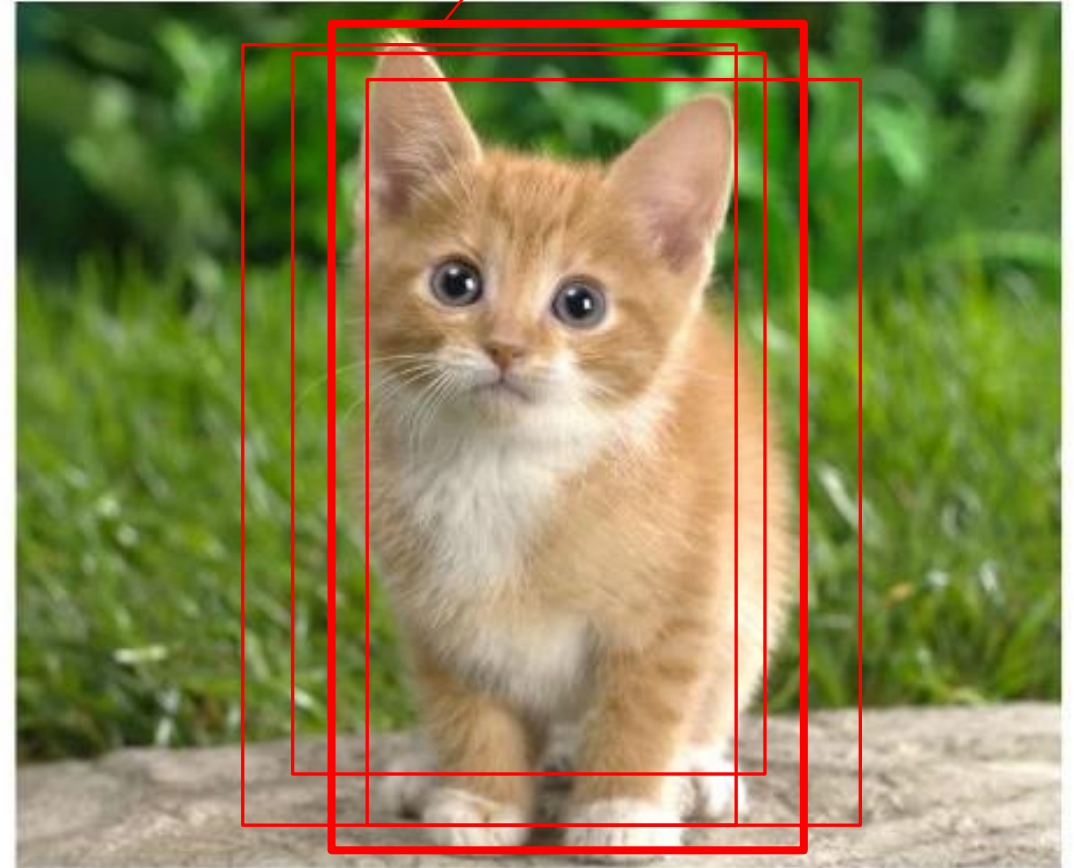


BBox

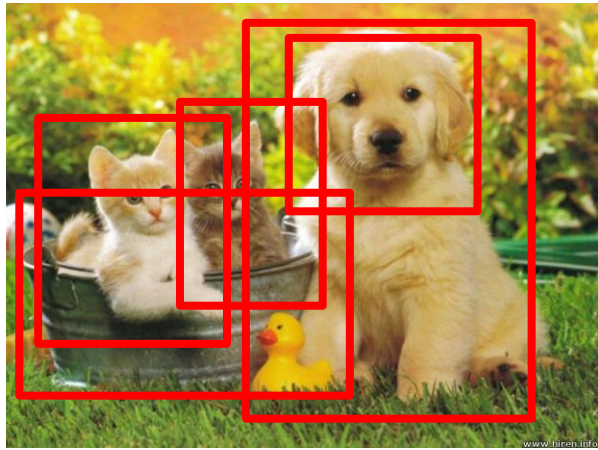
# Inference: Non Maximal Suppression (NMS)

- Sort detection by score (per class)
- Take most confident
- Remove all overlapping
- Repeat

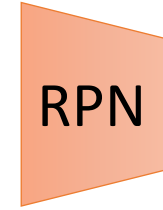
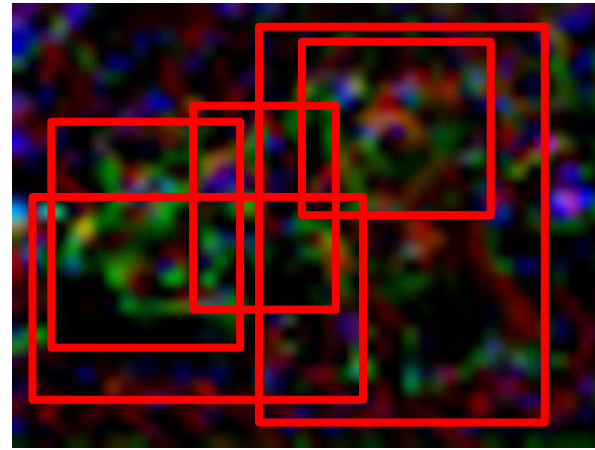
"Cat"  $p=0.98$



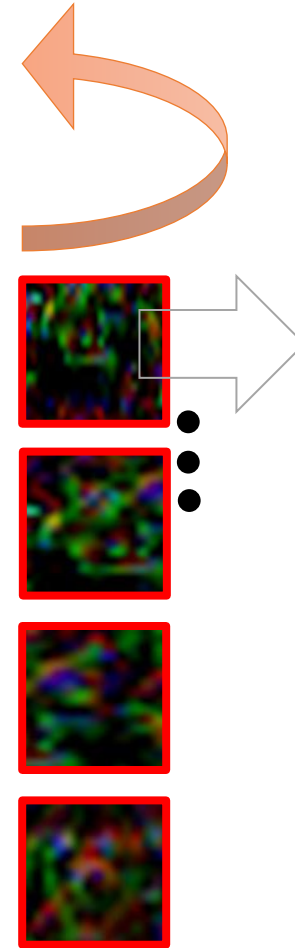
# Object Detection



backbone



RPN



MEN. WOMEN. BABIES. ELDERLY.  
ARE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow

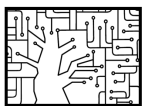
Class / BG  
BBox

Challenges:

- Multiple types of outputs
- Number of objects

# Object Detection: Pitfalls

- Imbalance
- Multiscale



# Imbalance

Number of “negative” anchors  $\sim O(10K)$

Number of “positive” anchors  $\sim O(10)$

**What happens to CE loss in this case?**

$$CE(p_i, y_i) = -y_i \log(p_i)$$

$$= -9 \cdot \log(0.5) - 9 \cdot \log(0.5)$$

$$=$$

Prediction  $p_i$

	True $y_i$	
	BG	Obj
BG	9981	9
Obj	9	1

Unconfident:  
 $p=0.5$

Confident:  
 $p=0.99$





# Imbalance

Number of “negative” anchors  $\sim O(10K)$

Number of “positive” anchors  $\sim O(10)$

**What happens to CE loss in this case?**

$$\begin{aligned} \nabla CE(p_i, y_i) &= p_i - y_i \\ &= 9 \cdot 0.5 + 9 \cdot (1 - 0.5) \end{aligned}$$

