Background Activation Suppression for Weakly Supervised Object Localization (Supplementary Materials)

Pingyu Wu[1,†] Wei Zhai[1,†] Yang Cao[1,2,*]
[1] University of Science and Technology of China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{wpy364755620@mail., wzhai056@mail., forrest@}ustc.edu.cn

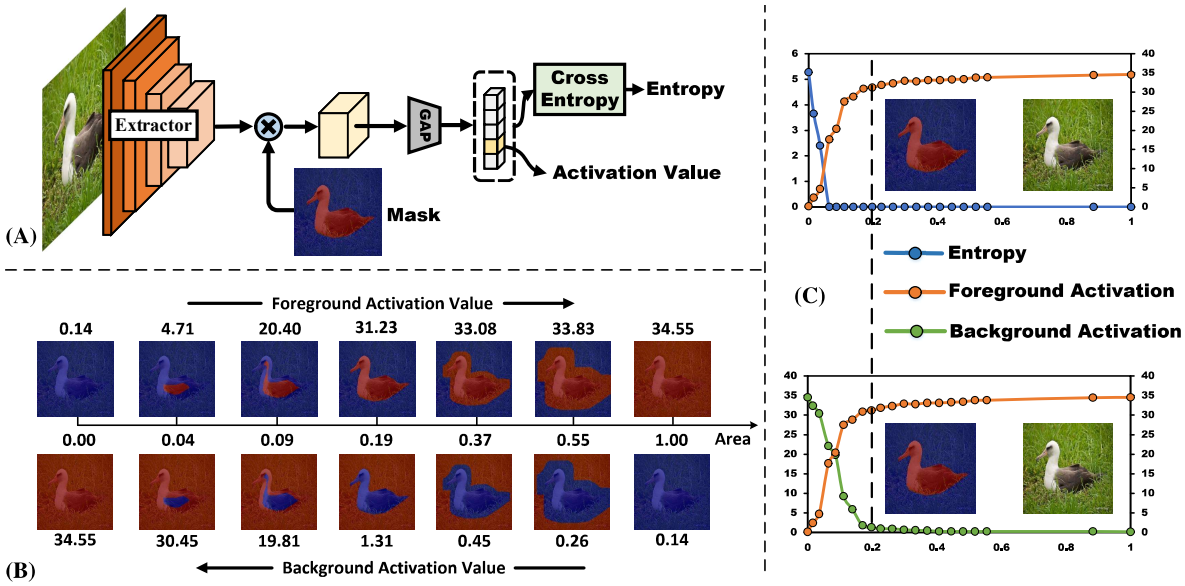

Figure 1. (A) Flow diagram of experiment. The activation value and cross-entropy corresponding to the mask are generated by masking the feature map. (B) The foreground activation value and background activation value are obtained by reserving the foreground area and background area, respectively. (C) The curve of entropy, foreground activation, and background activation with mask area, the dashed line represents the position of the ground-truth mask.

1. Exploratory Experiment

We introduce the implementation of the experiment, as shown in Fig. 1(A). For a given GT binary mask, the activation value (Activation) and cross-entropy (Entropy) corresponding to this mask are generated by masking the feature map. We erode and dilate the ground-truth mask with a convolution of kernel size $5n \times 5n$, obtain foreground masks with different area sizes by changing the value of n, and plot the activation value versus cross-entropy with the area as the horizontal axis, as shown in Fig. 1(B). By inverting the foreground mask, the corresponding background activation value for the foreground mask area is generated in the same way. In Fig. 1(C), we show the curves of entropy, foreground activation, and background activation with mask area. It can be noticed that both background activation and foreground activation value have a higher correlation with the mask compared to the entropy. We show more examples in Fig. 2 and Fig. 3 to illustrate the generality of this phenomenon. Fig. 2 reflects that there is a "**mismatch**" between entropy and ground-truth mask, while the activation

*Corresponding author. † Equal contributions.

value tends to "**saturate**" when mask expands to the object boundary. In Fig. 3, we compare the foreground and background activation curves, which show a "**symmetry**", indicating that using the background activation value to learn generator is equally effective.

