Supplementary Material for "Unveiling the Potential of Structure Preserving for Weakly Supervised Object Localization"

Xingjia Pan[1,3*] Yingguo Gao[1*] Zhiwen Lin[1*] Fan Tang[2†] Weiming Dong[3,4,5]
Haolei Yuan[1] Feiyue Huang[1] Changsheng Xu[3,4,5]
[1]Youtu Lab, Tencent [2]Jilin University [3]NLPR, Institute of Automation, CAS
[4]School of Artificial Intelligence, UCAS [5]CASIA-LLVision Joint Lab
{noahpan, yingguogao, xavierzwlin,harryyuan,garyhuang}@tencent.com
tangfan@jlu.edu.cn,{weiming.dong, changsheng.xu}@ia.ac.cn

1. High-order Self Correlation

To apply the inherent structure-preserving ability of the network for accurate WSOL, we propose to use high-order self-correlation (HSC) to capture the structural information of objects. The h_{th} order similarity between f_i and f_i are formulated as:

$$S^h(f_i, f_j) = \frac{1}{(HW)^{h-1}} \sum_{k^0} \cdots \sum_{k^{h-2}} S(f_i, f_{k^0}) \cdots S(f_{k^{h-2}}, f_j), \tag{1}$$

where $k^0, k^{h-2} \in \Omega$, $h \in \{2, 3, 4 \cdots\}$ and $i \neq k^0 \neq k^{h-2}$. h denotes the order, and Ω denotes the set of all features. The $S^h(f_i, f_j)$ is then normalized to $[0, 1]$ following:

$$\hat{S}^h(f_i, f_j) = \frac{S^h(f_i, f_j) - min_{k\in\Omega} S^h(f_i, f_k)}{max_{k\in\Omega} S^h(f_i, f_k) - min_{k\in\Omega} S^h(f_i, f_k)}, \tag{2}$$

Then, we define HSC as:

$$SC^h(f) = \left[\hat{S}^h(f_i, f_j)|_{i,j}\right]. \tag{3}$$

Compared with first-order self-correlation, HSC can preserve the details of the object by considering long-range context. However, the HSC may introduce additional noise.

2. Quantitative Results

Bounding Box Localization We here show more results on bounding boxes localization. Table 1 and Table 2 are the localization results on ILSVRC validation and CUB-200-2011 testing sets, respectively. On ILSVRC validation set, we achieve the state-of-the-art with VGG16, and obtain comparable results with current state-of-the-art method I^2C [14] with Inception V3. Compared with baseline methods, the proposed SPA is much simpler and almost parameter-free without introducing additional convolutional layers. On CUB-200-2011 testing set, the proposed SPA surpassing all the baseline methods.

Error Analysis. To further reveal the effect of our method, we define five metrics to perform bad case analysis on the proposed SPA approach. Specifically, we divide the localization error into five cases: classification error (Cls), multi-instance error (M-Ins), localization part error (Part), localization more error (More), and others (OT). For better description, we also define *IoG* and *IoP*. Similar to *IoU*(intersection over Union), IoG means that intersection over ground truth box and IoP means that intersection over predict bounding box.

- *Cls* includes the predictions that are wrongly classified.
- *M-Ins* indicates that the prediction intersects with at least two ground truth boxes, and IoG > 0.3.
- *Part* indicates that the predicted bounding box only cover the parts of object, and IoB > 0.5.
- *More* indicates that the predicted bounding box is larger than the ground truth bounding box by a large margin, and IoG > 0.7.
- *OT* indicates other predictions that do not fall under above mentioned cases.

The five cases defined above are mutually exclusive. We show the details of the definition of five metrics in Algorithm 1 Each metric calculates the percentage of images belonging to corresponding error in the validation/testing set. Table 3 lists localization error statistics. Localization error *Cls* caused by wrong classification dominates the error. But the classification is not the main concern for WSOL. We mainly focus on *M-Ins*, *Part*, and *More* metrics. For

