In the paper, we propose a novel framework for WSOL problem by augmenting CAM that is generated from traditional recognition networks. The performance of our model on ILSVRC [1] and CUB-200 [4] both outperform previous methods, becoming the new state-of-the-art.

In the supplementary material, we first show the detail structure of our backbone network for appending multiple classifiers to generate CAMs. In addition, we generate more visualization results to demonstrate that the combined CAM from our framework is more complete and precise compared with each individual CAM. Furthermore, we compare our method with SPG [6], a previous method which also utilizes background parts of CAM but sets fixed thresholds in an one-size-fit-all manner. The better localization results of our method indicates that the learned sample-adapted thresholds during training perform much better than the unique value predefined.

A. Network Structure

We show the backbone networks used for our framework in Fig. 1, which are based on VGGnet [2] and GoogLeNet [3], respectively.

For VGGnet, we remove the last fully connected layer and append our two classifiers, $\mathcal{W}_{\mathcal{L}}$ and $\mathcal{W}_{\mathcal{F}}$ after fourth and last pooling layer. In addition, we change the last two pooling layers to keep the resolution of the feature map, which follows the configure in [6]. For GoogLeNet, we remove the convolutional blocks after $Mixed\_6e$ and append two classifiers after $Mixed\_6b$ and $Mixed\_6e$, respectively. For more details about our two classifiers, please refer to our paper.

B. Visualization Result

We show more powerful visualization results in Fig. 3 and Fig. 4. In most cases, the combined CAM can generate more complete and precise localization results compared with each individual CAM. Besides, we also generate more visualizations to compare our framework with SPG [6] in Fig. 2. We can see the CAM augmented by our framework can cover more precise foreground object rather than only a small discriminative part, which demonstrate the advantages of our proposed method.

References

[1] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[2] K Simonyan, A Vedaldi, and A Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. pages 1–8. ICLR, 2014.

[3] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[4] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

[5] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S. Huang. Adversarial complementary learning for weakly supervised object localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *The European Conference on Computer Vision (ECCV)*, September 2018.

