Category-aware Allocation Transformer for Weakly Supervised Object Localization — *Supplementary Material*

Zhiwei Chen[1] Jinren Ding[1] Liujuan Cao[1*] Yunhang Shen[2]
Shengchuan Zhang[1] Guannan Jiang[3] Rongrong Ji[1]
[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China. [2]Tencent Youtu Lab, China. [3]CATL, China.
zhiweichen.cn@gmail.com, dingjinren@stu.xmu.edu.cn, odysseyshen@tencent.com,
{caoliujuan, zsc_2016, rrji}@xmu.edu.cn, jianggn@catl.com

In the supplementary material, we present additional results for CATR on the CUB-200-2011 [13] and ILSVRC [12] datasets in Appendix A. Additionally, we provide additional ablation studies on ILSVRC [12] in Appendix B. Moreover, we compare the parameter complexity of CATR with other methods on CUB-200-2011 [13] in Appendix C. Finally, we analyze the limitation of our method in Appendix D and present more localization visualizations of the proposed CATR in Appendix E.

A. Additional results

A.1. Complete Performance Comparison

Tab. 2 and Tab. 3 present additional performance comparisons of CATR with state-of-the-art methods on the CUB-2011-200 test set and the ILSVRC validation set, respectively. The results demonstrate that CATR achieves competitive classification accuracy. Specifically, on the CUB-200-2011 dataset, CATR reaches 83.72% Top-1 Cls and 96.82% Top-5 Cls, which are slightly lower than LCTR [2]. On the ILSVRC dataset, our method achieves the best 93.64% Top-5 Cls and is slightly lower than FAM [10] in Top-1 Cls (77.25% vs 77.63%). While fewer images are correctly classified, our method still achieves higher localization accuracy than LCTR [2] and FAM [10], which further suggests the effectiveness of CATR. Note that Top-1 and Top-5 Loc consider both localization and classification accuracy, *i.e.*, a prediction is correct only if both localization and classification are correct.

A.2. Additional Performance Comparison

Tab. 4 presents a comparison of CATR with methods that adopt a separate localization-classification pipeline. It is worth noting that these multi-stage approaches achieve remarkable results, but require separate networks for localization and classification that must undergo distinct training phases. For instance, SPOL [14] employs three distinct networks for WSOL. The first two separate modified ResNet50 are used for generating class activation maps and foreground segmentation, respectively. An additional DenseNet161/EffcientNet-B7 is then employed for classification. In contrast, the proposed CATR utilizes only one network, offering significant advantages in terms of efficiency. Additionally, CATR achieves competitive localization performance, *e.g.*, 79.62% vs 80.68% Top-1 Loc on the CUB-200-2011 dataset.

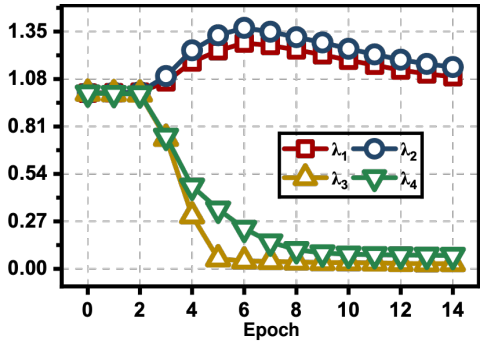

Figure 1. Analyses of loss weights in the training phase.

Method	Input size	#Params. (M)	MACs (G)	Top-1 Loc. (%)
VGG16-CAM [22]	224^2	19.64	16.31	44.15
DeitS-TS-CAM [4]	224^2	25.10	5.29	71.30
DeitS-TS-CAM∗ [4]	224^2	22.36	4.38	73.10
DeitS-LCTR [2]	224^2	25.76	4.50	79.20
DeitS-CATR (Ours)	224^2	22.92	4.49	**79.62**

Table 1. Comparison of parameters and MACs. Note that Top-1 Loc. is evaluated on the CUB-200-2011 test set, ∗ indicates the re-implement method.

*Corresponding author.

