Harmony Potentials:

Fusing Global and Local Scale for Semantic Image Segmentation

J. M. Gonfaus
X. Boix
F. S. Khan
J. van de Weijer
A. Bagdanov
M. Pedersoli
J. Serrat
X. Roca
J. Gonzàlez

Universitat Autònoma de Barcelona
Centre de Visió per Computador
CONSOLIDER INGENIO 2010

Motivation (I)

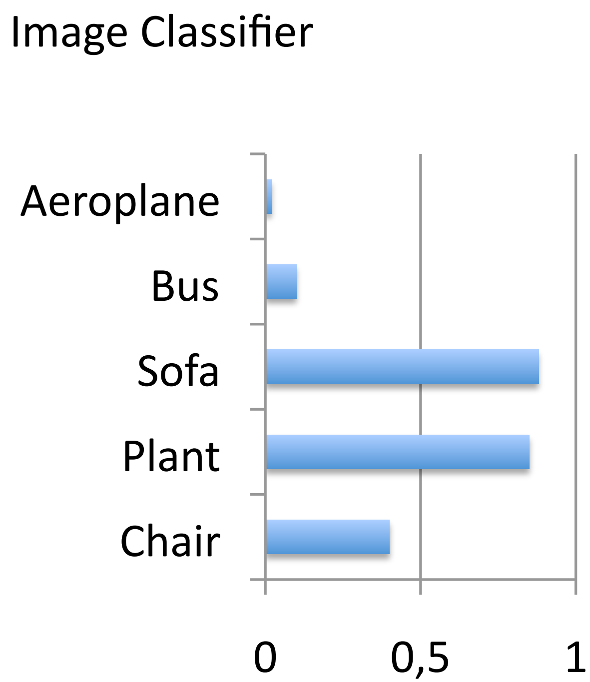
- Why combine global and local scale?

Motivation (I)

- Why combine global and local scale?

Motivation (I)

- Classification is often impossible based on local appearance only.

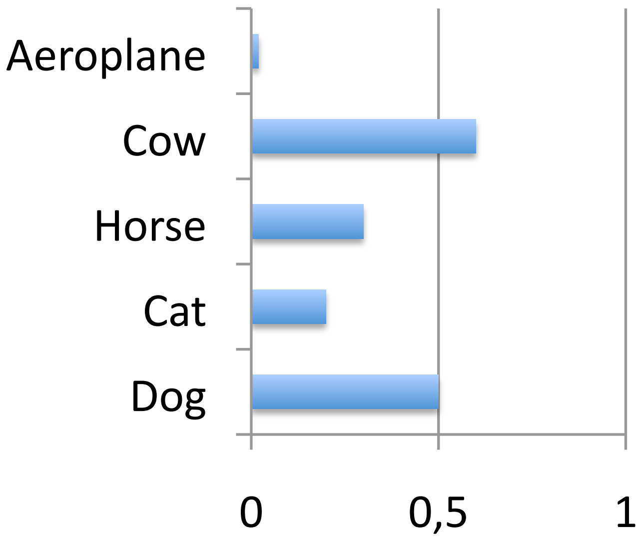


Context is a powerful and distinctive cue

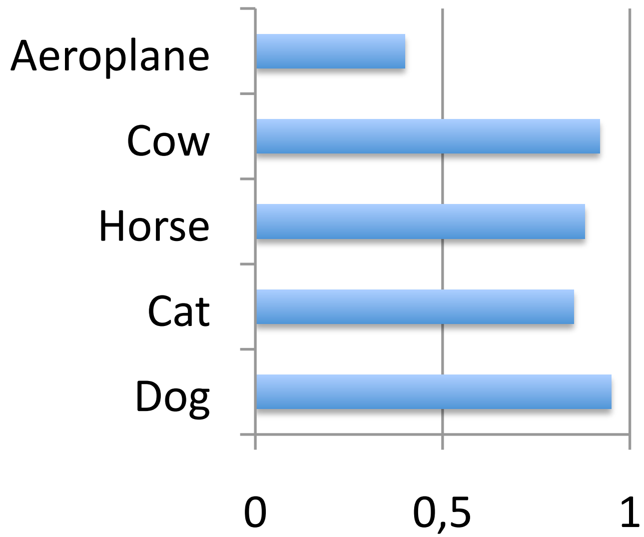
Motivation (II)

- How can we improve local classifiers?

Is this object X or some other object



Is this the foreground or the background of…



Inaccurate segmentation

Good class discrimination

Why not combine them?

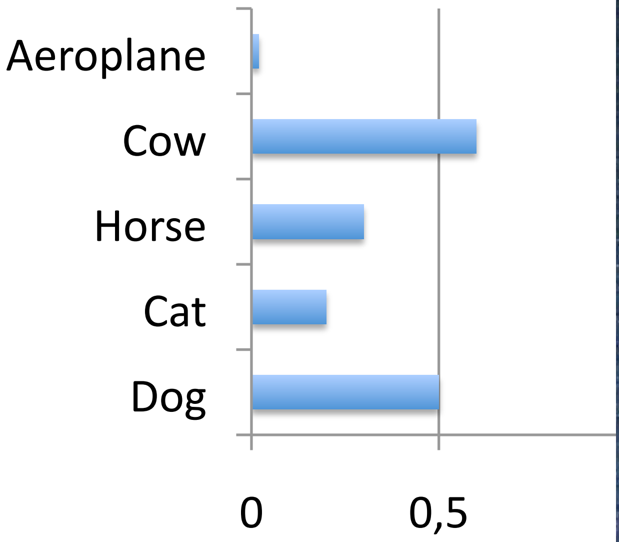
Good figure segmentation

Bad class discrimination

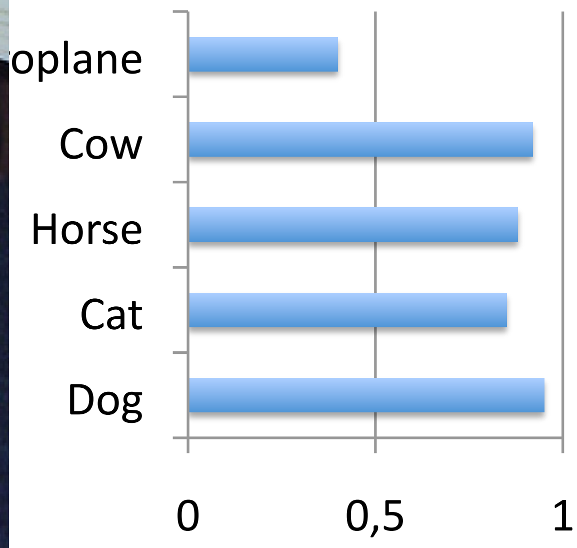
Motivation (II)

- How can we improve local classifiers?

Is this object X or
some other object



Is this the foreground
or the background of…



Inaccurate segmentation

Good class discrimination

Why not combine them?

Good figure segmentation

Bad class discrimination

Motivation (II)

- How can we improve local classifiers?
  - More information sources
    - Mid-level information through object detectors

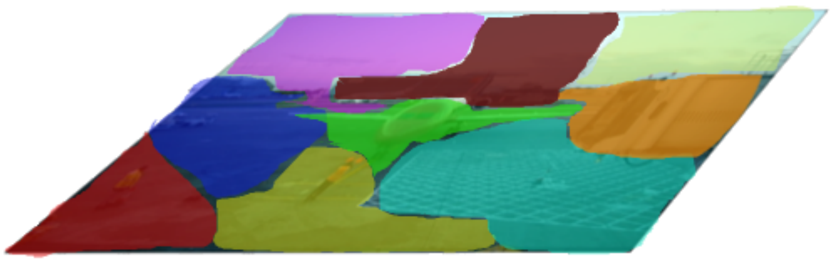
Outline

- Overview of our method
- How to fuse local and global scale
  - Harmony Potentials*
  - CVC_Harmony submission (35.4% on test)
- Improving local classifiers
  - CVC_Harmony+Det submission (40.1% on test)
- Results
- Conclusions

*J.M. Gonfaus, X. Boix, J. Van de Weijer, A. D. Bagdanov, J. Serrat, J. Gonzàlez
"Harmony Potentials for Joint Classification and Segmentation", in CVPR 2010

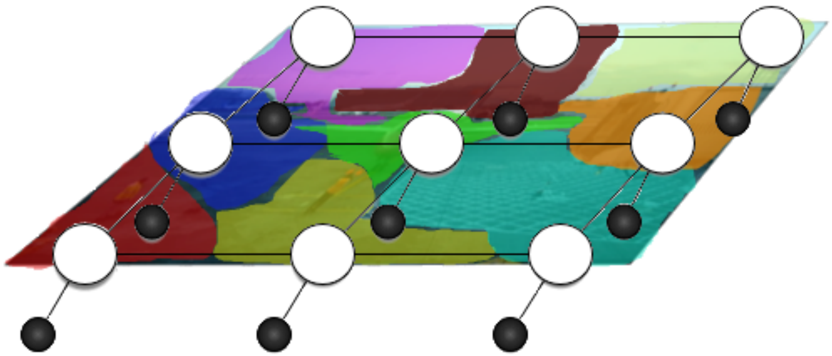
Overview of our method

Overview of our method



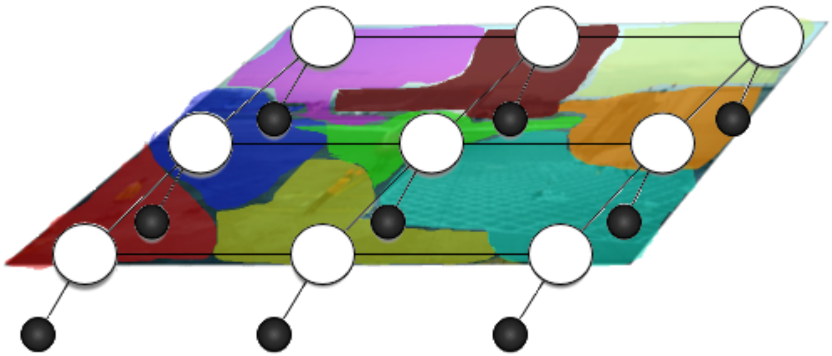
- Unsupervised segmentation.
  - Around 500 superpixels/image

Overview of our method



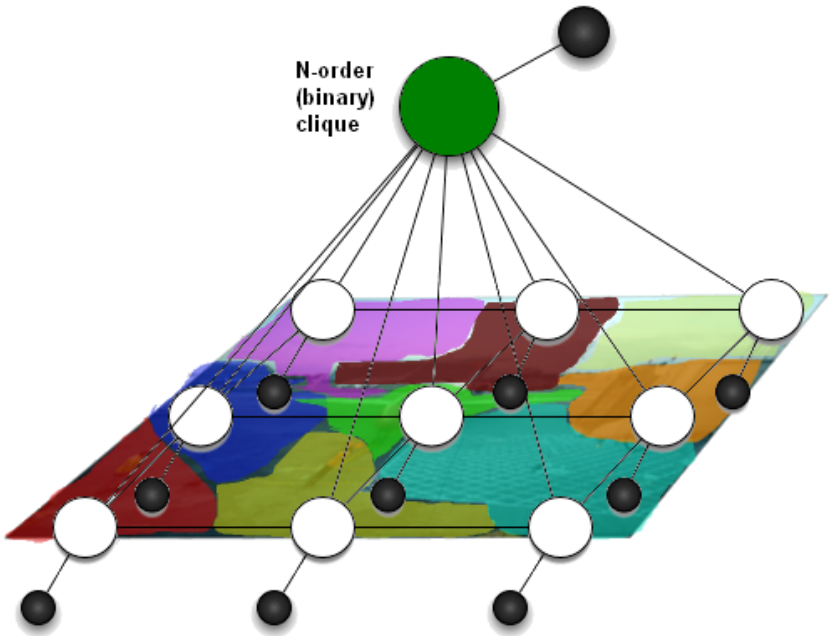
- Unsupervised segmentation.
- Superpixel nodes
  - Unary potential (CVC_Harmony)
    - BoW inside AND neighborhood
  - Smoothness potential
    - Pairwise Potts potential
  - BoW
    - SIFT, RGB Histogram, SSIM
    - Multiscale: 12, 24, 36, 48 square patches
    - Step size 50% of the patch
    - Quantized to 1000, 400, 300 words
    - Learned on SVM with 8000 samples + retraining

