Local Path Integration for Attribution
(Supplementary Material)

Abstract

In the supplementary material, we first provide more quantitative results on two additional metrics. Then, an ablation study using the DiffID metric is presented for further testing the fidelity of the proposed Local Path Integration. Moreover, additional visualization results are also provided for qualitative inspection. The results reported in this document will also be made public with full code of the submission through a project page after acceptance.

1 Experiments on ROAR and DiffROAR

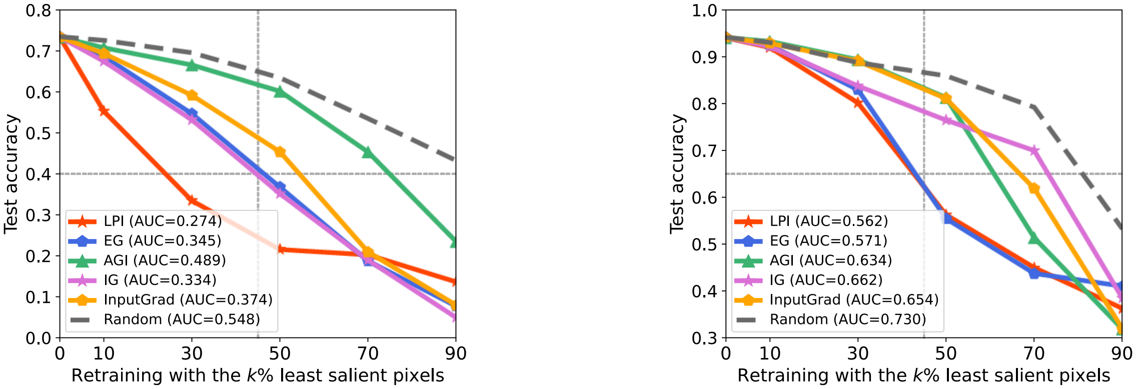
In this section, we provide quantitative results using the existing metrics ROAR and DiffROAR.

1.1 ROAR and DiffROAR

Although pixel perturbation [7] can provide an understanding of the interpretability of the attribution methods, Hooker et al. [4] report that it is unclear whether the degradation in model performance comes from the distribution shift or removing informative features. Therefore, they propse RemOve And Retrain (ROAR) to eliminate the effect of the input distribution shift. ROAR removes the most important pixels from input and retrains the model using the perturbed images. The performance of the retrained model is expected to degrade significantly by removing important pixels. Similar to ROAR, DiffROAR [8] separately retrains the models using images with the least and the most salient pixels removed. Then, DiffROAR measures the accuracy difference between the two models. In DiffROAR, a more reliable attribution method is expected to have a higher difference between the model accuracies. Our proposed DiffID also measures the difference in accuracy, which mitigates the effect of the input distribution shift. Additionally, DiffID does not need to retrain the model, which can ensure the invariance of the explained model and the efficiency of evaluation. These are highly desirable qualities of DiffID. However, since the literature still performs evaluation with ROAR and DiffROAR, we also provide analysis with these metrics to better contextualize our performance in the literature. As discussed shortly, our method generally also outperforms the existing related methods on these metrics as well.

