Supplementary Materials: Combinational Class Activation Maps for Weakly Supervised Object Localization

1. Visualization

We visualize more results on the ILSVRC [4] validation set and the CUB-200-2011 [5] test set in Fig. 1. The CAM [6] method tends to highlight small parts of the object, whereas our 1st map (M^{c_1}) tends to highlight more parts of the object compared to CAM. This is because using non-local modules helps to find relevant parts of the most discriminative parts of the object. Both CAM and the 1st map, which use the activation map of the highest probability class, highlight background regions, resulting in an inaccurate bounding box (see CAM and M^{c_1} in Fig. 1). We observe that the activation map of the lowest probability class catches background regions, *i.e.*, non-discriminative parts (see M^{c_K} in Fig. 1). To use this property effectively, we exploit a specific function to combine all activation maps. This leads the localization map to find entire object parts more accurately and suppress background regions, resulting in an accurate bounding box (see NL-CCAM in Fig. 1).

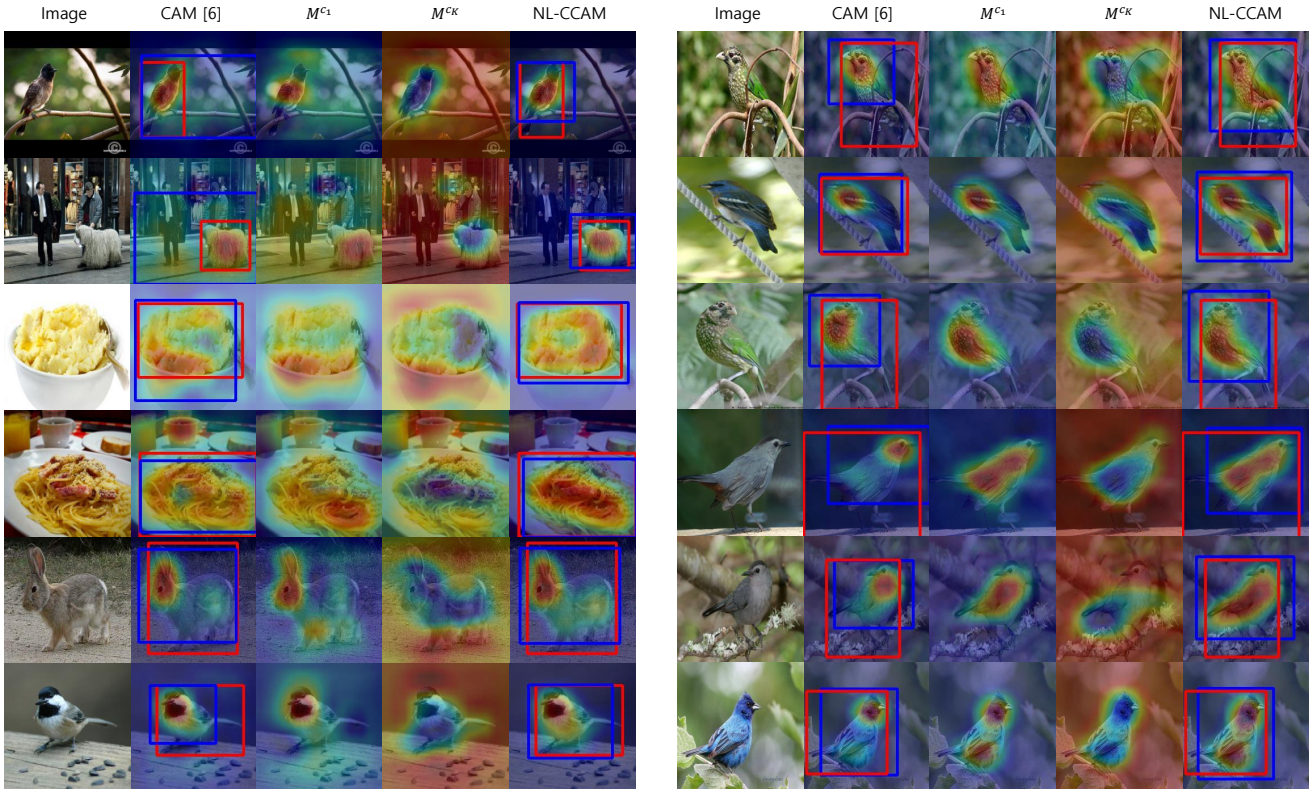

Figure 1. Qualitative object localization results compared with the CAM method on the ILSVRC and CUB-200-2011 datasets.

