Online Refinement of Low-level Feature Based Activation Map for Weakly Supervised Object Localization

Jinheng Xie, Cheng Luo, Xiangping Zhu, Ziqi Jin, Weizeng Lu, Linlin Shen*
School of Computer Science & Software Engineering, Shenzhen University, China
Shenzhen Institute of Artificial Intelligence of Robotics of Society, Shenzhen, China
Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, China
xiejinheng2020@email.szu.edu.cn, xiangping.zhu2010@gmail.com, llshen@szu.edu.cn

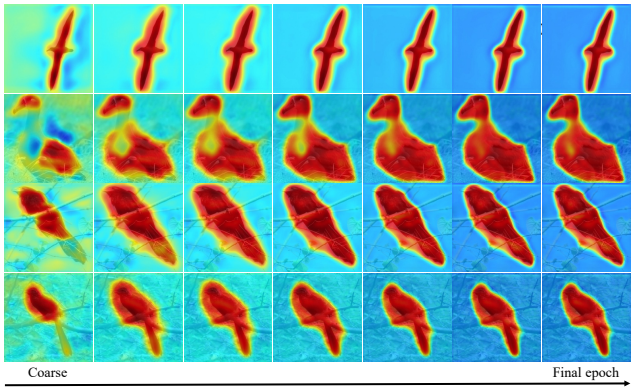

Figure 1: Activation map evolution along with the training epochs in the second stage, *i.e.*, activation map refinement. From left to right, it can be found that the complete and well-separated object regions are gradually derived with the going of the training process.

Strategies	$\mathcal{RE}$	$\mathcal{ER}$	$\mathcal{AE}$
CorLoc	84.07	80.55	**86.19**

Table 1: The performance comparisons among different dropout methods.

f_{max}	0.4	0.5	0.6
CorLoc	85.92	**86.19**	85.74

Table 2: The performance comparisons among different maximum dropout area settings.

1. Activation Map Evolution in Training Process

From left to right, Fig. 1 presents the evolution of activation maps in the second stage, *i.e.*, activation map refinement. In the second and third row of the figure, the non-activated regions, *i.e.*, bird neck and tail, are gradually derived. The refined activation maps are more well-separated than the coarse one generated by the first stage. This further proves the effectiveness of our proposed entropy-guided refinement.

2. Visual Results on Additional Datasets

To further validate the effectiveness and generalization ability of our approach, additional experiments on other datasets are also conducted. The localization results of our method are based on the well-derived activation map. Thus, we will only show the visual results of derived activation maps in the following.

Person re-identification datasets. Fig. 2 shows the derived activation maps of our method on different person re-identification datasets (*i.e.* Duke-MTMC [3], Market-1501 [5], and MSMT17 [4]). The first two rows belong to Duke-MTMC, in which it can be found that the human parts are completely activated in the activation maps without background regions, *e.g.* the car and road. Besides, the bag, which is the discriminative feature for person re-identification, can be successfully discovered in the activation maps. Instead of relying on a specific human parsing model, our approach can offer an alternative pedestrian segmentation model with only image-level labels (*i.e.* person ID), which can provide supports for the researches in person re-identification.

Other Fine-grained Classification Datasets. Fig. 3 presents the visual results of our method on Standford Dogs [1] and FGVC-Aircraft [2] datasets. As shown in the figure, the obtained activation maps can capture complete object regions. Besides, it can be found in the figure (eighth column of Standford Dog) that our model can get perfect activation maps even for the case of multiple objects. This further verifies the effectiveness of our method.

3. Erasing Strategy Comparisons

Table 1. provides the comparison between various erasing strategies (*i.e.*, random erasing $\mathcal{RE}$, erasing regions within a restricted rectangle $\mathcal{ER}$ and the proposed Attentive

*Corresponding Author

Figure 2: Examples of the derived activation maps on person re-identification datasets, *i.e.*, Duke-MTMC [3], Market-1501 [5], and MSMT17 [4].

erasing $\mathcal{AE}$). The proposed attentive erasing $\mathcal{AE}$ achieves improvements of 2.12% and 5.64% on CorLoc compared to $\mathcal{RE}$ and $\mathcal{ER}$.

4. Failure Cases

Fig. 4 presents failure cases of our method on CUB-200-2011 and ImageNet-1K datasets. We find that failure cases mainly fall into two categories: (1) Excited object regions are larger than the predicted bounding boxes, which reduces the final localization accuracy; (2) In ImageNet-1K, there are images with multiple objects. Our method treats multiple objects, which are close to each other, as the single object and thus leads to the failure localization results. However, in most cases, our method can locate objects correctly and surpass the existing methods by a large margin on several tested datasets as shown in our experiments.

The defect of the proposed method is mainly caused by the high-frequency noise in the low-level features, e.g., twigs and gravels, which is usually hard to be suppressed in the generation of activation maps. This provides a potential direction that we can try to add some constraints to suppress this noise and generate more robust and fine-grained activation maps.