*Equal contribution
†Corresponding author

Methods	Backbone	Loc Err.			Cls Err.	
		Top-1	Top-5	Gt-Known	Top-1	Top-5
Backprop [5]	VGG16 [6]	61.12	51.46	-	-	-
CAM [15]	VGG16 [6]	57.20	45.14	-	31.2	11.4
CutMix [11]	VGG16 [6]	56.55	-	-	-	-
ADL [1]	VGG16 [6]	55.08	-	-	-	-
ACoL [12]	VGG16 [6]	54.17	40.57	37.04	32.5	12.0
I^2C [14]	VGG16 [6]	52.59	41.49	36.10	30.6	10.7
MEIL [3]	VGG16 [6]	53.19	-	-	29.73	-
Ours	VGG16 [6]	**50.44**	**38.68**	**34.95**	**29.49**	**9.95**
CAM [15]	InceptionV3 [8]	53.71	41.81	37.32	26.7	8.2
SPG [13]	InceptionV3 [8]	51.40	40.00	35.31	30.3	9.9
ADL [1]	InceptionV3 [8]	51.29	-	-	27.2	-
ACoL [12]	GoogLeNet [7]	53.28	42.58	-	29.0	11.8
DANet [10]	GoogLeNet [7]	52.47	41.72	-	27.5	8.6
MEIL [3]	InceptionV3 [8]	50.52	-	-	26.69	-
I^2C [14]	InceptionV3 [8]	**46.89**	35.87	**31.50**	26.7	8.4
GC-Net [2]	InceptionV3 [8]	50.94	41.91	-	**22.6**	**6.4**
Ours	InceptionV3 [8]	47.29	**35.73**	31.67	26.74	8.19

Table 1. Comparison between our method and the state-of-the-art on the ILSVRC [4] validation set. Our method outperforms all other methods by a large margin for object localization. Here 'ClsErr', 'LocErr' and 'Gt-Known' are short for classification error, location error and Gt-known location error, respectively.

Methods	Backbone	Loc Err.			Cls Err.	
		Top-1	Top-5	Gt-Known	Top-1	Top-5
CAM [15]	GoogLeNet [7]	58.94	49.34	44.9	26.2	8.5
SPG [13]	GoogLeNet [7]	53.36	42.8	-	-	-
DANet [10]	InceptionV3 [8]	50.55	39.54	33.0	28.8	9.4
ADL [1]	InceptionV3 [8]	46.96	-	-	**25.45**	-
Ours	InceptionV3 [8]	**46.41**	**33.50**	**27.86**	26.49	8.61
CAM [15]	VGG16 [6]	55.85	47.84	44.0	**23.4**	**7.5**
ADL [1]	VGG16 [6]	47.64	-	-	34.73	-
ACoL [12]	VGG16 [6]	54.08	43.49	45.9	28.1	-
DANet [10]	VGG16 [6]	47.48	38.04	32.3	24.6	7.7
SPG [13]	VGG16 [6]	51.07	42.15	41.1	24.5	7.9
I^2C [14]	VGG16 [6]	44.01	31.6	-	-	-
MEIL [3]	VGG16 [6]	42.54	-	-	25.23	-
Ours	VGG16 [6]	**39.73**	**27.5**	**22.71**	23.89	7.85

Table 2. Comparison between our method and the state-of-the-art on the CUB-200-2011 test set. Here 'ClsErr', 'LocErr' and 'Gt-Known' are short for classification error, location error and Gt-known location error, respectively.

Methods	ILSVRC(%)					CUB-2011-200(%)				
	Cls	M-Ins	Part	More	OT	Cls	M-Ins	Part	More	OT
VGG16	29.47	10.65	3.85	9.58	0.2	23.28	-	21.91	10.53	1.76
Ours	29.49	9.97	2.83	7.66	0.5	23.89	-	9.25	6.33	0.26
InceptionV3	26.73	10.36	3.22	9.49	0.20	26.39	-	23.09	5.52	0.64
Ours	26.74	9.48	2.89	7.80	0.37	26.49	-	12.81	6.83	0.03

Table 3. Localization error statistics.

ILSVRC, the proposed SPA effectively reduces the three kinds of localization error. Although there is no explicit solution for multi-instance problem, our method reduces the *M-Ins* error. From Table 3, it shows that the *M-Ins* and *More* are the most important problems to solve. It it worth to explore CAM-based methods for *weakly supervised instance*

Algorithm 1 Error Analysis Algorithm.

