Railroad is not a Train: Saliency as Pseudo-pixel Supervision for Weakly Supervised Semantic Segmentation (*Supplementary Material*)

Seungho Lee*
Yonsei University
seungholee@yonsei.ac.kr

Minhyun Lee*
Yonsei University
lmh315@yonsei.ac.kr

Jongwuk Lee
Sungkyunkwan university
jongwuklee@skku.edu

Hyunjung Shim†
Yonsei University
kateshim@yonsei.ac.kr

A. Appendix

A.1. Implementation Details

Classification networks. As mentioned in Section 4 (Experimental setup), we choose ResNet38 [48] as the backbone network of our method with an output stride of 8 for both PASCAL VOC 2012 [12] and MS COCO 2014 [30]. For COCO 2014 dataset, our method is trained up to 50 epochs with a batch size of 16, where input images are randomly cropped by 321×321 and λ is 0.9. Other hyperparameters are the same as those for Pascal VOC 2012 dataset.

Segmentation networks. We adopt five segmentation networks for evaluating PASCAL VOC 2012: 1) VGG16 based DeepLab-V1, 2) VGG16 based DeepLab-V2, 3) ResNet101 based DeepLab-V1, 4) ResNet101 based DeepLab-V2, and 5) ResNet38 based DeepLab-V1. Following the common practice in developing the segmentation model, input images are randomly scaled to $[0.5, 0.75, 1.0, 1.25, 1.5]$ and cropped to 321×321 for training. For inference, test images are scaled to 513×513 without cropping. We use the SGD optimizer with a batch size of 10 (20 for COCO 2014) and train the networks until 20k iterations(24k for COCO 2014). The initial learning rate is $1e-3$ for segmentation networks with VGG16 and ResNet38 and $2.5e-4$ for segmentation networks with ResNet101. We follow the poly policy $lr_{iter} = lr_{init}(1 - \frac{iter}{max_iter})^{\gamma}$ with $\gamma = 0.9$ for a learning rate decay. We set the momentum as 0.9 and the weight decay term as $5e-4$. All the segmentation networks are implemented on PyTorch [36] and trained using two NVIDIA GeForce RTX 2080Ti GPUs.

*indicates an equal contribution.
†Hyunjung Shim is a corresponding author.

λ	0.0	0.25	0.5	0.75	1.0
mIoU (%)	68.2	68.5	69.4	**69.9**	69.5

Table A.1. Effects of λ on the pseudo-mask generation. The performance (mIoU) of pseudo-masks is reported under various λ. The results are evaluated on the PASCAL VOC 2012 train set.

τ	0.2	0.3	0.4	0.5
mIoU (%)	68.3	68.9	**69.4**	69.1

Table A.2. Effects of τ on the pseudo-mask generation. The performance (mIoU) of pseudo-masks is reported under various τ. The results are evaluated on the PASCAL VOC 2012 train set.

A.2. Effect of the Hyperparameters

In this section, we analyze how two hyperparameters, λ and τ, affect the quality of the pseudo-masks. From Equation (1), we used λ to adjust the ratio between the foreground map M_{fg} and the background map M_{bg} in computing the estimate of saliency map $\hat{M}_s$. Specifically, the larger the λ, the more the foreground map affects estimating the saliency map. (When λ is set to 1.0, only the foreground map affects estimating the saliency map.) Table A.1 shows the quality of the pseud-masks on PASCAL VOC 2012 depending on λ. These results indicate that the best performance is achieved at 0.75, and the performance is robust as long as the value is within $0.5 < \lambda < 1.0$. This implies that relying more on the foreground map is generally a good choice, but the overall performance is not sensitive to the choice of λ. For all the experiments in this paper, we set λ to 0.5 for PASCAL VOC 2012 (not optimal but sufficiently good) and 0.9 for MS COCO 2014 in our experiments.