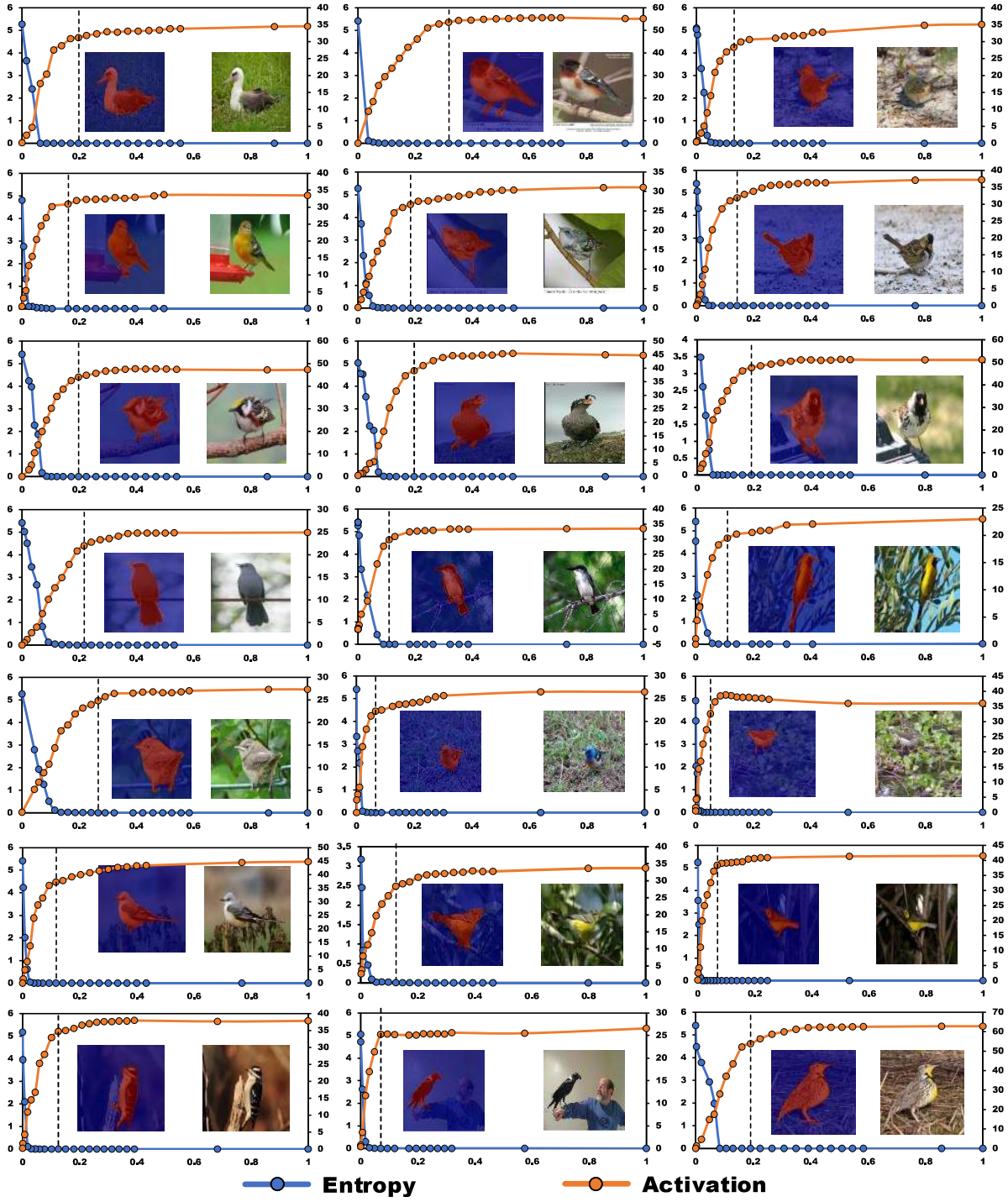

Figure 2. Examples about the entropy value of CE loss $w.r.t$ foreground mask and foreground activation value $w.r.t$ foreground mask.

