Seeking the strongest rigid detector, supplementary material

Rodrigo Benenson*†‡ Markus Mathias*† Tinne Tuytelaars† Luc Van Gool†
† ESAT-PSI-VISICS/IBBT,
Katholieke Universiteit Leuven, Belgium
firstname.lastname@esat.kuleuven.be
‡ Max Planck Institut für Informatik
Saarbrücken, Germany
benenson@mpi-inf.mpg.de

Abstract

This supplement provides additional curves for the different sections of the main paper. We provide, side by side, the results for all experiments both on the INRIA and ETH datasets. This allows the reader to judge the variability and trends across different datasets. Some of the plots here are copies of the one in the main paper.

Also, in section 1.1 we present in more detail the structure of the learned classifiers. This allows to visually notice the difference with traditional HOG regular cells.

1. Which feature pool?

Figures 1 and 2 present the results of the feature pool experiments, for different datasets and different scales. See corresponding section 4 in the paper for a description of each curve.

When observing all the sub-figures of figure 1 it can be seen that `Random++` is competitive with `RandomSymmetric++`, the experiments provide no strong support for choosing one over the other.

On figure 2 (and 1) we confirm that the "++" systematically improve quality.

1.1. Which features are selected?

The main paper presents for the first time experiments for a integral channel features detector trained using *all* the possible rectangular features. It is interesting to analyse which features are picked amongst all possible ones. This is certainly of interest to design pedestrian detectors and possibly for other classes too. In figure 3 we present a set of initial visualizations of the learned `AllFeatures` model. Some of the figures of this section are generated using the tools provided in the open source release of the `VeryFast` detector [1].

*Indicates equal contribution

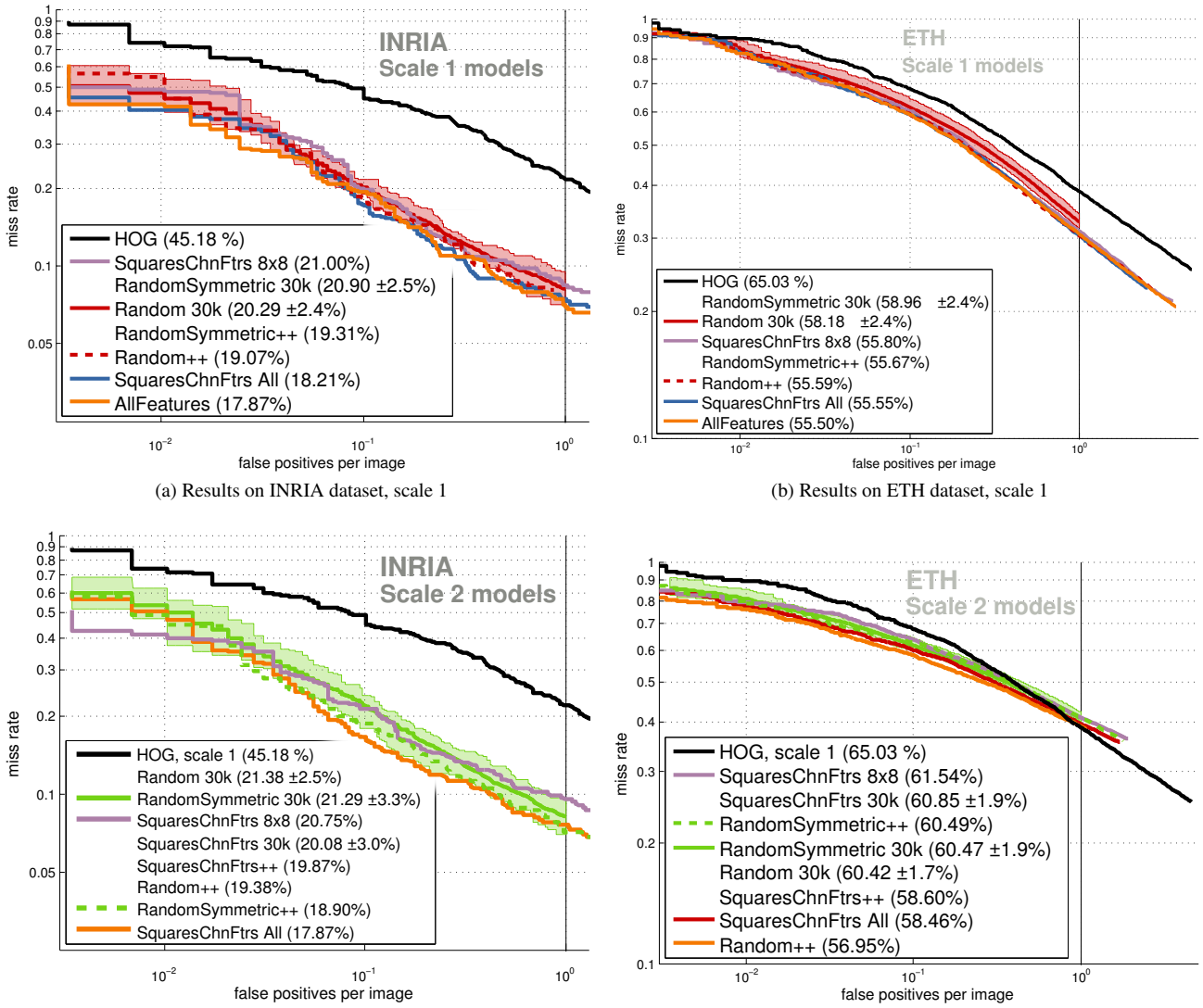
Channels distribution In figure 3a we observe that all feature channels used seem equally informative, since Adaboost selects roughly the same number of features in each channel (during the construction of the ensemble of level-2 decision trees). The vertical, horizontal, and gradient magnitude channels seems slightly more informative than the others. Consult the top row of figure 5 for an example of the computed channels.