Input: Predicted bounding box $B_p \in \mathbb{R}^{1\times 4}$; Predicted label C_p; ground truth boxes list $B_g \in \mathbb{R}^{N\times 4}$; ground truth label C_g; Maximum IoU $\hat{IoU}$ between the B_p and B_g;
Output: Cls, M-Ins, Part, More, OT;

Set Cls = M-Ins = Part = More = OT = 0
if $C_p \neq C_b[0]$ **then**
    Cls =1
    return Cls, M-Ins, Part, More, OT
end if
if $\hat{IoU} > 0.5$ **then**
    return Cls, M-Ins, Part, More, OT
end if
Calculate $IoG \in \mathbb{R}^{1\times N}$ between the B_p and B_g
if $Count(IoG > 0.3) > 1$ **then**
    M-Ins = 1
    return Cls, M-Ins, Part, More, OT
end if
Calculate the maximum $\hat{IoG}$ and $\hat{IoP}$
if $\hat{IoP} > 0.5$ **then**
    Part = 1
    return Cls, M-Ins, Part, More, OT
end if
if $\hat{IoG} > 0.7$ **then**
    More = 1
    return Cls, M-Ins, Part, More, OT
end if
OT = 1
return Cls, M-Ins, Part, More, OT

localization in future work. For CUB-200-2011, there is only one instance in each image. Our method effectively reduces *Part* and *More* errors on VGG16, which indicates that our localization maps are much accurate. On Inception V3, the proposed SPA significantly reduce *Part* by 10.2%.

3. More Examples

Self-correlation Maps. We here show more visualization examples for self-correlation maps in Fig. 1, Fig. 2, Fig. 3, Fig. 4. In each figure, we show five self-correlation maps for corresponding points marked by green cross in original images. The middle row shows the first-order self-correlation maps and the bottom row shows the second-order self-correlation maps.

localization Results. We also show more visualization examples of localization maps in Fig. 5, Fig. 6, Fig. 7. In each figure, the original images with ground truth bounding boxes (in red) are shown in top row. The localization maps with CAM and the proposed SPA are shown in middle and bottom row, respectively.

M_{bg}				M_{obj}			
τ(bg)	IoU	IoG	IoP	τ + σ(fg)	IoU	IoG	IoP
0.1	0.42	0.44	0.92	0.0	0.30	1.00	0.30
0.2	0.54	0.59	0.89	0.1	0.33	1.00	0.33
0.3	0.60	0.69	0.86	0.2	0.36	0.99	0.35
0.4	0.65	0.77	0.83	0.3	0.37	0.98	0.37
0.5	0.67	0.83	0.80	0.4	0.38	0.97	0.38
0.6	0.69	0.87	0.77	0.5	0.38	0.95	0.40
0.7	0.69	0.90	0.76	0.6	0.38	0.93	0.41
0.8	0.70	0.92	0.76	0.7	0.39	0.90	0.43
0.9	0.71	0.94	0.75	0.8	0.39	0.86	0.34

Table 4. Ablation study of τ and σ

α	Loc Err.			Cls Err.	
	Top-1	Top-5	Gt-Known	Top-1	Top-5
0.0	54.71	43.35	38.76	30.60	10.60
0.1	54.56	43.13	38.50	30.65	10.62
0.2	53.97	42.53	38.05	30.01	10.15
0.3	53.25	41.50	36.97	30.22	10.23
0.4	53.46	41.43	37.03	30.93	10.73
0.5	**52.71**	**40.82**	**36.40**	30.19	10.21
0.6	53.31	41.44	37.24	30.84	10.71

Table 5. Ablation study of α when fixing τ=0.4, σ=0.1.

τ	Loc Err.			Cls Err.	
	Top-1	Top-5	Gt-Known	Top-1	Top-5
0.2	53.50	41.61	36.81	30.87	10.70
0.3	53.24	41.21	36.61	30.92	10.71
0.4	**52.71**	**40.82**	**36.40**	30.19	10.21
0.5	52.94	41.46	37.63	30.18	10.21
0.6	54.95	44.51	41.98	30.90	10.73

Table 6. Ablation study of τ when fixing α=0.5, σ=0.1.

Ablation study of τ and σ in Equs. 3 and 4. M_{bg} and M_{obj} are calculated by Equs. 3 and 4 under different τs and σs. Average mask IoU, IoG (Intersection over GT), and IoP (Intersection over Prediction) are reported in Table 4. One can see that larger IoG implies the generated masks covering most of the desired region while low IoP means most of the masks are corrected. As the initial masks, M_{bg} and M_{obj}, are good enough to guide the model to suppress backgrounds and active full object extent, statistically.

Ablation study of α, τ and σ. We fix two of these variables and report the results with the variety of the left one on ILSVRC in Tables 5, 6 and 7.

