Appendix for "Bridging the Gap between Classification and Localization for Weakly Supervised Object Localization"

Eunji Kim[1] Siwon Kim[1] Jungbeom Lee[1] Hyunwoo Kim[2] Sungroh Yoon[1,3*]
[1] Department of Electrical and Computer Engineering, Seoul National University [2] LG AI Research
[3] Interdisciplinary Program in AI, AIIS, ASRI, INMC, and ISRC, Seoul National University
{kce407, tuslkkk, jbeom.lee93}@snu.ac.kr, hwkim@lgresearch.ai, sryoon@snu.ac.kr

A. Societal Impact

As deep neural networks require large amounts of data, data-related industries are expanding. One of the prevalent business models of the industries is data annotation. However, the cost of data annotation is burdensome for general users. To reduce the cost, approaches for weakly supervised learning have been proposed, which only requires weaker supervision than fully supervised learning. Since we propose a method of weakly supervised object localization, image-level annotation is sufficient. Our method may threaten the business model of data labeling companies that provide fine-grained labels such as pixel-level annotations and bounding box annotations.

B. Experimental Details

We employ SGD optimizer with momentum 0.9 and weight decay 5×10^{-4}. Following the work of Choe *et al*. [2], the networks are divided into two parts, and the learning rate is set differently for those two. VGG16 [8] is divided into old layers and newly added layers when modifying it to VGG16-GAP [11], and ResNet50 [4] is divided into the layers prior to the fourth layer and the others. On the CUB-200-2011 dataset [9], the learning rate is set to 4×10^{-3} and 2×10^{-3} for the former part of VGG16 and ResNet50, respectively. It is set to 2×10^{-2} for the latter part of both backbones. On the ImageNet-1K dataset [7], the learning rate is set to 2×10^{-5} and 1×10^{-5} for the former part of VGG16 and ResNet50, respectively. It is set to 1×10^{-4} for the latter part of both backbones. The proposed method is implemented using PyTorch [6].

Following the work of Choe *et al*. [2], we use `train-fullsup` [2] of each dataset as a validation set to select the best model for `MaxBoxAccV2` [2] scores. Please refer to the work of Choe *et al*. [2] for the details of the dataset.

*Correspondence to: Sungroh Yoon (sryoon@snu.ac.kr).

C. Additional Results and Discussions

Additional results and discussions are presented to support the experimental results in the main paper.

C.1. Sensitivity to Bounding Box Threshold

A threshold is required to draw a bounding box around an object from a continuous localization map. Fig. A1 shows that the change of localization accuracy with various IoUs when varying the threshold. In each plot with IoU δ, the maximum value becomes `MaxBoxAccV2(`δ`)` score. For all δ, the curve of our method is consistently above the curve of the vanilla method, which shows that the superiority of the localization performance of our method does not depend on the threshold. When δ is 0.3 and 0.5, the curve of our method nearby a maximum value is flatter than the curve of the vanilla method. This shows that our method is less sensitive to the threshold for a bounding box than the vanilla method. When δ is 0.7, our method is sharper than the vanilla method, but the localization accuracy of our method is more than twice that of the vanilla method, so the flatness comparison is meaningless.

C.2. Feature Direction Alignment

Fig. A3 shows some examples of CAM, $\mathcal{F}$, and $\mathcal{S}$ from the vanilla method and our method on the CUB-200-2011 and ImageNet-1K datasets. In $\mathcal{S}$ from the vanilla method, the overall values are similar and some regions that belong to the object have low values. For instance, in the 'car' example (in the second row and second column of the ImageNet-1K dataset), the middle part of the object has low similarity, resulting in low activation in the CAM. Different from the vanilla method, the values of $\mathcal{S}$ from our method are high in the object regions and low in the background regions. Furthermore, the values of $\mathcal{F}$ are high across the entire object region. This makes the CAM that captures more object region.

