Weakly Supervised Learning of Instance Segmentation with Inter-pixel Relations
—Supplementary Material—

Jiwoon Ahn
DGIST, Kakao Corp.
jyun@dgist.ac.kr

Sunghyun Cho*
DGIST
scho@dgist.ac.kr

Suha Kwak*
POSTECH
suha.kwak@postech.ac.kr

This supplementary material provides contents omitted in the main paper due to space limitation. Section 1 describes the centroid detection algorithm (Section 5.1 of the main paper) in more detail, and Section 2 introduces the instance and semantic segmentation models trained with our synthetic labels for the final evaluation. Additional qualitative results are then presented in Section 3.

1. Details of the Centroid Detection Algorithm

As discussed in Section 5.1 of the main paper, a small group of neighboring pixels, instead of a single coordinate, are considered as a centroid in practice. To this end, we first identify pixels whose displacement vectors in $\mathcal{D}$ have magnitudes smaller than a certain threshold, and consider them as candidate centroids. Specifically, the set of candidate centroids are defined as:

$$\mathcal{C} = \left\{ \mathbf{x} \mid \|\mathcal{D}(\mathbf{x})\|_2 < 2.5 \right\} = \hat{\mathcal{C}}_1 \cup \hat{\mathcal{C}}_2 \cup \cdots \cup \hat{\mathcal{C}}_K, \quad (1)$$

where $\hat{\mathcal{C}}_i$ is a connected component of pixels in $\mathcal{C}$ and K is the number of connected components. Then a class-agnostic instance map I is obtained by assigning each pixel a connected component index in the following manner:

$$I(\mathbf{x}) = k, \ \text{ if } \left(\mathbf{x} + \mathcal{D}(\mathbf{x})\right) \in \hat{\mathcal{C}}_k, \ \ \forall \mathbf{x}. \quad (2)$$

2. Details of Our Segmentation Networks

As our framework aims to generate synthetic labels for instance and semantic segmentation, we evaluated the efficacy of our framework by learning fully supervised models for the two tasks with our synthetic labels. Specifically, we adopt Mask R-CNN [5] for instance segmentation and DeepLab v2 [1] for semantic segmentation. Both of them are first pretrained on ImageNet [2] then finetuned with the synthetic labels instead of groundtruth segmentation masks. The rest of this section describes details of the two models.

*Co-corresponding authors.

2.1. Mask R-CNN for instance Segmentation

We use Detectron [4], which is the official implementation of [5], to implement Mask R-CNN [5] with ResNet-50-FPN [8] as its backbone. We directly adopt the default training setting given in the provided source code, except the number of training steps that is adjusted for better adaptation to the PASCAL VOC 2012 dataset [3].

2.2. DeepLab v2 for Semantic Segmentation

We manually implement DeepLab v2 [1] in PyTorch [9]. Its architecture consists of ResNet-50 [6] followed by an atrous spatial pyramid pooling module [1]. The training setting of ours is identical to that of the original model. We also employ the ensemble of multi-scale prediction during evaluation. Specifically, a single input image is converted to a set of 8 images through resizing with 4 different scales $\{0.5, 1.0, 1.5, 2.0\}$ and horizontal flip, and fed into the segmentation network so that the 8 outputs are aggregated by pixel-wise average pooling.

We also reproduce the performance of the fully supervised DeepLab v2, which is the *upperbound* our segmentation model can achieve. Note that, as summarized in Table 4 of the main paper, *upperbound* we measured is lower than the performance reported in the original paper [1] as we did not tune the parameters of dense CRF [7] carefully. Thanks to the accurate segmentation labels synthesized in our framework, the DeepLab trained with our synthetic labels achieves 89.4% of its fully supervised one on the PASCAL VOC 2012 *test* set.

3. More Qualitative Results of Our Approach

In this section, we provide additional qualitative results of our framework on the PASCAL VOC dataset. Although IRNet is trained with image-level supervision only, it successfully finds accurate class boundary and displacement field to instance centroids which are not directly available in CAMs, and synthesizes accurate instance segmentation masks from CAMs incorporating those two additional in-

formation as illustrated in Figure 1.

Figure 2 and Figure 3 show additional instance segmentation and semantic segmentation results of our models, respectively. Thanks to synthetic labels that are able to differentiate attached instances, our models not only find fine object shape, but also detect independent instances that are adjacent and of the same class.

