Probable-class Nearest-neighbor Explanations Improve AI & Human Accuracy

Giang (Dexter) Nguyen
Computer Science Department
Auburn University
nguyengiangbkhn@gmail.com

Valerie Chen
Machine Learning Department
Carnegie Mellon University
vchen2@andrew.cmu.edu

Mohammad Reza Taesiri
Electrical & Computer Engineering Department
University of Alberta
mtaesiri@gmail.com

Anh Totti Nguyen
Computer Science Department
Auburn University
anh.ng8@gmail.com

Abstract

Nearest neighbors (NN) have traditionally been used both for making final decisions—such as in Support Vector Machines or k-NN classifiers—and for providing users with explanations of a model's decisions. In this paper, we introduce a novel set of nearest neighbors to enhance the predictions of a frozen, pretrained image classifier $\mathbf{C}$, thereby integrating performance improvement with explainability. We leverage an image comparator $\mathbf{S}$ that (1) compares the input image with NN images from the top-K most ***probable*** classes given by $\mathbf{C}$; and (2) uses the similarity scores from $\mathbf{S}$ to weight and refine the confidence scores of $\mathbf{C}$. Our method not only consistently improves fine-grained image classification accuracy of $\mathbf{C}$ on datasets such as CUB-(Birds)-200, Cars-196, and Dogs-120 but also enhances the human interpretability of the model's decisions. Through human studies conducted on CUB-200 and Dogs-120 datasets, we demonstrate that presenting users with relevant examples from multiple probable classes help users gain better insight into the model's reasoning process, which improves their decision accuracy compared to prior methods that visualize only the top-1 class training examples Nguyen et al. (2021); Taesiri et al. (2022).

1 Introduction

k-nearest neighbors (NNs) are traditionally considered explainable classifiers by design Papernot & McDaniel (2018). Yet, only recent human studies have found concrete evidence that showing the NNs to humans improves their decision-making accuracy Nguyen et al. (2021); Liu et al. (2022); Chen et al. (2023); Chan et al. (2023); Kenny et al. (2022, 2023); Chiaburu et al. (2024); Nguyen et al. (2024), even more effectively than feature attribution maps in the image domain Nguyen et al. (2021); Kim et al. (2022). These studies typically presented users with the input image, a model's top-1 prediction, and the training NNs from the top-1 class. However, examples from the top-1 class are not always beneficial to users.

One such setting where top-1 NNs can actually hurt human decision-making accuracy is in fine-grained image classification. When users are asked to accept or reject the model's decision—a *distinction* task, as shown in Fig. 1(a), examples from the top-1 class easily fooled users into incorrectly accepting wrong predictions at an excessively high rate—81.5% on CUB-200; Table A6 in Taesiri et al. (2022). This is because NNs from the top-1 class often looked deceivingly similar to the input (e.g., input Elegant vs. Caspian; Fig. 1).

Interpretable AI: Past, Present and Future Workshop at NeurIPS 2024

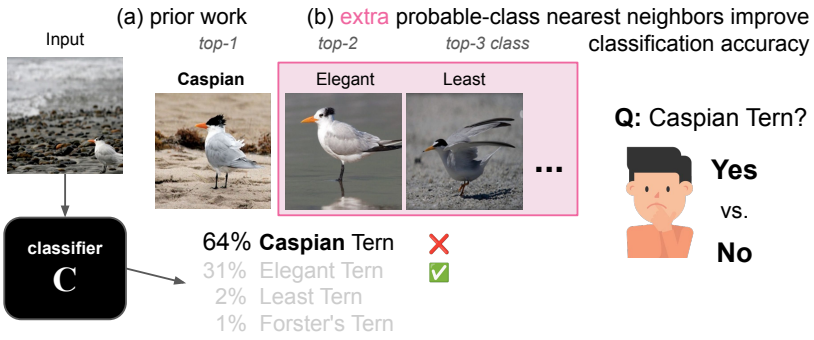


Figure 1: Given an input image x and a black-box, pretrained classifier $\mathbf{C}$ that predicts the label for x, prior work (a) often shows only the NNs from the top-1 predicted class as explanations for the decision, which often *fools* humans into accepting *incorrect* AI recommendations (here, **Caspian** Tern) due to the similarity between the input and top-1 class examples. Instead, including extra nearest neighbors (b) from top-2 to top-K classes improves not only human accuracy on this binary distinction task but also AI's accuracy on fine-grained image classification problems (see Fig. 2).

