TriLite: Efficient Weakly Supervised Object Localization with Universal Visual Features and Tri-Region Disentanglement

Arian Sabaghi, José Oramas
University of Antwerp, sqIRL/IDLab, imec
Antwerp, Belgium
{arian.sabaghikhameneh, jose.oramas}@uantwerpen.be

1. Overview

This document provides additional qualitative results and technical details to support the findings presented in the main submission.

2. TriHead Module Details

The TriHead module transforms ViT patch tokens into spatially feature maps using a single convolutional layer with a kernel size of 3×3, followed by batch normalization (BN) and a spatial softmax activation. To obtain the feature vectors $\mathbf{f}^{fg}$ and $\mathbf{f}^{ang}$, we perform a weighted channel-wise multiplication (as detailed in Eq.$\tilde{(1)}$) between reshaped patch features $\mathbf{F} \in \mathbb{R}^{D \times w \times h}$ and corresponding 2D maps $\mathbf{M}^c \in \mathbb{R}^{w \times h}$:

$$\mathbf{f}^c = \frac{\sum_{i=1}^{wh} \mathbf{M}_i^c \mathbf{F}_i}{\sum_{i=1}^{wh} \mathbf{M}_i^c + \epsilon}, \quad c \in \{fg, bg\}, \tag{1}$$

Subsequently, we pass the feature vectors $\mathbf{f}^{fg}$ and $\mathbf{f}^{bg}$ through a single fully connected layer, matching the number of output classes (e.g., 1000 classes for Imagenet-1k [3], 200 classes for CUB-200-2011 [4]) and 100 classes for OpenImages [1, 2], resulting in logits $\mathbf{z}^{fg}$ and $\mathbf{z}^{bg}$ corresponding to foreground and background image regions, respectively. During training, we apply dropout to $\mathbf{f}^{fg}$ and $\mathbf{f}^{bg}$—with a rate of 0.25 for the Imagent-1k, 0.5 for CUB-200-2011 and no dropout for OpenImages.

Thus, the trainable parameters in the TriHead module comprises only one convolutional layer and one fully connected layer.

3. Role of the Ambiguous Channel

TriLite introduces a third output channel, referred to as the *ambiguous* channel, in addition to the foreground and background channels. This channel can be interpreted as a slack variable that relaxes the rigid partitioning between foreground and background.

The ambiguous channel captures regions that cannot be reliably assigned to either category. These include contextual cues correlated with the target class (e.g., surrounding environment or co-occurring objects), as well as inherently uncertain areas where semantic boundaries are unclear (Fig. 1).

By allocating such regions to a dedicated channel, TriLite avoids enforcing a hard foreground/background split and reduces interference between the two representations, resulting in a more stable feature decomposition. Consequently, the foreground maps become more spatially precise, with reduced leakage into background regions, which improves localization performance.

This effect is further supported by the ablation study in the main paper (Table 3), where removing the ambiguous channel consistently leads to a drop in accuracy.

4. Inference and Training Efficiency

Table 1 provides extended efficiency measurements obtained on a single NVIDIA A100 GPU under a fixed software environment. In addition to the metrics reported in the main paper, we include peak memory usage during training and evaluate additional baselines and model variants.

Inference latency is computed over 1,000 runs after 20 warm-up iterations. Training efficiency is reported in images/sec along with peak memory usage, using a batch size of 128 (GenPromp uses batch size 8 due to memory constraints). These results should be interpreted comparatively across methods. TriLite remains efficient across all settings, combining high training throughput with low inference latency, while also exhibiting favorable memory usage compared to competing approaches.

5. Sensitivity to Adversarial Loss Weight α

The adversarial loss in TriLite explicitly supervises background regions. As a result, its optimal weighting α is inherently dataset-dependent. For example, CUB typically

Table 1. Training and inference efficiency comparison. Two values are reported for GenPromp due to its two-stage pipeline. † Inference at 224×224 resolution. ‡ Inference at 448×448 resolution. Both variants of DINOv2 only differ at inference resolution.

