Finer-CAM 🔍 : Spotting the Difference Reveals Finer Details for Visual Explanation

Supplementary Material

The supplementary material is organized into the following sections. Appendix A provides more details of the method implementation and experiment settings. Appendix B discusses experimental results on more datasets and model architectures, and Appendix C presents more visualizations of the proposed Finer-CAM.

A. More Implementation Details

A.1. Sorted Weight Similarity Distribution

We show the distribution of sorted weight similarity of three datasets in Fig. 1. Here we provide the implementation details for reproducing the curves. First, we train a linear classifier for each dataset on top of the pre-trained CLIP visual encoder [32]. The visual encoder is frozen during the classifier training. Next, we calculate the similarity matrix $\boldsymbol{S}$ between the weights of the linear classifier with each element S_{pq} defined by:

$$S_{pq} = \frac{\boldsymbol{w}^p \cdot \boldsymbol{w}^q}{\|\boldsymbol{w}^p\|_2 \|\boldsymbol{w}^q\|_2}, \tag{11}$$

where $\boldsymbol{w}^p$ and $\boldsymbol{w}^q$ represent the linear classifier weights for class p and class q, respectively. The diagonal elements are subtracted by 1 to eliminate self-similarity. The similarity matrix is then sorted in descending order for each class:

$$\boldsymbol{S}^{\text{sorted}} = \text{sort\_rows}(\boldsymbol{S}), \tag{12}$$

such that the first element of each row has the largest similarity to the query class. Last, we compute the class-wise average of the sorted similarity values to generate the distribution curve. The shaded regions in the figure stand for standard variation. Therefore, the leftmost point of each curve reflects the average similarity between one class and its most similar counterpart. Although after model training, the average similarity is low, for each class, there still exist certain other classes with high similarities.

A.2. Activation Faithfulness Examination

Based on the extension to multi-modal zero-shot models, the proposed Finer-CAM can be applied to verify if the prediction of a linear classifier faithfully aligns with real class attributes, as illustrated in Sec. 4.3. Here we provide more implementation details of the process. The CUB-200 dataset [40] provides continuous attribute labels for each class. Given one target class, we conduct subtraction between the attribute labels of the target class and that of the most similar class. The top 3 attributes with the largest value difference are selected as discriminative attributes, and are to be highlighted in the image.

Next, we generate two saliency maps for one given image of the specified class. The first saliency map is obtained based on Eq. (5) to maximize the difference between the target class and similar class prediction logits. It reflects the distinctions recognized by the classifier model. The second saliency map is obtained by setting text attribute labels and the general category "bird" as comparing pairs in the zero-shot classification setting. It shows the "ground truth" difference between the two classes. Subsequently, we can compare if the classifier-based saliency map activates similar regions as the attribute-based one. An aligned saliency map pair indicates that the classifier is looking at correct hints to distinguish the image. Oppositely, if the saliency maps misalign, either the classifier is not working as expected, or there are certain traits not labeled by the dataset.

Table 3. Classification accuracy (%) of linear probing on DINOv2 and CLIP backbones on five datasets.

Model	Birds-525	CUB-200	Cars	Aircraft	FishVista
CLIP	95.3	58.4	64.9	53.9	64.6
DINOv2	97.5	66.4	78.7	83.9	79.6

A.3. Dataset Information

We utilized five publicly available datasets to evaluate our method. Below, we summarize the key characteristics of each dataset, including the number of categories, sample distributions, and additional details provided by the respective dataset sources.

- **Birds-525** [31]: This dataset comprises 525 bird species with 84,635 training images and 2,625 validation images. It provides a diverse collection suitable for fine-grained classification tasks.
- **CUB-200** [40]: This dataset is a benchmark for fine-grained categorization with 11,788 images across 200 bird species. The dataset includes 5,994 training images and 5,794 testing images, with detailed annotations such as subcategory labels, part locations, and bounding boxes.
- **Cars** [20]: This dataset contains 16,185 images of 196 car models, split into 8,144 training images and 8,041 testing images. Categories include make, model, and year, making it ideal for subtle visual recognition tasks.
- **Aircraft** [23]: This dataset comprises 10,200 aircraft images annotated across 70 family-level categories. The

Table 4. The quantitative evaluation results on the proposed Finer-CAM and baseline CAM methods on FishVista and Aircraft. The abbreviations stand for deletion, relative drop, and localization, respectively.