Methods	Backbone	Loc. Acc			Cls. Acc	
		Top-1	Top-5	GT-known	Top-1	Top-5
CAM [22]	GoogLeNet	41.06	50.66	55.10	73.80	91.50
DANet [18]	GoogLeNet	49.45	60.46	67.03	71.20	90.60
ADL [3]	InceptionV3	53.04	–	–	74.55	–
SPA [11]	InceptionV3	53.59	66.50	72.14	73.51	91.39
FAM [10]	InceptionV3	70.67	–	87.25	81.25	–
CAM [22]	VGG16	44.15	52.16	56.00	76.60	92.50
ADL [3]	VGG16	52.36	–	75.41	65.27	–
ACoL [20]	VGG16	45.92	56.51	62.96	71.90	–
DANet [18]	VGG16	52.52	61.96	67.70	75.40	92.30
MEIL [9]	VGG16	57.46	–	73.84	74.77	–
SPA [11]	VGG16	60.27	72.50	77.29	76.11	92.15
FAM [10]	VGG16	69.26	–	89.26	77.26	–
ORNet [17]	VGG16	67.73	80.77	86.20	77.00	93.00
BAS [16]	VGG16	71.33	85.33	91.00	77.49	93.18
BGC [6]	VGG16	70.83	88.07	93.17	–	–
TS-CAM [4]	Deit-S	71.30	83.80	87.70	80.30	94.80
LCTR [2]	Deit-S	79.20	89.90	92.40	**85.00**	**97.10**
SCM [1]	Deit-S	76.40	91.60	**96.60**	78.50	94.50
CATR (Ours)	Deit-S	**79.62**	**92.08**	94.94	83.72	96.82

Table 2. Comparison of CATR with the state-of-the-art methods on the CUB-200-2011 [13] test set.

Methods	Backbone	Loc. Acc			Cls. Acc	
		Top-1	Top-5	GT-known	Top-1	Top-5
CAM [22]	VGG16	42.80	54.86	59.00	66.60	88.60
ADL [3]	VGG16	44.92	–	–	–	–
ACoL [20]	VGG16	45.83	59.43	62.96	67.50	88.00
I^2C [21]	VGG16	47.41	58.51	63.90	69.40	89.30
MEIL [9]	VGG16	46.81	–	–	70.27	–
FAM [10]	VGG16	51.96	–	**71.73**	70.90	–
ORNet [17]	VGG16	52.05	63.94	68.27	71.60	90.40
BAS [16]	VGG16	52.96	65.41	69.64	70.84	90.46
BGC [6]	VGG16	49.94	63.25	68.92	–	–
CAM [22]	InceptionV3	46.29	58.19	62.68	73.30	91.80
ADL [3]	InceptionV3	48.71	–	–	72.83	–
DANet [18]	GoogLeNet	47.53	58.28	–	72.50	91.40
I^2C [21]	InceptionV3	53.11	64.13	68.50	73.30	91.60
GC-Net [8]	InceptionV3	49.06	58.09	–	77.40	93.60
SPA [11]	InceptionV3	52.73	64.27	68.33	73.26	91.81
FAM [10]	InceptionV3	55.24	–	68.62	**77.63**	–
TS-CAM [4]	Deit-S	53.40	64.30	67.60	74.30	92.10
LCTR [2]	Deit-S	56.10	65.80	68.70	77.10	93.40
SCM [1]	Deit-S	56.10	66.40	68.80	76.70	93.00
CATR (Ours)	Deit-S	**56.90**	**66.64**	69.25	77.25	**93.64**

Table 3. Comparison of CATR with state-of-the-art methods on the ILSVRC [12] validation set.

B. Additional ablation studies

In Fig. 1, we further show the changes in four parameters using the automatic weighted loss mechanism (ALM) [7] on the ILSVRC [12] dataset. We also observe that λ_2 is consistently the largest during training, which supports the argument that CSM establishes the connection between the attention maps and the specific classes. Moreover, the results suggest that OCM plays a supporting role in refining the object regions, as the values of λ_3 and λ_4 decrease in the training phase.

C. Parameter Complexity

In Tab. 1, we present a comparison of the parameter complexity with other methods. Our method (with 22.92M parameters and 4.49MACs) outperforms the CNN-based VGG16-CAM [22] (with 19.64M parameters and 16.31MACs) by 35.47% (79.62% vs 44.15%). Furthermore, our method performs much better than the benchmark method [4] (79.62% vs 73.10%) with a slightly increased number of parameters (22.92M vs 22.36M). Notebly, compared with LCTR [2], CATR obtains 0.42% improvement of TOP-1 Loc. with a 2.84M parameter reduction.