Instead of focusing on top-1 examples, we propose to diversely sample the NN examples from the top-K predicted, most-probable classes to more fully represent the query and better inform users to make decisions. That is, we propose **Probable-Class Nearest Neighbors** (PCNN), a novel explanation type consisting of K nearest training-set images, where each image is taken from a class among the top-K classes, as illustrated in Fig. 1(b). We show that PCNN not only improves human decisions on the distinction task over showing top-1 examples but can also be leveraged to improve AI-alone accuracy by **re-ranking** the predicted labels of a pretrained, frozen classifier $\mathbf{C}$ (Fig. 2).

Assumptions As is the case for many real-world applications, we assume that there exists a pretrained, black-box classifier $\mathbf{C}$, e.g., a foundation model Bommasani et al. (2021), responsible for a large amount of information processing in the pipeline. Due to computation and algorithm constraints, $\mathbf{C}$ may not be easily re-trained to achieve better accuracy. Therefore, like Bansal et al. (2021), we assume that $\mathbf{C}$ is frozen—humans or other models would interact with $\mathbf{C}$ to make final decisions (Fig. 1).

To leverage PCNN for re-ranking $\mathbf{C}$'s predicted labels, we train an image comparator $\mathbf{S}$, which is a binary classifier that compares the input image with each PCNN example and outputs a sigmoid value that is used to weight the original confidence scores of $\mathbf{C}$ (Fig. 2). Then, $\mathbf{C}$ and $\mathbf{S}$ together form a $\mathbf{C} \times \mathbf{S}$ model that outperforms $\mathbf{C}$ alone.

Our experiments on 10 different classifiers $\mathbf{C}$ across 3 fine-grained classification tasks for bird, car, and dog species reveal:[1].

- Our $\mathbf{C} \times \mathbf{S}$ consistently improves upon the original $\mathbf{C}$ accuracy on all three domains: CUB-200, Cars-196, and Dogs-120 (Sec. 3.1).
- In human studies with 60 participants on CUB-200 and 32 on Dogs-120, we find that PCNN explanations significantly help users calibrate their reliance on AI predictions, resulting in substantial accuracy improvements on the distinction task. Specifically, user accuracy increased by nearly 10 percentage points on CUB-200 (from 54.55% to 64.58%) and by over 5 percentage points on Dogs-120 (from 63.55% to 69.21%) when using PCNN explanations, compared to showing only top-1 class examples (Sec. 3.2).

2 Methods

2.1 Datasets and pretrained classifiers $\mathbf{C}$

We train and test our method on three standard fine-grained image classification datasets of birds, cars, and dogs. To study the generalization of our findings, we test a total of 10 classifiers $\mathbf{C}$ (4 bird, 3 car, and 3 dog classifiers) of varying architectures and top-1 performance.

CUB-200 (CUB-200-2011) Wah et al. (2011) has 200 bird species, with 5,994 images for training and 5,794 for testing. We test four different classifiers: a ResNet-50 pretrained on iNaturalist Van Horn et al. (2018) and finetuned on CUB-200 (85.83% accuracy) by Taesiri et al. (2022); and three

[1]Code and data are available at `https://github.com/anguyen8/nearest-neighbor-XAI`

