Supplementary Materials
CASENet: Deep Category-Aware Semantic Edge Detection

Zhiding Yu*
Carnegie Mellon University
yzhiding@andrew.cmu.edu

Chen Feng* Ming-Yu Liu† Srikumar Ramalingam†
Mitsubishi Electric Research Laboratories (MERL)
cfeng@merl.com, mingyul@nvidia.com, srikumar@cs.utah.edu

1. Multi-label Edge Visualization

In order to effectively visualize the prediction quality of multi-label semantic edges, the following color coding protocol is used to generate results in Fig. 1, Fig. 4, and Fig. 6 in the main paper. First, we associate each of the K semantic object class a unique value of Hue, denoted as $\mathsf{H} \triangleq [\mathsf{H}_0, \mathsf{H}_1, \cdots, \mathsf{H}_{K-1}]$. Given a K-channel output $\mathbf{Y}$ from our CASENet's fused classification module, where each element $\mathbf{Y}_k(\mathbf{p}) \in [0, 1]$ denotes the pixel $\mathbf{p}$'s predicted confidence of belonging to the k-th class, we return an HSV value for that pixel based on the following equations:

$$\mathbf{H}(\mathbf{p}) = \frac{\sum_k \mathbf{Y}_k(\mathbf{p})\mathsf{H}_k}{\sum_k \mathbf{Y}_k}, \tag{1}$$

$$\mathbf{S}(\mathbf{p}) = 255 \max\{\mathbf{Y}_k(\mathbf{p}) | k = 0, \cdots, K-1\}, \tag{2}$$

$$\mathbf{V}(\mathbf{p}) = 255, \tag{3}$$

which is also how the ground truth color codes are computed (by using $\hat{\mathbf{Y}}$ instead). Note that the edge response maps of testing results are thresholded with 0.5, with the two classes having the strongest responses selected to compute hue based on Eq. (1).

For Cityscapes, we manually choose the following hue values to encode the 19 semantic classes so that the mixed Hue values highlight different multi-label edge types:

$$\mathsf{H} \triangleq [359, 320, 40, 80, 90, 10, 20, 30, 140, 340, 280, 330, 350, 120, 110, 130, 150, 160, 170] \tag{4}$$

The colors and their corresponding class names are illustrated in following Table 1. The way Hue is mixed in equation 1 indicates that any strong false positive response or incorrect response strength can lead to hue values shifted from ground truth. This helps to visualize false prediction.

road	sidewalk	building	wall
fence	pole	traffic light	traffic sign
vegetation	terrain	sky	person
rider	car	truck	bus
train	motorcycle	bicycle	

Table 1. The adopted color codes for Cityscapes semantic classes.

2. Additional Results on SBD

2.1. Early stage loss analysis

Fig. 1 shows the losses of different tested network configurations between iteration 100-500. Note that for Fig. 3 in the main paper, loss curves between iteration 0-8000 is not available due to the large averaging kernel size. One can see CASENet's fused loss is initially larger than its side5 loss. It later drops faster and soon become consistently lower than the side5 loss (see Fig. 3 in the main paper).

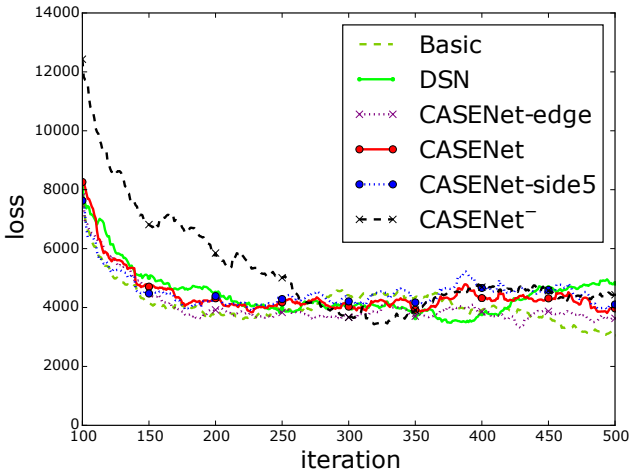

Figure 1. Early stage losses (up to 500 iterations) of different network configurations with a moving average kernel length of 100.

*The authors contributed equally.
†This work was done during the affiliation with MERL.