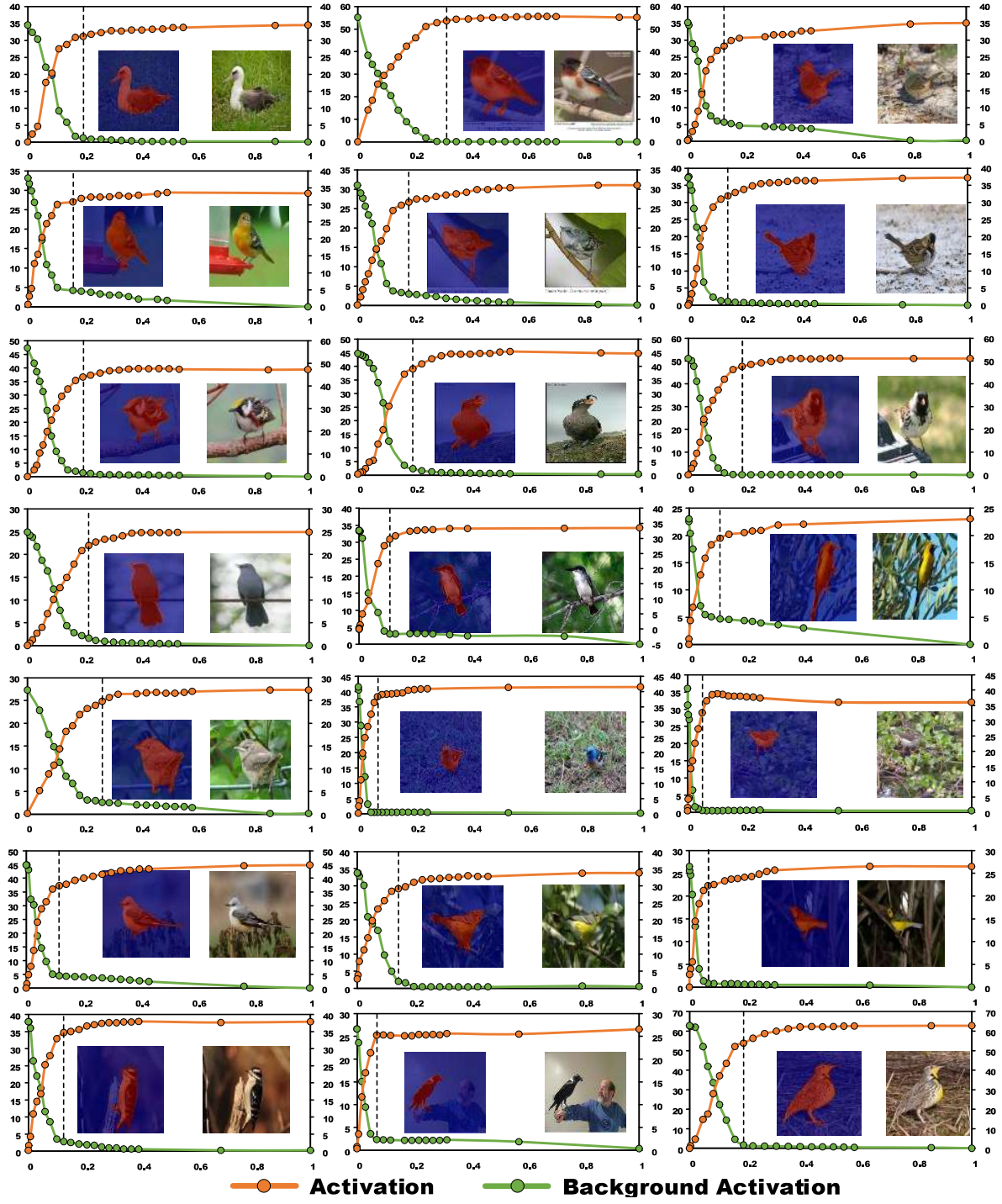

Figure 3. Examples about foreground activation value $w.r.t$ foreground mask and background activation value $w.r.t$ foreground mask.

2. Experiment

2.1. Hyperparameter

Hyperparameter α in total loss. α denotes the factor of $\mathcal{L}_{FRG}$. Foreground region guidance loss can guide the activation map learning to the approximate location, which is necessary when the backbone is ResNet50, MobileNetV1, and InceptionV3, but is not required for VGG16. As shown in Table 1, on CUB-200-2011, the best results are obtained at $\alpha = 0$ when the backbone is VGG16.