References

[1] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceed-*

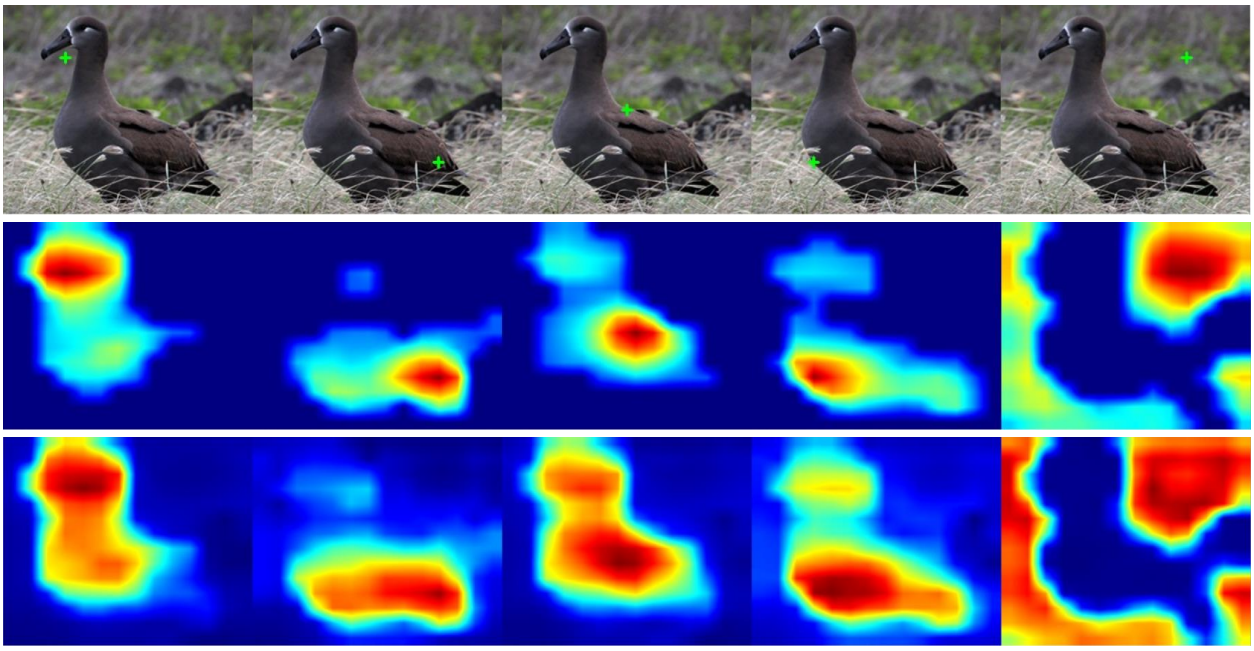

Figure 1. Visualization of the self-correlation maps with first-order similarity (middle row) and second-order similarity(bottom row). The images are from the CUB-200-2011 [9] testing set.