Spatial distribution In figure 3b we plot the centre of all features' rectangles. We notice that the model area is completely covered. This indicates that spatial coverage is an important characteristic for the feature pool.
It can also be noticed that there is some asymmetry in the features (also visible in figures 5 and 4). Because we do not inject any class specific knowledge into our training, there is nothing enforcing the model to respect the reflection symmetry that human bodies exhibit.

Irregular cells As pointed out in the introduction of the main paper, one of the differentiating characteristics of our `Roerei` detector with respect to `HOG+SVM` is the use (learning) of irregular cells. In figure 3c, we sort the features by their weights magnitude and show the top 12 features (12 being an arbitrary number). It can be clearly seen that, in general, the features follow an irregular pattern. Compare figure 3c (based on data) with the illustration of figure 2 in the main paper (created for explanation purposes).

Features aspect ratio In figure 3d we plot the cumulative distribution of the features' aspect ratios. The distribution is split in three: the wide features (width $>$ height), the long features (height $>$ width), and the square features (height $=$ width). The square features all have aspect ratio 1, and thus the corresponding curve represents a single value.
We observe that most selected features are not squares (despite all squares being available in the features pool).

Features area In figure 3e we plot the cumulative distribution of the features' area. This figure clearly indicates



(a) Results on INRIA dataset, scale 1

(b) Results on ETH dataset, scale 1

(c) Results on INRIA dataset, scale 2

(d) Results on ETH dataset, scale 2

Figure 1: Detector quality on INRIA using different feature pool settings. Lower average miss-rate is better.

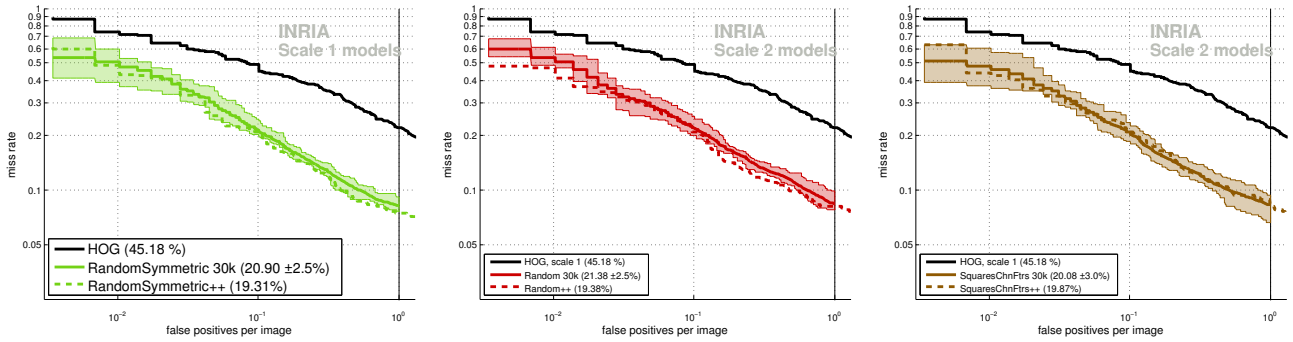
that most features are small (area smaller than 30 pixels^2, but keep in mind that we use shrinking factor 4). From the study here presented, we cannot tell how important are the few large features for the overall quality of the detector.

Relation between area and aspect ratio To give more insight beyond the cumulative distributions of figures 3d and 3e, we design one more visualization. In figure 4 we stack together all features that are in a specific range of area and aspect ratio values. In this plot the colours encode the count of overlapping features at a particular pixel (and each entry of the table is separately normalized, for better visualization).

