Spatial-Aware Token for Weakly Supervised Object Localization (Supplementary Materials)

Pingyu Wu[1], Wei Zhai[1,†], Yang Cao[1,3], Jiebo Luo[2], Zheng-Jun Zha[1]
[1] University of Science and Technology of China [2] University of Rochester
[3] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{wpy3647556200@mail., wzhai056@mail., forrest@}ustc.edu.cn
jluo@cs.rochester.edu zhazj@ustc.edu.cn

1. Ablation Study

Number of spatial aware transformer blocks. We fix the whole network to 12 blocks and adjust the number of spatial aware transformer blocks, denoted as N. As shown in Table 1, the best results are achieved when N is set to 3, which indicates that fusing the localization maps M^l learned from different blocks is helpful to obtain a complete localization result. However, when N is too large, it increases the optimization difficulty of the normalization loss thus reducing the localization performance.

N	CUB-200			ImageNet		
	Top-1 Cls	Top-1 Loc	GT-k. Loc	Top-1 Cls	Top-1 Loc	GT-k. Loc
1	81.45	79.82	97.48	78.16	59.04	71.84
2	81.62	80.17	98.02	78.33	59.90	72.77
3	**82.05**	**80.96**	**98.45**	**78.41**	**60.15**	**73.13**
4	81.93	80.76	98.36	78.23	59.87	72.92
5	80.69	78.75	97.12	78.16	58.67	71.41

Table 1. **Number of spatial aware transformer blocks.** We select the 10-th block as the spatial aware transformer block when N = 1, and the 10-th and 11-th blocks when N=2. When N $>$ 2, the last N blocks are adopted as spatial aware transformer blocks.

Dot position. We explore the impact of the position of the dot product in the spatial-query attention module, as shown in Table 2. Quantitative experiments show that performing the dot product before softmax will reduce the performance of classification and localization, mainly because the exponential form in softmax makes the semantic prediction M^l learning insufficient. Therefore, the dot product after the softmax function enables semantic prediction M^l to better capture the localization knowledge from the self-attention mechanism.

Dot Position	CUB-200			ImageNet		
	Top-1 Cls	Top-1 Loc	GT-k. Loc	Top-1 Cls	Top-1 Loc	GT-k. Loc
Before softmax	81.67	80.39	98.19	77.78	59.57	72.85
After softmax	**82.05**	**80.96**	**98.45**	**78.41**	**60.15**	**73.13**

Table 2. **Dot position.** Ablation experiments on the position of the dot product in the spatial-query attention module.

2. Analysis

Class token $w.r.t.$ **spatial-aware token.** To analyze the differences between spatial-aware token and class token, we implement the exploratory experiment as shown in Table 3. From Table 3 (a), it can be analyzed that sharing the same token

† Corresponding author.

between class token and spatial token will bring optimization conflict between classification and localization tasks, thus decreasing both classification and localization performance. In addition, as illustrated in Table 3 (b), using separate tokens and initializing the weights of the spatial token to the pre-trained weights of class token also results in reduced localization accuracy, suggesting that the information learned by the spatial token and class token are significantly different. As a result, it is necessary to learn a separate spatial token from scratch.

	Initial Weights		Cls Acc.		Loc Acc.		
	Class token	Spatial token	Top-1	Top-5	Top-1	Top-5	GT-k.
(a)	Pre-trained (shared)		77.83	93.92	58.31	68.35	71.08
(b)	Pre-trained	Pre-trained	78.34	94.14	59.81	69.96	72.60
(c)	Pre-trained	Random initial	**78.41**	**94.46**	**60.15**	**70.52**	**73.13**

Table 3. **Class token** $w.r.t.$ **spatial-aware token.** (a) Class token and spatial token share the same token, and its initial weights are the pre-trained weights of class token. (b) Class token and spatial token use separate tokens, and their initial weights are the pre-trained weights of class token. (c) Class token and spatial token use separate tokens, where the initial weights of the spatial token are randomly initialized.

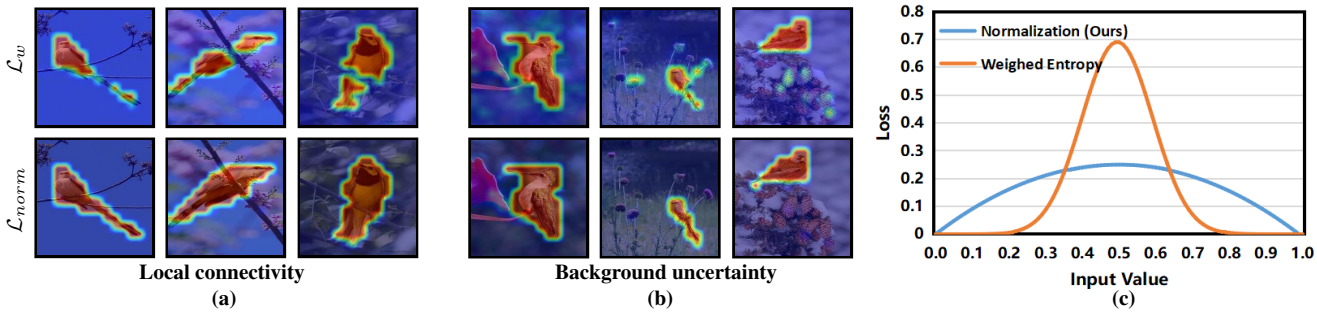

Figure 1. **Comparison between** $\mathcal{L}_w$ **and** $\mathcal{L}_{norm}$**.** (a) Visual comparison of SAT w/ $\mathcal{L}_w$ and SAT w/ $\mathcal{L}_{norm}$ on local connectivity. (b) Visual comparison of SAT w/ $\mathcal{L}_w$ and SAT w/ $\mathcal{L}_{norm}$ on background uncertainty. (c) Loss-Input value curves for different loss functions.

