This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SIC: Similarity-Based Interpretable Image Classification with Neural Networks

Tom Nuno Wolf Emre Kavak Fabian Bongratz Christian Wachinger
Lab for Artificial Intelligence in Medical Imaging, Technical University of Munich, Germany
Munich Center for Machine Learning (MCML), Germany
tn.wolf@tum.de

Abstract

The deployment of deep learning models in critical domains necessitates a balance between high accuracy and interpretability. We introduce SIC, an inherently interpretable neural network that provides local and global explanations of its decision-making process. Leveraging the concept of case-based reasoning, SIC extracts class-representative support vectors from training images, ensuring they capture relevant features while suppressing irrelevant ones. Classification decisions are made by calculating and aggregating similarity scores between these support vectors and the input's latent feature vector. We employ B-Cos transformations, which align model weights with inputs, to yield coherent pixel-level explanations in addition to global explanations of case-based reasoning. We evaluate SIC on three tasks: fine-grained classification on Stanford Dogs and FunnyBirds, multi-label classification on Pascal VOC, and pathology detection on the RSNA dataset. Results indicate that SIC not only achieves competitive accuracy compared to state-of-the-art black-box and inherently interpretable models but also offers insightful explanations verified through practical evaluation on the FunnyBirds benchmark. Our theoretical analysis proves that these explanations fulfill established axioms for explanations. Our findings underscore SIC's potential for applications where understanding model decisions is as critical as the decisions themselves.

1. Introduction

Deep learning is increasingly becoming part of everyday life, e.g., targeted advertisements, gaming, and high-stakes areas like banking and medicine. In the latter, identifying errors in the model's decision-making process is crucial to detect failure cases, which can be addressed by adding transparency to model predictions using explainable AI (XAI). Explanations are typically produced post-hoc for black-box models using approximation methods, which estimate the contribution of each input feature to the prediction [3, 19, 23, 24, 28, 30]. While these methods can accurately compute feature contributions for low-dimensional inputs [19], they introduce approximation errors for high-dimensional inputs such as images [1].

As an alternative, *inherently interpretable models* provide explanations by design. For instance, logistic regression can be considered an inherently interpretable model since evaluating its coefficients directly explains the effect of each input feature towards the prediction. Translating these properties to complex functions implemented by neural networks was further fueled by the argument that if a technique is able to summarize the decision-making of a black-box without any error, the black-box could be replaced by an inherently interpretable model in the first place [26]. As a result, a growing research community is dedicated to developing inherently interpretable models, e.g., for decision-critical tasks like medicine [4, 16, 33].

The explanations provided by these models can be categorized into local and global explanations, and evaluated in terms of axioms to verify if they are faithful [30] [1]. Local explanations focus on individual predictions, revealing how the model arrived at a specific outcome for a particular instance. In contrast, global explanations provide insight into the model's overall behavior. In logistic regression, global explanations are derived from the model's coefficients, which describe how each feature influences the prediction across all instances. To produce local explanations, a sample's specific feature values are multiplied by their corresponding coefficients, illustrating how each feature contributed to the particular prediction.

Related Work As outlined in Tab. 1, *locally* interpretable deep learning models include the work by Wang and Sabuncu [31], who trained a non-parametric classification head with metric learning to classify an image by summing the softmax-normalized similarity scores between training

[1] Throughout this manuscript, we define faithful explanations as explanations that satisfy all of the axioms proposed by Sundararajan et al. [30]: Completeness, Sensitivity, Implementation Invariance, Dummy, Linearity, and Symmetry-Preserving.

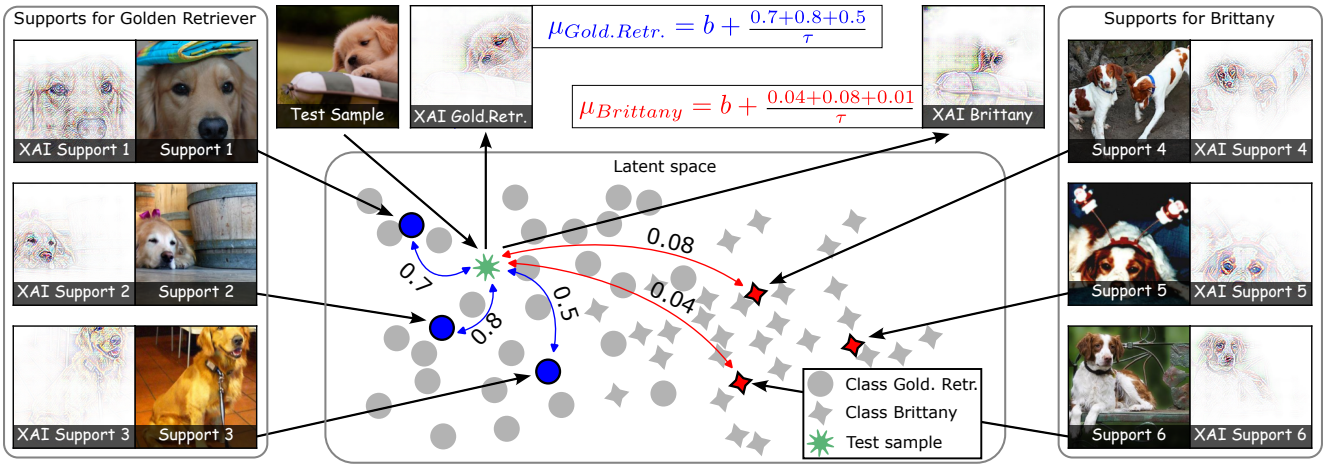


Figure 1. SIC trains a neural network to extract class-representative support feature vectors (red and blue) from training images. It computes class logits μ as the sum of temperature-normalized similarity scores between a class's support feature vectors and a test sample's feature vector (green), as shown for the predicted *Golden Retriever* class and non-predicted *Brittany* class. Its B-cos backbone permits the computation of faithful local and global explanations. Stars and circles denote samples of different classes.

