FAM: Visual Explanations for the Feature Representations from Deep Convolutional Networks (Appendix)

Yuxi Wu, Changhuai Chen, Jun Che, Shiliang Pu*
Hikvision Research Institute, China
{wuyuxi, chenchanghuai, chejun, pushiliang.hri}@hikvision.com

A. Explanation Map

Our proposed generation method of explanation maps is adaptive in different situations. To match the visual sense, we divide the activation values into 3 grades based on the corresponding colors in visualization results. The sub-regions with high activation values (Red) would be preserved completely, and the areas with low activation (blue) would be masked. Meanwhile, to distinguish the sub-regions with medium activation (green) from high activation, the information of these pixels would be partly retained. As for the retained proportion, we found 90% is low enough to show the difference. As showed in Figure 1, masking the information with 10% would already cause significant influence on the retrieval results for Person Re-ID.

In this way, the metrics based on explanation maps can better discern the difference among different visualization techniques. For instance, the visualization result of Grad-FAM in Figure 2 (c) indicated that the person body got medium activation and was less important than the shoes for the Re-ID model. Meanwhile, Score-FAM generated a similar visualization result, but the person body also got high activation. The explanation maps of two visualization techniques are visually similar, but there is a huge gap between the retrieval results. And the results demonstrate that Score-FAM provide a more reliable visual explanation for this image.

B. Comparison between Grad-FAM and Score-FAM

In this work, we propose two visualization techniques, Grad-FAM and Score-FAM. The experimental results show that Score-FAM outperforms Grad-FAM on three metrics by large scale in both datasets. And the examples in Figure 2 show the difference of two techniques more clearly. The sub-regions that Grad-FAM focused on also got high activation for Score-FAM in all examples. However, Score-FAM concerned more areas that were less important for Grad-FAM, and the difference caused a huge gap between the retrieval results of explanation maps. The results demonstrate that these sub-regions are also very important for the feature representations, and thus the visual explanations generated by Score-FAM are more precise and reliable than Grad-FAM.

However, the main drawback of Score-FAM is the heavy computational cost. Score-FAM costs much more time than Grad-FAM. The computational cost of Score-FAM is acceptable for few samples, but it is unsuitable to employ Score-FAM to process large amounts of images. In the latter case, due to the high efficiency, Grad-FAM is the first choice. And few interested samples can be processed afterwards by Score-FAM to get more precise visual explanations. Therefore, these two FAM-based methods have their own advantages, and are applicable to different situations.

C. Applicationn for Domain Gap

The applications of Person Re-ID are restricted by a fact that the trained model might fail in practical scenarios. The performance drops are usually caused by domain gap, including different lighting conditions, backgrounds, seasons, etc. However, it is hard to confirm the exact cause of domain gap.

Since Score-FAM interprets the attention of models, the visualization results can be employed to analyze the cause of bad performance. In this section, we would use Score-FAM to find out the main factors of domain gap.

As an example, the model trained on Market1501 would fare poorly on CUHK03, which means there is a huge gap between these two datasets. For this model, we randomly select 200 images from CUHK03 to generate visual explanation by ScoreFAM.

By analyzing the visualization results, we found out three main factors of the domain gap between CUHK03 and Market1501. As showed in Figure 3, the spotted floor, striped floor and tiled wall in background got high atten-

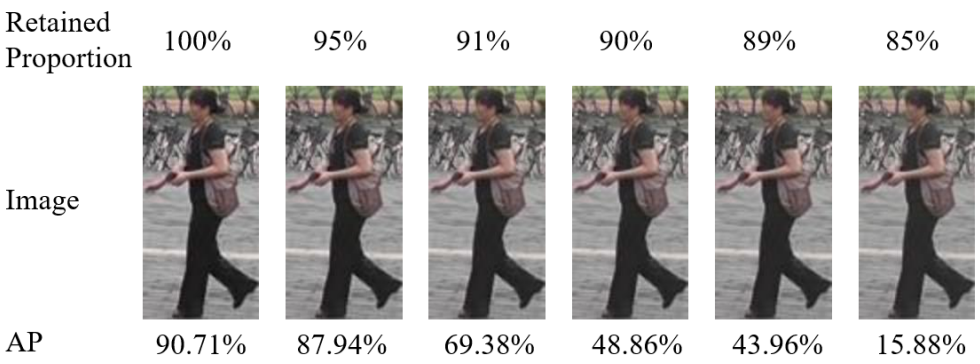


Figure 1. For a query sample of Market1501, a series of images are generated by specified retained proportions. For every pixel of the image, the information is masked with the specific ratio. Average Precision (AP) is a popular metric for evaluating retrieval results in test, and higher is better. Even though the images are almost same visually, the generated images with little mask would get much worse retrieval results for Person Re-ID.