Prediction  $p_i$

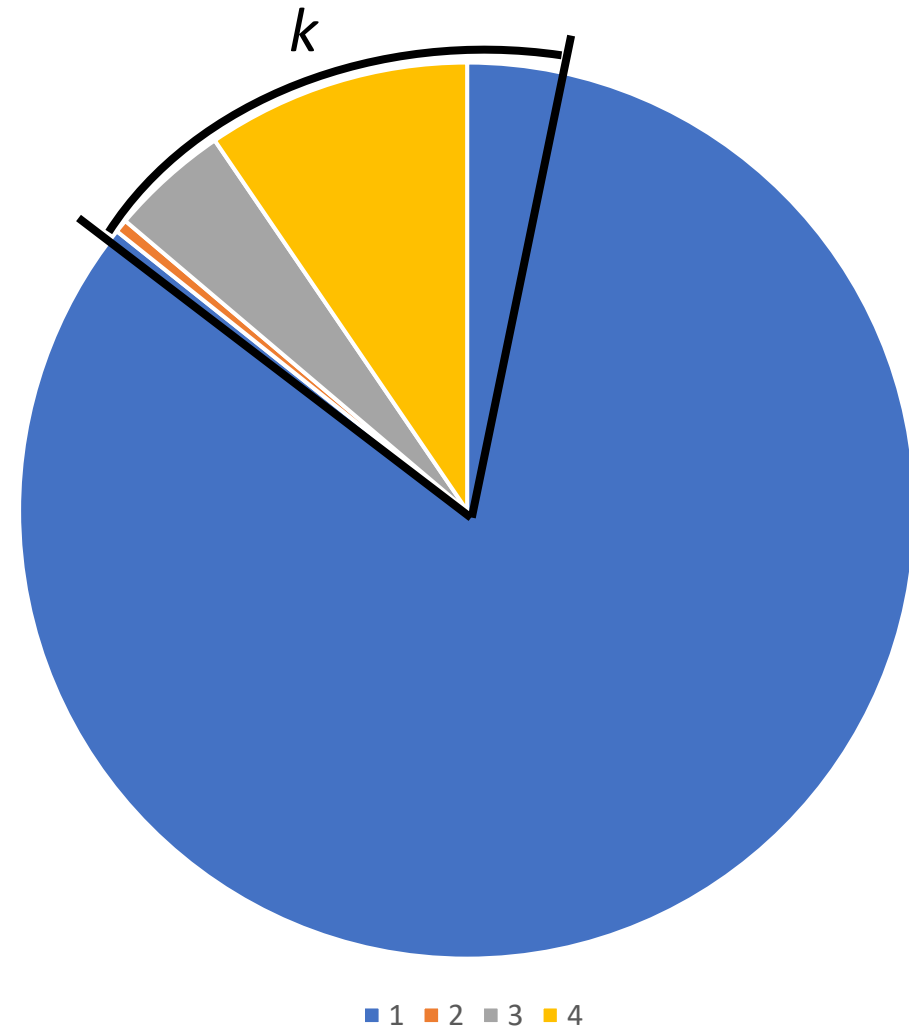
	True $y_i$	
	BG	Obj
BG	9981	9
Obj	9	1

Unconfident:  $p=0.5$

Confident:  $p=0.99$

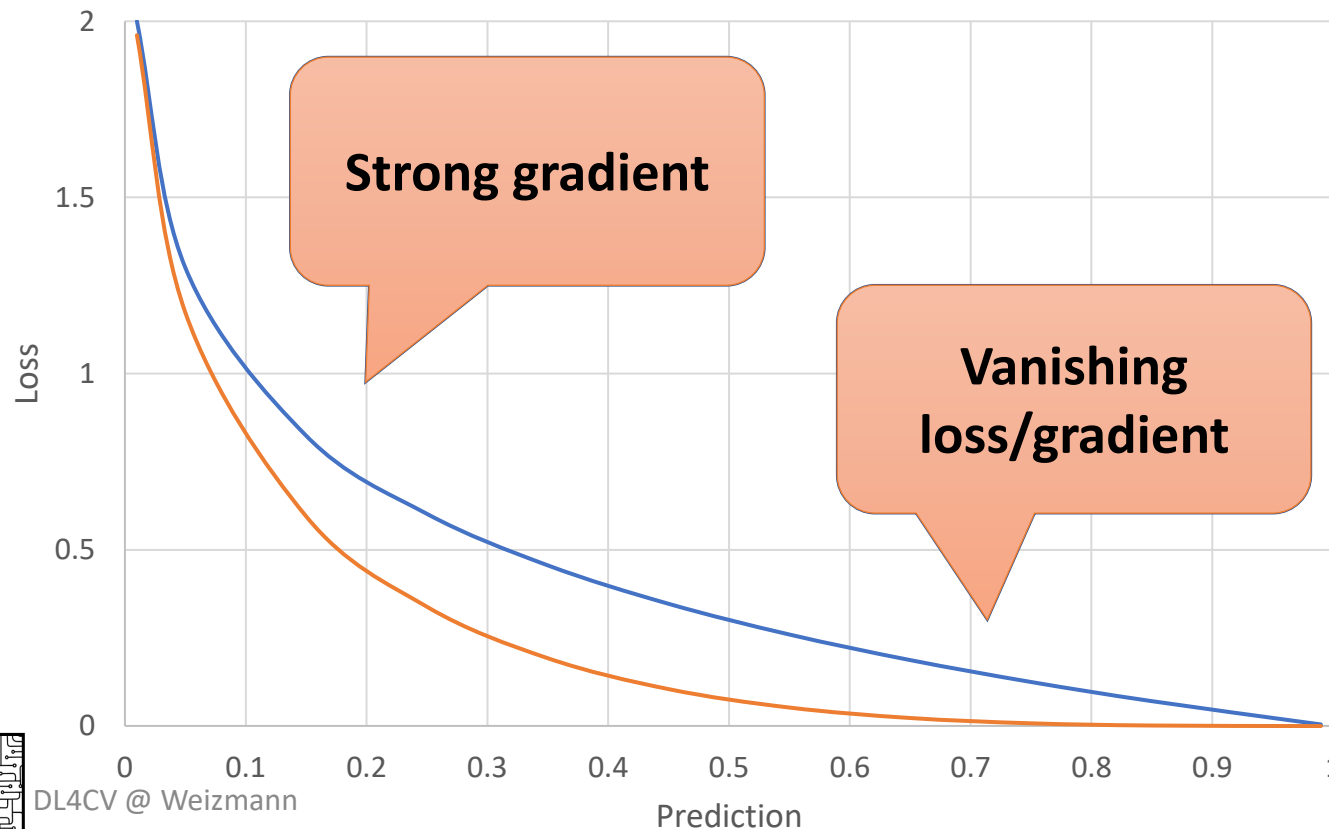
# Imbalance – Hard Negative Mining

Compute loss for all  $N$  anchors  
Select top  $k$  “hard” examples  
Compute gradient for hard  $k$  **only**



# Imbalance – Focal Loss

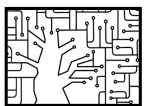
Lin, Goyal, Girshick, He, and Dollár  
Focal loss for dense object detection (PAMI 2018)



