Supplementary Material

Krishna Kanth Nakka
krishna.nakka@epfl.ch
Mathieu Salzmann
mathieu.salzmann@epfl.ch

Computer Vision Lab
EPFL, Switzerland

In this document, we provide additional experiments and qualitative results to further support those in the main paper.

1 Influence of λ on Classification Accuracy

We first evaluate the influence of the hyper-parameter λ in Eq. 7 of the main paper, which defines the strength of the attention module loss, on the classification accuracy. To this end, we evaluate our approach for different values of λ on the CUB-200 bird dataset, after 20 training epochs and without performing any data augmentation (image flipping) at inference time. The results of this experiment are provided in Table 1. Note that accuracy is stable over a very large range of values, thus showing that our approach is robust to the choice of this parameter.

λ	0.0001	0.01	0.4	1
Accuracy	83.3	83.7	83.7	83.1

Table 1: Influence of λ on the final classification accuracy.

2 Attentional Global Average Pooling

While our main goal was to introduce an attention mechanism in structured representations, our approach also applies to unstructured pooling strategies, such as global average pooling (GAP). To illustrate this, we implemented an attentional GAP layer using our attention map. As shown in Table 2, this also typically outperforms the standard GAP strategy, thus further showing the benefits of attention when performing feature aggregation. Note, however, that our attentional VLAD strategy still significantly outperforms the GAP one.

Pooling	Anno.	Birds	Cars	Aircrafts
GAP	BBox	79.8	89.3	86.6
Attentional-GAP	BBox	76.3	91.1	88.3
NetVLAD	BBox	82.4	89.8	88.0
Attentional-NetVLAD	BBox	85.5	93.5	89.2
GAP		78.6	86.2	84.5
Attentional-GAP		77.8	89.6	85.5
NetVLAD		80.6	89.4	86.4
Attentional-NetVLAD		84.3	92.8	88.8

Table 2: Attentional Global Average Pooling on fine-grained datasets.

3 Additional Qualitative Results

Below, we show the attention maps obtained with our approach for additional randomly-sampled images from the four datasets used in the main paper.