Overview of our method



- Unsupervised segmentation.
- Superpixel nodes
  - Unary potential (CVC_Harmony+det)
    - BoW inside AND neighborhood
    - Detection scores
    - Location prior
  - Smoothness potential
    - Pairwise Potts potential
  - BoW
    - SIFT, RGB Histogram, SSIM
    - Multiscale: 12, 24, 36, 48 square patches
    - Step size 50% of the patch
    - Quantized to 1000, 400, 300 words
    - Learned on SVM with 8000 samples + retraining

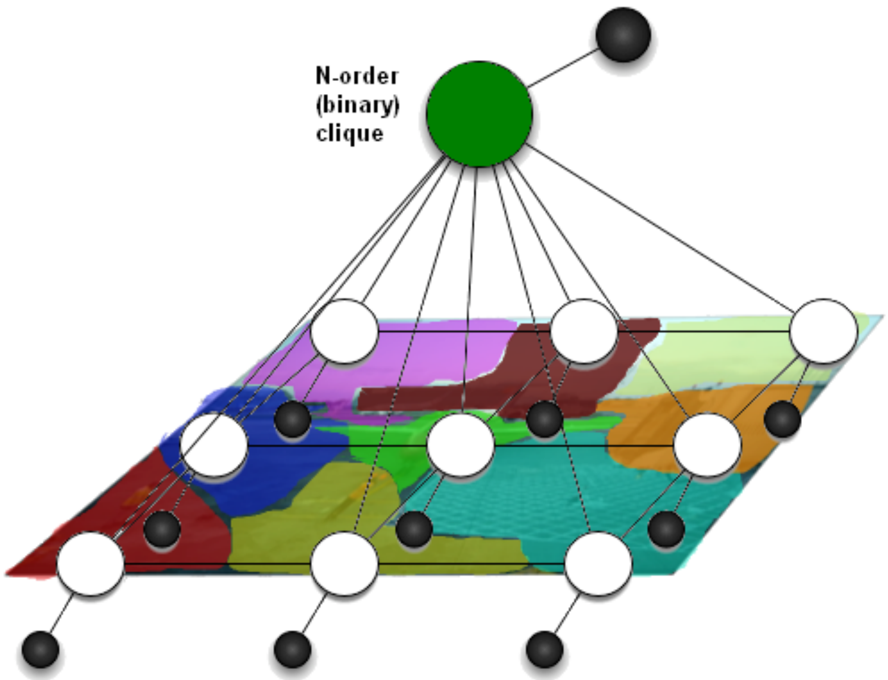
Overview of our method



- Unsupervised segmentation.
- Superpixel nodes
- Global Node
  - Unary potential:
    - Global classifier method
    - CVC_flat submission:

      mAP: 61% for classification task
  - Consistency potential
    - From global node to each sp
    - Harmony Potential

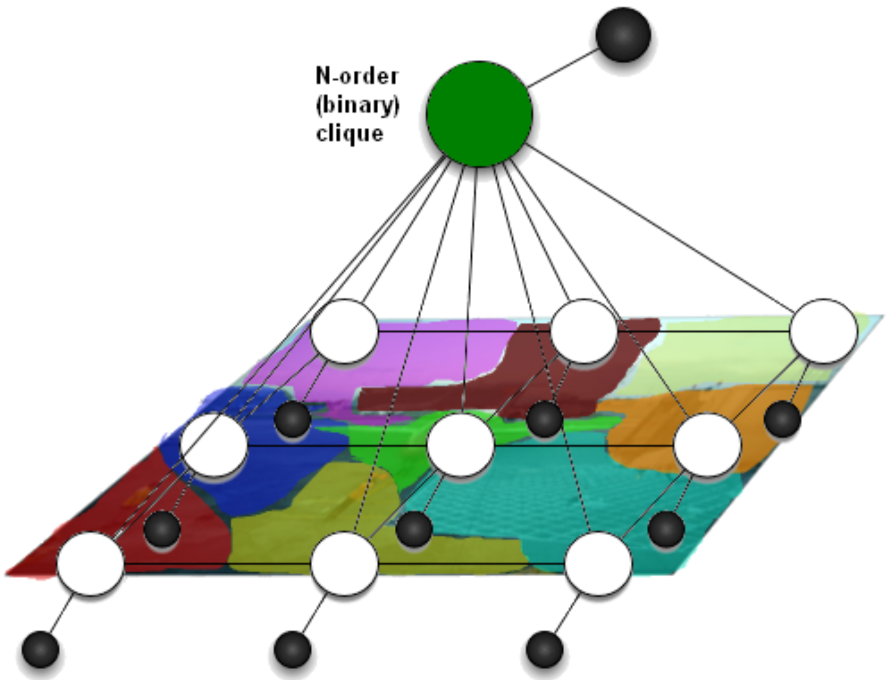
Model



$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \phi(x_i) + \sum_{(i,j) \in \mathcal{E}_L} \psi_L(x_i, x_j) + \sum_{(i,g) \in \mathcal{E}_G} \psi_G(x_i, x_g).$$

Unary Potential Smoothness Potential Consistency Potential

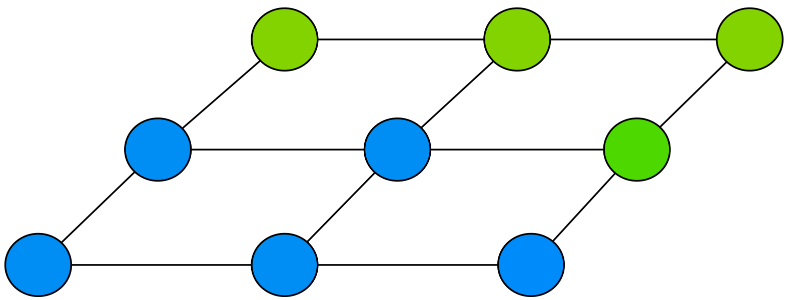
Model



$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \phi(x_i) + \sum_{(i,j) \in \mathcal{E}_L} \psi_L(x_i, x_j) + \boxed{\sum_{(i,g) \in \mathcal{E}_G} \psi_G(x_i, x_g)}.$$

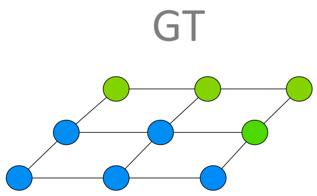
Consistency Potential

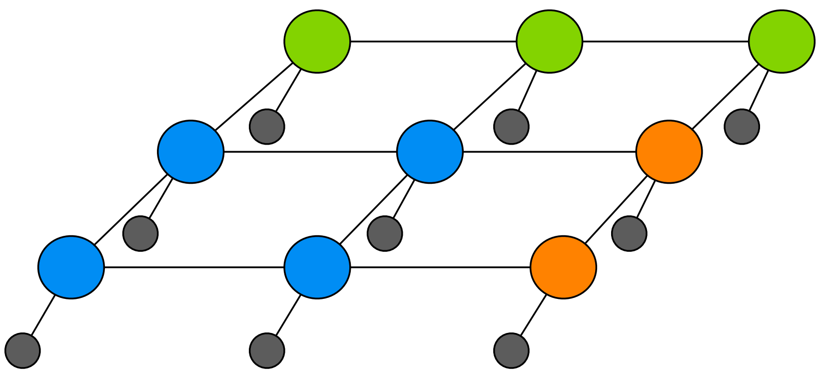
Consistency potential



- Ground-Truth
- Unary Potentials
- Potts-based Potentials
- Robust P^N Potentials
- Harmony Potentials

Consistency potential

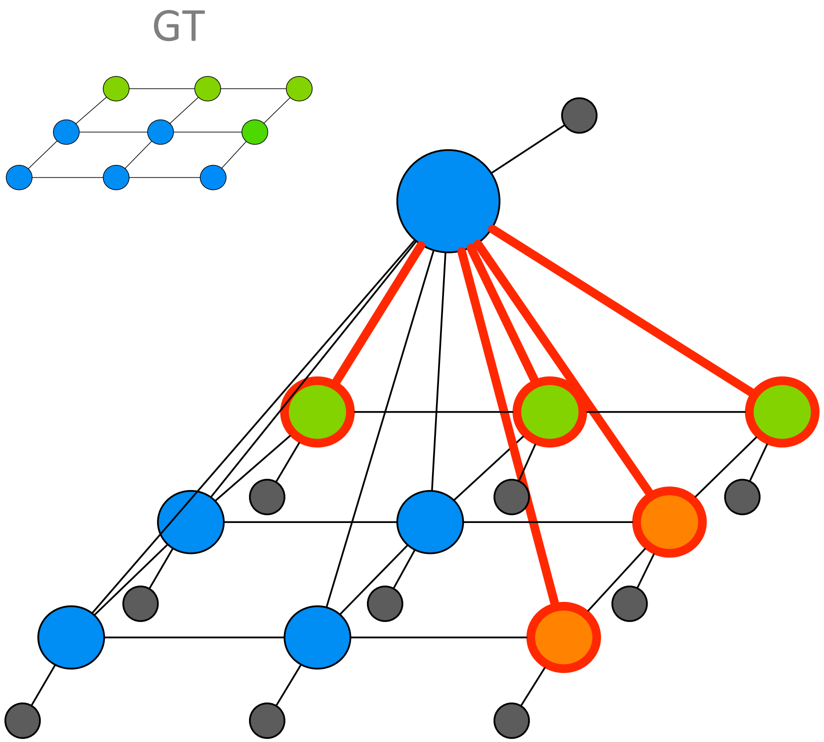




- Ground-Truth
- Unary Potentials
- Potts-based Potentials
- Robust P^N Potentials
- Harmony Potentials

$$\psi_G(x_i, x_g) = 0.$$

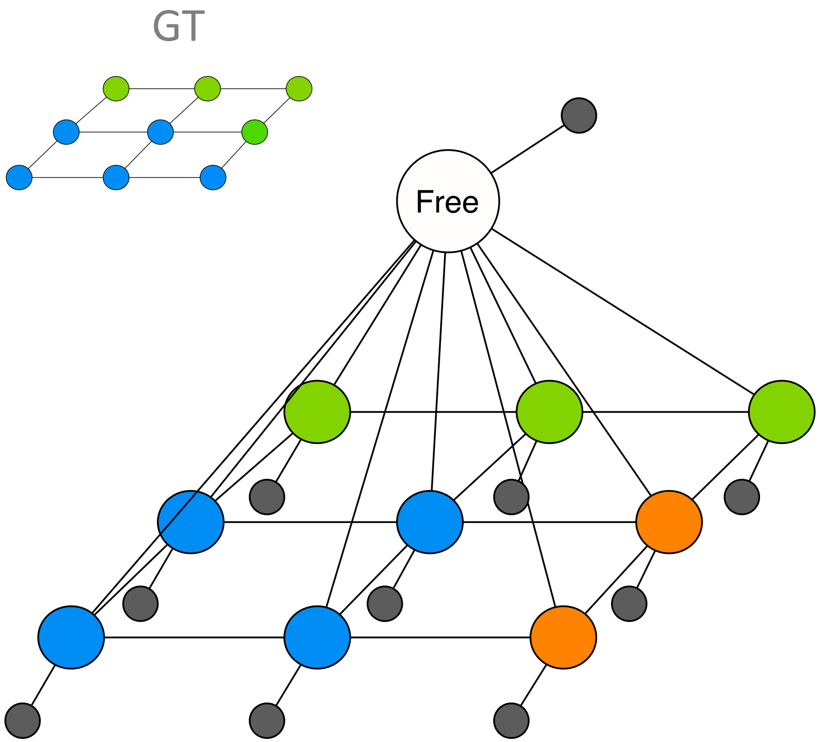
Consistency potential



- Ground-Truth
- Unary Potentials
- Potts-based Potentials
- Robust P^N Potentials
- Harmony Potentials

