Pairwise Matching of Intermediate Representations for Fine-grained Explainability

Anonymous authors
Paper under double-blind review

Abstract

The differences between images belonging to fine-grained categories are often subtle and highly localized, and existing explainability techniques for deep learning models are often too diffuse to provide useful and interpretable explanations. We propose a new explainability method (PAIR-X) that leverages both intermediate model activations and backpropagated relevance scores to generate fine-grained, highly-localized pairwise visual explanations. We use animal and building re-identification (re-ID) as a primary case study of our method, and we demonstrate qualitatively improved results over a diverse set of explainability baselines on 35 public re-ID datasets. In interviews, animal re-ID experts found PAIR-X to be a meaningful improvement over existing baselines for deep model explainability, and suggested that its visualizations would be directly applicable to their work. We also propose a novel quantitative evaluation metric for our method, and demonstrate that PAIR-X visualizations appear more plausible for correct image matches than incorrect ones even when the model similarity score for the pairs is the same. By improving interpretability, PAIR-X enables humans to better distinguish correct and incorrect matches.[1]

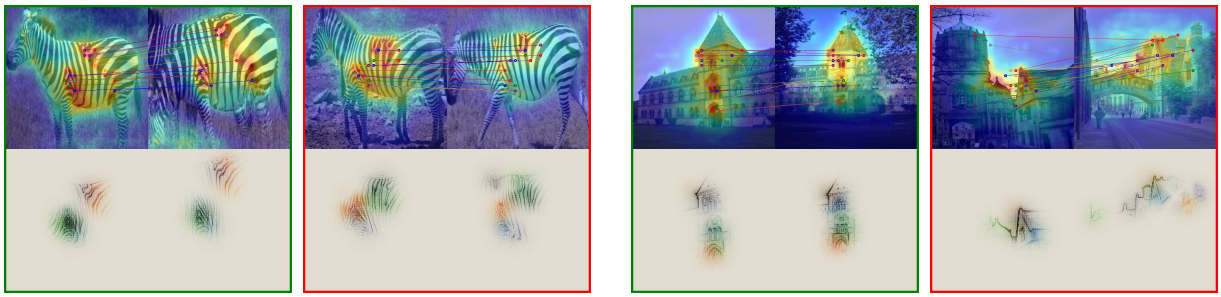
1 Introduction

Similarity-based deep metric learning has proven to be highly effective for a variety of tasks, particularly fine-grained tasks including image retrieval (Wang et al., 2014), facial recognition (Hu et al., 2014; Schroff et al., 2015), and open-set categorization problems such as animal re-identification (re-ID) (Čermák et al., 2023; Schneider et al., 2018; Haurum et al., 2020; Andrew et al., 2021). However, for robust, trustworthy deployment of these systems, interpretability is key. Existing explainability techniques are often insufficient, producing coarse visualizations that do not adequately capture the fine-grained details important to many tasks (Achtibat et al., 2023). As shown in Figure 2, this makes it difficult to precisely interpret which factors contribute most to the predicted similarity between a given pair of images (Zhu et al., 2021).

One application where explainability for deep metric learning models on fine-grained images is *necessary* for trustworthy deployment is animal re-identification (re-ID)—the task of distinguishing individual members of a species. Animal re-ID is a crucial tool in ecology and conservation, serving as the foundation for key applications such as monitoring population trends and analyzing both individual and collective behaviors (Tan et al., 2022). It is an active area of machine learning research, and recent years have seen steady growth in the performance of deep vision models on this task (Schneider et al., 2018). Animal re-ID typically relies on subtle, highly-localized fine-grained features such as subtle variations in spot or stripe patterns, or contours of fins, flukes, or ears. In contrast, explainability techniques like Grad-CAM frequently highlight broad image regions relevant to identification across all individuals (*e.g.* highlighting the entire side of a giraffe, as in Figure 2), but fail to capture localized details that vary between individuals (*e.g.* the placement of specific patterns on the giraffe).

