Supplementary material for Normalization Matters in Weakly Supervised Object Localization

Jeesoo Kim[1] Junsuk Choe[3] Sangdoo Yun[2] Nojun Kwak[1]

[1] Department of Intelligence and Information, Seoul National University [2] NAVER AI Lab
[3] Department of Computer Science and Engineering, Sogang University

A. Details about datasets

In the work of [2], Caltech-UCSD Birds-200-2011 (CUB) consists of 5994 "train" images and 5794 "test" images for 200 classes and they are used as the `train-weaksup` and `test` set respectively. The `train-fullsup` has been additionally collected from Flickr with extra annotated bounding boxes and used as the validation set. Similarly, 1.2M "train" images and 10K "validation" images for 1000 classes in ImageNet are used as the `train-weaksup` and `test` set respectively. Images with annotated bounding boxes from the ImageNetV2 [4] are used as the `train-fullsup` set. In both datasets, we select the best percentile for IVR from the `train-fullsup` set both in CUB and ImageNet. Additionally, we use OpenImages [3] which is reorganized in [5] with 29819, 2500, and 5000 images of `train-weaksup, train-fullsup` and `test`. Even though the evaluation metric in OpenImages is different from that of CUB and ImageNet, we still use `train-fullsup` as a validation set to find the optimal percentile for IVR.

B. Visualization of localized samples

Fig. 1,2 show the localization results in CUB and Fig. 3,4 show the localization results in ImageNet. The red boxes are bounding boxes upon the threshold of IoU 70. In each method, correctly localized samples in all normalization methods are placed on the first row. Using the same threshold, samples on the second row show some failure cases in each normalization method. Even an optimal threshold searched throughout a dataset fails localization in many samples.

C. Evaluation with Negative Weight Clamping (NWC)

NWC is a very powerful method for WSOL as it prevents contradictory features from disturbing each other. Tab. 1 shows the results of all methods combined with NWC. To find out the best performance score, IVR has used the best percentile verified from the validation set in all individual experiments.

In ImageNet, the performance improvement from min-max normalization is not significant, but PaS has shown the best result in CAM, HaS, ADL, and CutMix when combined with NWC. Also, unlike the results without NWC, max normalization has been no more better than min-max normalization. IVR is still better than min-max normalization and outperforms PaS only in ACoL and SPG. Among all cases, the top localization accuracy in IoU 50 recorded 68.73% in CutMix-Inception with PaS while the best score reported in [1] is 64.44%.

In CUB, IVR outperforms all other methods in a great deal. Min-max normalization is also comparable with max normalization. Meanwhile, the performance in PaS significantly drops in all cases. Using the individual best percentile acquired from the validation set, 89.14% of localization accuracy in IoU 50 can be acquired in CutMix-VGG with IVR.

In OpenImages, PaS shows the greatest performance degradation in all cases. Instead, contrary to the results in the main paper, min-max normalization shows the best result in OpenImages. IVR remains in the second best normalization method with a slightly lower score compared to min-max normalization.

According to this result, removing negative values in the weight makes max normalization unnecessary. However, IVR still shows comparable performance in ImageNet and OpenImages, and far more better performance in CUB. Also in Tab. 1, we provide the performance drop from the top score among the four normalization methods in red. The variation of these values among three datasets is reported in the last column and we can easily see that IVR with NWC shows the most stable and

