A Protocol for Evaluating Model Interpretation Methods from Visual Explanations Supplementary Material

Hamed Behzadi-Khormouji José Oramas
University of Antwerp, imec-IDLab

Implementation Details

Fine-tuning. Following [2], in order to fine-tune on the CUB-70 dataset a VGG19 model pre-trained on ImageNet dataset, we superseded the last Max-Pooling layer to a new one with stride 4 and kernel size 4. Additionally, we changed the probability parameter of the Drop-Out layer in the classifier part of VGG19 to 0.2. We applied a similar change for the feature extraction part of the VGG19 backbone of the ProtoPNet architecture. In both cases, we replaced the last Fully-Connected layer of the models with a new one, initialized randomly, with the output size equal to the number of classes.

Training Optimizer. For the binary CelebAMask-HQ classification task, we utilized the Adam optimizer with the learning rate $1e-4$. To adjust the learning rate during training phase, we decay the learning rate with a factor equal to 0.95 after each epoch. This setting was applied for the training of the three CNN architectures as well as the Topic-based interpretation method.

For the classification on the CUB-70 dataset, we used the SGD optimizer with learning rate $1e-3$, momentum 0.9, and weight decay $1e-4$. Besides, we decay the learning rate with factor 0.1 every 30 epochs. These setting were used for training the three base CNN architectures. Furthermore, Adam optimizer with learning rate $1e-3$ and weight decay 0.95 was used for training phase of the Topic-based interpretation method. Following [1], we adjusted the optimizer settings for training the ProtoPNet architecture on both datasets. We selected and adjusted these optimizers through different experiments in order to reach high classification performance for each model.

Topic-based interpretation. The hyper-parameter in Topic-based interpretation method is the number of topic vectors. [4] evaluated their method on a synthetic dataset and the AwA dataset [3]. However, for determining the number of topic vectors for CelebAMask-HQ and CUB-70 dataset we examined different values on which the method can classify images with high accuracy and close to those of ProtoPNet as well as base CNNs VGG19, Densenet121, and Resnet50. Therefore, for the CelebAMask-HQ classification task we selected 20 as the number of topic vectors for the three CNNs. In a similar manner, the hyper-parameter was set to 300 in CUB-70 classification task. Since the CUB-70 dataset has fine-grained categories, then we needed to increase the complexity of the Topic-based interpretation method to be trained with high accuracy close to those of ProtoPNet as well as base CNNs VGG19, Densenet121, and Resnet50.

ProtoPNet. The shape of the prototype parameter from the different ProtoPNet-based methods was defined following several tests. For models trained on ClebAMask-HQ, they were set to $(20, 128, 1, 1)$, $(24, 128, 1, 1)$, and $(30, 1024, 1, 1)$ for the considered VGG19, Densenet121, and Resnet50, respectively.

For the three CNN models trained on the CUB-70 dataset, this parameter was set to $(350, 128, 1, 1)$. In these settings, the first element (i.e., 20, 24, 30, and 350) indicates the number of prototypes in the prototype layer and the rest of the elements (i.e., $[128, 1, 1]$ and $[1024, 1, 1]$) indicate the shape of prototype vectors.

Visual Explanations with High / Low Coverage

Figures 1-4 illustrates visual explanations with the highest and lowest Intersection-over-Union (IoU) scores. More specifically, Figures 1 and 2 show the visual explanations with the highest and lowest IoU scores for each of the interpretation methods over CNNs trained on CUB-70 dataset. Similarly, Figures 3 and 4 illustrate the results on CelebA dataset.

As can be seen in Figures 1 (top) and 2 (top), the visual explanations with the highest coverage rates highlight the entire or different parts of birds. In contrast, the visual explanations with the lowest coverage rate (Figures 1 (bottom) and 2 (bottom)) highlight the foliage of the trees and plants in the background.

The similar trend can be seen in the CelebA dataset. The visual explanations with the highest coverage rates showed in Figures 3 (top) and 4 (top) highlight small parts such as

nose, *lips*, and bigger parts such *hair* and *skin*. Figures 3 (bottom) and 4 (bottom), in contrast, show that visual explanations with the lowest coverage rate highlight the background.

Illustrations with a Higher Resolution

We illustrate some figures related to the visual explanations and the quantitative evaluation result, presented in the main paper, in a larger size to be able to see the details in them.

Figures 5, 6, and 7 show the visual explanations of the investigated interpretation methods over the CNNs VGG19, Densnet121, and Resnet50 trained on the CUB-70 dataset, respectively. Figures 8, 9, and 10 illustrate the visual explanations results over the CNNs VGG19, Densnet121, and Resnet50 trained on the CelebA dataset, respectively.

Figure 11 illustrates the part-level coverage rate comparison, presented in the main paper with a higher resolution. As can be seen, VEBI has the higher coverage rate in the higher number of semantic parts in models trained on CUB-70 and CelebA datasets, such as Densenet121-CUB70, ResNet50-CUB70, Resnet50-CelebA, and VGG19-CelebA.

More visual explanation results on CelebA dataset

In this section, we illustrate more visual explanation results of VEBI, ProtoPNet, and Topic-based interpretation methods on CelebA dataset. Figures 12, 13, and 14, show the results on CNNs VGG19, Densenet121, and Resnet50, respectively.

Visualization results of Original VEBI

According to the section 4 (*Compared methods*), original VEBI considers the activation maps computed by all the convolutional layers. Hence, in this section, we illustrate the visual explanation results of the original VEBI generated by the proposed evaluation protocol on the considered CNNs and datasets. Figures 15, 16, and 17 show the visual explanation examples on VGG19, Densenet121, and Resnet50 trained on CelebA dataset, respectively. Also, Figs. 18, 19, and 20 show 36 examples of visual explanations from 18 classes of CUB70 on CNNs VGG19, Densenet121, and Resnet50, respectively.

The captions in Figs. 15-20 show the pairs convolutional layer and filter identified by the original VEBI. For example, $L6\_F10$, $L11\_F261$, $L15\_F334$ indicates that the proposed evaluation protocol generates the visual explanation (section 3.1. *Interpretation with Explanation Capability*) through combining the interpretation heatmaps from pairs convolution layers and filters (6,10), (11,261), and (15,334) identified by original VEBI.

References

[1] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[2] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4476–4484, 2017.

[3] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009.

[4] Chih-Kuan Yeh, Been Kim, Sercan Ömer Arik, Chun-Liang Li, Pradeep Ravikumar, and Tomas Pfister. On concept-based explanations in deep neural networks. *CoRR*, abs/1910.07969, 2020.

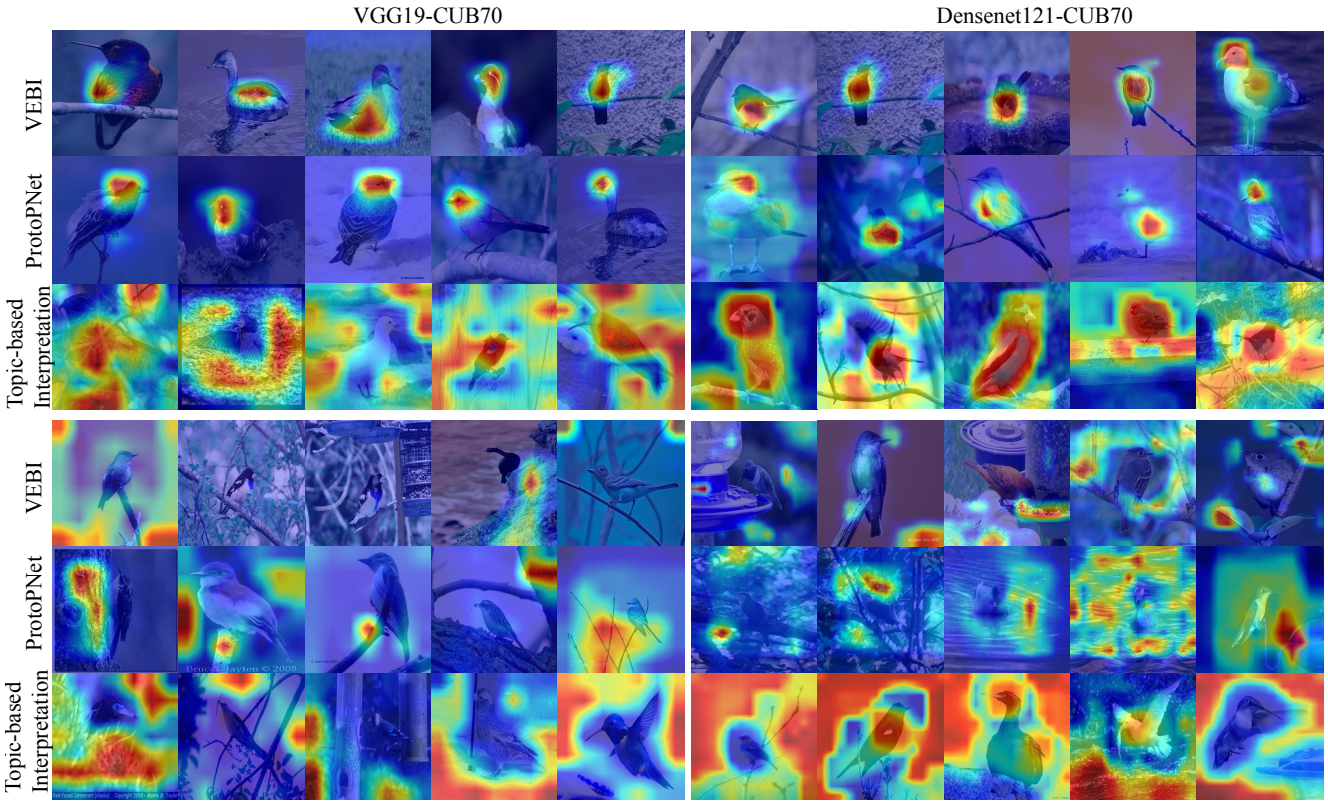

Figure 1. Visual explanations with the highest (top) and lowest (bottom) IoU coverage for each of VEBI, ProtoPNet, and Topic-based interpretation methods over CNNs VGG19 and Densenet121 trained on CUB-70 dataset.