image latent feature vectors and the test image latent feature vector. They proposed to provide local explanations by showing the most influential images alongside their corresponding scores for a given test sample, similar to the k-nearest-neighbor algorithm. While intuitive, the approach lacks contribution maps in pixel space to identify important input features. To this end, BagNet [6] classify local patches of an image individually and sum the probabilities of all patches for the final classification. Providing the class-probability maps of the individual patches' predictions as explanations shows which local structures showed evidence of a respective class. However, they lack faithful fine-grained attributions computed on a pixel-level, which are provided by B-cos networks [5, 7]. In these models, each forward pass is summarized as a linear transformation, and its resulting weight matrix then serves to compute local explanations in image space. However, B-cos models lack global interpretability because the weight matrix is dynamically computed for each input.

Model	Local	Global	Faithful
NW-Head [31]	✓	✓	✗
ProtoPNet [8]	✓	✓	✗
BagNet [6]	✓	✗	✓
B-Cos [5]	✓	✗	(✓)
SIC (ours)	✓	✓	✓

Table 1. Comparison of the ability of inherently interpretable models to provide local, global, and faithful explanations.

Models that provide global interpretability include concept bottleneck models [18], which train an encoder to predict concepts present in an image and a classifier to predict the target class from these concepts, which serve as local explanations during a prediction. However, they require access to concepts, i.e., meta information on each concept present in an image, e.g., the color of the tail, legs, or head of a bird. While the concepts themselves provide global and local explanations and are manually pre-defined, concept bottleneck models lack explanations in the image domain. The most prominent inherently interpretable model is ProtoPNet [8] and its extensions [9, 16], which provide local and global explanations. Its key idea is to learn class-specific prototypical parts across the training set to facilitate reasoning as "this part of an unseen test image looks like that part from my training images." To provide image space explanations, it upsamples distance maps from the latent space into the input domain. However, Hoffmann et al. [13] found that the explanations provided by ProtoPNet can be misleading and called for more rigorous evaluations. Sacha et al. [27] proposed a benchmark to evaluate spatial misalignment of explanations given by ProtoPNet, which was also found by the theoretical evaluation of Wolf et al. [34]. They showed that spatial relationship between the input space and latent space can not be guaranteed for ProtoPNets with deep convolutional neural networks encoders, generalizing to other prototypical networks like Pip-Net [21], ProtoTrees [20], XProtoNet [16], or the Deformable ProtoPNet [9]. While prototypical part-based case-based reasoning has shown to be the most human comprehensible explanation [15, 17, 22], the number of explanations to be evaluated in its current implementations grows with the number of prototypical parts per class learned by the model.

Contribution We propose SIC, a transparent neural network for **s**imilarity-based **i**nterpretable image **c**lassifciation, that fulfills all three properties: local, global, and faithful explanations (see Tab. 1). We inject case-based reasoning into SIC's decision-making to provide local and global explanations that reduce the complexity of explanations to be evaluated. SIC can implement any neural network architecture in its feature extractor without limiting their computational capacity. Specifically, the model extracts support vectors from training images that capture only class-representative features by suppressing negative contributions from other classes. Classification is performed by computing and summing similarity scores between these support vectors and the latent feature vector of the input image (see Fig. 1). Leveraging B-cos networks ensures that both the support vectors and their similarity to test images are directly and faithfully explainable in the input space, enabling an intuitive understanding of the model's reasoning. In summary, the support samples serve two purposes: (i) they help the model capture the high variability between different classes in the data, and (ii) they provide the user with examples that illustrate which parts of the class-representative training images the feature extractor focuses on. Our contributions are as follows:

- We propose SIC, an inherently interpretable network that provides faithful local and global explanations for image classification with either single or multiple labels.
- We prove that explanations provided by SIC fulfill the axioms required to be faithful.
- We practically evaluate SIC's explanations based on the criteria proposed in the FunnyBirds framework and its associated dataset.
- We empirically evaluate SIC and its explanations on three tasks (fine-grained single-label image classification on Stanford Dogs, pathology prediction on RSNA, multi-label image classification on Pascal VOC), and three architectures with different numbers of parameters (DenseNet121: ∼8M, ResNet50: ∼26M, a hybrid vision transformer: ∼81M).
- We demonstrate how the explanations provided by SIC enable model debugging.

2. Background

Learning Class-Discriminative Support Vectors Wang and Sabuncu [31] proposed to classify images with the non-parametric Nadaraya-Watson head and a convolutional neural network $\mathcal{F}_\theta$ with parameters θ extracting latent vector of dimension d. During training, they randomly sample a number of support vectors v_k^c for every batch per class from the training set $\{(\mathcal{I}_k, y_k)\}_{k=1}^N$, $V^c = \{v_k^c = \mathcal{F}_\theta(\mathcal{I}_k) \mid y_k = c\}$, with $\Omega(V^c) = N_s$. This selection is executed for every training batch, thereby directing the optimization of the feature extractor $\mathcal{F}_\theta$ to form distinct, class-specific clusters within the latent feature space.