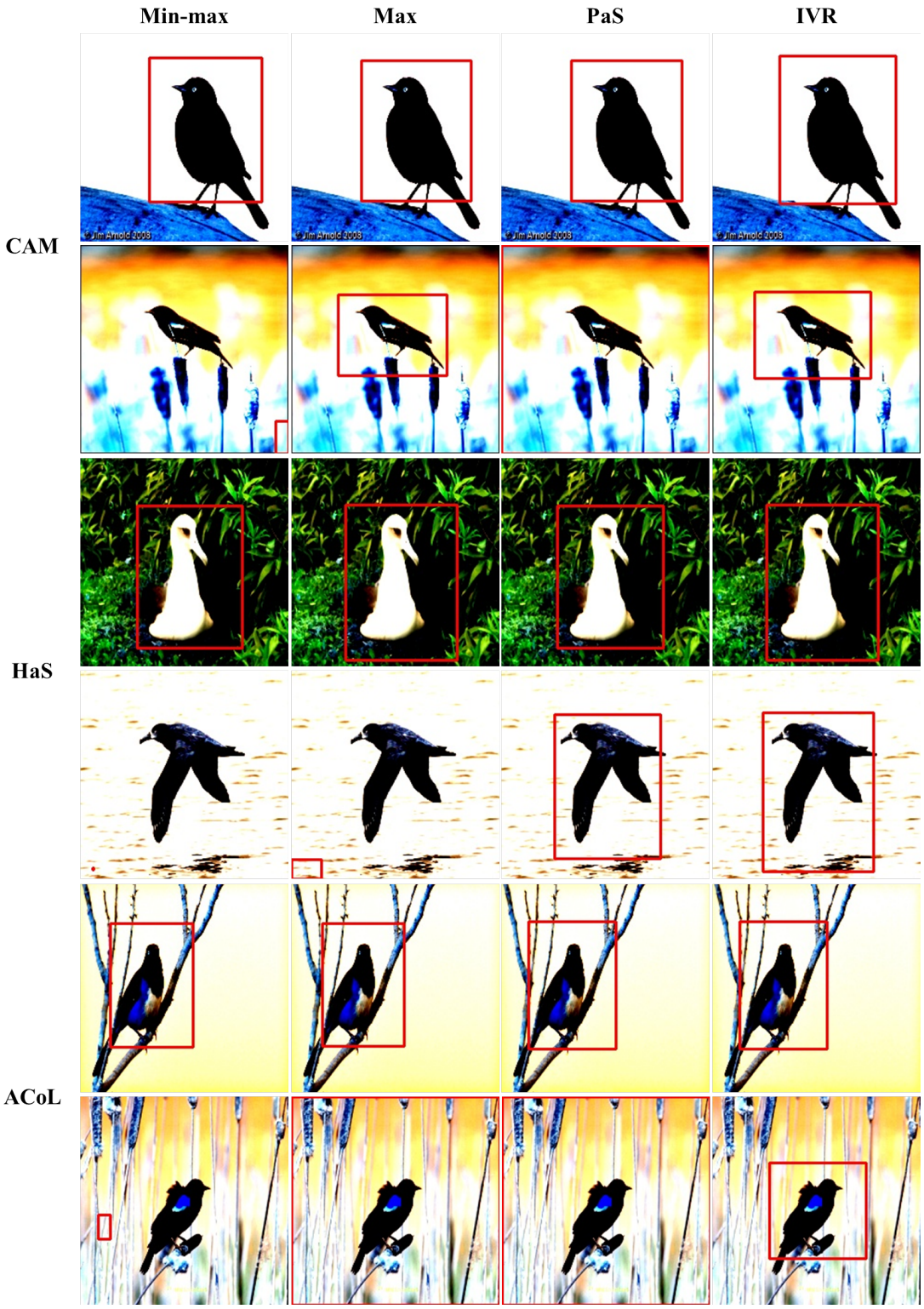

Figure 1: Localized examples of CAM, HaS and ACoL in CUB.