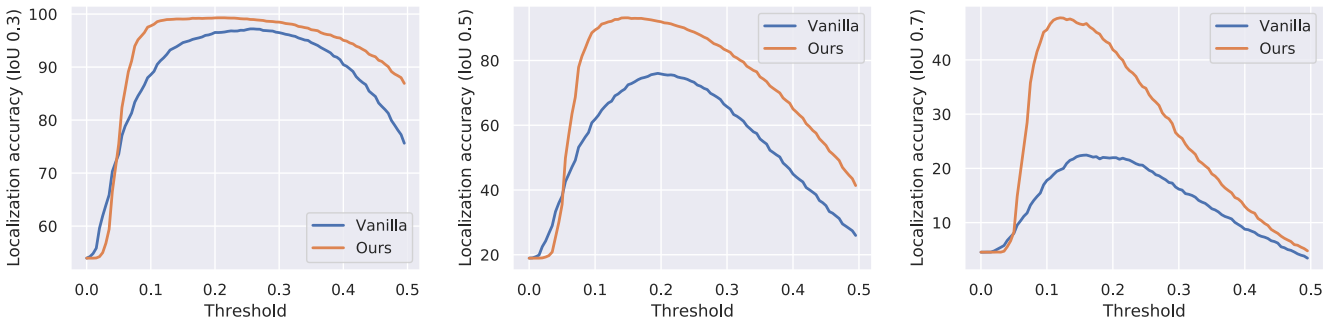

Figure A1. Comparisons of the localization performance when varying threshold on the CUB-200-2011, using VGG16 as a backbone. Three plots show the localization accuracy with IoU 0.3, 0.5, and 0.7. The maximum value of each curve is `MaxBoxAccV2` (δ).

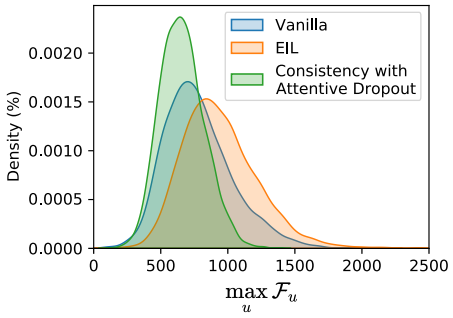

Figure A2. Comparison of density histogram on $\max_u \mathcal{F}_u$ with the vanilla method, EIL, and consistency with attentive dropout. The analyzes are performed on the CUB-200-2011 test set using VGG16 as a backbone.

C.3. Consistency with Attentive Dropout

We compare consistency with attentive dropout with EIL [5], the most recent one among the previous erasing methods [3, 5, 10], on the CUB-200-2011 using VGG16. As mentioned in the main paper, attentive dropout directly regularizes feature activation, whereas EIL indirectly influences the activation through class prediction. Fig. A2 shows the different effect on feature activation of consistency with attentive dropout and EIL. To encourage a model to predict correct class without highly activated region, EIL enhances the maximum value. In contrast, consistency with attentive dropout reduces the maximum value through regularization. As a result, attentive dropout distributes the activations more effectively than EIL (Fig. 7(b) in the main paper).

C.4. Localization Results

Fig. A4 compares the localization results from the vanilla method [11] and our method on the CUB-200-2011 and ImageNet-1K datasets. While the vanilla method misses the less discriminative parts, *e.g.*, wings and tails of birds and bodies of animals, our method successfully captures the entire object region.

Method	`PxAP`
CAM [28] CVPR '16	58.3
ADL [3] CVPR '19	58.7
CutMix [28] ICCV '19	58.1
CAM+PaS [1] ECCV '20	59.6
ADL+IVR [9] ICCV '21	59.3
CALM [10] ICCV '21	61.3
Ours	**63.7**

Table A1. Comparison of `MaxBoxAccV2` scores on the OpenImages dataset using VGG16 as a backbone.

C.5. Sensitivity to Hyperparameters

In the main paper, we mention as a limitation that there are several hyperparameters to be decided in our method. We provide further analysis with a different dataset and backbone than that presented in the main paper.