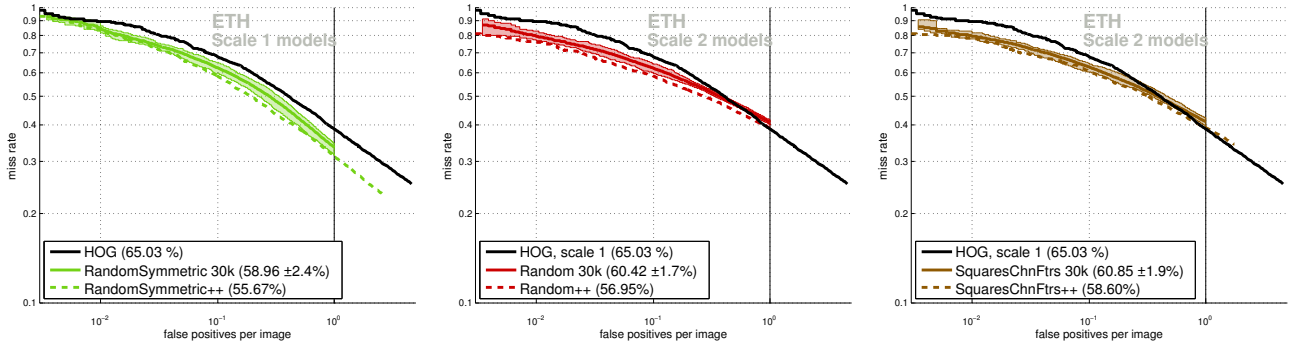
We can see that most features of small size are either slightly elongated or slightly widened. Small elongated features concentrate along the full body, while small widened features concentrate on the head and feet regions. Middle sized wide features seem to concentrate exclusively in the lower hips area. Most square features are of very small size, and concentrated on the face.

The `Roerei` detector In figure 5 we present an example of the employed channel features (top row), and present the learned model at each scale. It can be seen that as scale size increases, the model becomes more and more fine grained, presenting more details of the human anatomy.

Overall, the different channels seem to focus mainly on shoulders, head, and feet; which seems a natural choice. Scale 2 is sparser on the torso area, which might an side effect of being the only one trained using only square features



(a) Results on INRIA dataset, scale 1 (b) Results on INRIA dataset, scale 2 (c) Results on INRIA dataset, scale 2



(d) Results on ETH dataset, scale 1 (e) Results on ETH dataset, scale 2 (f) Results on ETH dataset, scale 2

Figure 2: Curves omitted in figure 1 for legibility, and their "++" counterparts.

(see section 9 of the main paper).
When comparing figure 5 with figure 5 of [1], we observe that the areas of high score contribution seem both larger and more focused.

2. Which feature normalization?

Figure 6 presents the results of the feature normalization experiments, for both INRIA and ETH datasets. See corresponding section 5 in the paper for a description of each curve.

3. Which weak classifier?

Figure 7 presents the results of the weak classifier experiments, for both INRIA and ETH datasets. See corresponding section 6 in the paper for a description of each curve.

4. Which training method?

Figure 8 presents the results of the training method experiments, for both INRIA and ETH datasets. See corresponding section 7 in the paper for a description of each curve.

5. Which training set?

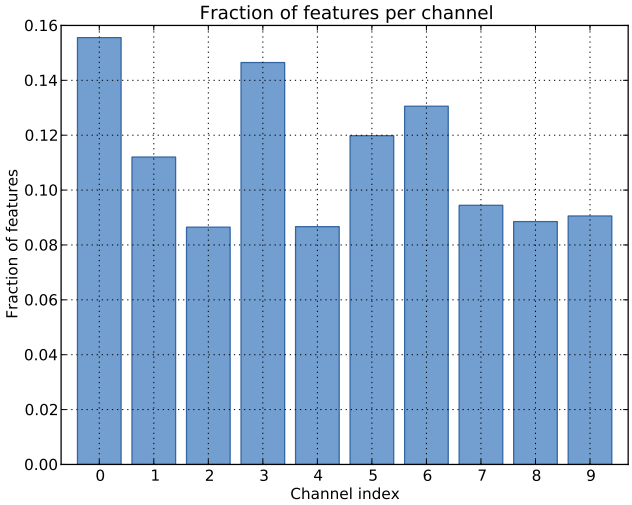
Figure 9 presents the results of the training set experiments, for both INRIA and ETH datasets. See corresponding section 8 in the paper for a description of each curve.

6. How does quality improve at each stage?

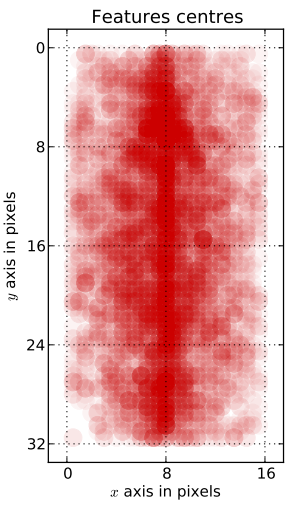
In figure 10 we have the visual counter-part of table 1 from the main paper, for INRIA and ETH datasets.