1.2 Experimental Results

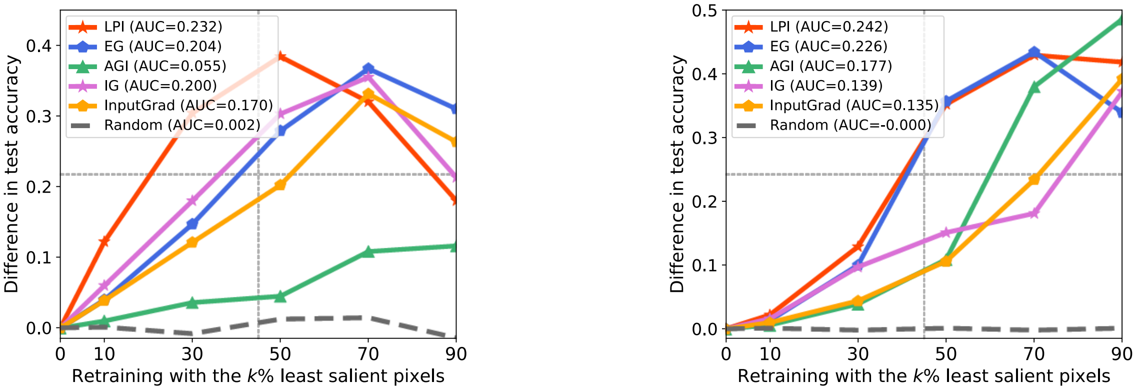
Figure 1 and Figure 2 show the experimental results on both ROAR and DiffROAR benchmarks. We compare the proposed Local Path Integration (LPI) with other attribution methods including a random baseline, InputGrad [9], Integrated Gradients (IG) [11], Expected Gradients (EG) [1], Adversarial Gradient Integration (AGI) [6]. Specifically, we first train a PreActResNet [3] on CIFAR-10 [5] and CIFAR-100 datasets [5] separately. Then we retrain the model with the perturbed images followed by ROAR and DiffROAR. For different attribution methods, we retrain the vanilla PreActResNet with the same epochs (10 epochs) and initial learning rate (0.01) for a fair comparison. Figure 1 shows the results on the ROAR benchmark, which measures the accuracy of the model trained with the least $k\%$ salient input pixels. Figure 2 shows the results on the DiffROAR benchmark, which measures the difference in the accuracy of models trained with the most and least salient $k\%$ pixels. Compared with CIFAR-10, CIFAR-100 with more classes can obtain results with small dispersion. In summary, our LPI method generally outperforms the other attribution methods on both the metrics for the two datasets. These results ascertain the effectiveness of the proposed technique.



(a) ROAR on CIFAR-100 dataset.

(b) ROAR on CIFAR-10 dataset.

Figure 1: Experimental results using the ROAR metric. The curves indicate the test accuracy change with different models trained with input images removed $k\%$ most salient pixels. **Lower** values indicate **better** performance. EG's performance is comparable to our LPI method for CIFAR-10. However, with more classes in CIFAR-100, our method outperforms EG by a large margin.



(a) DiffROAR on CIFAR100 dataset.

(b) DiffROAR on CIFAR10 dataset.

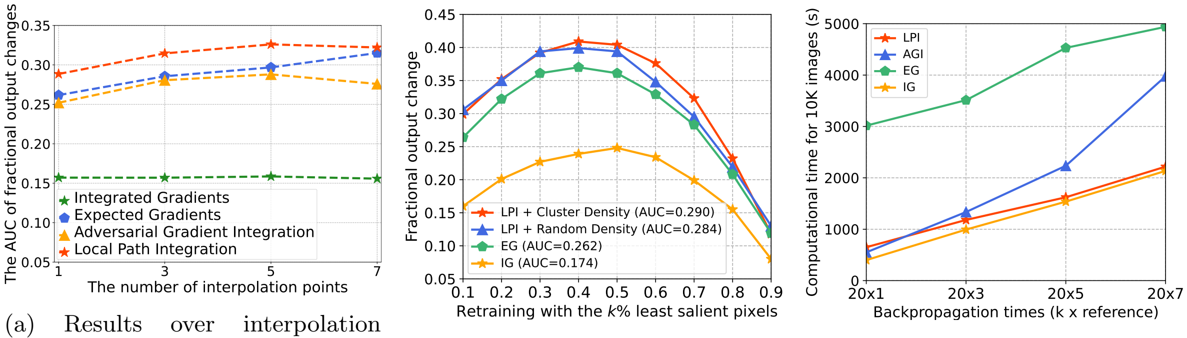
Figure 2: Experimental results using the DiffROAR metric. The curves plot the difference in test accuracy of the models trained with images from which $k\%$ most and least salient pixels are removed. **Higher** values indicate **better** performance.