Methods	Backbone		CUB-200-2011 [13] Loc. Acc			ILSVRC [12] Loc. Acc		
	Localization	Classification	Top-1	Top-5	GT-known	Top-1	Top-5	GT-known
PSOL [19]	InceptionV3	InceptionV3	65.51	83.44	–	54.82	63.25	65.21
PSOL [19]	ResNet50	ResNet50	70.68	86.64	90.00	53.98	63.08	65.44
PSOL [19]	DenseNet161	DenseNet161	74.97	89.12	93.01	55.31	64.18	66.28
PSOL [19]	DenseNet161	EfficientNet-B7	77.44	89.51	93.01	58.00	65.02	66.28
SLT-Net [5]	VGG16	VGG16	67.80	–	87.60	51.20	62.40	67.20
SLT-Net [5]	InceptionV3	InceptionV3	66.10	–	86.50	55.70	65.40	67.60
SPOL [14]	ResNet50†	DenseNet161	79.74	93.69	96.46	56.40	66.48	69.02
SPOL [14]	ResNet50†	EfficientNet-B7	80.12	93.44	96.46	59.14	67.15	69.02
ISIC [15]	ResNet50	ResNet50	**80.68**	**94.08**	**97.32**	**59.61**	**67.84**	**70.01**
CATR (Ours)	Deit-S		79.62	92.08	94.94	56.90	66.64	69.25

Table 4. Comparison with the methods based on a separate localization-classification pipeline. Note that '†' indicates the backbone is modified.

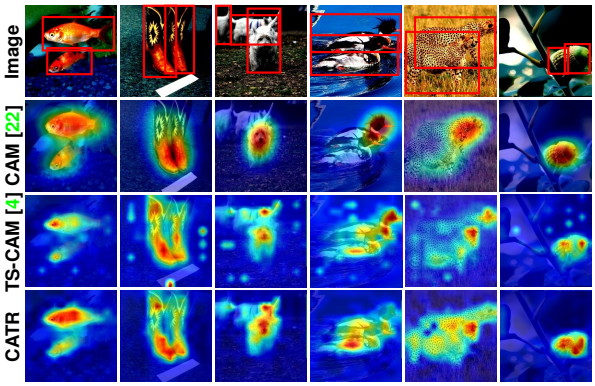

Figure 2. The failure cases on ILSVRC with CAM [22], TS-CAM [4], and our proposed CATR. The ground-truth bounding boxes are depicted in red. All methods demonstrate poor localization performance when there are multiple instances in an image.

D. Limitation

As depicted in Fig. 2, our method exhibits suboptimal performance in separating multiple instances within an image. This is primarily due to the lack of instance-level supervision, which makes it difficult for the weakly supervised object localization task to differentiate between different objects. Therefore, in order to address this issue, it would be worth exploring new methods in future research.

E. Visualization

In Fig. 3, we present additional visualizations of the pseudo maps generated by self-attention maps and utilized to supervise category-aware map learning. It is evident from the visualizations that the pseudo maps contain class-specific features, which effectively highlight the robust object regions.

In Fig. 4, we visualize pure localization maps of our method. We can observe that our method preserves long-range feature dependency well and covers the complete extent of the objects.

In Fig. 5 and Fig. 6, we show the additional localization results on the CUB-2011-200 test set and ILSVRC validation set, respectively. We can observe that the proposed CATR maintains long-range feature dependency effectively and accurately localizes the entire object.