References

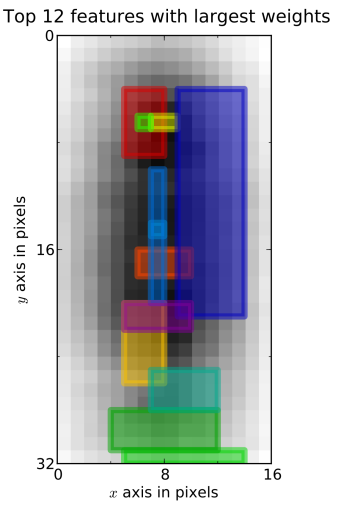
[1] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. In *CVPR*, 2012. 1, 3



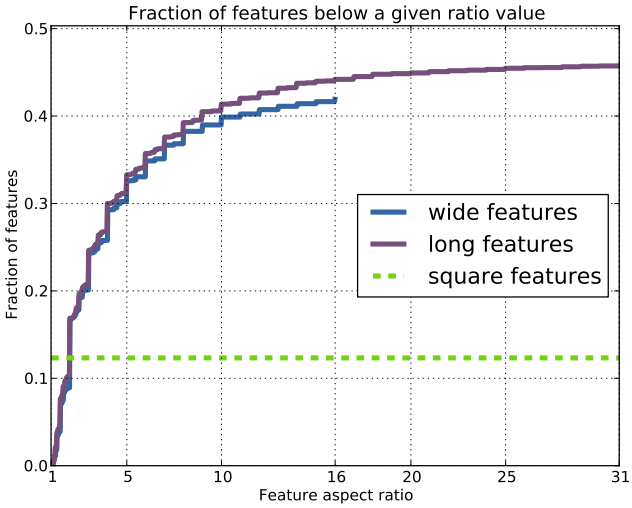
(a) Proportion of features in each channel. Channel 0 is vertical lines, channel 3 horizontal ones, channel 6 is the gradient magnitude, and channels 7,8, and 9 encode colour as LUV respectively. See also top row of figure 5



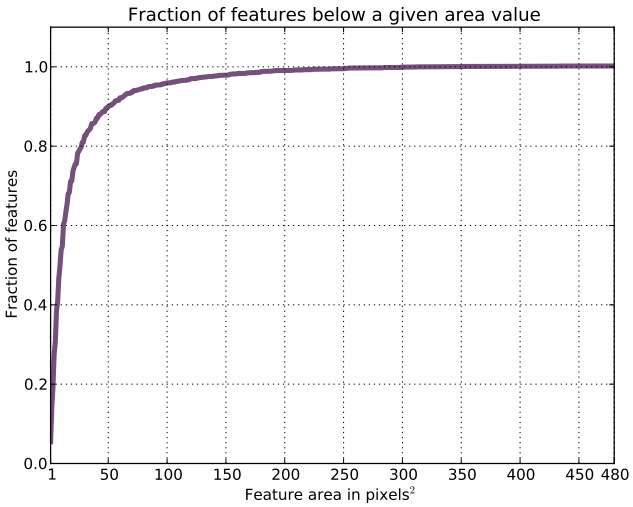
(b) Spatial distribution of the features centres, all feature channels together



(c) Top 12 features with the largest score weight. Feature colours set randomly, full model in grey background



(d) Histogram of the features aspect ratio, all feature channels together



(e) Cumulative sum of the features area distribution, all feature channels together

Figure 3: Different statistics of the `AllFeatures` model, scale 1 (16×32 pixels).