$$\psi_G(x_i, x_g) = \gamma_i^l \mathrm{T}[x_i \neq x_g]$$

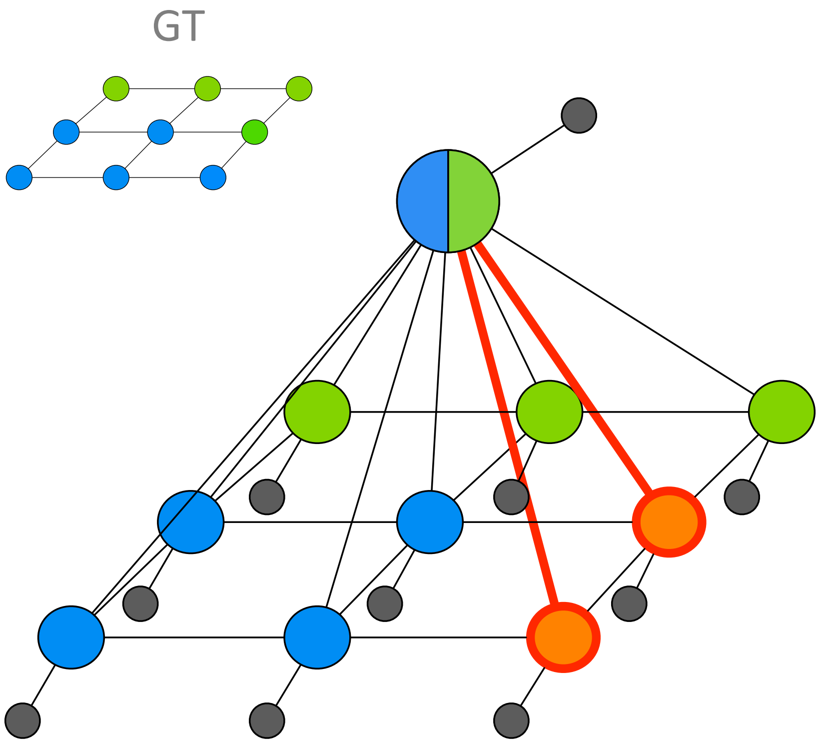
Consistency potential



- Ground-Truth
- Unary Potentials
- Potts-based Potentials
- Robust P^N Potentials
- Harmony Potentials

$$\psi_G(x_i, x_g) = \begin{cases} 0 & \text{if } x_g = l_F \text{ or } x_g = x_i \\ \gamma_i^l & \text{otherwise, where } l = x_i \end{cases}$$

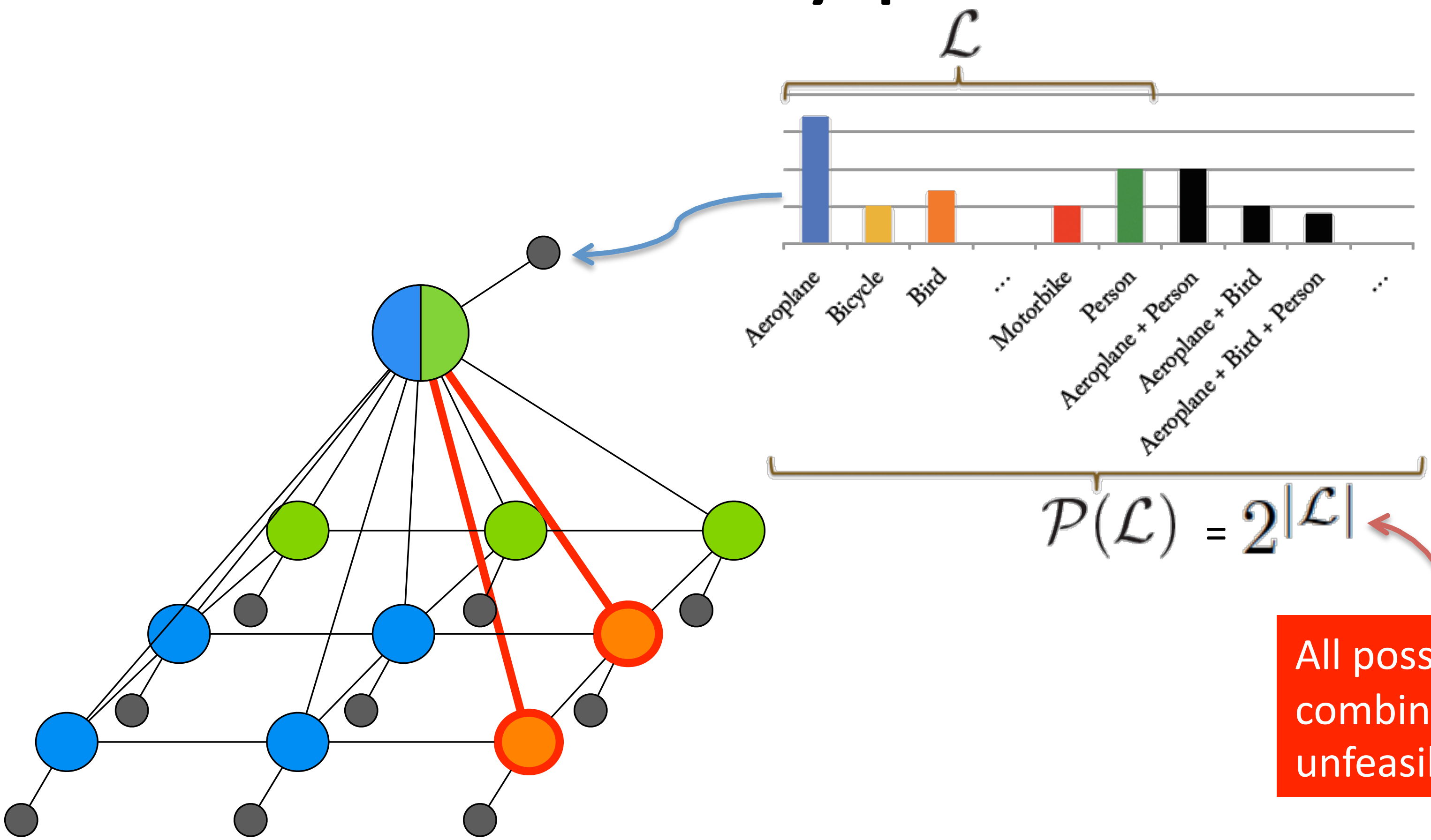
Consistency potential



- Ground-Truth
- Unary Potentials
- Potts-based Potentials
- Robust P^N Potentials
- Harmony Potentials

$$\psi_G(x_i, x_g) = \gamma_i^l \mathrm{T}[x_i \notin x_g]$$

Consistency potential



$$\mathcal{P}(\mathcal{L}) = 2^{|\mathcal{L}|}$$

All possible label combinations is unfeasible

Consistency potential

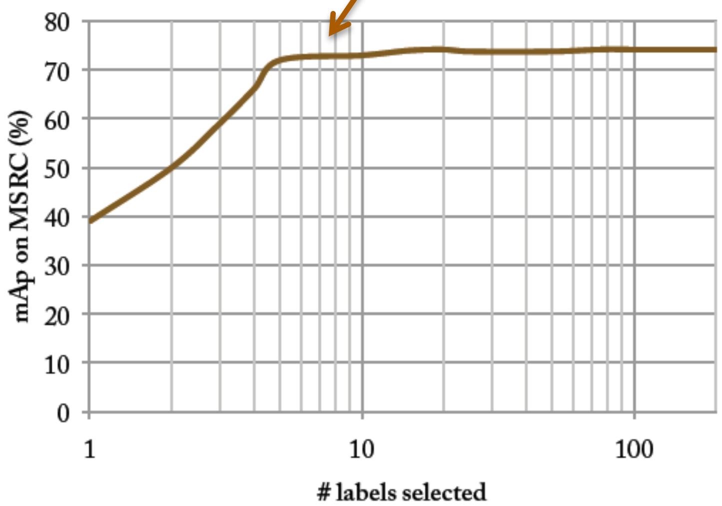
- Ranked subsampling of $\mathcal{P}(\mathcal{L})$

$$P(\ell \subseteq x_g^*|\mathbf{O}) \propto \boxed{P(\ell \subseteq x_g^*)}\boxed{P(\mathbf{O}|\ell \subseteq x_g^*)}$$

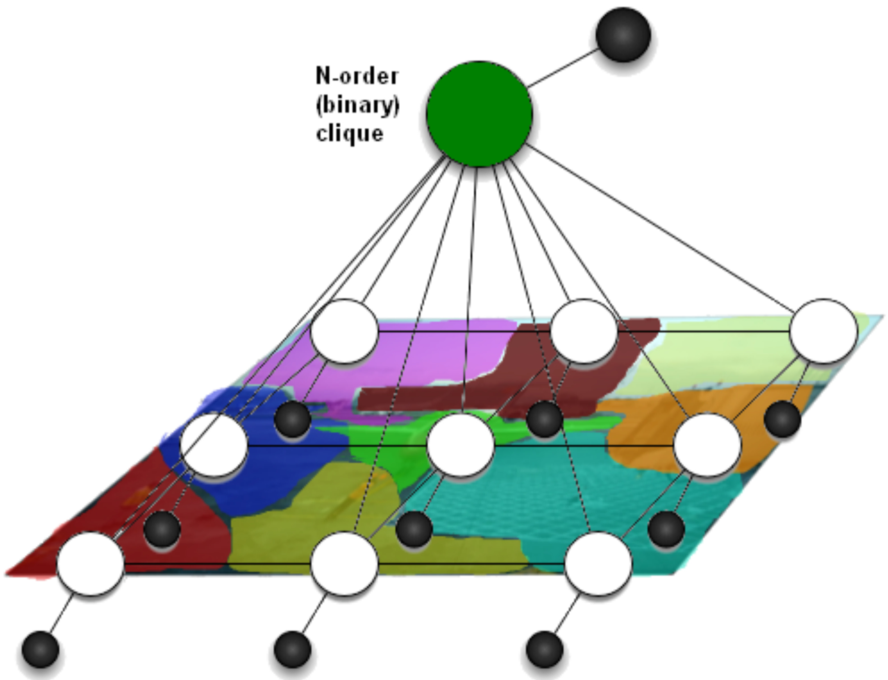
Prior
From the training data we extract the co-occurrence statistics of labels

Likelihood
Image classification scores each combination

Few best combinations are required to saturate the performance



Model



$$E(\mathbf{x}) = \boxed{\sum_{i \in \mathcal{V}} \phi(x_i)} + \sum_{(i,j) \in \mathcal{E}_L} \psi_L(x_i, x_j) + \sum_{(i,g) \in \mathcal{E}_G} \psi_G(x_i, x_g).$$

Unary Potential

Unary potential

- Local classifiers are weak classifiers
  - Too ambiguous because little information is used
- Combining multiple classifiers makes our local unary potential stronger.
- Features:
  - foreground/background
  - class versus others
  - object detections
  - spatial location prior

$F_{fg\text{-}bg}$: Fore-Background

- Easy to identify whether the superpixel belongs to the object class or to its common background

