Global average pooling in deep ConvNets

Matthieu Cord
Joint work with Thibaut Durand and Nicolas Thome*

Sorbonne Universities, UPMC Paris 6, CNRS
*CEDRIC, CNAM

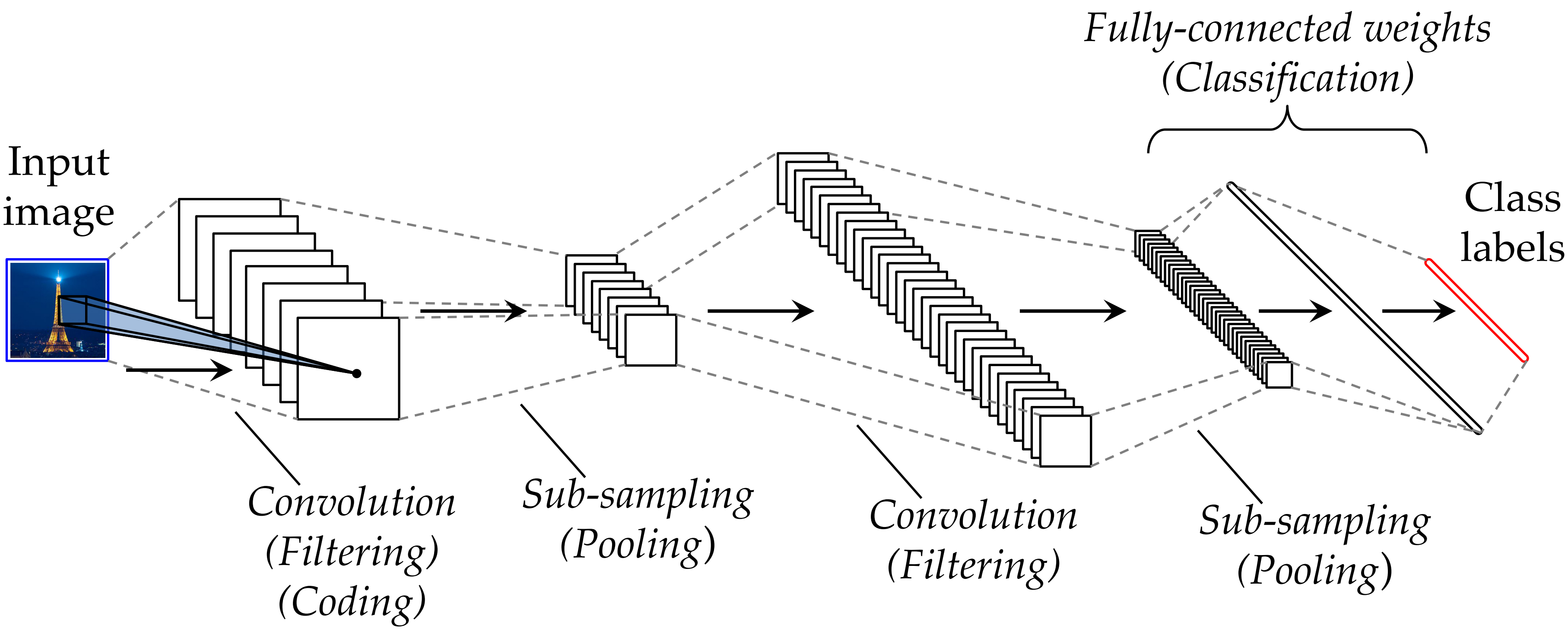
Outline

1. Deep net framework
2. Fully Convolutional Nets
3. Where is Pooling inside the architecture?
4. How to pool?

Deep Convolutional Neural Networks (Deep ConvNets)

[Fukushima 79]

[LeCun-89]

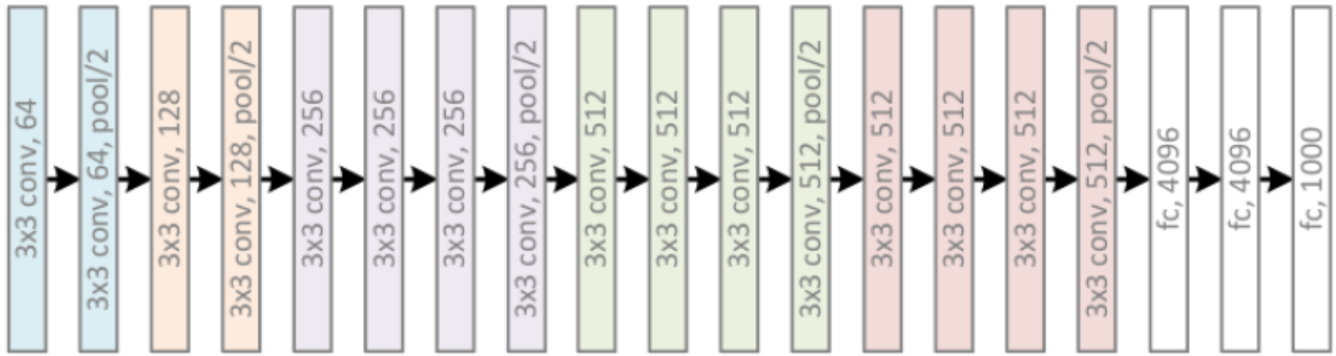


- **Convolution** uses local weights shared across the whole image
- **Pooling** shrinks the spatial dimensions

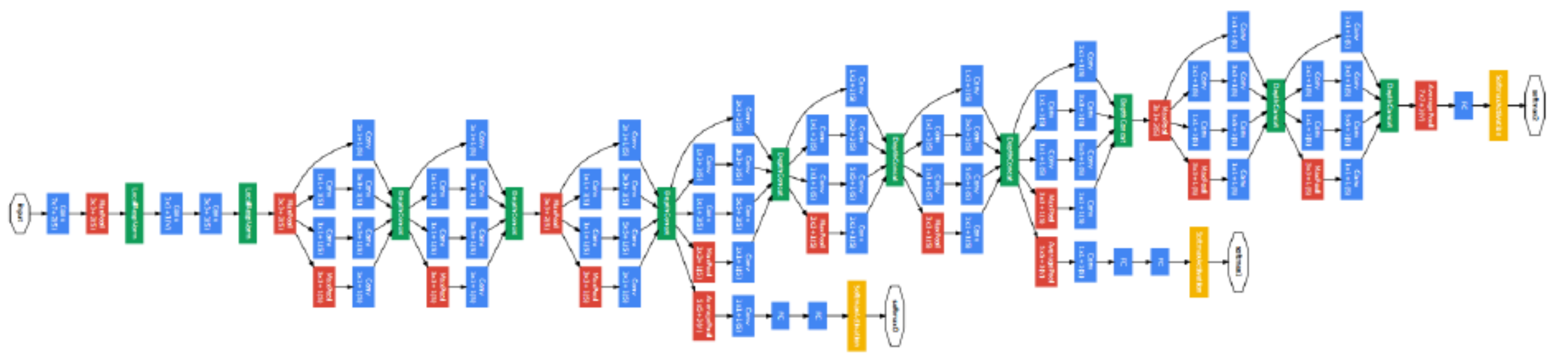
[Hinton-12]

Post 2012 deep architectures

VGG, 16/19 layers, 2014



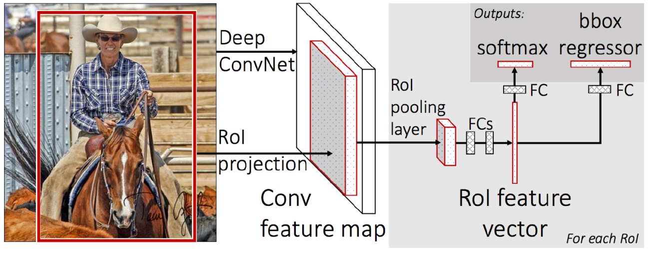
GoogleNet, 22 layers, 2014



ResNet, 152 layers, 2015

Key issues for Deep&Vision

- Computer Vision: from the ImageNet Object recognition task
  - **Classification: How to do for large and complex scenes?**
  - Detection: R-CNN Fast/Faster R-CNN
    [Girshick, CVPR14, ICCV 15, NIPS 15]
  - Segmentation



Girshick. *Fast R-CNN*. ICCV 2015

- Supervised/Unsupervised – learning generic data representation
- Theoretical support to understand deep: convergence, why it works,…
- Vision and Language
- Connection to Computational/informational Neurosciences
- Compression/Embedded/Green nets
- Deep generative models,
- …

How to deal with complex scenes?

ImageNet style

Pascal VOC style

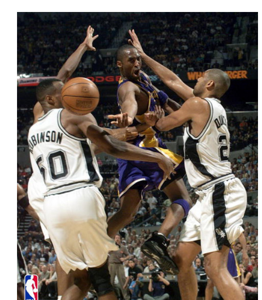
- Working on datasets with complex scenes (large and cluttered background), not centered objects, variable size, ...

VOC07/12

MIT67

15 Scene

COCO



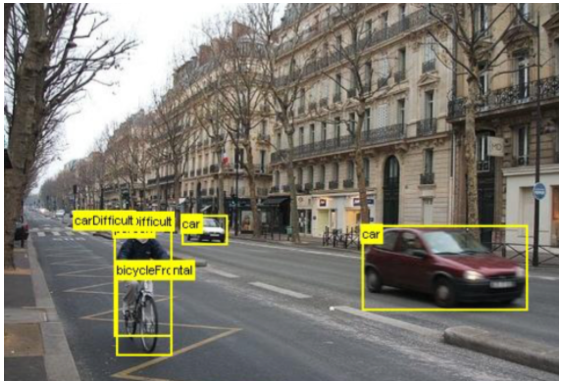
VOC12 Action

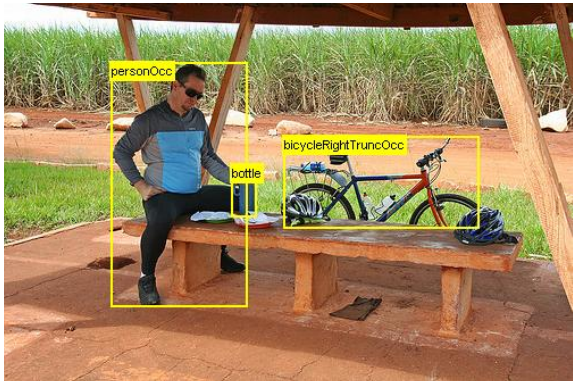
From ImageNet to complex scenes?

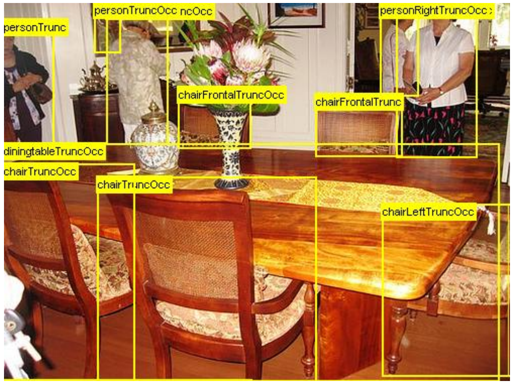
- Naive approach: resize the image
- Region based approach: use regions to have images that look like ImageNet [Oquab, CVPR14]

	Naive	Region
VOC 2012 (AP)	70.9 %	78.7 %

- Regions → better prediction







- Full annotations expensive → training with weak supervision

From ImageNet to complex scenes?

- Working on datasets with complex scenes (large and cluttered background), not centered objects, variable size, ...