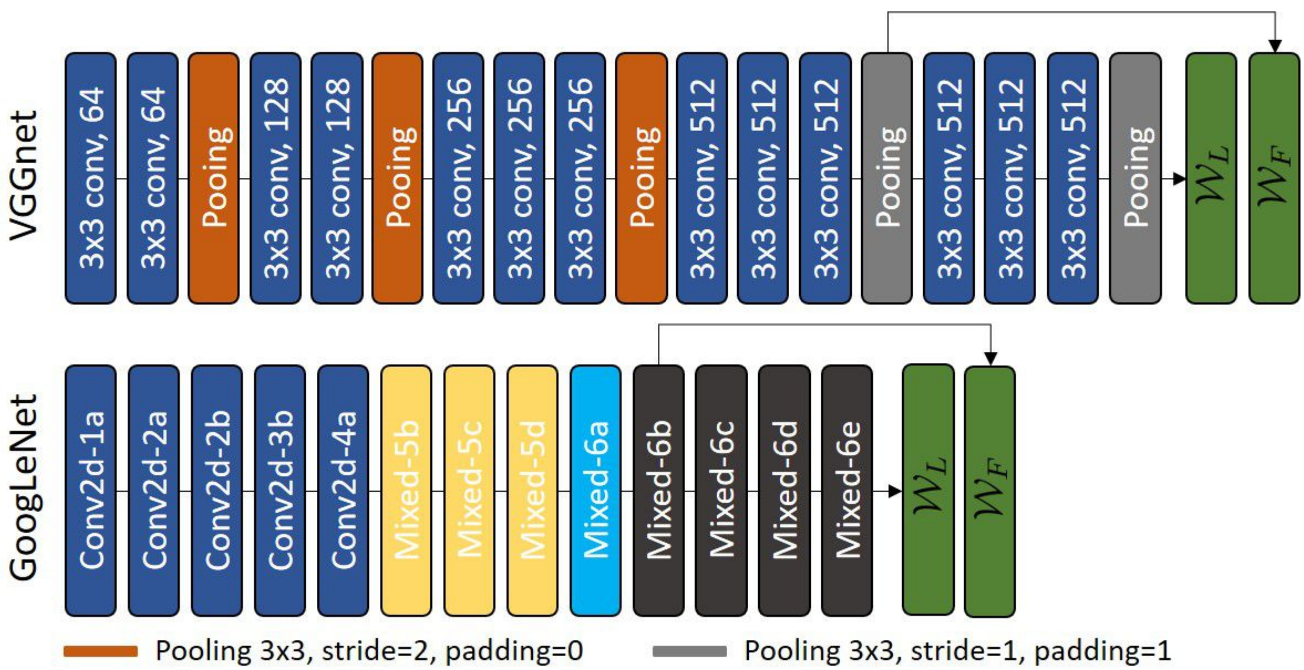

Figure 1. The structure of our backbone networks. We keep them same with SPG [6] and ACoL [5] to make a fair comparison.

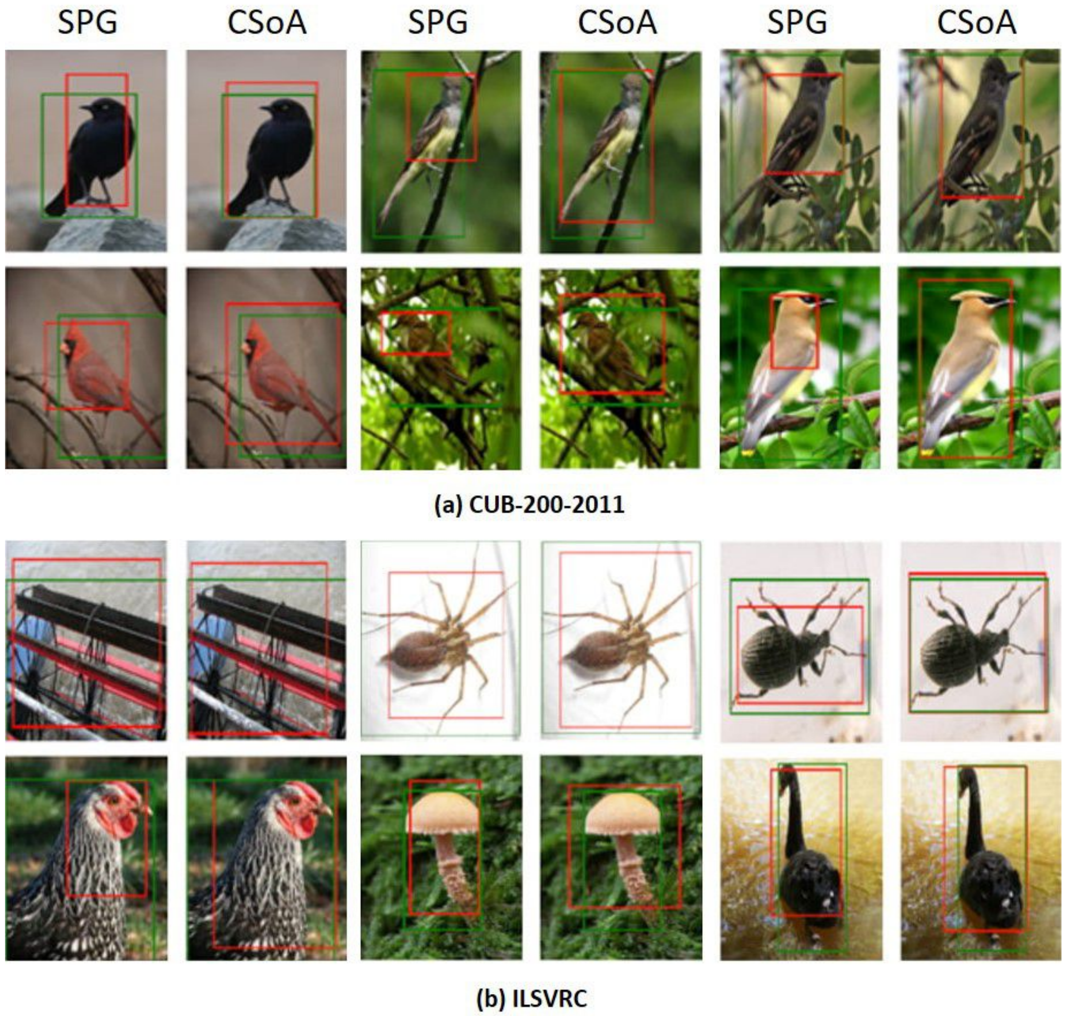

Figure 2. Comparison between our model and SPG [6]. We keep the backbone network and related configures same.