$F_{fg\text{-}bg}$: Fore-Background

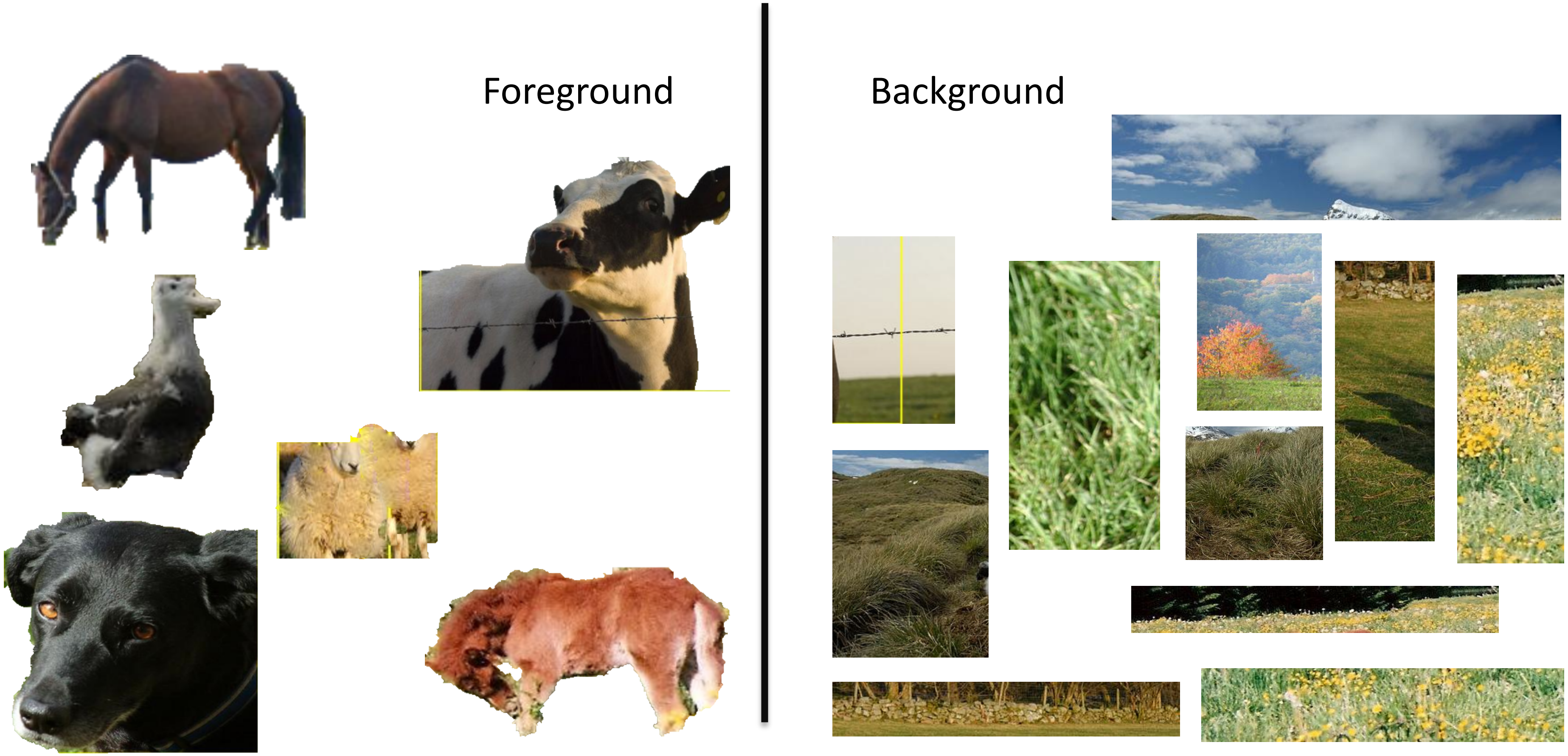
- Easy to identify whether the superpixel belongs to the object class or to its common background

$F_{fg\text{-}bg}$: Fore-Background

- Easy to identify whether the superpixel belongs to the object class or to its common background

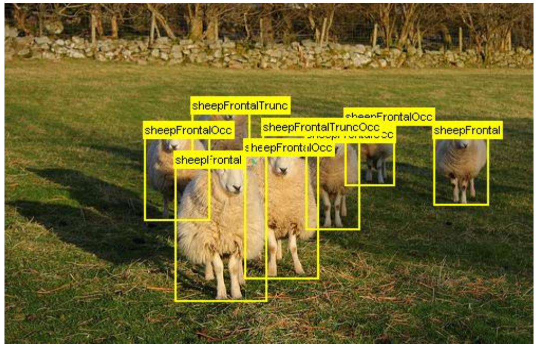
F_{class}: Class vs. other classes

- Learning how different an object is from its common background becomes difficult for certain class combinations



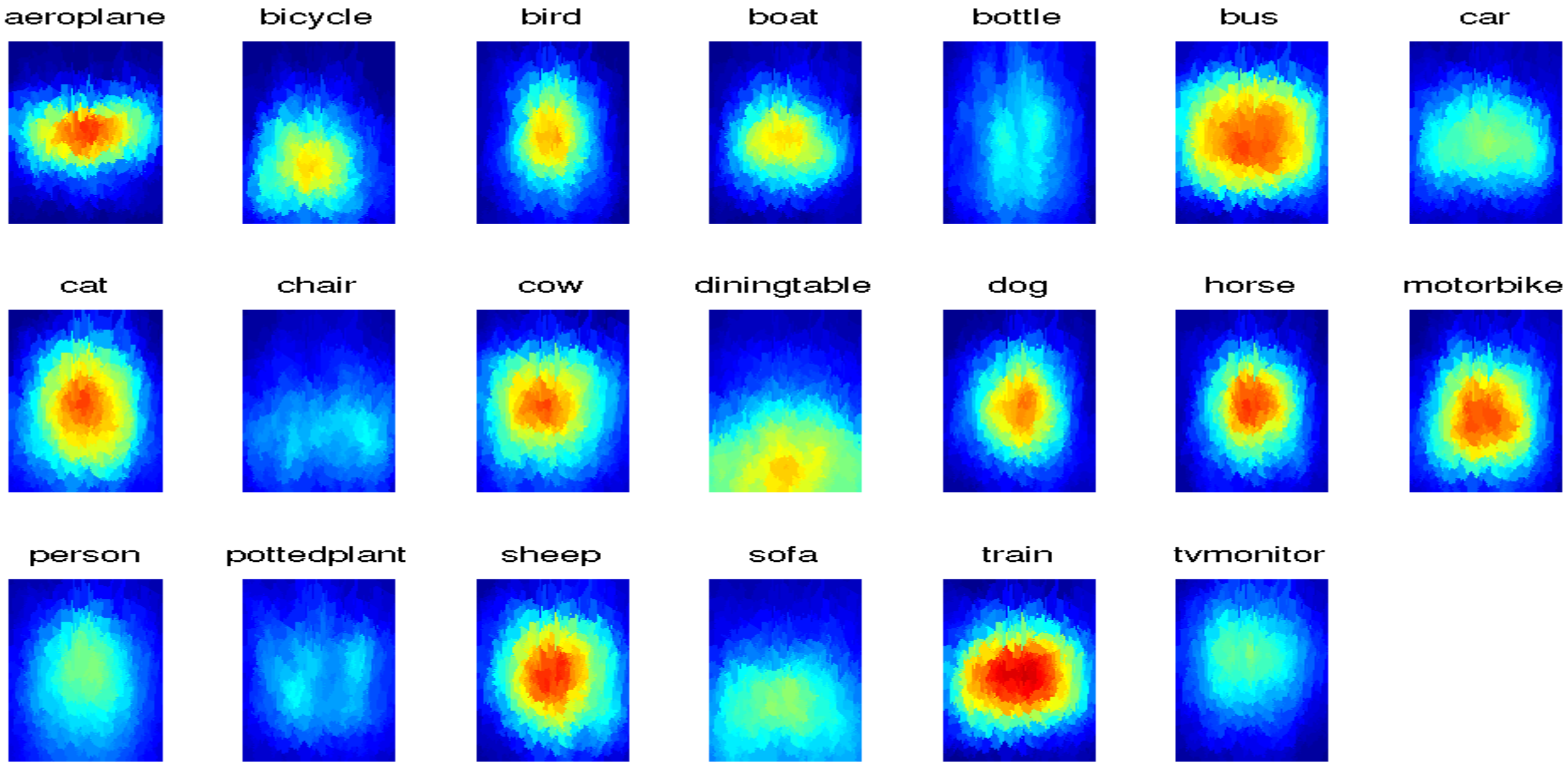
F_{class}: Class vs. other classes

- Learning how different an object is from its common background becomes difficult for certain class combinations



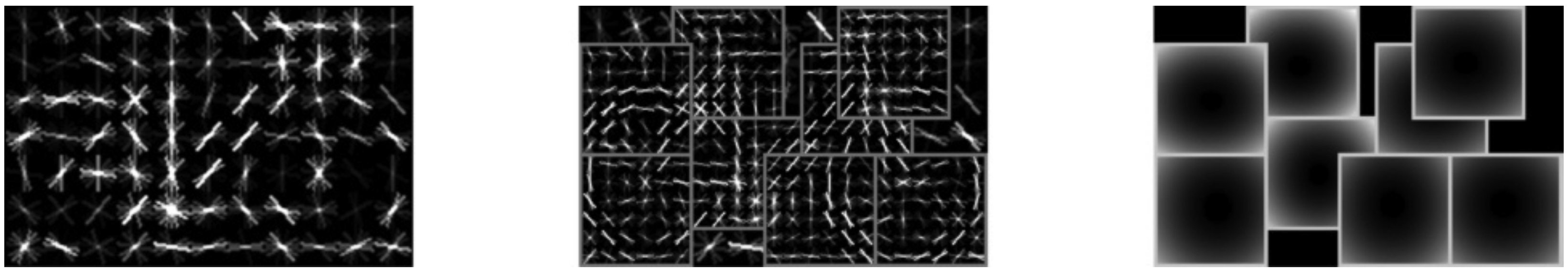
$F_{position}$: Location prior

- Objects tend to appear in class-specific, particular locations (and not at the borders)



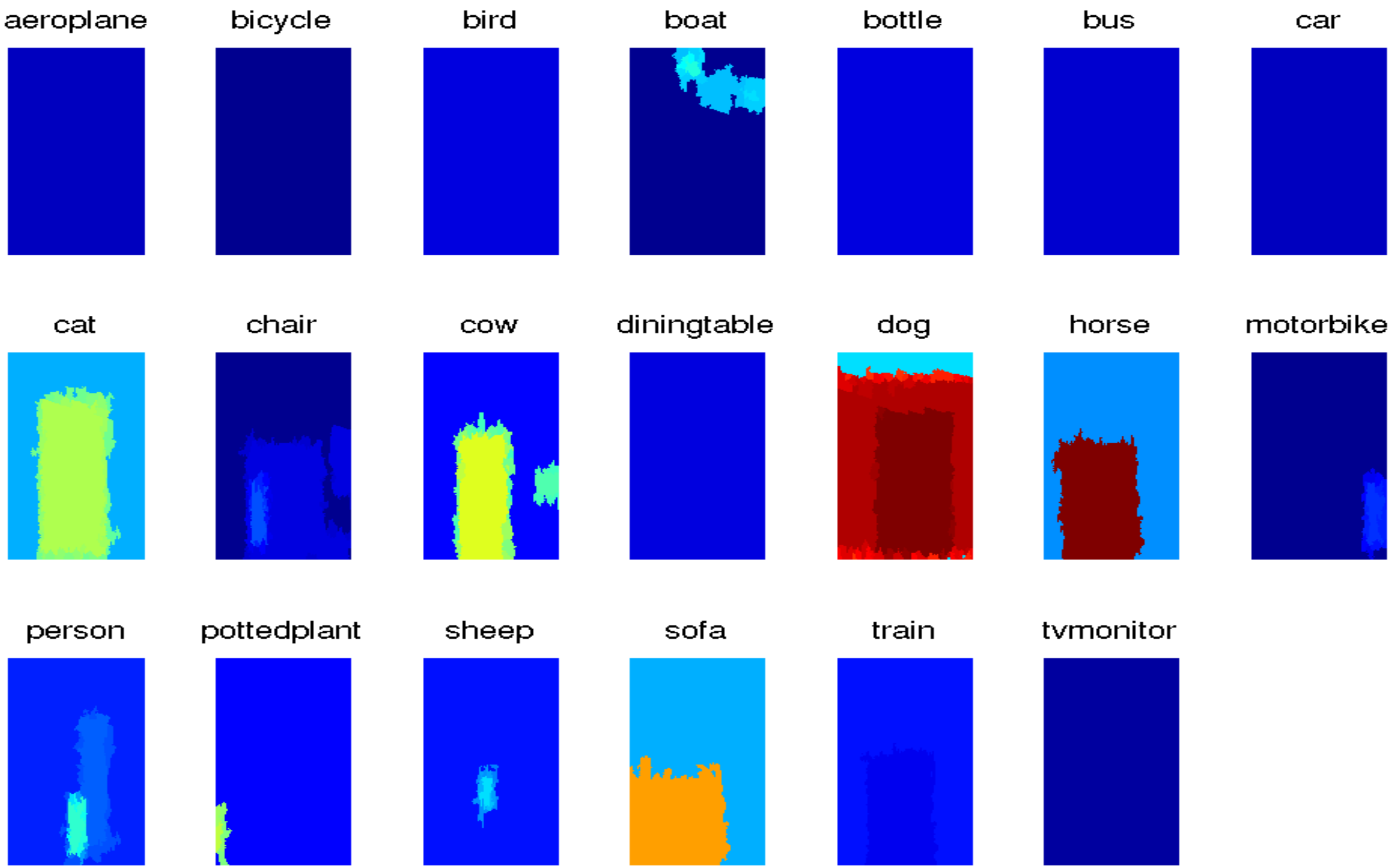
F_{det}: Object detector* scores

- Mid-level information is added by considering object detections [Felzenszwalb et al. 2010].
- Average over superpixel area with maximum detection score at each pixel.
- Scores = [-1 , ∞)
- Class specific "No detection" score is learned.
- Keeps the CRF and the model simple.