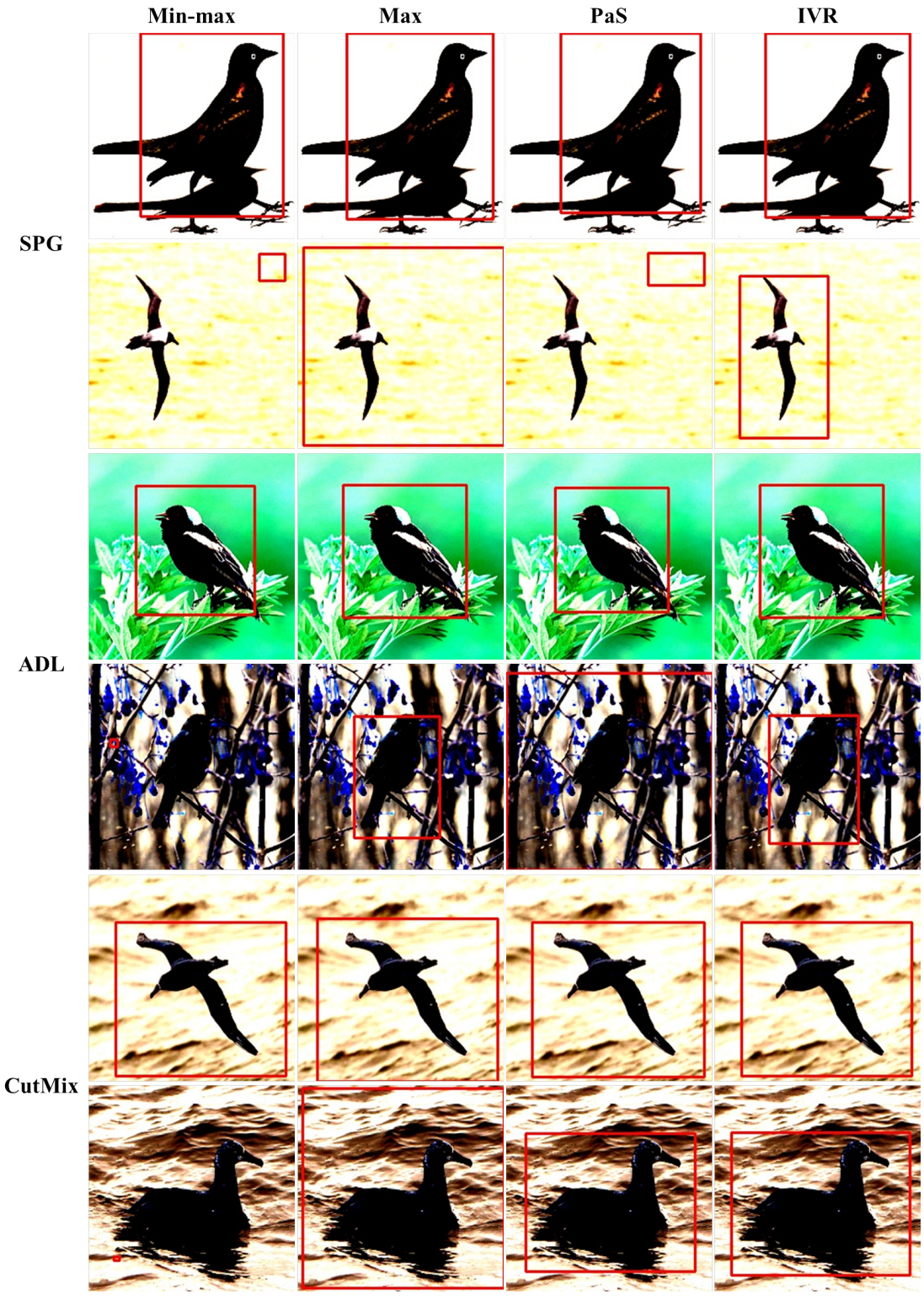

Figure 2: Localized examples of SPG, ADL and CutMix in CUB.