Method	FishVista			Aircraft			
	Del. ↓	RD.$_{@0.05}$ ↑	RD.$_{@0.1}$ ↑	Del. ↓	RD.$_{@0.05}$ ↑	RD.$_{@0.1}$ ↑	Loc. ↑
Grad-CAM [37]	**0.037**	0.177	0.205	**0.039**	0.097	0.112	0.608
+ Finer	0.039	**0.193**	**0.217**	**0.039**	**0.113**	**0.127**	**0.614**
Layer-CAM [17]	**0.049**	0.163	0.181	**0.037**	0.101	0.113	0.662
+ Finer	**0.049**	**0.196**	**0.210**	0.039	**0.113**	**0.124**	**0.664**
Score-CAM [41]	**0.051**	0.158	0.188	**0.050**	0.074	0.086	0.595
+ Finer	0.052	**0.174**	**0.203**	**0.050**	**0.085**	**0.094**	**0.602**

Table 5. The quantitative evaluation results on the proposed Finer-CAM and baseline CAM methods **with DINOv2 as the backbone**. The abbreviations stand for deletion, relative drop, and localization, respectively.

Method	Birds525			CUB				Cars			
	Del. ↓	RD.$_{@0.05}$ ↑	RD.$_{@0.1}$ ↑	Del. ↓	RD.$_{@0.05}$ ↑	RD.$_{@0.1}$ ↑	Loc. ↑	Del. ↓	RD.$_{@0.05}$ ↑	RD.$_{@0.1}$ ↑	Loc. ↑
Grad-CAM [37]	0.252	0.041	0.069	0.171	0.124	0.157	0.500	**0.088**	0.222	0.280	0.619
+ Finer	**0.250**	**0.049**	**0.080**	**0.165**	**0.151**	**0.185**	**0.530**	0.091	**0.243**	**0.306**	**0.632**
Layer-CAM [17]	**0.254**	0.047	0.075	**0.143**	0.174	0.210	0.682	**0.105**	0.210	0.270	0.618
+ Finer	0.258	**0.055**	**0.079**	0.148	**0.192**	**0.230**	**0.729**	0.108	**0.236**	**0.294**	**0.647**
Score-CAM [41]	**0.282**	**0.042**	0.062	**0.174**	0.125	0.157	0.630	**0.152**	0.127	0.173	0.579
+ Finer	0.284	0.036	**0.064**	0.176	**0.137**	**0.168**	**0.640**	**0.152**	**0.141**	**0.191**	**0.586**

Table 6. The quantitative evaluation results on the proposed Finer-CAM and baseline CAM methods on FishVista and Aircraft **with DINOv2 as the backbone**. The abbreviations stand for deletion, relative drop, and localization, respectively.

Method	FishVista			Aircraft			
	Del. ↓	RD.$_{@0.05}$ ↑	RD.$_{@0.1}$ ↑	Del. ↓	RD.$_{@0.05}$ ↑	RD.$_{@0.1}$ ↑	Loc. ↑
Grad-CAM [37]	**0.132**	0.206	0.270	**0.178**	0.242	0.309	0.561
+ Finer	0.135	**0.224**	**0.290**	**0.178**	**0.270**	**0.339**	**0.586**
Layer-CAM [17]	**0.129**	0.215	0.278	**0.168**	0.286	0.367	0.729
+ Finer	0.134	**0.220**	**0.288**	0.170	**0.312**	**0.383**	**0.749**
Score-CAM [41]	**0.154**	0.159	0.210	**0.198**	0.182	0.257	0.611
+ Finer	0.159	**0.173**	**0.229**	0.203	**0.194**	**0.264**	**0.653**

dataset is divided into training, validation, and test subsets of 3,334 images each, with hierarchical annotations for classification.