tion for the model trained on Market1501. The results indicate that this model would be unreliable under these backgrounds. In this way, the reliability of Re-ID models on new domains can be preliminarily judged without labeling.

Under the guidance of the known background interference, the model can be also improved pertinently. For instance, the black edge of wall got high activation in all samples of Figure 3 (c), and thus different persons would get high similarity in this position. The main reason is that this type of background did not exist in the training process of this model. Thus it cannot distinguish between the tiled wall and person body. To solve this problem directly, the tiled wall need to appear in the train set. Cut-and-paste can achieve it based on few target images. This method help to improve the model more efficiently in practical applications.

D. Application on Self-Supervised Learning

Figure 4 shows more examples for application of FAM on self-supervised representations learning. Since the model is randomly initialized, the visualization results begin from 10 epochs. At this point, the model ignores the foreground object in most cases. After the training of 30 epochs, part of foreground regions get high activations in saliency maps. In the case of simple background, such as the third line of Figure 4, the salient regions have been localized to the foreground. In the subsequent training process, the salient regions are gradually localized to a small region, but part of backgrounds still get medium activation. By analyzing the change of visualization results, the observers can get a better understanding of the employed pretext task.

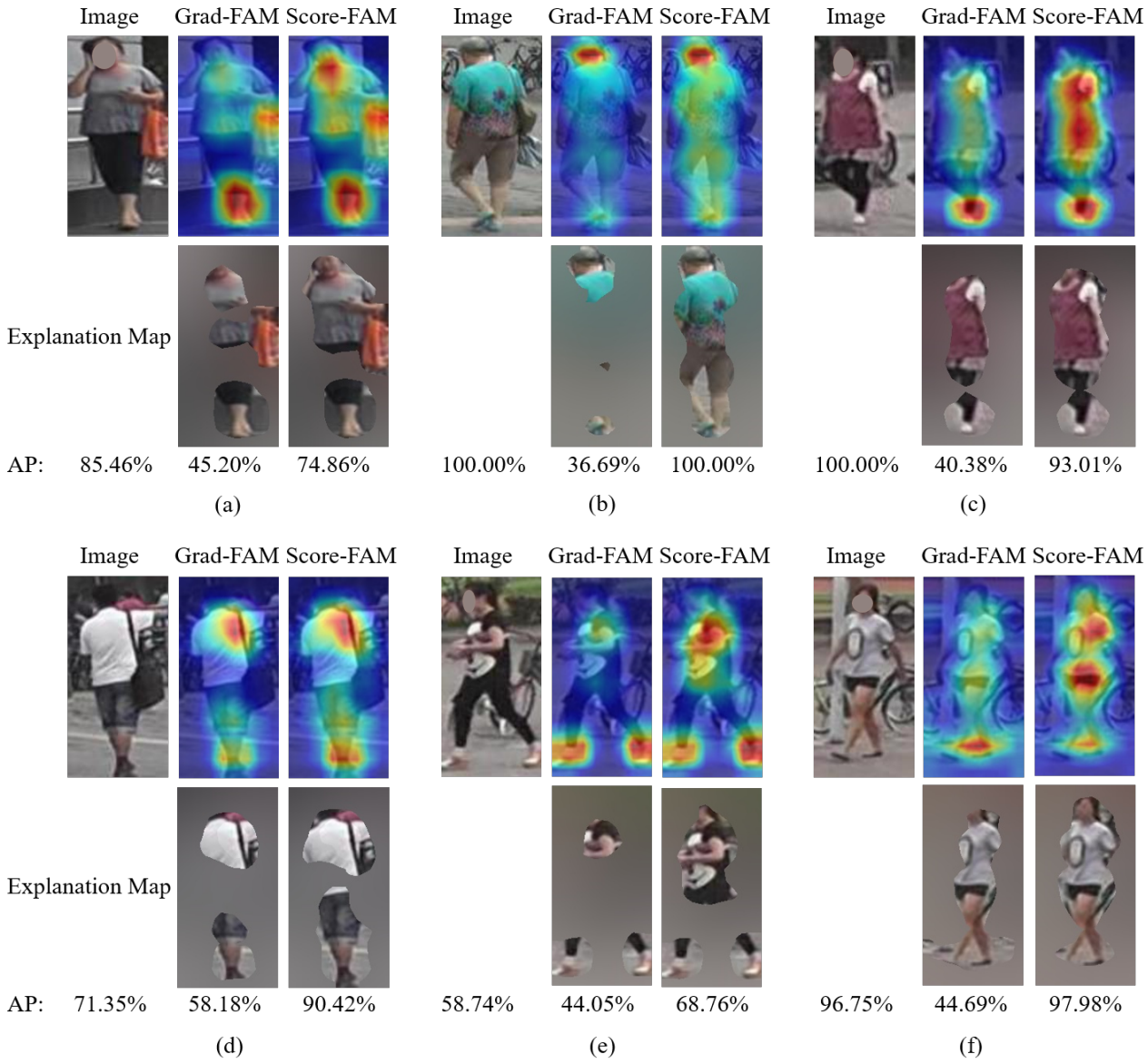
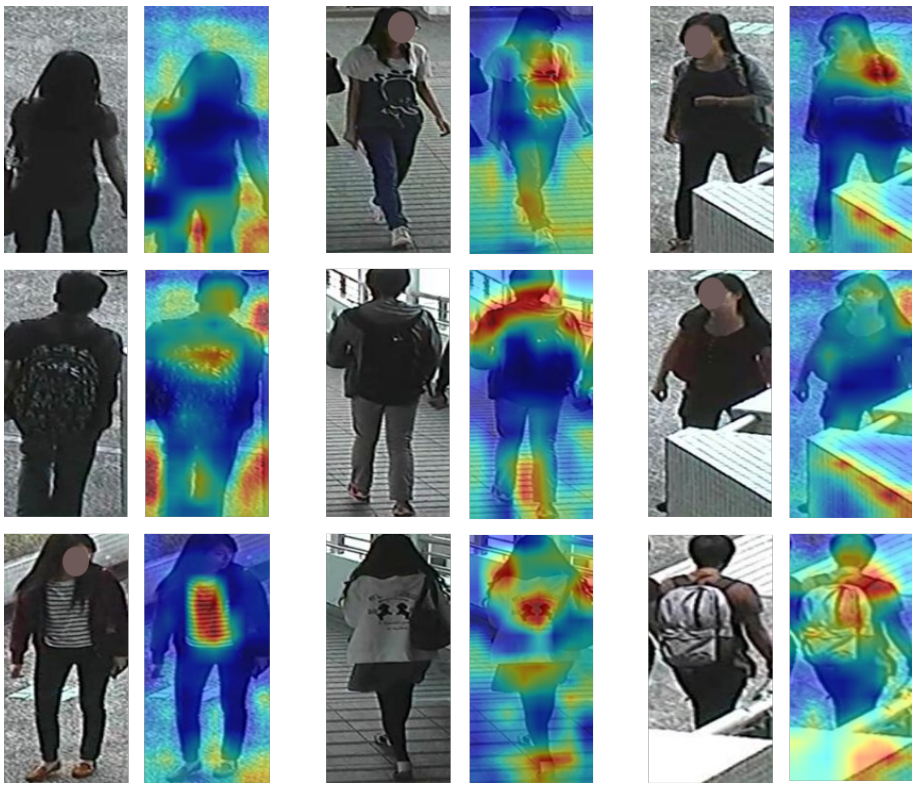


Figure 2. Examples for Grad-FAM and Score-FAM, and the AP of corresponding explanation maps. Although Grad-FAM got similar visualization results with Score-FAM, the retrieving results of corresponding explanation maps were much worse. In some examples with background interference, the explanation maps of Score-FAM got higher AP than original inputs.



(a) Spotted floor. (b) Striped floor. (c) Tiled wall.

Figure 3. Examples for the poor performance of model trained on Market1501. The images are from the query set of CUHK03. By observing the visualization results, we analyze 3 types of significant background interference.