CUB-200-2011 on ResNet50. We find the best localization performance at 0.6 for λ_{sim}, 0.07 for λ_{norm}, and 2 for λ_{drop}, respectively. The thresholds τ_{fg} and τ_{bg} for $\mathcal{L}_{\text{sim}}$ are set to 0.4 and 0.2, respectively. The hyperparameters γ and p for $\mathcal{L}_{\text{drop}}$ are set to 0.8 and 0.25, respectively. Fig. A5(a) shows the change of GT Loc when varying the hyperparameters. The sensitivity to each hyperparameter is similar to that on the CUB-200-2011 dataset using VGG16 as a backbone. λ_{sim} affects the localization performance the most among hyperparameters.

ImageNet-1K on VGG16. The best localization performance is found at 0.5 for λ_{sim}, 0.2 for λ_{norm}, and 3 for λ_{drop}, respectively. The hyperparameters τ_{fg}, τ_{bg}, γ, and p are set to 0.5, 0.3, 0.8, and 0.5, respectively. Fig. A5(b) shows the GT Loc at various hyperparameter values. Different from the CUB-200-2011 dataset, the localization performance is most affected by λ_{drop}. The sensitivities to the other hyperparameters are similar to those with a different dataset and backbone.

Discussion. The hyperparameters are set differently on the two datasets. This is because their tasks are somewhat dif-

ferent; the classification on the CUB-200-2011 dataset is a fine-grained classification. We additionally evaluate our methods on OpenImages30K [1, 2], where the task is similar to that on the ImageNet-1K dataset. We use VGG16 as a backbone and set the hyperparameters the same as those on the ImageNet-1K dataset. Note that the OpenImages30K dataset is annotated with a mask, and we use `PxAP` metric for evaluation, following the work of Choe *et al*. [2]. As shown in the Tab. A1, our method outperforms the recent methods by a large margin on the OpenImages30K dataset as well, which shows the hyperparameters used on the ImageNet-1K dataset can be applied successfully to a different dataset.

References

[1] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11700–11709, 2019. 3

[2] Junsuk Choe, Seong Joon Oh, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluation for weakly supervised object localization: Protocol, metrics, and datasets. *arXiv preprint arXiv:2007.04178*, 2020. 1, 3

[3] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8766–8775, 2020. 2

[6] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 1

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[9] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 1