2.2. Class-wise prediction examples

We illustrate 20 typical examples of the class-wise edge predictions of different comparing methods in Fig. 2 and 3, with each example corresponding to one of the SBD semantic category. One can observe that the proposed CASENet slightly but consistently outperforms ResNets with the basic and DSN architectures, by overall showing sharper edges and often having stronger responses on difficult edges.

Meanwhile, Fig. 4 shows several difficult or failure cases on the SBD Datasets. Interestingly, while the ground truth says there is no "aeroplane" in the first row and "dining table" in the second, the network is doing decently by giving certain level of edge responses, particularly in the "dining table" example. The third row shows an example of the false positive mistakes often made by the networks on small objects. The networks falsely think there is a sheep while it is in fact a rock. When objects become smaller and lose details, such mistakes in general happen more frequently.

2.3. Class-wise precision-recall curves

Fig. 5 shows the precision-recall curves of each semantic class on the SBD Dataset. Note that while post-processing edge refinement may further boost the prediction performance [1], we evaluate only on the raw network predictions to better illustrate the network performance without introducing other factors. The evaluation is conducted fully based on the same benchmark code and ground truth files released by [2]. Results indicate that CASENet slightly but consistently outperforms the baselines.

2.4. Performance at different iterations

We evaluate the Basic, DSN, CASENet on SBD for every 2000 iterations between 16000-30000, with the MF score shown in Fig. 6. We found that the performance do not change significantly, and CASENet consistently outperforms Basic and DSN.

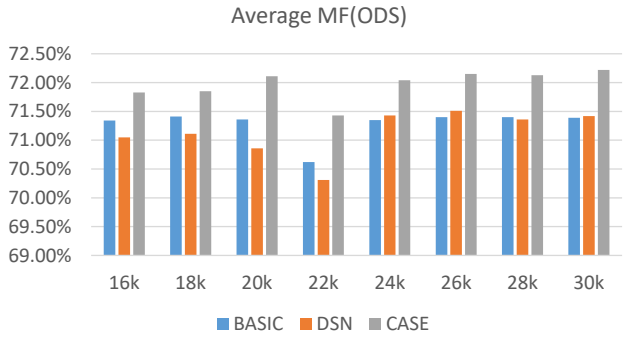

Figure 6. Testing Performance vs. different iterations.

2.5. Performance with a more standard split

Considering that many datasets adopts the training + validation + test data split, we also randomly divided the SBD training set into a smaller training set and a new validation set with 1000 images. We used the average loss on validation set to select the optimal iteration number separately for both Basic and CASENet. Their corresponding MFs on the test set are 71.22% and 71.79%, respectively.

3. Additional Results on Cityscapes

3.1. Additional qualitative results

For more qualitative results, the readers may kindly refer to our released videos on Cityscapes validation set, as well as additional demo videos.

3.2. Class-wise precision-recall curves

Fig. 7 shows the precision-recall curves of each semantic class on the Cityscapes Dataset. Again the evaluation is conducted only on the raw network predictions. Since evaluating the results at original scale (1024×2048) is extremely slow and is not necessary, we bilinearly downsample both the edge responses and ground truths to 512×1024. Results indicate that CASENet consistently outperforms the ResNet with the DSN architecture.

References

[1] G. Bertasius, J. Shi, and L. Torresani. High-for-low, low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision. In *ICCV*, 2015. 2