2 Ablation Study on LPI

In this section, we analyze the effects of underlying components of LPI. We employ ImageNet 2012 Validation Set and ResNet-34 in this analysis.

2.1 Interpolation Points

Path attribution methods calculate the attributions by integrating the gradients along a path from the reference to the input. As the number of calculated interpolation points on this path increases, the results may slightly change. Here, we first test the AUC of the DiffID as the number of interpolation points grows for different methods. The results are shown in Figure 3(a). For EG and LPI, we set fixed number images as references, i.e. 20. In AGI, we accumulate the adversarial noise with 20 steps. In IG, the interpolations points are set from 20 to 140. As such, we are ensuring that the number of backpropagations required are the same for all the methods for a fair comparison. We can observe that the proposed LPI is able to perform the best consistently when the number of interpolation points are varied. This affirms that our intuition of defining local paths is correct.



(a) Results over interpolation points.

(b) Density function comparison.

(c) Computation time comparison.

Figure 3: Ablation study results on the DiffID benchmark. **(a)** The AUC of the fractional output changes on DiffID benchmark along with the interpolation points grows. **Higher** values are more desirable. Our method consistently outperforms other methods. **(b)** The density function comparison on the DiffID benchmark. **Higher** values indicate **better** performance. Our random density variant LPI + Random Density is still able to outperform EG, which testifies that integration over local paths is more effective than the global ones. **(c)** The computation time comparison over different number of backpropagations. Our method achieves competitive performance with regards to efficiency. **Lower** values indicate **better** efficiency.

2.2 Density Function

For LPI, we propose a technique to efficiently estimate a distribution by clustering. We employ the density of each cluster as the density function. In Figure 3(b), we test the effectiveness of proposed techniques on the DiffID benchmark. Figure 3(b) shows that the performance of LPI with clustering density can outperform the LPI with random density. In addition, both the clustering density and random density can outperform EG, which also demonstrates that integrating gradients in the local neighborhood is still more effective than integrating them globally, even when we do not particularly account for the densities of the local clusters.

2.3 Computational Overhead

In Figure 3(c), we evaluate the computation time for 10K images sampled from ImageNet Val set over different numbers of backpropagations. The results show that the efficiency of LPI approaches IG for 20 backpropgations, which are used in our experiments. EG has a high memory overhead as it needs to load large dataset (e.g., ImagNet) to sample the reference. Hence, it has the worst efficiency among the solutions. AGI has similar efficiency as LPI when based on one targeted adversarial class.

3 Visual Inspection

Figure 4 and Figure 5 show the attribution maps for ResNet-34 [2] and VGG-16 [10]. We compare the proposed LPI with other feature attribution methods including InputGrad, IG, EG and AGI. In addition, we choose the same images on both ResNet-34 and VGG-16 for visual inspection in Figure 6 for convenient performance comparison across models. We can observe that the proposed LPI attribution consistently aligns well with the foreground object for both models. Also, LPI's maps are more consistent as compare to EG.

4 Experimental Platform

In the specific implementation, we employ the PyTorch deep learning framework (v1.12.1) with Python language. We train and test all models on one NVIDIA GTX 3090Ti GPU with 24GB memory.