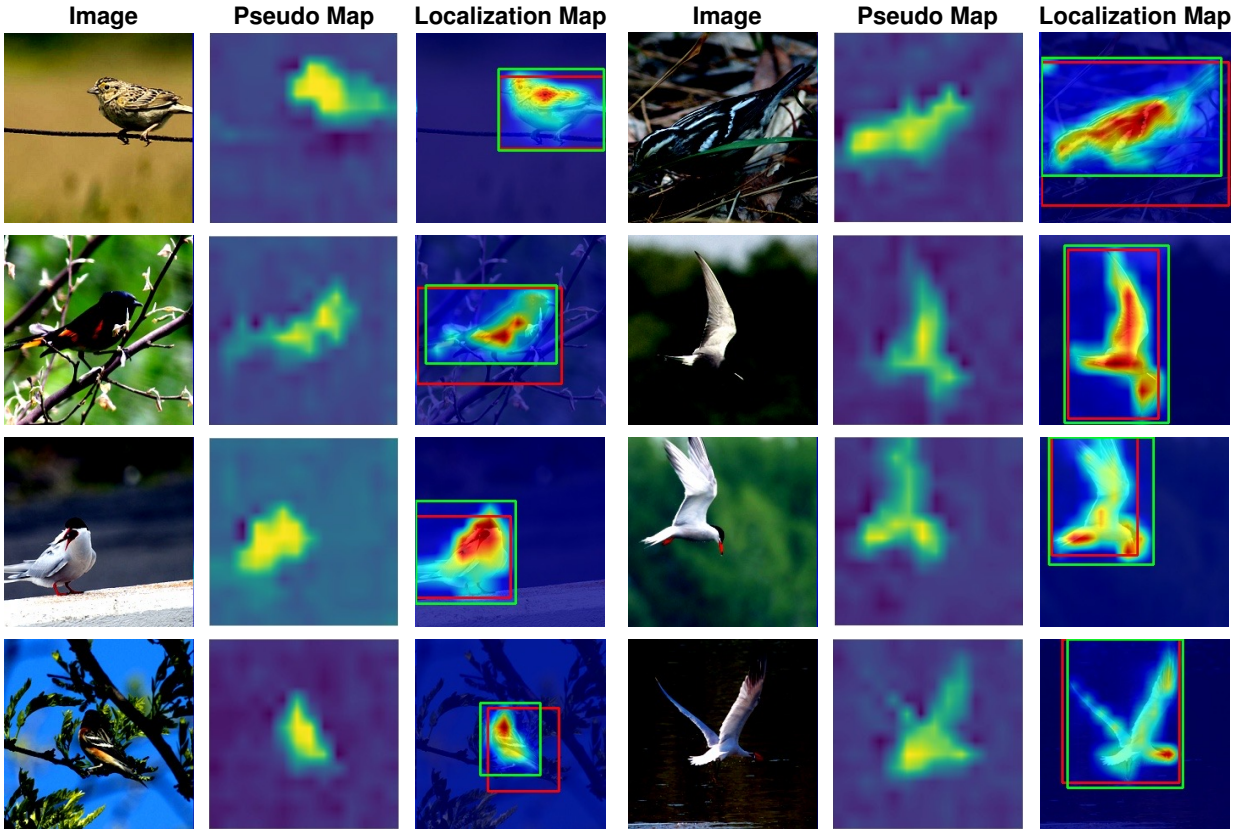

Figure 3. Visualization of the pixel-level pseudo maps $\mathbf{M}_{\mathrm{ocm}}$ and localization maps $\mathbf{M}_{\mathrm{fuse}}$. The ground-truth bounding boxes are highlighted in red, and the predicted bounding boxes are highlighted in green.