From Equation (2), τ is defined as the threshold to control whether the localization map should be assigned to the

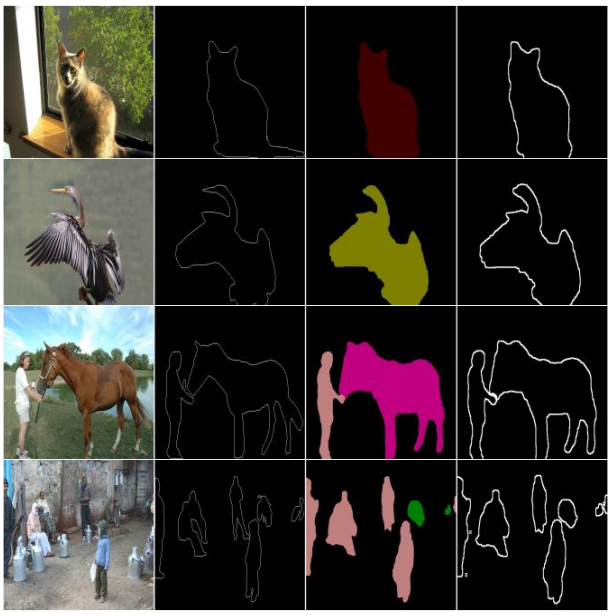
Class	Naïve	Pre-defined	Our adaptive
bkg	89.3	89.8	90.1
aero	78.9	80.1	82.9
bike	42.0	43.8	42.5
bird	79.4	79.7	80.5
boat	72.3	72.7	72.8
bottle	70.2	70.4	69.4
bus	79.9	80.3	82.1
car	75.8	75.3	78.9
cat	78.2	77.9	82.8
chair	29.6	31.0	33.1
cow	77.5	78.5	80.1
table	30.2	47.1	43.4
dog	78.7	78.9	79.6
horse	81.5	81.3	81.4
mbk	79.0	79.5	80.5
person	76.0	76.5	76.7
plant	46.8	46.1	52.0
sheep	80.6	79.8	80.8
sofa	28.6	37.6	39.9
train	71.4	71.5	77.8
tv	50.4	48.6	50.7
mean	66.5	67.9	69.4

Table A.3. Effects of different map selection strategies. The performances (mIoU) of pseudo-masks per class are evaluated on PASCAL VOC 2012 train set.

foreground or the background map. The larger the τ, the more likely the localization map is assigned to the background map. In Table A.2, we present the performance of the pseudo-masks over τ. These results clearly showcase that our method is robust against the choice of τ, especially between 0.2 and 0.5. By default, we set $\tau = 0.4$ for both PASCAL VOC 2012 and MS COCO 2014 in our experiments.

A.3. Effect of the Map Selection Strategy

In Section 5.2, we showed the effectiveness of our map selection strategy. For the detailed analysis, we evaluate the per-class IoU of the pseudo-masks upon map selection strategies. The results are summarized in Table A.3. As witnessed by the results in Table 3 of Section 5.2, the naïve strategy shows poor performances in several classes (*i.e. chair*, *dining table*, and *sofa*). Because the pre-defined class strategy excludes above mentioned classes (having poor performances when assigned to the foreground map), it performs better than the naïve strategy. However, it requires a manual selection and thus less practical. Lastly, our adaptive method generally achieves more accurate results without any manual interactions.



(a) (b) (c) (d)

Figure A.1. Boundary quality of pseudo-masks. We qualitatively evaluate our method in terms of the boundary quality on SBD trainval set. (a) Input images, (b) groundtruth, (c) pseudo-masks from our EPS and (d) boundary maps from our EPS.

A.4. Examples of Boundary Maps

To show the boundary quality of pseudo-masks, we focus on the object boundaries for evaluation. We compare boundary maps extracted from the pseudo-masks. Specifically, we obtain the boundary maps for all target objects by applying the Laplacian edge detector, and then aggregate them for generating a class-agnostic boundary map. Figure A.1 shows the results of boundary maps and their corresponding pseudo-masks from our EPS. These results show that our EPS is more effective to learn the boundaries.

A.5. Analysis on the Co-occurrence Problem

In Section 5.1, we showed that our method resolves the co-occurrence problem. We quantitatively analyze 1) which background classes frequently occur with the target objects and 2) whether or not our EPS alleviates the co-occurrence problem in additional co-occurring pairs.