$$CE = -y_i \log p_i$$
$$FL = -y_i (1 - p_i)^\gamma \log p_i$$

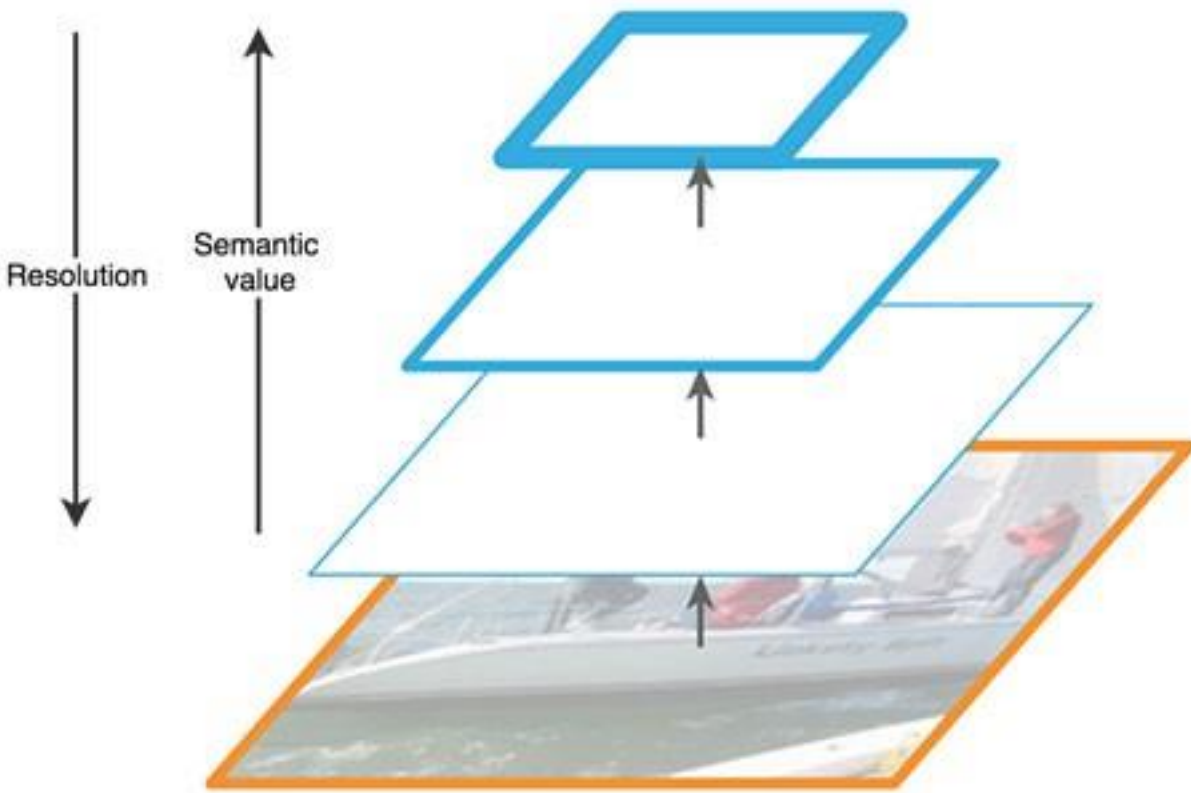
# Object Detection: Pitfalls

- Imbalance
- Multiscale



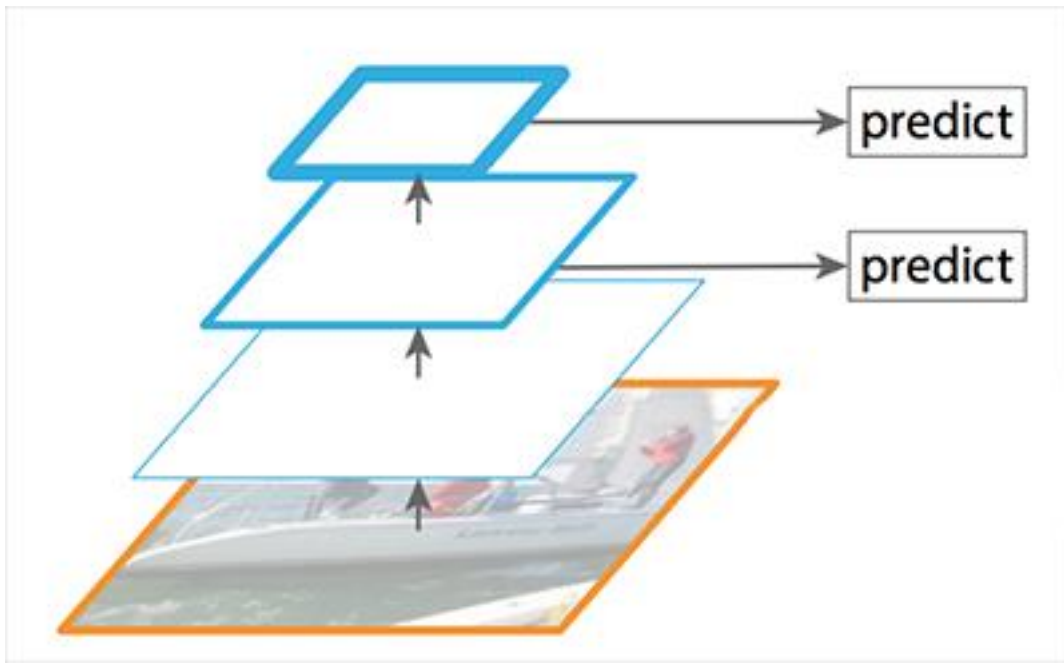
# Resolution vs Semantic Value

How to handle multiscale predictions?



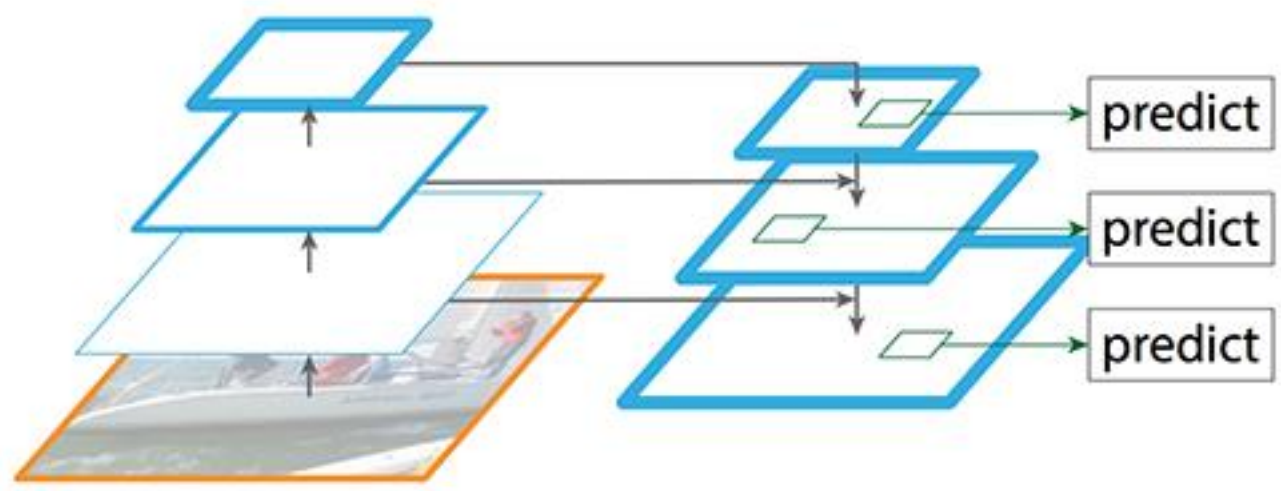
# Resolution vs Semantic Value

How to handle multiscale predictions?



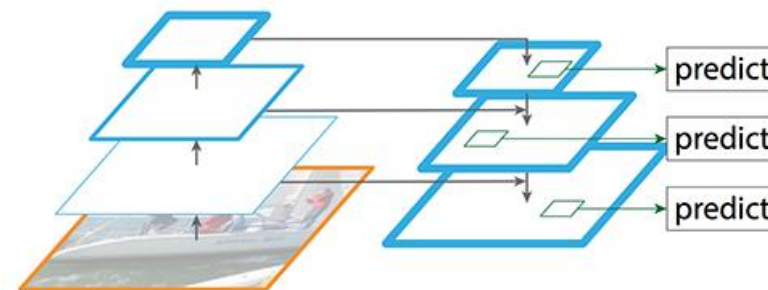
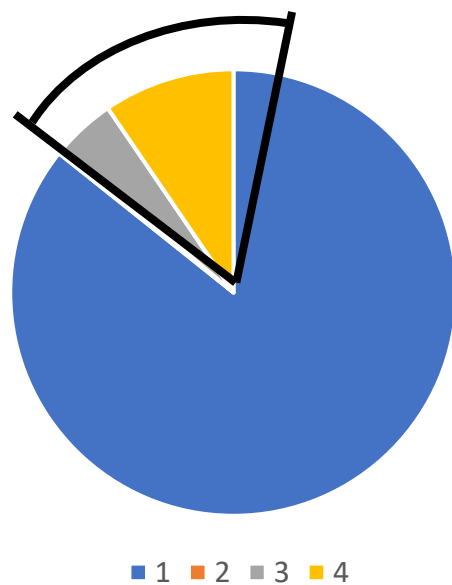
# Feature Pyramid Network (FPN)

How to handle multiscale predictions?



# Object Detection

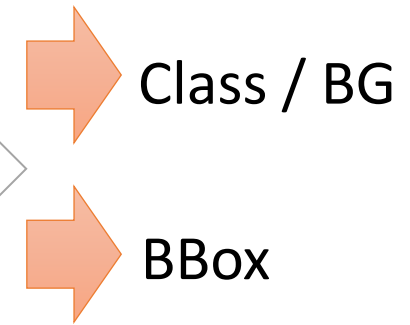
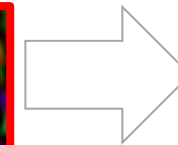
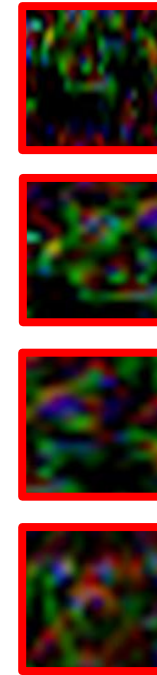
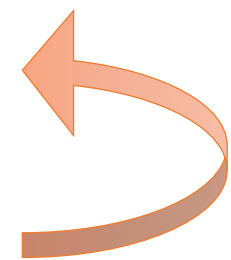
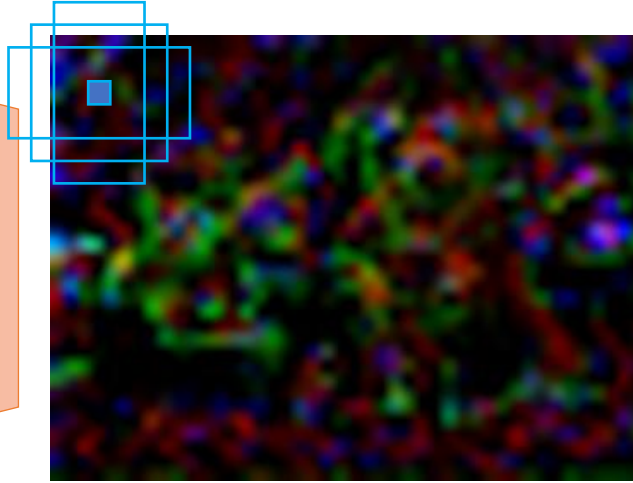
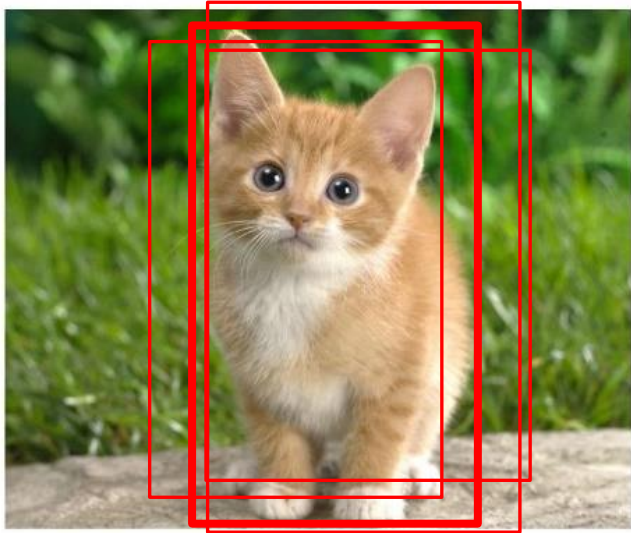
- Imbalance data
- Multiscale