VOC07/12

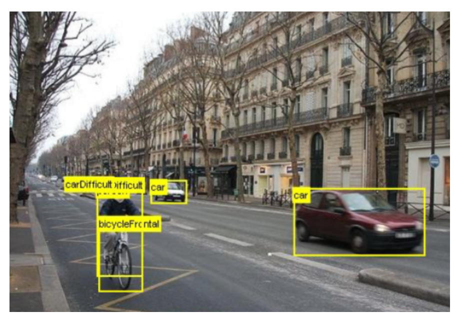

MIT67


15 Scene


COCO


VOC12 Action

- Select relevant regions $\rightarrow$ better prediction

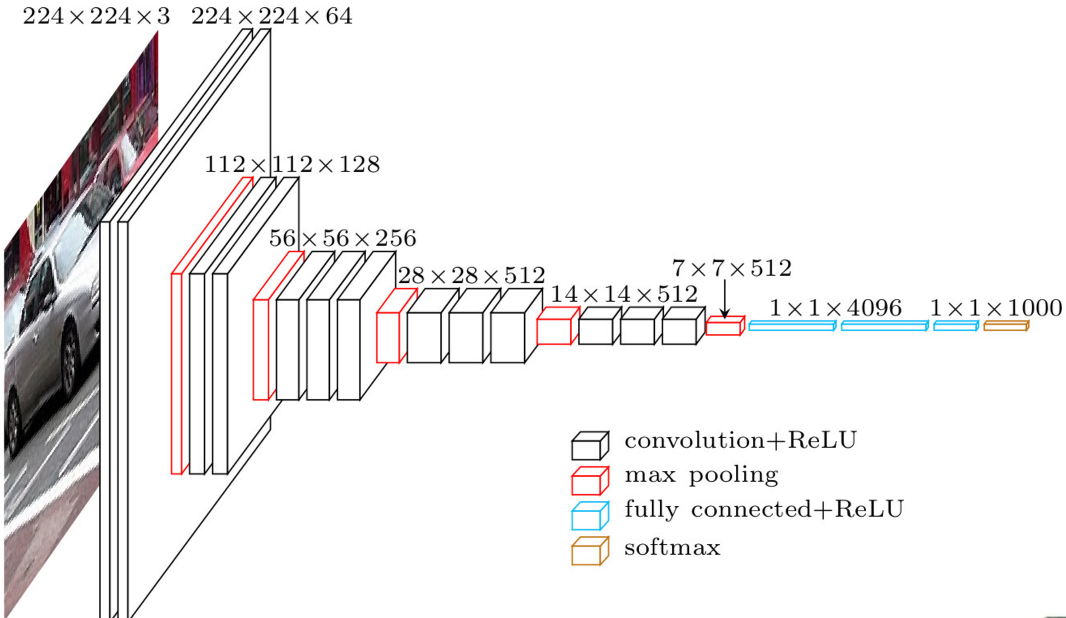


- Full annotations expensive $\Rightarrow$ training with weak supervision

Outline

1. Deep net framework
2. **Fully Convolutional Nets**
3. Where is Pooling inside the architecture?
4. How to pool?

VGG-16 [Simonyan, ICLR15]

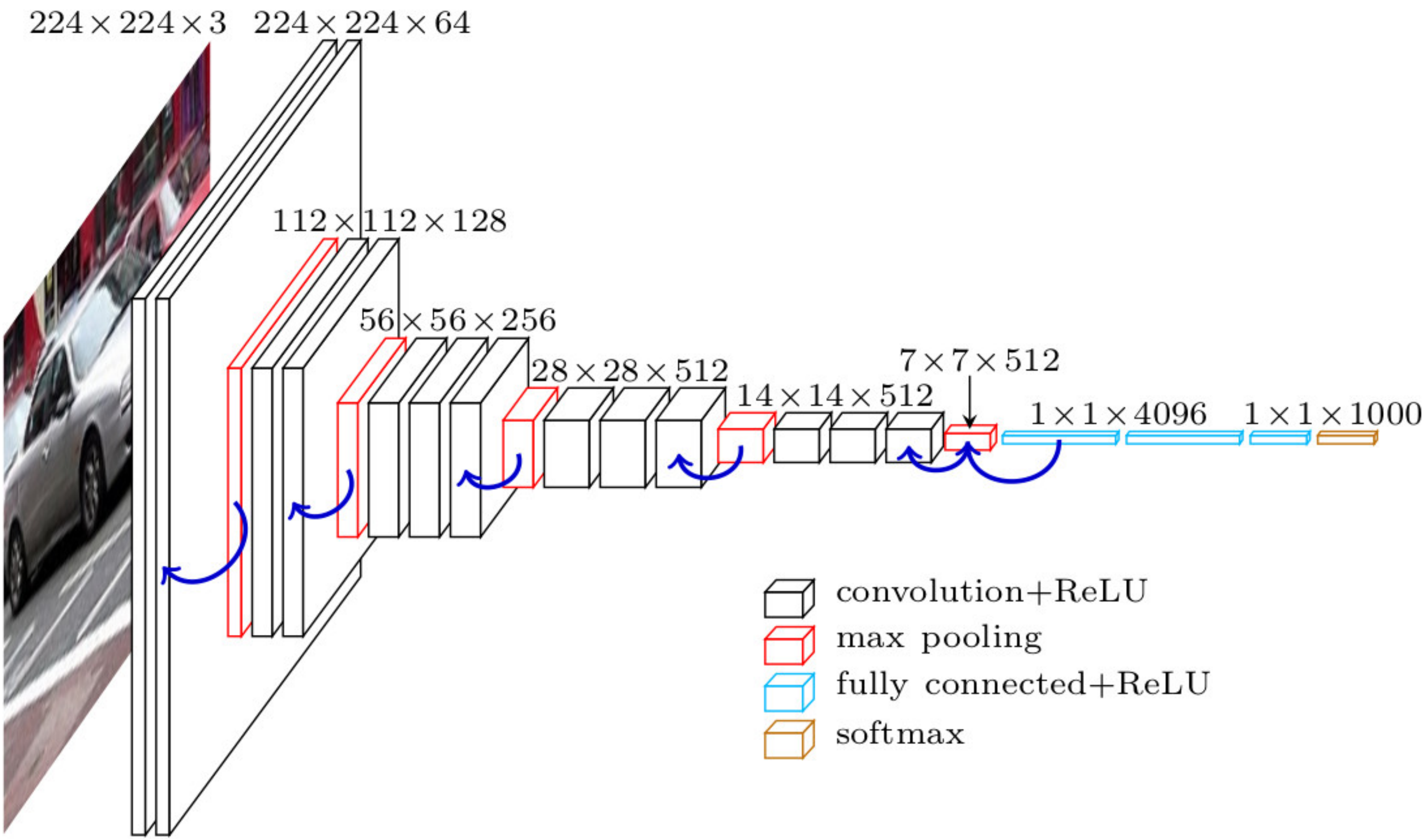


How to adapt VGG scheme for large images?

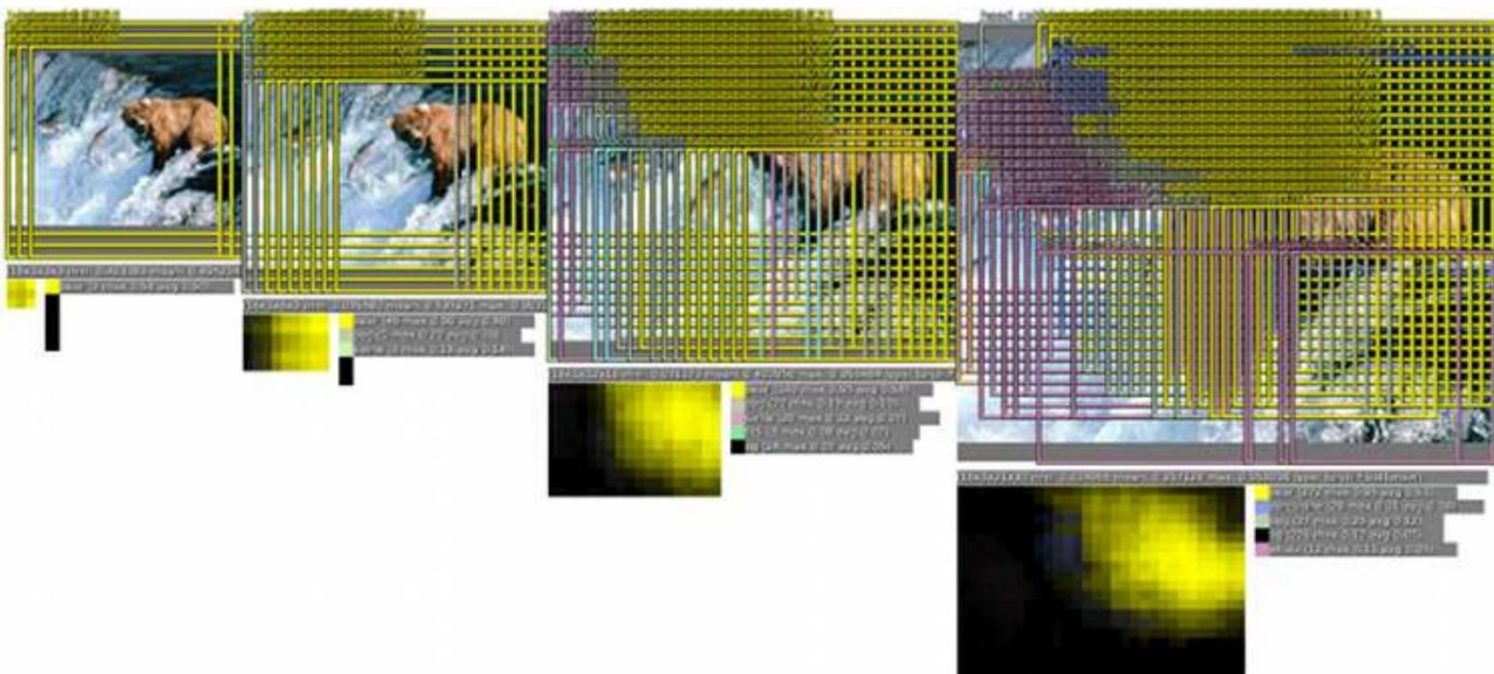
Simonyan et al. *Very deep convolutional networks for large-scale image recognition.* ICLR 2015

VGG-16

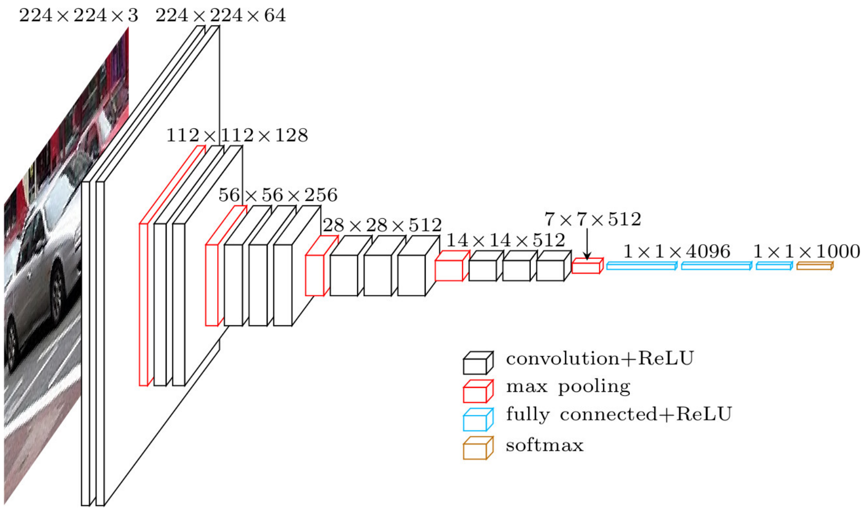
Input image: fixed size

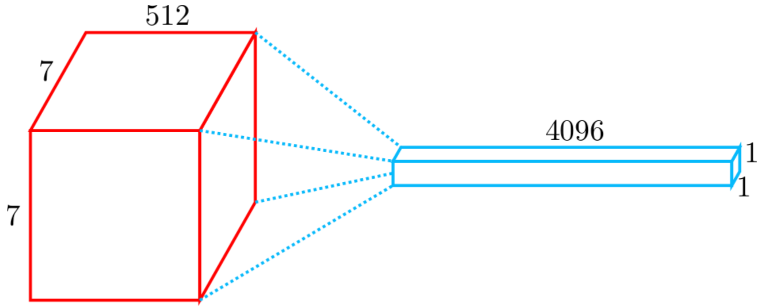


Sliding window [Sermanet, OverFeat14]