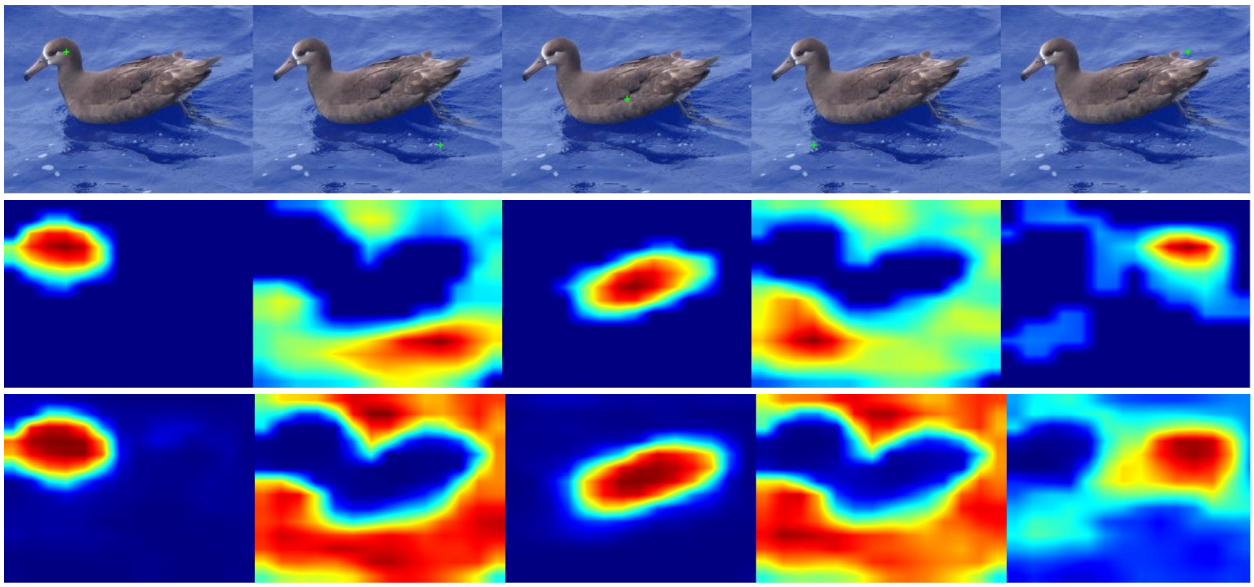

Figure 2. Visualization of the self-correlation maps with first-order similarity (middle row) and second-order similarity(bottom row). The images are from the CUB-200-2011 [9] testing set.

ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2219–2228, 2019.

[2] Weizeng Lu, Xi Jia, Weicheng Xie, Linlin Shen, Yicong Zhou, and Jinming Duan. Geometry constrained weakly supervised object localization. *arXiv preprint arXiv:2007.09727*, 2020.

[3] 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.

[4] 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.

[5] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman.

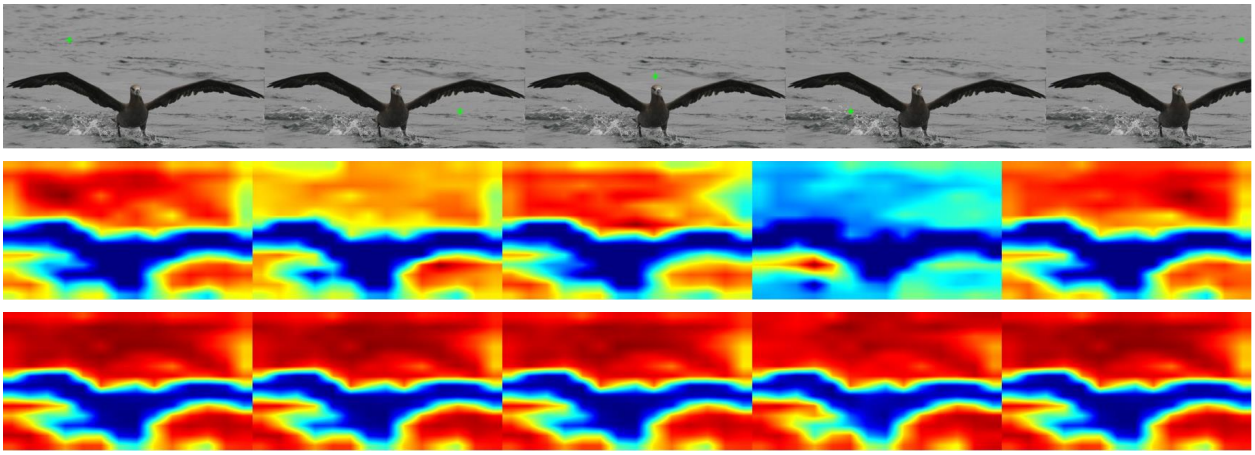

Figure 3. Visualization of the self-correlation maps with first-order similarity (middle row) and second-order similarity(bottom row). The images are from the CUB-200-2011 [9] testing set.