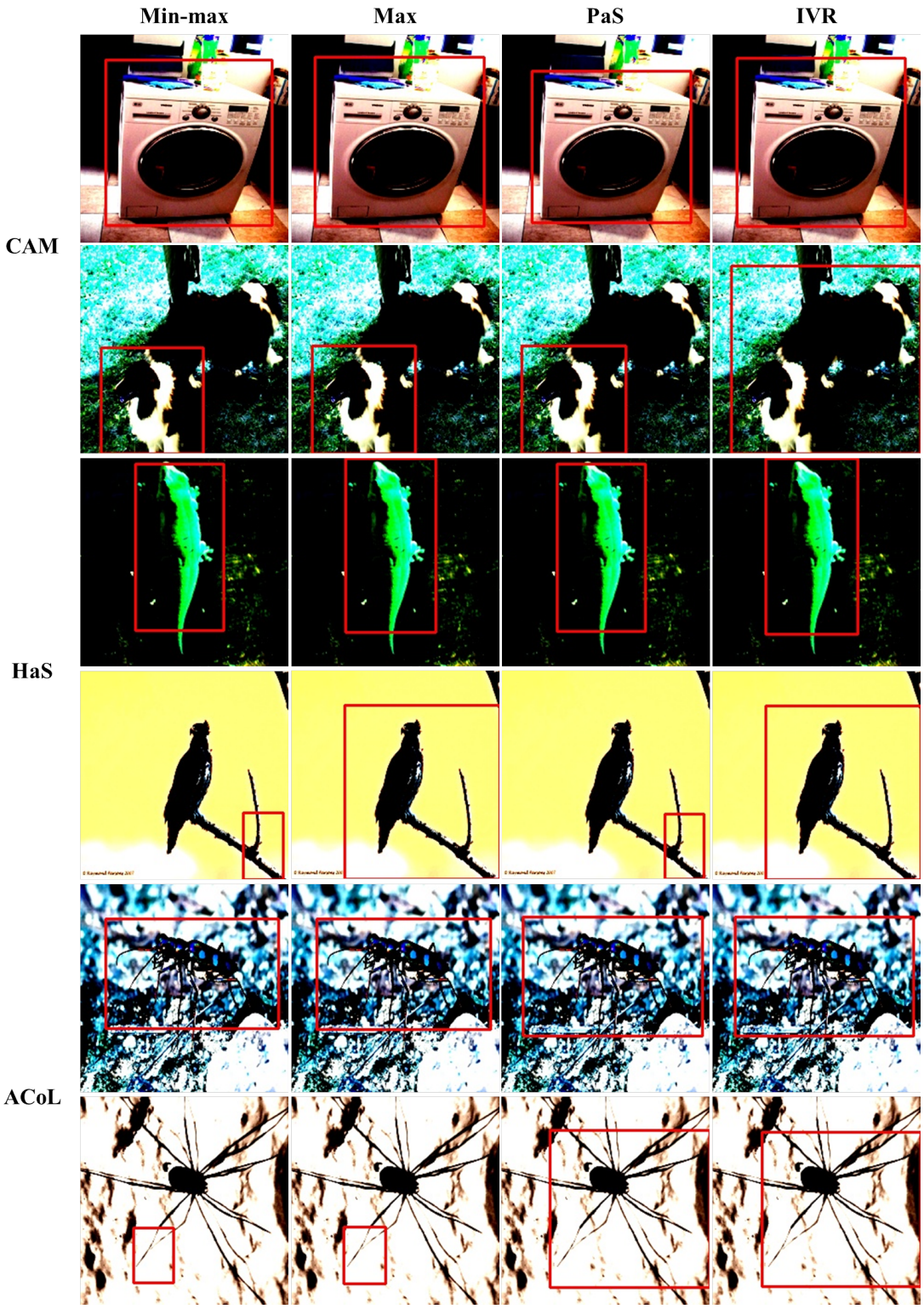

Figure 3: Localized examples of CAM, HaS and ACoL in ImageNet.