For classification, a test image $\mathcal{I}$ is first transformed into its latent representation $\mathcal{F}_\theta(\mathcal{I})$. The probability that $\mathcal{I}$ belongs to class c is computed by applying the softmax function to the negative squared L2 distances between $\mathcal{F}_\theta(\mathcal{I})$ and each support vector v_k^c:

$$p(y = c \mid \mathcal{I}) = \sum_k \frac{\exp\left(-\|\mathcal{F}_\theta(\mathcal{I}) - v_k^c\|_2^2\right)}{\sum_{k'=1}^{N_s}\sum_{c'=1}^{C} \exp\left(-\left\|\mathcal{F}_\theta(\mathcal{I}) - v_{k'}^{c'}\right\|_2^2\right)}.$$

This classification strategy leverages the learned latent space structure to assign class probabilities based on proximity to the nearest support clusters.

Local explanations show the images of the most influential support samples in terms of softmax probabilities. At first glance, these explanations seem intuitive, but the main limitations are two-fold: (i) Computing probabilities from (inverted) Euclidean distances with softmax: Regardless of the input (e.g., in an adversarial attack), the model explanations will suggest that at least one support sample is similar with sufficient probability. Additionally, the softmax probabilities are inapplicable to multi-label classification tasks out of the box since they sum to one. Thus, the model cannot predict three classes simultaneously under a standard decision threshold of 0.5. (ii) No pixel-level explanations: While explanations using similar input images are intuitive for humans, the lack of attribution maps hinders a further understanding of the network's mechanisms and the detection of biased predictions.

Faithful Explanations with B-cos Networks B-cos networks proposed by Böhle et al. [5] impose weight-input alignment between the layer weights and input by reformulating the scalar product used by neural network computations, removing biases and standard non-linearities. This forces the network to focus on the most salient features across the training set and facilitates the summary of each forward pass as an input-dependent linear equation. Each input feature's contribution can directly be traced by its corresponding weight in the corresponding transformation matrix.

Mathematically, the B-cos transformation computes a modified dot-product, that increases weight-input alignment with an exponent B, between an input vector x and a layer weight w:

$$\text{B-cos}(x; w) = \|\hat{w}\|\|x\||\cos(x, \hat{w})|^B \times \text{sgn}(\cos(x, \hat{w})),$$

with $\hat{w} = w/\|w\| \implies \|\hat{w}\| = 1$, cos the cosine similarity, sgn the sign function, and $\times$ the real-valued multiplication. While it is bounded by the magnitude of $\|x\|$, it effectively computes an input-dependent linear equation, in

matrix form expressed by B-cos$(x, \mathbf{W}) = \tilde{\mathbf{W}}(x)x$, with $\tilde{\mathbf{W}}(x) = |\cos(x, \hat{\mathbf{W}})|^{B-1} \odot \hat{\mathbf{W}}$, where $|\cos(x, \hat{\mathbf{W}})|^{B-1}$ scales the elements of the fixed weight matrix $\hat{\mathbf{W}}$ ($\odot$ the elementwise multiplication). A neural network $\mathcal{F}_\theta$ with parameters θ that computes a series of these equations is called piece-wise linear and its forward pass is summarized by the single input-dependent weight matrix:

$$\begin{aligned}\mathcal{F}_\theta(x) &= \tilde{\mathbf{W}}_L(a_L)\tilde{\mathbf{W}}_{L-1}(a_{L-1})\dots\tilde{\mathbf{W}}_1(a_1 = x)x \\ &= \left(\prod_{j=1}^{L}\tilde{\mathbf{W}}_j(a_j)\right)x = \mathbf{W}_{1\to L}(x)x. \qquad (1)\end{aligned}$$

Thus, the contribution maps in terms of a pixel location (m, n) in image space for a single forward pass are computed with $\phi_j^l(x)_{(m,n)} = \sum_{ch}\left[[\mathbf{W}_{1\to l}]_j^T \odot x\right]_{(ch,m,n)}$, with l any layer and j the index of the row (neuron) in $\mathbf{W}_{1\to l}$ (note that this allows to explain any intermediate layer's l neurons j). Since B-cos networks are trained with binary cross-entropy loss, features negatively attributing the class logit cannot increase its probability and are not visualized.

Böhle et al. [5] proposed to compute RGBA explanations by normalizing each color channel pair[2] to sum to 1 to maintain the angle of each color pair and to then scale RGBA values into the range [0..1]. The alpha value (pixel opacity) is computed with the 99.9th percentile $p_{99.9}$ of the resulting weight w as $\min\left(\left[\|w_{(m,n)}\|_2\right] / \left[p_{99.9}(\|w_{(m,n)}\|_2)\right], 1\right)$ and smoothed with an average kernel of size nine.

B-cos models offer the notable advantage of generating local explanations from a single forward pass that resembles the network's transformation of the input image. However, whether B-cos meets the axioms of faithful explanations remains unclear. In Appendix A.1, we prove that B-cos satisfies these necessary axioms.

3. Methods

We propose SIC, which combines the strengths of B-cos networks and the Nadaraya-Watson head, to obtain an *inherently interpretable* neural network for similarity-based classification that provides *faithful local* and *global* explanations with a focus on minimizing the number of explanations requiring review. SIC achieves this by extracting class-representative support vectors from training images for similarity-based classification and by leveraging the B-cos transform to faithfully explain its decision. Our implementation is available at `www.github.com/ai-med/SIC`.