First, in order to analyze which background pixels are coincident with target objects, we utilize the PASCAL-CONTEXT dataset, which has thorough labels of PASCAL VOC 2011 dataset. We adopt the labels (*e.g., sky, water, and road* more than 400) of PASCAL-CONTEXT and use those labels to compute the co-occurring frequencies between the target classes and the background pixel in PASCAL VOC dataset. (Here, we term the additional classes only in the PASCAL-CONTEXT dataset as context classes.) As shown in Figure A.2, we calculate the co-occurrence frequencies

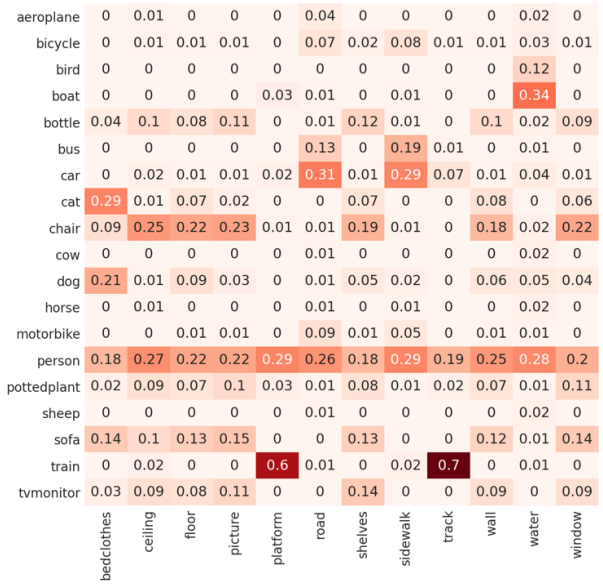


Figure A.2. Co-occurrence matrix between labels of PASCAL VOC 2012 and those of PASCAL-CONTEXT. Each entry represents the co-occurrence ratio for the appearance of the target label in PASCAL VOC 2012.

Method	*car* w/ *road*	*car* w/ *sidewalk*	*cat* w/ *bedclothes*
CAM [55]$_{\text{CVPR'16}}$	0.18 (46.4)	0.04 (51.4)	0.12 (41.9)
SEAM [44]$_{\text{CVPR'20}}$	0.15 (54.6)	0.06 (51.1)	0.09 (62.2)
ICD [13]$_{\text{CVPR'20}}$	0.09 (58.9)	0.03 (59.1)	0.14 (74.0)
SGAN [50]$_{\text{ACCESS'20}}$	0.10 (41.9)	0.03 (45.7)	0.04 (68.3)
Our EPS	0.04 (69.7)	0.01 (72.1)	0.06 (76.3)

Table A.4. Comparison with representative existing methods handling the co-occurrence problem. Here, we present additional pairs which are not covered in the main text. Each entry is $m_{k,c}$ in blue (the lower the better) and IoU in the bracket (the higher the better).

between the target classes and the context classes. We choose 12 context classes which have target classes with thg high co-occurrence frequencies and a low standard deviations. (A low standard deviation indicates that the context label does not co-occur with a particular target label.) Note that each entry in Figure A.2 represents a ratio of images in which co-occurrence occurs among images of target class. Note that we exclude *dining table* because it is divided into more detailed labels in the PASCAL-CONTEXT dataset. Finally, we choose three pairs of classes used in Section 5.1 (*train* and *railroad*, *train* and *platform*, and *boat* and *water*) because they show highest co-occurring frequencies. Additionally, *car* and *road*, *car* and *sidewalk*, and *cat* and *bedclothes* also co-occur. In Table A.4, we present that the confusion ratio and the class IoU of the additional co-occurring pairs. Our method still shows the lowest confusion ratio except for the pair, *cat* and *bedclothes* and the highest IoU. This demonstrates that our method effectively handles the co-occurrence problem.

Method	Sup.	val	test
AffinityNet [2]$_{\text{CVPR'18}}$	I.	61.7	63.7
SEAM [44]$_{\text{CVPR'20}}$	I.	64.5	65.7
Our EPS	I.+S.	**68.9**	**70.0**

Table A.5. Segmentation results (mIoU) on PASCAL VOC 2012. All the results are based on ResNet38.