MEN. WOMEN. BABIES. ELDERLY.  
ARE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow

# Object Detection



Can we do better?  
“Set” prediction

**MEN. WOMEN. BABIES. ELDERLY.**  
**ARE STILL HELD HOSTAGE BY HAMAS**  
**#BringThemHomeNow**

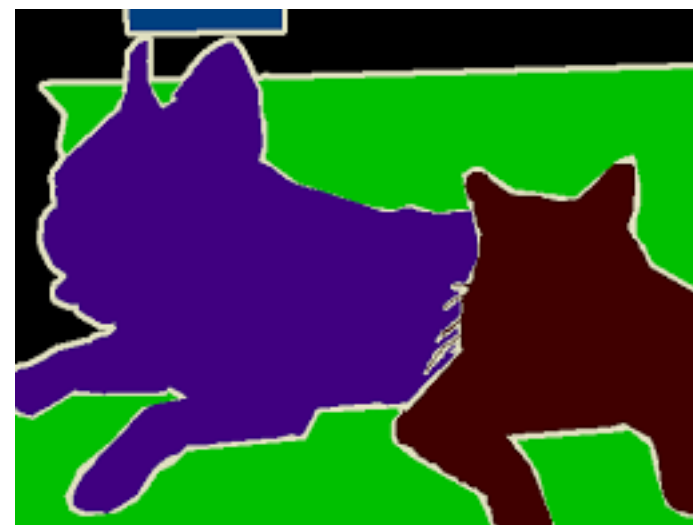
# Semantic Segmentation



# Semantic Segmentation

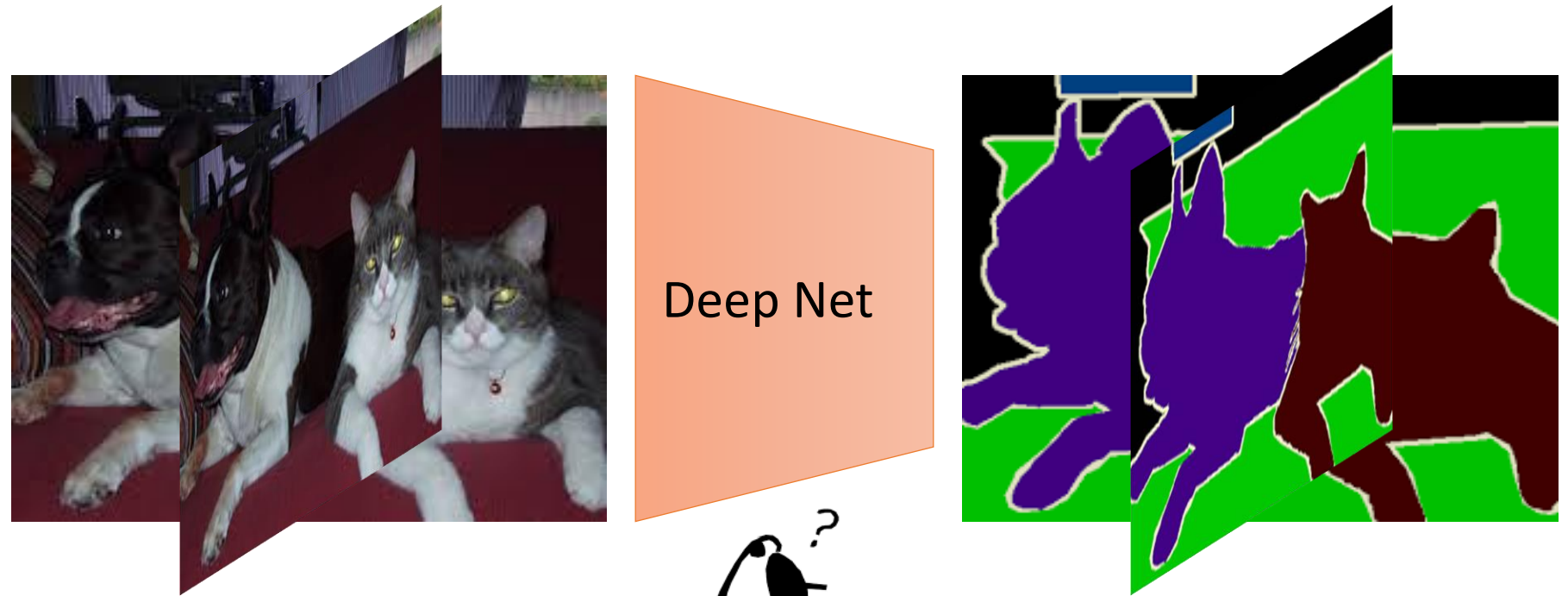


Deep Net



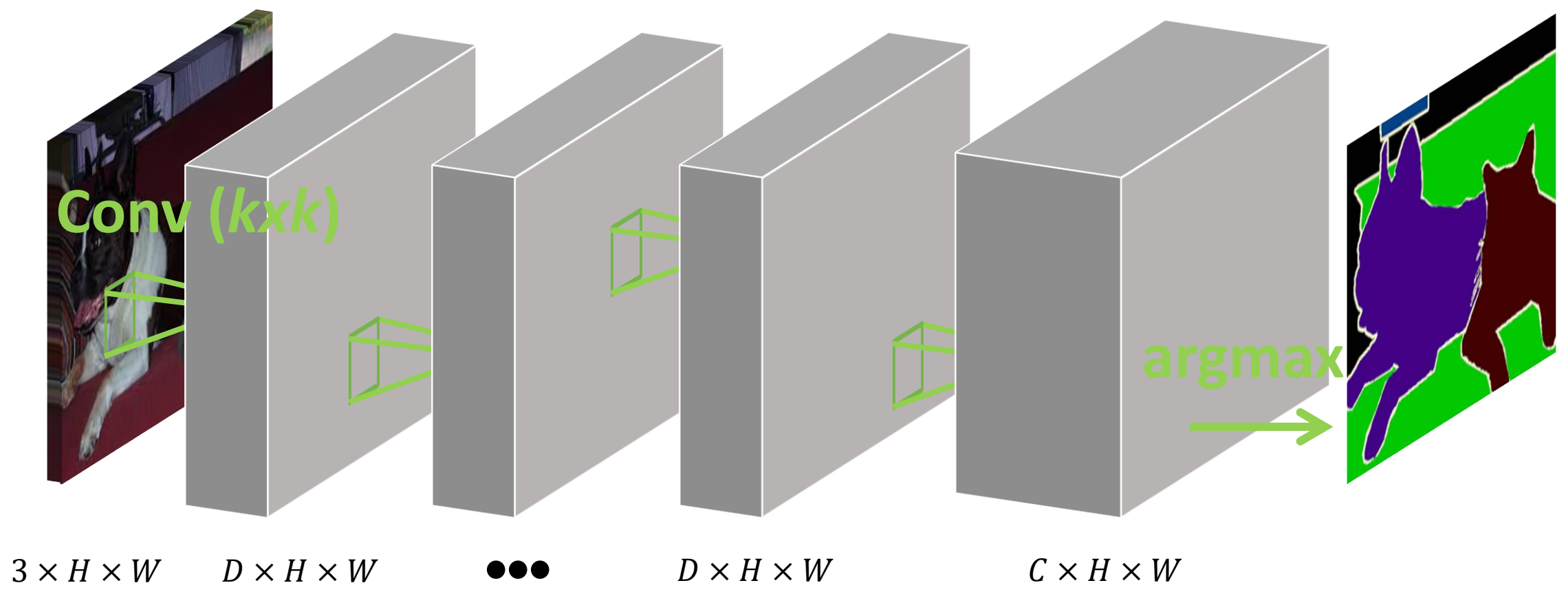
MEN. WOMEN. BABIES. ELDERLY.  
ARE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow

# Semantic Segmentation



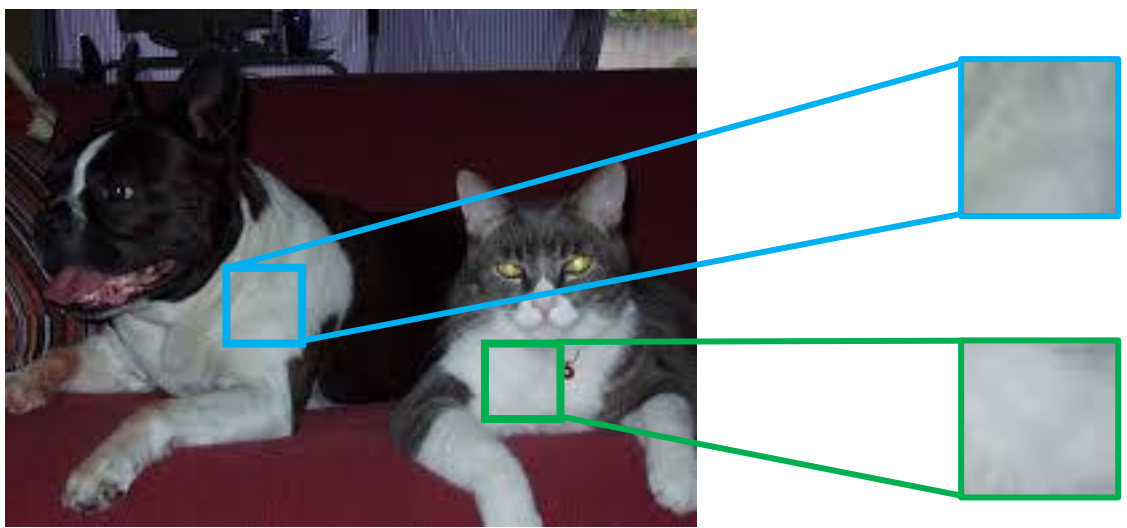
MEN. WOMEN. BABIES. ELDERLY.  
ARE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow

# Semantic Segmentation (v0)



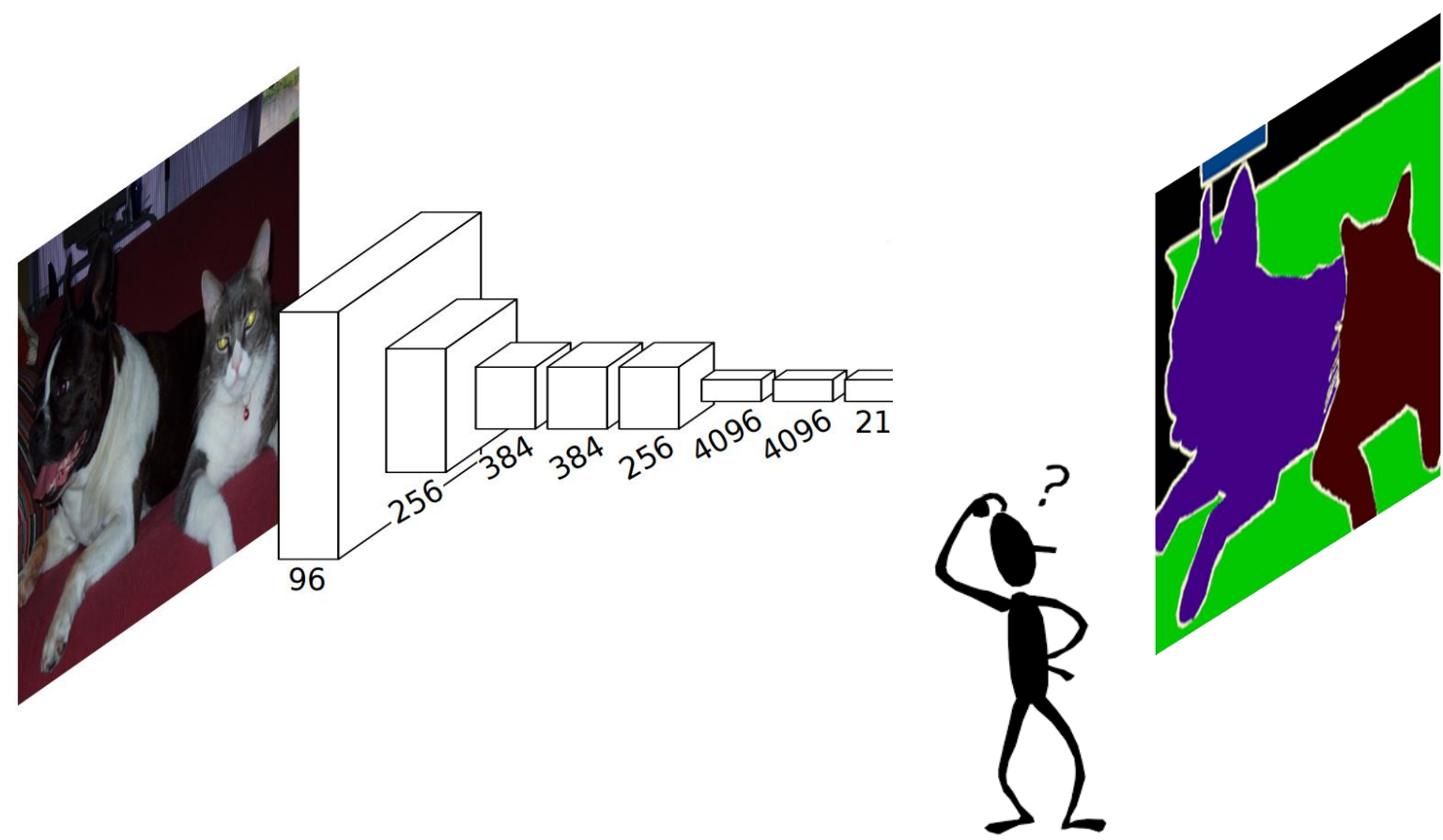
# Semantic Segmentation (v0)

## Challenge: Context (RF) vs Resolution



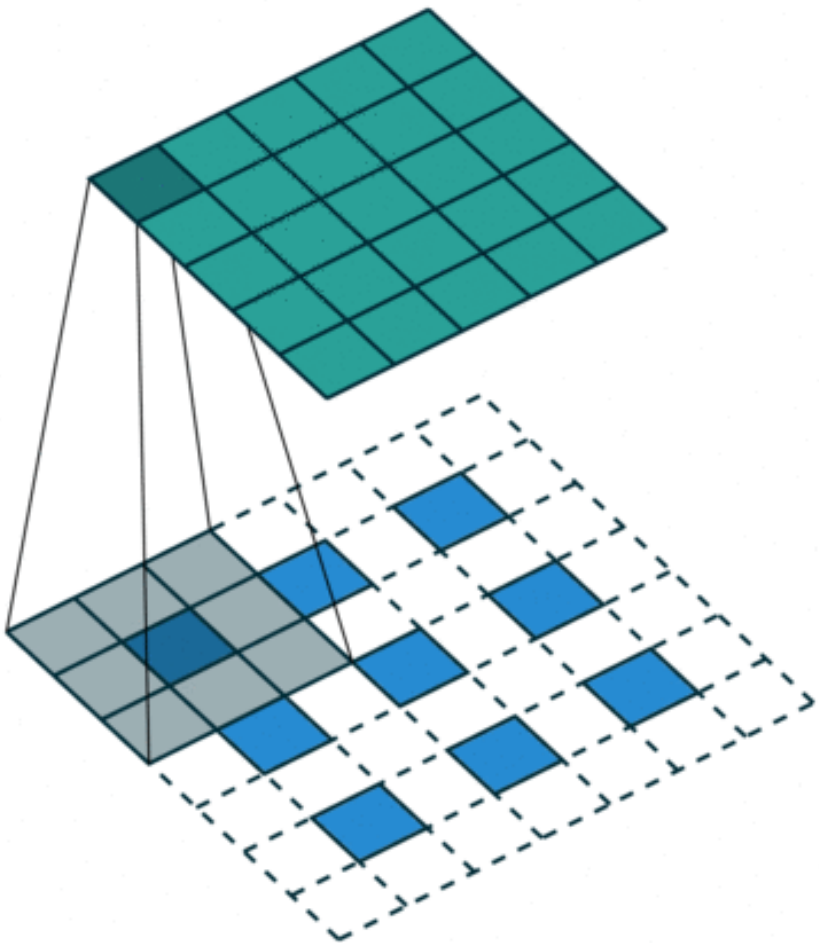
# Semantic Segmentation - FCN

Replace FC layers with conv – “sliding window” classification



MEN. WOMEN. BABIES. ELDERLY.  
ARE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow