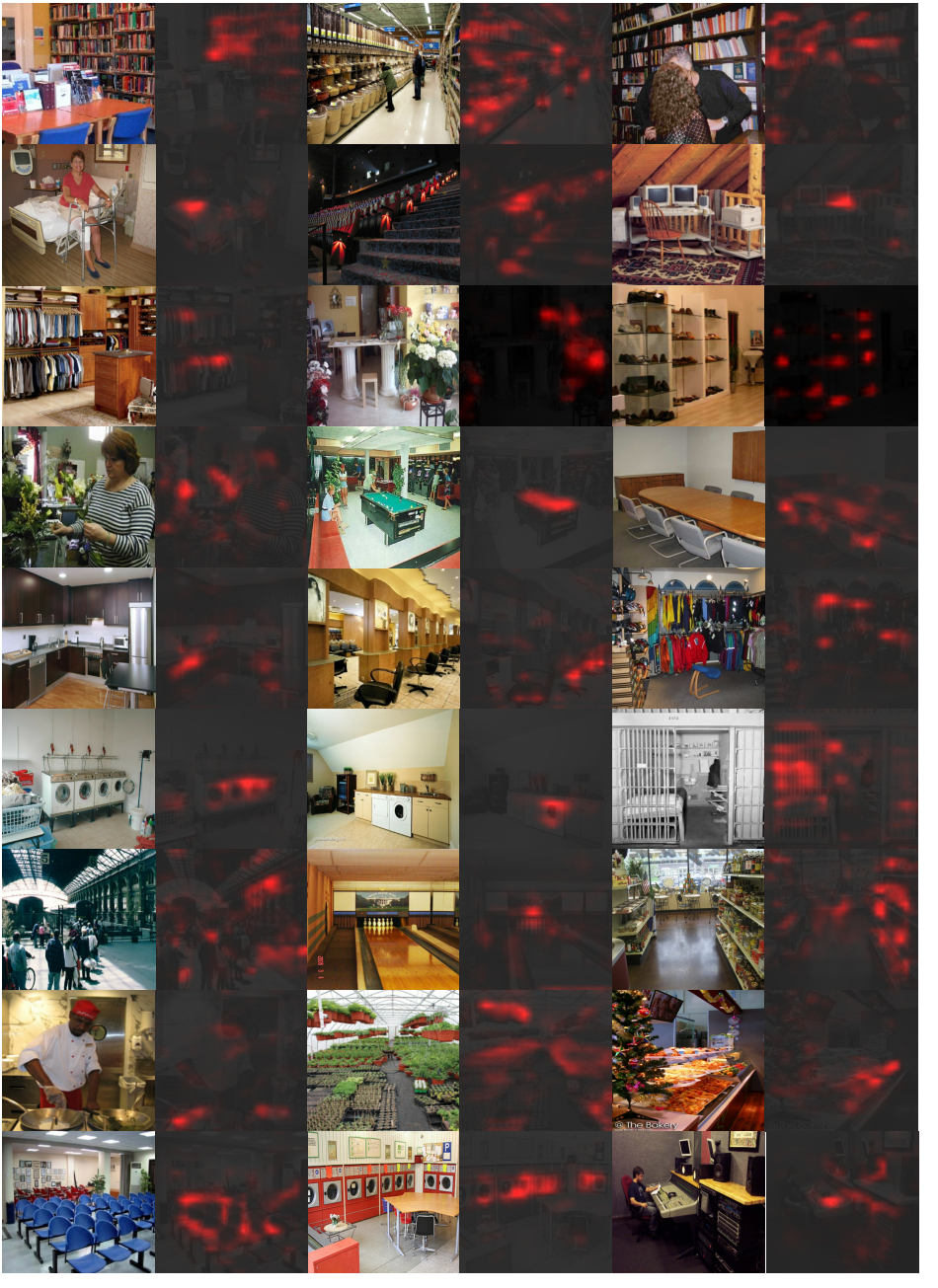

Figure 1: Generated attention maps on the MIT-Indoor dataset.

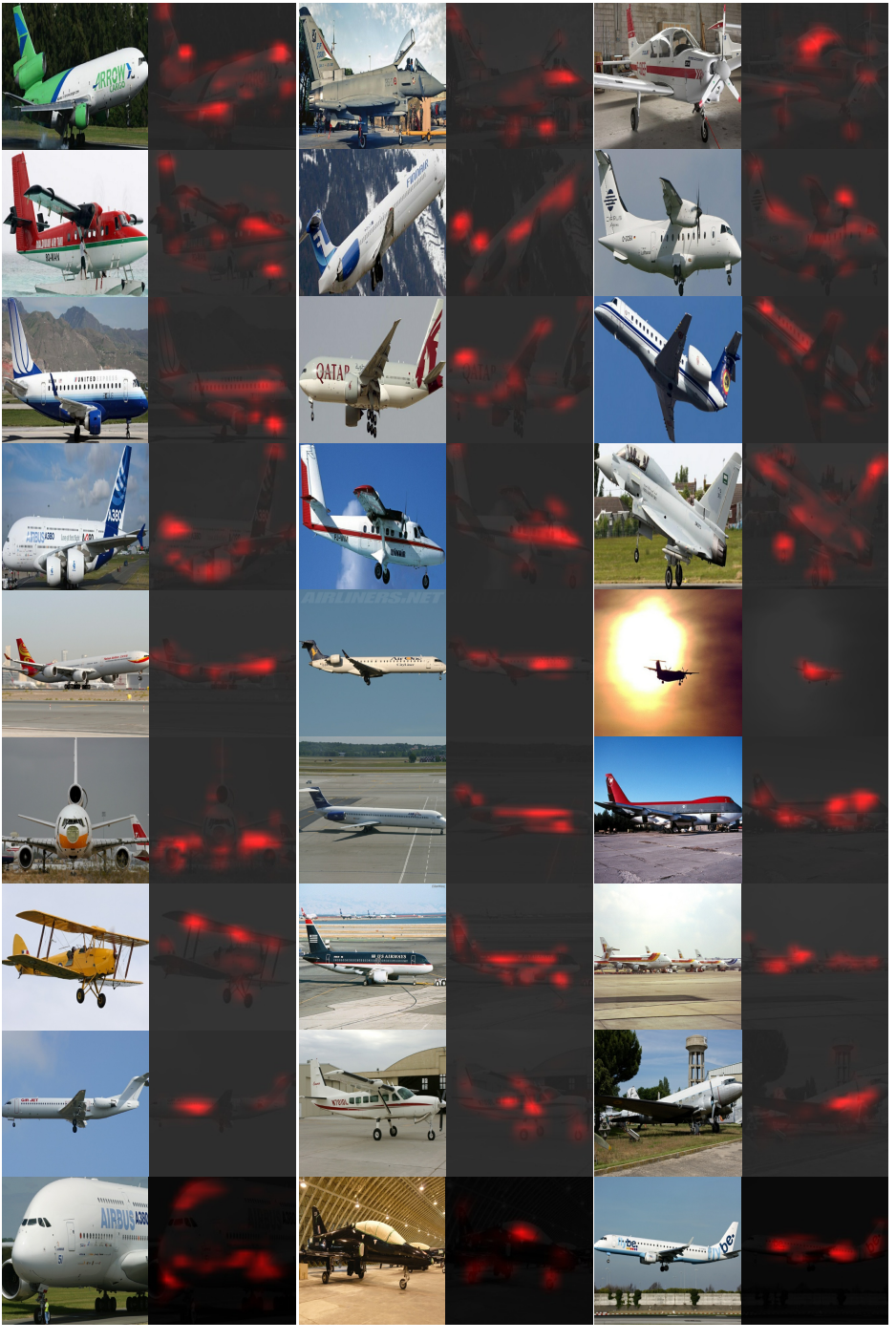

Figure 2: Generated attention maps on the Aircrafts dataset.

Figure 3: Generated attention maps on the Birds dataset.