Normalization loss $w.r.t.$ **weighed entropy loss.** We compare the weighed entropy loss ($\mathcal{L}_w$) in ORNet [21] with our proposed normalization loss ($\mathcal{L}_{norm}$) in Table 4. Both losses aim to provide pixel-level supervision to increase the distinction between foreground and background of the localization map, but the effects are somewhat different, as shown in Fig. 1. 1) The proposed $\mathcal{L}_{norm}$ includes a gaussian filtering operation to incorporate the values of adjacent patches in the calculation of the loss, thus encouraging the local continuity of the localization map. As illustrated in Fig. 1 (a), the localization map generated by SAT w/ $\mathcal{L}_{norm}$ has better connectivity compared to SAT w/ $\mathcal{L}_w$. 2) Fig. 1 (c) shows the loss curves of the two loss functions versus the input values. Compared to $\mathcal{L}_{norm}$, $\mathcal{L}_w$ is already close to zero at input values of 0.2 or 0.8, which indicates that $\mathcal{L}_w$ allows the background to be activated with low response, as presented in Fig. 1 (b). While using a higher visualization threshold to filter out the background region will reduce the connectivity of localization map generated by SAT w/ $\mathcal{L}_w$ in Fig. 1 (a), resulting in decreased localization performance. Therefore, $\mathcal{L}_{norm}$ is more suitable for the proposed SAT and SAT w/ $\mathcal{L}_{norm}$ achieves the best results in Table 4.

	Method	Cls Acc.		Loc Acc.		
		Top-1	Top-5	Top-1	Top-5	GT-k.
(a)	SAT w/ $\mathcal{L}_w$	81.64	95.22	78.68	91.82	96.27
(b)	SAT w/ $\mathcal{L}_{norm}$	**82.05**	**95.56**	**80.96**	**94.13**	**98.45**

Table 4. **Normalization loss** $w.r.t.$ **weighed entropy loss.** The accuracy of our method using normalization loss $\mathcal{L}_{norm}$ and weighed entropy $\mathcal{L}_w$ loss on CUB-200, respectively.

Batch area loss $w.r.t.$ **area loss.** In Table 5, we compare the proposed batch area loss $\mathcal{L}_{ba}$ with the area loss $\mathcal{L}_{area}$ in FPM-based [13, 20, 21] on CUB-200. Experiments show that $\mathcal{L}_{area}$ cannot be well applied to SAT either in one-stage or two-stage. This is because the generation and learning of the localization map in the SAT occur in the attention module, while in transformer, the token sequence input to the attention module can be propagated to the next layer by the skip-connection, which makes the area loss not suitable for SAT. For this reason, we propose batch area loss, which not only provides a

	Method	Stage	Cls Acc.		Loc Acc.		
			Top-1	Top-5	Top-1	Top-5	GT-k.
(a)	SAT w/ $\mathcal{L}_{area}$	one-stage	79.39	95.41	25.68	32.83	34.92
(b)	SAT w/ $\mathcal{L}_{area}$	two-stage	80.19	94.51	57.01	66.79	70.54
(c)	SAT w/ $\mathcal{L}_{ba}$	one-stage	**82.05**	**95.56**	**80.96**	**94.13**	**98.45**

(a) (b) (c)

Table 5. **batch area loss** $w.r.t.$ **area loss.** One-stage indicates training the model in an end-to-end manner. Two-stage means first training the network with classification losses only. Then the weights of the backbone are fixed and only the spatial token is trained with all losses.

sparse area supervision for the localization maps, but also guarantees the tolerance of area variation between instances. The visualization results and accuracy verify the effectiveness of the proposed batch area loss.

3. Performance

Tunable parameters. We detail the tunable parameters for freezing different parts in Table 6, where we follow the freezing settings of Table 8 in the main text. When freezing 81% of the parameters, only 1.4M parameters are tunable on the backbone network, which is **6%** of the parameters in the entire backbone network (21.7M). In this case, SAT still exceeds the existing transformer-based approaches in both classification and localization with only 4.8M tunable parameters , which verifies the efficiency and effectiveness of the proposed method.