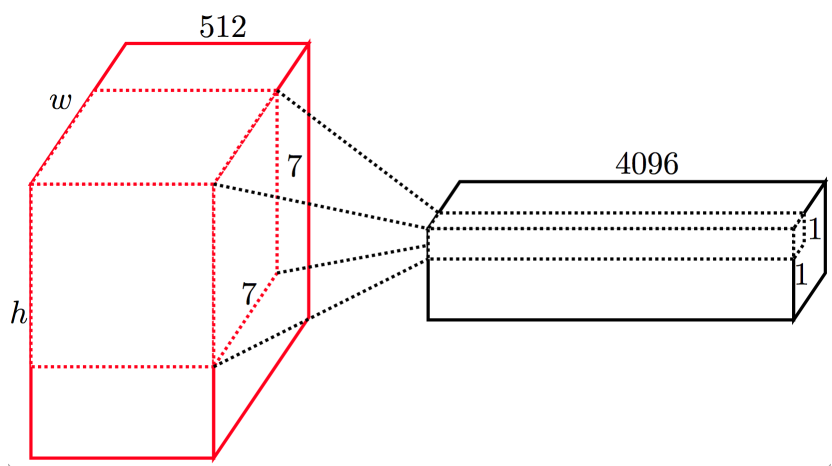


Sliding window => Convolutional Layers

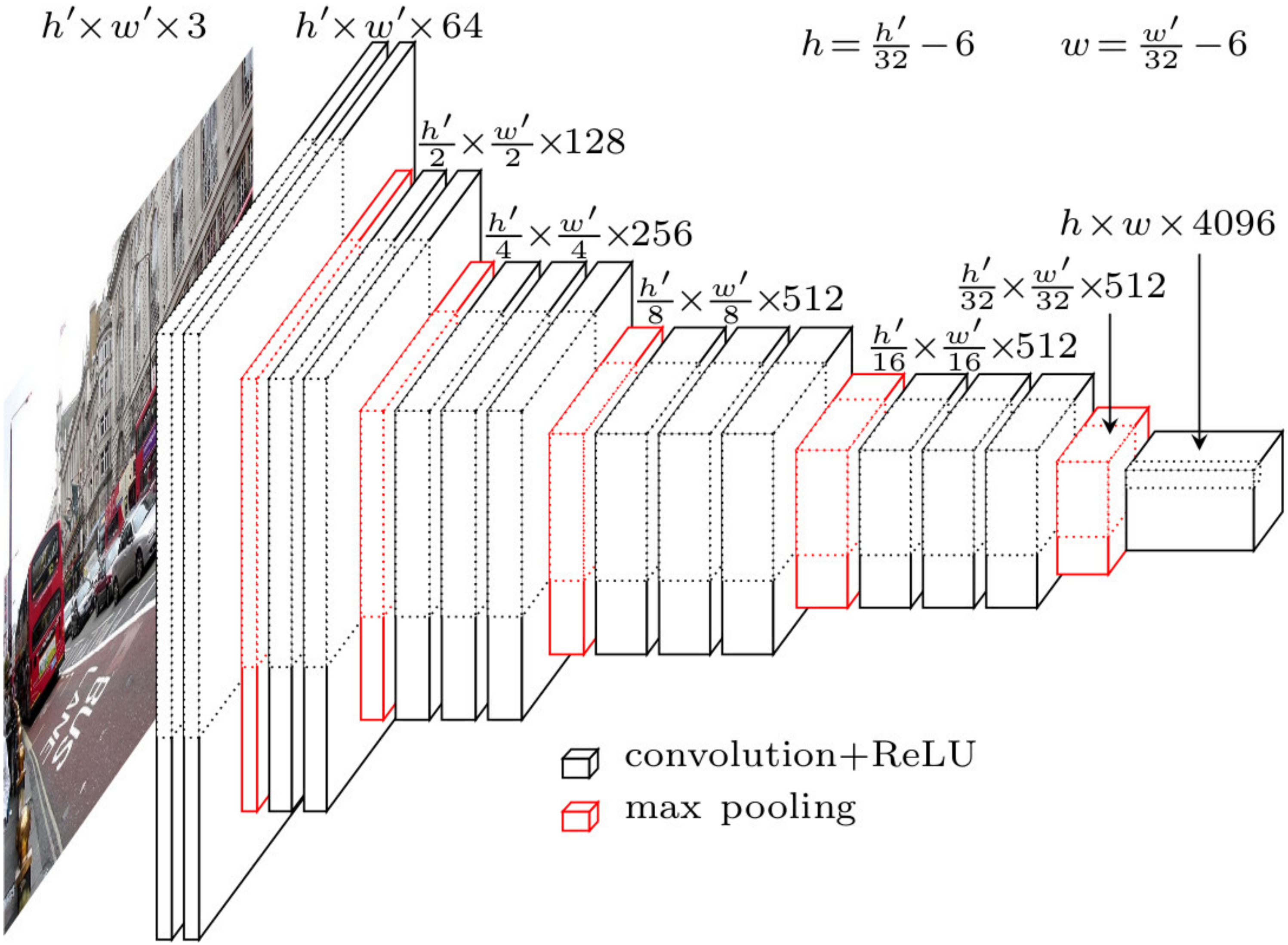




Fully connected as convolutional layer (here 4096 conv. filters 7x7x512)



Sliding window => Convolutional Layers



Fully connected layers as conv layers

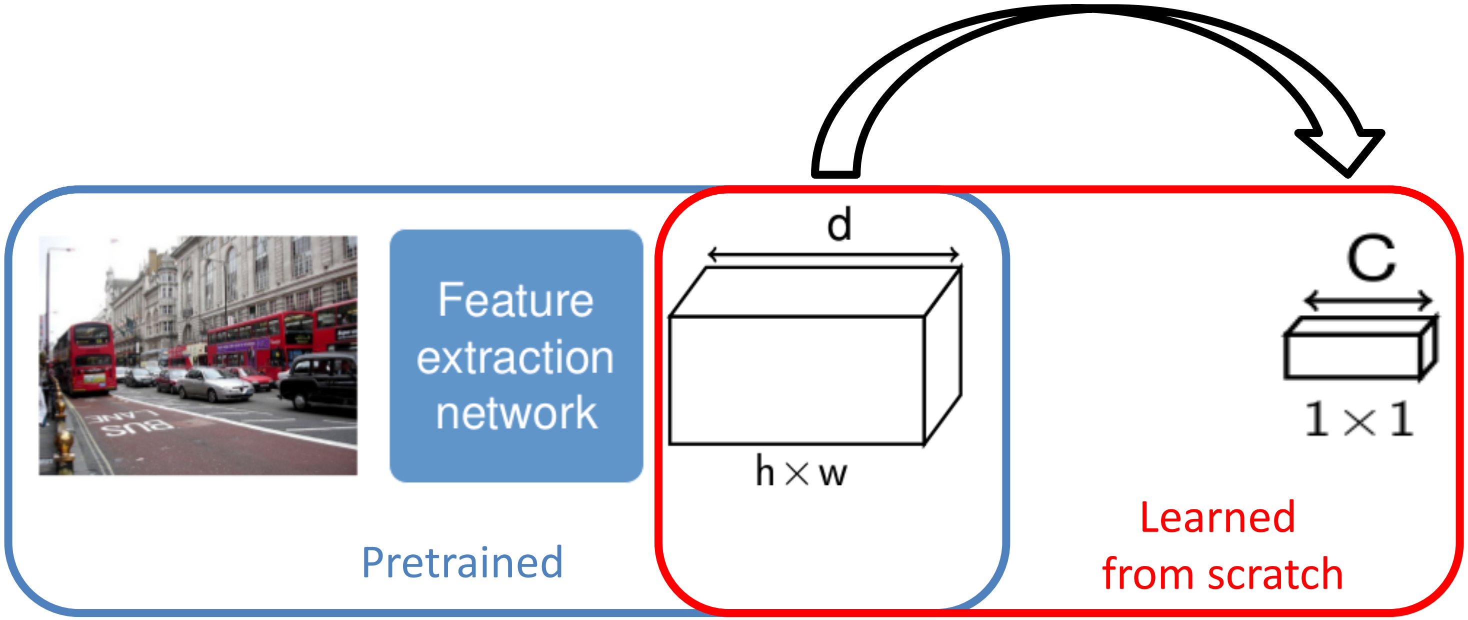
In many archi to process large images/datasets

- OverFeat (Sermanet)
- Fast R-CNN (Girshick)
- Weldon (Durand)
- SPLeap++ (Kulkarni)

Outline

1. Deep net framework
2. Fully Convolutional Nets
3. **Where is Pooling inside the architecture?**
4. How to pool?

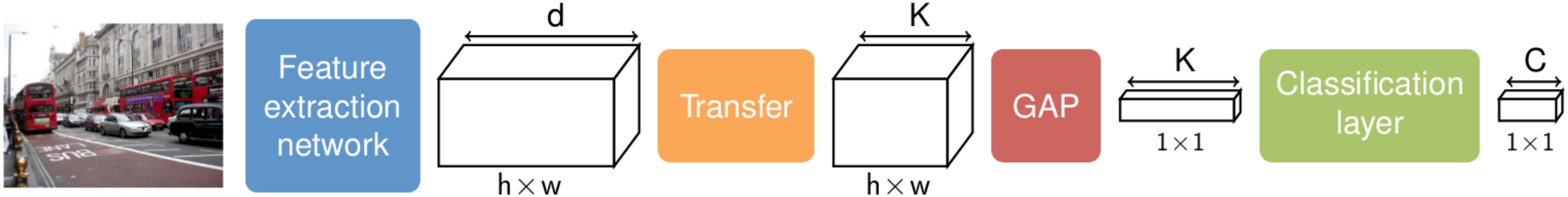
Transfer/Pooling/Classify



- Image-based strategy
- Region-based strategy

Transfer/Pooling

Global Average Pooling [Zhou, 2016], (ResNet)



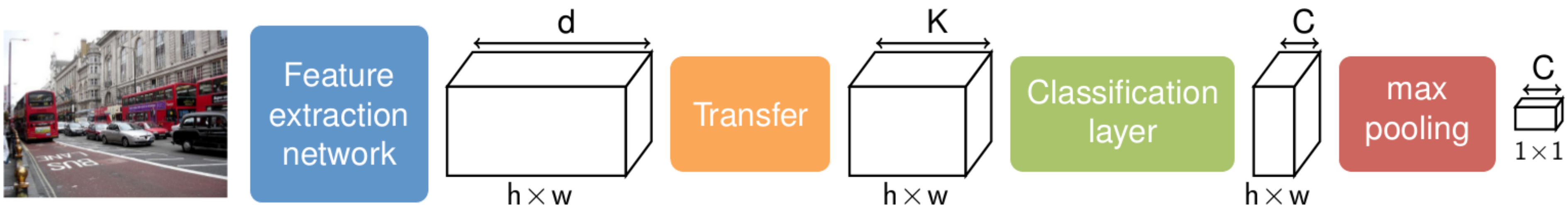
- ➢ Image-based strategy

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba.