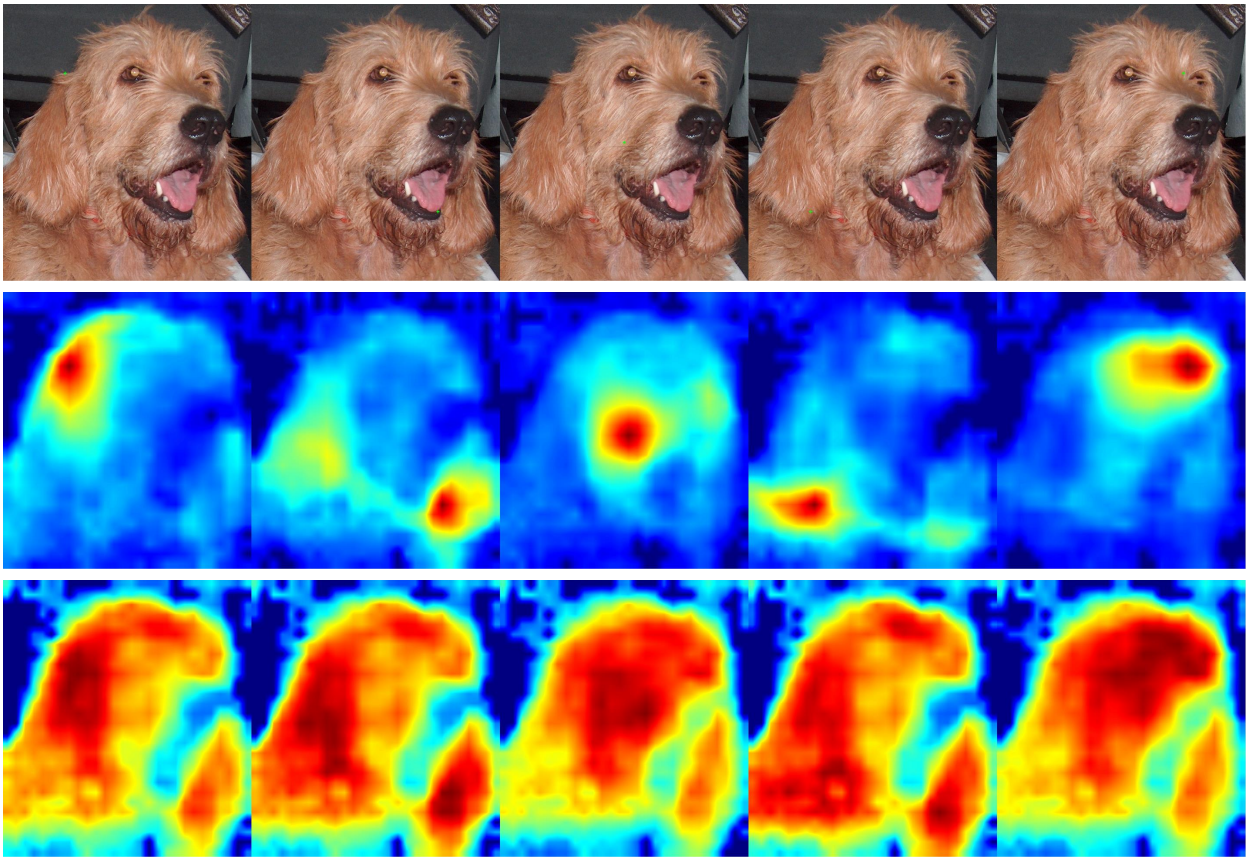

Figure 4. Visualization of the self-correlation maps with first-order similarity (middle row) and second-order similarity(bottom row). The images are from the ILSVRC [4] validation set.

Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

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

[7] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on*

Figure 5. Visualization of the localization maps with CAM [15] (middle row) and the proposed SPA(bottom row). The images are from the ILSVRC [4] validation set.