- **FishVista** [24]: This dataset is a large collection of 60,000 fish images spanning 1,900 species, designed for species classification and trait identification. We use a subset of 414 species, with 35,328 training images, 4,996 validation images, and 7,556 test images. It includes fine-grained annotations and pixel-level segmentations for 2,427 images.

B. More Experimental Results

B.1. Model Accuracy

We present the classification accuracy of linear probing on two backbones, CLIP [32] and DINOv2 [28], on five datasets. Tab. 3 summarizes the results. Generally, DINOv2 provides visual embeddings with better quality and achieves higher classification accuracy. We use OpenCLIP ViT-B-16 (pre-trained on LAION-400M) in all the experiments.

B.2. Results on FishVista and Aircraft

In addition to Tab. 1, we also conduct the quantitative evaluation on the FishVista [24] and Aircraft [20] datasets in Tab. 4. Finer-CAM yields similar performance on the deletion AUC as baselines while performing much better in terms of relative drop and localization metrics. The performance superiority further supports the effectiveness of the proposed Finer-CAM method.

B.3. Results on DINOv2

We adopt the pre-trained CLIP model [32] as the backbone in the previous experiments. Here, we further employ DINOv2 [28] to extract visual embeddings for generating

Table 7. The comparison of different aggregation strategies. Del. and RD.$_{@0.05}$ represent deletion AUC and relative drop when masking out the top 5% activated pixels, respectively.

Aggregation	Before `ReLU`		After `ReLU`
	Max	Avg	Avg
Del. ↓	0.081	**0.080**	0.081
RD.$_{@0.05}$ ↑	0.184	**0.192**	0.191

saliency maps. We report the results on the five adopted datasets in Tab. 5 and Tab. 6. Similarly, the proposed Finer-CAM achieves higher relative drop and localization performance compared with baselines. It indicates that Fine-CAM can be applied to a variety of architectures and provide effective interpretation. It can be observed in Tab. 3 that the linear classifier trained on top of DINOv2 achieves higher accuracy than that on CLIP. As a result, it requires deleting more pixels to decrease the confidence of the target class, leading to larger deletion AUC values and smaller relative drop in some cases compared with CLIP.

B.4. Aggregation strategy.

There are multiple potential strategies to aggregate the activations from different comparison references (cf. Eq. (8)). Tab. 7 summarizes the comparison of three aggregation ways. Generally, averaging the activation weights from difference references before the `ReLU` operation yields the best performance.

C. More Visualizations

C.1. Failure Cases

We include some failure cases in Fig. 10. In these examples, the baseline Grad-CAM highlights large portions of the images that are not the target objects. Through further analysis, it often happens when the classifier fails to provide correct prediction. Under this circumstance, Finer-CAM also cannot interpret the decision effectively. Finer-CAM may degenerate to baseline methods when the logit similarity does not reflect visual similarity, *i.e.*, the target class is significantly different from others.

C.2. Multi-modal Interaction

We demonstrate that in addition to interpreting classifiers, the proposed Finer-CAM can also be applied to multi-modal scenarios to localize concepts in the images. We provide more examples in Fig. 11. Using Grad-CAM with the target concept alone often leads to inaccurate or wrong activations. In comparison, with a base concept (*e.g.*, "bird" or "car") as reference, emphasizing their difference produces substantially more accurate localization of fine-grained traits or object parts. We also compare the localization capability with a recent method GEM [4]. GEM is