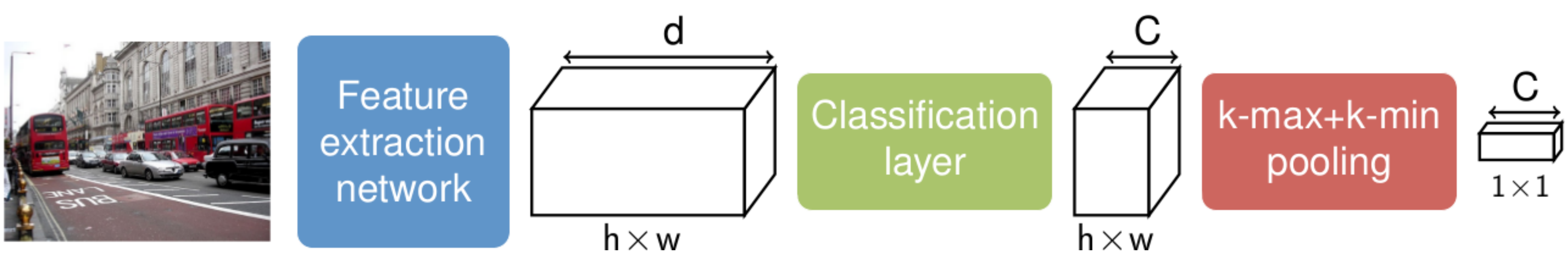
Learning Deep Features for Discriminative Localization. CVPR 2016

Transfer/Pooling

Deep MIL [Oquab, CVPR15]

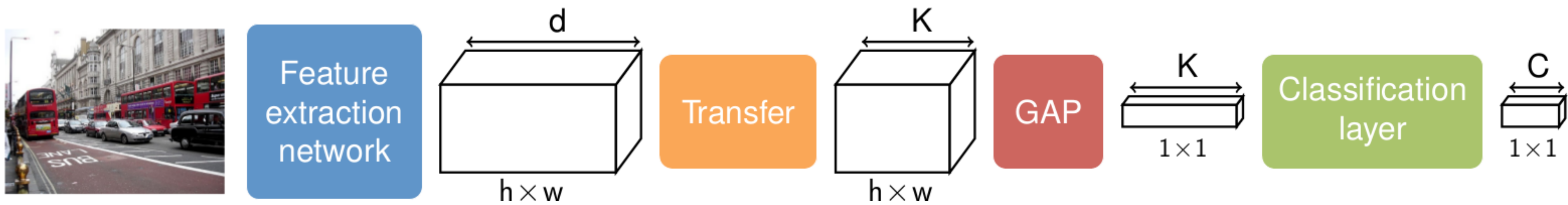


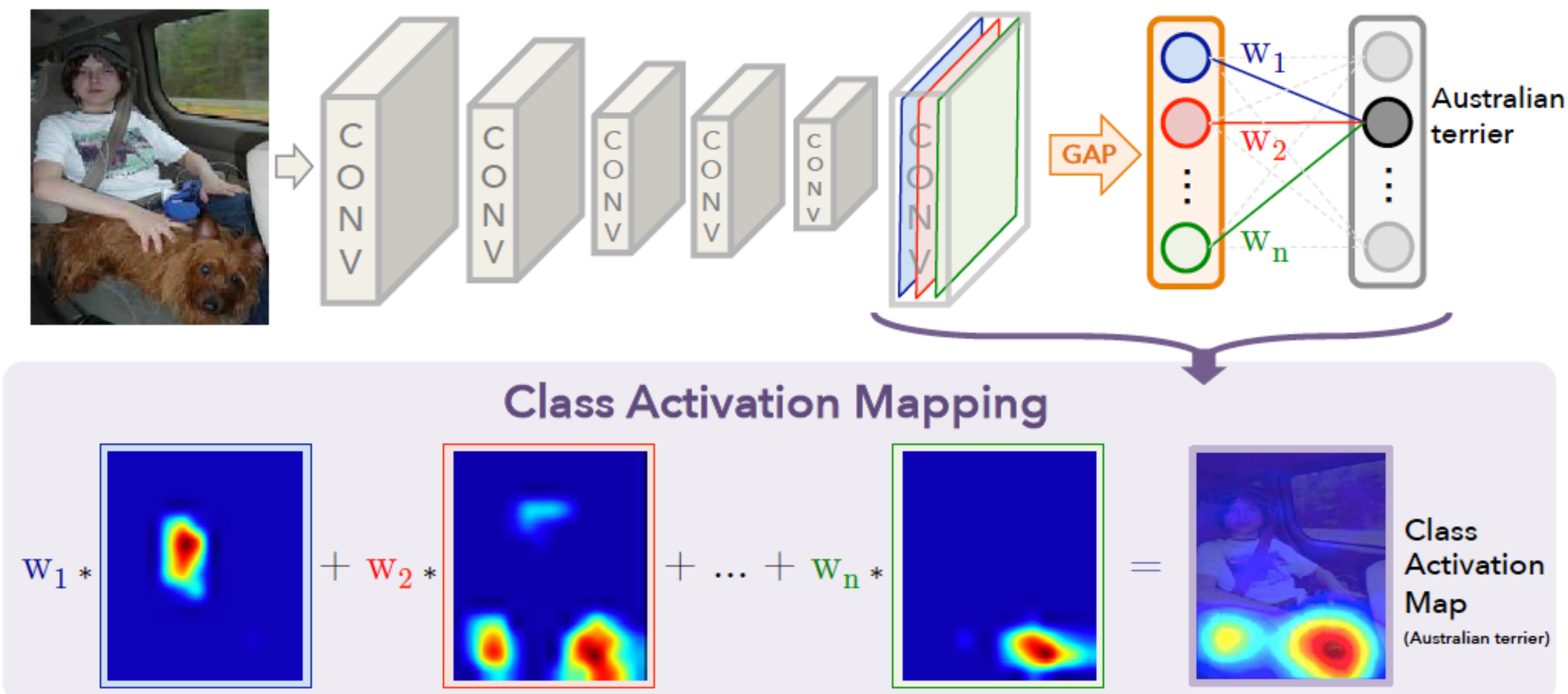
WELDON [Durand, CVPR16] (≈ProNet [Sun, CVPR16])



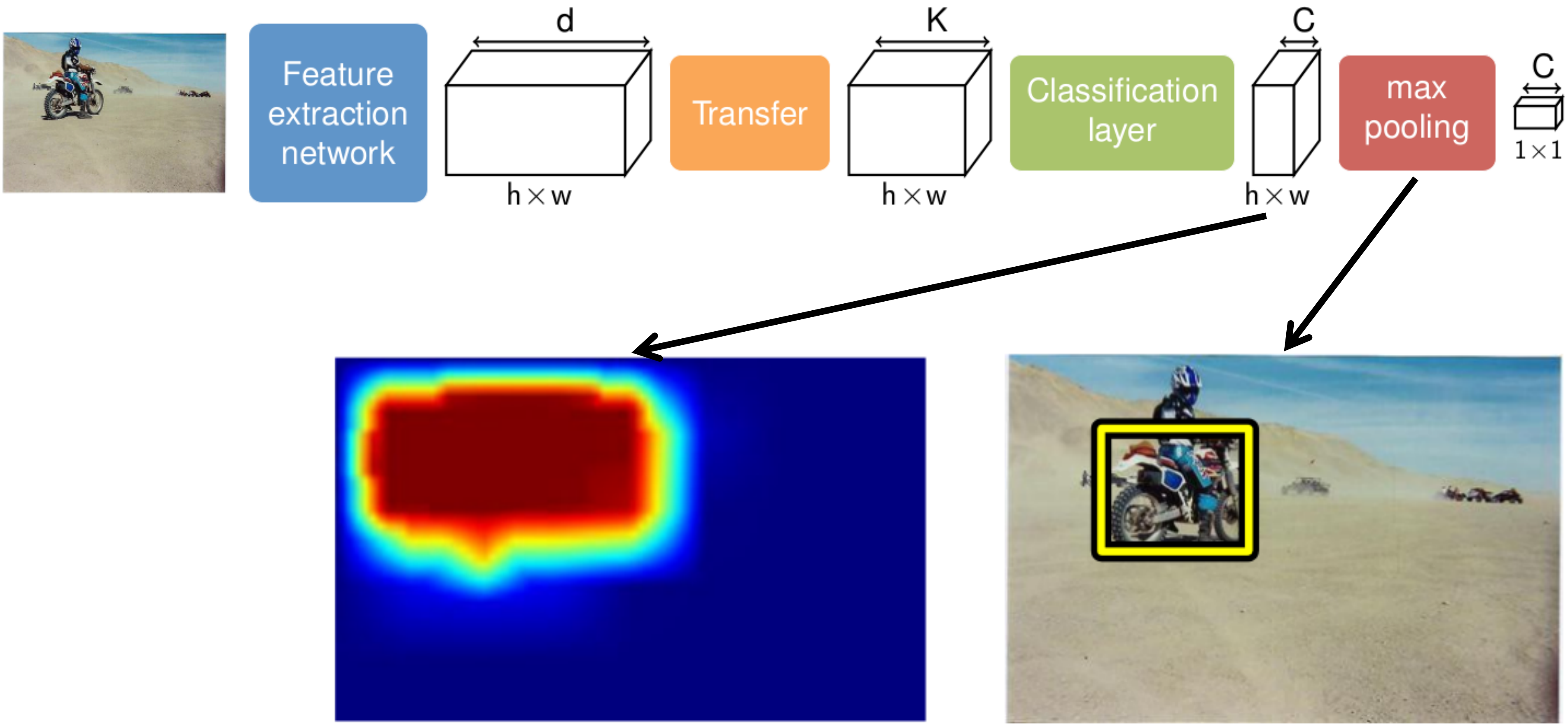
- ➢ Region-based strategy

Class Activation Mapping (CAM) for GAP [Zhou, CVPR16]

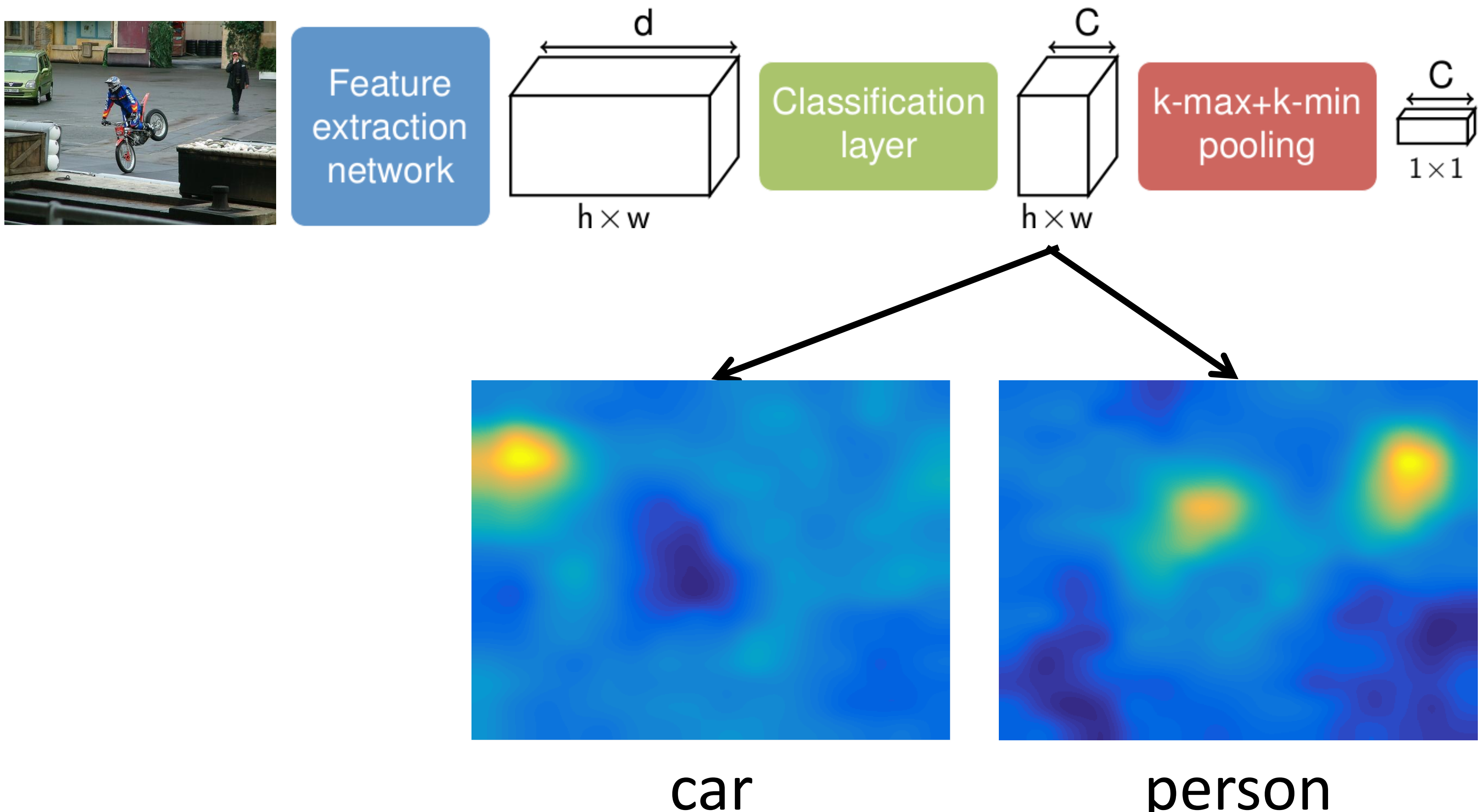




CAM
for [Oquab, CVPR15]