Method	**Params (M)**	**Infer. Time (ms)**	**GFLOPs**	**Images/s**	**Memory (GB)**
CAM (VGG16)	19.6	1.6	32.5	536	29.1
TS-CAM (DeiT-S)	22.3	7.5	4.75	462	9.1
LCTR (DeiT-S)	25.7	6.1	5.0	680	7.5
GenPromp (SD+EffB7)	1066.2	276	1272.7	10 / 7.1	17.0 / 72.3
TeD-Loc (ViT-EVA-L)	569.6	28.3	–	–	–
TriLite (DINO)	21.8	5.1	4.61	1380	0.9
TriLite (DeiT-S)	22.2	5.1	4.61	1408	0.7
TriLite (DINOv2)†	22.2	6.25	6.13	1180	0.8
TriLite (DINOv2)‡	22.2	11.34	31.7	1180	0.8

exhibits relatively homogeneous backgrounds, whereas ImageNet contains significantly more clutter and occlusions.

Despite this dependency, TriLite demonstrates strong robustness to the choice of α. As shown in Table 2, performance remains stable across a wide range of values, and the model remains competitive even when the adversarial loss is removed ($\alpha = 0$). On CUB, performance peaks at $\alpha = 100$, while remaining stable for $\alpha \in [0, 100]$ before degrading at larger values. On ImageNet, performance is highly stable across all tested values, with only minor variation.

Table 2. Effect of adversarial loss weight α on GT-Loc accuracy (%) across datasets.

α	0	50	100	200	300
CUB	97.0	97.9	98.5	91.3	72.2
ImageNet	77.5	77.6	77.6	77.9	77.7

6. Additional Qualitative Results

Figure 2, 3 and 4 shows more qualitative comparisons between a CAM-based baseline [5], existing state-of-the-art method and our proposed method on the CUB-200-2011 Imagenet-1k and OpenImages datasets.

References

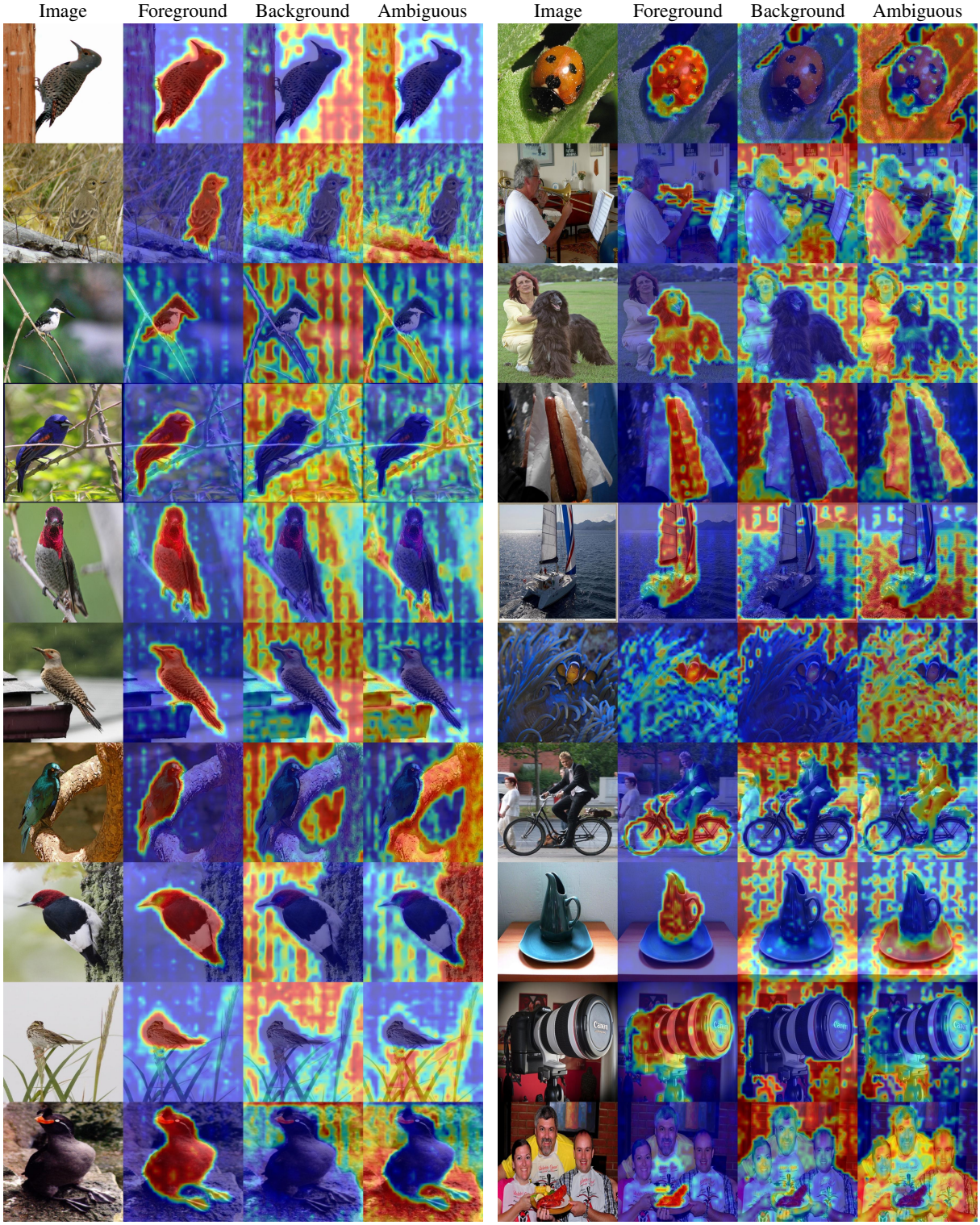
[1] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11700–11709, 2019. 1