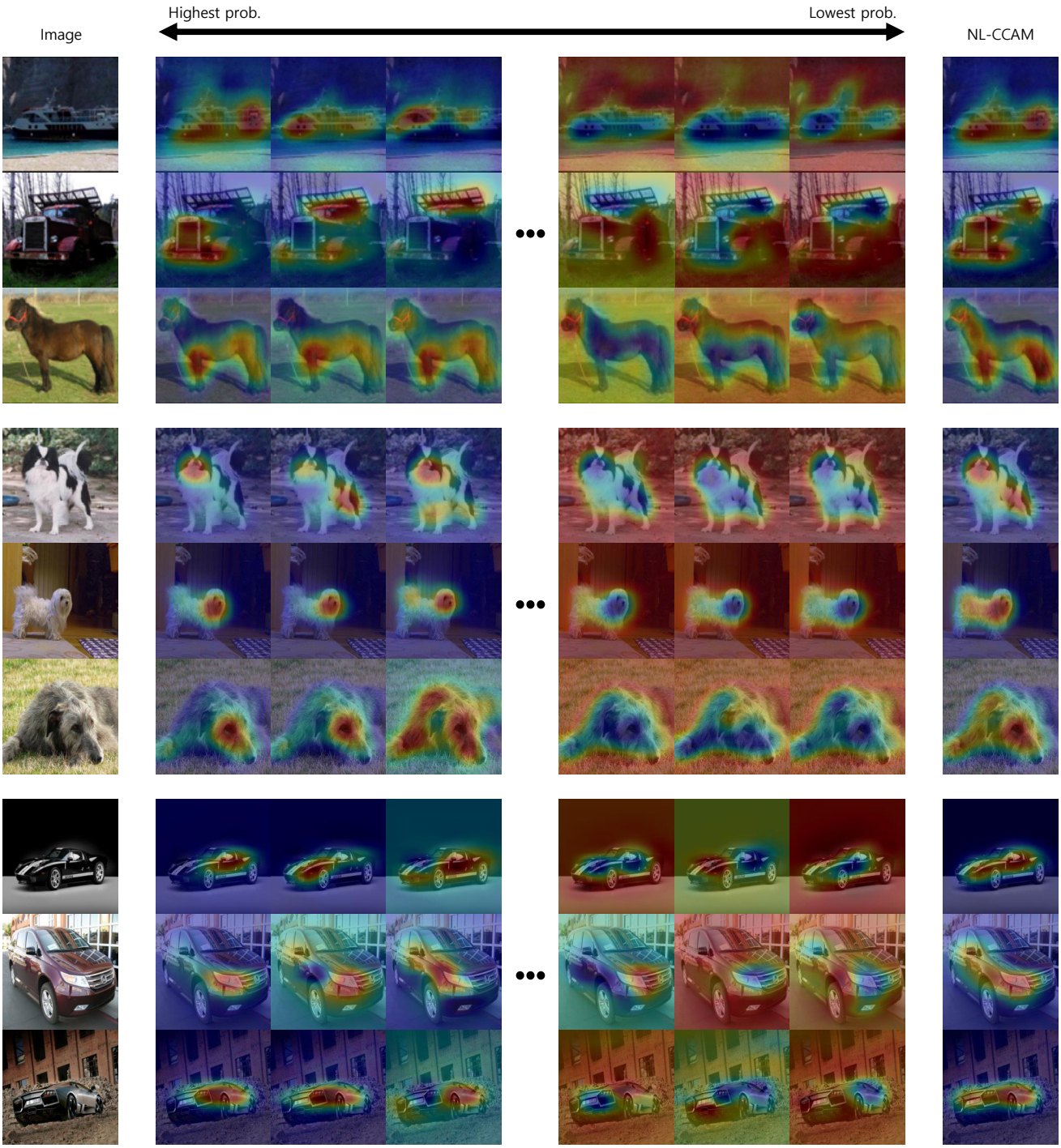

Figure 2. The activation maps of the proposed method on the STL-10, Stanford-Dogs, and Stanford-Cars datasets. The top rows are the activation maps on the STL-10 dataset, the middle rows are the maps on the Stanford-Dogs dataset, and the bottom rows are the maps on the Stanford-Cars dataset.

Furthermore, we visualize the activation maps from the highest to the lowest probability class on STL-10 [1], Stanford-Dogs [2], and Stanford-Cars [3] to prove our approach applies to various datasets. In Fig. 2, the maps of higher probability classes catch some parts of the object, while the maps of lower probability classes tend to highlight background regions.

Methods	GT-known loc. err.
VGGnet-GAP [6]	41.16
VGGnet-CCAM (ours)	31.02
NL-CCAM (ours)	**29.79**

Table 1. GT-known localization error on the CUB-200-2011 test set.

2. Localization

We compare the GT-known localization errors to eliminate the influence caused by classification results. Table 1 shows that NL-CCAM achieves 29.79% on the CUB-200-2011 dataset.

References

[1] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *PMLR*, 2011.

[2] A. Khosla, N. Jayadevaprakash, B. Yao, and F. Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPRW*, 2011.

[3] J. Krause, M. Stark, D. Jia, and F. Li. 3d object representations for fine-grained categorization.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[5] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. *Technical Report CNS-TR-2011-001*, 2011.

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.