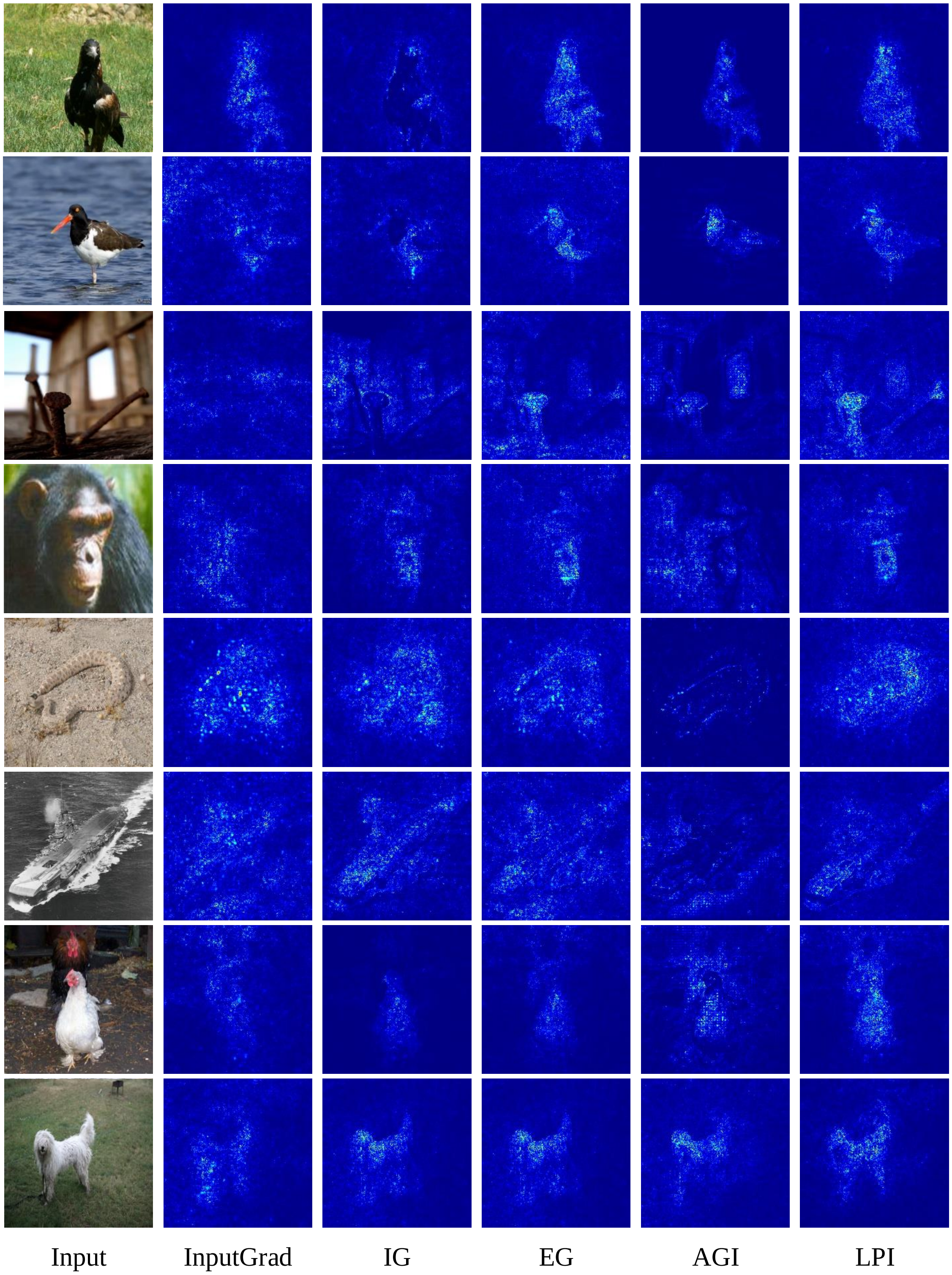


Figure 4: The attribution maps calculated by different attribution methods on ResNet-34.