3.1. Architecture

We visualize SIC in Fig. 2. It consists of a B-cos feature extractor $\mathcal{F}_\theta$, which transforms an image $\mathcal{I}$ into latent vectors $f = \mathcal{F}_\theta(\mathcal{I}) \in \mathbb{R}^d$, with $\mathcal{I} \in \mathbb{R}^{ch\times H\times W}$, d the latent dimension, ch, H, and W the image's channels, height, and width, respectively, and θ the model weights ($\mathcal{F}_\theta$ can be any B-cos neural network architecture).

The evidence predictor $\mathcal{E}$ consists of a function $\oplus : \mathbb{R}^d \to \mathbb{R}^d_{\geq 0}$ that transforms the real-valued vectors f into non-negative vectors $f^+ = \oplus(f)$, and a similarity measure that computes similarities between support vectors $v_i^c \in F^c = \{f_k^+ = \mathcal{F}_\theta(\mathcal{I}_k) \mid y_k = c\}$ and f^+ as $sim(f^+, v_i^c) \in \mathbb{R}$, with $\Omega(F^c) = N_s$ and i denoting the i-th support vector of class c. Logits are computed for class c as the sum of its similarity scores $\mu_c = b + \sum_{v_i^c}\frac{sim(f^+, v_i^c)}{\mathcal{T}}$, where b represents a fixed bias term, $i = 1, \dots, N_s$ indexes the support samples, and $\mathcal{T}$ is a temperature to improve convergence by scaling the logits' magnitude appropriately.

3.2. Training and Optimization

Following Wang and Sabuncu [31] ("closest cluster"), we randomly sample the training set to extract N_s support vectors v_k^c per class to encourage the model to form distinct, class-specific clusters in latent space. After every epoch, we compute sets of latent vectors $\hat{V}^c = \{f_k^+ = \mathcal{F}_\theta(\mathcal{I}_k) \mid y_k = c\}_{k=1}^N$, which we each cluster using k-means to extract centroids γ_i^c. Then, we replace each centroid γ_i^c with the closest latent vector to yield support vectors from the same class, defined as $v_i^c = \arg\min_{f_k^+ | y_k = c}\|f_k^+ - \gamma_i^c\|_2$, which facilitates case-based reasoning similar to k-nearest-neighbors, as only features descriptive of the class can increase the probability of that class by definition of SIC.

We follow Böhle et al. [7] and optimize the model with binary cross-entropy loss, which effectively increases alignment pressure compared to training with standard multi-class cross-entropy loss.

3.3. Explainability

Global Explanations While global explanations provided by SIC are ideal for a developer to debug a model, they also facilitate case-based reasoning, which is most intuitive to understand for humans [15, 17, 22].

First, we evaluate what is encoded in each support vector v_i^c by computing explanations given the corresponding support image and its extracted feature vector. This feature vector equals v_i^c before its rescaling to the unit norm, simplifying the B-cos transformation to B-cos$(v_i^c, \hat{v}_i^c) = \|v_i^c\|$, since $\hat{v}_i^c = v_i^c / \|v_i^c\|$ and $\|\hat{v}_i^c\| = 1$ and $\cos(v_i^c, \hat{v}_i^c)^B = 1$ for any B.

[2] B-cos networks require an image encoding of [R, G, B, 1-R, 1-G, 1-B] in the channel dimension to uniquely encode colors and to mitigate favoring of brighter regions.

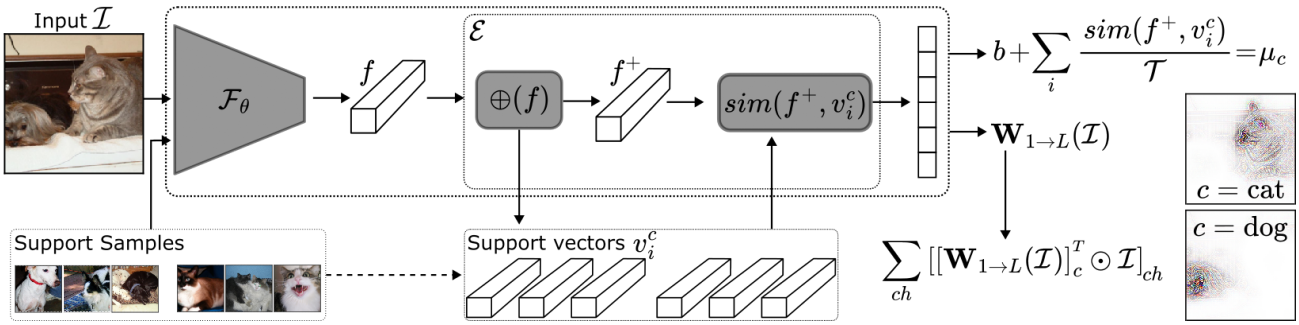


Figure 2. SIC consists of a feature extractor $\mathcal{F}$ with model weights θ that extracts latent vectors f. The evidence predictor $\mathcal{E}$ converts them into positive-valued vectors f^+, which, given other images from the training set, can also serve as support vectors v_i^c. Then, it computes the similarity between the input image's latent vector f and support vectors v_i^c. Summing the class-specific temperature-normalized similarity scores and adding a bias b yields the class logit μ_c. The B-cos transform facilitates the summary of a forward pass as a single weight-matrix $\mathbf{W}_{1\to L}(\mathcal{I})$, which is leveraged to compute faithful contribution maps.