Methods	Frozen Rate	Tunable Parameters		Inference		Accuracy		
		Backbone	Head	FLOPs	Parameters	Top-1 Cls	Top-1 Loc	GT-k. Loc
TS-CAM [5]	0%	21.7M	3.4M	4.9G	25.1M	74.30	53.40	67.60
LCTR [2]	0%	21.7M	15.1M	7.2G	36.8M	77.10	56.10	68.70
SCM [1]	0%	21.7M	3.4M	4.9G	25.1M	76.70	56.10	68.80
SAT (e)	81%	**1.4M**	3.4M	4.9G	25.1M	77.79	58.29	71.14
SAT (d)	64%	5.8M	3.4M	4.9G	25.1M	78.23	59.37	72.20
SAT (c)	42%	11.1M	3.4M	4.9G	25.1M	78.24	59.97	72.88
SAT (b)	21%	16.4M	3.4M	4.9G	25.1M	78.12	60.00	73.10
SAT (a)	0%	21.7M	3.4M	4.9G	25.1M	**78.41**	**60.15**	**73.13**

Table 6. **Tunable parameters.** The frozen parts are as follows: **(a)** None. **(b)** Attention layer of transformer blocks. **(c)** MLP layer of transformer blocks. **(d)** Transformer blocks. **(e)** Position embedding, projection, transformer blocks, and MLP layer of spatial aware transformer blocks.

Fine-grained. To further validate the effectiveness of SAT, we compare the accuracy of SAT with TS-CAM [5] on three fine-grained datasets, including Standford Dogs [7], FGVC-Aircraft [12], and Standford Cars [9], as shown in Table 7. On Standford Dogs, we achieve significant gains of **17.83%** and **17.47%** on Top-1 Loc and GT-known Loc compared to TS-CAM. Besides, we obtain **98.80%** and **99.76%** GT-known Loc on FGVC-Aircraft and Standford Cars, exceeding TS-CAM by **2.07%** and **4.14%**, respectively. Fig. 2 illustrates several visual comparisons between TS-CAM and our proposed method on three fine-grained datasets. Compared to TS-CAM, the localization results generated by the proposed method have better visualization and more complete coverage of the object.

Dataset	Method	Cls Acc		Loc Acc		
		Top-1	Top-5	Top-1	Top-5	GT-known
Standford Dog [7]	TS-CAM	81.24	97.25	65.14	77.19	78.67
	SAT	**86.03** (+4.79)	**98.61** (+1.36)	**82.97** (+17.83)	**94.92** (+17.73)	**96.14** (+17.47)
FGVC-Aircraft [12]	TS-CAM	81.28	95.41	79.69	93.22	96.73
	SAT	**82.66** (+1.38)	**95.89** (+0.48)	**82.18** (+2.49)	**95.23**(+2.01)	**98.80** (+2.07)
Standford Cars [9]	TS-CAM	83.16	96.48	79.74	92.43	95.62
	SAT	**85.92** (+2.76)	**97.55** (+1.07)	**85.79** (+5.95)	**97.35** (+4.98)	**99.76** (+4.14)

Table 7. **Fine-grained.** Comparison with TS-CAM method on three fine-grained datasets.

Error analysis. To further analyze the effect of the proposed method, we count all the localization errors (90 images) on CUB-200 [17] test set (5,794 images) and classify them according to the error causes. As listed in Table 8, we classify the

error causes into the following six categories, including **object occlusion** (36 images), **localization more** (28 images), **water reflection** (18 images), **localization part** (5 images), **multiple instances** (2 images), **label error** (1 image). Specifically, object occlusion causes the object to be split into two or more parts, resulting in incomplete localization results, as shown in Table 8. Localization more is often due to the positive effect of co-occurrence context on the classification network, leading to localizing confounding background regions. In addition, water reflection is an inherent challenge for weakly supervised object localization, and it is difficult to achieve correct localization results with only image-level labels. In this way, to achieve better localization performance, future work needs to take more into account the interaction between objects and background to overcome the problems of object occlusion and localization more.

Total Errors	Object Occlusion	Localization More	Water Reflection	Localization Part	Multiple Instances	Label Error
90	36	28	18	5	2	1