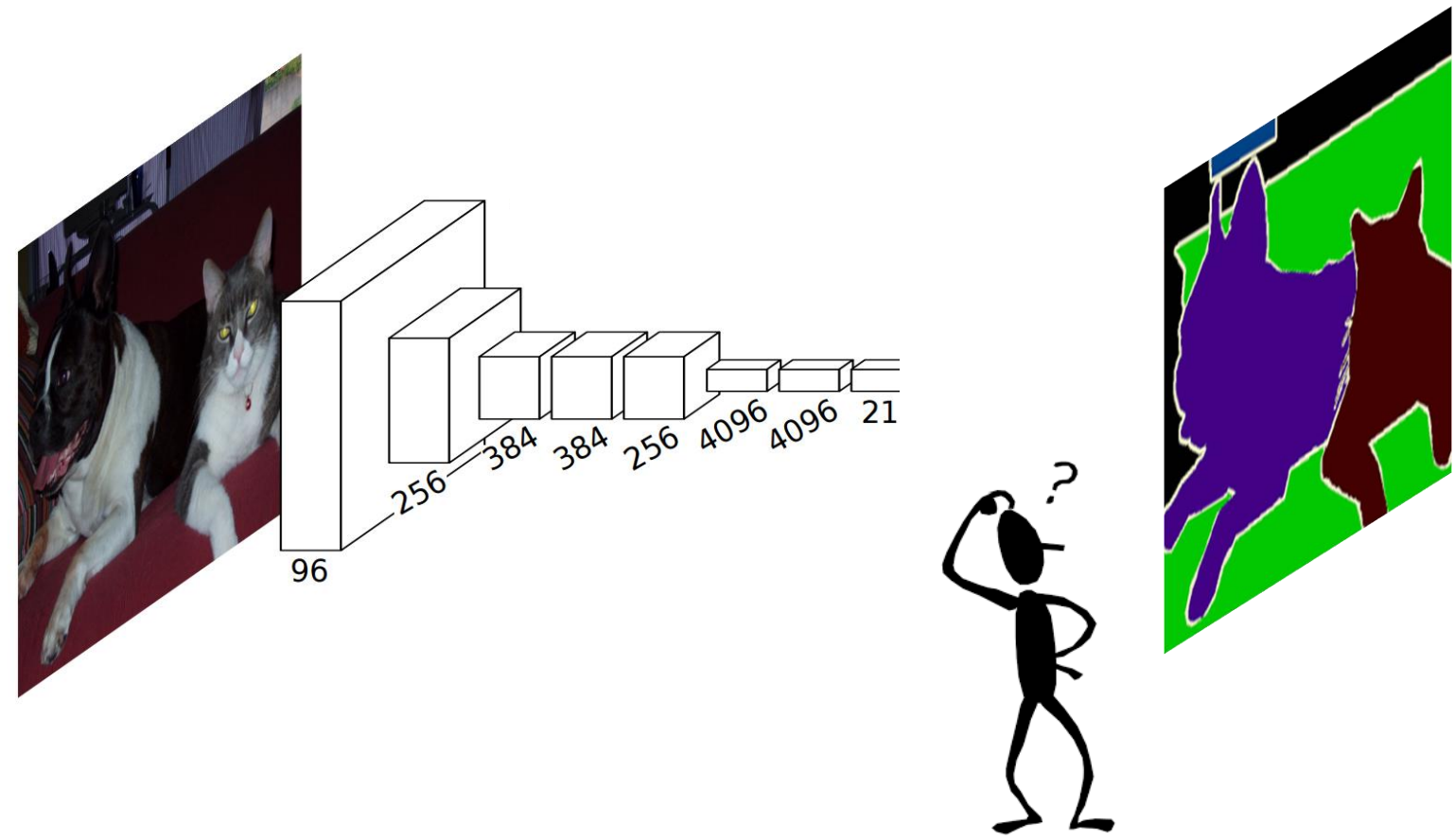
# Transposed Convolution





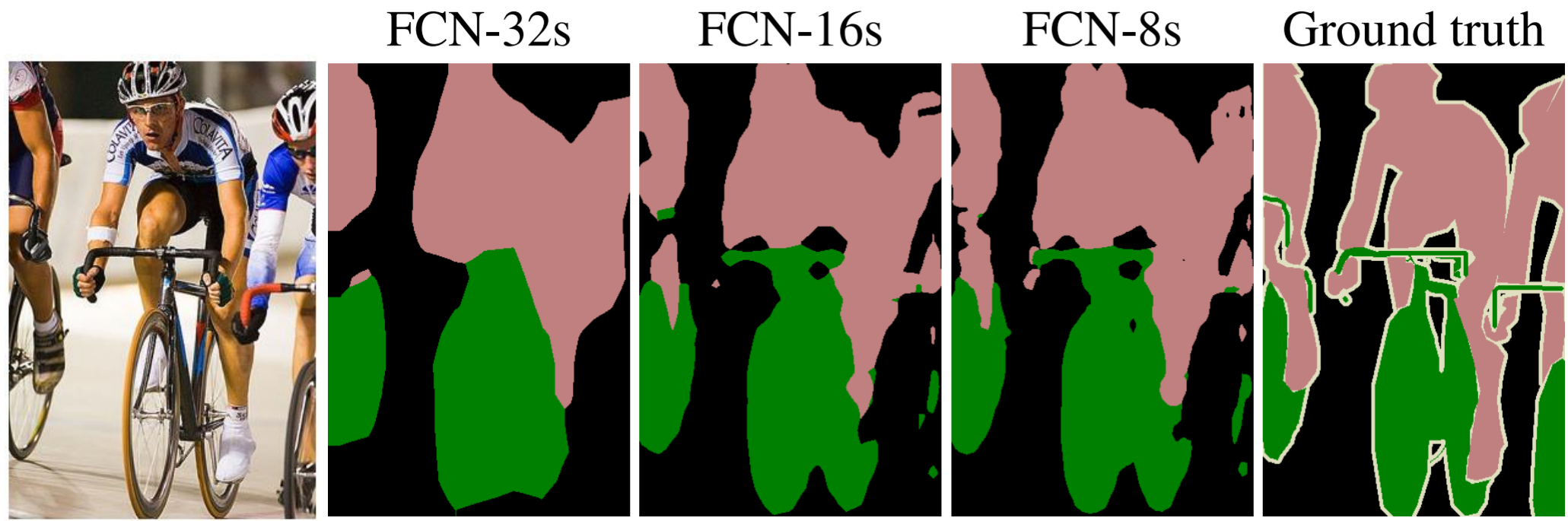
# Semantic Segmentation - FCN

Replace FC layers with conv – “sliding window” classification



MEN. WOMEN. BABIES. ELDERLY.  
ARE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow

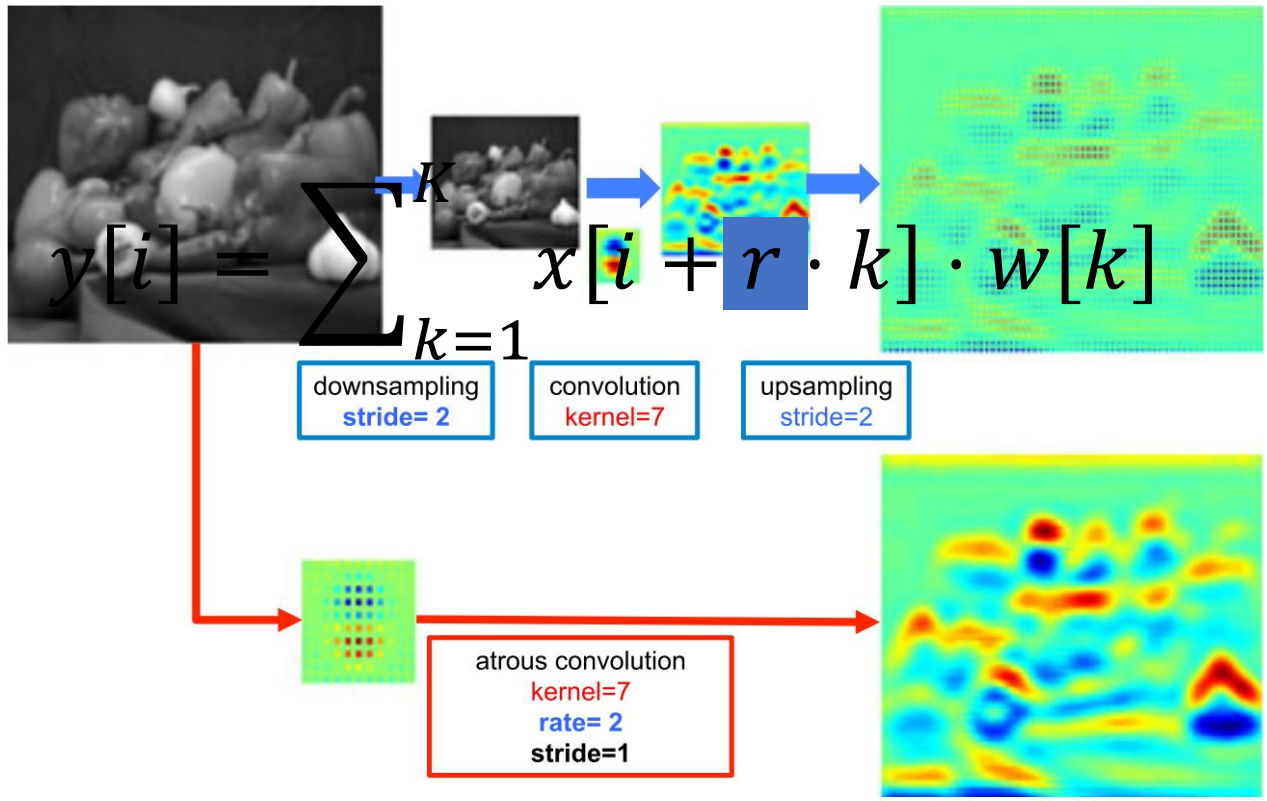
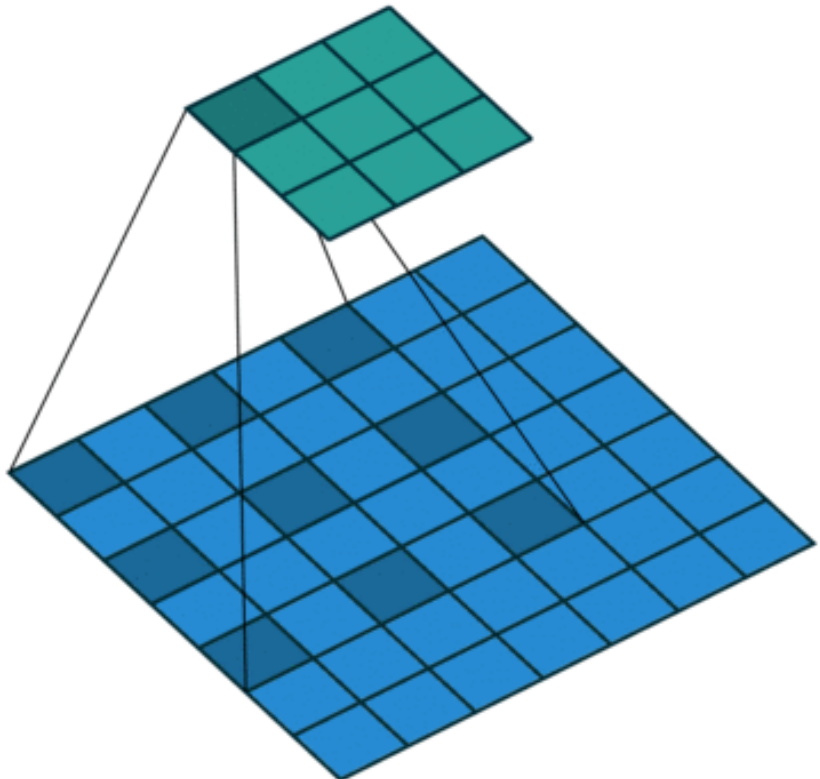
# Semantic Segmentation - FCN