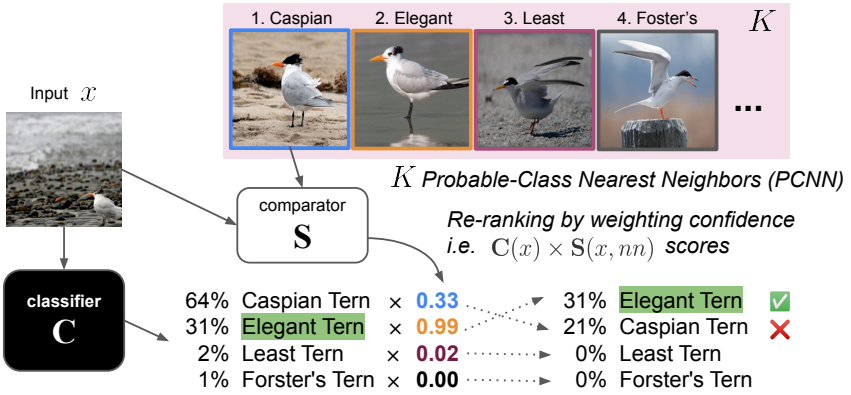


Figure 2: $\mathbf{C} \times \mathbf{S}$ re-ranking algorithm: From each class among the top-K predicted classes by $\mathbf{C}$, we find the nearest neighbor nn to the query x and compute a sigmoid similarity score $\mathbf{S}(x, nn)$, which weights the original $\mathbf{C}(x)$ probabilities to re-rank the labels.

ImageNet-pretrained ResNets (18, 34, and 50 layers) finetuned on CUB-200 with 60.22%, 62.81%, and 62.98% accuracy, respectively.

Cars-196 (Stanford Cars) Krause et al. (2013) includes 196 distinct classes, with 8,144 images for training and 8,041 for testing. We use ResNet-18, ResNet-34, and ResNet-50, all pretrained on ImageNet and then finetuned on Cars-196. Their top-1 accuracy scores are 86.17%, 82.99%, and 89.73%, respectively.

Dogs-120 (Stanford Dogs) Khosla et al. (2011) has a total 120 of dog breeds, with 12,000 images for training and 8,580 images for testing. We test three models: ResNet-18, ResNet-34, and ResNet-50, all pretrained on ImageNet and then finetuned on Dogs-120, achieving top-1 accuracy of 78.75%, 82.58%, and 85.82%.

2.2 Re-ranking using both image comparator S and classifier C

PCNN is a set of K *nearest*-neighbor images to the query. Each of PCNN examples is taken from one training-set class among the top-K predicted classes by $\mathbf{C}$ (see Fig. 2). We empirically test $K = \{1, 2, 3, 5, 10, 15\}$ and find $K = 10$ to be optimal for the top-1 classification accuracy.

The **distance metric** for finding nearest neighbors per class is L_2 (using `faiss` framework) Johnson et al. (2019) at the average pooling of the last conv features of $\mathbf{C}$.

Re-ranking algorithm Given a well-trained comparator $\mathbf{S}$ and the PCNN, we repeat the following for each class among the top-K classes: Multiply each original confidence score in $\mathbf{C}(x)$ by a corresponding score $\mathbf{S}(x, nn)$ where nn is the nearest neighbor from a corresponding predicted class (see Fig. 2). Based on the newly weighted scores $\mathbf{C}(x) \times \mathbf{S}(x, nn)$, we re-rank the top-$K$ labels.

2.3 Training image comparator S

Network architecture Our image comparator $\mathbf{S}$ follows closely the design of the CrossViT Chen et al. (2021), which takes in a pair of images (see Fig. 3). The image patch embeddings are initialized with convolutional features from a pretrained convolutional network.

Objective function We aim to train $\mathbf{S}$ to separate image pairs taken in the same class from those pairs where images are from two different classes. As standard in contrastive learning Chen et al. (2020), we first construct a set of positive pairs and a set of negative pairs from the training set, and then train $\mathbf{S}$ using a binary sigmoid cross-entropy loss. Note that training the comparator also finetunes the pretrained conv layers $\mathbf{f}$, which are part of the comparator model (Fig. 3).