[10] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018. 2

[11] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 1, 2

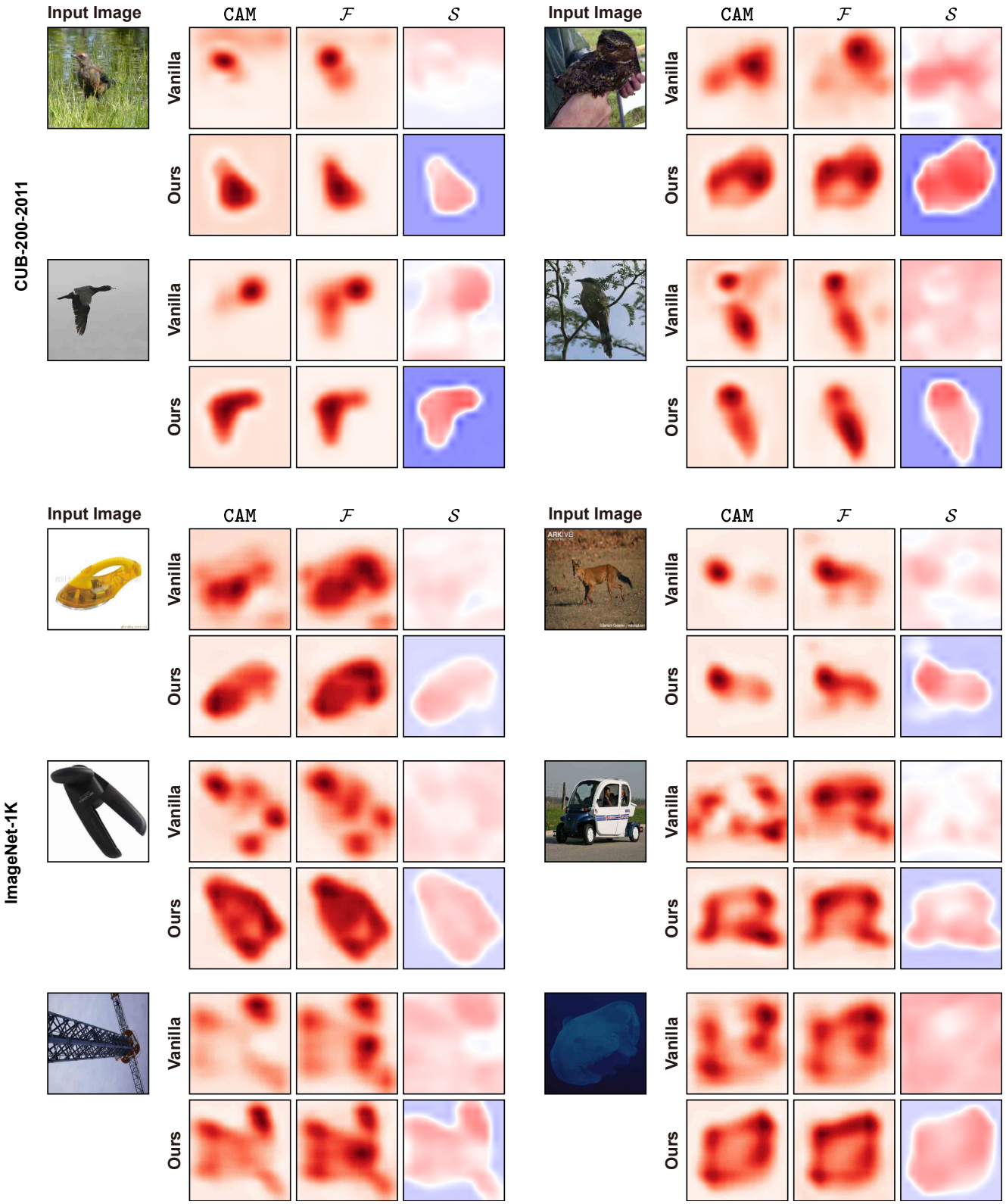

Figure A3. Comparisons of CAM, $\mathcal{F}$, and $\mathcal{S}$ between the vanilla method and our method on the CUB-200-2011 and ImageNet-1K datasets, using VGG16 as a backbone.