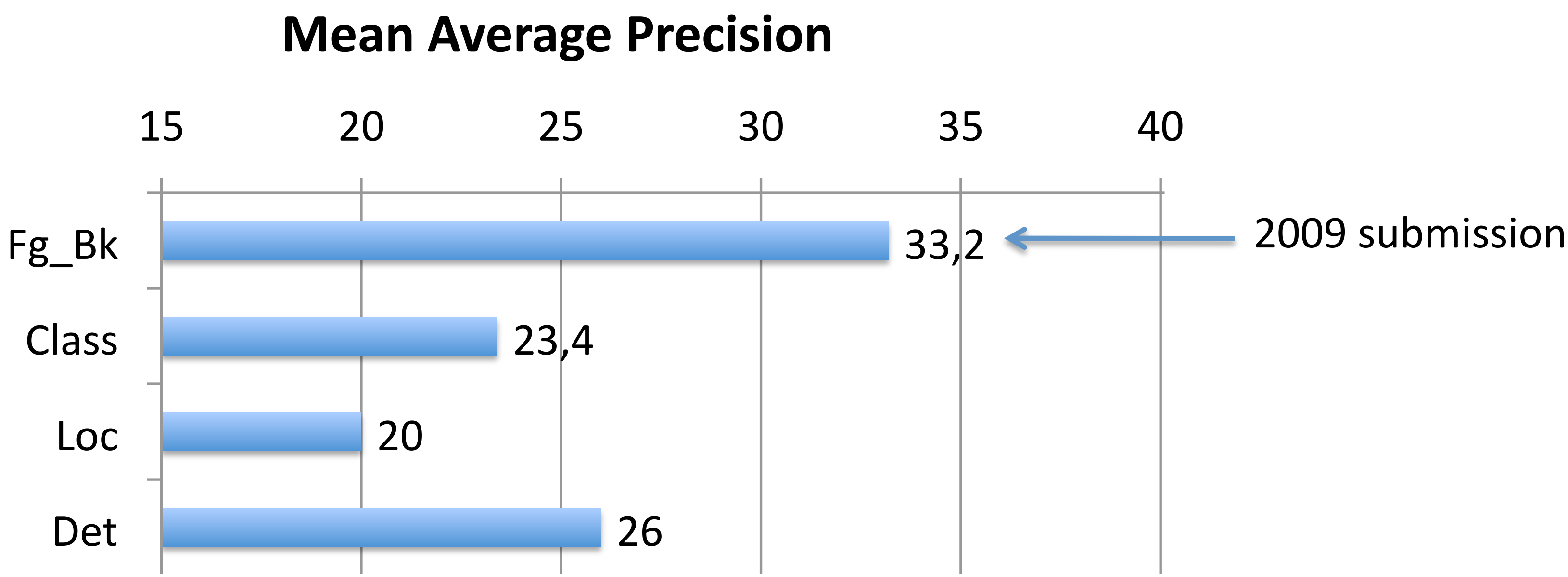


*Felzenszwalb, Girshick, McAllester, Ramanan, "Object Detection with Discriminately Trained Part based models", PAMI 2010

F_{det}: Object detector* scores



Results on validation set 2010



Combination of features

- Naïve Bayes approach
- Specific sigmoid per class and per classifier

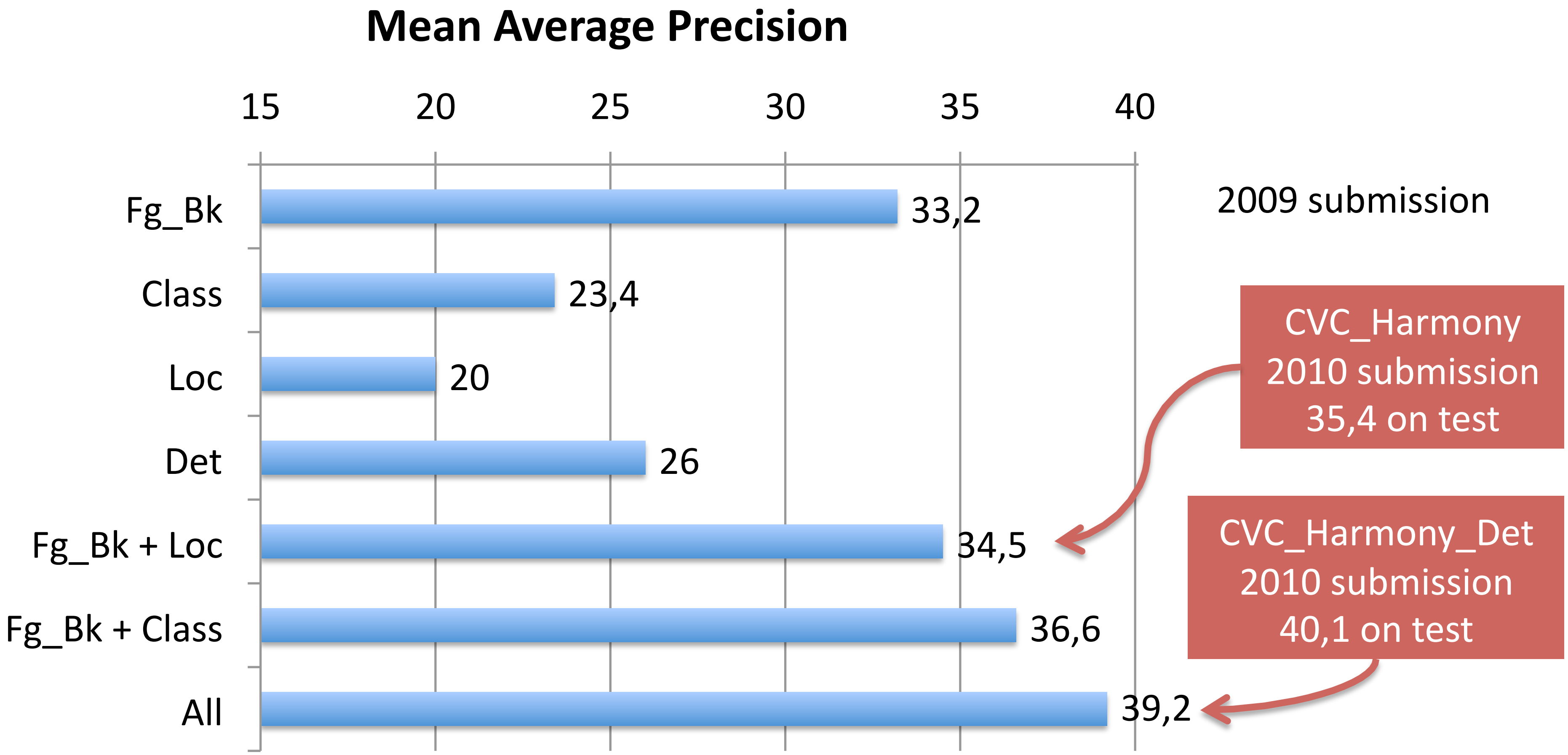
$$\phi(x_i) = \prod_{f \in F} \frac{1}{1 + \exp(-a^f x_i^f + b^f)}$$

- Total number of parameters to be learned:

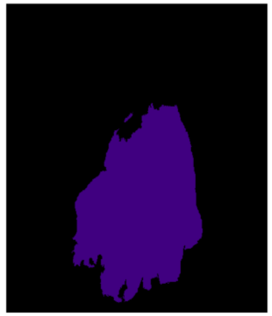
$$\underbrace{2\text{x}20\text{x}4}_{\text{feature sigmoids}} + \underbrace{20}_{\text{no\_detection score}} + \underbrace{4}_{\text{CRF weights}} + \underbrace{1}_{\text{background probability}} = 185 \text{ parameters}$$

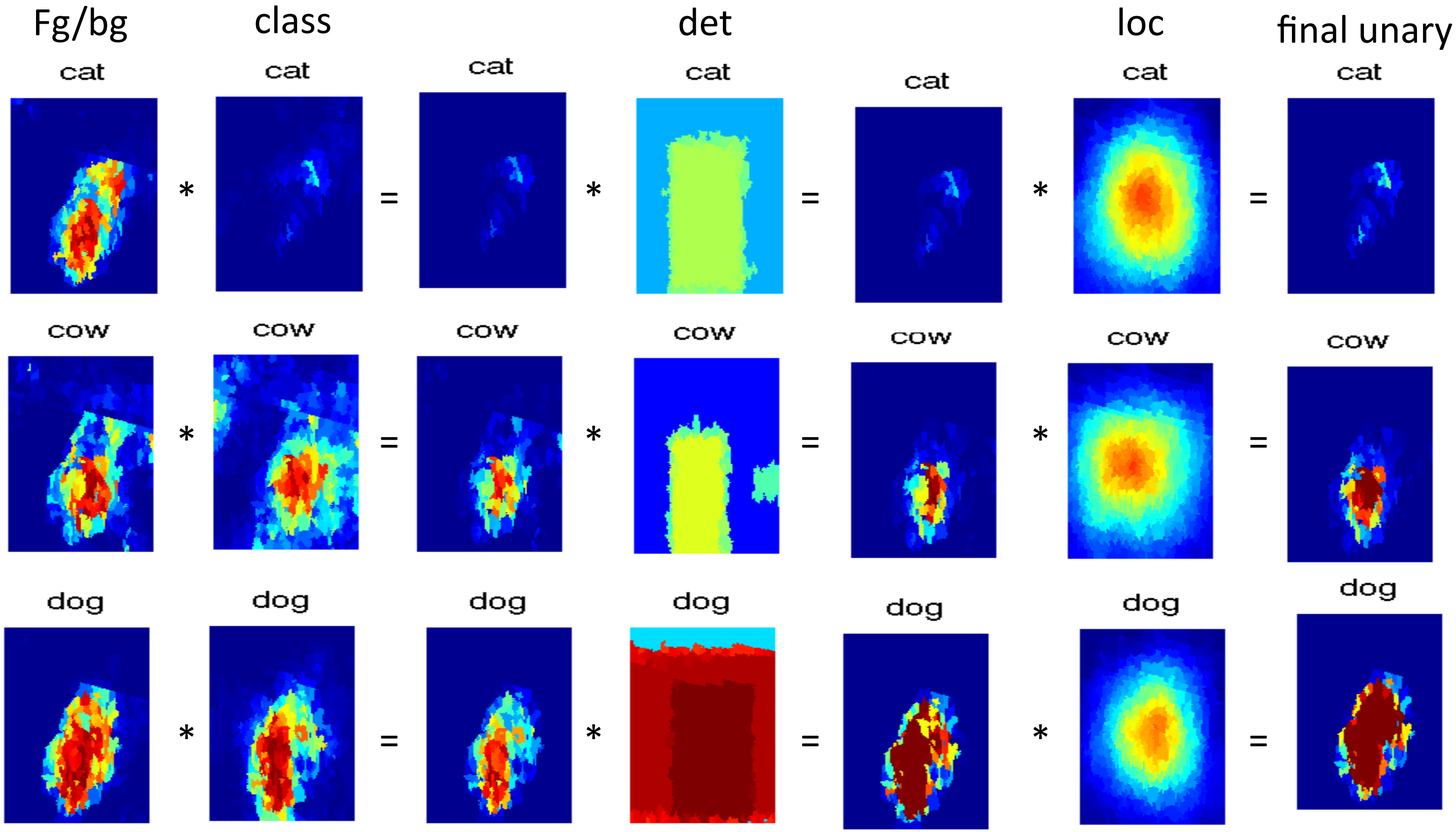
- All parameters are jointly optimized by stochastic steepest ascent