α	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Top-1	**71.33**	69.79	69.14	69.75	70.07	69.20	70.01	69.55	69.34	69.22	68.84
Top-5	**85.33**	83.57	83.00	83.86	83.41	82.74	83.91	83.32	82.83	83.75	82.76
GT-known	**91.07**	89.24	88.95	89.75	89.33	88.72	89.37	89.17	88.67	89.70	88.64

Table 1. **Performance $w.r.t$ α on CUB-200-2011.**

Hyperparameter β in total loss. β reflects the degree of constraint between foreground area and background suppression. when β is small, more areas in the foreground activation map are activated, while when β is too large, it will suppress the learning of the activation map. As shown in Table 2, our method achieves the best performance when β = 0.7 on VGG16.

β	**0.1**	**0.5**	**0.6**	**0.7**	**0.8**	**1.0**	**1.2**	**1.5**
Top-1	68.65	70.35	70.47	**71.33**	70.01	70.66	70.13	68.90
Top-5	82.32	82.32	84.78	**85.33**	84.11	84.73	84.00	83.01
GT-known	87.79	90.01	90.33	**91.07**	90.01	90.30	89.74	88.74

Table 2. **Performance $w.r.t$ β on CUB-200-2011.**

Selection of hyperparameter. We show the selection of hyperparameters and the corresponding localization accuracy of the proposed BAS for different backbones and datasets in Table 3.

Dataset	Backbone	α	β	λ	K	**Top-1**	**Top-5**	**GT-known**
CUB-200-2011 [11]	VGG16	0.0	0.7	1.0	80	71.33	85.33	91.07
	MobileNetV1	0.5	1.5	1.0	200	69.77	86.00	92.35
	ResNet50	0.5	1.2	1.0	200	77.25	90.08	95.13
	InceptionV3	0.5	1.0	1.0	200	73.29	86.31	92.24
ILSVRC [9]	VGG16	0.05	1.0	1.0	1	52.96	65.41	69.64
	MobileNetV1	0.5	1.5	1.0	1	52.97	66.59	72.00
	ResNet50	1.0	2.0	1.0	1	57.18	68.44	71.77
	InceptionV3	1.0	2.5	1.0	1	58.51	67.00	71.93

Table 3. **Selection of hyperparameters under different backbones and datasets and corresponding localization accuracy.**

Accuracy. We show more results in Table 4 and Table 5. It can be noted that the proposed method achieves excellent localization accuracy along with high accuracy on the classification task, which indicates that BAS can learn object localization results without affecting the classification ability.

3. More Examples

Visual Results. More visualization examples are shown in Fig. 4 and Fig. 5. As can be seen in Fig. 5, even in a noisy environment, BAS can still accurately localize objects, which indicates that the proposed BAS has robust localization capability.

Mask Annotation. We demonstrate part of the mask labels provided by CUB-200-2011 and compare the localization maps of SPA [8] and BAS on VGG16, as shown in Fig. 6. Compared to SPA, our localization maps are brighter on the object area and better localized at the edges of the object.