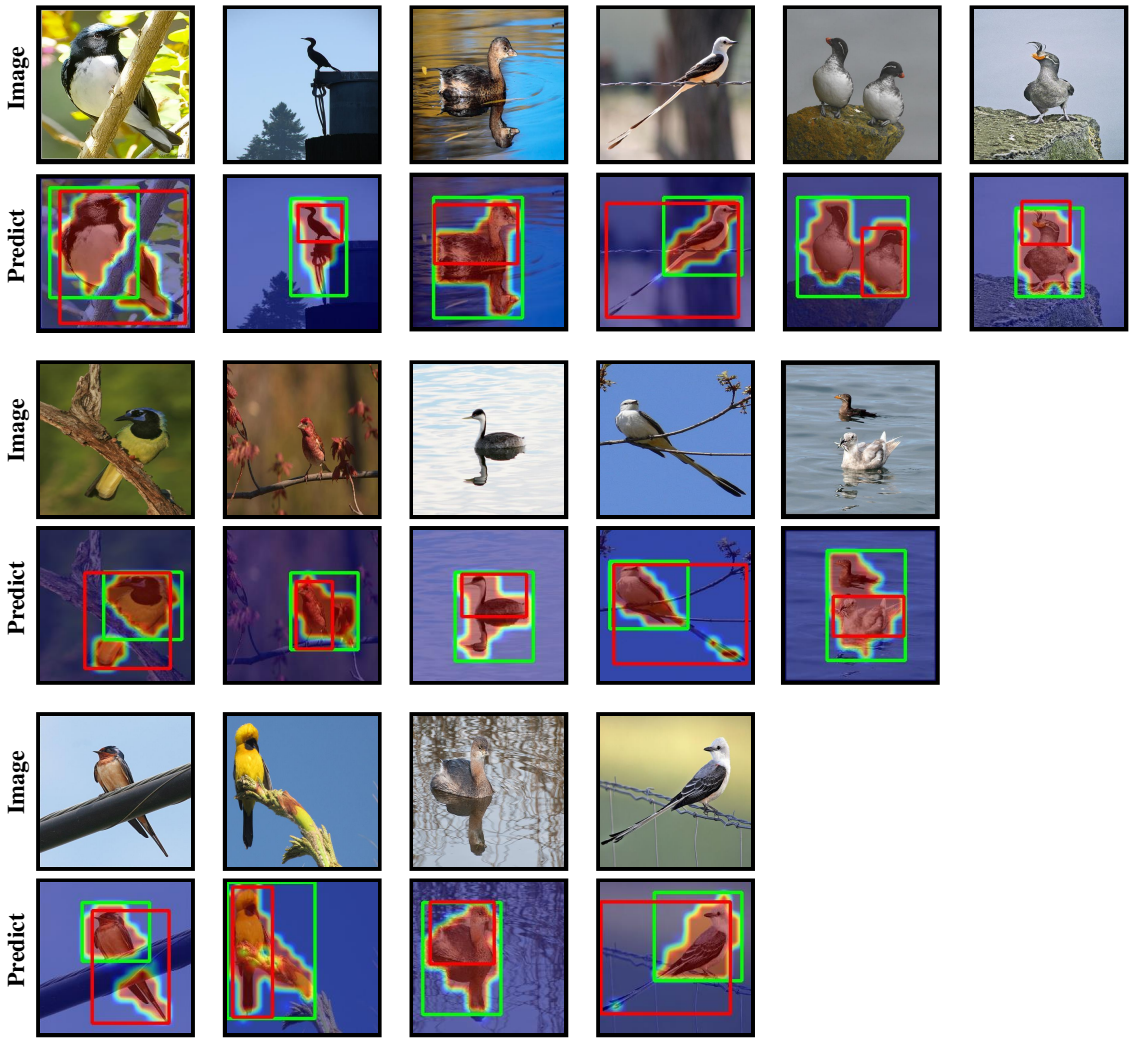

Table 8. **Localization error analysis** on CUB-200.

Methods	CUB-200 Loc Acc.			ImageNet Loc Acc.		
	Top-1	Top-5	GT-k.	Top-1	Top-5	GT-k.
PSOL*	72.45*	87.48*	90.00	54.71*	63.54*	65.44
SPOL*	80.73*	93.76*	96.46	59.89*	67.68*	69.02
SAT	**80.96**	**94.13**	**98.45**	**60.15**	**70.52**	**73.13**

Table 9. **Reproducing the convnet-based methods** on the Deit-S. * indicates the reproduced results.

Comparison with convnet-based methods. We replace the classifiers of PSOL [24] and SPOL [18] with Deit-S [16] backbone and report the reproduced localization results in the Table 9. Compared to the above methods, SAT still achieves the best localization results on both benchmarks.

Main results. In Table 10, we show the more complete comparison results with other SOTA methods on CUB-200 [17] and ImageNet [15]. It can be seen that the proposed SAT achieves the best performance on both datasets in terms of Top-1/Top-5/GT-known Loc three localization metrics.

Visual Results. More visualizations on OpenImages [3], CUB-200 [17], and ImageNet [15] datasets are shown in Fig. 3, Fig. 4, and Fig. 5, respectively. It can be noted that SAT demonstrates robust localization ability in various challenging scenarios, including different scaled objects, complex environments, and object occlusions.