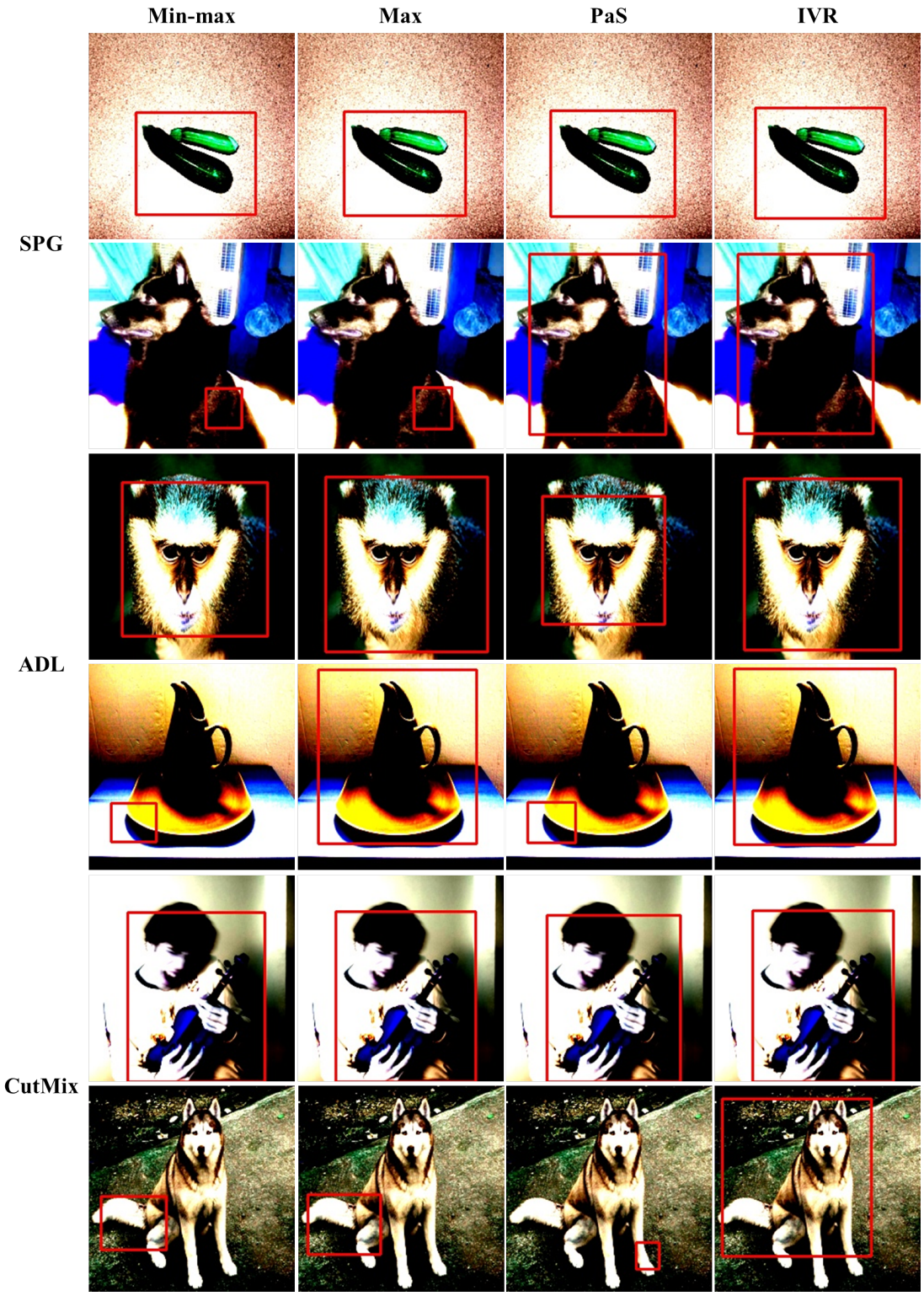

Figure 4: Localized examples of SPG, ADL and CutMix in ImageNet.

Table 1: Evaluating WSOL with different normalization methods with Negative Weight Clamping (NWC) [1]. Numbers in red are the performance drop from the top score among the four normalization methods. The last column represents the variance of these scores in one normalization method. IVR shows the smallest variance.