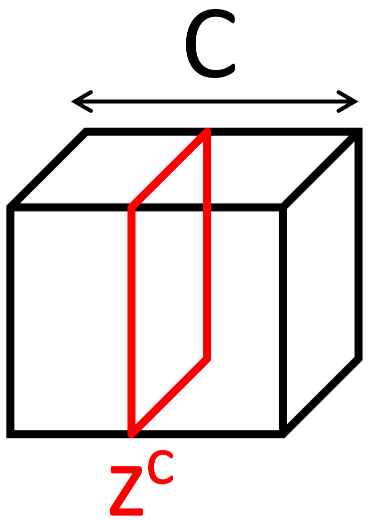
CAM
for WELDON [Durand, CVPR16]



Outline

1. Deep net framework
2. Fully Convolutional Nets
3. Where is Pooling inside the architecture?
4. **How to pool?**

Pooling schemes



- Max [Oquab, CVPR15]

$$y^c = \max_{i,j} \; z^c_{ij}$$

- GAP [Zhou, CVPR16]

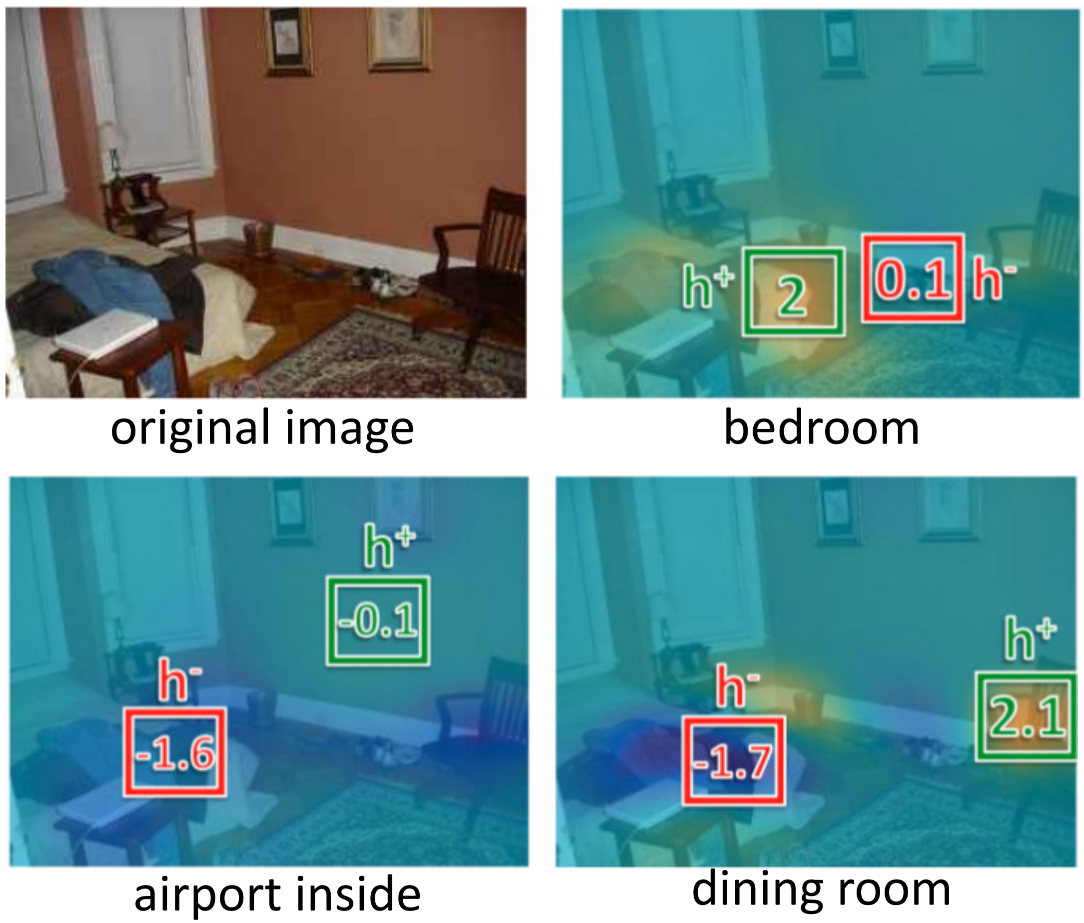
$$y^c = \frac{1}{N} \sum_{i,j} \; z^c_{ij}$$

- LSE [Pinheiro, CVPR15] / SPLeap [Kulkarni, ECCV16]

$$y^c = \frac{1}{\beta} \log \left(\frac{1}{N} \sum_{i,j} \; \exp(\beta \cdot z^c_{ij}) \right)$$

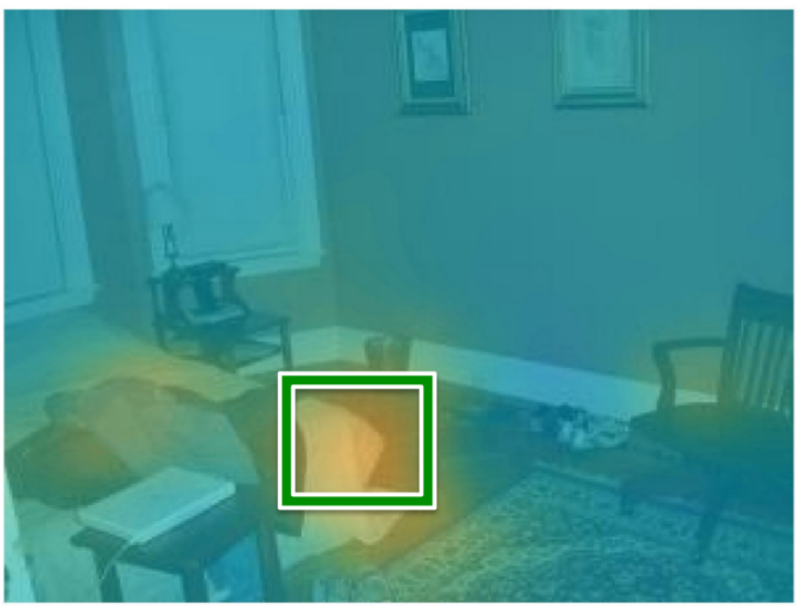
WELDON: max+min pooling

- h^+: presence of the class → high h^+
- h^-: localized evidence of the absence of class

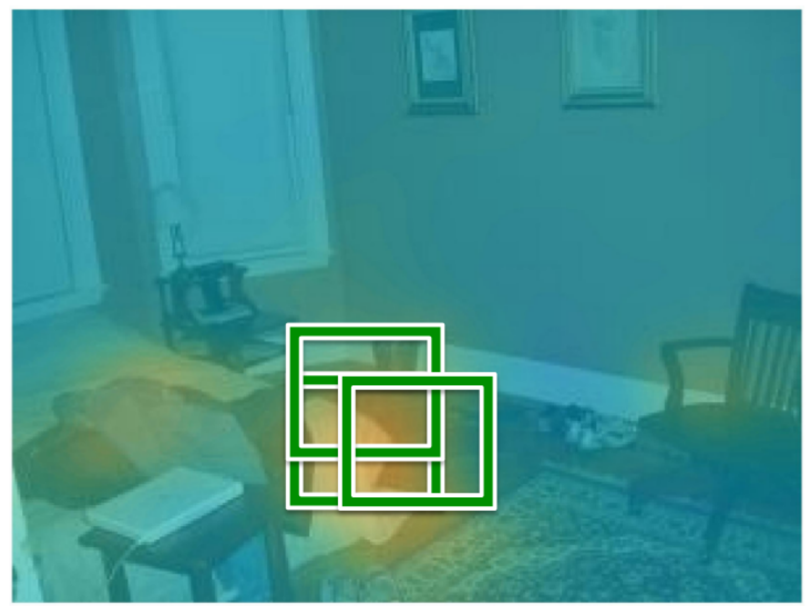


WELDON Pooling

- max + min strategy
- Top instances: using several regions, more robust region selection [Vasconcelos, CVPR15]



k=1



k=3

WELDON Pooling

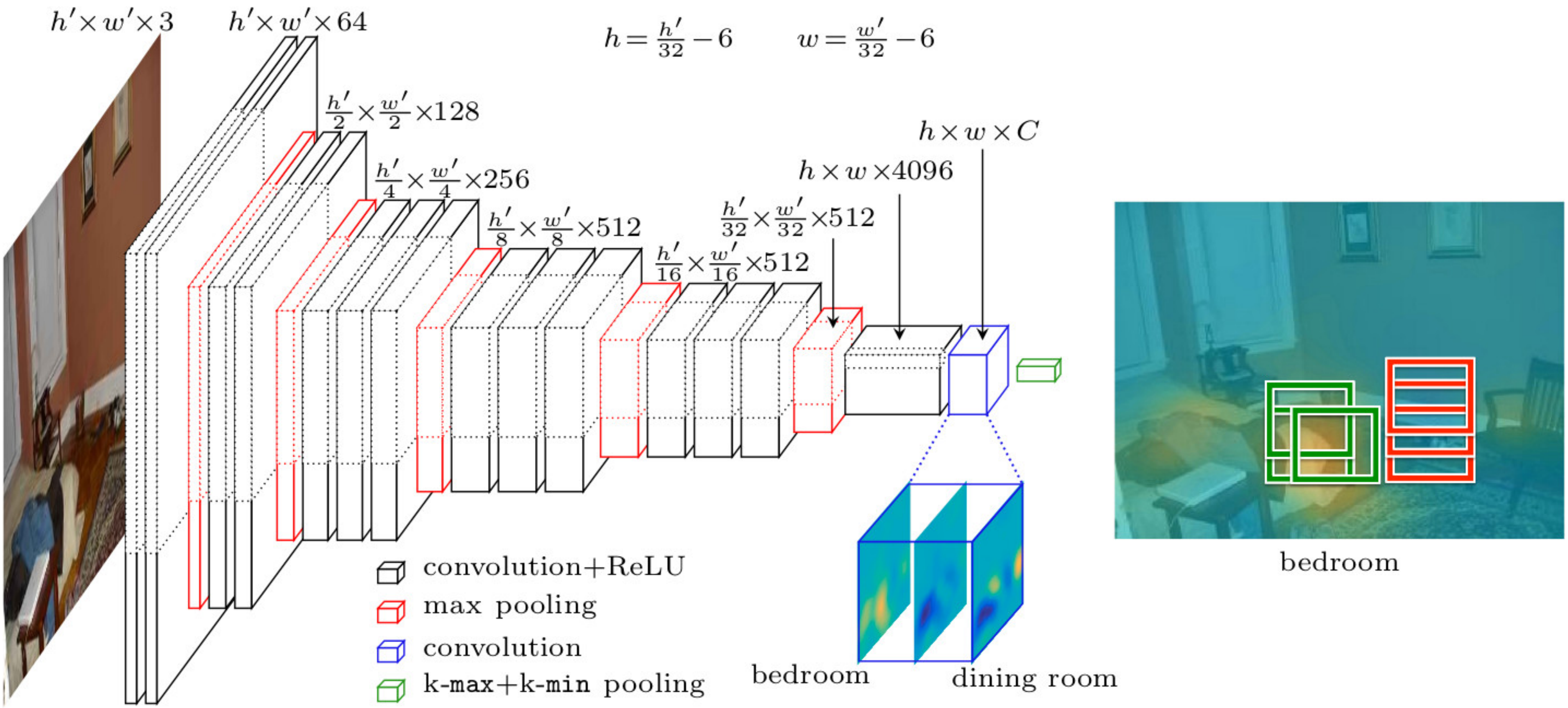
- max + min strategy
- Top instances: using several regions, more robust region selection [Vasconcelos, CVPR15]