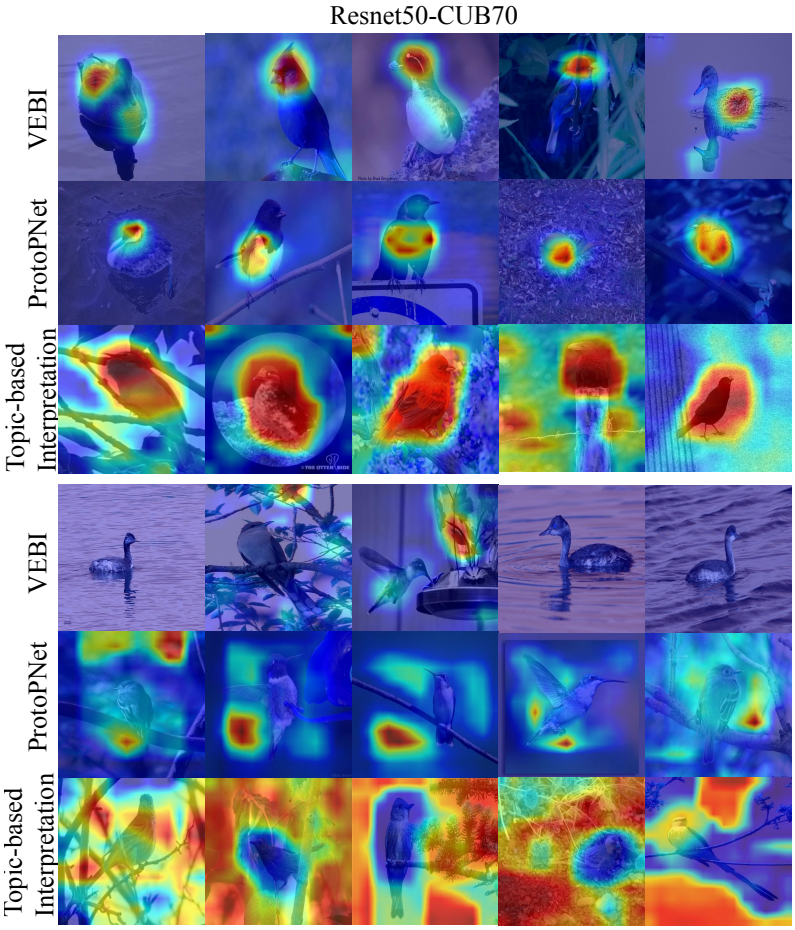

Figure 2. Visual explanations with the highest (top) and lowest (bottom) IoU coverage for each of VEBI, ProtoPNet, and Topic-based interpretation methods on Resnet50 trained on CUB-70 dataset.

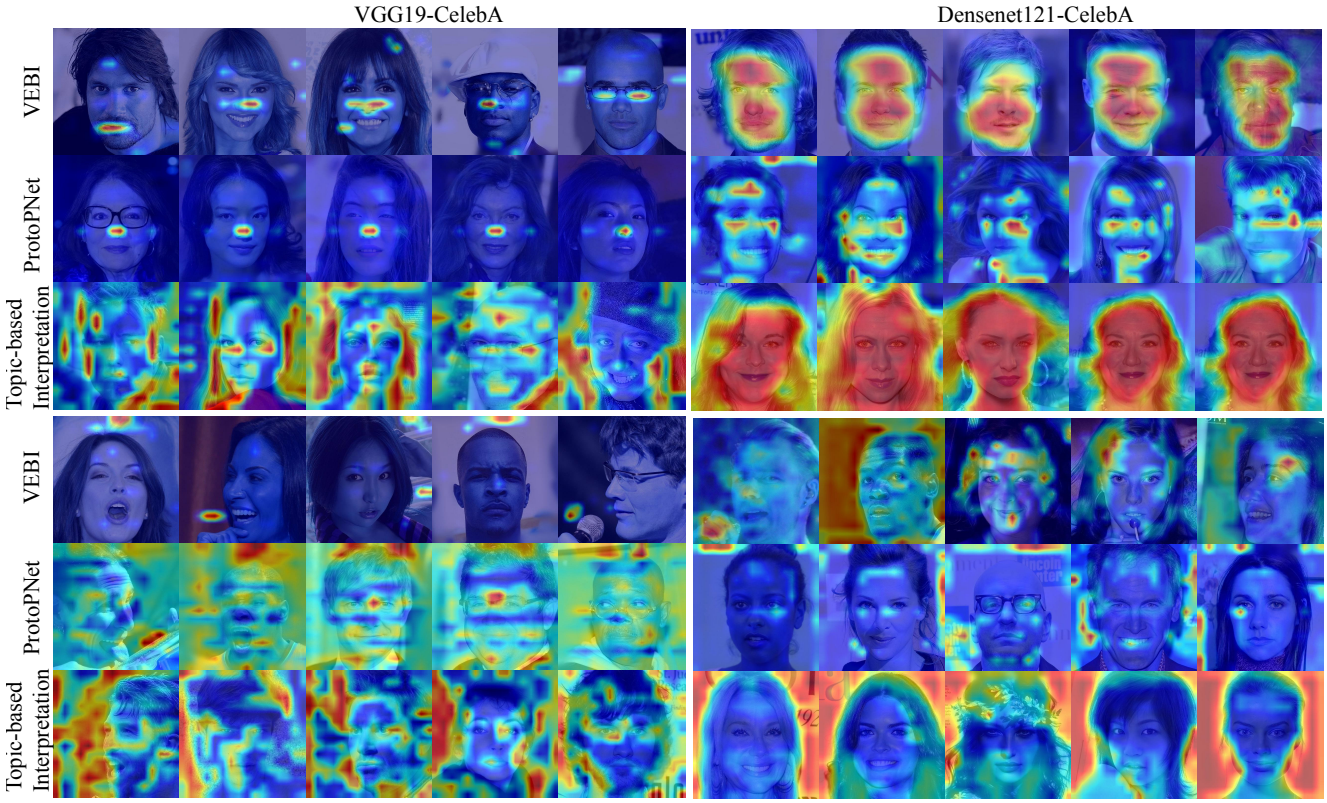

Figure 3. Visual explanations with the highest (top) and lowest (bottom) IoU coverage for each of VEBI, ProtoPNet, and Topic-based interpretation methods over CNNs VGG19 and Densenet121 trained on CelebA dataset.

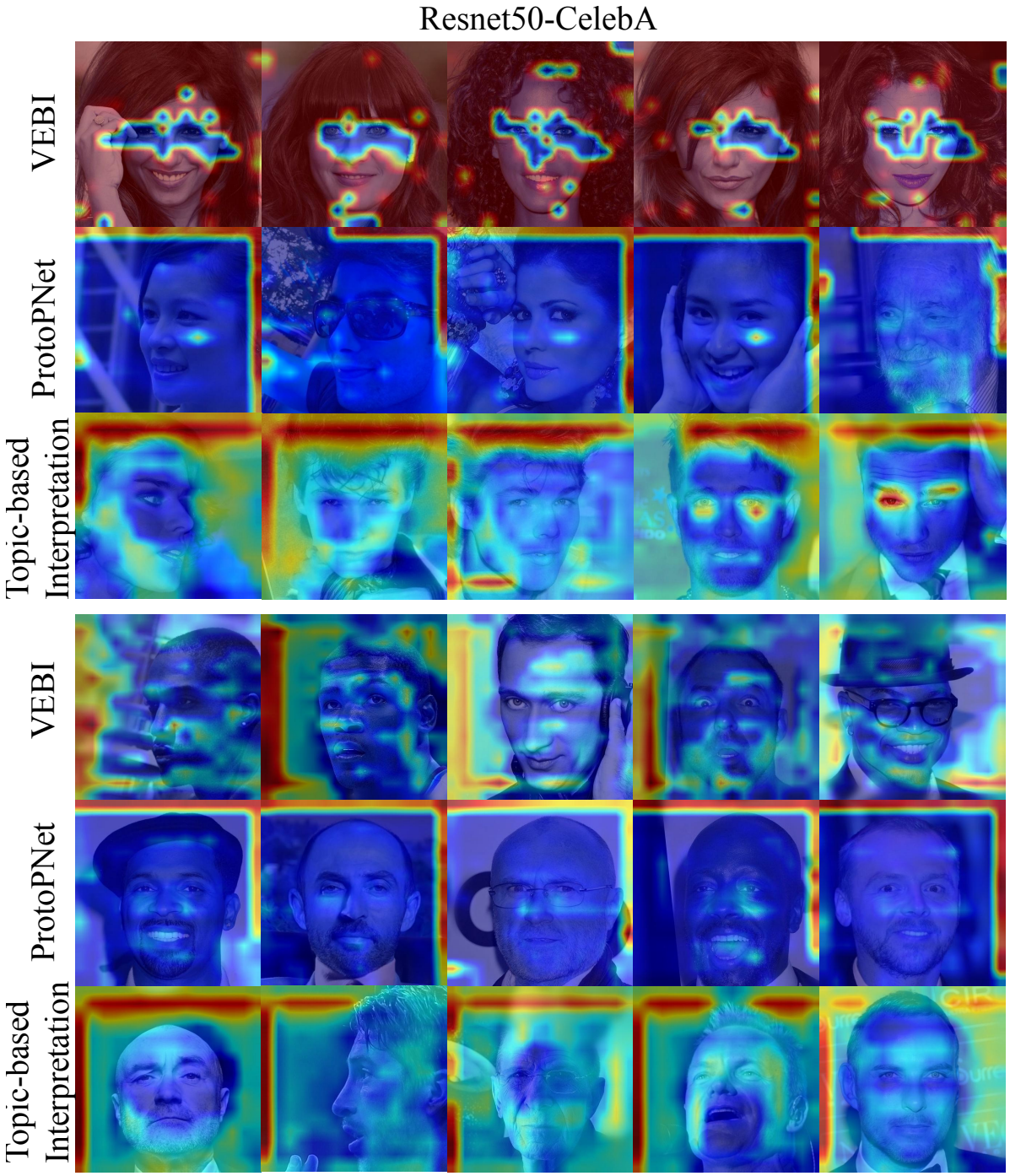

Figure 4. Visual explanations with the highest (top) and lowest (bottom) IoU coverage for each of VEBI, ProtoPNet, and Topic-based interpretation methods on Resnet50 trained on CelebA dataset.