Methods	Venue	Backbone	CUB-200 [17] Loc Acc.			ImageNet [15] Loc Acc.		
			Top-1	Top-5	GT-known	Top-1	Top-5	GT-known
CAM [28]	CVPR16	VGG16	41.06	50.66	55.10	42.80	54.86	59.00
ACoL [25]	CVPR18	VGG16	45.92	56.51	62.96	45.83	59.43	62.96
ADL [4]	CVPR19	VGG16	52.36	–	75.41	44.92	–	–
DANet [23]	ICCV19	VGG16	52.52	61.96	67.70	–	–	–
I2C [27]	ECCV20	VGG16	55.99	68.34	–	47.41	58.51	63.90
MEIL [11]	CVPR20	VGG16	57.46	–	73.84	46.81	–	–
SLT [6]	CVPR21	VGG16	67.80	–	87.60	51.20	62.40	67.20
ORNet [21]	ICCV21	VGG16	67.73	80.77	86.20	52.05	63.94	68.27
BAS [20]	CVPR22	VGG16	71.33	85.33	91.07	52.96	65.41	69.64
Kim et al. [8]	CVPR22	VGG16	70.83	88.07	93.17	49.94	63.25	68.92
CREAM [22]	CVPR22	VGG16	70.44	85.67	90.98	52.37	64.20	68.32
CAM [28]	CVPR16	InceptionV3	41.06	50.66	55.10	46.29	58.19	62.68
SPG [26]	ECCV18	InceptionV3	46.64	57.72	–	48.60	60.00	64.69
DANet [23]	ICCV19	InceptionV3	49.45	60.46	67.03	47.53	58.28	–
I2C [27]	ECCV20	InceptionV3	55.99	68.34	72.60	53.11	64.13	68.50
GCNet [10]	ECCV20	InceptionV3	58.58	71.00	75.30	49.06	58.09	–
SPA [14]	CVPR21	InceptionV3	53.59	66.50	72.14	52.73	64.27	68.33
FAM [13]	ICCV21	InceptionV3	70.67	–	87.25	55.24	–	68.62
CREAM [22]	CVPR22	InceptionV3	71.76	86.37	90.43	56.07	66.19	69.03
BAS [20]	CVPR22	InceptionV3	73.29	86.31	92.24	58.51	<u>69.00</u>	71.93
BagCAMs [29]	ECCV22	InceptionV3	60.07	–	89.78	53.87	–	71.02
CAM [28]	CVPR16	ResNet50	46.71	54.44	57.35	38.99	49.47	51.86
ADL [4]	CVPR19	ResNet50	62.29	–	–	48.53	–	–
I2C [27]	ECCV20	ResNet50	–	–	–	51.83	64.60	68.50
PSOL [24]	CVPR20	ResNet50	70.68	86.64	90.00	53.98	63.08	65.44
FAM [13]	ICCV21	ResNet50	73.74	–	85.73	54.46	–	64.56
SPOL [18]	CVPR21	ResNet50	80.12	93.44	96.46	59.14	67.15	69.02
DA-WSOL [30]	CVPR22	ResNet50	66.65	–	81.83	55.84	–	70.27
BAS [20]	CVPR22	ResNet50	77.25	90.08	95.13	57.18	68.44	71.77
Kim et al. [8]	CVPR22	ResNet50	73.16	86.68	91.60	53.76	65.75	69.89
CREAM [22]	CVPR22	ResNet50	76.03	–	89.88	55.66	–	69.31
BagCAMs [29]	ECCV22	ResNet50	69.67	–	94.01	44.24	–	<u>72.08</u>
ISIC [19]	ECCV22	ResNet50	<u>80.68</u>	<u>94.08</u>	<u>97.32</u>	<u>59.61</u>	67.84	70.01
TS-CAM [5]	ICCV21	Deit-S	71.30	83.80	87.70	53.40	64.30	67.60
LCTR [2]	AAAI22	Deit-S	79.20	89.90	92.40	56.10	65.80	68.70
SCM [1]	ECCV22	Deit-S	76.40	91.60	96.60	56.10	66.40	68.80
SAT (ours)	This Work	Deit-S	**80.96**	**94.13**	**98.45**	**60.15**	**70.52**	**73.13**

Table 10. **Comparison with state-of-the-art methods.** The best results are highlighted in **bold**, second are <u>underlined</u>.