# DeepLab: Atrous Convolution

Chen, Papandreou, Kokkinos, Murphy and Yuille

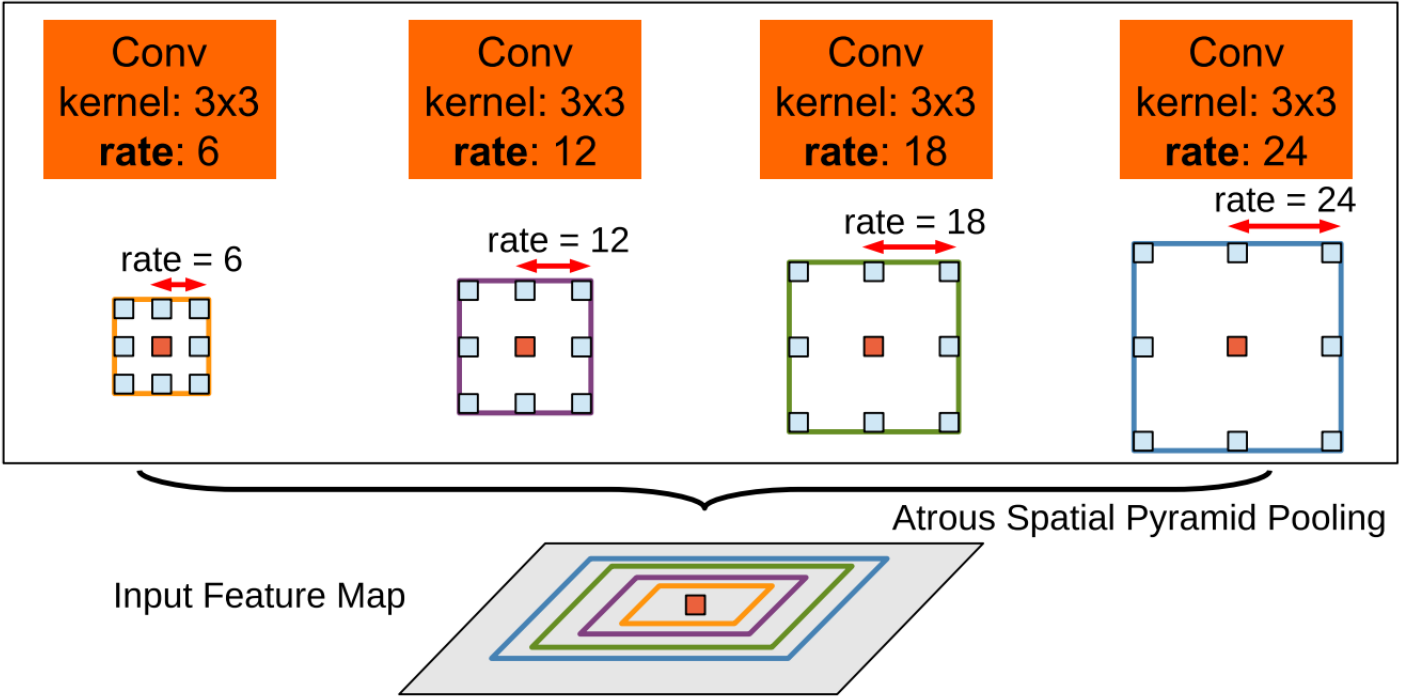
[Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs \(PAMI 2018\)](#)



# DeepLab: Atrous Convolution

Chen, Papandreou, Kokkinos, Murphy and Yuille

[DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs \(PAMI 2018\)](#)