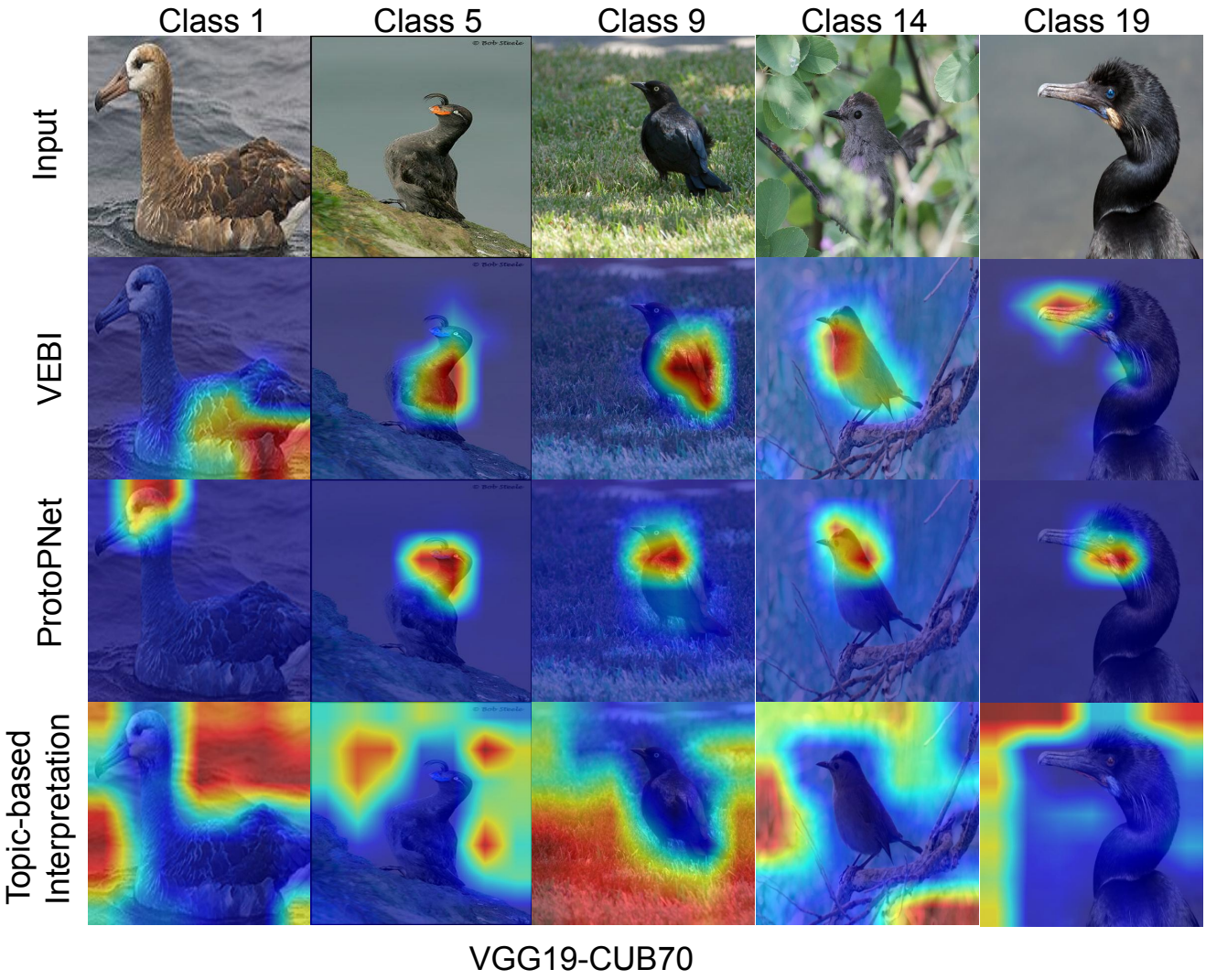

Figure 5. Visual explanations of VEBI, ProtoPNet, and Topic-based interpretation methods on VGG19 trained on the CUB-70 dataset.

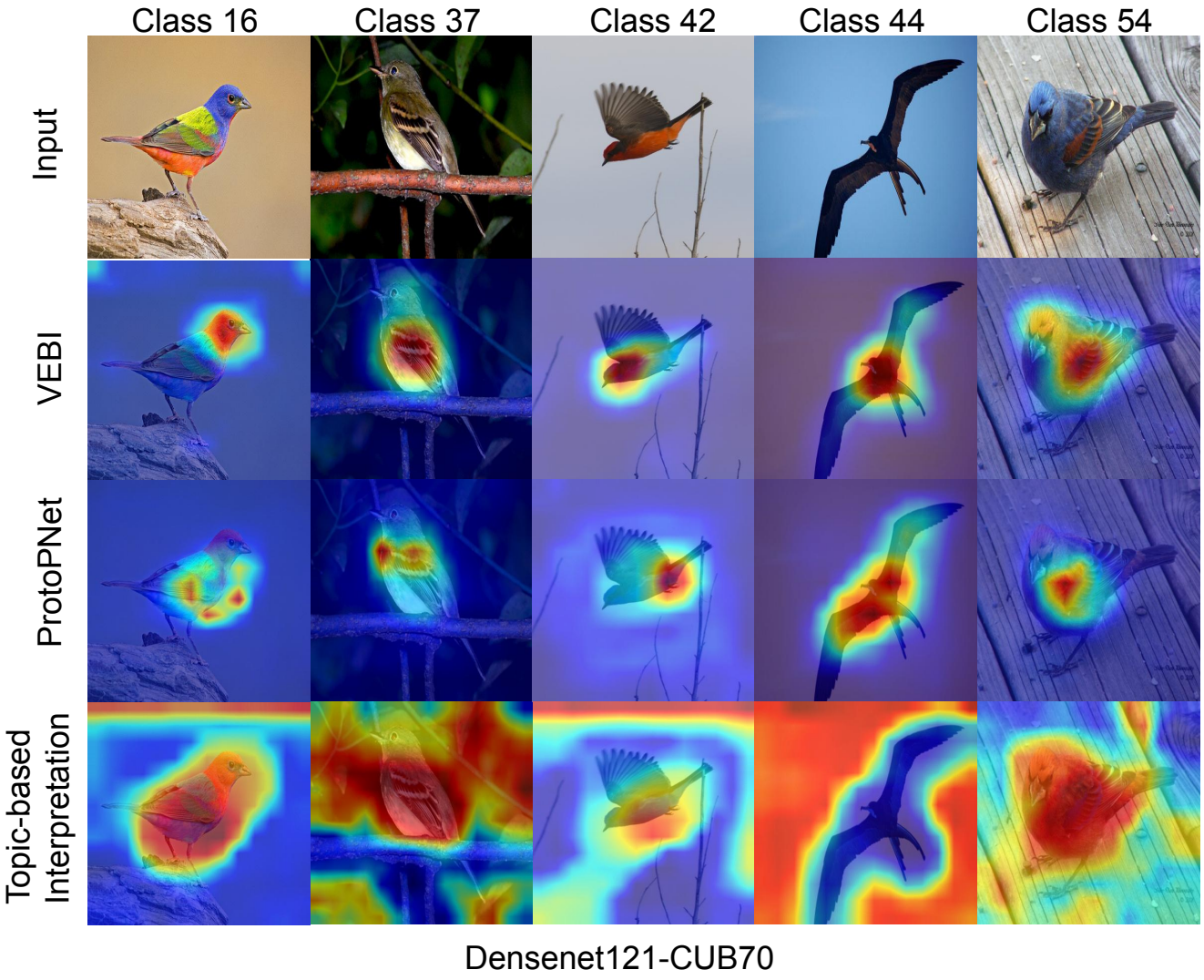

Figure 6. Visual explanations of VEBI, ProtoPNet, and Topic-based interpretation methods on Densenet121 trained on the CUB-70 dataset.