Method	Norm	ImageNet (MaxBoxAccV2)				CUB (MaxBoxAccV2)				OpenImages (PxAP)				Var
		VGG	Incep.	ResNet	Mean	VGG	Incep.	ResNet	Mean	VGG	Incep.	ResNet	Mean	
CAM	Minmax	60.28	64.47	64.77	63.17(-1.51)	68.90	62.39	69.26	66.85(-0.60)	59.73	64.30	59.90	**61.31**(0.00)	0.58
	Max	59.58	64.41	64.50	62.83(-1.85)	68.82	62.24	68.58	66.54(-0.91)	59.20	64.44	59.87	61.17(-0.14)	0.74
	PaS	62.45	65.21	66.38	**64.68**(0.00)	67.86	60.51	68.66	65.68(-1.78)	55.96	59.69	55.42	57.02(-4.29)	4.64
	IVR	60.98	65.32	65.51	63.94(-0.74)	68.95	62.83	70.57	**67.45**(0.00)	59.66	63.88	59.57	61.04(-0.27)	**0.20**
HaS	Minmax	61.16	64.40	64.72	63.43(-1.39)	75.31	61.74	74.48	70.51(-2.16)	59.16	62.73	56.67	**59.52**(0.00)	1.20
	Max	60.75	64.35	64.47	63.19(-1.63)	74.89	62.02	74.01	70.31(-2.37)	58.58	62.67	56.78	59.34(-0.18)	1.24
	PaS	62.89	65.36	66.21	**64.82**(0.00)	69.99	60.17	71.20	67.12(-5.56)	55.95	57.11	53.21	55.42(-4.10)	8.30
	IVR	61.56	65.22	65.16	63.98(-0.84)	77.59	64.15	76.29	**72.67**(0.00)	59.09	62.62	56.03	59.24(-0.28)	**0.25**
ACoL	Minmax	55.42	64.45	61.20	60.36(-0.60)	64.18	60.63	74.78	66.53(-0.88)	53.93	57.04	57.66	**56.21**(0.00)	0.20
	Max	55.25	64.44	61.17	60.28(-0.67)	64.22	60.58	74.66	66.49(-0.93)	53.91	57.06	57.63	56.20(-0.01)	0.23
	PaS	56.42	64.68	61.76	**60.95**(0.00)	63.83	60.30	74.74	66.29(-1.13)	51.35	53.11	54.20	52.89(-3.32)	2.85
	IVR	55.43	65.08	61.20	60.57(-0.39)	64.48	60.83	76.94	**67.42**(0.00)	53.62	56.89	57.07	55.86(-0.35)	**0.06**
SPG	Minmax	59.54	64.45	63.62	62.54(-1.42)	67.74	64.17	71.98	67.96(-2.18)	58.98	63.68	58.33	**60.33**(0.00)	1.22
	Max	59.22	64.35	63.40	62.32(-1.64)	66.94	63.65	70.73	67.11(-3.04)	58.64	63.64	58.51	60.26(-0.07)	2.20
	PaS	61.40	65.16	65.30	**63.95**(0.00)	64.46	61.07	69.39	64.97(-5.17)	55.22	58.84	55.22	56.43(-3.91)	7.26
	IVR	60.22	65.38	64.09	63.23(-0.72)	71.06	65.01	74.36	**70.14**(0.00)	58.81	63.59	57.72	60.04(-0.29)	**0.19**
ADL	Minmax	63.67	62.55	64.74	63.65(-1.10)	69.40	66.66	69.03	68.36(-0.27)	58.93	57.01	57.01	**57.65**(0.00)	0.33
	Max	63.43	62.47	64.44	63.45(-1.31)	69.28	66.45	69.06	68.26(-0.37)	58.42	57.04	56.84	57.43(-0.22)	0.35
	PaS	64.52	63.58	66.17	**64.76**(0.00)	68.12	63.78	67.17	66.36(-2.28)	55.48	53.86	54.21	54.52(-3.13)	2.62
	IVR	64.28	64.07	65.44	64.60(-0.16)	69.50	66.94	69.46	**68.63**(0.00)	58.84	56.75	56.49	57.36(-0.29)	**0.01**
CutMix	Minmax	59.10	65.01	64.61	62.90(-1.45)	78.22	63.68	71.17	71.02(-1.01)	59.37	64.41	60.63	**61.47**(0.00)	0.56
	Max	58.39	64.97	64.33	62.56(-1.79)	77.95	63.48	70.55	70.66(-1.38)	58.89	64.58	60.66	61.38(-0.09)	0.78
	PaS	61.18	65.80	66.10	**64.36**(0.00)	74.39	61.71	70.77	68.96(-3.08)	55.78	59.73	55.85	57.12(-4.35)	5.00
	IVR	59.33	65.95	65.18	63.49(-0.87)	78.93	64.78	72.40	**72.04**(0.00)	59.24	63.98	60.20	61.14(-0.33)	**0.27**

robust performance. This implies that the minimum value in the class activation map must be selected adaptively. Researchers should remember that each WSOL method requires a specific normalization method that fits best in different datasets.

References

[1] 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. 1, 6

[2] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3133–3142, 2020. 1

[3] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 1

[4] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 1

[5] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 1