Class	our EPS (VGG16)	our EPS (ResNet38)	our EPS (ResNet101)
bkg	90.7	91.4	91.7
aero	86.1	87.3	89.4
bike	34.5	38.4	40.6
bird	82.8	85.9	84.7
boat	65.3	65.4	67.0
bottle	65.6	72.0	71.6
bus	82.4	86.5	87.8
car	77.3	79.2	82.7
cat	83.0	86.1	87.4
chair	30.1	31.2	33.6
cow	73.9	75.1	81.9
table	40.1	36.1	37.3
dog	77.6	80.9	82.5
horse	74.4	76.9	82.9
mbk	70.6	74.5	76.6
person	78.9	80.6	82.8
plant	46.7	55.0	54.0
sheep	75.2	76.8	79.7
sofa	36.2	38.5	39.1
train	81.4	82.1	85.4
tv	45.3	47.2	51.7
mean	66.6	68.9	71.0

Table A.6. Per-class segmentation results (mIoU) on PASCAL VOC 2012 val set.

A.6. Additional Segmentation Network

In Section 5.3, we evaluate the segmentation accuracies on the four segmentation backbone networks. Here, we adopt an additional backbone network *i.e.*, ResNet38 [48] for the segmentation model to compare with SEAM [44]. As shown in Table A.5, EPS outperforms the AffinityNet [2] and SEAM with a large margin and this demonstrates the superiority of EPS.

A.7. Per-class Performance

For PASCAL VOC 2012 validation set and test set, Table A.6 and Table A.7 show the per-class IoU of the segmentation results over various backbone models (*i.e.*, VGG16, ResNet38 and ResNet101). Here, we adopt DeepLab-LargeFOV (V1) as a segmentation network. When our EPS employs a strong backbone model, *i.e.*,

Class	our EPS (VGG16)	our EPS (ResNet38)	our EPS (ResNet101)
bkg	91.1	91.6	91.9
aero	85.3	87.7	89.0
bike	36.9	37.8	39.3
bird	84.5	87.1	88.2
boat	54.2	59.5	58.9
bottle	64.0	68.7	69.6
bus	82.5	84.1	86.3
car	79.2	80.1	83.1
cat	84.9	83.9	85.8
chair	31.7	34.0	35.0
cow	69.4	73.8	83.6
table	46.3	43.1	44.1
dog	80.2	80.8	82.4
horse	73.2	80.1	86.5
mbk	79.1	79.0	81.2
person	78.5	80.1	80.8
plant	56.0	63.0	56.8
sheep	81.5	83.0	85.2
sofa	44.1	47.4	50.5
train	77.0	76.4	81.2
tv	45.4	48.2	48.4
mean	67.9	70.0	71.8

Table A.7. Per-class segmentation results (mIoU) on PASCAL VOC 2012 test set.

ResNet101, it tends to show the best performance in both validation and test sets.

A.8. Qualitative Examples

Figure A.3 shows more examples of estimated saliency maps from our EPS. Despite the noisy of the saliency maps, the estimated saliency maps capture the full extent of the target objects. But, it fails at capturing indistinguishable objects and captures confusing objects. Figure A.4 and Figure A.5 show more examples and failure cases of pseudo-masks and segmentation results from our EPS on PASCAL VOC 2012. Our method effectively addresses the three challenges of WSSS, but there are some failure cases: 1) shape bias, 2) confusing objects, and 3) indistinguishable shape. Figure A.6 shows more examples and failure cases of segmentation results in our EPS on MS COCO 2014. Our method shows fine results, but it fails at separating the confusing objects or capturing indistinguishable object, and hardly captures the small objects in images with a large number of objects.