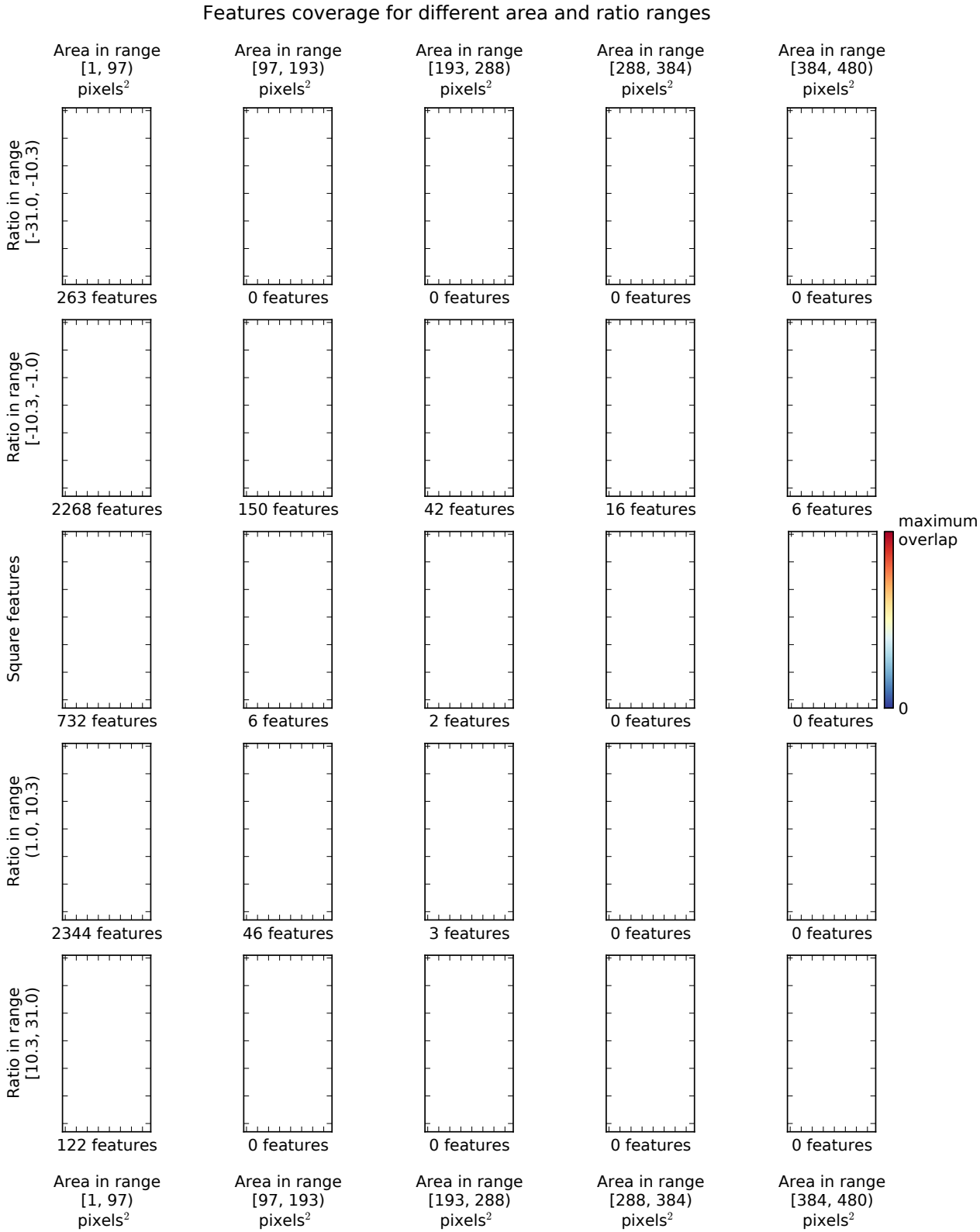


Figure 4: Spatial distribution of the selected features for the `AllFeatures` model, scale 1 (16×32 pixels). Each rectangle shows the selected features, for specific range of feature areas (columns) and aspect ratio (rows), all feature channels merged. Each rectangle is normalized so that deep blue indicates "no feature covers this area", and strong red "maximum overlap on this area".