References

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. 1

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 1

[3] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010. 1

[4] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He. Detectron. https://github.com/facebookresearch/detectron, 2018. 1

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770 – 778, 2016. 1

[7] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 109–117, 2011. 1

[8] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[9] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *AutoDiff, NIPS Workshop*, 2017. 1

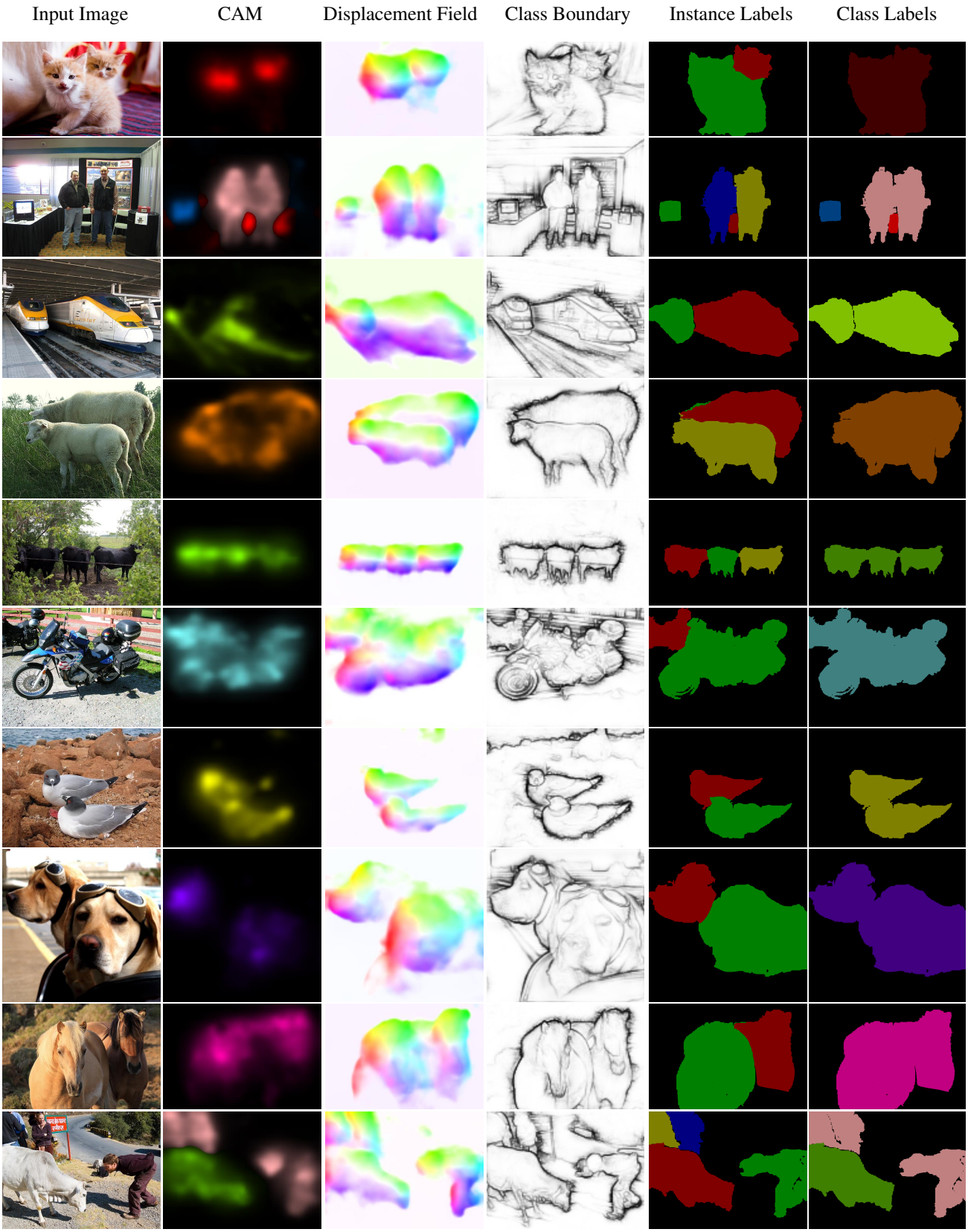

Figure 1. Qualitative results of our instance segmentation model on the PASCAL VOC 2012 *train* set.