The demand for better explainability techniques for animal re-ID is driven by an ongoing shift from classical re-ID techniques, such as HotSpotter or CurvRank (Crall et al., 2013; Weideman et al., 2020), towards deep models that improve predictive accuracy and scalability to large datasets (Otarashvili et al., 2024; Čermák

[1] Our code is available at: `https://github.com/pairx-explains/pairx`



(a) Closest correct match (b) Closest incorrect match (c) Closest correct match (d) Closest incorrect match

Figure 1: **PAIR-X provides interpretable, fine-grained, and highly-localized explainations which enable both correct and incorrect matches to be quickly identified.** The top half of each explanation shows pairwise-matched high-contribution deep features, and the bottom half shows a color-coded backpropagation to the original image pixels, highlighting plausible or implausible orientation shifts between fine-grained features.

et al., 2023; Schneider et al., 2018; Haurum et al., 2020; Andrew et al., 2021). Classical techniques for re-ID typically rely on algorithmic matching of localized features (*e.g.* SIFT (Lowe, 2004)), and are inherently explainable, as the local feature matches used for identification can be explicitly visualized. The resulting explanations can facilitate efficient manual review of model predictions, which is particularly necessary for high-profile applications such as population estimation for endangered species using visual mark-recapture, where incorrect labels can lead to significant errors in population size estimates (Stevick et al., 2001). In interviews with giraffe re-ID experts, we found that the shift from classical techniques to deep models had resulted in a significant loss of explainability, which made it more challenging and thus slower for experts to verify predictions.

To address this gap, we propose **P**airwise m**A**tching of **I**ntermediate **R**epresentations for e**X**plainability (**PAIR-X**), a post-processing method that produces fine-grained visual explanations for deep models that mimic and extend those of classical techniques, without retraining or architectural changes (see Figure 1). PAIR-X combines techniques from classic local feature matching with insights from modern explainability techniques for deep models (Bach et al., 2015; Zhu et al., 2021; Achtibat et al., 2023; Amir et al., 2022), and produces visualizations with the following key qualities:

- **Local pairwise correspondences.** PAIR-X mimics the explainability visualizations produced by classical feature matching techniques (as shown in Figure 2). Correspondences between local image regions can be explicitly visualized.
- **Fine-grained resolution.** PAIR-X produces explanations in the resolution of the original input image, thus capturing details in full resolution and shifting focus from broadly relevant regions to highly discriminative details.
- **Quantifiable metrics.** Previous explainability approaches have often relied solely on manual, qualitative review to measure performance. For PAIR-X, we additionally propose a set of quantitative metrics (see Section 3.3) to compare its performance across models and datasets. Our metrics are designed to approximate how plausible a given visualization will appear to a user.

We evaluate PAIR-X across 34 public datasets for animal re-ID from WildlifeDatasets (Čermák et al., 2023) as well as the Oxford Building 5k dataset as a proof-of-concept outside of animal re-ID (Philbin et al., 2007). We find that PAIR-X performs best on fine-grained, pattern-based tasks, rather than cases where identity depends on more global or gestalt characteristics. Qualitatively, PAIR-X enables the visualization of interpretable local correspondences between image pairs, which are useful for both efficiently verifying correct matches and flagging high-scoring incorrect matches. We additionally propose a novel metric for our method which quantitatively demonstrates that PAIR-X produces measurably more plausible explanations for matching image pairs than for non-matching image pairs. PAIR-X can distinguish correct and incorect pairs **even in some cases where the deep metric model fails** (*e.g.* high model match score for an incorrect pair).