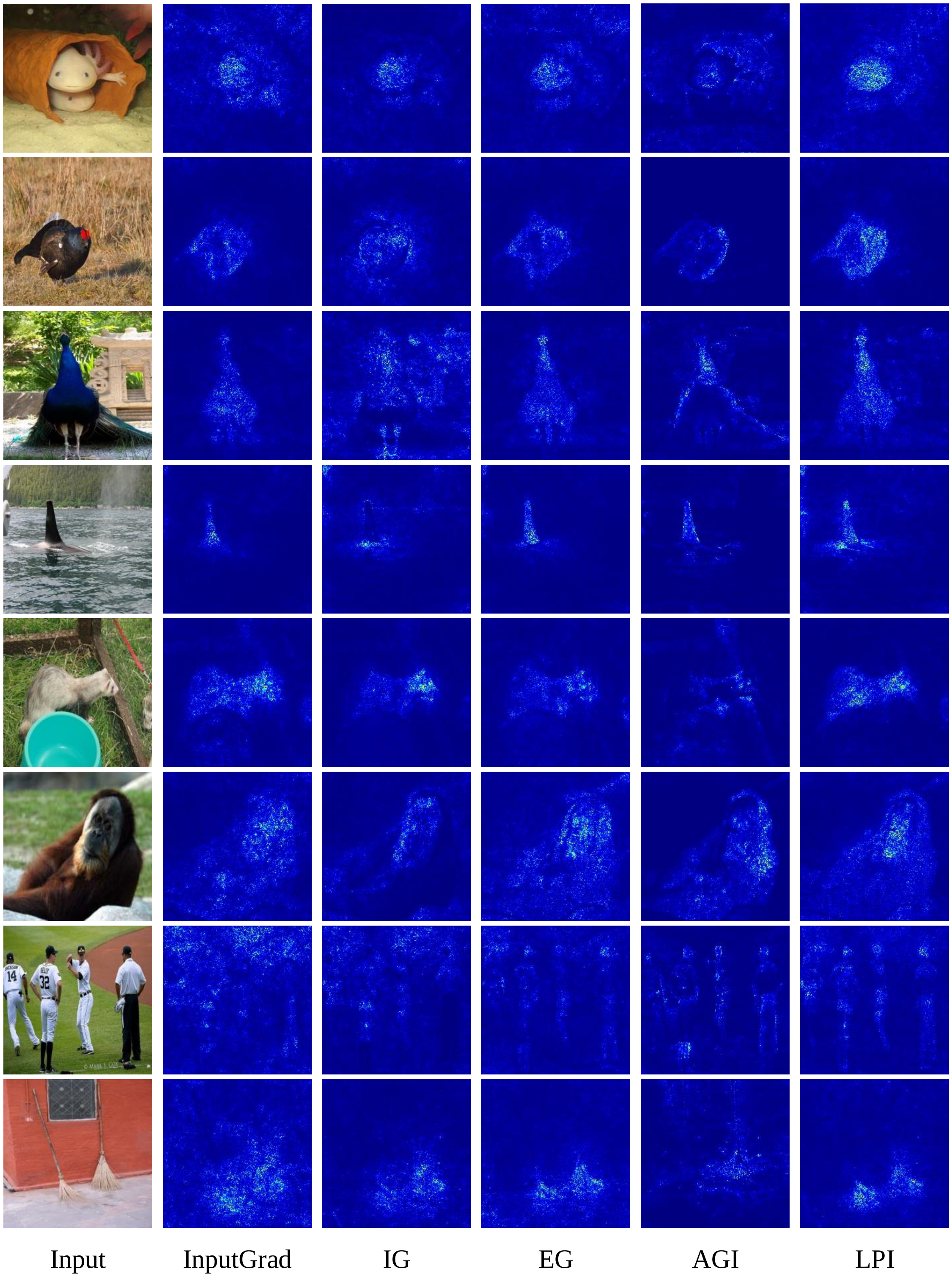


Figure 5: The attribution maps calculated by different attribution methods on VGG-16.