We observed negative attributions when computing contribution maps this way during initial experiments. While only visualizing features with positive class contributions as explanations in RGBA space is reasonable during the classification of an input image, these negative attributions can still serve as a means of debugging a model if carefully evaluated. The definition B-cos$(v_i^c, \hat{v}_i^c) = \|v_i^c\|$ implies that negative attributions are a means of the model to downscale the range of the B-cos similarity (accordingly, we observed the negative attributions in the corners of the images). This indicates the temperature scaling τ was chosen too small, and re-training with a more significant temperature helped to overcome this issue.

Additionally, SIC's decision-making is based on a fixed set of latent support vectors. Hence, we can compute the inter- and intra-class similarities to gain insights into the latent space learned by the feature extractor.

Local Explanations Local explanations are threefold: First, the temperature-scaled similarity scores $\frac{sim(f^+, v_i^c)}{\mathcal{T}}$ (support evidence) provided by the evidence predictor $\mathcal{E}$ quantify the alignment of a test image with a support sample, and their proportion of the log-probability. Second, since a test image is assigned to a class if it is similar to that class's support sample, understanding which parts of the support sample's image are compressed in the support vectors themselves is crucial, i.e., explaining the model mechanisms. The B-cos properties of SIC allow us to faithfully compute the RGBA explanations in terms of the similarity measure. Since each prediction can be summarized by a linear equation, we compute contribution maps in terms of the class logit (test contribution) following the linearity axiom [30] ($\phi = \alpha\phi_1 + \beta\phi_2$ iff $\mathcal{F}(x) = \alpha\mathcal{F}_1(x) + \beta\mathcal{F}_2(x)$). Therefore, the test sample contributions highlight the image features that aligned with the support vectors. Third, the global RGBA explanations (support contributions) of the support samples allow the user to check for the intersection of support contributions and test contributions, similar to k-nearest-neighbor.

Theoretic Evaluation To thoroughly assess the interpretability of SIC, we conduct a theoretical evaluation of its explanations based on the axioms defined by Sundararajan et al. [30], namely *Completeness, Sensitivity, Implementation Invariance, Dummy, Linearity,* and *Symmetry-Preserving*. Our proofs (see App. A.1), confirm that SIC's explanations adhere to these principles, thereby validating the model's capability to provide faithful and comprehensive insights into its decision-making processes. Thus, SIC solves the limitations of previous work, which either does not provide pixel-level explanations, like the NW-Head [31] or lack global explanations like B-cos neural networks [7].

4. Experiments and Results

We evaluate SIC on three tasks with three backbone architectures as feature extractors. The tasks are (i) multi-label image classification on Pascal VOC [10], (ii) fine-grained image classification with mutually exclusive labels on Stanford Dogs [2], and (iii) pathology prediction with mutually exclusive labels on RSNA [29]. Additionally, we evaluate the explanations given by SIC on the FunnyBirds framework [12]. The backbone architectures are B-cos versions of DenseNet121 [14], ResNet50 [11], and a Hybrid Vision Transformer [35] (VitC), with ∼8M, ∼26M, and ∼81M trainable parameters, respectively.

First, we compare the performance of SIC in terms of the accuracy (balanced accuracy on RSNA) to black-box neural networks and established inherently interpretable neural networks, namely ProtoPNet [8], a BagNet17 [6], the NW-Head [31], and B-Cos neural networks [7] in Fig. 3 and Tab. A.2. Compared to black-box versions of the backbones, SIC performs on par on Pascal VOC and RSNA

($\geq +0.29\%$ and -0.65% Acc, respectively), and worse on Stanford Dogs (-3.7% Acc). Overall, ProtoPNet and BagNets perform worse than all competing methods, and SIC performs better than nine out of 14 models.

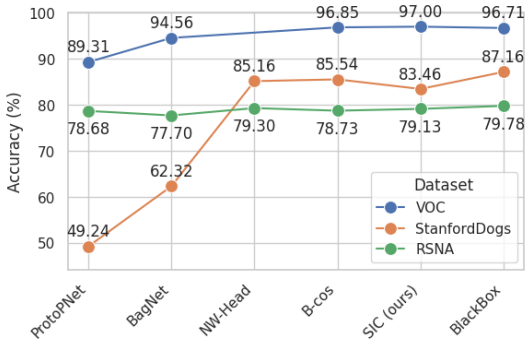
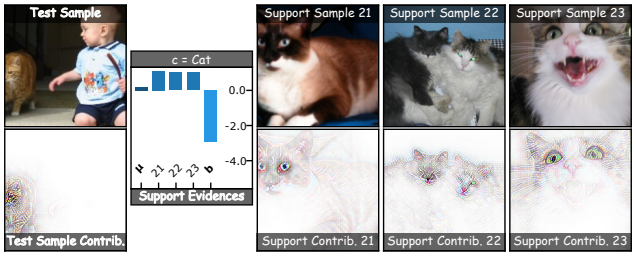


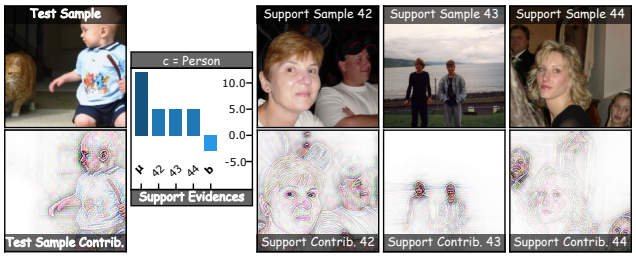
Figure 3. Results in terms of accuracy for ResNet backbones on Pascal VOC and Stanford Dogs, and DenseNet backbones on RSNA. SIC performs superior to nine out of 14 models. Refer to Tab. A.2) for exhaustive results.