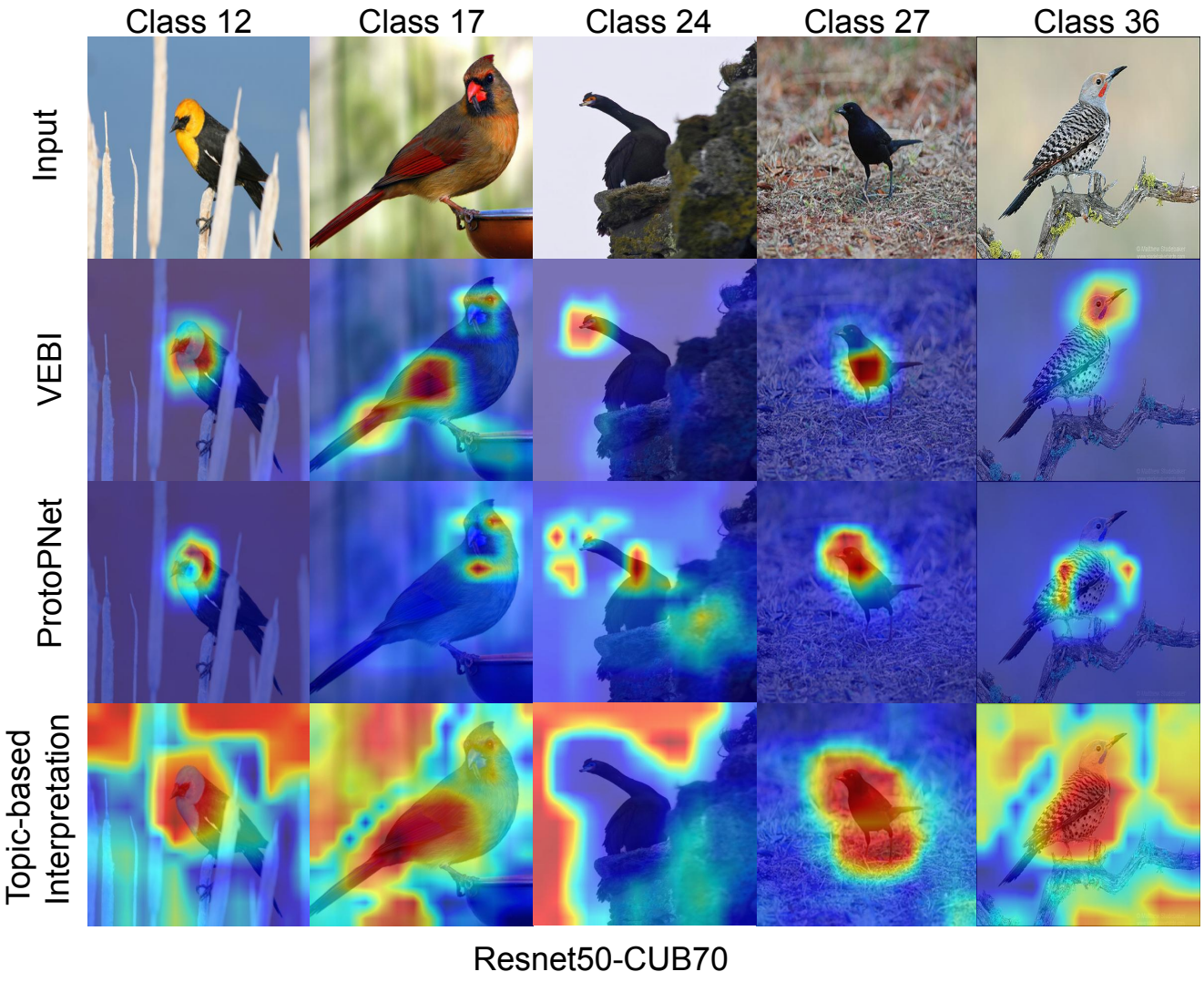

Figure 7. Visual explanations of VEBI, ProtoPNet, and Topic-based interpretation methods on Resnet50 trained on the CUB-70 dataset.

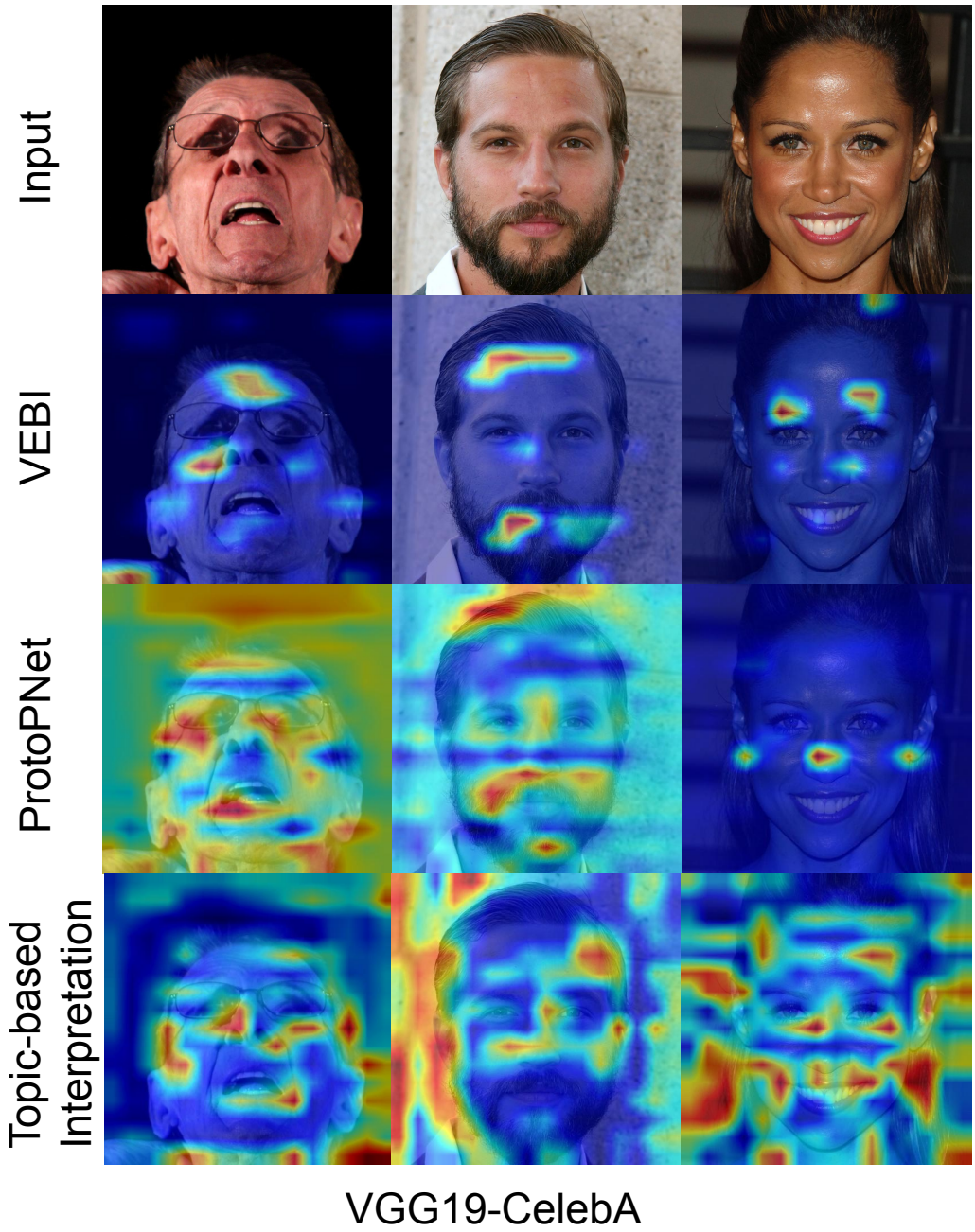

Figure 8. Visual explanations of VEBI, ProtoPNet, and Topic-based interpretation methods on VGG19 trained on the CelebA dataset.

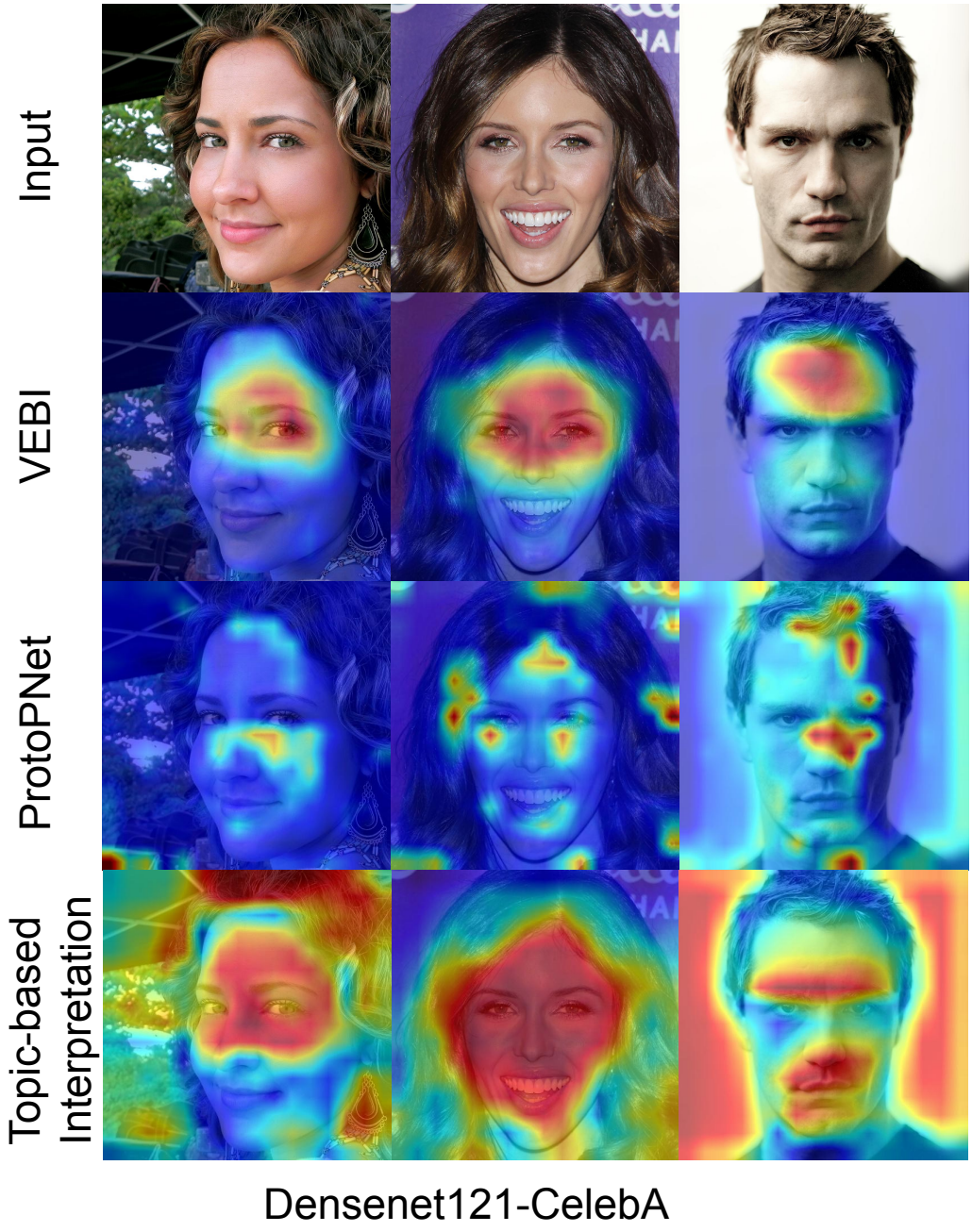

Figure 9. Visual explanations of VEBI, ProtoPNet, and Topic-based interpretation methods on Densenet121 trained on the CelebA dataset.