Methods	Venue	Backbone	Loc. Acc.			Cls. Acc.	
			Top-1	Top-5	GT-known	Top-1	Top-5
CAM [19]	CVPR16	VGG16	41.06	50.66	55.10	76.60	92.50
ACoL [16]	CVPR18	VGG16	45.92	56.51	62.96	71.90	–
ADL [3]	CVPR19	VGG16	52.36	-	75.41	65.27	–
DANet [14]	ICCV19	VGG16	52.52	61.96	67.70	75.40	92.30
I2C [18]	ECCV20	VGG16	55.99	68.34	–	–	–
MEIL [6]	CVPR20	VGG16	57.46	–	73.84	74.77	–
GCNet [5]	ECCV20	VGG16	63.24	75.54	81.10	76.80	92.30
PSOL [15]	CVPR20	VGG16	66.30	84.05	89.11	–	–
SPA [8]	CVPR21	VGG16	60.27	72.50	77.29	76.11	92.15
SLT [4]	CVPR21	VGG16	67.80	–	87.60	76.60	–
FAM [7]	ICCV21	VGG16	69.26	–	89.26	77.26	–
ORNet [13]	ICCV21	VGG16	67.73	80.77	86.20	77.00	93.00
BAS (Ours)	This Work	VGG16	**71.33**	**85.33**	**91.07**	**77.49**	**93.18**
CAM [19]	CVPR16	MobileNetV1	48.07	59.20	63.30	73.25	91.50
HaS [10]	ICCV17	MobileNetV1	46.70	–	67.31	65.98	–
ADL [3]	CVPR19	MobileNetV1	47.74	–	–	70.43	–
RCAM [2]	ECCV20	MobileNetV1	59.41	–	78.60	73.51	–
FAM [7]	ICCV21	MobileNetV1	65.67	–	85.71	**76.38**	–
BAS (Ours)	This Work	MobileNetV1	**69.77**	**86.00**	**92.35**	74.67	**92.60**
CAM [19]	CVPR16	ResNet50	46.71	54.44	57.35	80.26	–
ADL [3]	CVPR19	ResNet50-SE	62.29	–	–	80.34	–
PSOL [15]	CVPR20	ResNet50	70.68	86.64	90.00	–	–
WTL [1]	WACV21	ResNet50	64.70	–	77.35	77.28	–
FAM [7]	ICCV21	ResNet50	73.74	–	85.73	**82.72**	–
SPOL [12]	CVPR21	ResNet50	**80.12**	**93.44**	**96.46**	–	–
BAS (Ours)	This Work	ResNet50	77.25	90.08	95.13	80.84	**94.39**
CAM [19]	CVPR16	InceptionV3	41.06	50.66	55.10	73.80	91.50
SPG [17]	ECCV18	InceptionV3	46.64	57.72	–	–	–
DANet [14]	ICCV19	InceptionV3	49.45	60.46	67.03	71.20	90.60
I2C [18]	ECCV20	InceptionV3	55.99	68.34	72.60	–	–
GCNet [5]	ECCV20	InceptionV3	58.58	71.00	75.30	76.80	**93.40**
PSOL [15]	CVPR20	InceptionV3	65.51	83.44	–	–	–
SPA [8]	CVPR21	InceptionV3	53.59	66.50	72.14	73.51	91.39
SLT [4]	CVPR21	InceptionV3	66.10	–	86.50	76.40	–
FAM [7]	ICCV21	InceptionV3	70.67	–	87.25	**81.25**	–
BAS (Ours)	This Work	InceptionV3	**73.29**	**86.31**	**92.24**	79.01	93.10

Table 4. Comparison of localization accuracy and classification accuracy with state-of-the-art methods on CUB-200-2011. Best results are highlighted in **bold**, second are underlined.