Pascal VOC [10] Fig. 4 illustrates explanations of SIC for multi-label image classification on Pascal VOC with a DenseNet121 backbone. As seen in Figs. 4 (a) and (b), SIC shifts its focus in the test image depending on the class to predict, i.e., the corresponding column c in $[\mathbf{W}_{1\to L}]_c^T$: To predict the cat class, only the area around the cat is present in the contribution map. When predicting the human class, the complementary area of the image is contained in the contribution map. The support contributions show that the model is particularly sensitive to the eyes of a cat. In the test sample in Fig. 4 (a) and (b), only the left half and one eye of the cat are visible, which explains the comparably low support evidences in this particular instance. Arguably, the support contributions of the human class in Fig. 4 (b) suggest that the model learned to look for human eyes in support sample 42, which shows high contributions from the two lights on the ceiling. For support samples 43 and 44, the whole heads of the humans are the primary contributions to the support vectors.

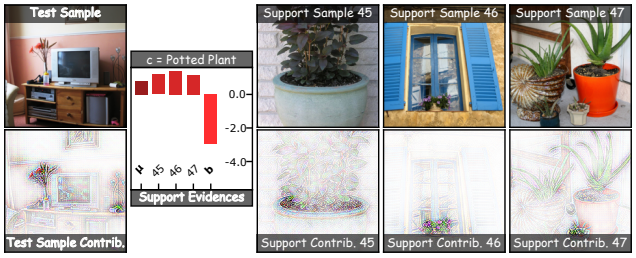
Observing the wrong prediction presented in Fig. 4 (c) for the class potted plant, we see that the support contribution map of support sample 45 focuses on the soil in the pot, while the support samples 46 and 47 focus on the actual plants themselves. We see in the contribution map of the test sample that the model mistakes the flower on the left-hand side and the artificial flowers on the right-hand side with this class. The model correctly classified the TV/monitor class for the same sample, presented in Fig 4 (d), in which the test contribution map primarily highlights the TV. Here, the support contributions of sample 57 show that the support vector carries information on all three monitors present in the image. In the case of support sample 58, the support contribution shows high contributions from the keyboard in addition to the monitor, indicating that the model picked up this bias from the data. Support contribution 59 focuses on the TV instead of background objects.



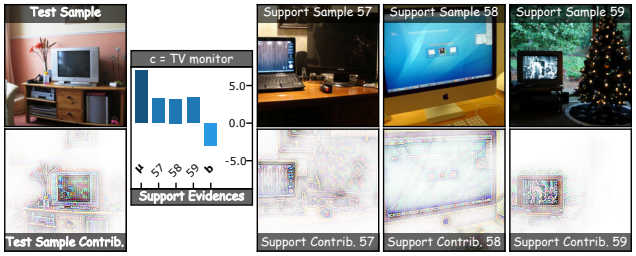
(a) Prediction for the cat class.



(b) Prediction of the person class.



(c) A wrong prediction: The model mistook the flower with potted plants.



(d) However, the model correctly identified the TV.

Figure 4. Correct classifications are depicted with blue support evidence bars, incorrect ones with red. The model trained on Pascal VOC with a DenseNet121 backbone barely classified the test sample in (a) as a cat, because all support images suggest that the model learned to look for two cat eyes. However, the test sample only contains one eye of the cat, and its support contributions suggest that the model is susceptible to detecting two eyes. Computing the contributions for the person class in (b) shows how the model utilizes the other portion of the input features. This shift is also present in the example (c) and (d).

We present additional results for other test images and backbones in Figs. A.8-A.10. In many test sample instances (e.g., compare Fig. A.8 (a) vs. (b), Fig. A.8 (c) vs. (d), Fig.A.10 (a) vs. (b)), we observe a strong shift in contributions depending on the class prediction. This demonstrates that the latent vectors of samples with multiple labels efficiently encode features for all classes that need to align with the class support vectors. In Fig. A.9 (a)-(d), we see that the model predicted four classes, for which, according to the dataset labels, only a single class was actually contained (chair). However, the explanations look very reasonable. Indeed, there is a table in the scene; whether it is a dining table is obviously debatable. There is also a child in the image, so predicting humans is not unreasonable. Lastly, the potted plant class logit is very low, and the test contribution map highlights the flower on the table.

For the test sample predicted in Fig. A.10 (d) and (e), we see the correct classification of humans with high support evidence scores (note how the model extracted the human for support sample 44 in very dark lighting conditions), and SIC believed that the martial arts club logo on the person's outfit resembled the logo of a drink.

Stanford Dogs [2] We evaluate SIC with a ResNet50 backbone on the Stanford dogs dataset. The contribution maps for the test samples in Figs. A.11, A.12,and A.13 show that SIC consistently focuses on the dogs and discards background elements such as humans, especially in Figs. A.11 (a) and (b).

The wrong classification of the test sample presented in Fig. A.11 (c) shows that primarily the head of the dog contributed to an increased similarity with the support vector of support sample 46, whose support contribution focuses on the head, too. In contrast, the support contributions of support samples 45 and 47 both indicate that the support vectors contain significant portions of the body of the dog. Evaluating the explanations of the correct class in Fig. A.11 (d) shows that the support sample 33 containing two dogs has high support evidence, as there are two dogs of its class in the test image. However, the lower support evidence of 34 and 35, which both contain one dog, indicate that the number of instances is encoded into support vector 33.