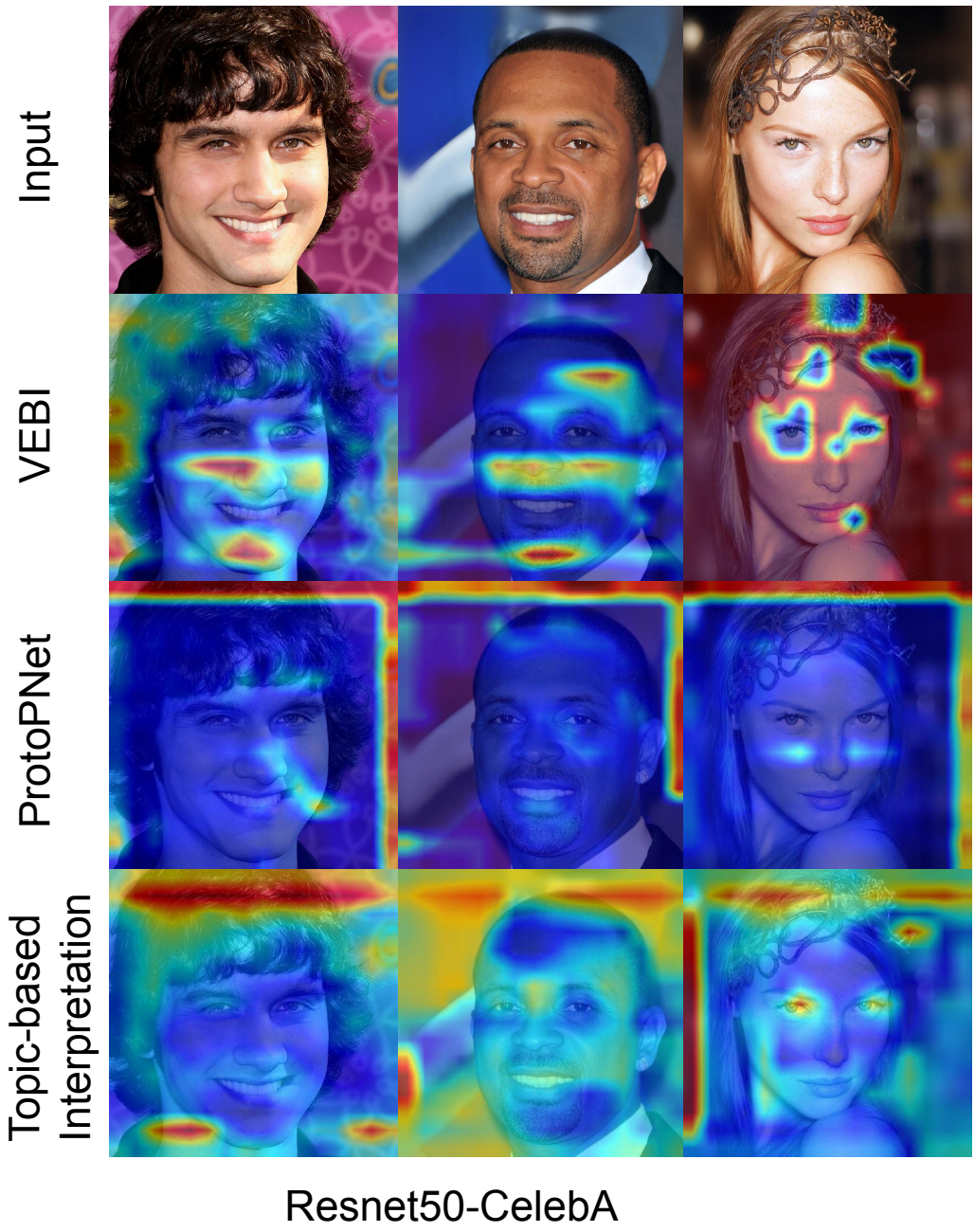

Figure 10. Visual explanations of VEBI, ProtoPNet, and Topic-based interpretation methods on Resnet50 trained on the CelebA dataset.

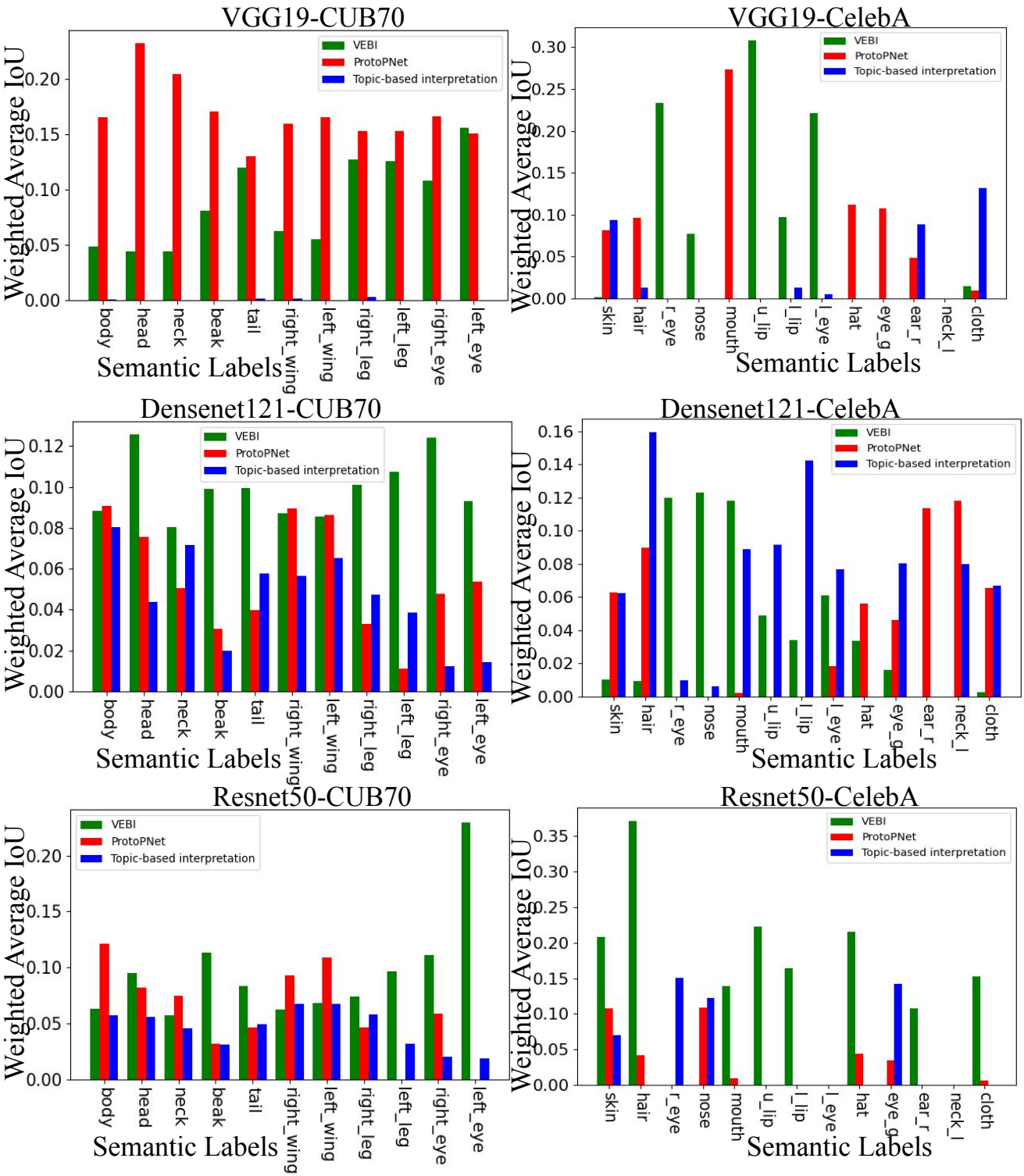

Figure 11. Semantic part-level comparison among model interpretation methods VEBI, ProtoPNet, and Topic-based over CNNs VGG19, Densenet121, and Resnet50 trained on the CelebA and CUB-70 datasets.

Figure 12. Visual explanations of VEBI, ProtoPNet, and Topic-based interpretation methods on VGG19 trained on the CelebA dataset.