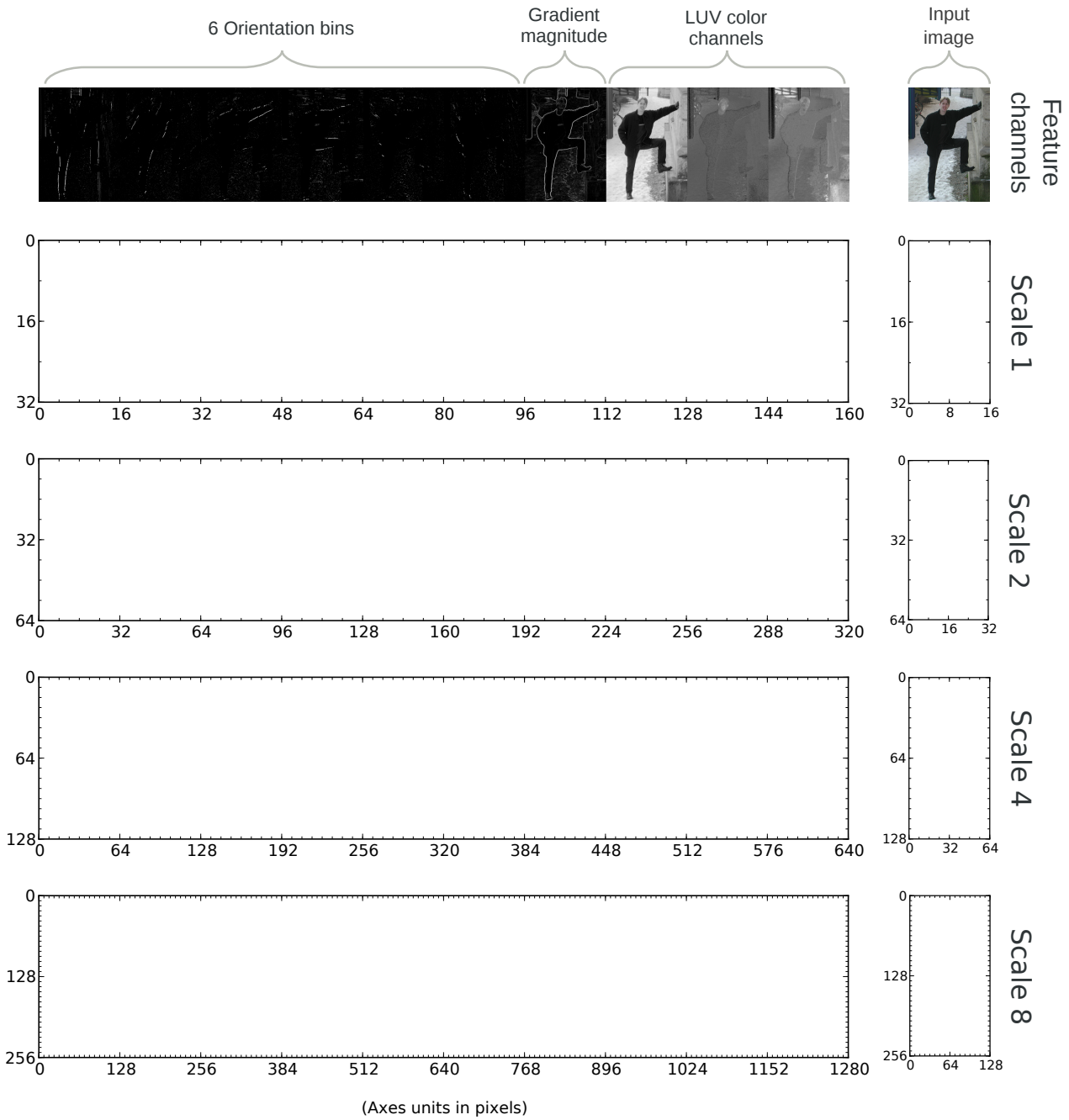
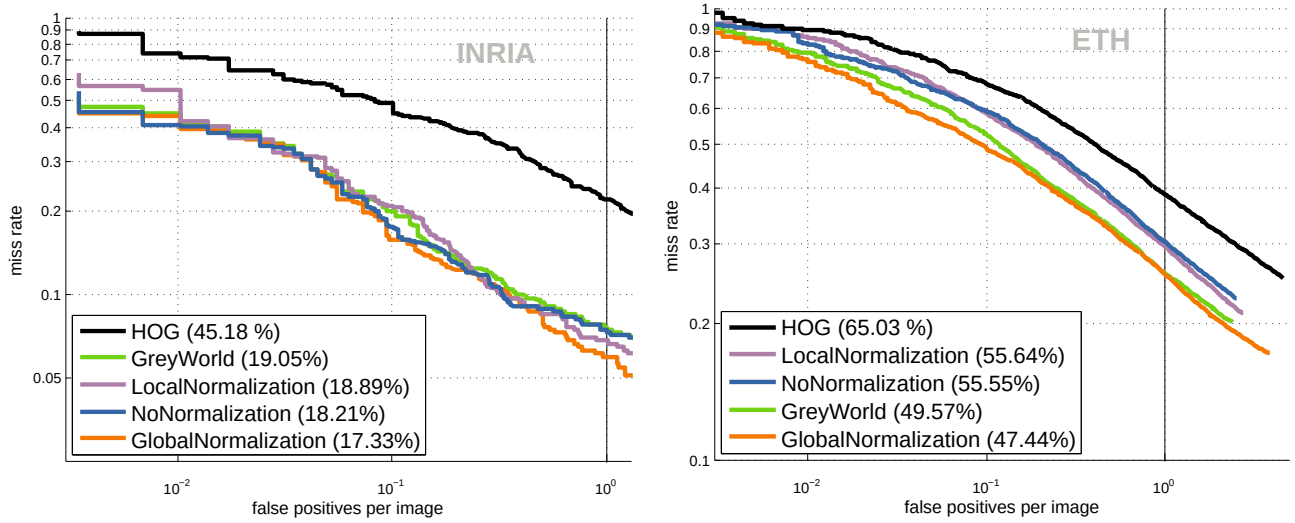