Considering the support evidence of support sample 162 in Figs. A.13 (a) and (b), its support evidence for the test sample in Fig. A.13 (a) is lower compared to the test sample in Fig. A.13 (b), which only contains the lower portion of the head like support sample 162. The support contributions of the other support samples that focus on the eyes and the body show the inverse behavior of the two respective test samples. However, we observe alike support evidence for all support samples for other classes, e.g., Fig. A.12 (a) and (b). This exemplifies that SIC is able to capture the distribution of more challenging classes.

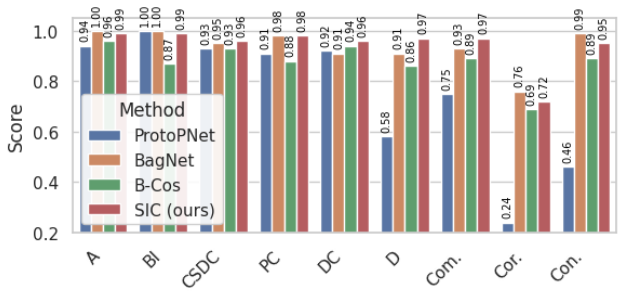


Figure 5. Results on the FunnyBirds Framework [12]. Note that *Completeness* (Com.) is the mean across Controlled Synthetic Data Check (CSDC), Preservation Check (PC), Deletion Check (DC), and Distractability (D); Accuracy (A), Background Independence (BI), Correctness(Cor.), Contrastivity (Con.). Refer to Tab. A.4 for full results.

RSNA [29] We utilize the RSNA dataset to showcase SIC on a decision-critical task to differentiate between images with lung opacity and others (healthy and unhealthy without lung opacity, indicated by the presence of bounding boxes) from chest X-ray images. Here, we want to verify reasonable decision-making by evaluating explanations in terms of bounding boxes. An explanation of a classifier trained to predict lung opacity should focus on an area inside the bounding box, illustrated in Figs. A.14 and A.15.

Here, the model has to learn very subtle intra- and inter-class differences. Evaluating the prototypes in Fig. A.16 shows only positive contributions as desired. In the case of the three non-occluded support samples on the left, we immediately see that the sample is likely an outlier in that it has some disease without opacity. The second and third support contributions are primarily highlighted within the lungs and bounding boxes for the respective classes. Interestingly, the model seems to exclude atypical objects from the healthy support vectors: Consider the third non-occluded support contribution that does not contain contributions for the dark clamp present in the support sample. However, the first two support contributions of each class contain most of the input and are, thus, hard to interpret.

The predictions in Figs. A.14 (a)-(d) exemplify how the support evidence varies across the class's support samples and suggest that the increasing number of support samples allowed SIC to reproduce the class's latent space with a likely high variability. Additionally, we observe a high degree of intersection between significant test sample contributions and bounding boxes for the occluded class in Figs. A.14 (a)-(d) and that the focus of the model is within the lungs for non-occluded samples in Fig. A.15, indicating that the model indeed learned to differentiate occlusion from no-occlusion using medically reasonable regions.

FunnyBirds [12] This framework contains a dataset of synthetic renderings of 50 bird species that are defined by the five concepts beak, wings, feet, eyes, and tail. Removing or replacing a concept with one of another class enables analysis of explanations in terms of *completeness*, *background independence*, *contrastivitiy*, and *correctness*.

The results, presented in Figs. 5 and A.6, and Tab. A.4, show that SIC outperforms ProtoPNet [8] and B-cos [5] in eight out of nine explainability metrics. However, BagNet [6] has a lead in terms of *accuracy* (+1%), *correctness* (+4%), and *contrastivity* (+4%). The explanations for the representative test sample in Fig. A.7 show that SIC extracted a diverse range of poses with each support sample offering a distinct view of the relevant concepts. Moreover, SIC demonstrates near-perfect resilience to background noise, evidenced by a *background independence* of 0.99 and the support contributions shown in Fig. A.7. A *contrastivity* of 0.95 further confirms that only class-discriminative features were learned, which aligns with our model design intent. *Correctness* reflects the assumption that iterative removal of the most important parts from a test image produces a corresponding decline in the classification probability. SIC achieves 0.72 *correctness*, which is in line with the scores (< 0.80) reported by Hesse et al. [12]. A potential explanation is that the part importance is computed by summing attributions scores within a part without normalization for the part's size. Finally, a *completeness* of 0.97 suggests that SIC encoded all class-representative features within its support vectors, as outlined in the support contributions in Fig. A.7.

Debugging SIC While the experiments primarily highlighted the explainability features from the view of a user, we now exemplify how the global and local explanations help to identify flaws in the optimization process. Revisiting the model trained on RSNA, we detail how to use the global explanations of SIC to gain more insights into the model behavior. We observe that the support evidence of support sample 0 of the correct classifications in Figs. A.15 (a) and (b) is almost absent compared to the wrong classification in Fig. A.15 (c). Considering that support contribution 0 contains large parts of the input, this raises suspicion. In addition, we see that the only significant evidence for the correct class depicted in Fig. A.15 (d) comes from support sample 0, which suffers from contributions across the whole input as well. As a result, we evaluate the latent space learned by SIC in Fig. A.17 and find that support vectors 0 and 3, although of different classes, are very close in the t-SNE projection. We attribute this to a weak latent representation, suggesting that the model has yet to learn a more class-separable latent space.