Sampling positive and negative pairs For each training-set example x, we construct a set of positive pairs $\{(x, nn_+)\}$ and negative pairs $\{(x, nn_-)\}$ (Fig. 4). To find *nearest* images, we use the distance metric described in Sec. 2.2.

positive pairs We take Q nearest images nn_+ to the query x from the same class of x (e.g., Elegant Tern in Fig. 4).

negative pairs One can also take nn_- nearest images from the random non-groundtruth classes. However, in the preliminary experiments, we find that taking nn_- from a random class (e.g., among

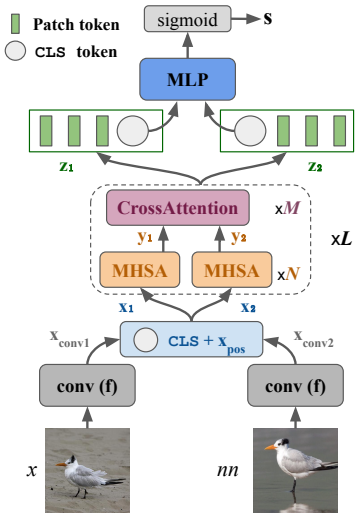


Figure 3: Our comparator takes in a pair of images (x, nn) and outputs a sigmoid score $\mathbf{s} = \mathbf{S}(x, nn) \in [0, 1]$ indicating whether two images belong to the same class. L, M, and N are the depths of the respective blocks.

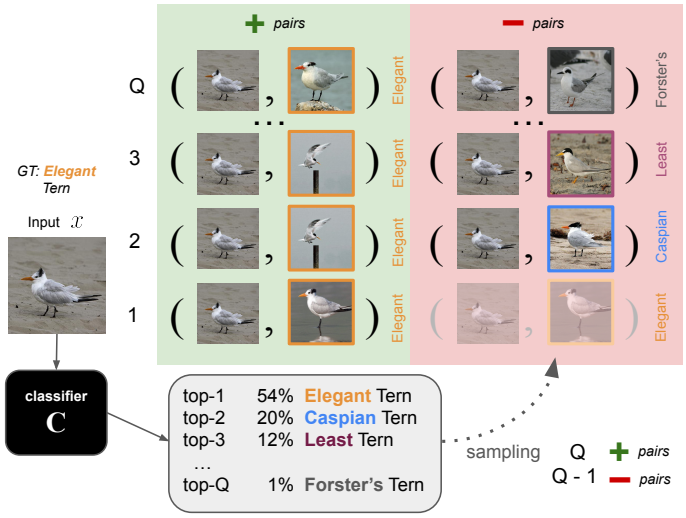


Figure 4: For each training-set image x, we sample Q nearest images from the groundtruth class of x to form Q positive pairs $\{(x, nn_+^i)\}_Q^{i=1}$. To sample Q *hard*, negative pairs: Per non-groundtruth class among the top-Q predicted classes from $\mathbf{C}(x)$, we take the nearest image to the input. Here, when the groundtruth label (Elegant Tern) is among the top-Q labels, there would be only $Q - 1$ negative pairs.

200 bird classes) produces "easy" negative pairs (x, nn_-), i.e., images are often too visually different, not strongly encouraging $\mathbf{S}$ to learn to focus on subtle differences between fine-grained species as effectively as our "hard" negatives sampling.

Sampling using classifier C First, we observe that pretrained classifiers $\mathbf{C}$ often have a very **high top-10 accuracy** (e.g., 98.64% on CUB-200 for ResNet-50) and therefore tend to place species visually similar to the ground-truth class among the top-Q labels. Therefore, we leverage the predictions $\mathbf{C}(x)$ of the classifier $\mathbf{C}$ on x to sample hard negatives. That is, we sample Q nn_- nearest images to the query. Yet, each nn_- is from a class among the top-Q predicted labels for x, i.e., from the $\mathbf{C}(x)$. As illustrated in Fig. 4, if the groundtruth class appears in the top-Q labels, we exclude that corresponding negative pair, arriving at $Q - 1$ negative pairs. In this case, we will remove one positive pair to make the data balance. In sum, if the groundtruth class is in the top-Q labels, we will produce Q positive and $Q - 1$ negative pairs. Otherwise, we would produce Q positive and Q negative pairs. Empirically, we try $Q \in \{3, 5, 10, 15\}$ and find $Q = 10$ to yield the best comparator based on its test-set binary-classification accuracy.