[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] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 1

[4] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1

[5] 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. 2



(a) CUB-200-2011 (b) ILSVRC

Figure 1. Visualization of all output channels of TriLite. The ambiguous channel activates on salient regions that are correlated with the target class but do not belong to the object itself.

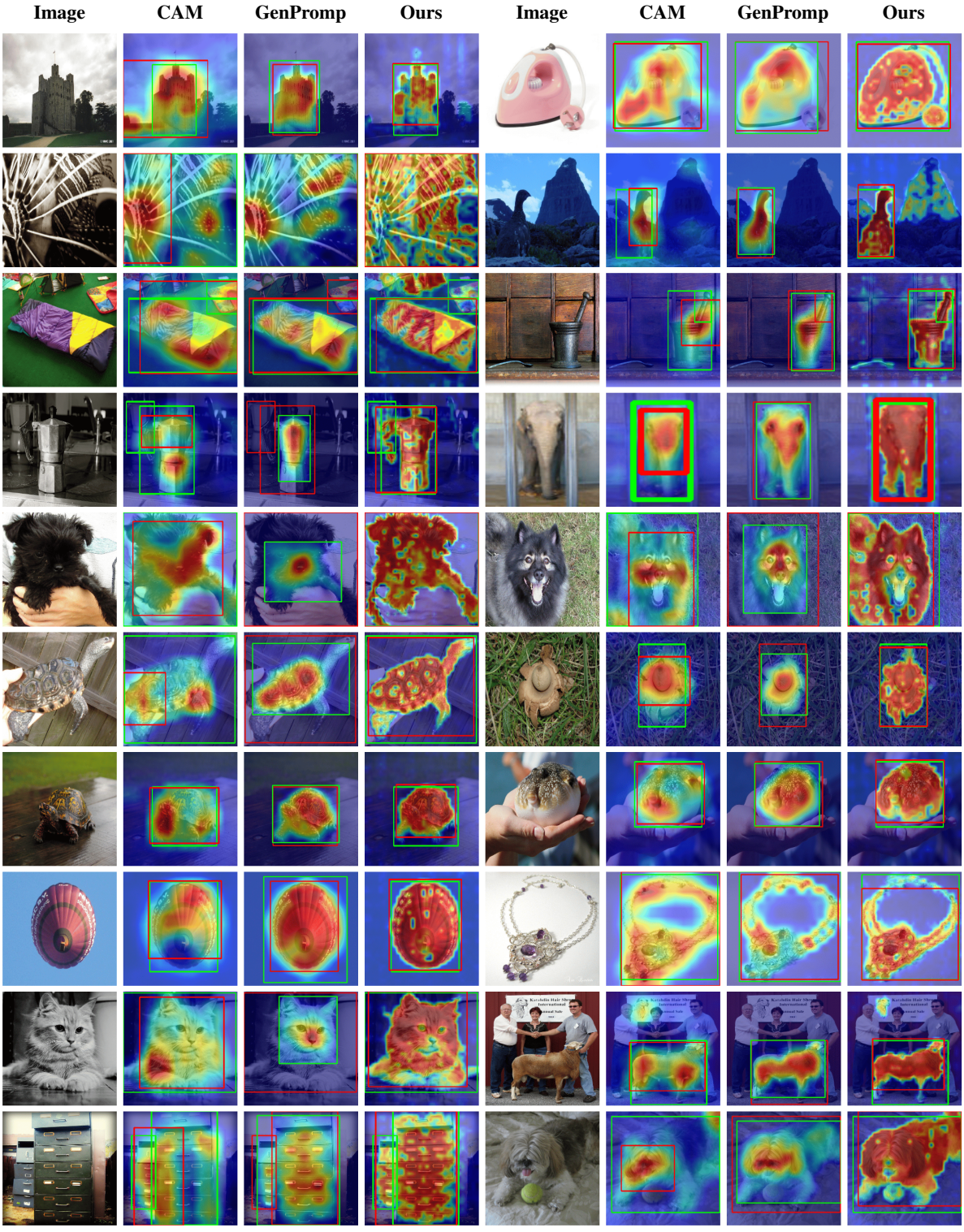


Figure 2. Qualitative comparison on the **Imagenet-1k** dataset. Each row shows two image sets. For every set we display the original image, a CAM heatmap, the GenPromp heatmap, and our TriLite heatmap.

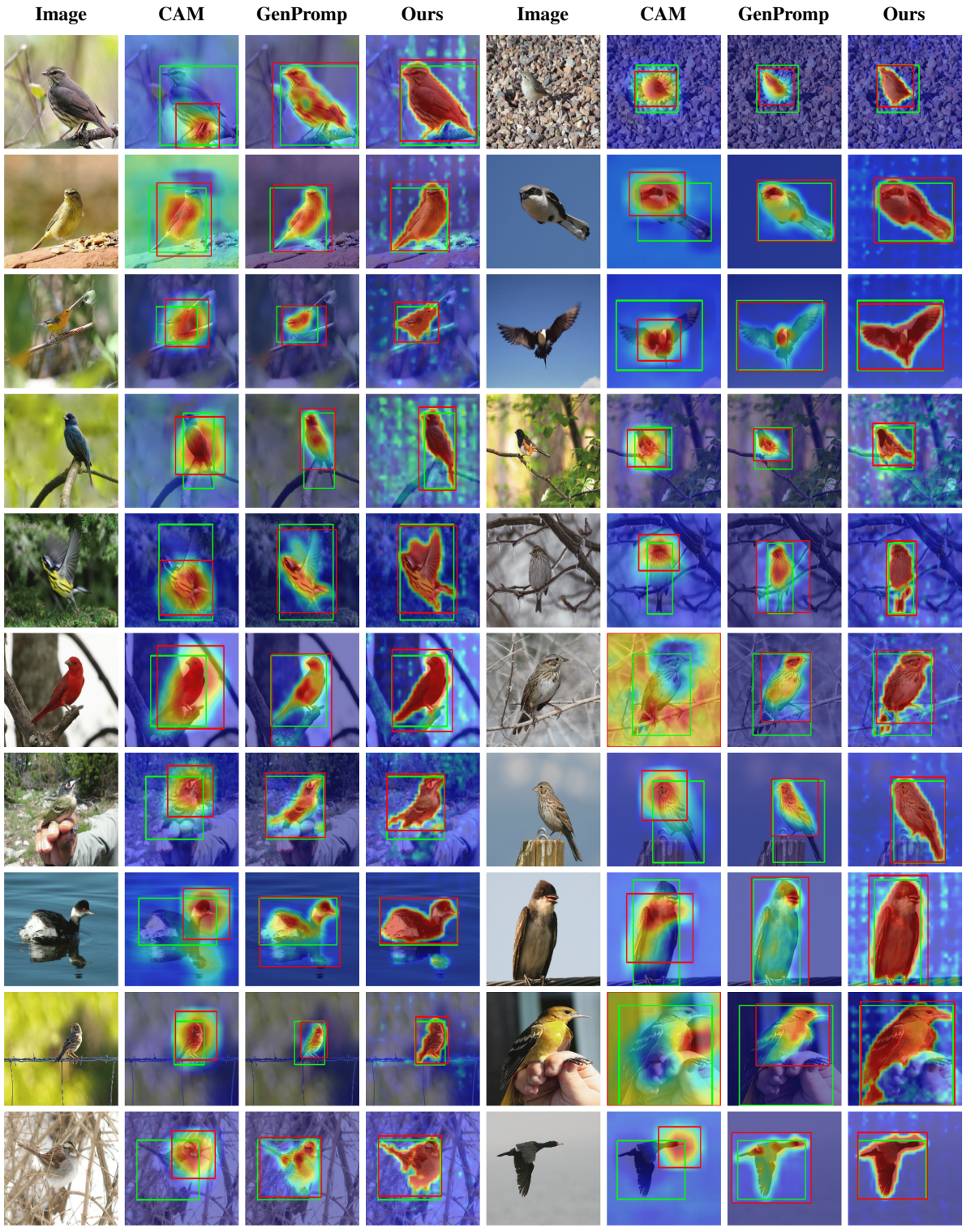


Figure 3. Qualitative comparison on the **CUB** dataset. Each row shows two image sets. For every set we display the original image, a CAM heatmap, the GenPromp heatmap, and our TriLite heatmap.