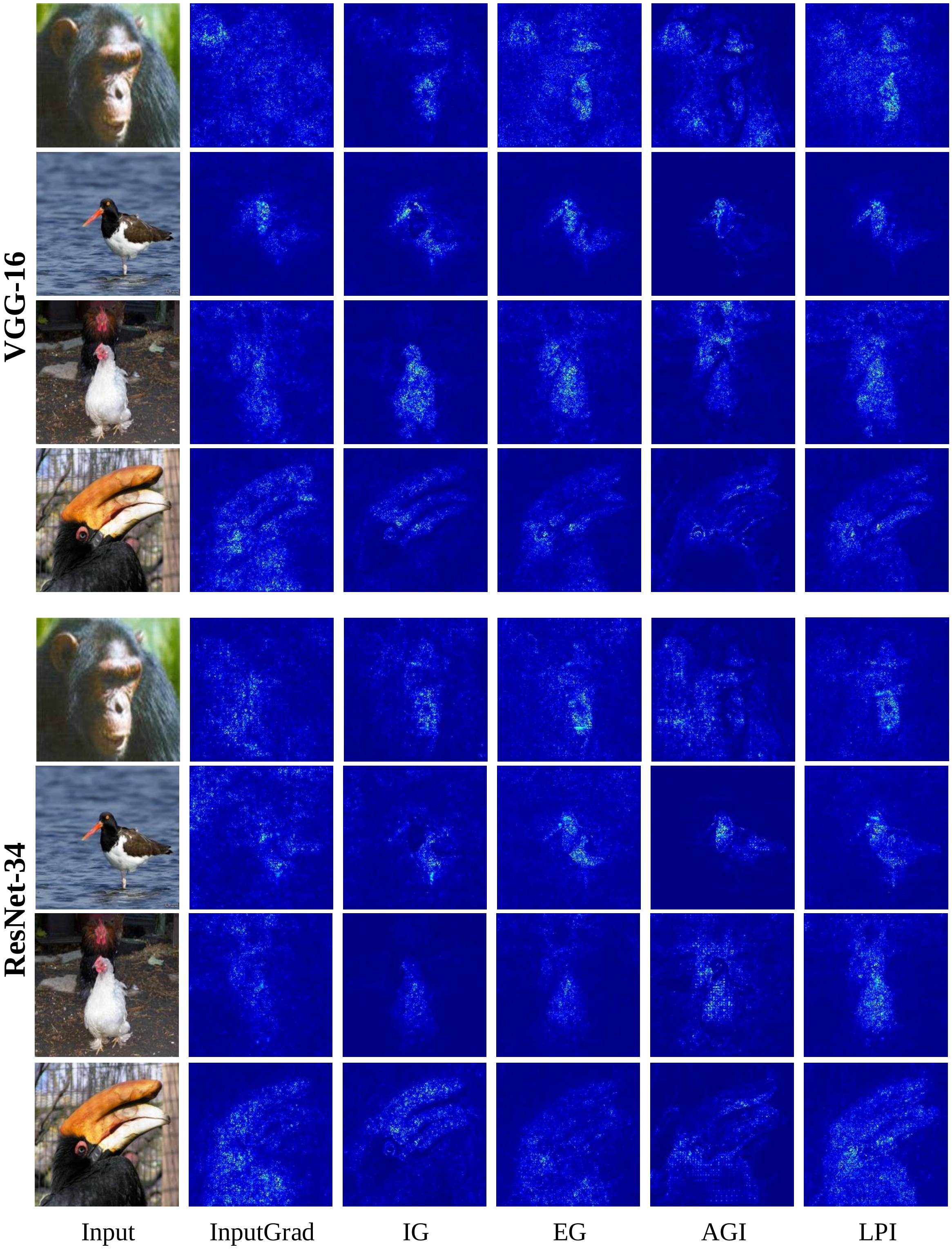


Figure 6: The attribution maps calculated by different attribution methods on both VGG-16 and ResNet-34. Same images are used for a convenient cross-model comparison.

References

[1] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 2021.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision, ECCV*, 2016.

[4] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in NeuralInformation Processing Systems, NeurIPS*, 2019.

[5] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

[6] Deng Pan, Xin Li, and Dongxiao Zhu. Explaining deep neural network models with adversarial gradient integration. In *International Joint Conference on Artificial Intelligence, IJCAI*, 2021.

[7] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 2016.

[8] Harshay Shah, Prateek Jain, and Praneeth Netrapalli. Do input gradients highlight discriminative features? *Advances in NeuralInformation Processing Systems, NeurIPS*, 2021.

[9] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop on International Conference on Learning Representations, ICLR*, 2014.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations, ICLR*, 2015.

[11] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning, ICML*, 2017.