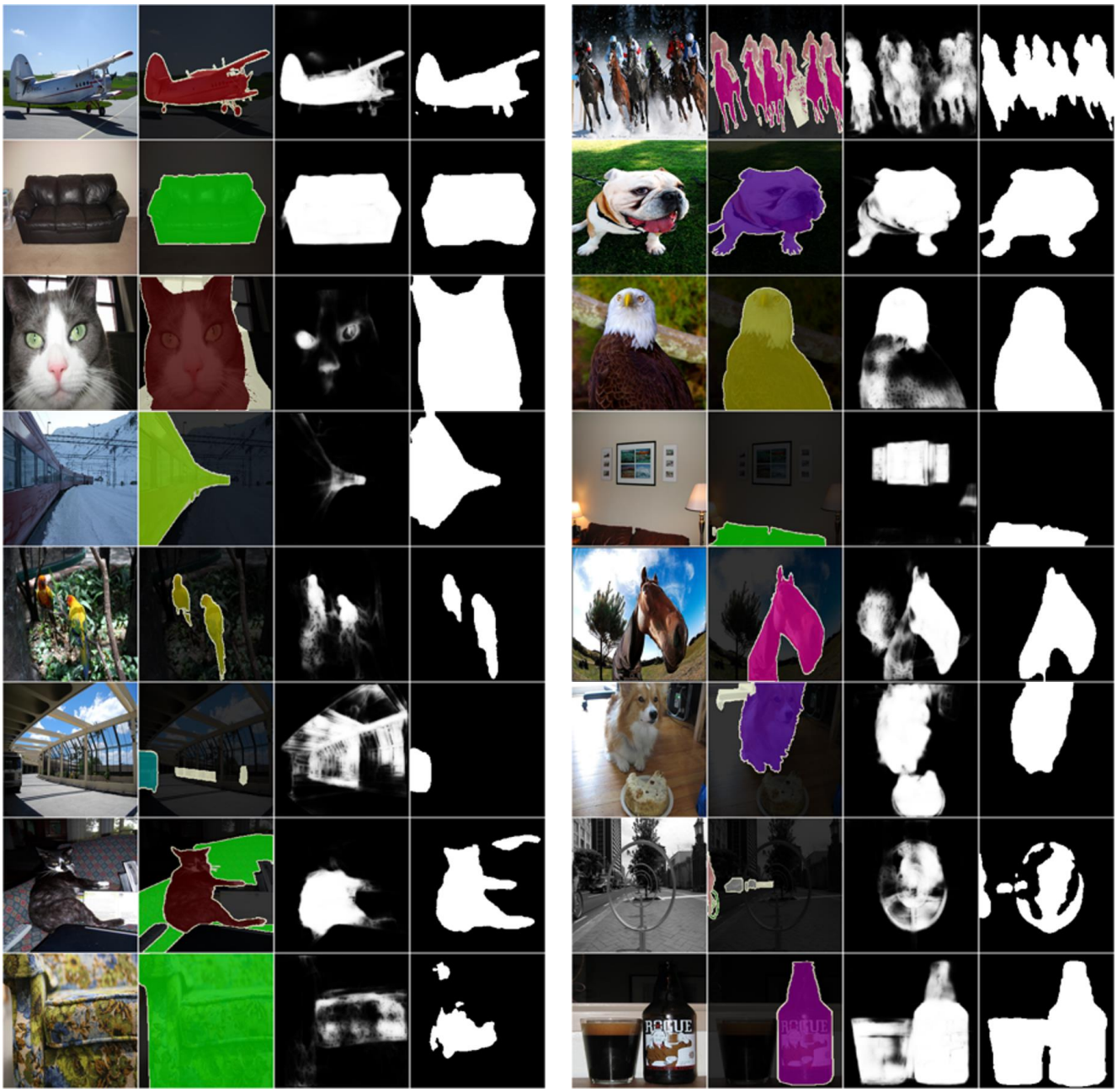


Figure A.3. Quality of estimated saliency maps on PASCAL VOC 2012. (a) Input images, (b) groundtruth, (c) saliency maps from PFAN [54] and (d) our estimated saliency maps. Our estimated saliency maps contain the target objects well along the saliency map (1st-2nd rows). Although the saliency maps have missing and noisy information, our results successfully restore missing objects (3rd-4th rows) and remove the noise (5th-6th rows). The last two rows show some failure cases.