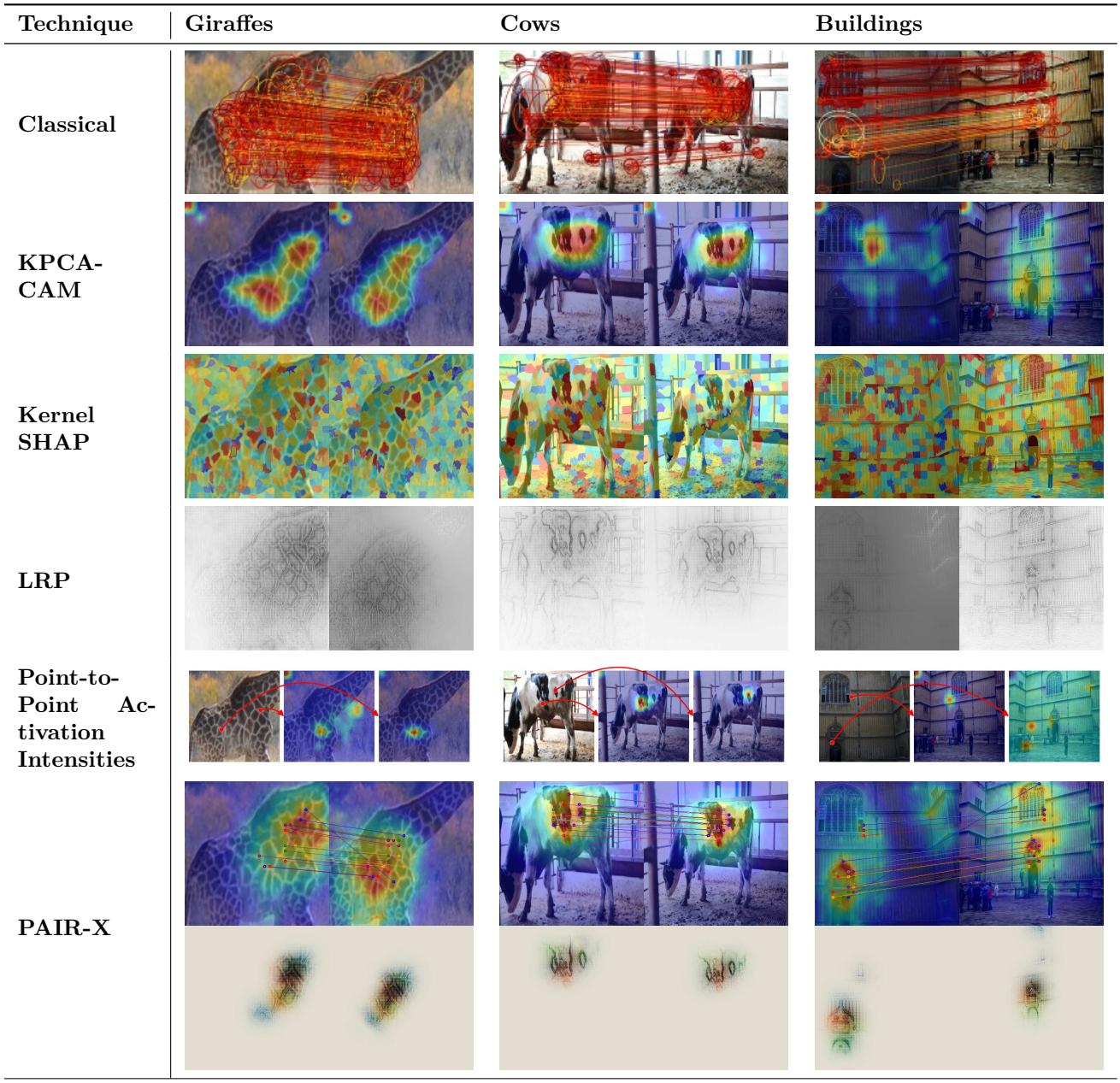


Figure 2: Qualitative comparison of explainability techniques across image pairs selected randomly from three datasets. See Supplementary for a more expansive justification of baselines, including an ablation over 12 CAM-based techniques and a hyperparameter search for SHAP.