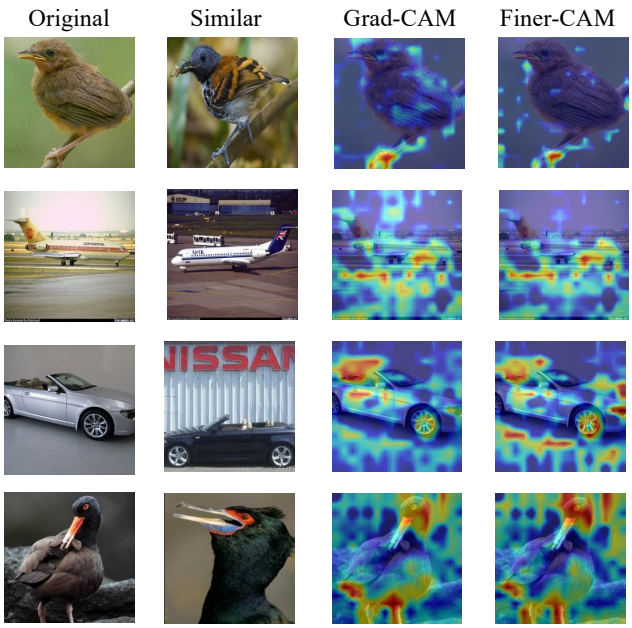

Figure 10. **Visualization of some failure cases** where Finer-CAM cannot produce better saliency maps than the Grad-CAM baseline.

capable to ground the target object in the images. However, when asked to localize fine-grained traits or object parts, it still yields activations over the entire object. Finer-CAM, comparatively, is a better tool to highlight details.

C.3. Qualitative Comparison

We visualize more examples in Fig. 12 on different datasets. The comparison also includes two XAI methods RISE [30] and Mask [12]. The results are obtained with DINOv2 [28] as the backbone. Comparatively, the proposed Finer-CAM activates the most discriminative image regions that can tell the difference between the target class and similar classes, and also suppresses the noise in the background.

C.4. Extrapolation

The proposed Finer-CAM highlights those discriminative image regions that maximize the prediction difference $y^c - \gamma y^d$ between the target class c and the similar class d. We have tested different γ settings from 0.0 to 1.0 in the main text. We also visualize the extrapolation case when $\gamma = 2.0$ in Fig. 13. Generally, the activations are more tensely highlighting subtle details.

C.5. Reverse Comparing

The reverse comparing aims to look for features predictive of the similar class from the target image, which maximize $y^d - y^c$. The visualization examples are shown in the last column of Fig. 13. The generated saliency maps can locate some traits that are predictive of the similar class, instead of the traits highlighted by Finer-CAM.