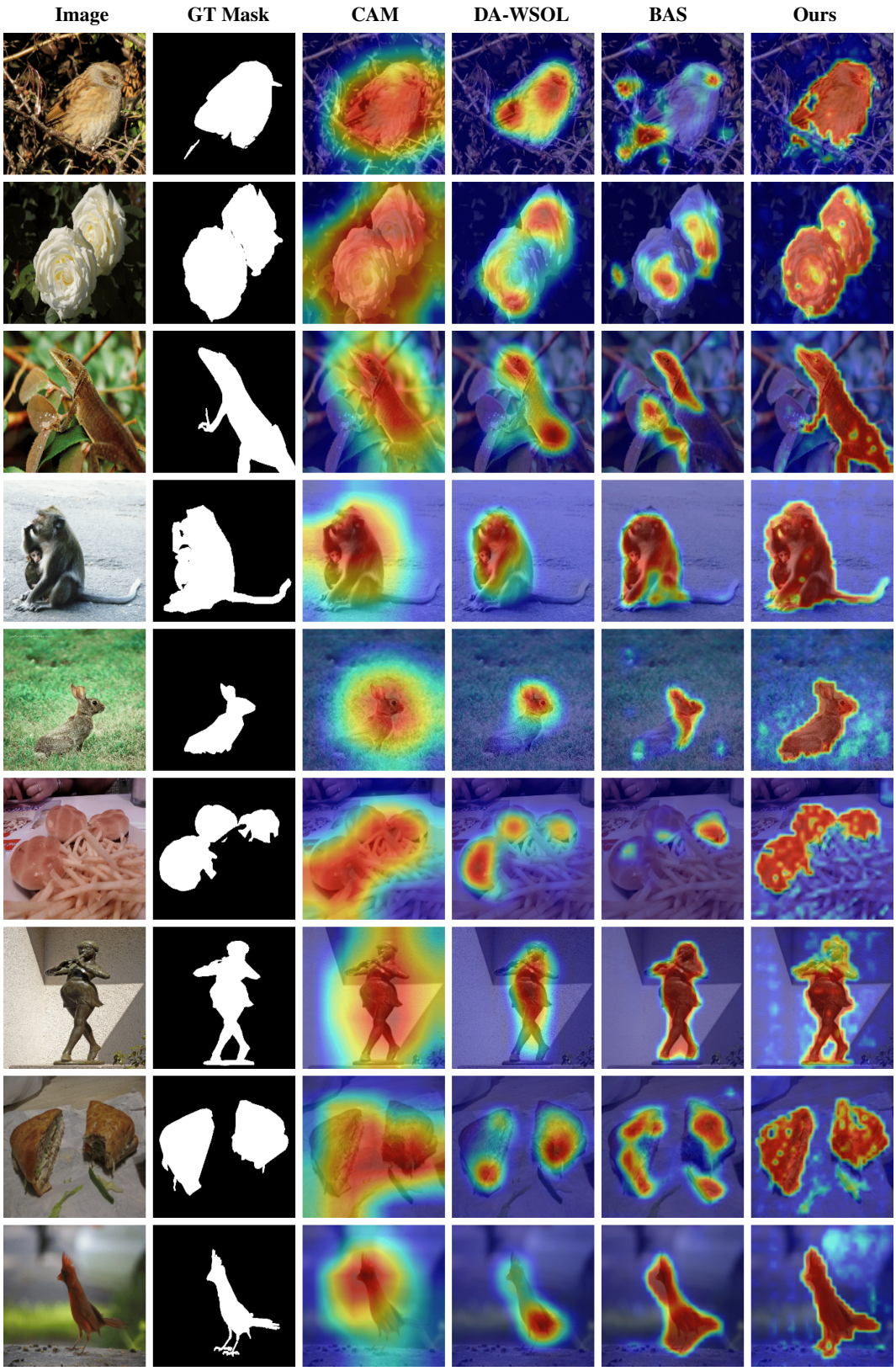


Figure 4. Qualitative comparison on the **OpenImages** dataset.