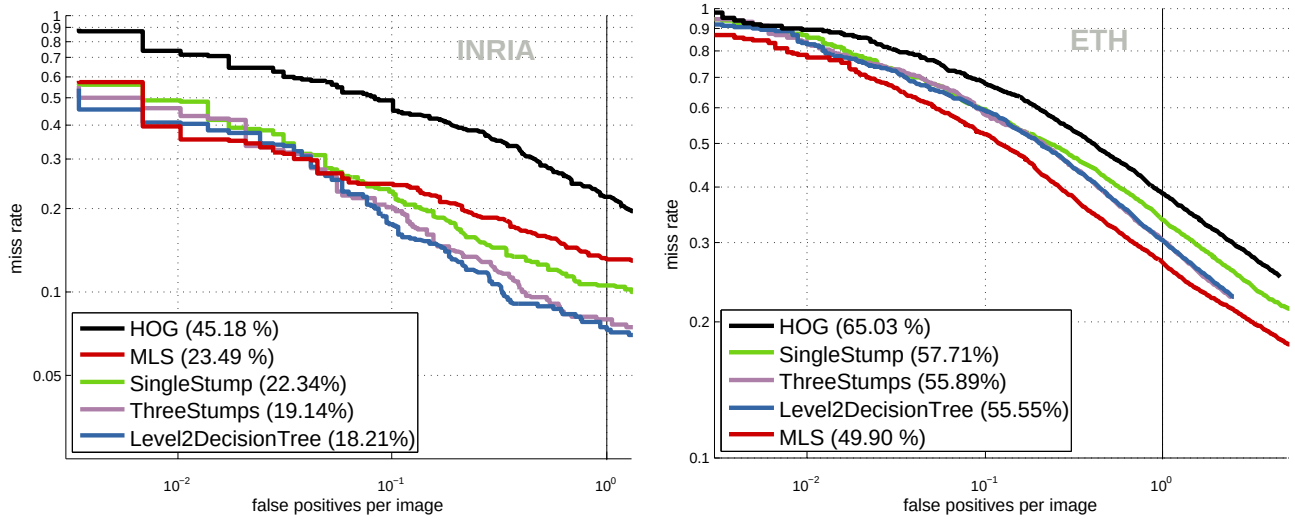
Figure 5: Visualization of the learned multi-scales `Roerei` model. The first row show an example of the computed features (and the example input image). Each column shows per channel the per-pixel detection score contribution of each model (weighted stacking of all features). Red and blue indicate high and (relatively) low contributions to the detection score, respectively. The right-most column shows the sum over all channels, for each model.



(a) Results on INRIA dataset

(b) Results on ETH dataset

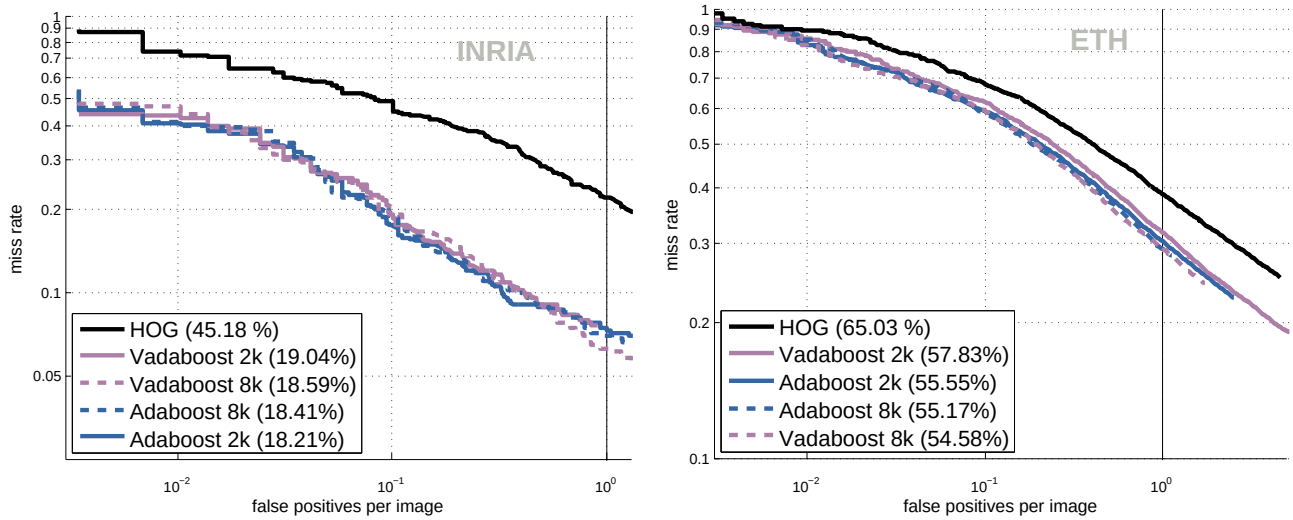
Figure 6: Detector quality when using different normalization schemes.