Figure 2. Qualitative results of our instance segmentation model on the PASCAL VOC 2012 *val* set.

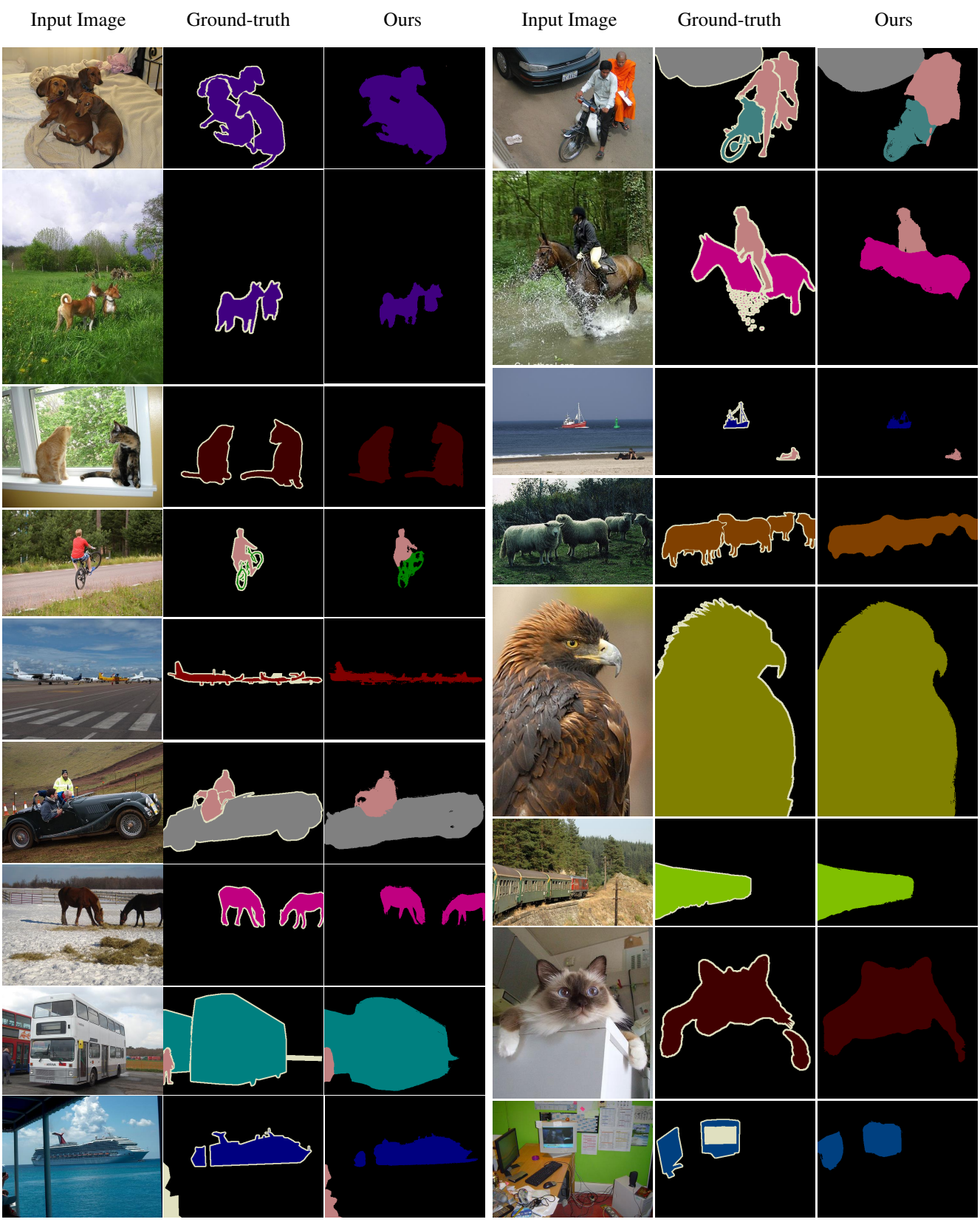

Figure 3. Qualitative results of our semantic segmentation model on the PASCAL VOC 2012 *val* set.