[2] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 2

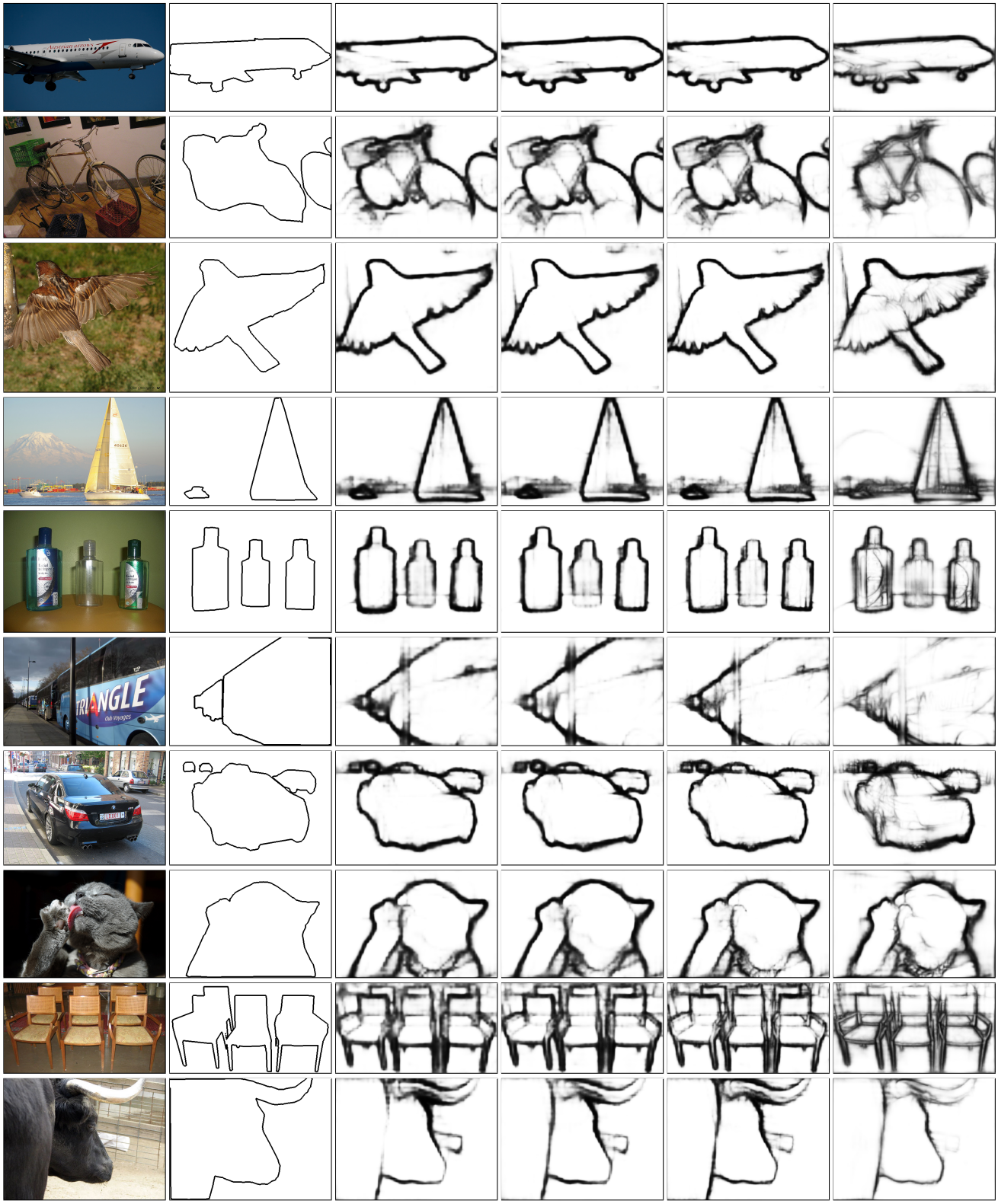

Figure 2. Class-wise prediction results of comparing methods on the SBD Dataset. Rows correspond to the predicted edges of “aeroplane”, “bicycle”, “bird”, “boat”, “bottle”, “bus”, “car”, “cat”, “chair” and “cow”. Columns correspond to original image, ground truth, and results of Basic, DSN, CASENet and CASENet-VGG.