(a) Results on INRIA dataset

(b) Results on ETH dataset

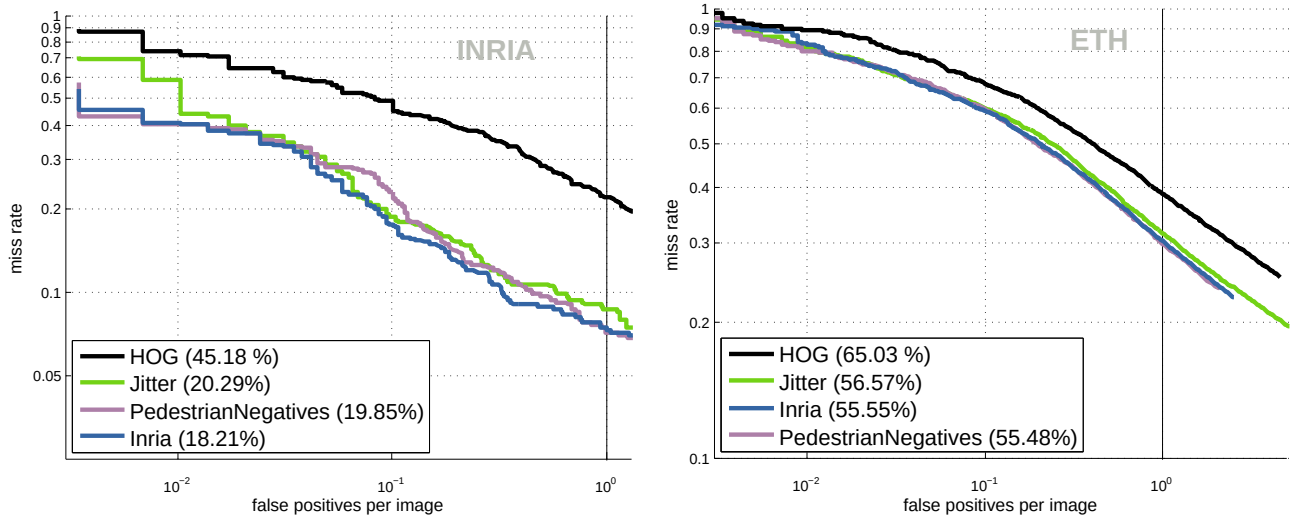
Figure 7: Which weak classifier to use?



(a) Results on INRIA dataset

(b) Results on ETH dataset

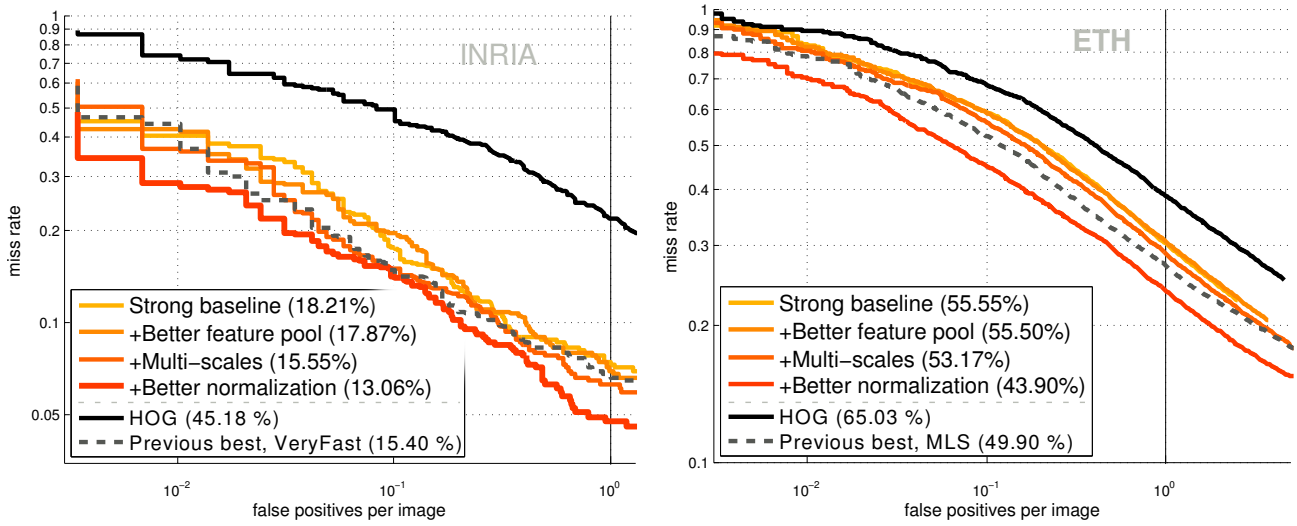
Figure 8: Which training method to use?



(a) Results on INRIA dataset

(b) Results on ETH dataset

Figure 9: Which training set to use?



(a) Step-by-step progress on INRIA dataset

(b) Step-by-step progress on ETH dataset

Figure 10: How does quality improve at each stage?