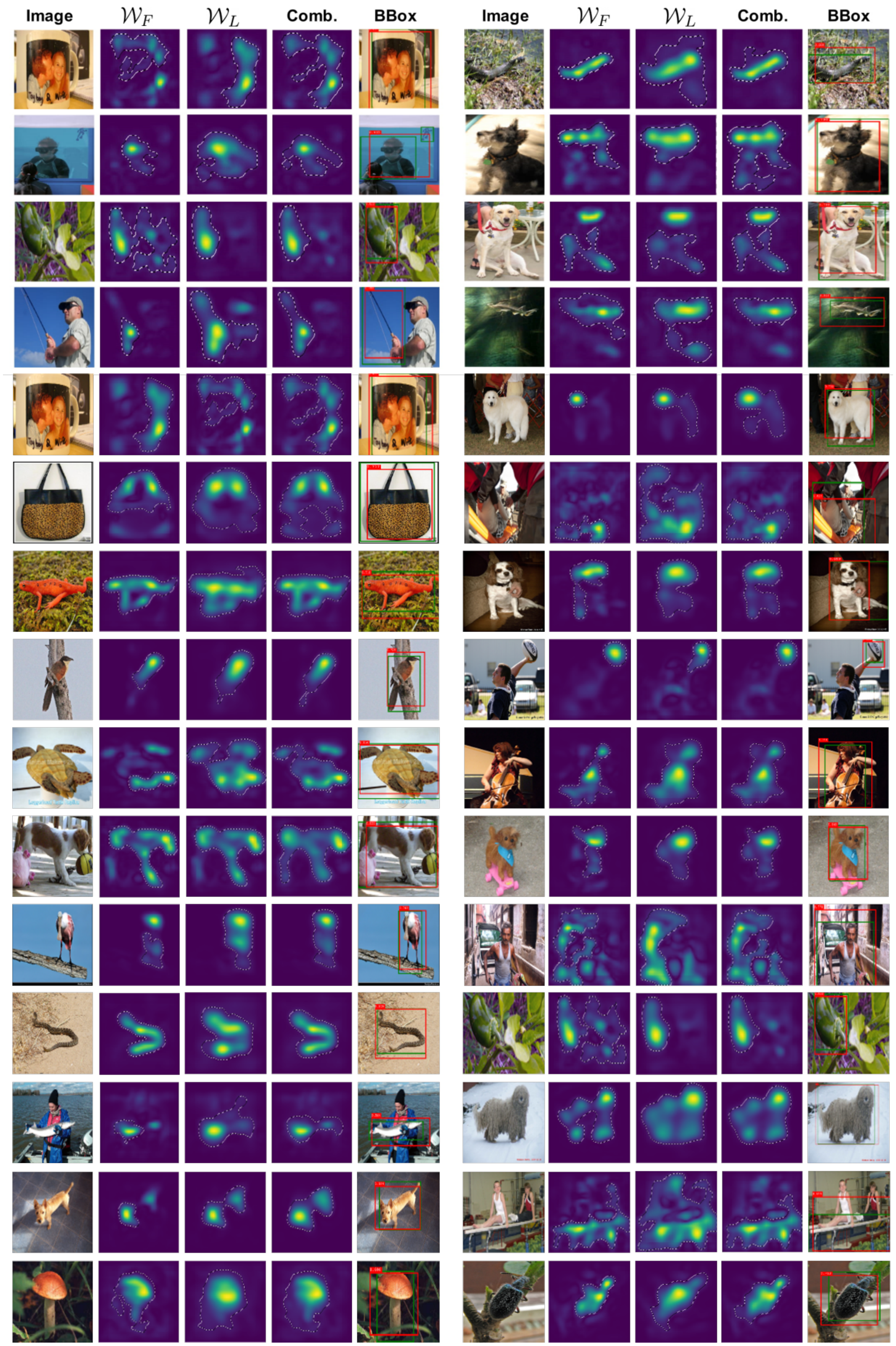

Figure 3. Visual examples from our CSoA framework on ILSVRC [1] dataset. The red box is predicted results while the green ones are ground truth labels.