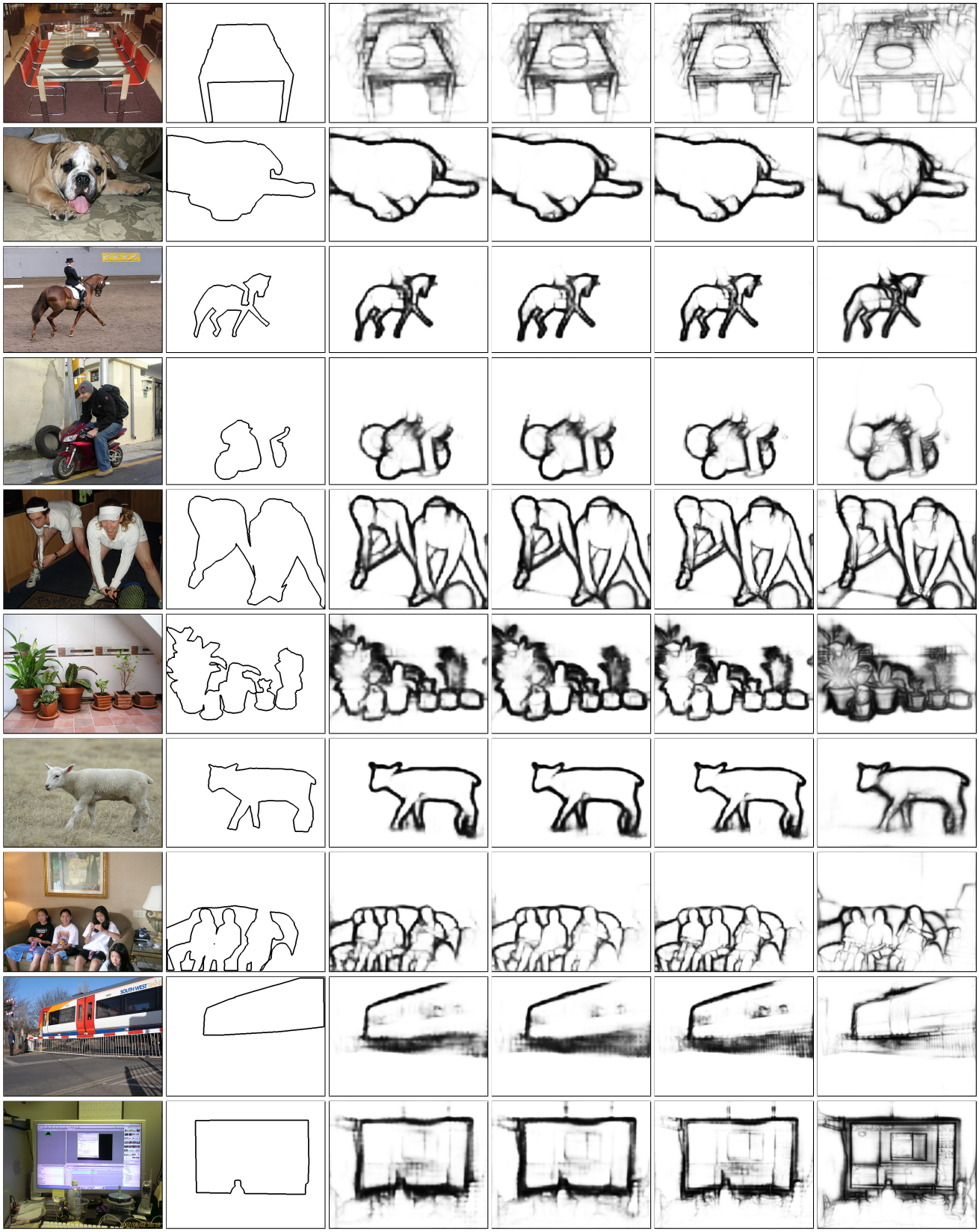

Figure 3. Class-wise prediction results of comparing methods on the SBD Dataset. Rows correspond to the predicted edges of “dining table”, “dog”, “horse”, “motorbike”, “person”, “potted plant”, “sheep”, “sofa”, “train” and “tv monitor”. Columns correspond to original image, ground truth, and results of Basic, DSN, CASENet and CASENet-VGG.