Results on validation set 2010

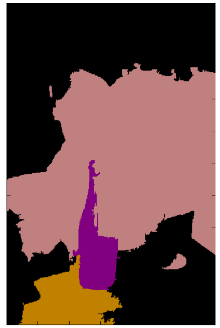


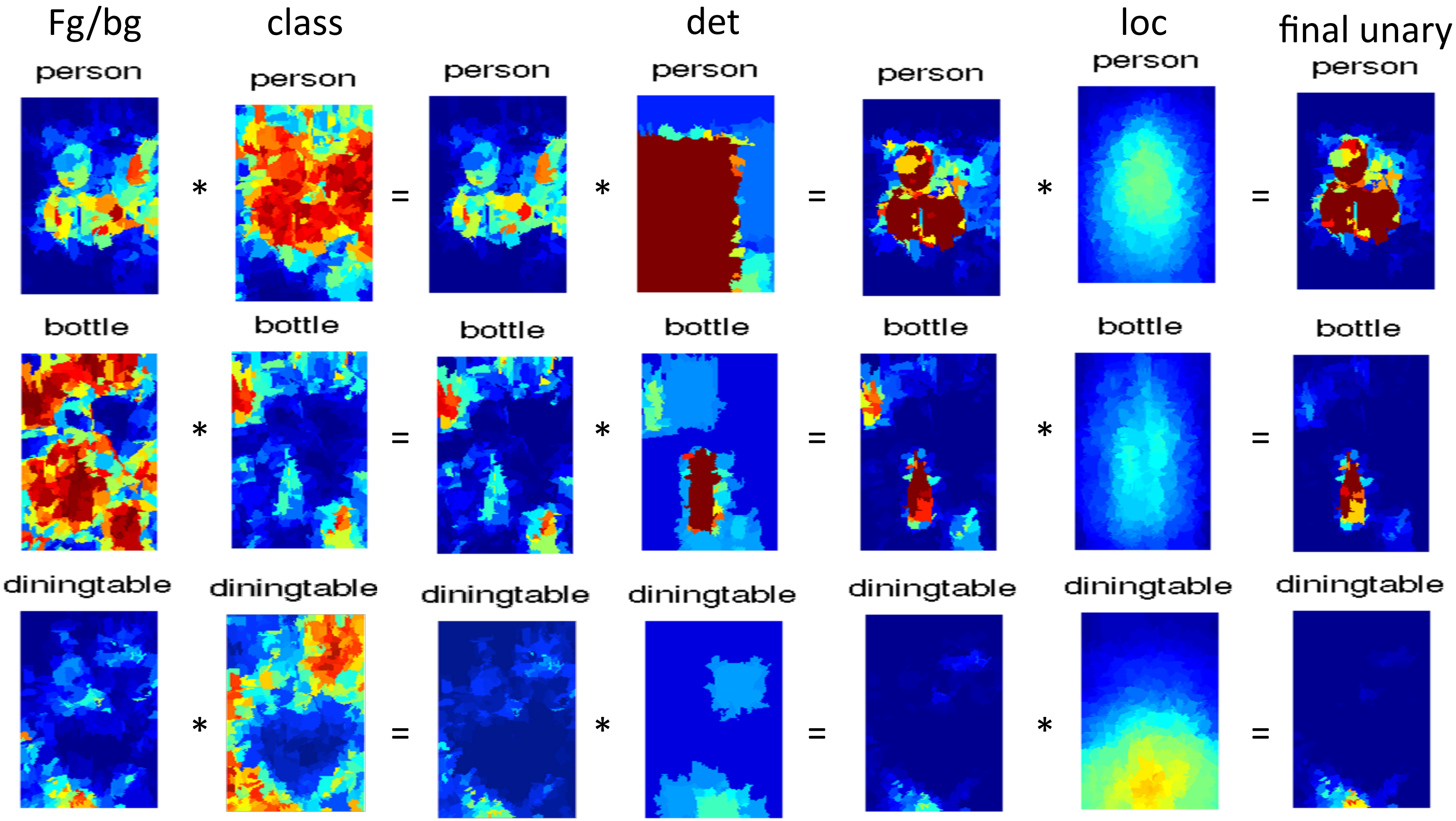
Illustrative examples



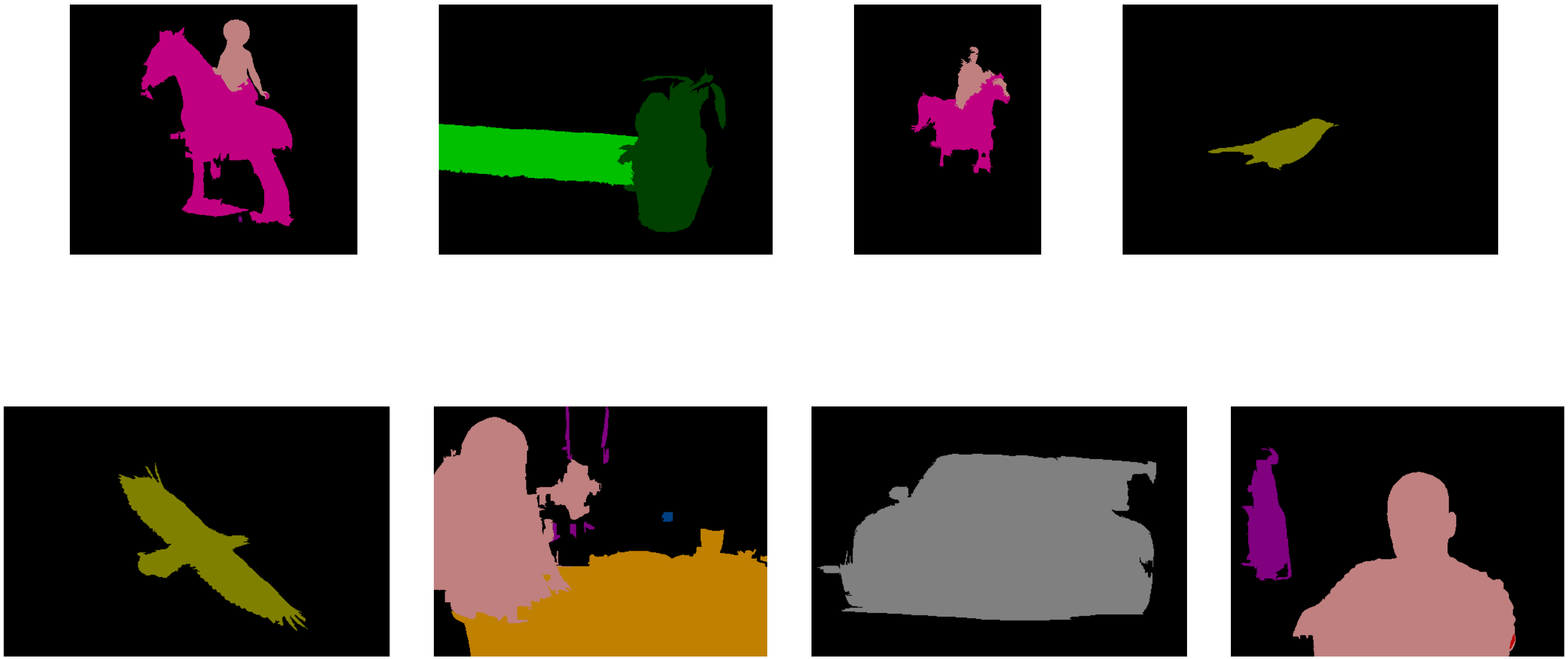


Illustrative examples





Final results



Conclusions

- Harmony potential is an effective way to fuse global and local scales for semantic image segmentation.
- We have focused on improving the local classifiers
- Baseline: 29%

  + combining fg/bg and multiclass classifiers (+2%)

  + object detection (+3%)

  + location prior (+1%)

  + per class parameter optimization (+5%)

more details: http://iselab.cvc.uab.es/pvoc2010

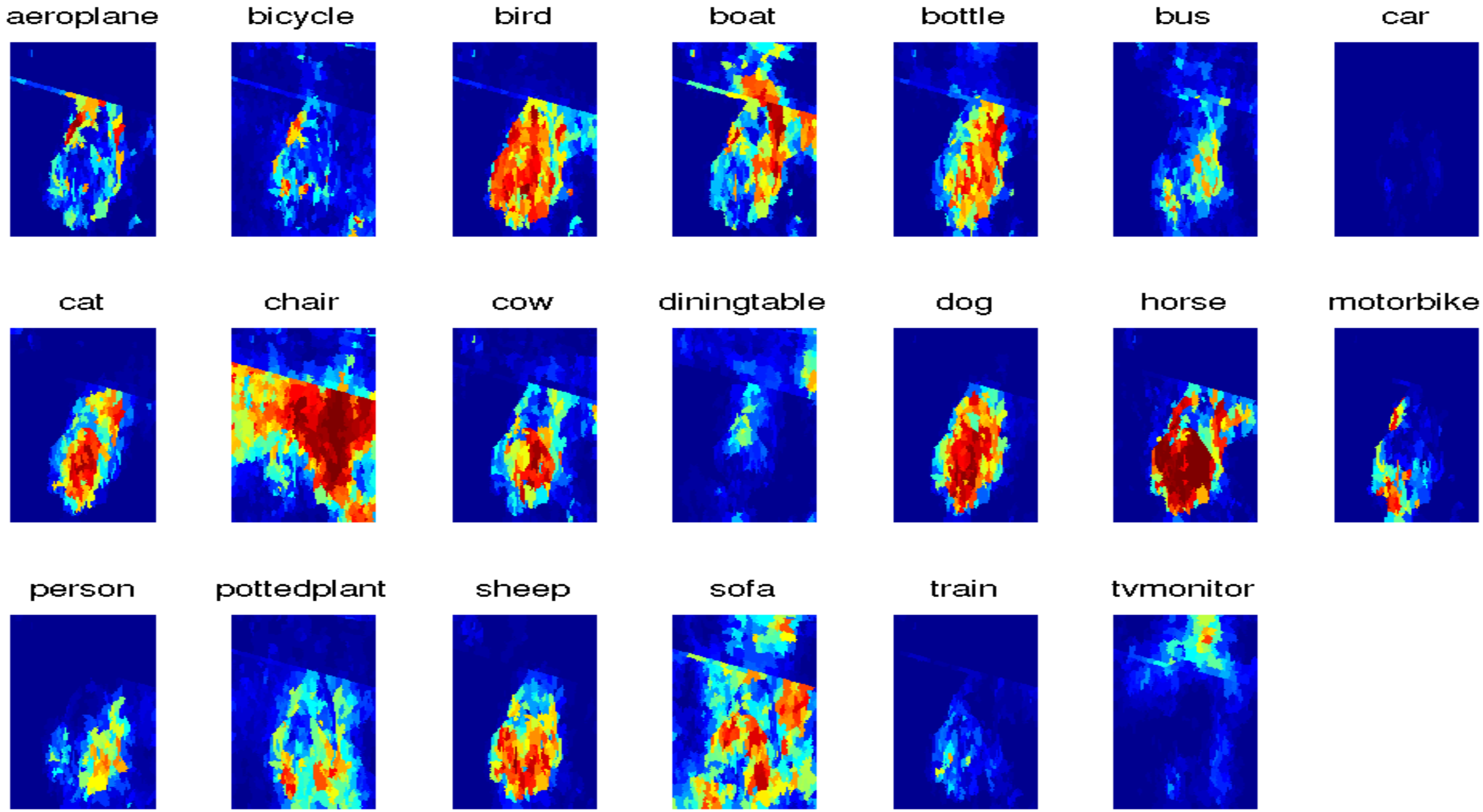
Thanks for your attention!

Gràcies per la vostra atenció!