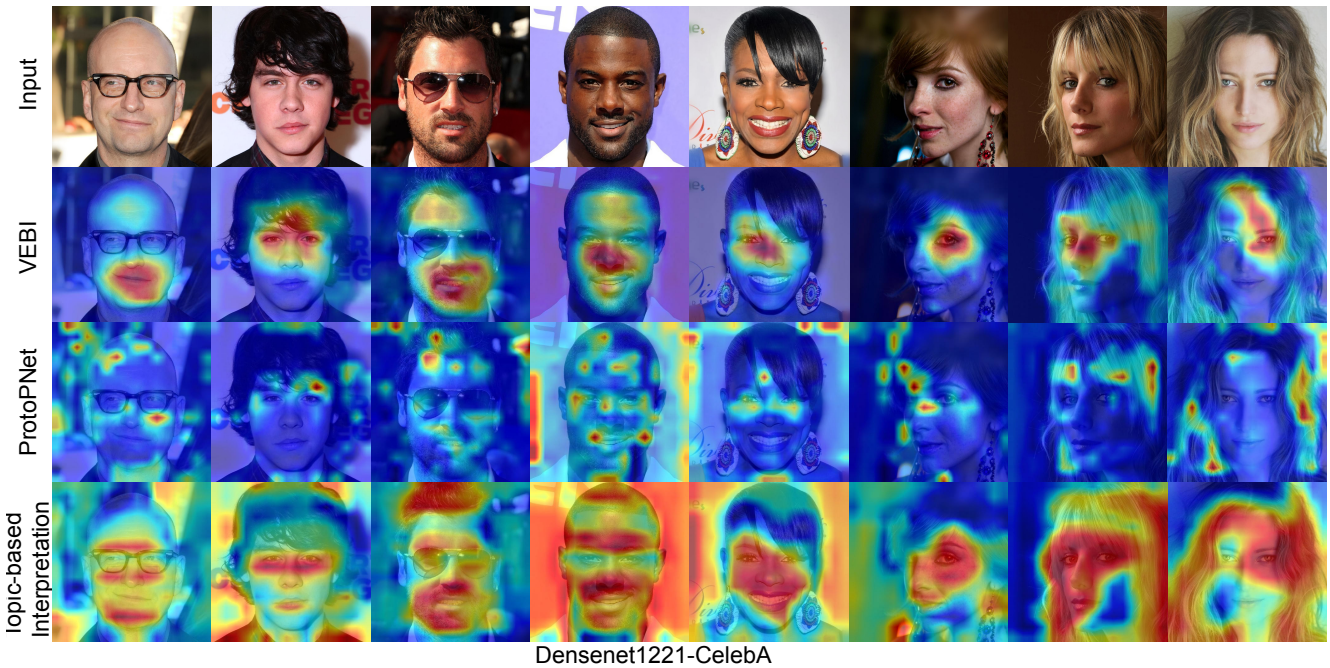

Figure 13. Visual explanations of VEBI, ProtoPNet, and Topic-based interpretation methods on Densenet121 trained on the CelebA dataset.

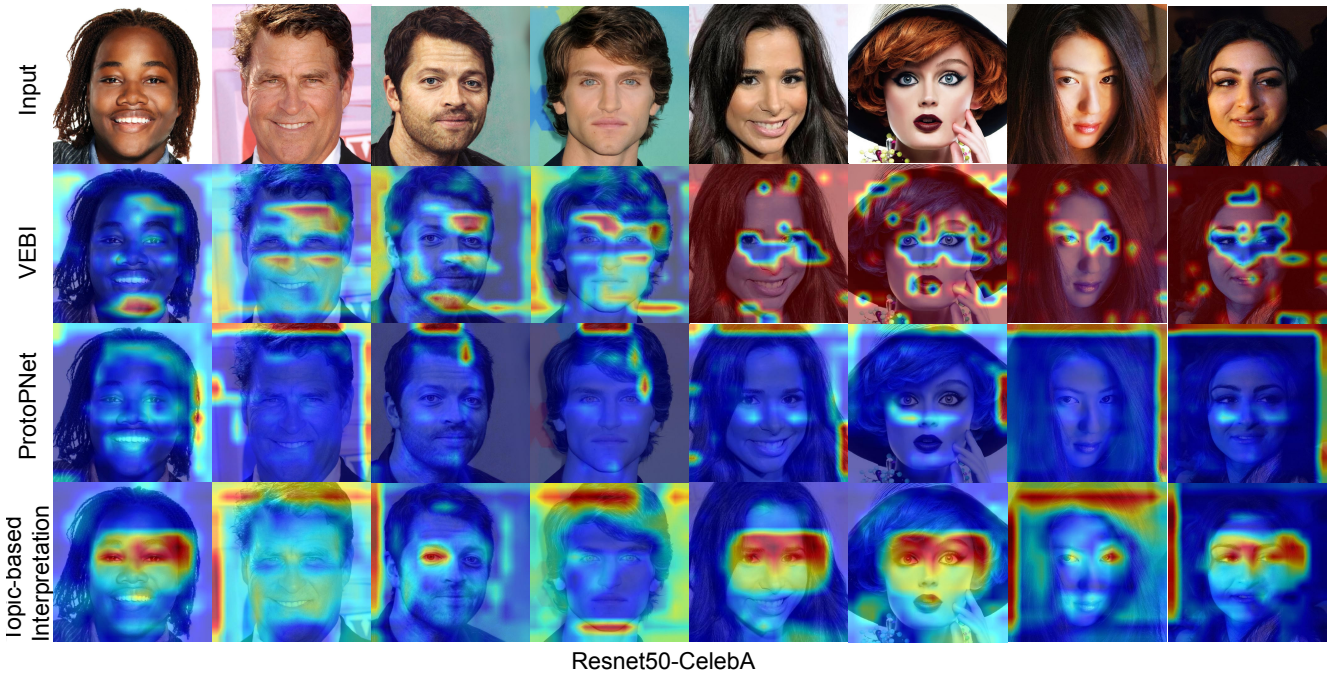

Figure 14. Visual explanations of VEBI, ProtoPNet, and Topic-based interpretation methods on Resnet50 trained on the CelebA dataset.

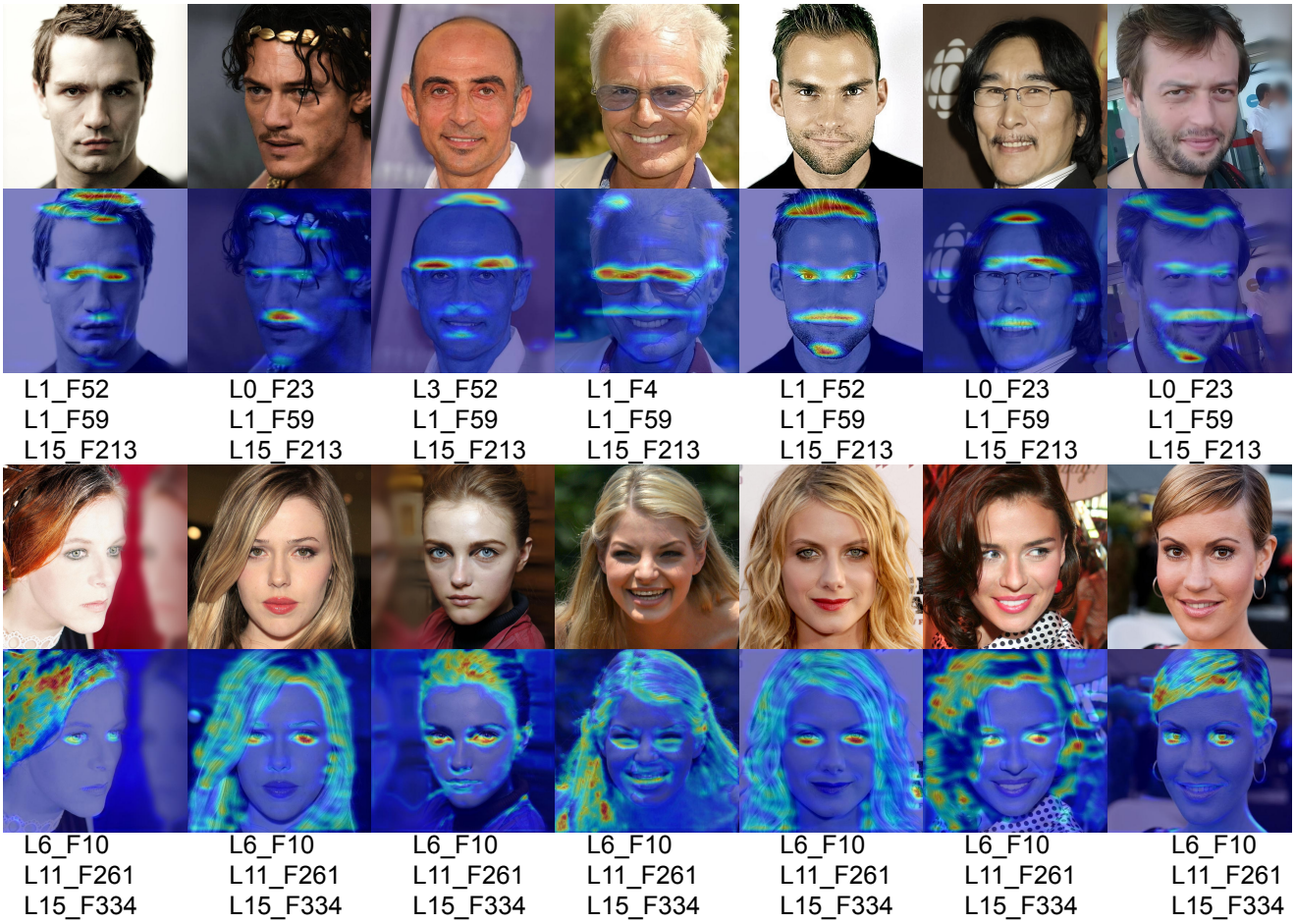

Figure 15. Visual explanations of original VEBI on VGG19 trained on CelebA dataset. The caption shows the pairs convolutional layers and filters identified by the original VEBI.