$$y^c = s_{k^+}^{top}(z^c) + s_{k^-}^{low}(z^c)$$

$$s_{k^+}^{top}(z^c) = \frac{1}{k^+} \sum_{i,j} h_{ij}^c z_{ij}^c \quad \text{with} \quad \mathbf{h}^c = \underset{\mathbf{h}=[h_{ij}\in\{0,1\}]_{i,j}}{arg\,max} \sum_{i,j} h_{ij} z_{ij}^c \quad \text{s.t.} \quad \sum_{i,j} h_{ij} = k^+$$

$$s_{k^-}^{low}(z^c) = \frac{1}{k^-} \sum_{i,j} \bar{h}_{ij}^c z_{ij}^c \quad \text{with} \quad \bar{\mathbf{h}}^c = \underset{\mathbf{h}=[h_{ij}\in\{0,1\}]_{i,j}}{arg\,min} \sum_{i,j} h_{ij} z_{ij}^c \quad \text{s.t.} \quad \sum_{i,j} h_{ij} = k^-$$

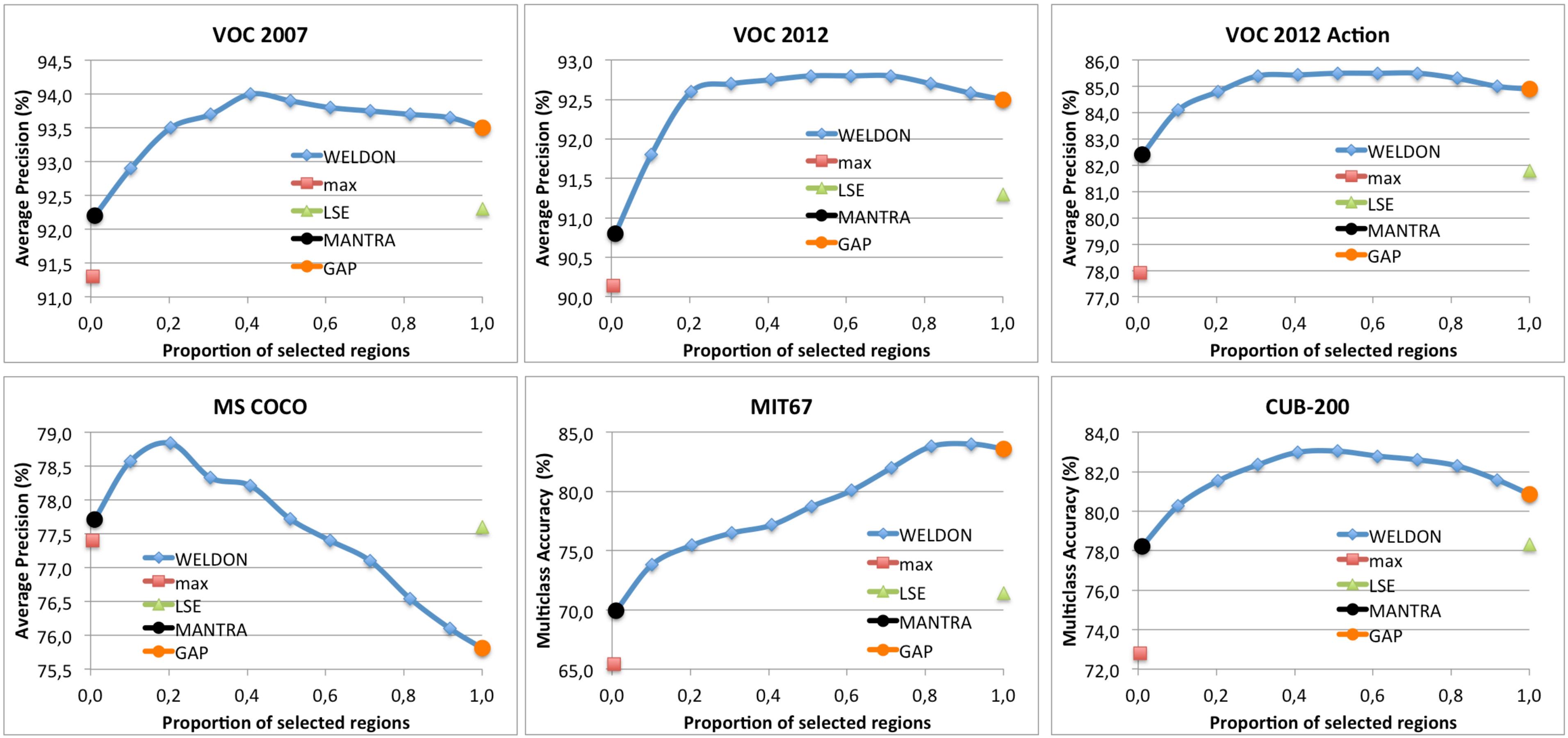
WELDON [Durand, CVPR16]

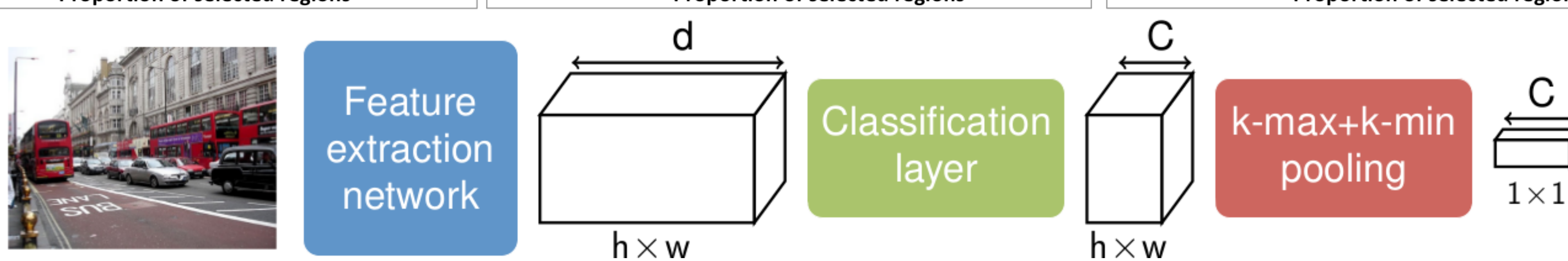


Outline

1. Deep net framework
2. Fully Convolutional Nets
3. Where is Pooling inside the architecture?
4. How to pool?
5. **Visualization and Experiments**

Pooling Analysis



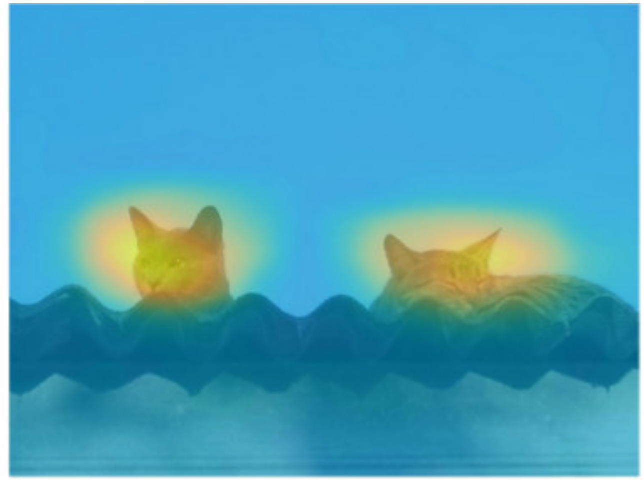


ImageNet (single model)

Model	Top-1 error	Top-5 error
VGG16 (144 crops)	24.4	7.2
GoogleNet (144 crops)	-	7.89
GoogleNet-GAP	35.0	13.2
VGG16-GAP	33.4	12.2
Inception-ResNet-v2 (12 crops)	**18.7**	**4.1**
ResNeXt-101 (1 crop)	19.1	4.4
ResNet-101 (1 crop)	22.44	6.21
ResNet-101 (10 crops)	21.08	5.35
ResNet-152 (10 crops)	20.69	5.21
ResNet-200 (10 crops)	20.15	4.93
FCN-WELDON	19.21	4.23

WELDON Visual results (VOC12)