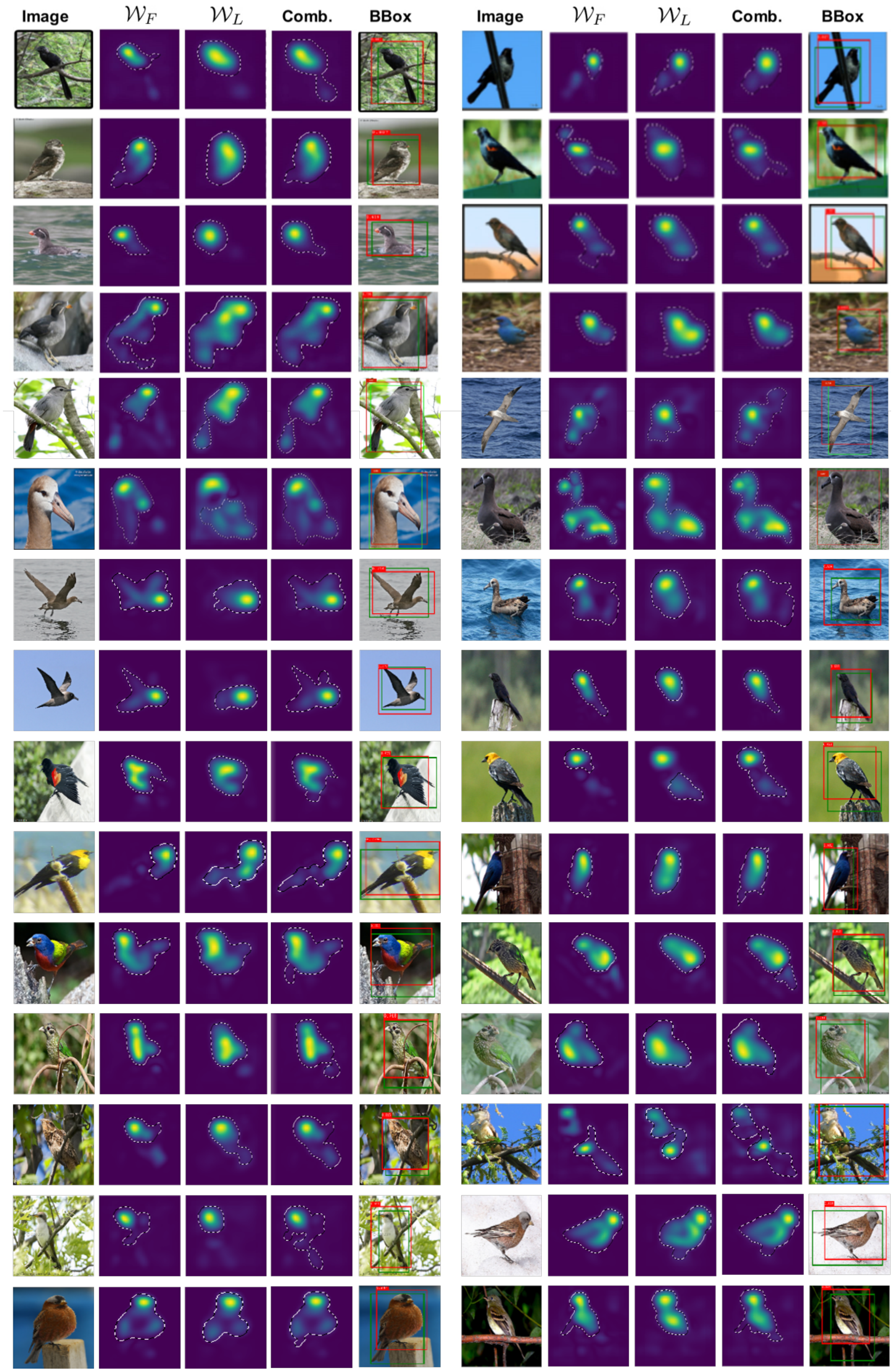

Figure 4. Visual examples from our CSoA framework on CUB-200 [4] dataset. The red box is predicted results while the green ones are ground truth labels.