Next, we evaluate whether the model indeed learned class-representative support vectors (see Fig. A.18) by computing the Silhouette score (0.655) [25] and plotting the inter- and intra-class B-cos similarities of SIC trained on Pascal VOC with a DenseNet121 backbone in Fig. A.19. We observe that the intra-class similarities on the diagonal are indeed higher than those off-diagonal, suggesting good class distinction in latent space. Additionally, the average distance between some classes is a bit higher, e.g., dining table, chair, and bottle. This is expected, as images containing a dining table often include a chair and a bottle, requiring the model to extract a latent vector capable of predicting all three classes with sufficient confidence.

5. Discussion and Conclusion

Current inherently interpretable models either lack global explanations, fail to explain the internal mechanisms of the deep learning model, or cannot provide explanations that faithfully explain their decision-making. To this end, we proposed SIC for similarity-based interpretable classification by providing faithful local, global, and pixel-level explanations. Our theoretical analysis proved that our proposed method's explanations satisfy established axioms for explanations, namely *Completeness*, *Sensitivity*, *Impementation Invariance*, *Dummy*, *Linearity*, and *Symmetry-Preserving*.

We evaluated our method, SIC, across three tasks using three distinct backbone architectures, demonstrating that the complex mechanisms learned by neural networks can be decomposed into intuitive, low-complexity explanations. Our practical evaluation employing the FunnyBirds framework further illustrated that the careful design of SIC manifests in its explanations. Notably, our method outperformed established inherently interpretable approaches on the majority of evaluation metrics, while its classification accuracy remained comparable to that of black-box models in two out of three real-world datasets.

Our results suggest that its intuitive global and local explanations may not only facilitate effective model debugging for developers but may also simplify the understanding of its decision-making process for end-users.

While k-means clustering is fast to optimize, SIC introduces additional overhead by requiring the computation of support vectors, which is a potential limitation in cases with many classes. However, as demonstrated with the Stanford Dogs dataset comprising 120 classes, SIC is readily applicable to datasets with hundreds of classes. Moreover, we note that the number of classes in decision-critical fields such as medicine is generally smaller.

In summary, our results demonstrated that embedding interpretability within complex deep neural networks not only achieves competitive performance on challenging benchmarks but also paves the way for more transparent and trustworthy computer vision systems.

Acknowledgments

This research was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). The authors gratefully acknowledge the computational and data resources provided by the Leibniz Supercomputing Centre (www.lrz.de). We thank Sofiène Boutaj for outstanding efforts in experimenting with B-cos models on medical data.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 1

[2] Khosla Aditya, Jayadevaprakash Nityananda, Yao Bangpeng, and Fei-Fei Li. Novel dataset for fine-grained image categorization. In *Proceedings of the First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Springs, USA*, 2011. 5, 7

[3] 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. 1

[4] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. Interpretable mammographic image classification using case-based reasoning and deep learning. *arXiv preprint arXiv:2107.05605*, 2021. 1

[5] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10329–10338, 2022. 2, 3, 4, 8, 13, 15

[6] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019. 2, 5, 8, 13, 15

[7] Moritz Böhle, Navdeeppal Singh, Mario Fritz, and Bernt Schiele. B-cos alignment for inherently interpretable cnns and vision transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4504–4518, 2024. 2, 4, 5

[8] 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, 5, 8, 13, 14, 15

[9] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable protopnet: An interpretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10265–10275, 2022. 2, 13

[10] M. Everingham, L. Gool, C. K. Williams, J. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2009. 5, 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[12] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3981–3991, 2023. 5, 7, 8, 13, 15

[13] Adrian Hoffmann, Claudio Fanconi, Rahul Rade, and Jonas Kohler. This looks like that... does it? shortcomings of latent space prototype interpretability in deep networks. *arXiv preprint arXiv:2105.02968*, 2021. 2

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[15] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, 33:4211–4222, 2020. 2, 4

[16] Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. Xprotonet: diagnosis in chest radiography with global and local explanations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15719–15728, 2021. 1, 2

[17] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. "help me help the ai": Understanding how explainability can support human-ai interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023. 2, 4

[18] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020. 2

[19] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 1

[20] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14933–14943, 2021. 2

[21] Meike Nauta, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2744–2753, 2023. 2

[22] 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. 2, 4

[23] V Petsiuk. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 1

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 1

[25] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 8

[26] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*, 1(5):206–215, 2019. 1

[27] Mikołaj Sacha, Bartosz Jura, Dawid Rymarczyk, Łukasz Struski, Jacek Tabor, and Bartosz Zieliński. Interpretability benchmark for evaluating spatial misalignment of prototypical parts explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21563–21573, 2024. 2

[28] 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 *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

[29] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1): e180041, 2019. 5, 7

[30] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1, 5, 11

[31] Alan Q Wang and Mert R Sabuncu. A flexible nadaraya-watson head can offer explainable and calibrated classification. *arXiv preprint arXiv:2212.03411*, 2022. 1, 2, 3, 4, 5, 13

[32] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 14

[33] Tom Nuno Wolf, Sebastian Pölsterl, and Christian Wachinger. Don't panic: Prototypical additive neural network for interpretable classification of alzheimer's disease. In *Information Processing in Medical Imaging–28th International Conference, IPMI 2023, San Carlos de Bariloche, Argentina, June 18–23, 2023, Proceedings*. Springer, 2023. 1

[34] Tom Nuno Wolf, Fabian Bongratz, Anne-Marie Rickmann, Sebastian Pölsterl, and Christian Wachinger. Keep the faith: Faithful explanations in convolutional neural networks for case-based reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5921–5929, 2024. 2

[35] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021. 5