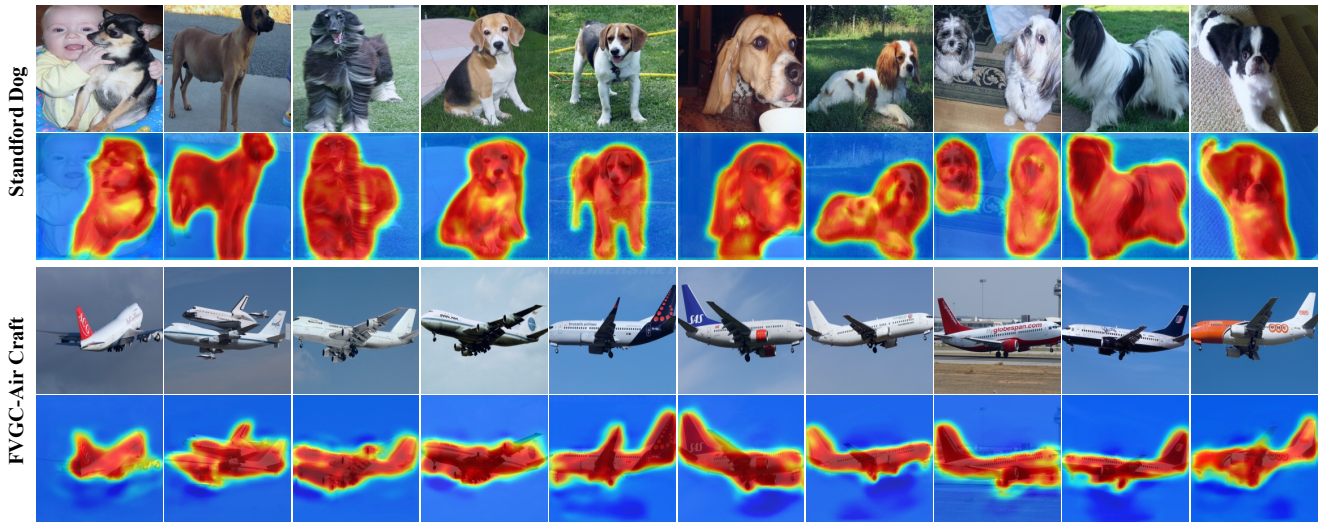

Figure 3: Activation maps derived from our method on Standford Dogs [1] and FGVC-Aircraft [2] datasets.

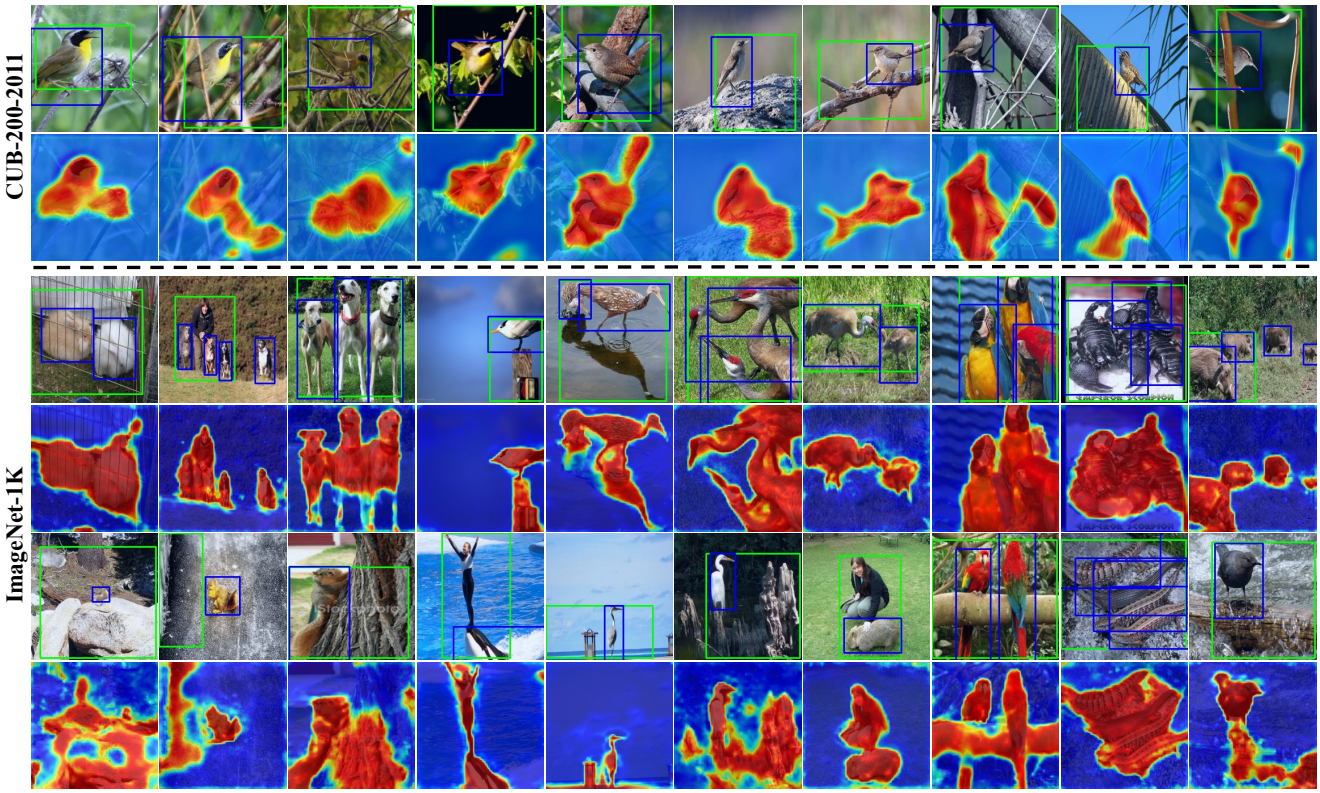

Figure 4: Failure cases. Ground-truth and predicted bounding boxes are highlighted in blue and green, respectively.

References

[1] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2011. 1, 3

[2] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1, 3

[3] Ergys Ristani, Francesco Solera, Roger S. Zou, R. Cucchiara, and Carlo Tomasi. Performance measures and a data set for

multi-target, multi-camera tracking. In *The European Conference on Computer Vision Workshop (ECCVW)*, 2016. 1, 2

[4] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 79–88, 2018. 1, 2

[5] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. 1, 2