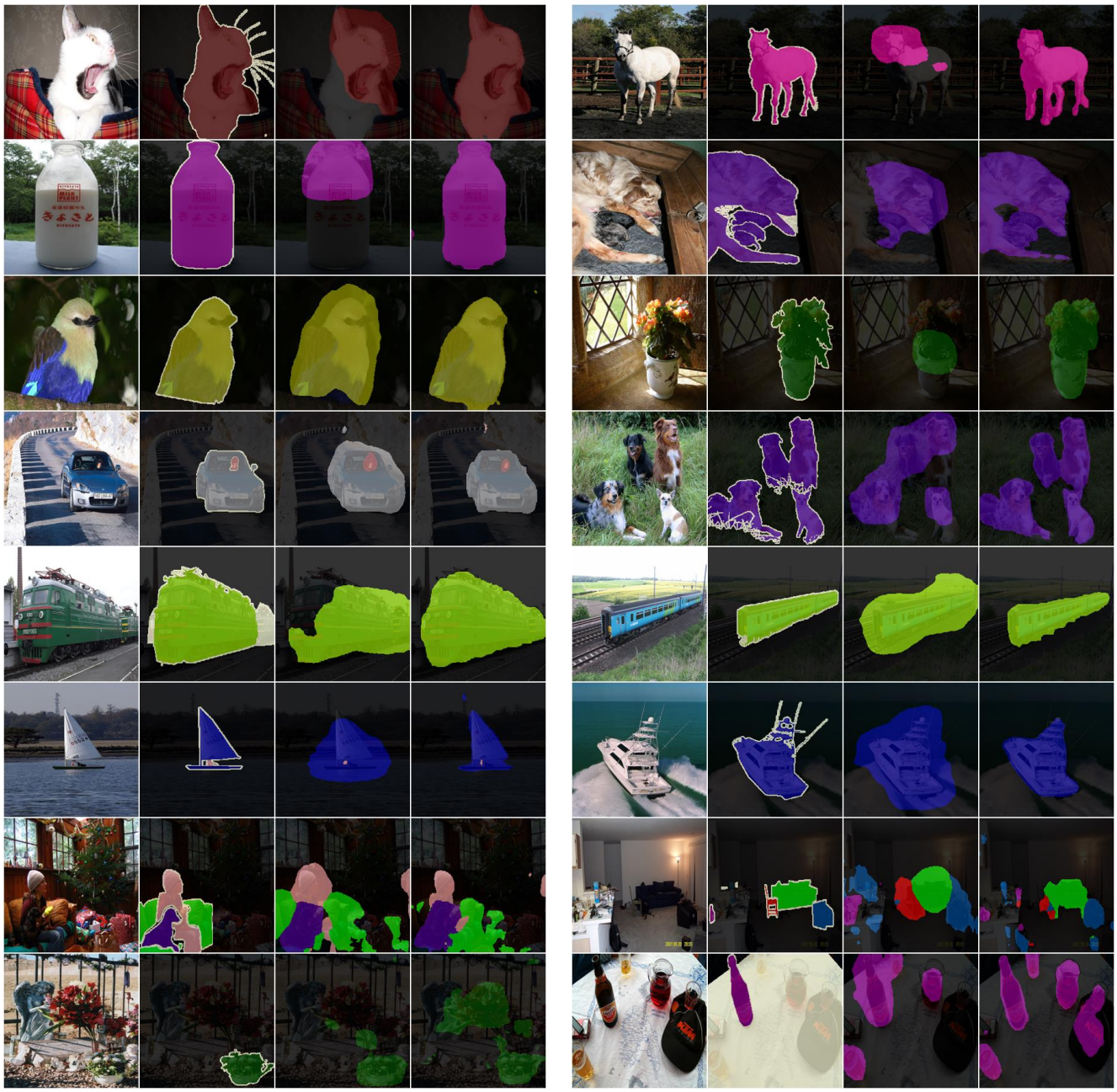


Figure A.4. Quality of pseudo-masks resolving the key challenges of WSSS on PASCAL VOC 2012 train set. (a) Input images, (b) groundtruth, (c) CAM and (d) our EPS. Our EPS effectively addresses the three challenges of WSSS: 1) sparse object coverage (1st-2nd rows), 2) boundary mismatch (3rd-4th rows), and 3) co-occurrence problem (5th-6th rows). The last two rows show some failure cases of pseudo-masks.

Figure A.5. Qualitative examples of segmentation on PASCAL VOC 2012. (a) Input images, (b) groundtruth, (c) our EPS with USPS [34] (the unsupervised saliency detection model) and (d) our EPS with PFAN [54] (the fully supervised saliency detection model). In each cases, both results are close to each other and do not differ much from the groundtruth. In general, the results with PFAN performs slightly better than the results with USPS. However, in some cases, the results with USPS are better (*e.g.*, a person's arm or leg). The last three rows show some failure cases: 1) shape bias (*e.g.*, *tv/monitor*), 2) confusing objects (*e.g.*, *cow* and *horse*, *chair* and *sofa*), and 3) indistinguishable shape (*e.g.*, *sofa* on which a *cat* sits).