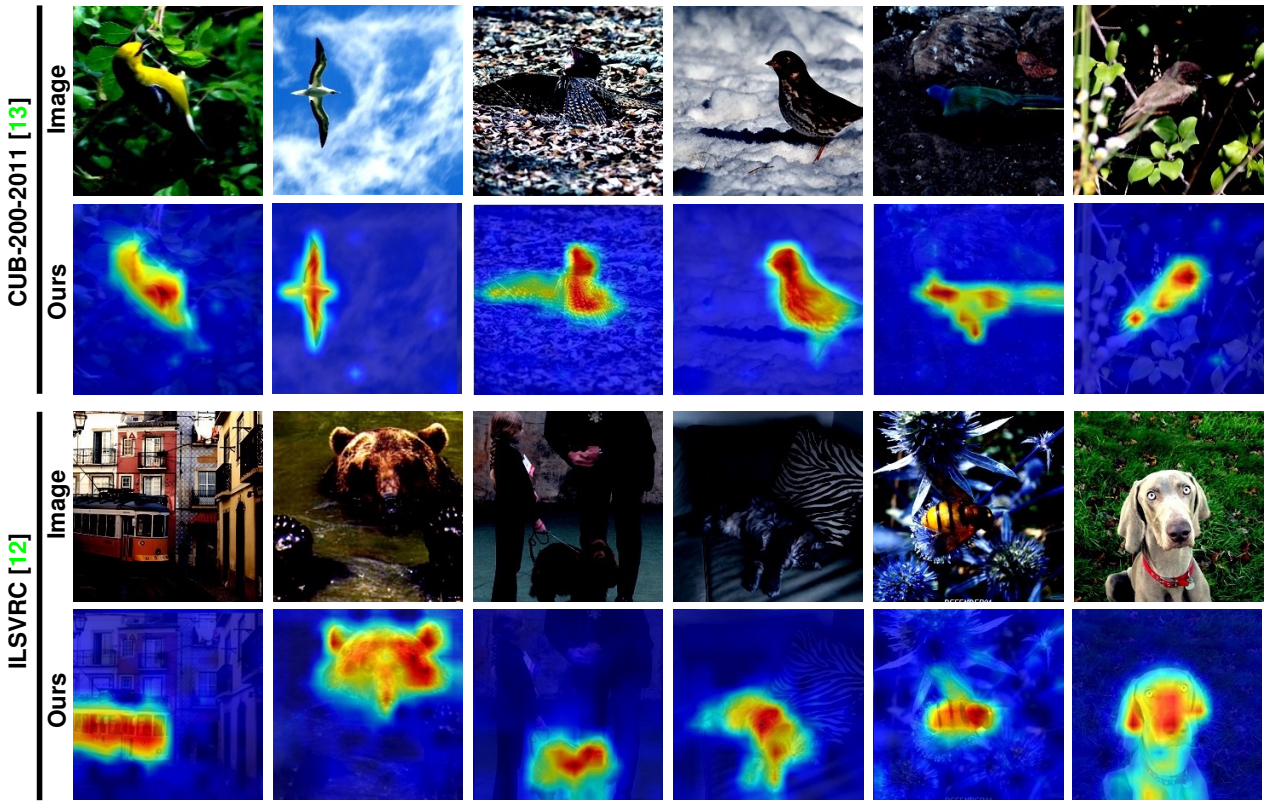

Figure 4. Visualization of localization maps on the CUB-200-2011 [13] and ILSVRC [12] datasets.