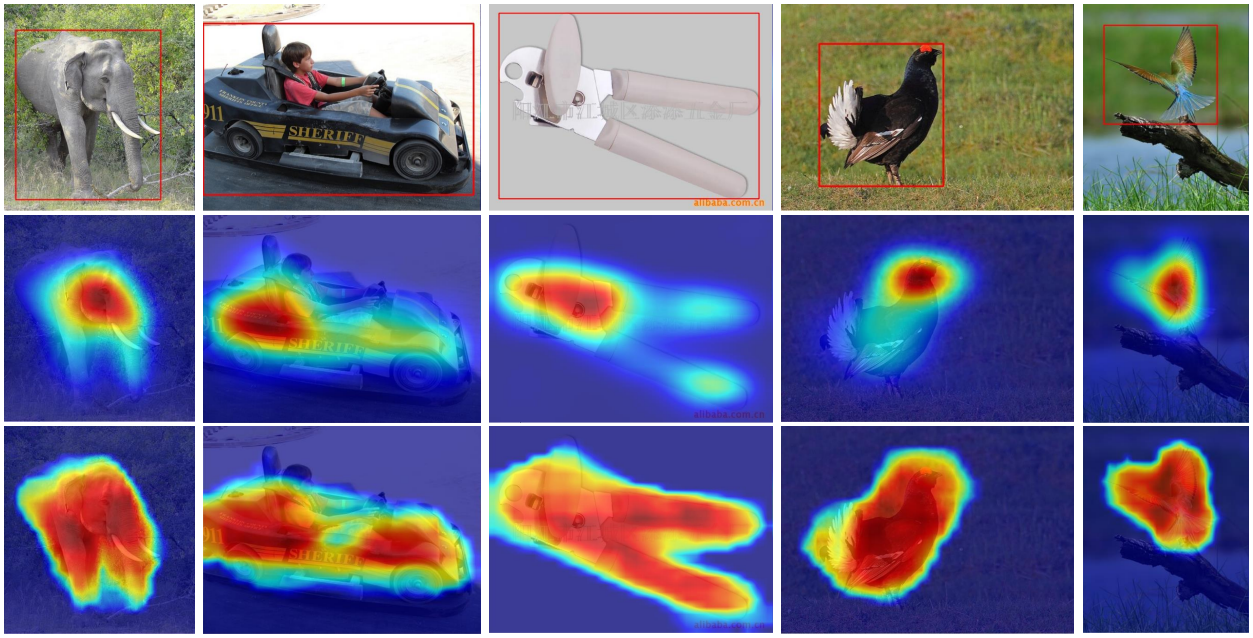

Figure 6. Visualization of the localization maps with CAM [15] (middle row) and the proposed SPA(bottom row). The images are from the ILSVRC [4] validation set.

σ	Loc Err.			Cls Err.	
	Top-1	Top-5	Gt-Known	Top-1	Top-5
0.0	52.82	40.88	36.35	30.21	10.20
0.1	**52.71**	**40.82**	**36.40**	30.19	10.21
0.2	53.32	41.52	37.67	30.89	10.69
0.3	53.53	42.06	38.52	30.86	10.71

Table 7. Ablation study of σ when fixing α=0.5, τ=0.4.

computer vision and pattern recognition, pages 1–9, 2015.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

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

[10] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly

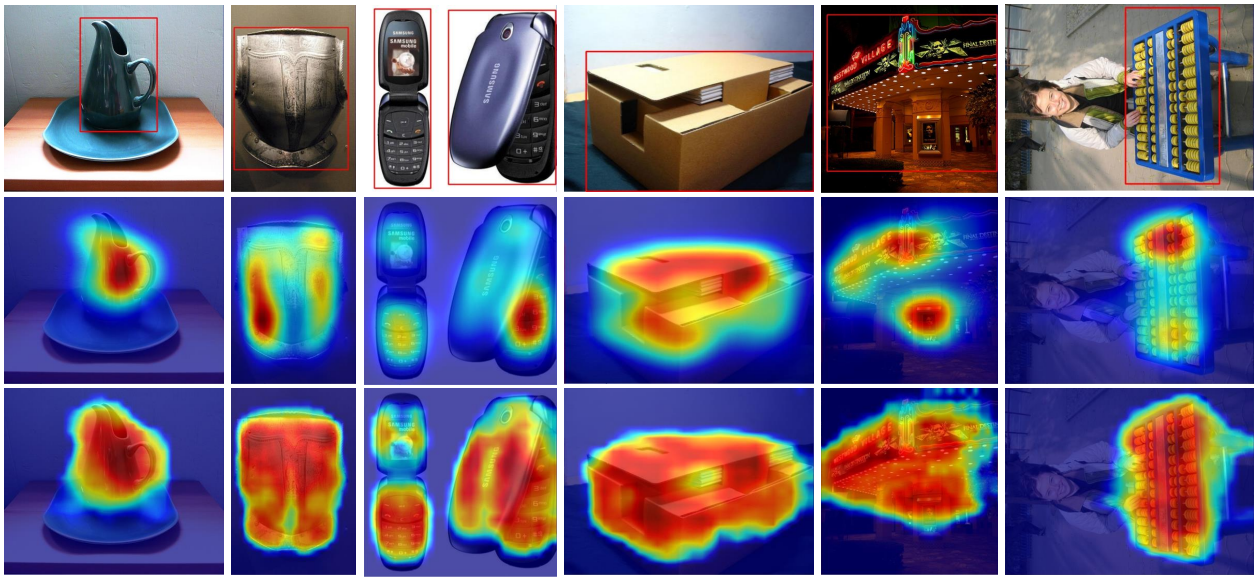

Figure 7. Visualization of the localization maps with CAM [15] (middle row) and the proposed SPA(bottom row). The images are from the ILSVRC [4] validation set.

supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6589–6598, 2019.

[11] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019.

[12] 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.

[13] 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.

[14] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[15] 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.