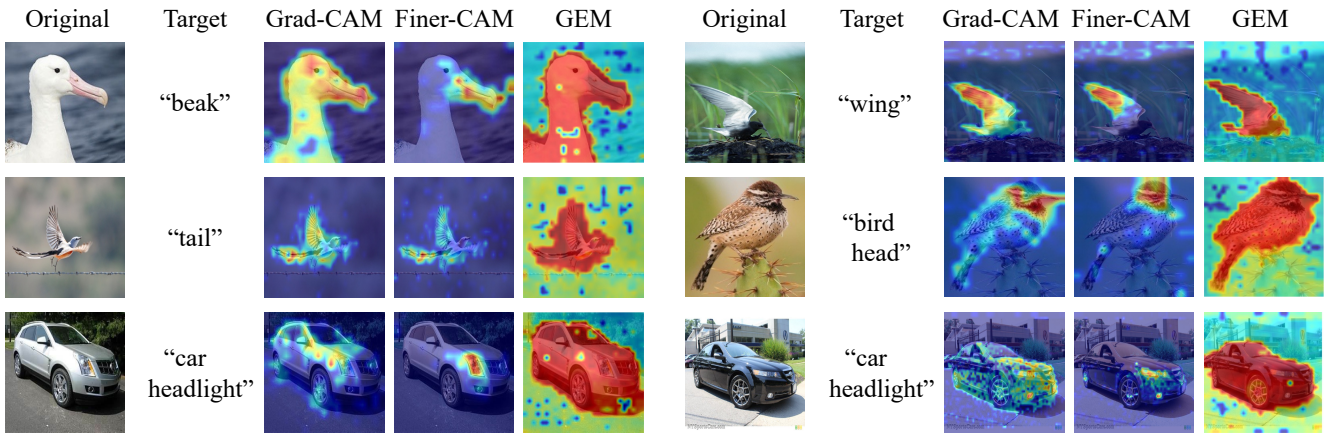

Figure 11. **Visualization of multi-modal localization of fine-grained traits or object parts**. For each original image, we aim to locate the target concept. By emphasizing the difference between the target concept and the original concept (“bird” or “car”), Finer-CAM accurately localizes the target image regions.