bus



cat

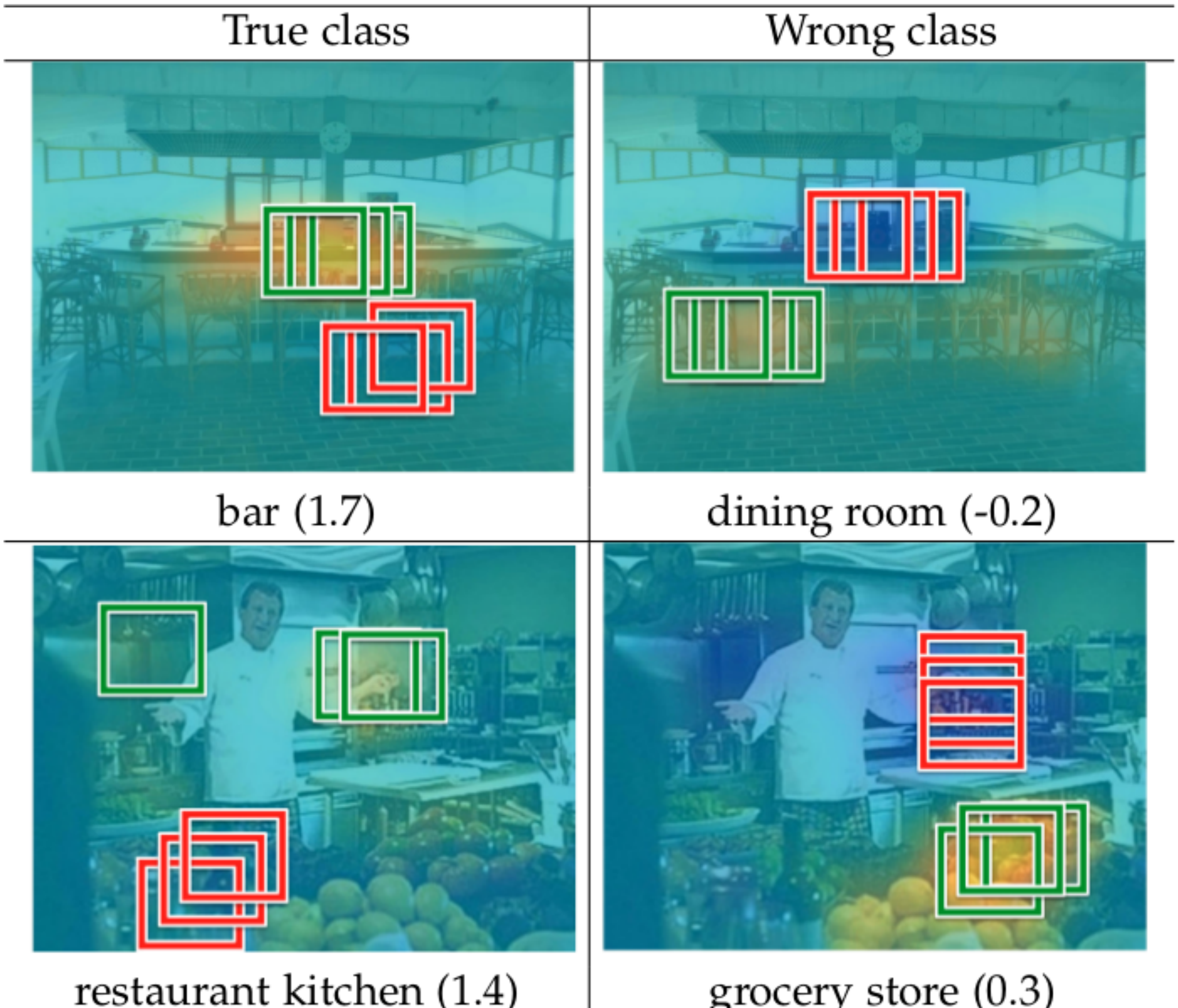
horse

aeroplane

bottle

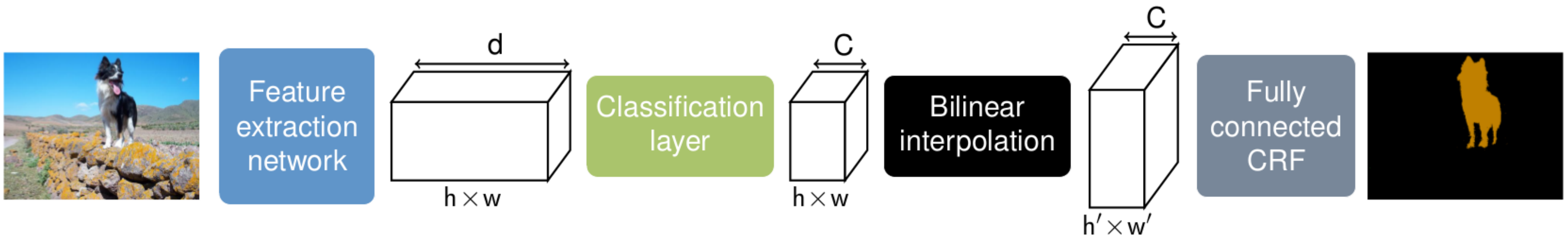
bicycle

Visual results (MIT67)



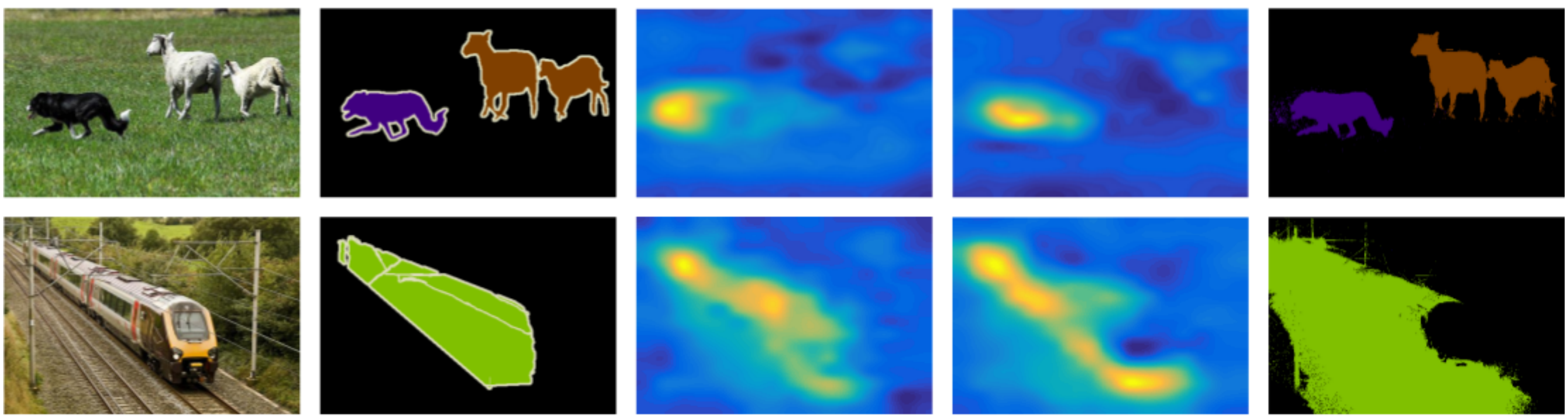
Extension: Segmentation

- WSL segmentation framework
  - Learning with image-level labels (presence/absence of the class)
  - Difficult task: no information about location and extend of objects
- Localized features in spatial maps
- Deep + fully connected CRFs



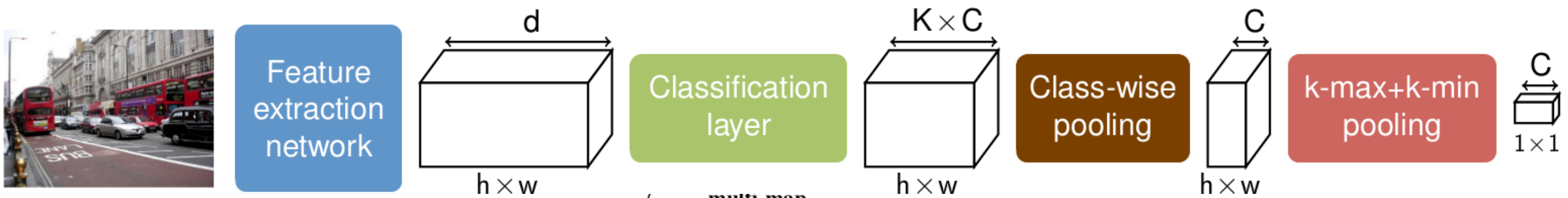
Extension: Segmentation

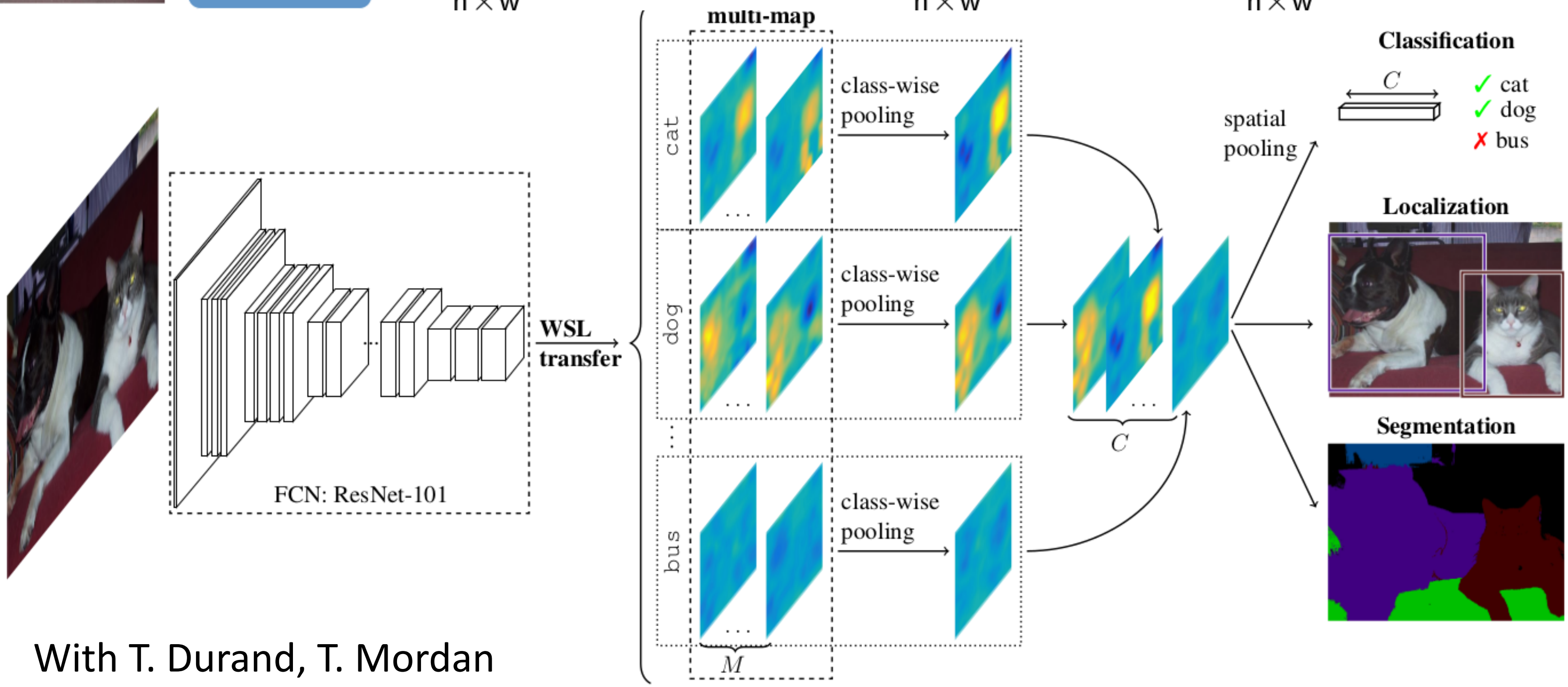
Method	Mean IoU
MIL-FCN [Pathak, ICLRW15]	24.9
MIL-Base+ILP+SP-sppxl [Pinheiro, CVPR15]	36.6
EM-Adapt +FC-CRF [Papandreou, ICCV15]	33.8
CCNN + FC-CRF [Pathak, ICCV15]	35.3
WILDCAT + FC-CRF	43.7



original image GT heatmap1 heatmap2 prediction

Extension: Wildcat (sub. CVPR17)





With T. Durand, T. Mordan

Share ideas of localized feature maps with R-FCN strategy of J. Dai, Yi Li, K. He, Jian Sun: R-FCN: Object Detection via Region-based Fully Convolutional Networks [NIPS 16]

Conclusion

Global Spatial Pooling: a major component in net design

Is there any learning trick behind this?