2 Related work

2.1 Animal re-ID

Historically, classical, inherently-explainable techniques based on homography-aligned local feature matching were used for individual re-ID of patterned (Crall et al., 2013; Lahiri et al., 2011; Kelly, 2001) and contoured species (Weideman et al., 2020; Hughes & Burghardt, 2017). Recently, deep models have been applied to a broad range of animal re-ID datasets (Otarashvili et al., 2024; Schneider et al., 2018; Čermák et al., 2023; Andrew et al., 2021; Haurum et al., 2020; Deb et al., 2018; Dlamini & Zyl, 2020). Deep metric learning

methods (Wang et al., 2014), including both convolutional neural network (CNN) (Otarashvili et al., 2024) and transformer architectures (Čermák et al., 2023), enable recognition of animals not seen during training, and have been shown to both generalize across species (Čermák et al., 2023), and scale more efficiently to large populations (Otarashvili et al., 2024). Concepts from local feature matching have been applied to deep vision model features for use in co-segmentation (Li et al., 2019b) and semantic correspondence (Ufer & Ommer, 2017). Like local features, these deep features can be matched between image pairs (Fischer et al., 2015; Amir et al., 2022; Balntas et al., 2017), but exploration into their use for explainability has been limited.

2.2 Explainability for deep vision models

Methods for generating saliency heat maps (*e.g.* , Grad-CAM, Grad-CAM++, DiffCAM, etc.) (Selvaraju et al., 2016; Chattopadhyay et al., 2017; Karmani et al., 2024; Draelos & Carin, 2021; Li et al., 2025; Zheng et al., 2020) are used to highlight image regions that contribute to the prediction of a certain class or to the similarity of features between pairs of images, but struggle to capture fine-grained features (Achtibat et al., 2023). Explainability methods which repeatedly perturb an image and measure the change in model output (Ribeiro et al., 2016; Lundberg & Lee, 2017) can be both computationally expensive and qualitatively ineffective for fine-grained details (see Fig. 2). Methods that build explainability into new model architectures can yield high-quality explanations, but require training specialized models from scratch (Chen et al., 2019). Layer-wise relevance propagation (LRP) backpropagates relevance back to the original image pixels and can capture fine-grained details (Bach et al., 2015). The initial relevance value is defined according to the model output, and relevance to that output is propagated backwards through the model according to a defined set of rules. This yields explanations in the spatial resolution of the original input; every image pixel can be assigned a relevance value. However, as seen in Figure 2, LRP fails to demonstrate how or why these details contribute to the model prediction. Concept relevance propagation (CRP) (Achtibat et al., 2023) combines the fine-grained, localized explanations of LRP with encoded, interpretable "concepts". While CRP can offer precise insights into a model's internal workings, it requires substantial human review to identify relevant and understandable concepts. To directly compare pairwise feature correspondences, Zhu et al. (2021) compute point-to-point activation intensity between pairs of images, and Nguyen et al. (2023) use optimal transport to match image regions. Both approaches rely on final-layer features, which are often coarse, and do not visualize receptive fields.

3 Our method

PAIR-X assumes access to a pretrained deep metric learning model (*i.e.* Kaya & Bilge (2019)) for the fine-grained task of interest, and, taking inspiration from classical feature-matching techniques, constructs an interpretable, highly localized post-hoc explanation of similarity between image pairs by combining intermediate deep feature matching (Fischer et al., 2015) and layerwise relevance propagation (Bach et al., 2015). These explanations can be generated for top-ranked pairs and used for human review and validation of matches. A visual description of our method can be found in Figure 3, and additional methodological details are captured in Suppl. Sec. B.

3.1 Deep feature matching

Local feature matching techniques typically operate on sets of keypoints $\mathcal{K}$ and descriptors $\mathcal{D}$, where keypoints describe spatial locations within an image, and descriptors provide information about features present at the keypoints. PAIR-X uses a simple spatial decomposition of an intermediate activation matrix to produce keypoint-descriptor sets $\mathcal{K}$ and $\mathcal{D}$ using the intermediate activations of a model (Fischer et al., 2015; Amir et al., 2022). Given an intermediate activation matrix A^l at selected layer l of shape $w \times h \times c$, which can be viewed as a spatial grid of c-dimensional feature vectors, where $w \times h$ represents the spatial resolution at layer l, we decompose A^l into $N = w \times h$ descriptors of length c, thus $\mathcal{D} = \{A^l_{i,j} \mid i \in \{1, ..., w\}, j \in \{1, ..., h\}\}$. The keypoints $\mathcal{K}$ for each descriptor are simply defined according to the estimated location in the image, *i.e.* , $\mathcal{K} = \{(i, j) \mid i \in \{1, ..., w\}, j \in \{1, ..., h\}\}$. As this definition of keypoint locations does not fully capture