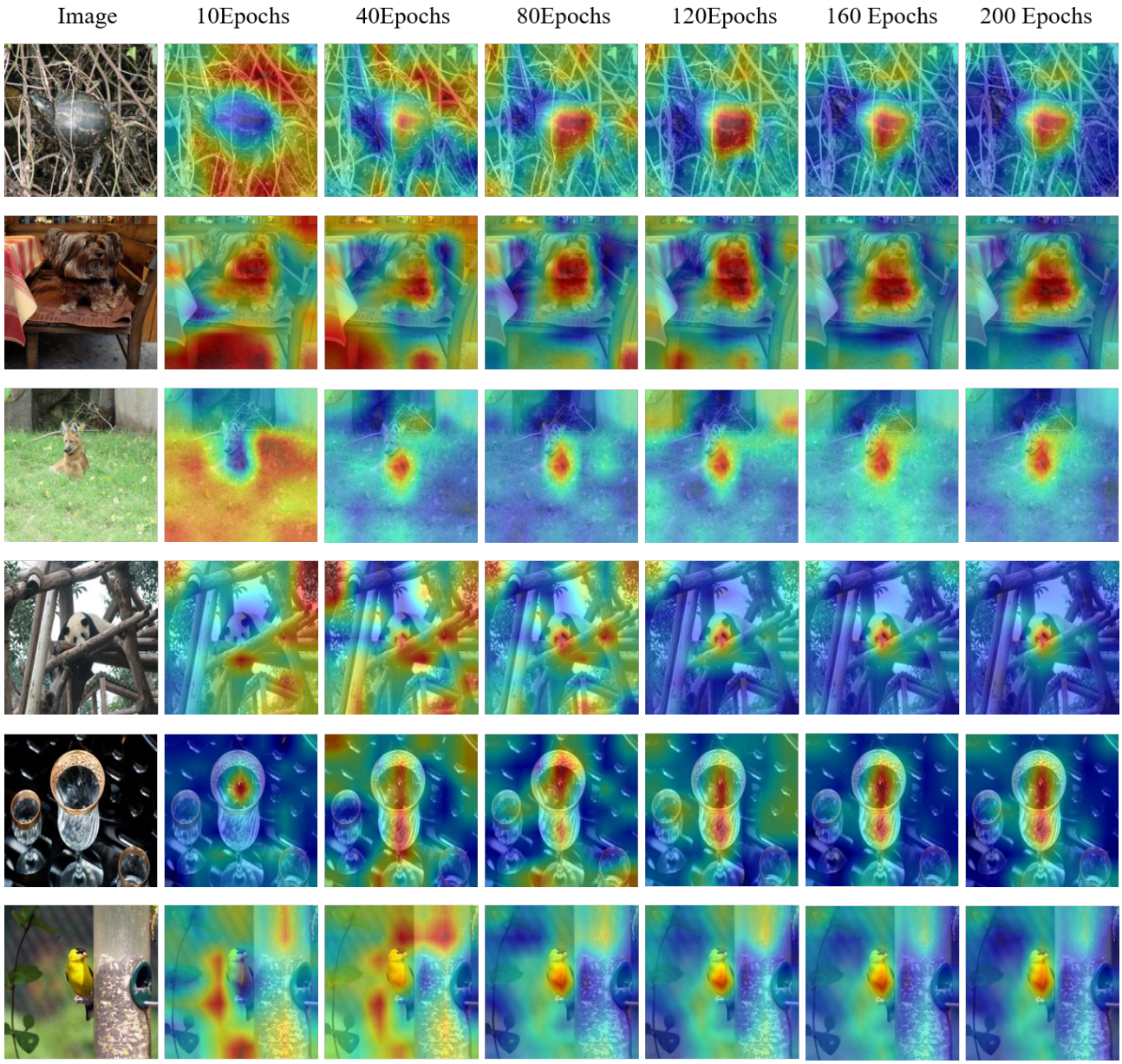


Figure 4. Visualizing the training process of the HRNet-W30 model trained by BYOL. Given 6 images for the models at 10, 40, 80, 120, 160, 200 epochs in self-supervised training, we visualize the change of salient regions generated by Grad-FAM. In the subsequent training process, the salient regions are gradually localized to the foreground objects.