3 Results

In this section, we demonstrate that PCNN examples enhance both AI and human accuracy. First, we use PCNN examples to train an image comparator $\mathbf{S}$, which improves classifier $\mathbf{C}$'s predictions via the re-ranking algorithm in Sec. 2.2. Second, when shown PCNN examples, human users increase their accuracy in distinguishing correct from incorrect predictions for both CUB-200 and Dogs-120.

3.1 $\mathbf{C} \times \mathbf{S}$ re-ranking consistently outperforms classifier $\mathbf{C}$

Here, we aim to test how our re-ranking algorithm (Sec. 2.2) improves upon the original classifiers $\mathbf{C}$.

Experiment For each of the 10 classifiers listed in Sec. 2.1, we train a corresponding comparator $\mathbf{S}$ (following the procedure described in Sec. 2.3) and form a $\mathbf{C} \times \mathbf{S}$ model.

Results Our $\mathbf{C} \times \mathbf{S}$ models outperform classifiers $\mathbf{C}$ consistently across all three architectures (ResNet-18, ResNet-34, and ResNet-50) and all three datasets (see Tab. 1). The largest gains on CUB-200, Cars-196, and Dogs-120 are +11.78, +3.03, and +1.04 percentage points (pp), respectively.

A trend is that when the original classifier $\mathbf{C}$ is weaker, our re-ranking often yields a larger gain. Intuitively, a weaker classifier's predictions benefit more from **revising** based on extra evidence (PCNN) and an external model (comparator $\mathbf{S}$). Yet, on CUB-200, we also improve upon the best

Table 1: On all three ResNet (RN) architectures and three datasets, our $\mathbf{C} \times \mathbf{S}$ consistently improves the top-1 classification accuracy (%) over the original classifiers $\mathbf{C}$ (e.g., by +11.48 on CUB-200) and also a baseline re-ranking $\mathbf{C} \rightarrow \mathbf{S}$ (which uses only $\mathbf{S}$ scores in re-ranking). "Pretraining" column specifies the datasets that $\mathbf{C}$ models were pretrained (before fine-tuning on the target dataset).

Classifier architecture		ResNet-18 (a)			ResNet-34 (b)			ResNet-50 (c)		
Dataset	Pretraining	$\mathbf{C}$	$\mathbf{C} \rightarrow \mathbf{S}$	$\mathbf{C} \times \mathbf{S}$	$\mathbf{C}$	$\mathbf{C} \rightarrow \mathbf{S}$	$\mathbf{C} \times \mathbf{S}$	$\mathbf{C}$	$\mathbf{C} \rightarrow \mathbf{S}$	$\mathbf{C} \times \mathbf{S}$
CUB-200	iNaturalist	n/a	n/a	n/a	n/a	n/a	n/a	85.83	87.72	88.59 (+2.76)
	ImageNet	60.22	66.78	71.09 (+10.87)	62.81	71.92	74.59 (+11.78)	62.98	71.63	74.46 (+11.48)
Cars-196	ImageNet	86.17	85.70	88.27 (+2.10)	82.99	83.57	86.02 (+3.03)	89.73	89.90	91.06 (+1.33)
Dogs-120	ImageNet	78.75	75.34	79.58 (+0.83)	82.58	80.82	83.62 (+1.04)	85.82	83.39	86.31 (+0.49)

model (iNaturalist-pretrained ResNet-50) by +2.76 (85.83% → 88.59%). Note that while the gains on Dogs-120 are modest (Tab. 1), dog images are the noisiest among the three tested image types, and therefore the small but consistent gains on Dogs-120 are valuable.