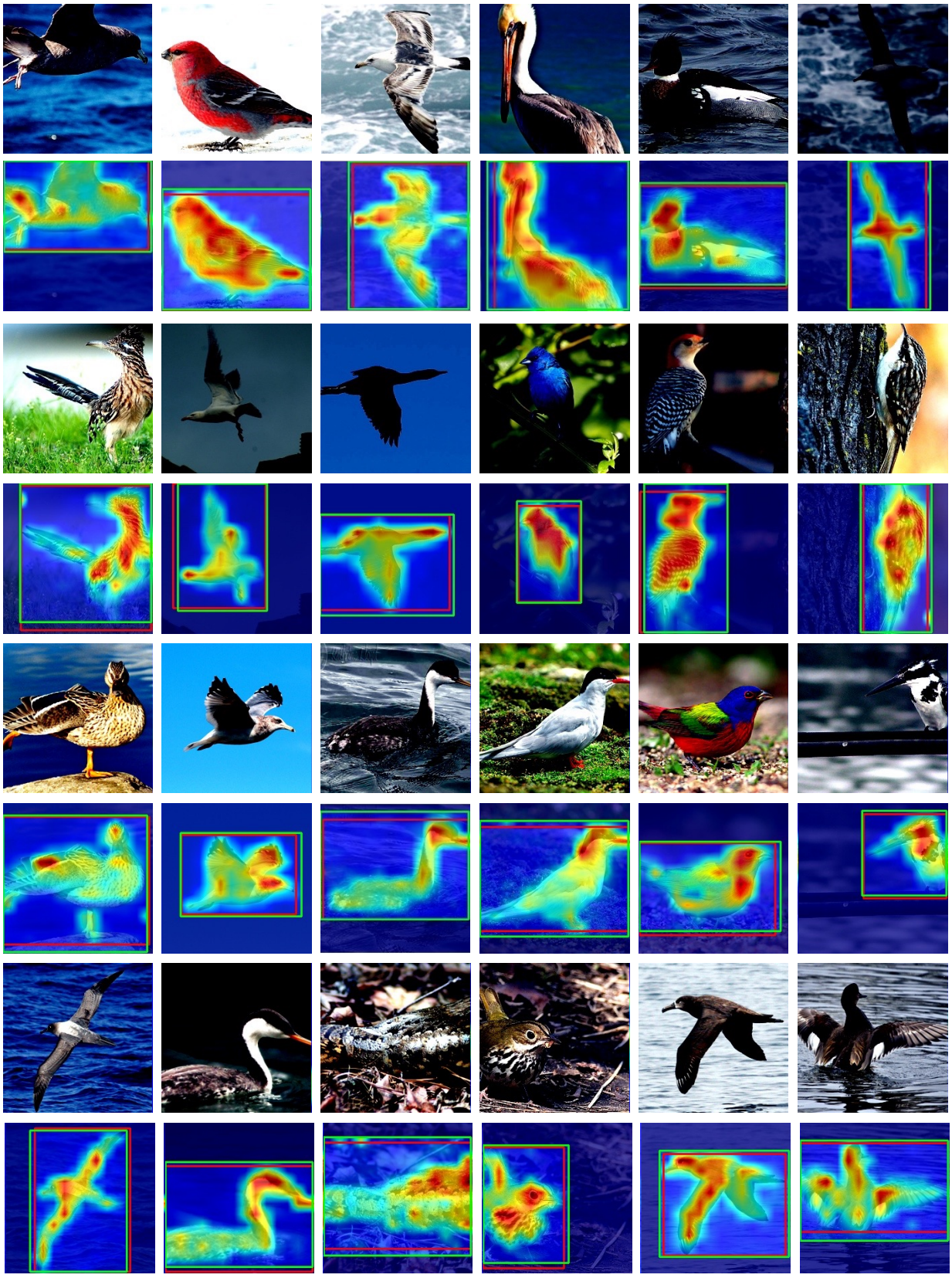

Figure 5. Visualization of the localization results on CUB-200-2011 [13]. The ground-truth bounding boxes are highlighted in red, and the predicted bounding boxes are highlighted in green.