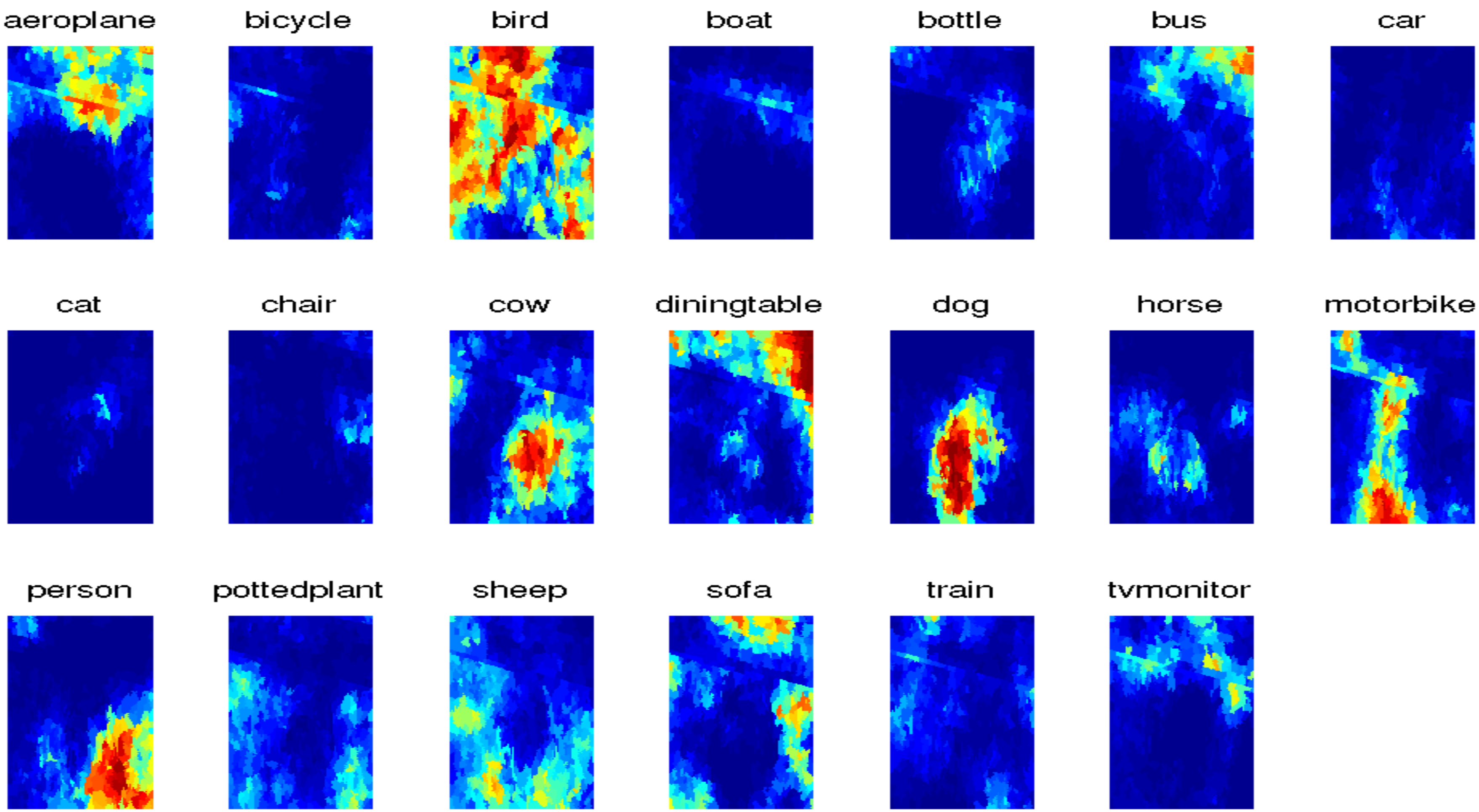
Ευχαριστώ για την προσοχή σας

Full Practical Example

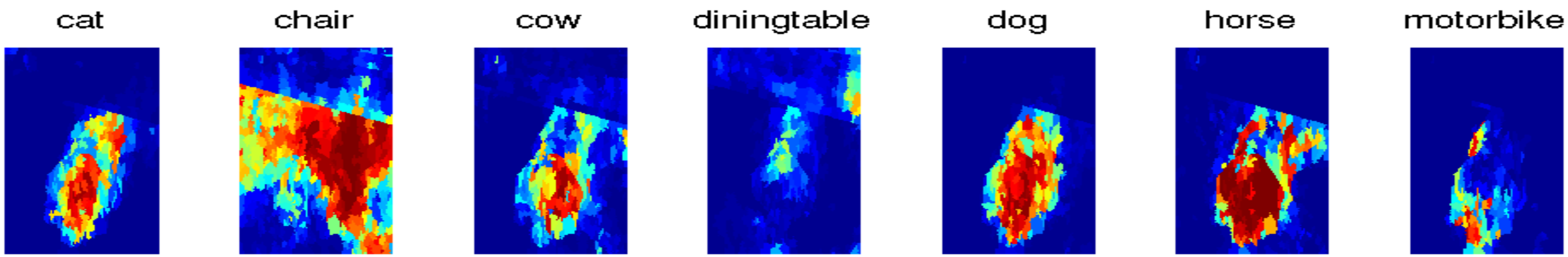
F_{fgbg}: Fore-Back ground



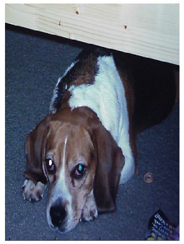
F_{class}: Class against other classes



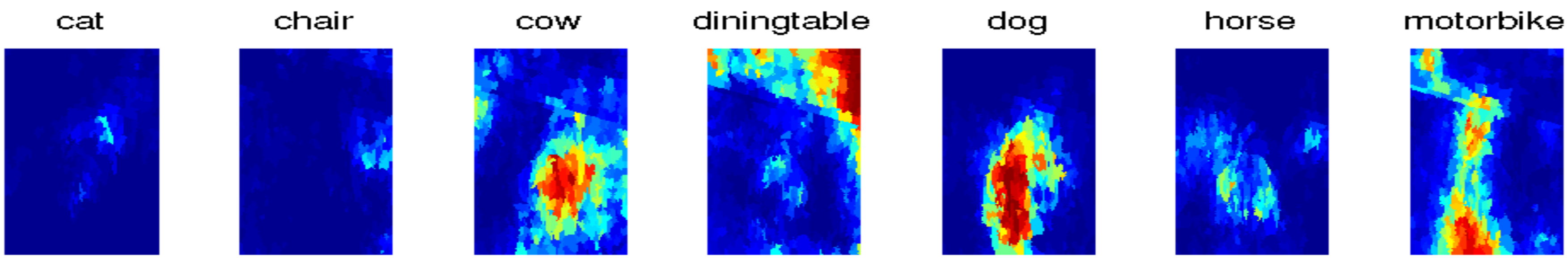
Close-up comparison



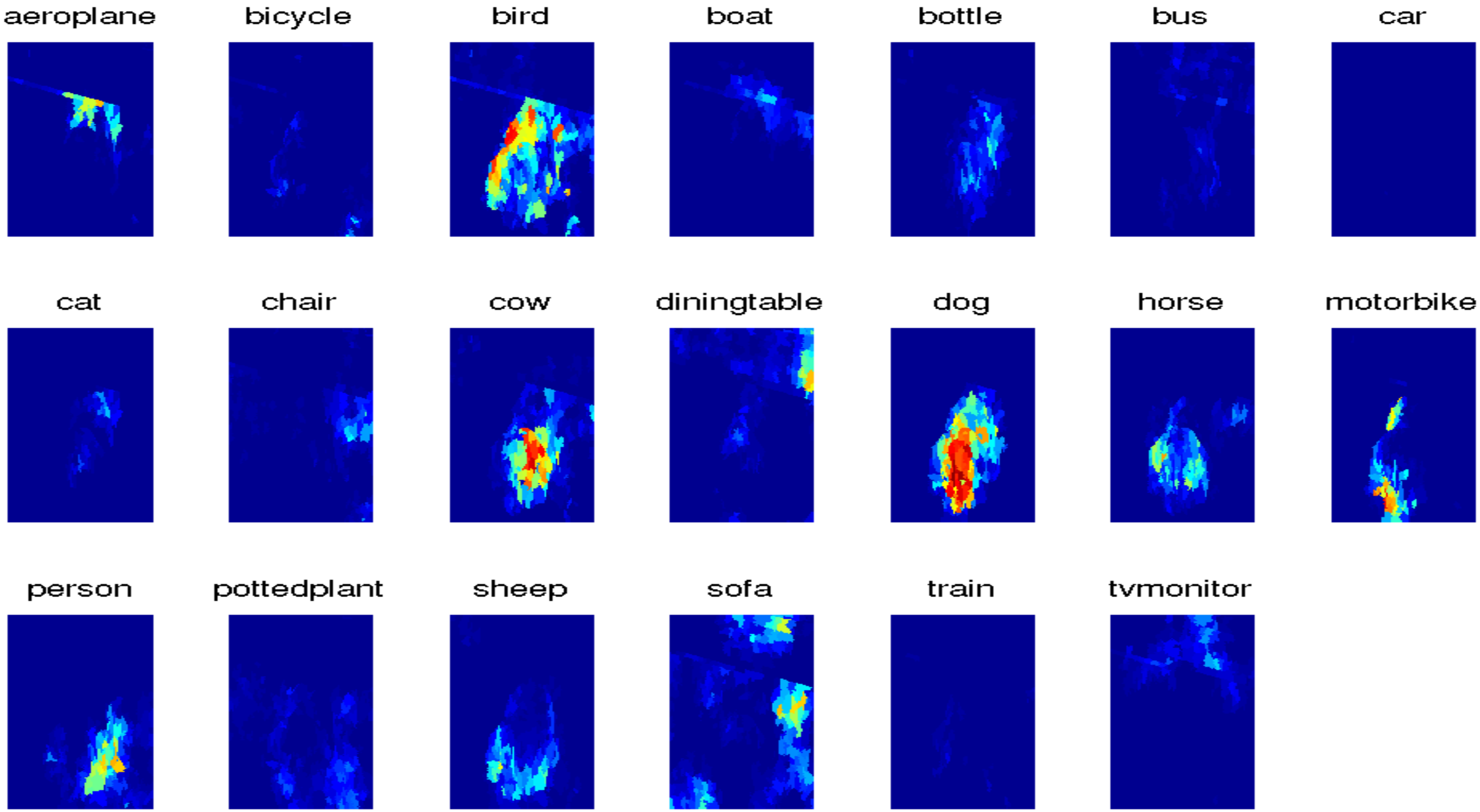
Fore-Back ground learning