- [Lampert, ECCV16]: seed strategy better than GAP for segmentation!
- GAP: AP better than Max pooling strategy, from 1 to 400 feedback updates

Matthieu Cord

http://webia.lip6.fr/~cord

N. Thome, T. Durand, T. Robert, T. Mordan, X. Wang, M. Blot, M. Carvahlo, H. BenYounes, R. Cadene

WELDON project page from Thibaut Durand 's Web page (Github)

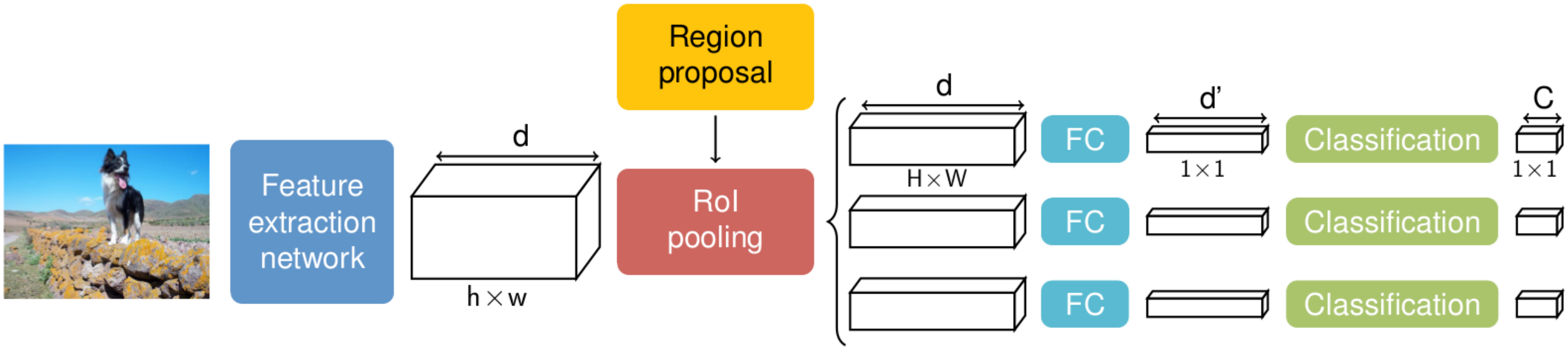
Our Deep Recipe Reco on your mobile: visiir.lip6.fr

Few Team's refs. on Deep learning for Visual Recognition

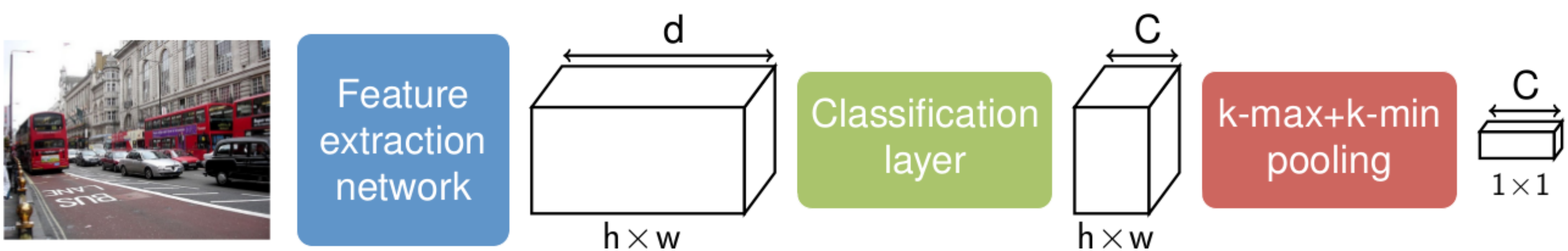
- WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks, T. Durand, N. Thome, M. Cord, CVPR 2016
- Deep Neural Networks Under Stress, M. Carvalho, M. Cord, S. Avila, N. Thome, E. Valle, ICIP 2016
- Max-Min convolutional neural networks for image classification, M. Blot, M. Cord, N. Thome, ICIP 2016
- MANTRA: Minimum Maximum Latent Structural SVM for Image Classification and Ranking, T Durand, N Thome, M Cord, ICCV 2015
- LR-CNN for fine-grained classification with varying resolution, M Chevalier+, ICIP 2015
- Top-Down Regularization of Deep Belief Networks, H. Goh, N. Thome, M. Cord, JH. Lim, NIPS 2013
- Sequentially generated instance-dependent image representations for classification, G Dulac-Arnold, L Denoyer, N Thome, M Cord, P Gallinari, ICLR 2014
- Learning Deep Hierarchical Visual Feature Coding, H. Goh+, IEEE Transactions on Neural Networks and Learning Systems 2014
- Unsupervised and supervised visual codes with Restricted Boltzmann Machines, H. Goh+, ECCV 2012
- Biasing Restricted Boltzmann Machines to Manipulate Latent Selectivity and Sparsity, H. Goh+, NIPS workshop 2010

Backup Slides

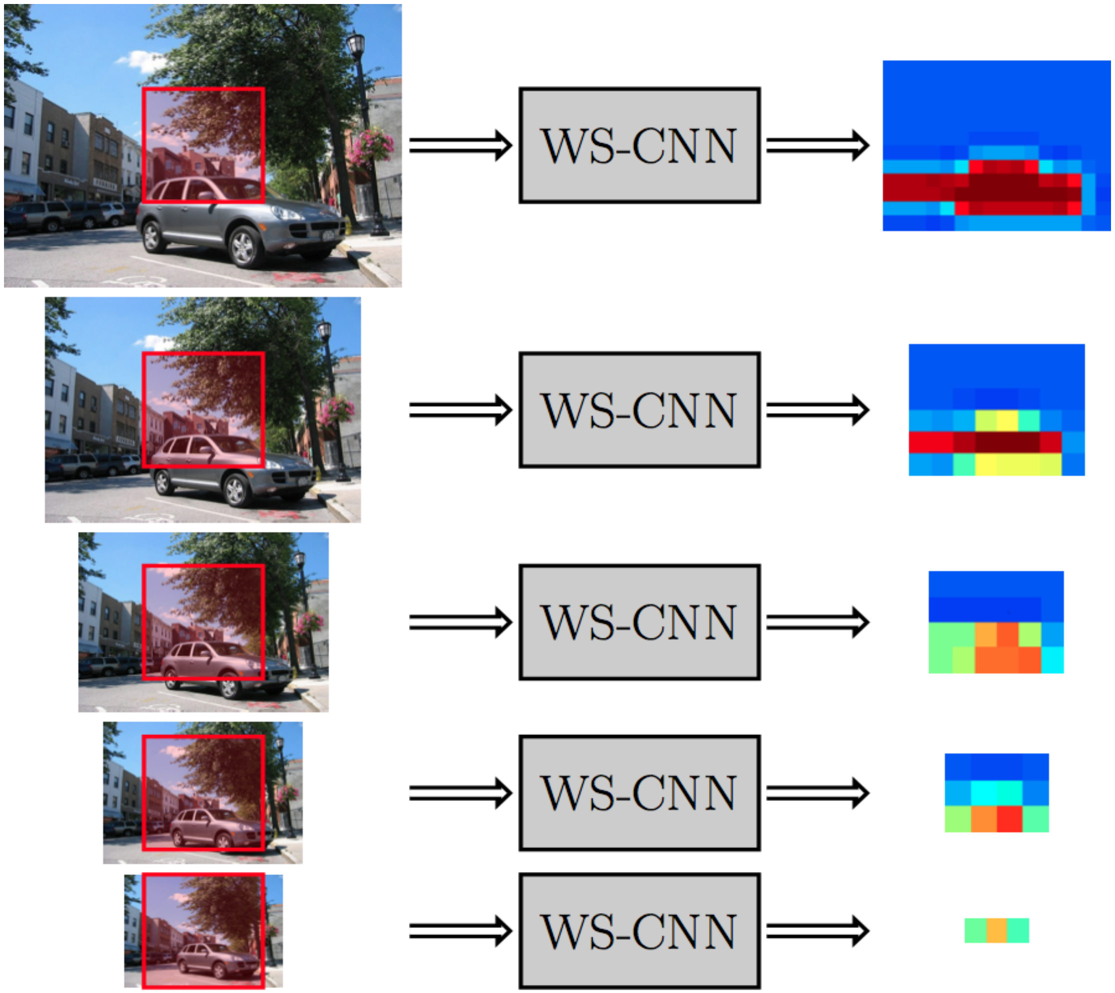
Fast(er) R-CNN



WELDON



- Multi-scale: 8 scales (combination with Object Bank strategy)

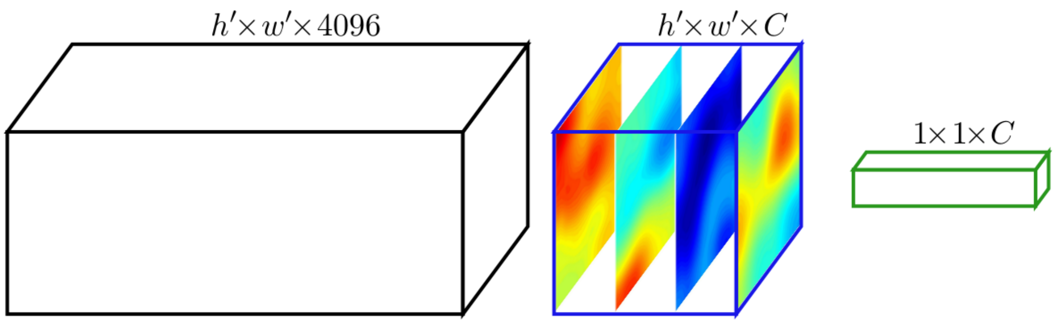


WELDON: learning

- Objective function for multi-class task and $k = 1$:

$$\min_{\mathbf{w}} \mathcal{R}(\mathbf{w}) + \frac{1}{N}\sum_{i=1}^{N} \ell(f_{\mathbf{w}}(\mathbf{x}_i), y_i^{gt})$$

$$f_{\mathbf{w}}(\mathbf{x}_i) = \arg\max_{y} \left(\max_{h} \mathbf{L}^{\mathbf{w}}_{\mathbf{conv}}(\mathbf{x}_i, y, h) + \min_{h'} \mathbf{L}^{\mathbf{w}}_{\mathbf{conv}}(\mathbf{x}_i, y, h') \right)$$



How to learn deep architecture ?

- Stochastic gradient descent training.
- Back-propagation of the **selecting windows** error.

WELDON: learning

Class is **present**

- **Increase** score of selecting windows.

Figure: Car map

WELDON: learning

Class is **absent**

- **Decrease** score of selecting windows.

Figure: Boat map

Conclusion: connections to others Latent Variables Models

- Hidden CRF (HCRF) [Quattoni, PAMI07]

$$\frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^{N} \log \sum_{(\mathbf{y},\mathbf{h})\in\mathcal{Y}\times\mathcal{H}} \exp\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}, \mathbf{h})\rangle - \log \sum_{\mathbf{h}\in\mathcal{H}} \exp\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h})\rangle$$

- Latent Structural SVM (LSSVM) [Yu, ICML09]

$$\frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^{N} \max_{(\mathbf{y},\mathbf{h})\in\mathcal{Y}\times\mathcal{H}} \{\Delta(\mathbf{y}_i,\mathbf{y}) + \langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y},\mathbf{h})\rangle\} - \max_{\mathbf{h}\in\mathcal{H}} \langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y}_i,\mathbf{h})\rangle$$

- Marginal Structural SVM (MSSVM) [Ping, ICML14]

$$\frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^{N} \max_{\mathbf{y}} \left\{\Delta(\mathbf{y}_i,\mathbf{y}) + \log\sum_{\mathbf{h}\in\mathcal{H}} \exp\langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y},\mathbf{h})\rangle\right\} - \log\sum_{\mathbf{h}\in\mathcal{H}} \exp\langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y}_i,\mathbf{h})\rangle$$

- WELDON

$$\frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^{N} \max_{\mathbf{y}} \left\{\Delta(\mathbf{y}_i,\mathbf{y}) + \sum_{\mathbf{h}\in\Omega\subseteq\mathcal{H}} \langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y},\mathbf{h})\rangle\right\} - \sum_{\mathbf{h}\in\Omega\subseteq\mathcal{H}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h})\rangle$$

MANTRA: model training

Learning formulation

- Loss function: $\ell_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) = \max_{\mathbf{y} \in \mathcal{Y}} [\Delta(\mathbf{y}_i, \mathbf{y}) + D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y})] - D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i)$
  - (Margin rescaling) upper bound of $\Delta(\mathbf{y}_i, \hat{\mathbf{y}})$, constraints:

$$\forall \mathbf{y} \neq \mathbf{y}_i, \qquad \underbrace{D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i)}_{\text{score for ground truth output}} \geq \underbrace{\Delta(\mathbf{y}_i, \mathbf{y})}_{\text{margin}} + \underbrace{D_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y})}_{\text{score for other output}}$$

- Non-convex optimization problem

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^{N} \ell_{\mathbf{w}}(\mathbf{x}_i, \mathbf{y}_i) \tag{3}$$

- Solver: non convex one slack cutting plane [Do, JMLR12]