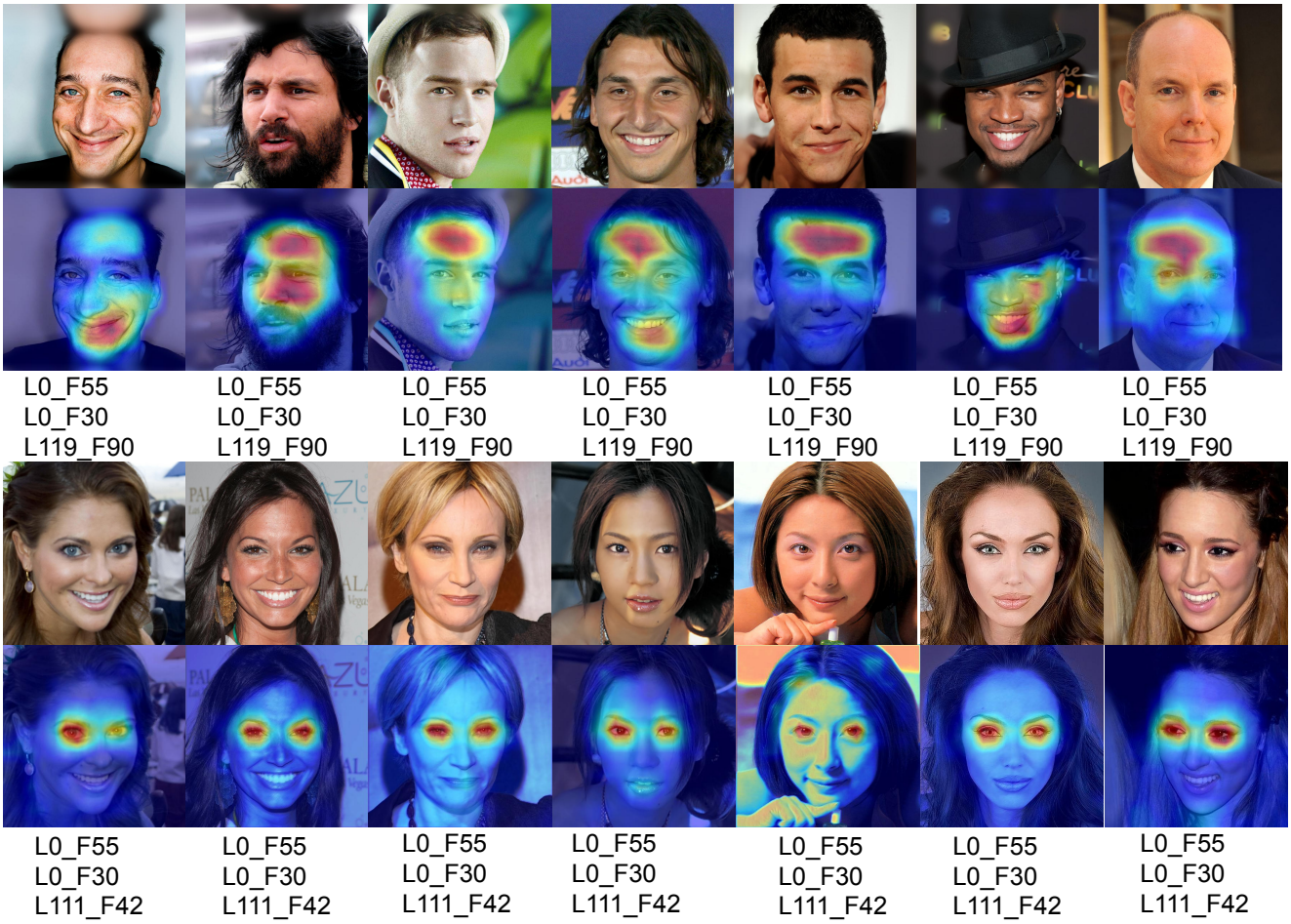

Figure 16. Visual explanations of original VEBI on Densenet121 trained on CelebA dataset. The caption shows the pairs convolutional layers and filters identified by the original VEBI.

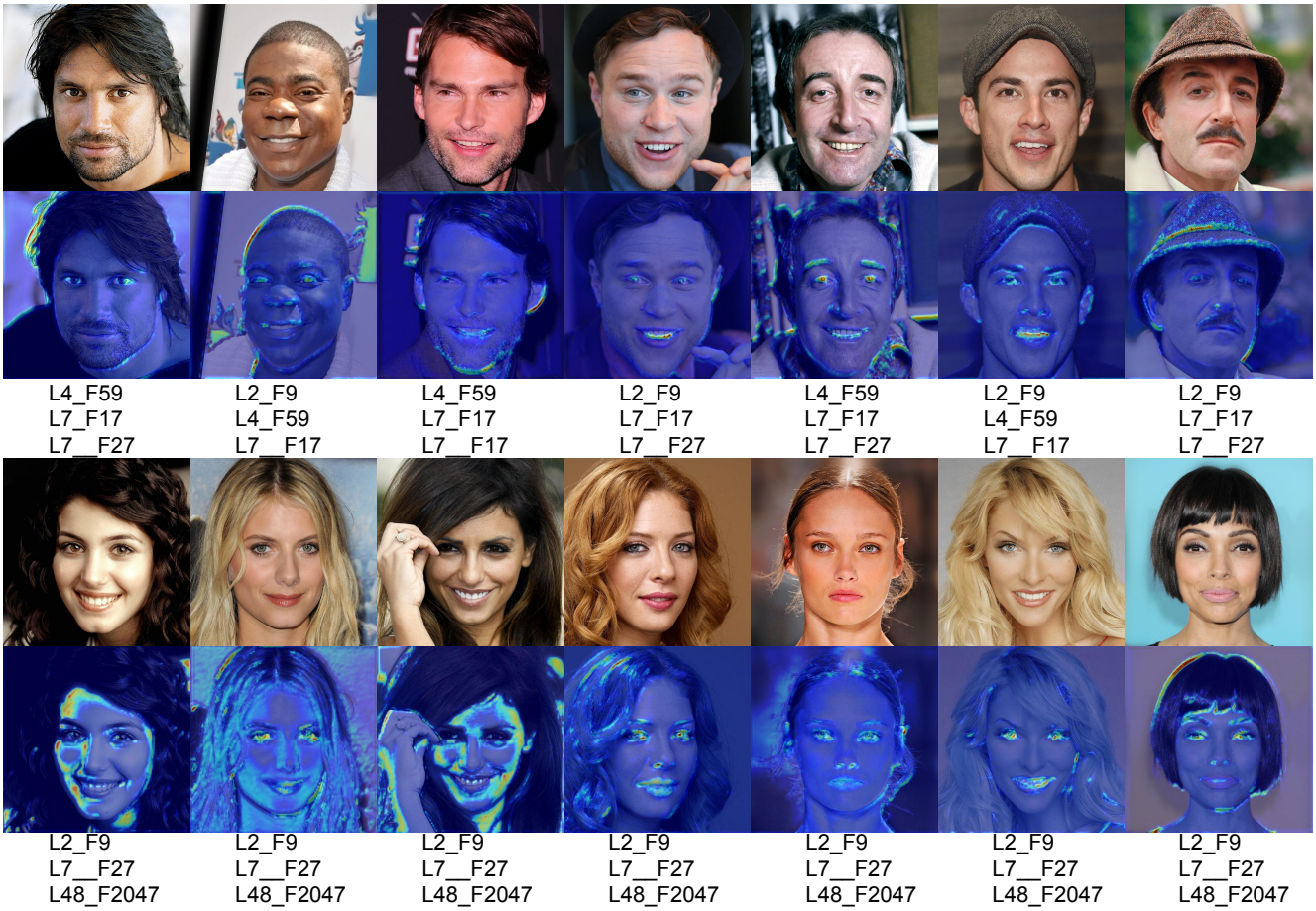

Figure 17. Visual explanations of original VEBI on Resnet50 trained on CelebA dataset. The caption shows the pairs convolutional layers and filters identified by the original VEBI.