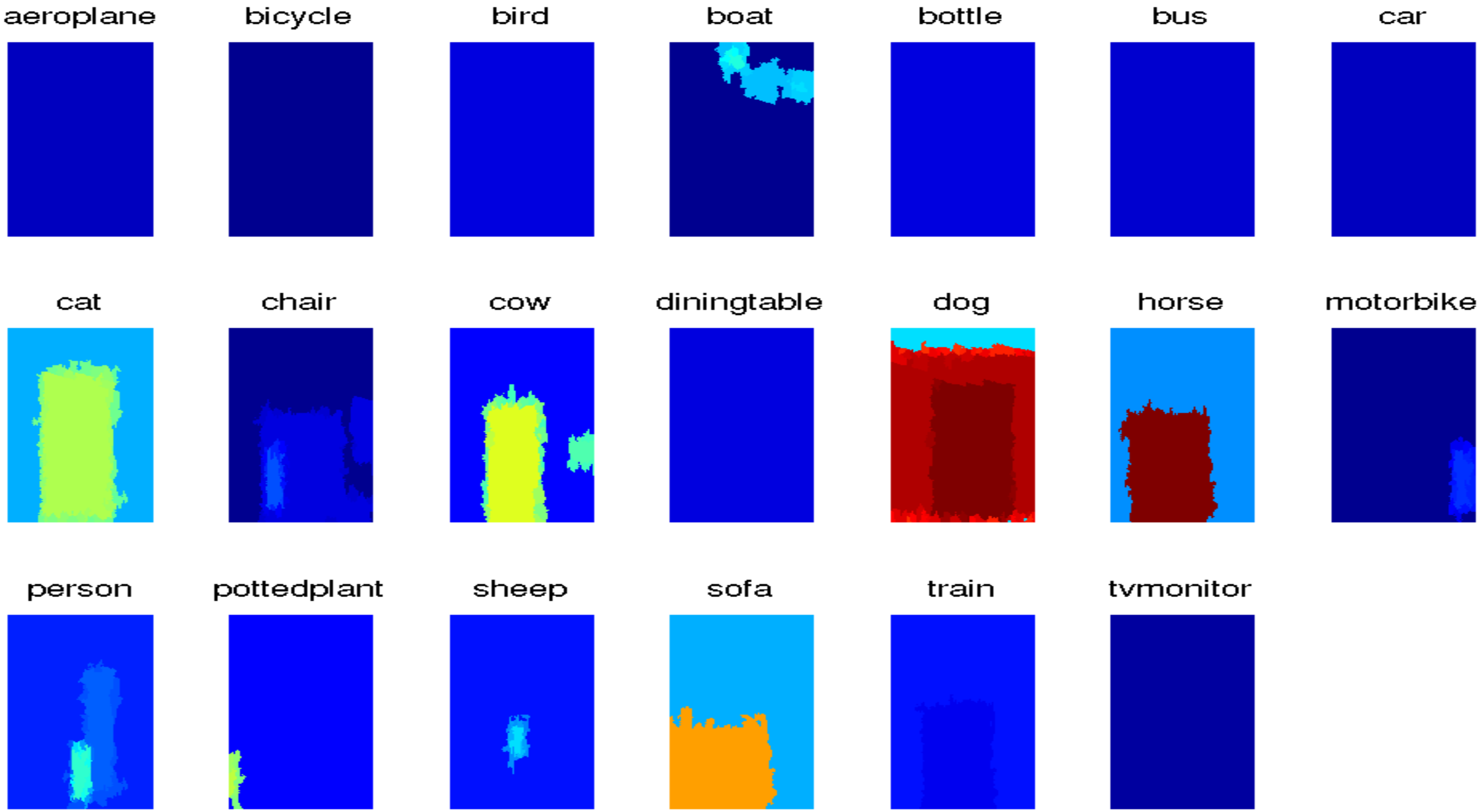
Class against others learning



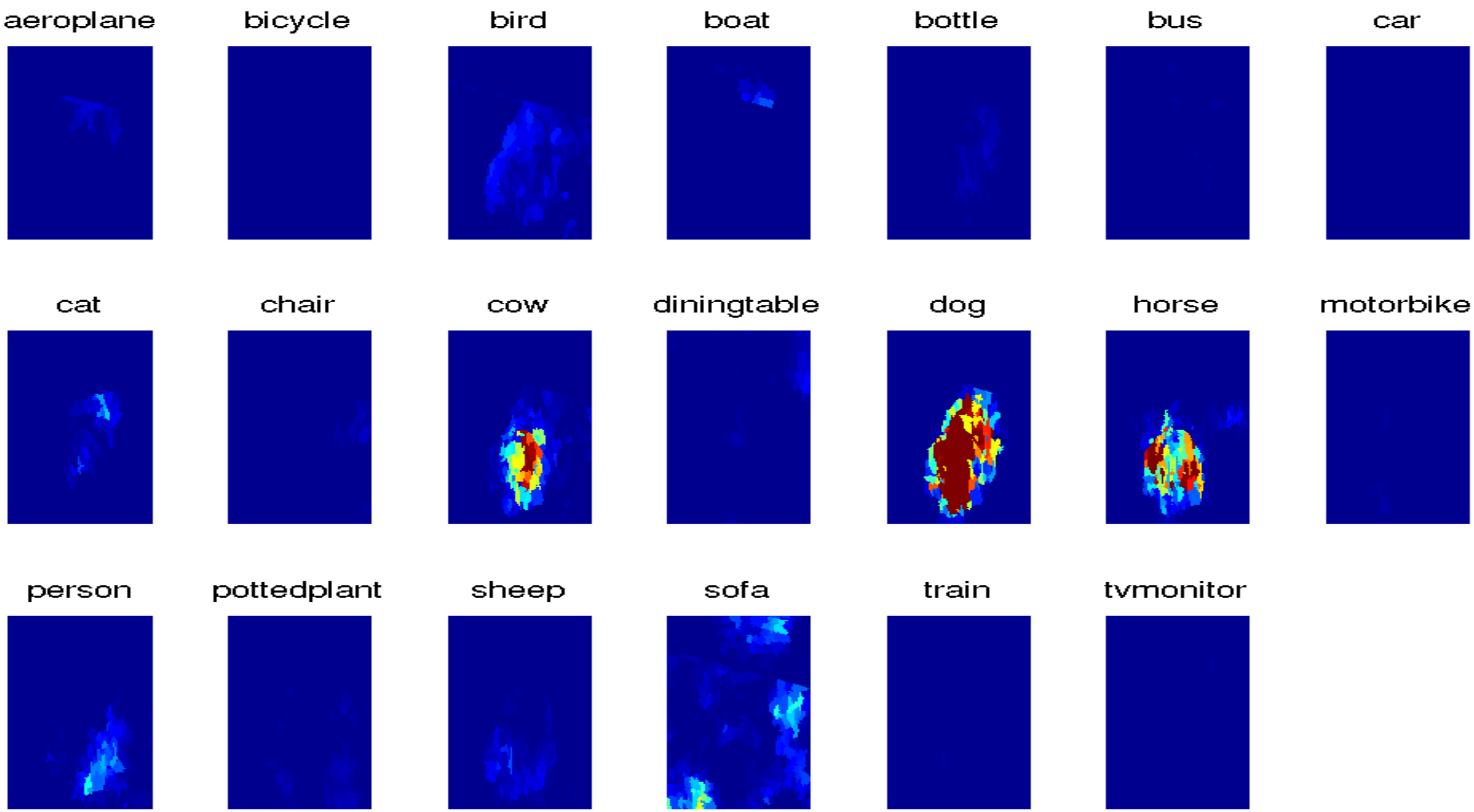
$$F_{fgbg} * F_{class}$$



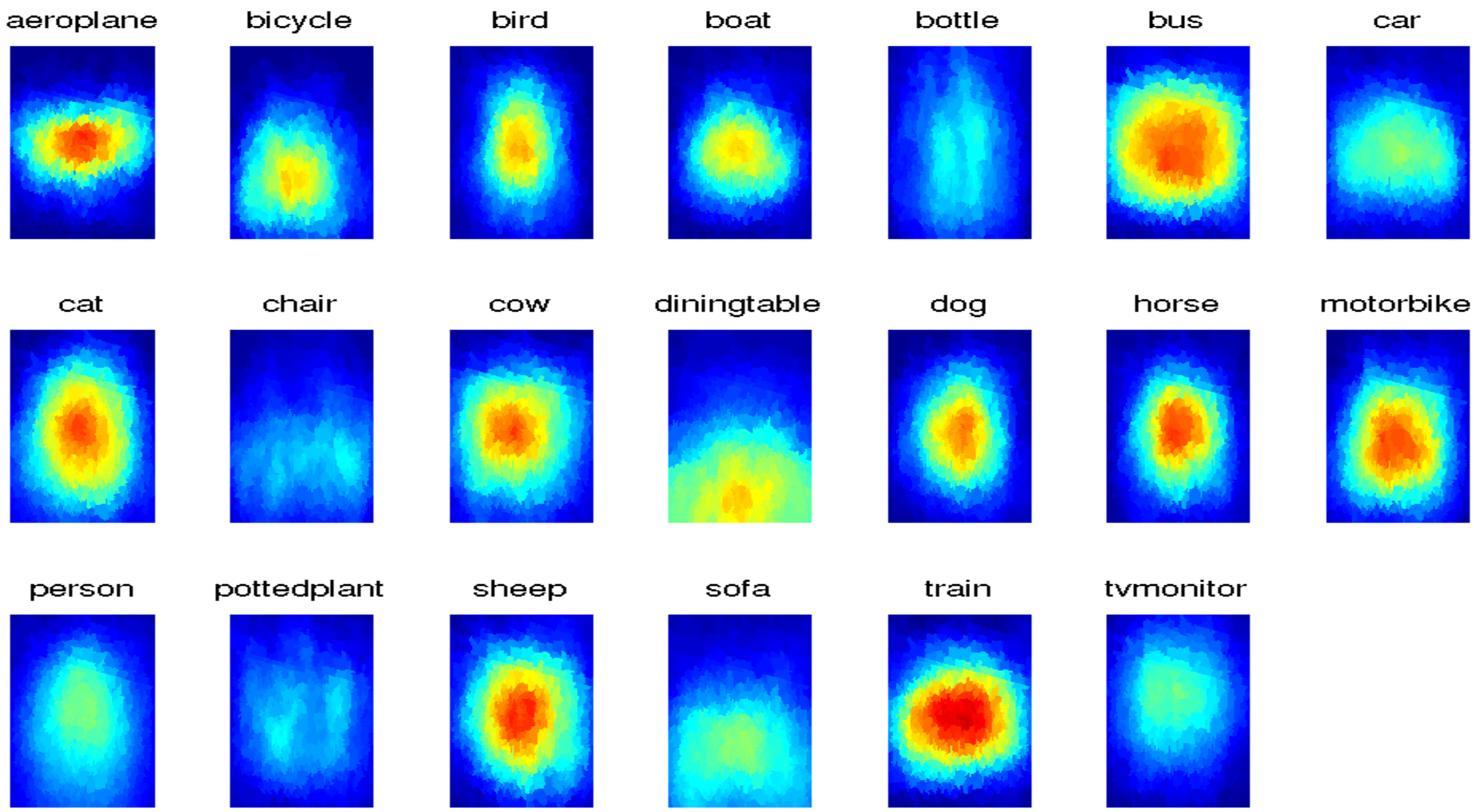
F_{det}: Detector Scores



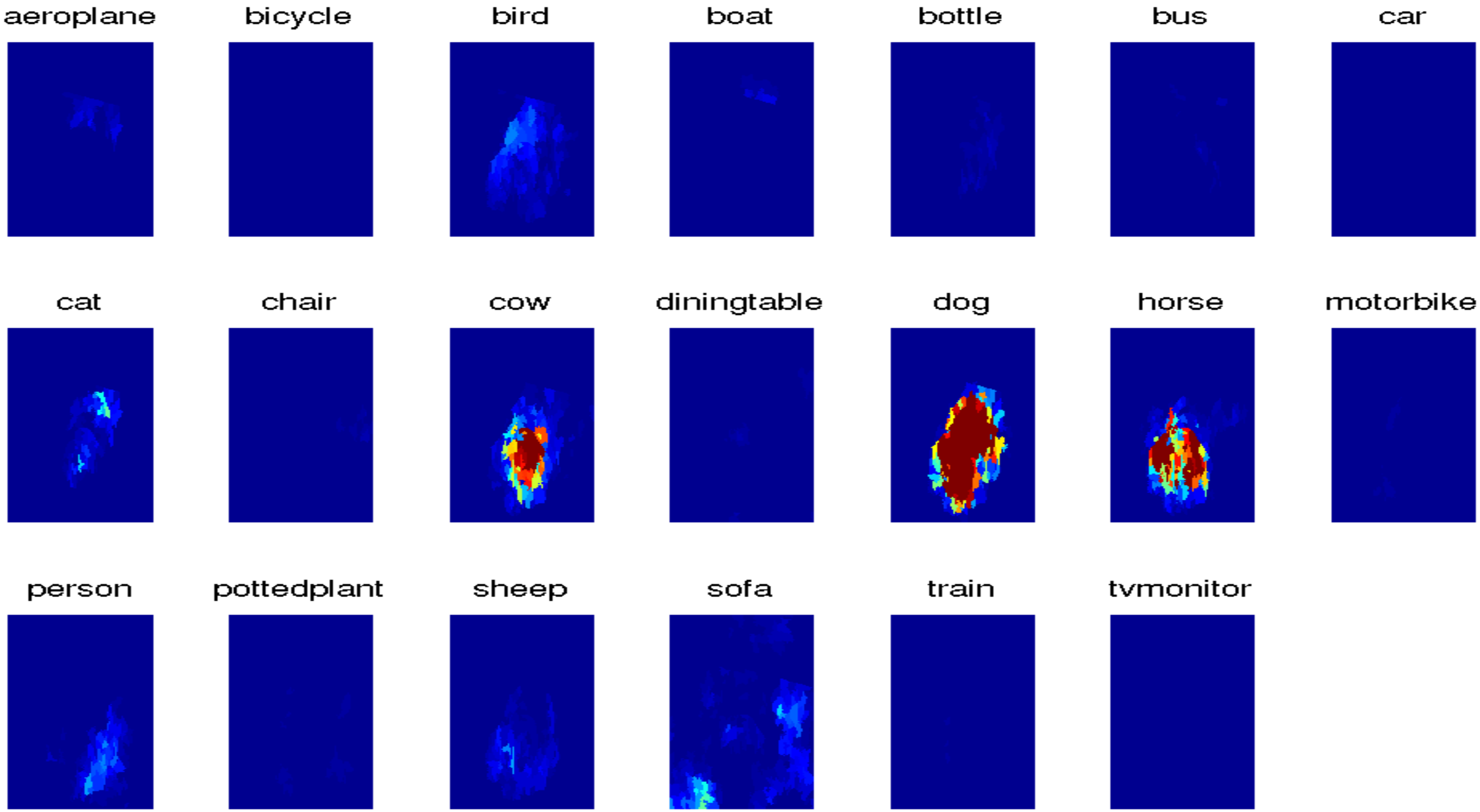
$$F_{fgbg} * F_{class} * F_{det}$$



$F_{location}$: Location Prior



$$F_{fgbg} * F_{class} * F_{det} * F_{loc}$$



Result