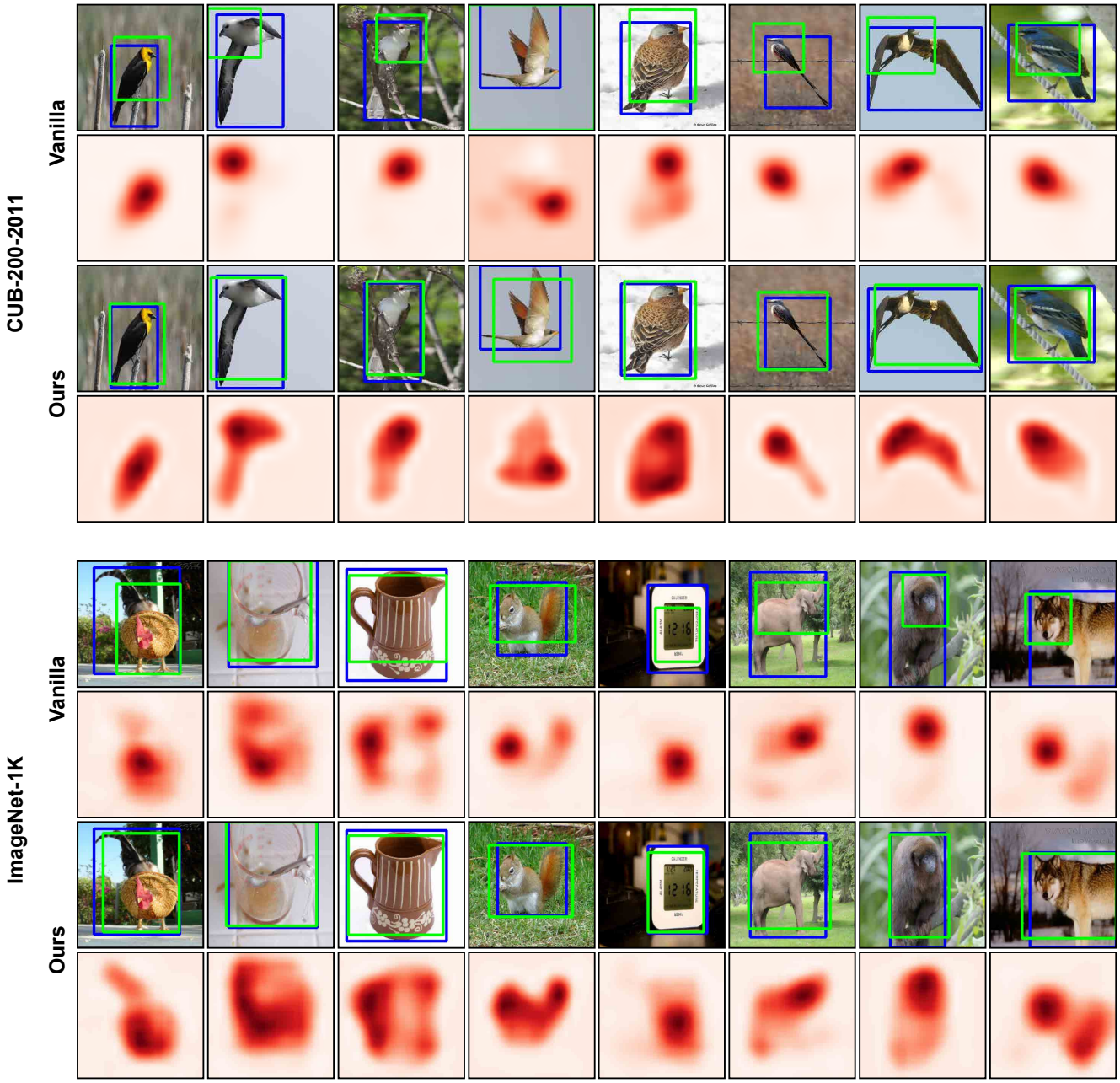

Figure A4. Comparison of localization results from the vanilla method and our method on CUB-200-2011 and ImageNet-1K datasets, using VGG16 as a backbone. Blue boxes denote the ground truth bounding boxes and green boxes denote the predicted bounding boxes.

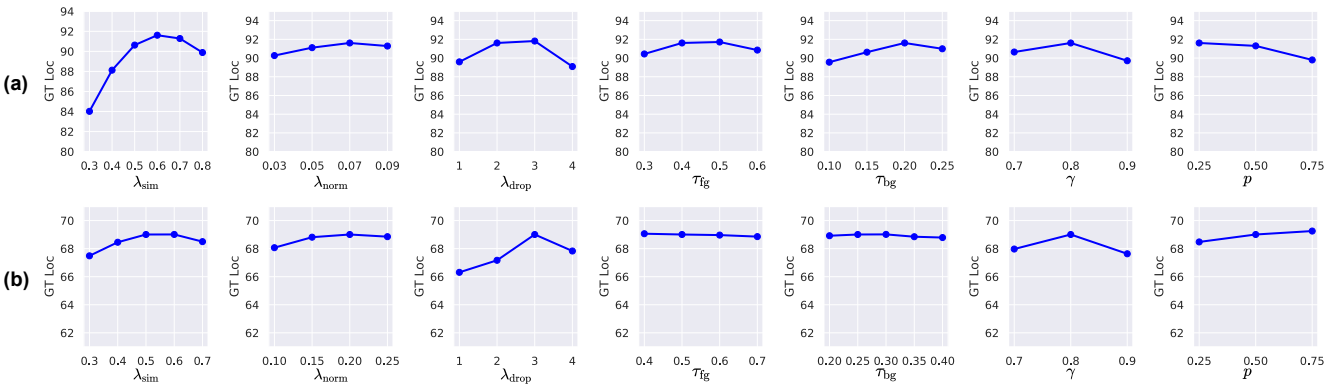

Figure A5. Effect of balancing factors for loss and various hyperparameters. The plots show the results on the CUB-200-2011 test set with ResNet50 and (b) those on the ImageNet-1K validation set with VGG16.