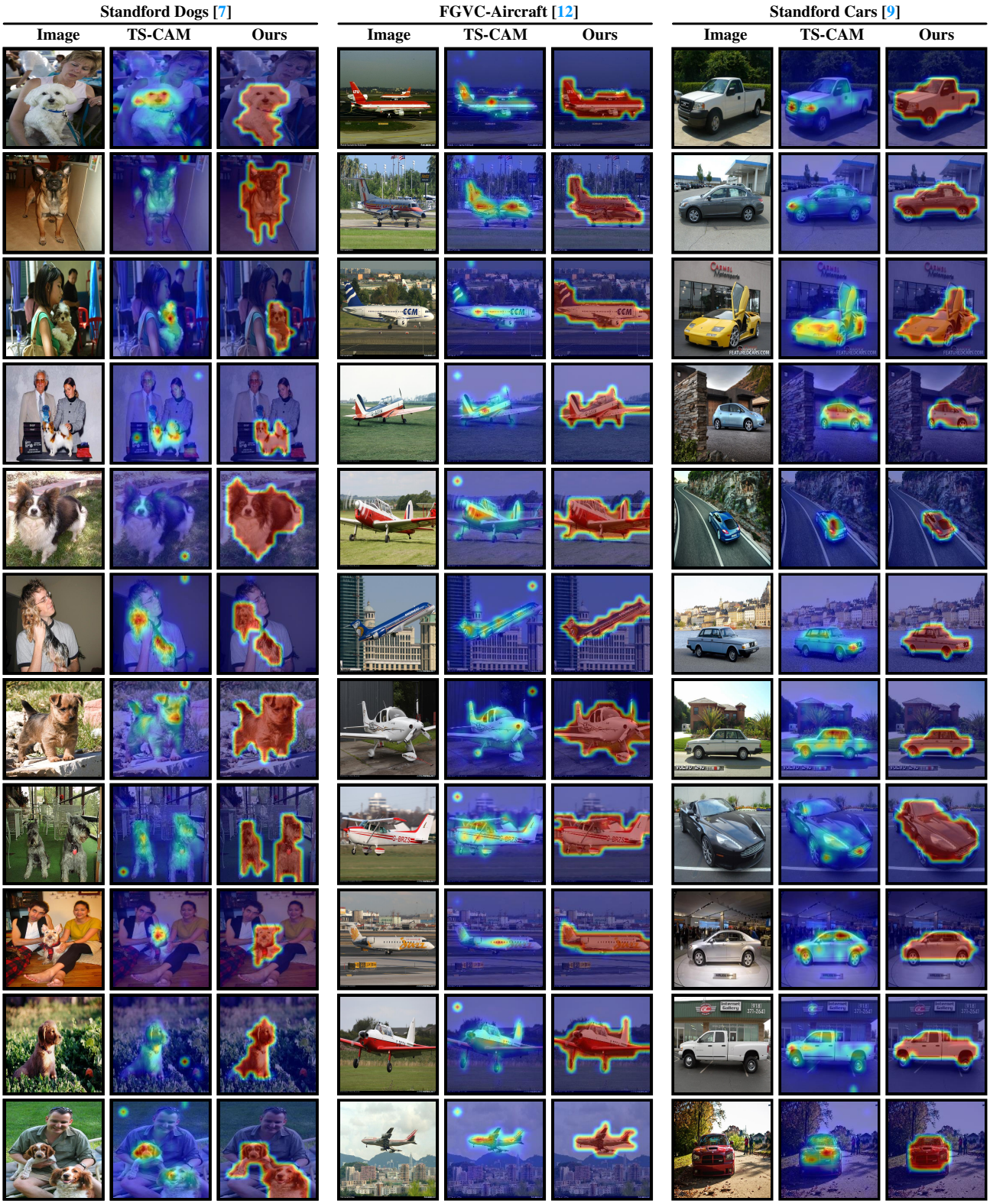

Figure 2. Visualization comparison with the baseline TS-CAM [5] method on Standford Dog [7], FGVC-Aircraft [12], and Standford Cars [9].

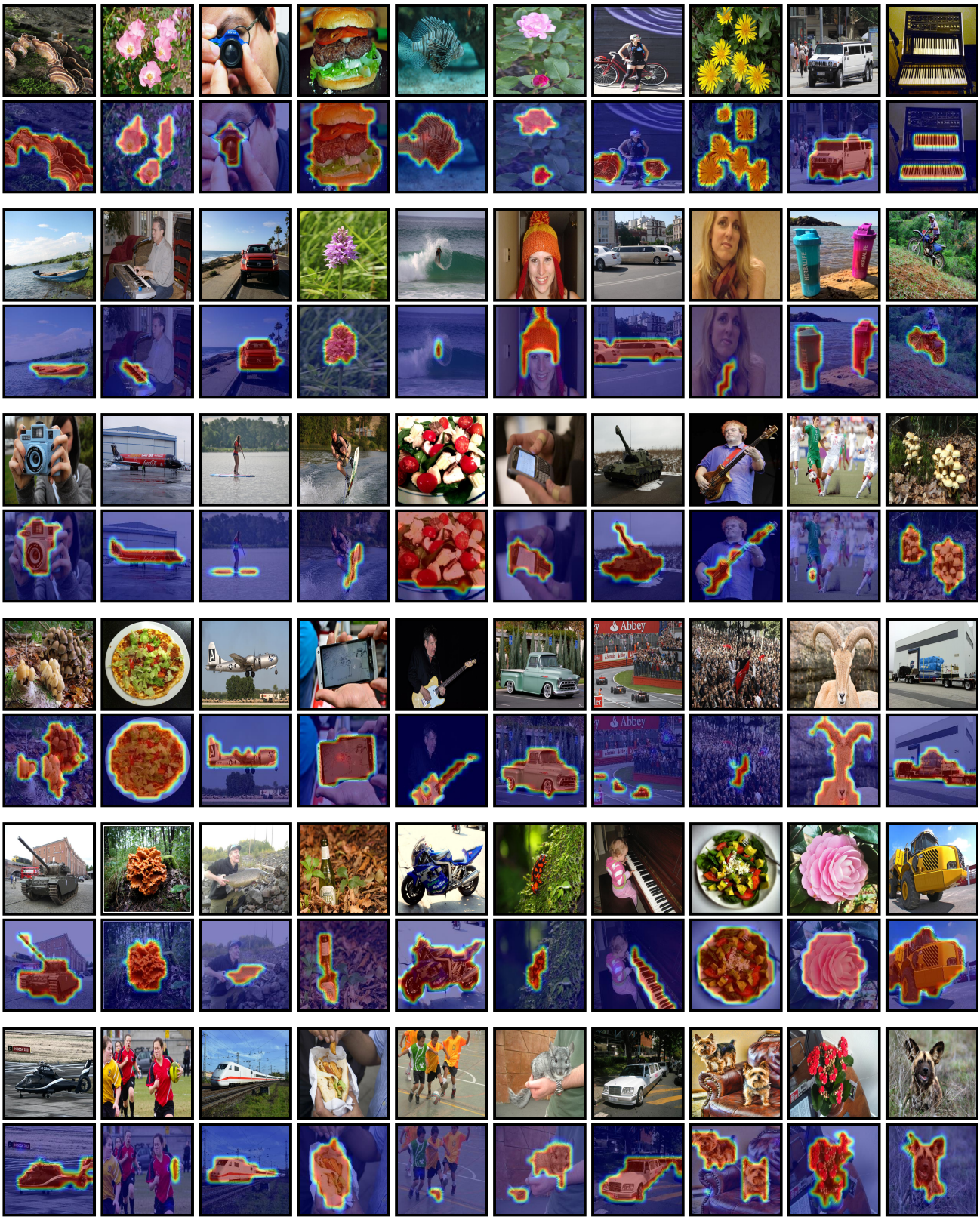

Figure 3. Visualization of the localization results on OpenImages [3].