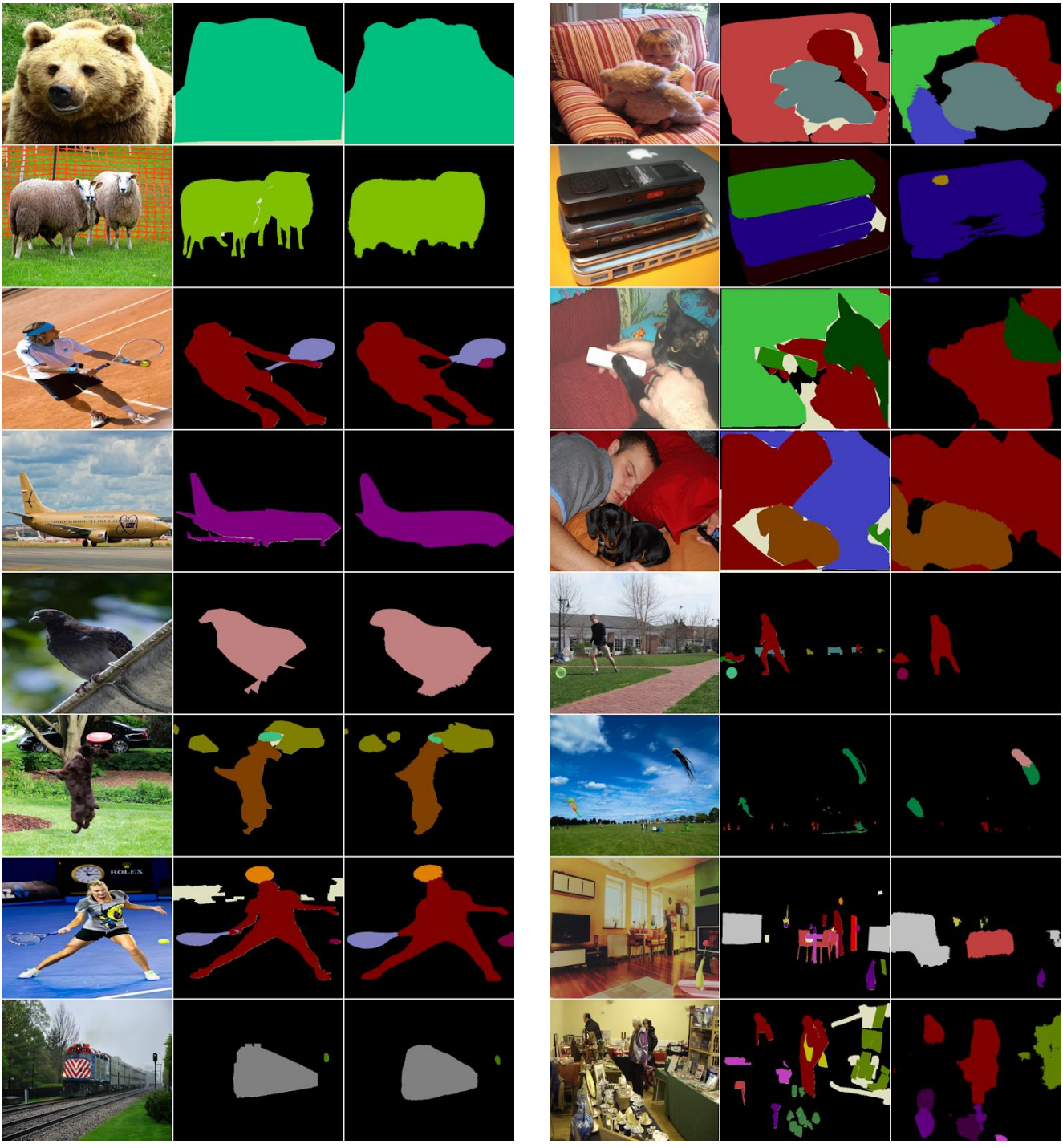


Figure A.6. Qualitative examples of segmentation on MS COCO 2014. (a) Input images, (b) groundtruth and (c) our EPS. The left column shows fine results of our EPS. The right column shows some failure cases: 1) confusing objects (*e.g.*, *couch* and *chair* and *remote* and *cell phone*), 2) indistinguishable objects (*e.g.*, *couch* where a *person* sits), 3) small objects, and 4) a large number of objects. Each case is represented on each of the two rows.