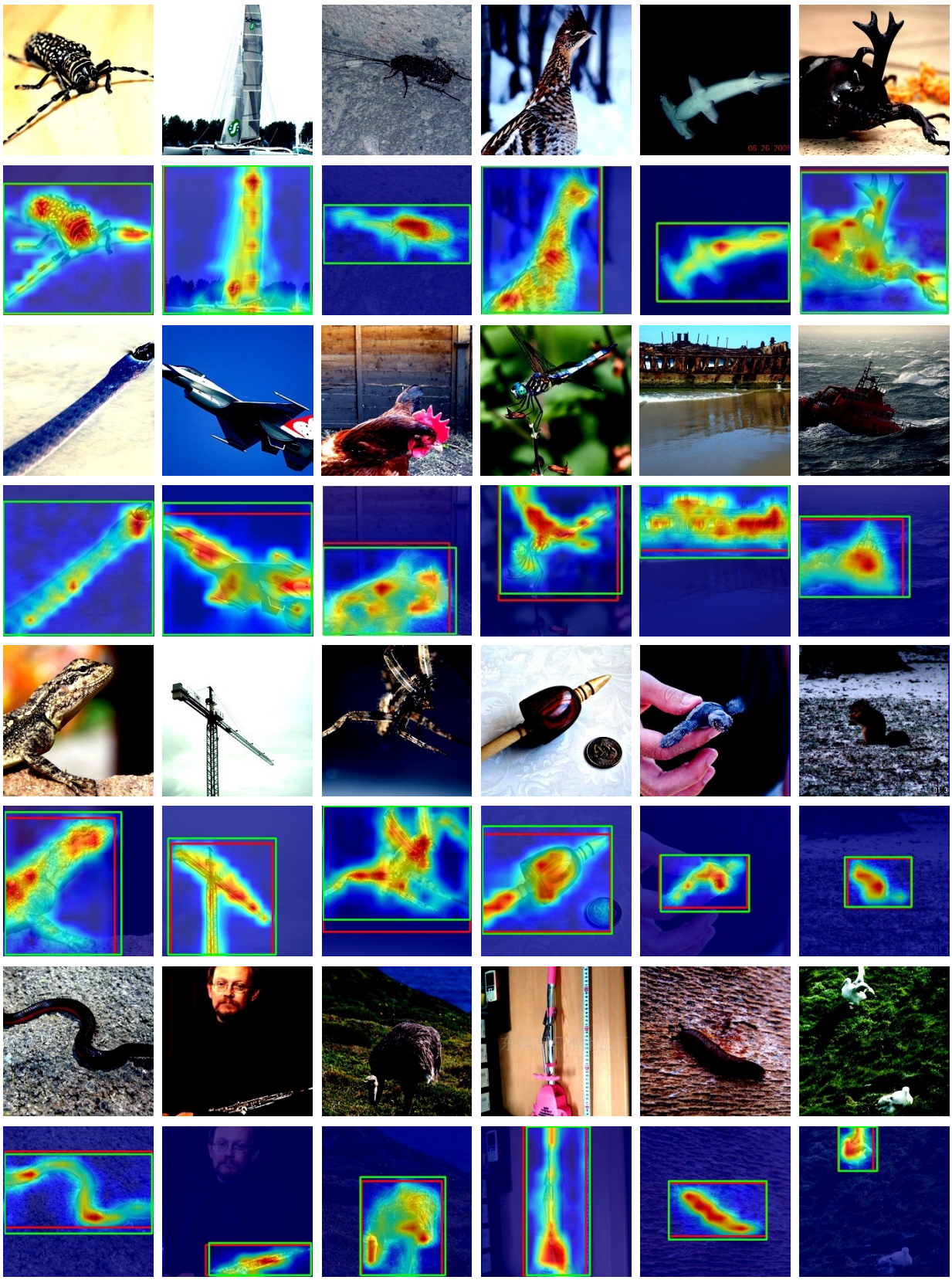

Figure 6. Visualization of the localization results on ILSVRC [12]. The ground-truth bounding boxes are highlighted in red, and the predicted bounding boxes are highlighted in green.

References

[1] Haotian Bai, Ruimao Zhang, Jiong Wang, and Xiang Wan. Weakly supervised object localization via transformer with implicit spatial calibration. *ECCV*, pages 612–628, 2022. 2

[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. In *AAAI*, volume 36, pages 410–418, 2022. 1, 2

[3] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *IEEE CVPR*, pages 2219–2228, 2019. 2

[4] 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 *IEEE ICCV*, pages 2886–2895, 2021. 1, 2, 3

[5] Guangyu Guo, Junwei Han, Fang Wan, and Dingwen Zhang. Strengthen learning tolerance for weakly supervised object localization. In *IEEE CVPR*, 2021. 3

[6] Eunji Kim, Siwon Kim, Jungbeom Lee, Hyunwoo Kim, and Sungroh Yoon. Bridging the gap between classification and localization for weakly supervised object localization. In *IEEE CVPR*, pages 14258–14267, 2022. 2

[7] Lukas Liebel and Marco Körner. Auxiliary tasks in multi-task learning. *arXiv preprint arXiv:1805.06334*, 2018. 2

[8] Weizeng Lu, Xi Jia, Weicheng Xie, Linlin Shen, Yicong Zhou, and Jinming Duan. Geometry constrained weakly supervised object localization. In *ECCV*, pages 481–496, 2020. 2

[9] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *IEEE CVPR*, pages 8766–8775, 2020. 2

[10] Meng Meng, Tianzhu Zhang, Qi Tian, Yongdong Zhang, and Feng Wu. Foreground activation maps for weakly supervised object localization. In *IEEE ICCV*, pages 3385–3395, 2021. 1, 2

[11] 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 *IEEE CVPR*, pages 11642–11651, 2021. 2

[12] 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. In *IJCV*, volume 115, pages 211–252. Springer, 2015. 1, 2, 3, 5, 7

[13] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. (CNS-TR-2011-001), 2011. 1, 2, 3, 5, 6

[14] Jun Wei, Qin Wang, Zhen Li, Sheng Wang, S Kevin Zhou, and Shuguang Cui. Shallow feature matters for weakly supervised object localization. In *IEEE CVPR*, pages 5993–6001, 2021. 1, 3

[15] 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 *ECCV*, pages 195–210, 2022. 3

[16] Pingyu Wu, Wei Zhai, and Yang Cao. Background activation suppression for weakly supervised object localization. In *IEEE CVPR*, pages 14228–14237, 2022. 2

[17] 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 *IEEE CVPR*, pages 132–141, 2021. 2

[18] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *IEEE ICCV*, pages 6589–6598, 2019. 2

[19] Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu. Rethinking the route towards weakly supervised object localization. In *IEEE CVPR*, pages 13460–13469, 2020. 3

[20] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *IEEE CVPR*, pages 1325–1334, 2018. 2

[21] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. In *ECCV*, pages 271–287, 2020. 2

[22] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE CVPR*, pages 2921–2929, 2016. 1, 2, 3