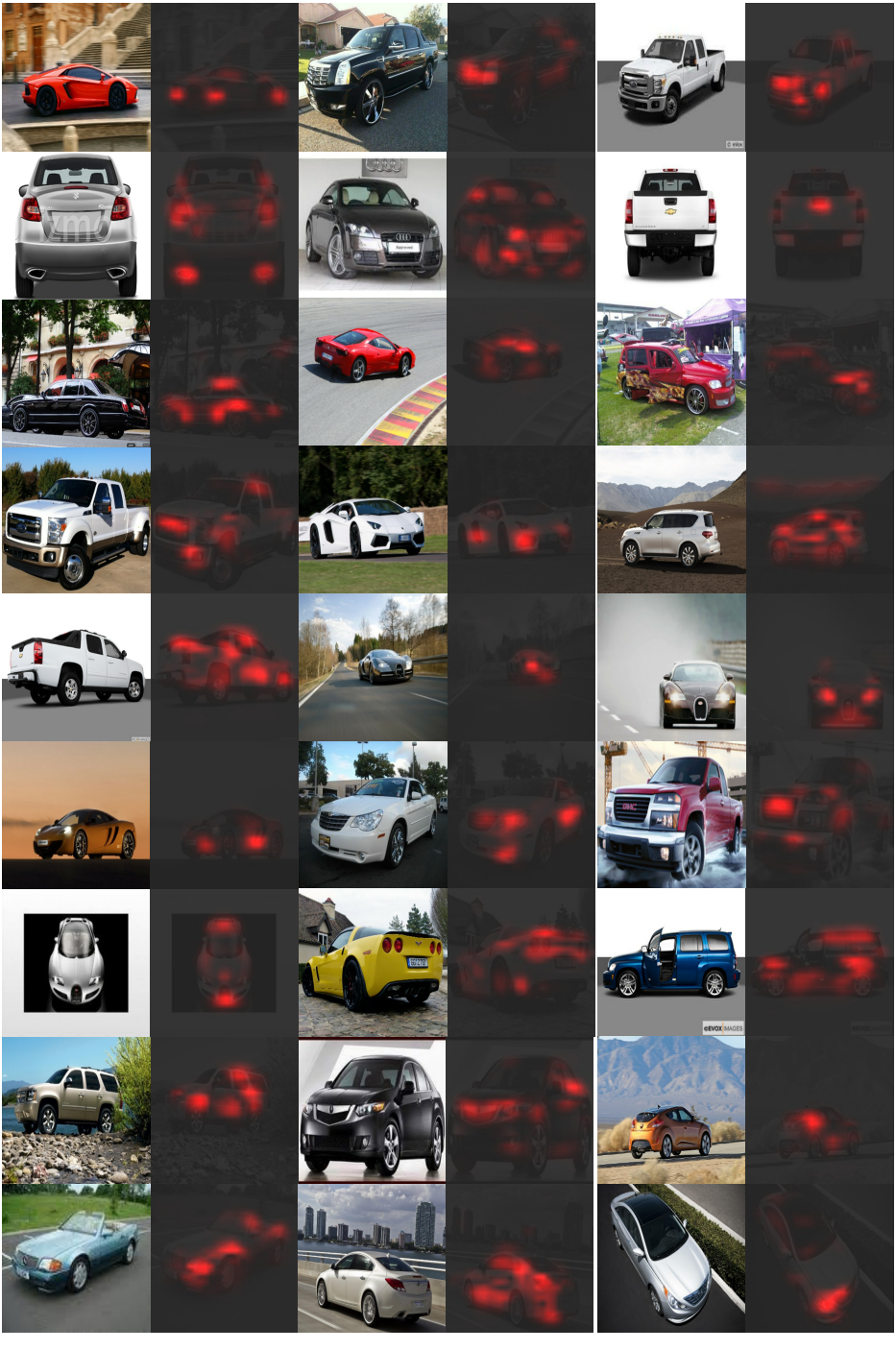

Figure 4: Generated attention maps on the Stanford-Cars dataset.

4 Failure Cases

Finally, in Fig. 5, we show some typical failure cases of our approach, such as attention to background regions on Birds dataset.

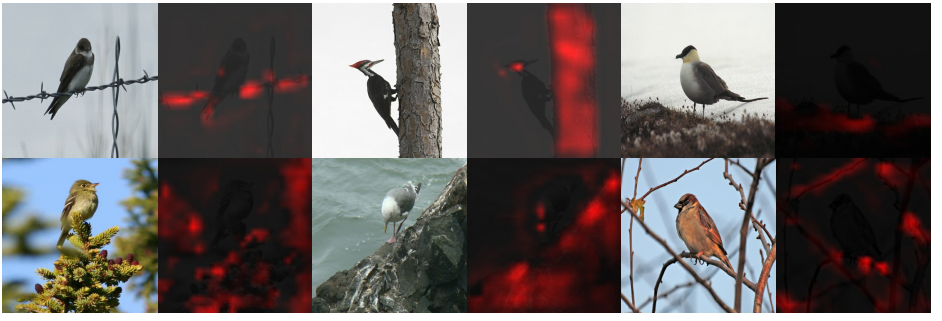

Figure 5: Failure cases of our model particularly due to incorrect attention.