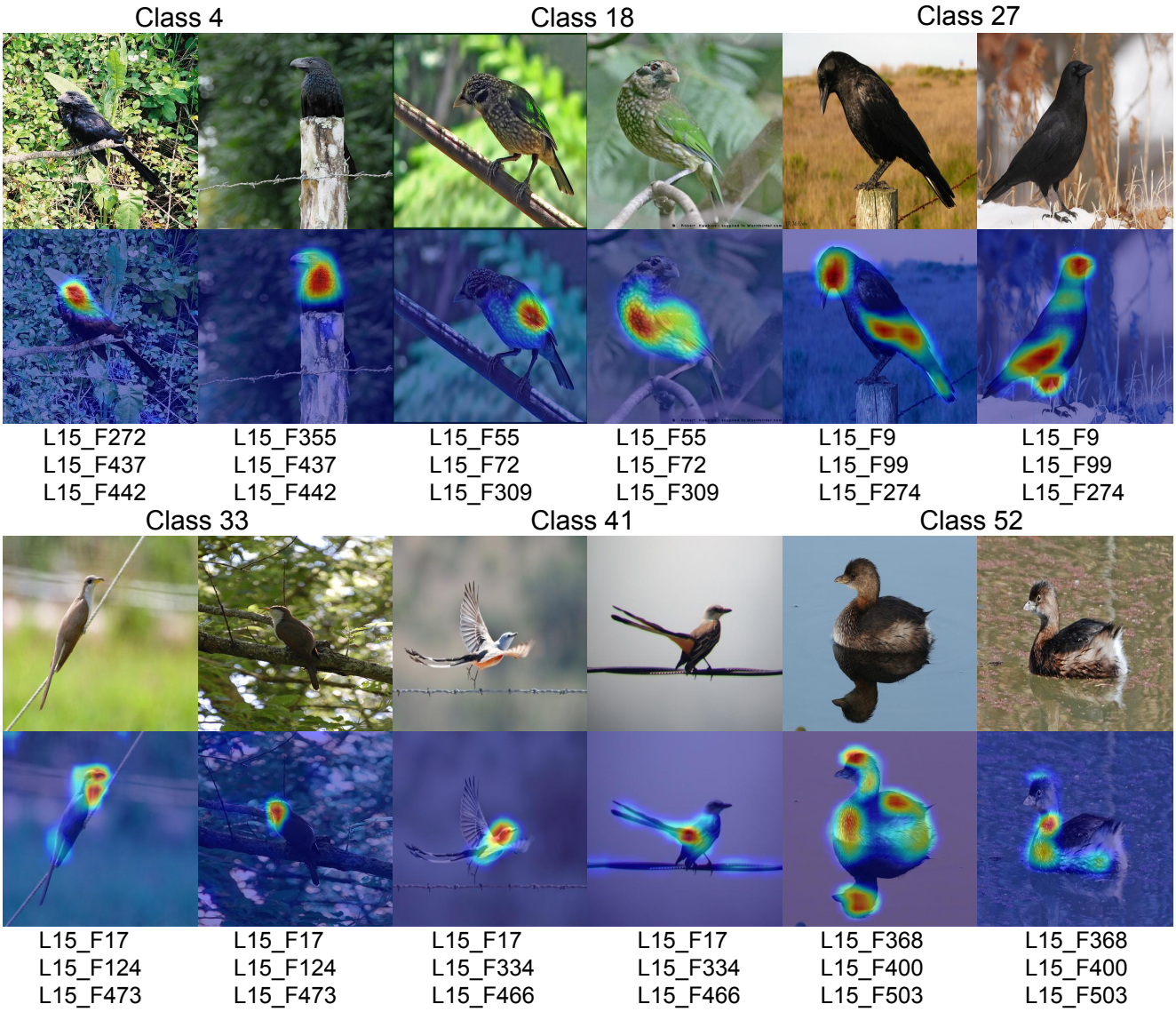

Figure 18. Visual explanations of original VEBI on VGG19 trained on CUB70 dataset. The caption shows the pairs convolutional layers and filters identified by the original VEBI.

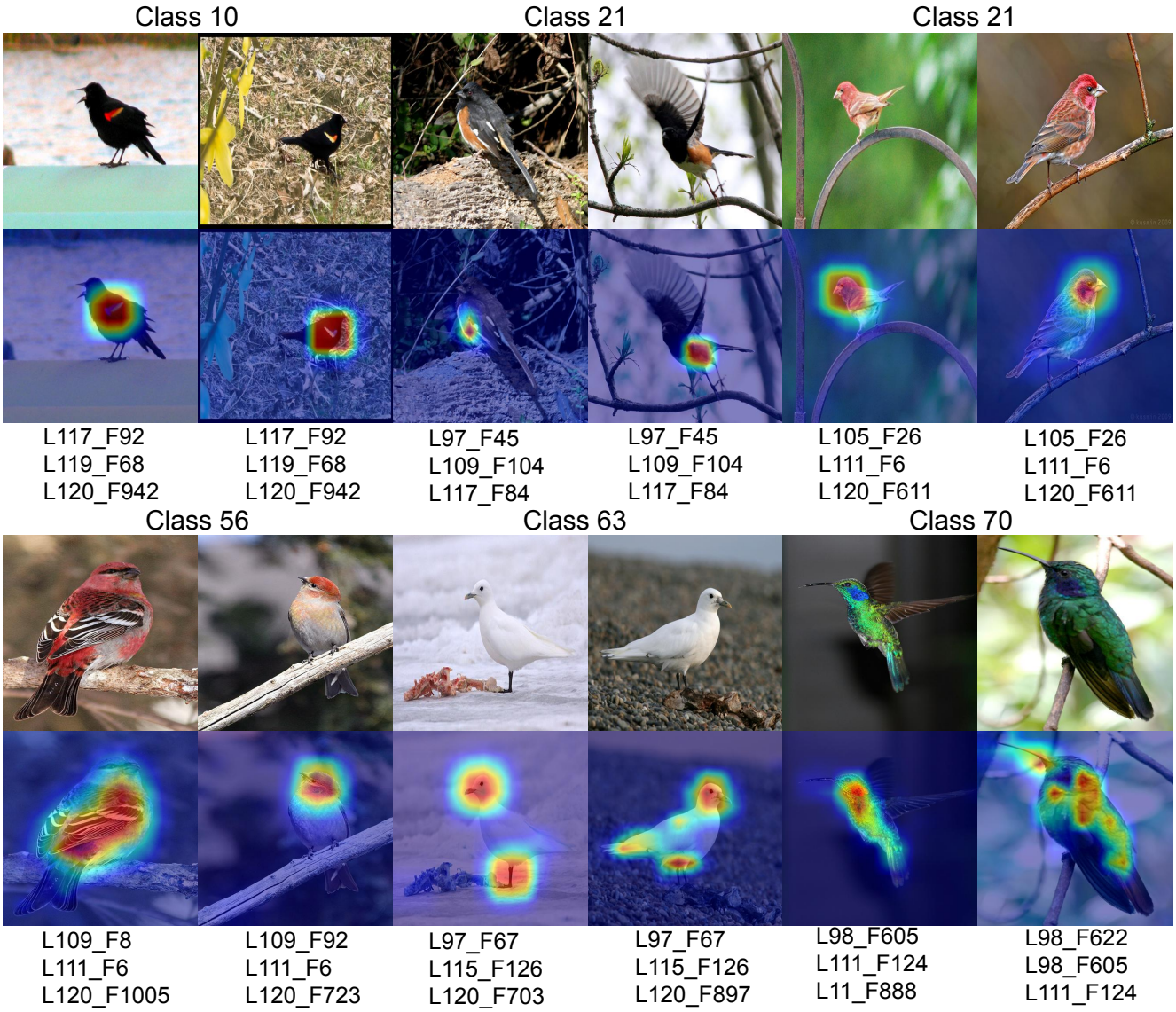

Figure 19. Visual explanations of original VEBI on Densenet121 trained on CUB70 dataset. The caption shows the pairs convolutional layers and filters identified by the original VEBI.

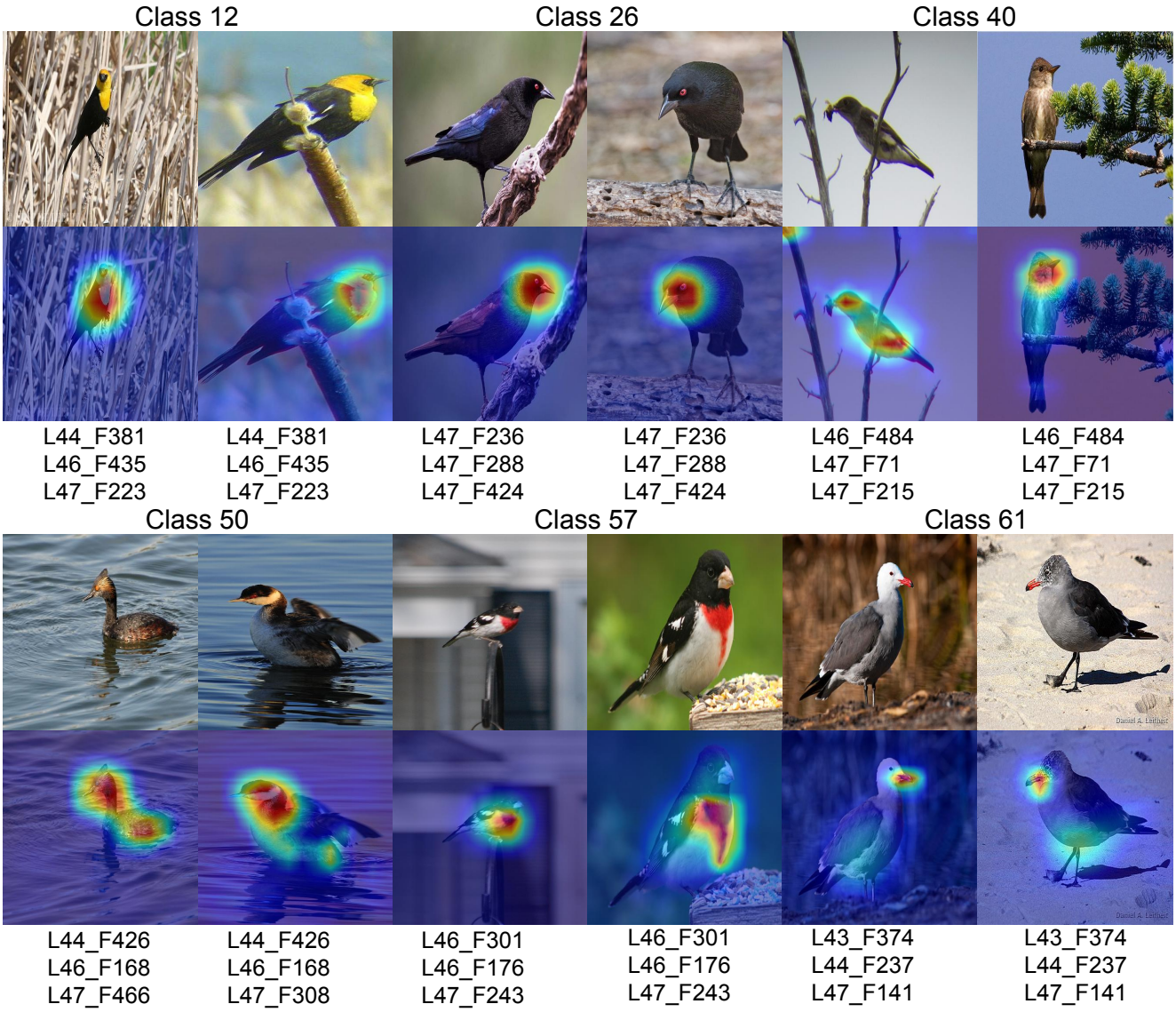

Figure 20. Visual explanations of original VEBI on Resnet50 trained on CUB70 dataset. The caption shows the pairs convolutional layers and filters identified by the original VEBI.