Methods	Venue	Backbone	Loc. Acc.			Cls. Acc.	
			Top-1	Top-5	GT-known	Top-1	Top-5
CAM [19]	CVPR16	VGG16	42.80	54.86	59.00	66.60	88.60
ACoL [16]	CVPR18	VGG16	45.83	59.43	62.96	67.50	88.00
ADL [3]	CVPR19	VGG16	44.92	–	–	69.48	–
I2C [18]	ECCV20	VGG16	47.41	58.51	63.90	69.40	89.30
MEIL [6]	CVPR20	VGG16	46.81	–	–	70.27	–
PSOL [15]	CVPR20	VGG16	50.89	60.90	64.03	–	–
SPA [8]	CVPR21	VGG16	49.56	61.32	65.05	70.51	90.05
SLT [4]	CVPR21	VGG16	51.20	62.40	67.20	**72.40**	–
FAM [7]	ICCV21	VGG16	51.96	–	**71.73**	70.90	–
ORNet [13]	ICCV21	VGG16	52.05	63.94	68.27	71.60	90.40
BAS (Ours)	This Work	VGG16	**52.96**	**65.41**	69.64	70.84	**90.46**
CAM [19]	CVPR16	MobileNetV1	43.35	54.44	58.97	66.20	87.23
HaS [10]	ICCV17	MobileNetV1	42.73	–	60.12	65.45	–
ADL [3]	CVPR19	MobileNetV1	43.01	–	–	67.77	–
RCAM [2]	ECCV20	MobileNetV1	44.78	–	61.69	67.15	–
FAM [7]	ICCV21	MobileNetV1	46.24	–	62.05	**70.28**	–
BAS (Ours)	This Work	MobileNetV1	**52.97**	**66.59**	**72.00**	68.94	**89.28**
CAM [19]	CVPR16	ResNet50	38.99	49.47	51.86	–	–
ADL [3]	CVPR19	ResNet50-SE	48.53	–	–	75.85	–
I2C [18]	ECCV20	ResNet50	51.83	64.60	68.50	**76.70**	–
PSOL [15]	CVPR20	ResNet50	53.98	63.08	65.44	–	–
WTL [1]	WACV21	ResNet50	52.36	–	67.89	–	–
FAM [7]	ICCV21	ResNet50	54.46	–	64.56	76.48	–
SPOL [12]	CVPR21	ResNet50	**59.14**	67.15	69.02	–	–
BAS (Ours)	This Work	ResNet50	57.18	**68.44**	**71.77**	76.06	**93.12**
CAM [19]	CVPR16	InceptionV3	46.29	58.19	62.68	73.30	91.80
SPG [17]	ECCV18	InceptionV3	48.60	60.00	64.69	69.70	90.10
DANet [14]	ICCV19	InceptionV3	47.53	58.28	–	72.50	91.40
I2C [18]	ECCV20	InceptionV3	53.11	64.13	68.50	73.30	91.60
GCNet [5]	ECCV20	InceptionV3	49.06	58.09	–	77.40	93.60
PSOL [15]	CVPR20	InceptionV3	54.82	63.25	65.21	–	–
SPA [8]	CVPR21	InceptionV3	52.73	64.27	68.33	73.26	91.81
SLT [4]	CVPR21	InceptionV3	55.70	65.40	67.60	**78.10**	–
FAM [7]	ICCV21	InceptionV3	55.24	–	68.62	77.63	–
BAS (Ours)	This Work	InceptionV3	**58.51**	**69.00**	**71.93**	77.99	**94.02**

Table 5. Comparison of localization accuracy and classification accuracy with state-of-the-art methods on ILSVRC. Best results are highlighted in **bold**, second are underlined.

Figure 4. Visualization of the localization results on CUB-200-2011 [11]. The ground-truth bounding boxes are in red, and the predictions are in green.

Figure 5. Visualization of the localization results on ILSVRC [9]. The ground-truth bounding boxes are in red, and the predictions are in green.