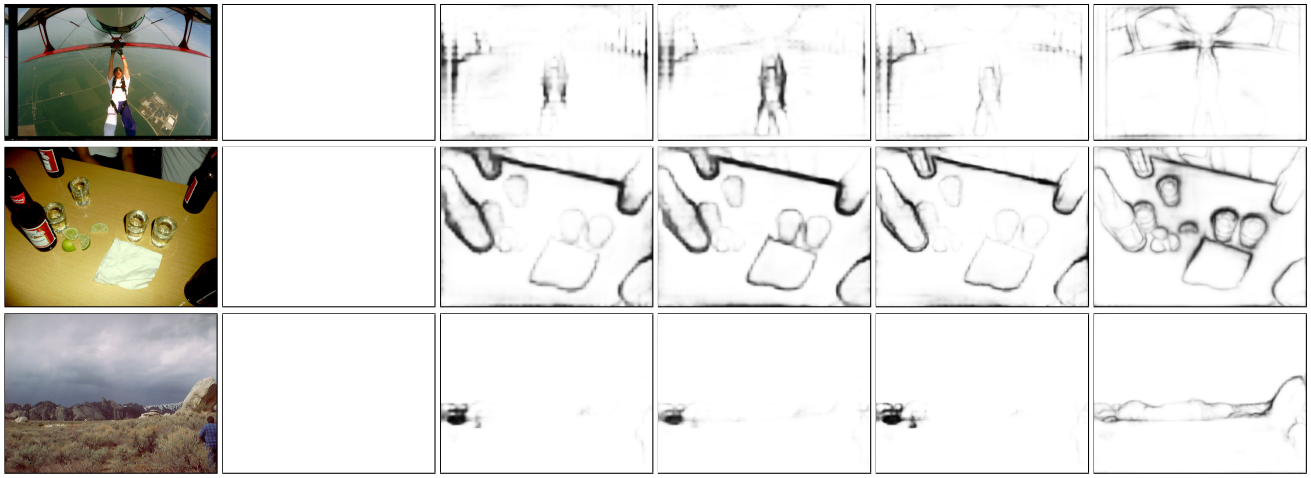

Figure 4. Difficult or failure cases on the SBD Dataset. Rows correspond to the predicted edges of “aeroplane”, “dining table” and “sheep”. Columns correspond to original image, ground truth, and results of Basic, DSN, CASENet and CASENet-VGG.

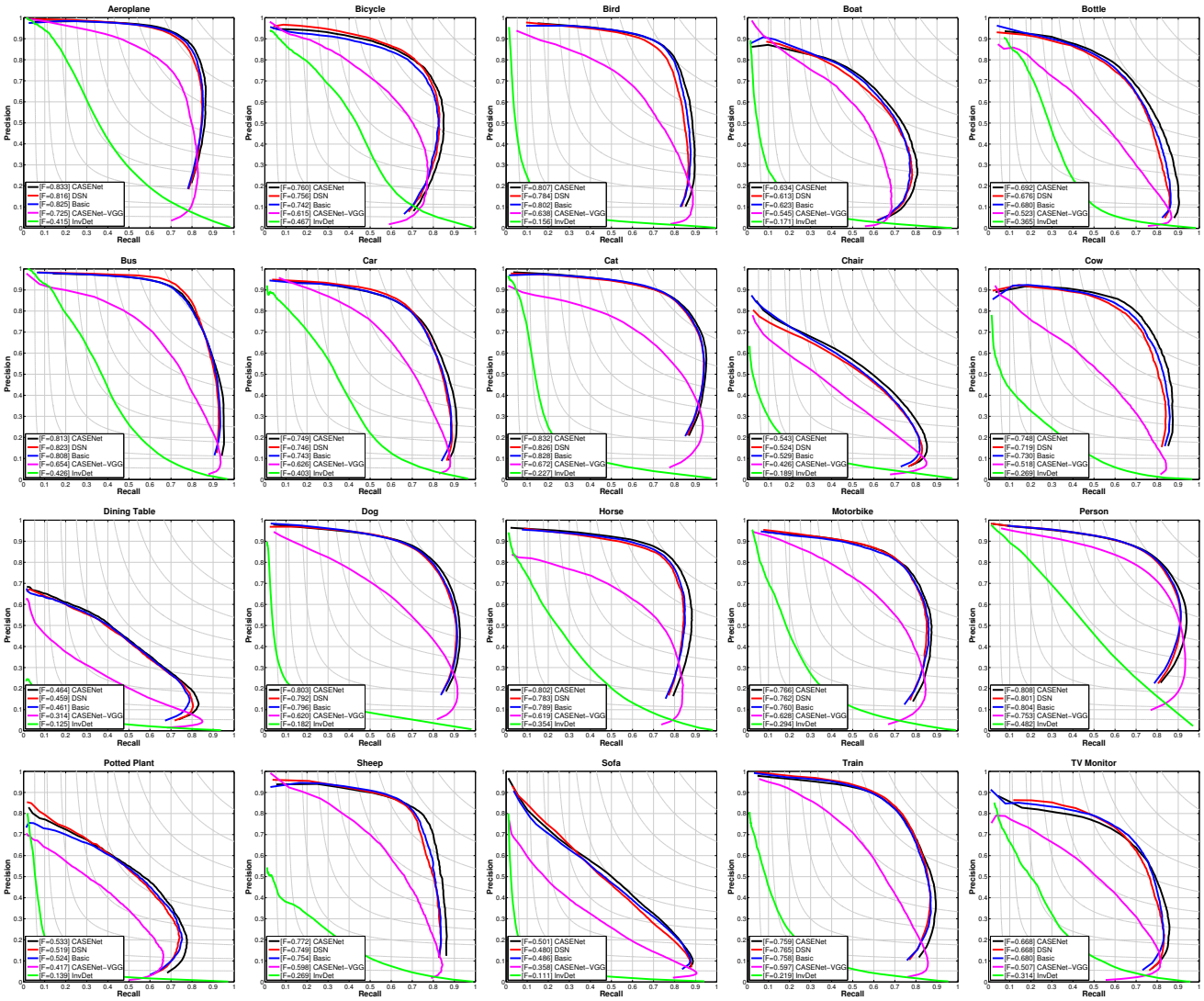

Figure 5. Class-wise precision-recall curves of the proposed methods and baselines on the SBD Dataset.