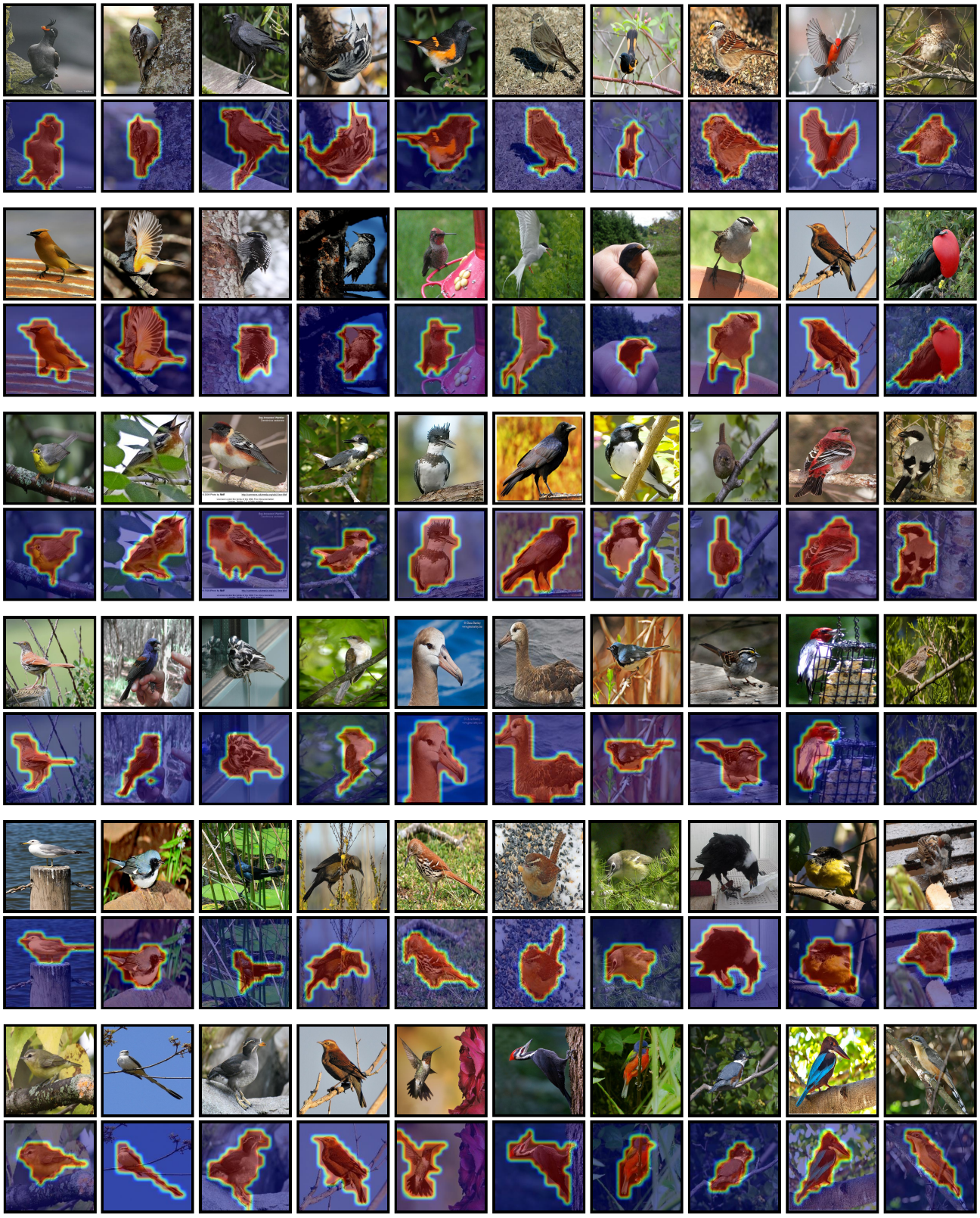

Figure 4. Visualization of the localization results on CUB-200 [15].