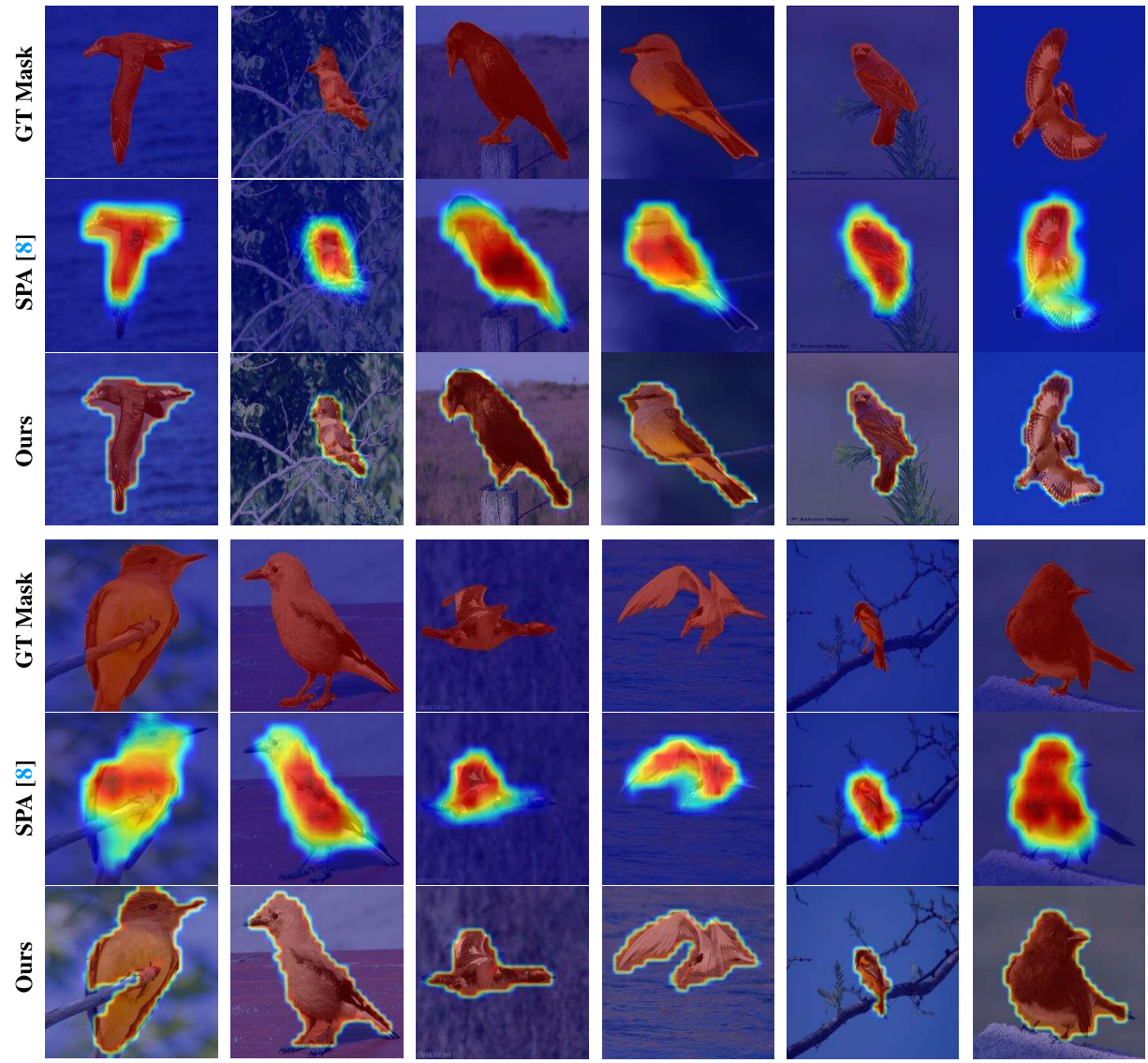

Figure 6. Illustration of the mask provided by CUB-200-2011, and comparison of the localization maps of SPA and our method on VGG16.

References

[1] Sadbhavana Babar and Sukhendu Das. Where to look?: Mining complementary image regions for weakly supervised object localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1010–1019, 2021. 5, 6

[2] Wonho Bae, Junhyug Noh, and Gunhee Kim. Rethinking class activation mapping for weakly supervised object localization. In *European Conference on Computer Vision*, pages 618–634. Springer, 2020. 5, 6

[3] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019. 5, 6

[4] Guangyu Guo, Junwei Han, Fang Wan, and Dingwen Zhang. Strengthen learning tolerance for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7403–7412, 2021. 5, 6

[5] Weizeng Lu, Xi Jia, Weicheng Xie, Linlin Shen, Yicong Zhou, and Jinming Duan. Geometry constrained weakly supervised object localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 481–496. Springer, 2020. 5, 6

[6] 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. 5, 6

[7] Meng Meng, Tianzhu Zhang, Qi Tian, Yongdong Zhang, and Feng Wu. Foreground activation maps for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3385–3395, 2021. 5, 6

[8] Xingjia Pan, Yingguo Gao, Zhiwen Lin, Fan Tang, Weiming Dong, Haolei Yuan, Feiyue Huang, and Changsheng Xu. Unveiling the potential of structure preserving for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11642–11651, 2021. 4, 5, 6, 9

[9] 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. 4, 8

[10] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE international conference on computer vision (ICCV)*, pages 3544–3553. IEEE, 2017. 5, 6

[11] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 4, 7

[12] Jun Wei, Qin Wang, Zhen Li, Sheng Wang, S Kevin Zhou, and Shuguang Cui. Shallow feature matters for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5993–6001, 2021. 5, 6

[13] Jinheng Xie, Cheng Luo, Xiangping Zhu, Ziqi Jin, Weizeng Lu, and Linlin Shen. Online refinement of low-level feature based activation map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 132–141, 2021. 5, 6

[14] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6589–6598, 2019. 5, 6

[15] Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu. Rethinking the route towards weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13460–13469, 2020. 5, 6

[16] 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 Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018. 5, 6

[17] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 597–613, 2018. 5, 6

[18] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 271–287. Springer, 2020. 5, 6

[19] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 5, 6