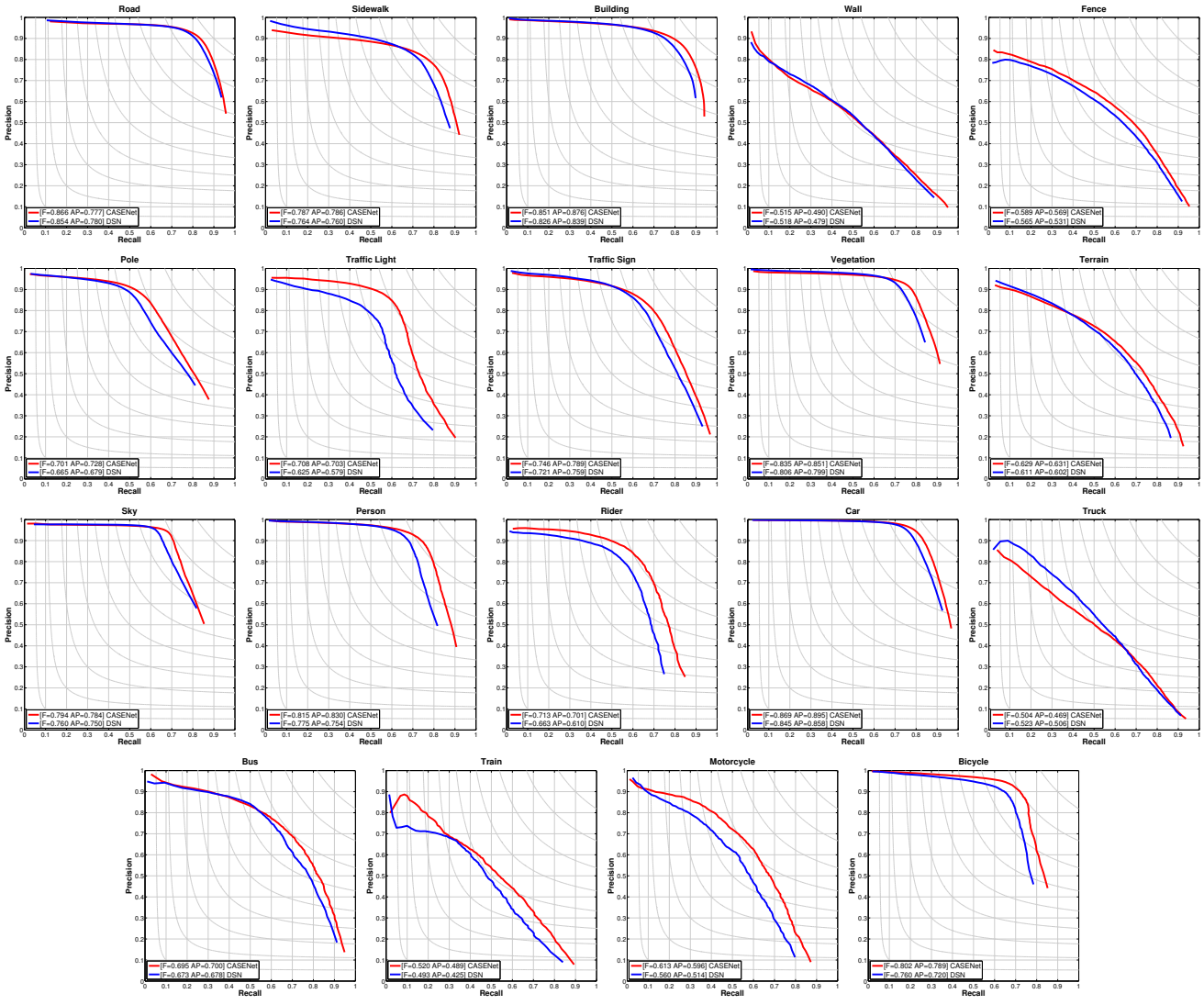

Figure 7. Class-wise precision-recall curves of CASENet and DSN on the Cityscapes Dataset.