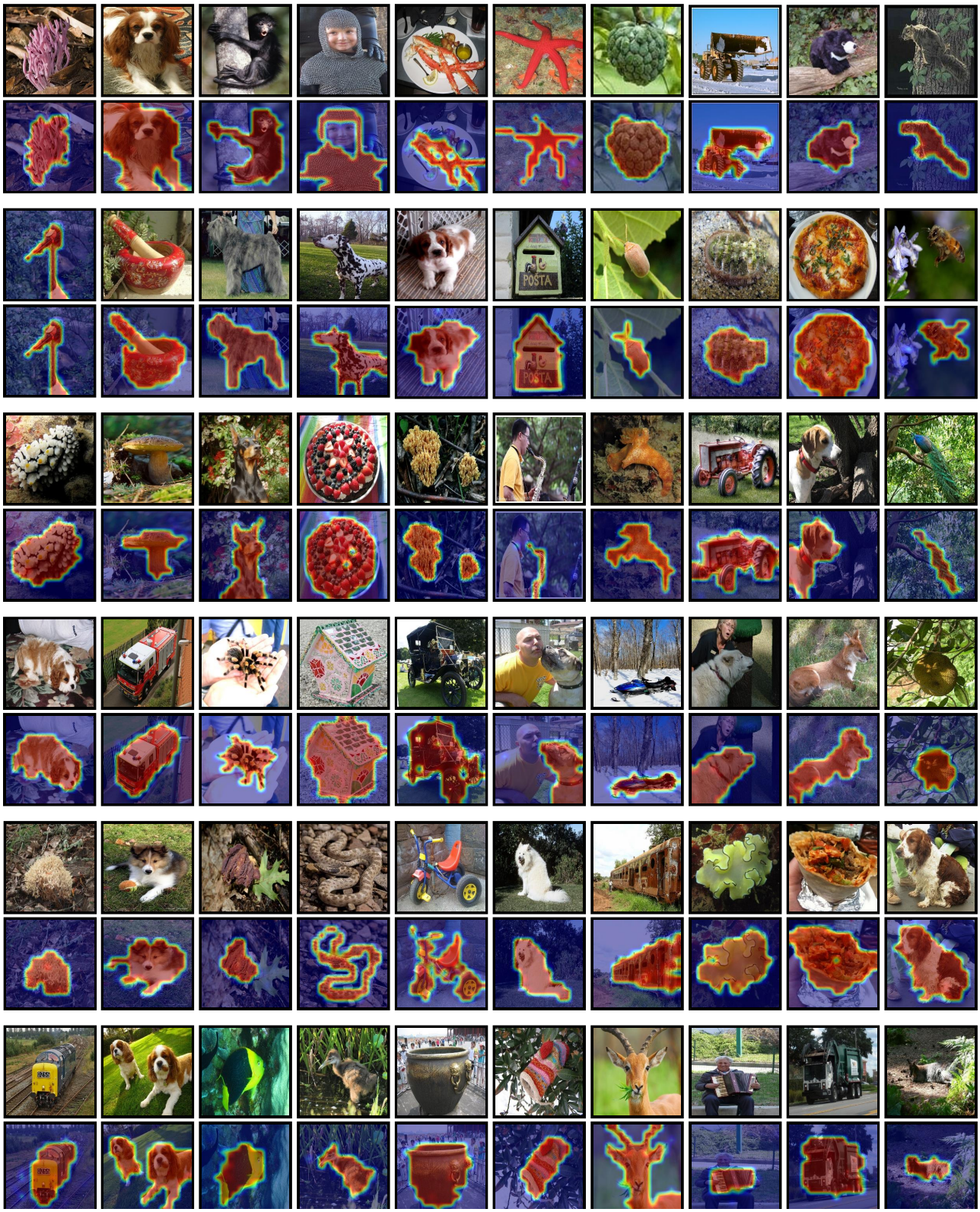

Figure 5. Visualization of the localization results on ImageNet [17].

References

[1] Haotian Bai, Ruimao Zhang, Jiong Wang, and Xiang Wan. Weakly supervised object localization via transformer with implicit spatial calibration. *arXiv preprint arXiv:2207.10447*, 2022. 3, 5

[2] Zhiwei Chen, Changan Wang, Yabiao Wang, Guannan Jiang, Yunhang Shen, Ying Tai, Chengjie Wang, Wei Zhang, and Liujuan Cao. Lctr: On awakening the local continuity of transformer for weakly supervised object localization. *arXiv preprint arXiv:2112.05291*, 2021. 3, 5

[3] 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. 5, 7

[4] 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

[5] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2886–2895, 2021. 3, 5, 6

[6] 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

[7] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2. Citeseer, 2011. 3, 6

[8] Eunji Kim, Siwon Kim, Jungbeom Lee, Hyunwoo Kim, and Sungroh Yoon. Bridging the gap between classification and localization for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14258–14267, 2022. 5

[9] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 3, 6

[10] 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

[11] 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

[12] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 3, 6

[13] 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. 2, 5

[14] 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. 5

[15] 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, 8

[16] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 5

[17] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 3, 5, 9

[18] 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

[19] Jun Wei, Sheng Wang, S Kevin Zhou, Shuguang Cui, and Zhen Li. Weakly supervised object localization through inter-class feature similarity and intra-class appearance consistency. In *European Conference on Computer Vision*, pages 195–210. Springer, 2022. 5

[20] Pingyu Wu, Wei Zhai, and Yang Cao. Background activation suppression for weakly supervised object localization. *arXiv preprint arXiv:2112.00580*, 2021. 2, 5

[21] 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. 2, 5

[22] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Rui-Wei Zhao, Tao Zhang, Xuequan Lu, and Shang Gao. Cream: Weakly supervised object localization via class re-activation mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9437–9446, 2022. 5

[23] 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

[24] 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

[25] 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

[26] 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

[27] 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

[28] 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

[29] Lei Zhu, Qian Chen, Lujia Jin, Yunfei You, and Yanye Lu. Bagging regional classification activation maps for weakly supervised object localization. *arXiv preprint arXiv:2207.07818*, 2022. 5

[30] Lei Zhu, Qi She, Qian Chen, Yunfei You, Boyu Wang, and Yanye Lu. Weakly supervised object localization as domain adaption. *arXiv preprint arXiv:2203.01714*, 2022. 5