3.2 PCNN improves human accuracy in predicting AI misclassifications on bird images

Given the effectiveness of PCNN examples for AIs, we are motivated to test them on human users. Specifically, we compare user accuracy in the *distinction* task Kim et al. (2022) (i.e., telling whether a given classifier $\mathbf{C}$ is correct or not) when presented with top-1 class examples as in prior work Nguyen et al. (2021); Taesiri et al. (2022) compared to when presented with PCNN (Fig. 6a vs. b).

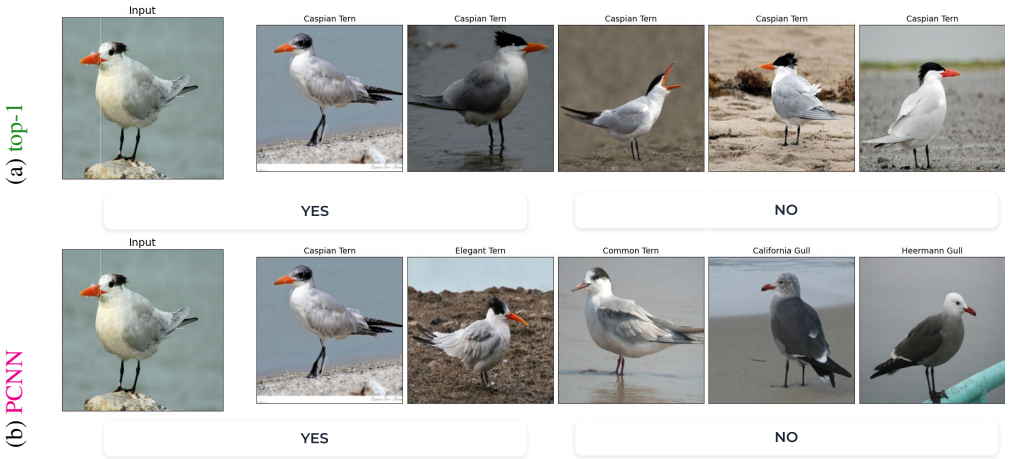


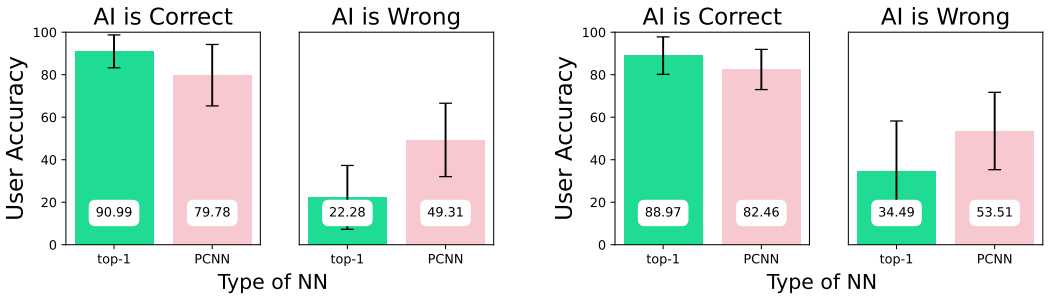
Figure 5: In both experiments, humans are asked whether the input image is Caspian Tern given that input, a model prediction, and either top-1 class examples (top) or PCNN explanations (bottom). When given only examples from the top-1 class, humans tend to accept the prediction, not knowing there are other very visually similar birds. Yet, the top-5 classes provide humans with a broader context which leads to better accuracy (64.58% vs. 54.55%; Sec. 3.2). We provide the interfaces here for public access:(bird and dogs).

Experiment We randomly sample 300 correctly classified and 300 misclassified images by the $\mathbf{C} \times \mathbf{S}$ model from the CUB-200 test set and similarly sample images from the Dogs-120 test set. From our institution, we recruit 33 lay users for the test with top-1 class examples and 27 users for the PCNN test on CUB-200, and 17 users for the top-1 test and 15 users for the PCNN test on Dogs-120. Per test, each participant is given 30 images, one at a time, and asked to predict (Yes or No) whether the top-1 predicted label is correct given the explanations (Fig. 5). To align with prior work, we choose $K = 5$ when implementing PCNN, i.e., we only show nearest examples from top-K (where $K = 5$) classes to keep the explanations readable to users.