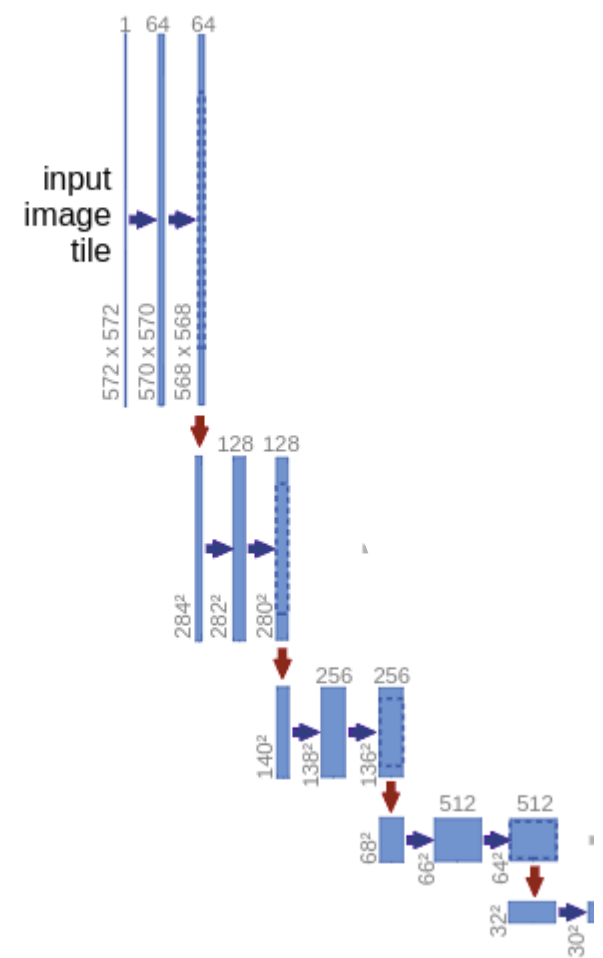


# DeepLab: Atrous Convolution

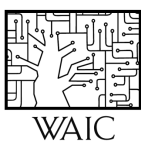
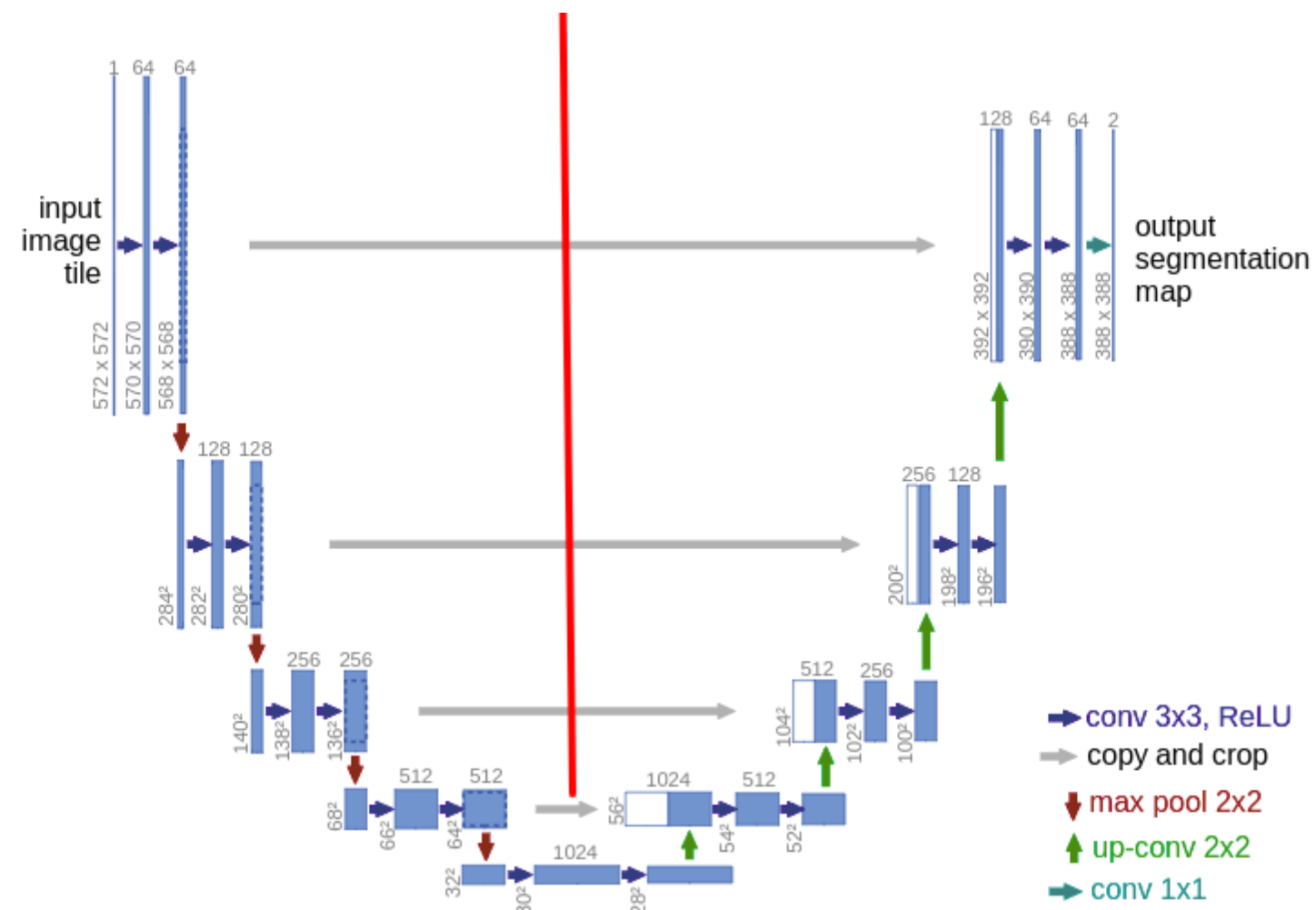
- Trade stride/pooling with “dilation” of kernel
- Increase receptive field without increase in parameters



# Semantic Segmentation – U-net



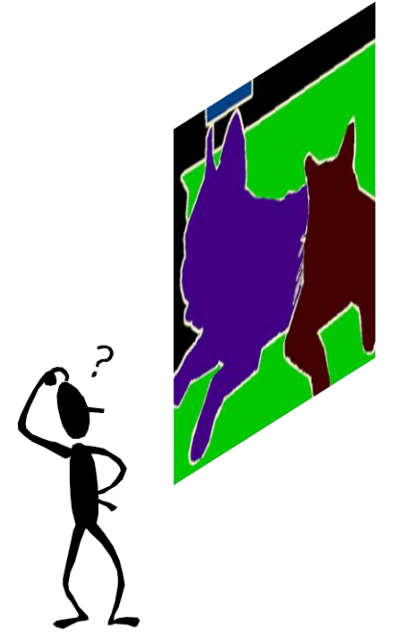
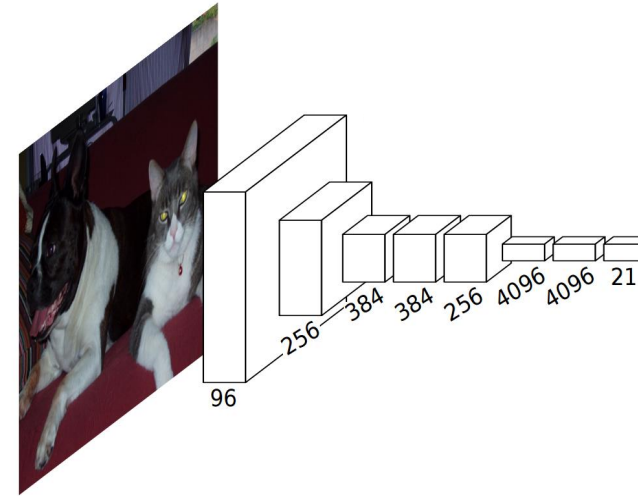
# Semantic Segmentation – U-net



# Semantic Segmentation

## Resolution vs. Semantic information

- FCN: using “transposed convolution”
- DeepLab: dilated convolution + simple interpolation
- U-net: skip connections





**MEN. WOMEN. BABIES. ELDERLY.**  
**ARE STILL HELD HOSTAGE BY HAMAS**  
**#BringThemHomeNow**

# Instance Segmentation



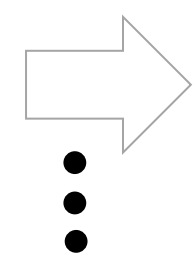
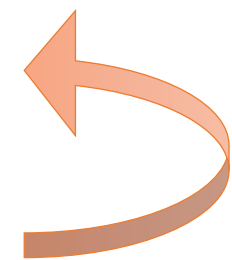
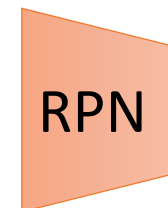
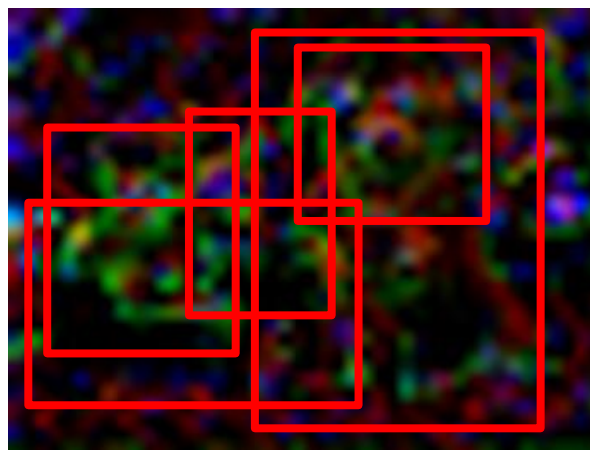
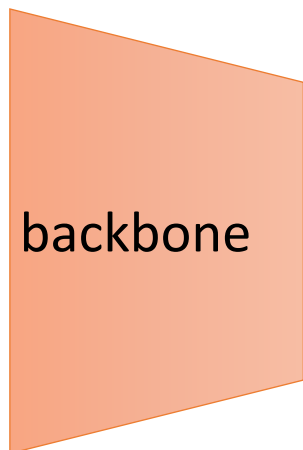
**MEN. WOMEN. BABIES. ELDERLY.**  
**ARE STILL HELD HOSTAGE BY HAMAS**  
**#BringThemHomeNow**

# Instance Segmentation: Detection or Segmentation!



# Instance Segmentation: Mask R-CNN

He, Gkioxari, Dollar and Girshick "Mask R-CNN" (ICCV 2017)



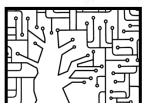
- Class / BG
- BBox
- Mask



MEN. WOMEN. BABIES. ELDERLY.  
ARE STILL HELD HOSTAGE BY HAMAS  
#BringThemHomeNow

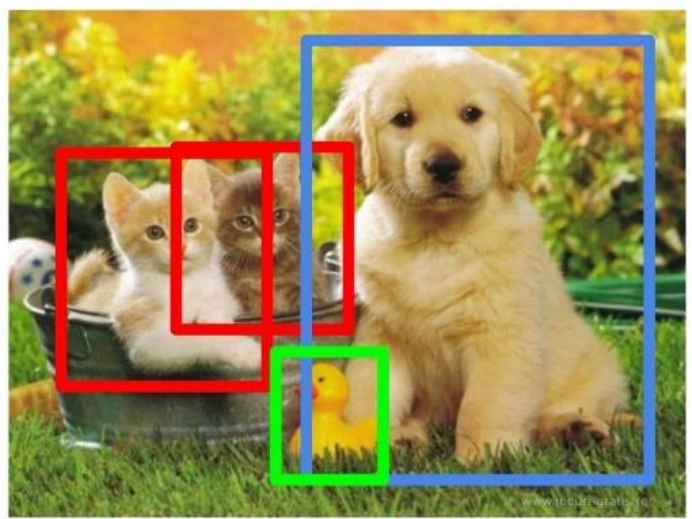
# Instance Segmentation: Mask R-CNN

He, Gkioxari, Dollar and Girshick “Mask R-CNN” (ICCV 2017)



# Summary

- Deep learning beyond image classification
- Classification “backbones” = “transfer learning”
- Same features - multitasking
- Handling varying number of predictions
- Coping with RF/resolution trade-offs



# What's next?

## Next week

Monday – Text & Images (Rafail)

Wednesday – Detection and Segmentation II (Lior)