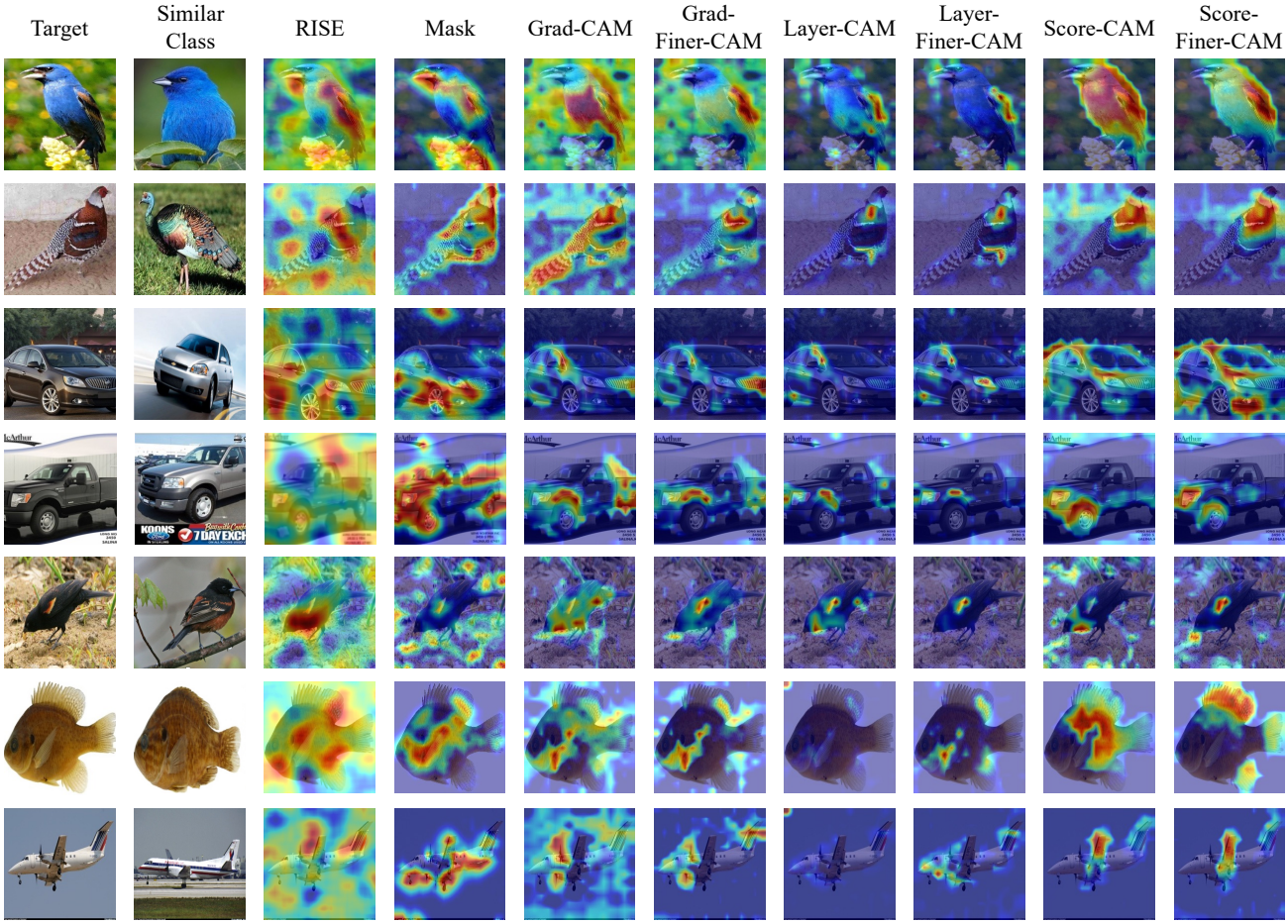

Figure 12. **Qualitative comparison of the saliency maps generated by baseline CAM methods** (Grad-CAM [37], Layer-CAM [17], and Score-CAM [41]), the proposed Finer-CAM applied on these three baselines, and other XAI methods (RISE [30] and Mask [12]).

References

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai):

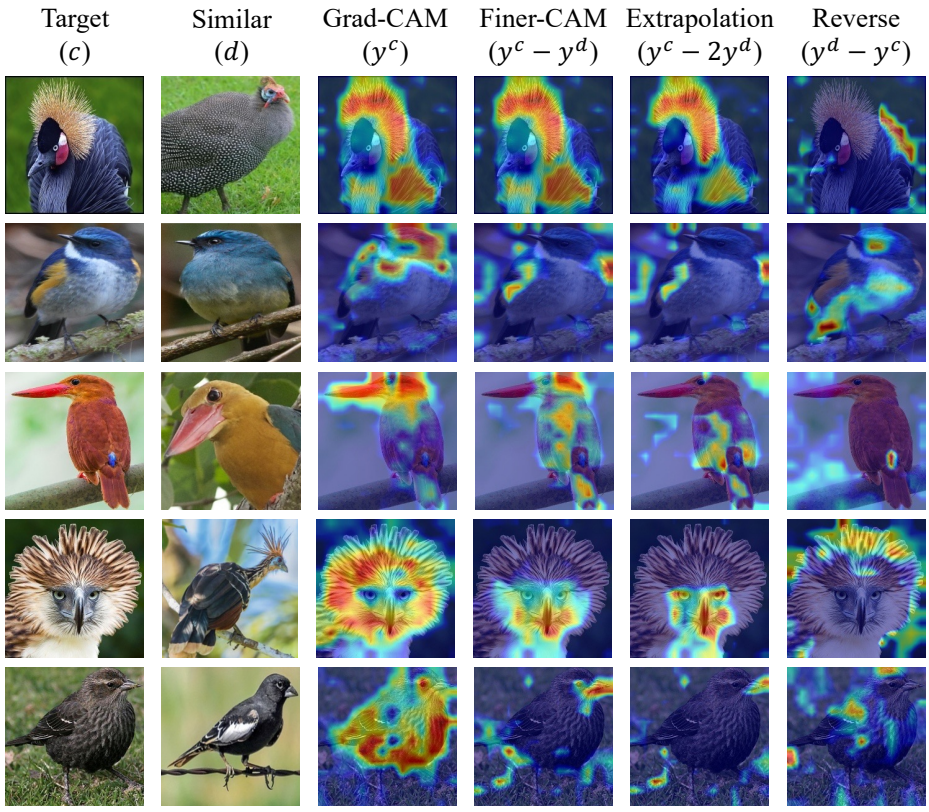

Figure 13. **Visualization of the extrapolation and reverse comparing cases with Grad-CAM as the baseline**. The first two columns show the target image from class c and an image from the similar class d. The next two rows show the saliency maps generated by Grad-CAM and Finer-CAM. Finer-CAM calculates the gradients of the difference between two prediction logits to acquire the activation weights. **Extrapolation** further emphasizes the difference, while **Reverse** tries to look for the traits of the similar class in the target image.

Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020. 2

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10 (7):e0130140, 2015. 5

[3] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *JMLR*, 11:1803–1831, 2010. 3

[4] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2024. 3

[5] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *ICML*, pages 1383–1391. PMLR, 2020. 3

[6] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, pages 839–847. IEEE, 2018. 3

[7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, pages 782–791, 2021. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[10] Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, Marcel van Gerven, and Rob van Lier. *Explainable and interpretable models in computer vision and machine learning*. Springer, 2018. 2

[11] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *ICCV*, pages 2950–2958, 2019. 3

[12] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, pages 3429–3437, 2017. 2, 3, 4

[13] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards

accurate visualization and explanation of cnns. In *BMVC*, pages 1–13, 2020. 3

[14] Naofumi Hama, Masayoshi Mase, and Art B Owen. Deletion and insertion tests in regression models. *Journal of Machine Learning Research*, 24(290):1–38, 2023. 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[17] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *TIP*, 30:5875–5888, 2021. 3, 5, 6, 2, 4

[18] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[19] Lukas Klein, Carsten T Lüth, Udo Schlegel, Till J Bungert, Mennatallah El-Assady, and Paul F Jäger. Navigating the maze of explainable ai: A systematic approach to evaluating methods and metrics. *arXiv preprint arXiv:2409.16756*, 2024. 3, 5, 7

[20] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013. 3, 5, 6, 1, 2

[21] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, page 4768–4777, 2017. 7

[22] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *NeurIPS*, 2017. 3

[23] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 3, 5, 1

[24] Kazi Sajeed Mehrab, M Maruf, Arka Daw, Harish Babu Manogaran, Abhilash Neog, Mridul Khurana, Bahadir Altintas, Yasin Bakis, Elizabeth G Campolongo, Matthew J Thompson, et al. Fish-vista: A multi-purpose dataset for understanding & identification of traits from images. *arXiv preprint arXiv:2407.08027*, 2024. 5, 2

[25] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66, 2022. 2

[26] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *IJCNN*, pages 1–7. IEEE, 2020. 3

[27] Youngrock Oh, Hyungsik Jung, Jeonghyung Park, and Min Soo Kim. Evet: enhancing visual explanations of deep neural networks using image transformations. In *WACV*, pages 3579–3587, 2021. 2

[28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3

[29] Dipanjyoti Paul, Arpita Chowdhury, Xinqi Xiong, Feng-Ju Chang, David Edward Carlyn, Samuel Stevens, Kaiya L Provost, Anuj Karpatne, Bryan Carstens, Daniel Rubenstein, et al. A simple interpretable transformer for fine-grained image classification and analysis. In *ICLR*, 2024. 2

[30] V Petsiuk, A Das, and K Saenko. Rise: Randomized input sampling for explanation of black-box models. In *BMVC*, pages 1–13, 2018. 3, 5, 6, 4

[31] Gerald Piosenka. Birds 525 species - image classification. 2023. 3, 5, 6, 1

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2, 4, 5, 1

[33] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *WACV*, pages 983–991, 2020. 3

[34] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *CVPR*, pages 8839–8848, 2020. 3

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *KDD*, pages 1135–1144, 2016. 2, 3

[36] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021. 1

[37] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 2, 3, 5, 6, 4

[38] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 3

[39] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015. 1

[40] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 3, 5, 6, 8, 1

[41] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *CVPRW*, pages 24–25, 2020. 2, 3, 5, 6, 4

[42] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *TPAMI*, 44(12):8927–8948, 2021. 3

[43] Mengqi Xue, Qihan Huang, Haofei Zhang, Lechao Cheng, Jie Song, Minghui Wu, and Mingli Song. Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. *arXiv preprint arXiv:2208.10431*, 2022. 2

[44] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I Inouye, and Pradeep Ravikumar. On the (in) fidelity and sensitivity of explanations. In *NeurIPS*, pages 10967–10978, 2019. 3

[45] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, 14(2):119–135, 2017. 3

[46] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, pages 5209–5217, 2017. 3

[47] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *CVPR*, pages 5012–5021, 2019. 3

[48] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 2, 3