Results We find that PCNN offers contrastive evidence for users to distinguish closely similar species, leading to better accuracy. That is, showing only examples from the top-1 class leads users to overly trust model predictions, rejecting only 22.28% (CUB-200) and 34.49% (Dogs-120) of the cases where the model misclassifies. In contrast, PCNN users correctly reject 49.31% (CUB-200) and 53.51% (Dogs-120) of model misclassifications (Fig. 6). Because users are given more contrastive

information to make decisions, they tend to doubt model recommendations more often, resulting in lower accuracy when the model is actually correct (79.78% vs. 90.99% on CUB-200; 82.46% vs. 88.97% on Dogs-120).

Yet, on average, compared to showing only examples from the top-1 class in prior works Nguyen et al. (2021); Taesiri et al. (2022), PCNN improves user accuracy by a large margin of nearly 10 points on CUB-200 (54.55% → 64.58%) and over 5 points on Dogs-120 (63.55% → 69.21%).



(a) CUB-200.

(b) Dogs-120.

Figure 6: Users often accept when top-1 class neighbors are presented, leading to high accuracy when the AI is correct and extremely poor accuracy when AI is wrong. PCNN mitigates this limitation of the top-1 examples by providing contrastive evidence.

4 Discussion and Conclusion

In this work, we propose PCNN, a novel explanation consisting of K images taken from the top-K predicted classes to more fully represent the query. We show that PCNN can be leveraged to improve the accuracy of fine-grained image classifiers without having to re-train them, which is increasingly a common scenario in this foundation model era Bommasani et al. (2021). We also find that showing PCNN also helps humans improve their decision-making accuracy compared to showing only top-1 class examples, which is a common practice in the literature. An important aspect to consider is that while PCNN provides a comprehensive explanation by presenting multiple probable-class examples, it is crucial to ensure this does not reduce user confidence to a degree that negatively affects decision-making.

References

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2021.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Alex J Chan, Alihan Huyuk, and Mihaela van der Schaar. Optimising human-ai collaboration by learning convincing explanations. *arXiv preprint arXiv:2311.07426*, 2023.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 357–366, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *arXiv preprint arXiv:2301.07255*, 2023.

Teodor Chiaburu, Frank Haußer, and Felix Bießmann. Copronn: Concept-based prototypical nearest neighbors for explaining vision models. In *World Conference on Explainable Artificial Intelligence*, pp. 69–91. Springer, 2024.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Eoin M Kenny, Mycal Tucker, and Julie Shah. Towards interpretable deep reinforcement learning with human-friendly prototypes. In *The Eleventh International Conference on Learning Representations*, 2022.

Eoin M Kenny, Eoin Delaney, and Mark T Keane. Advancing post-hoc case-based explanation with feature highlighting. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 427–435, 2023.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2. Citeseer, 2011.

Sunnie SY Kim, Nicole Meister, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Hive: evaluating the human interpretability of visual explanations. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pp. 280–298. Springer, 2022.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Han Liu, Yizhou Tian, Chacha Chen, Shi Feng, Yuxin Chen, and Chenhao Tan. Learning human-compatible representations for case-based decision support. In *The Eleventh International Conference on Learning Representations*, 2022.

Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems*, 34:26422–26436, 2021.

Giang Nguyen, Mohammad Reza Taesiri, Sunnie SY Kim, and Anh Nguyen. Allowing humans to interactively guide machines where to look does not always improve human-ai team's classification accuracy. *arXiv preprint arXiv:2404.05238*, 2024.

Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

Mohammad Reza Taesiri, Giang Nguyen, and Anh Nguyen. Visual correspondence-based explanations improve ai robustness and human-ai team accuracy. *Advances in Neural Information Processing Systems*, 35:34287–34301, 2022.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011.