the true receptive fields for each neuron, we additionally utilize LRP to create more precise visualizations (see Sec. 3.2). Given the complete sets of keypoints and descriptors, we perform brute-force matching with cross-checking, *i.e.* , descriptors (x, y) will only be returned for a match if x is the closest match to y, and vice versa. This yields a set of matches M.

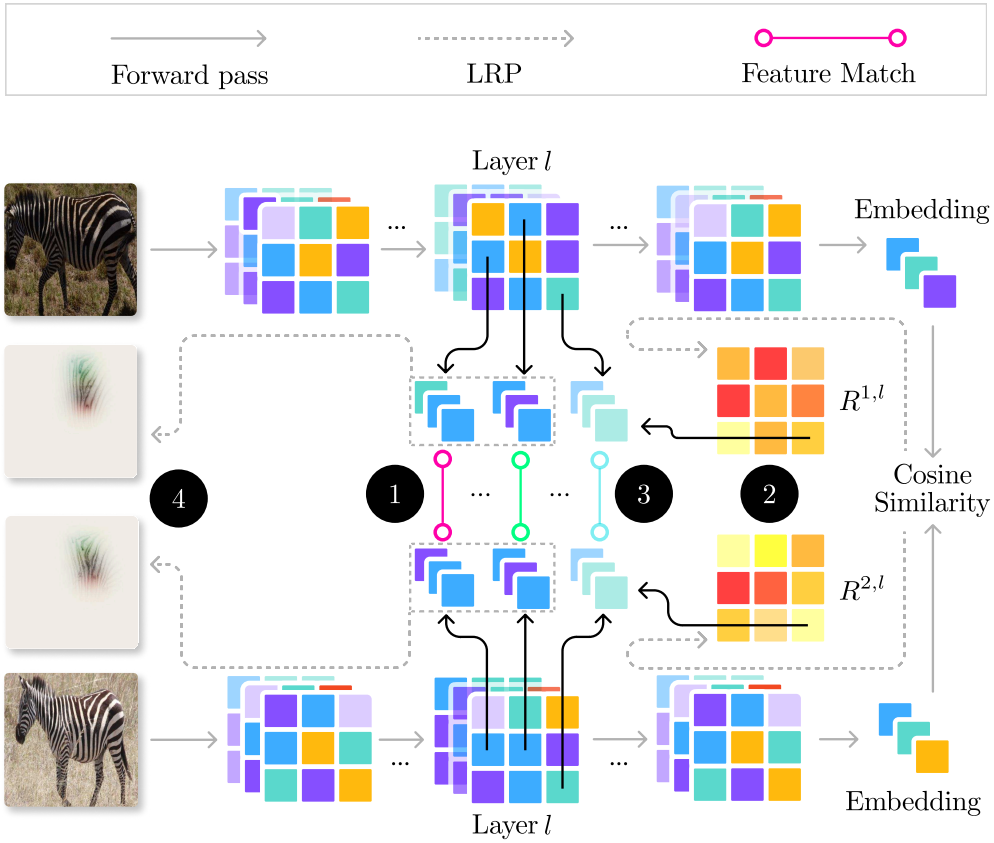


Figure 3: Overview of PAIR-X. In step 1, we match deep features derived from layer l. In step 2, we perform LRP to obtain the relevance of each feature to the final cosine similarity. In step 3, we filter the matched features acccording to their estimated relevance. Finally, in step 4, we use LRP to visualize which original image pixels are relevant to a filtered subset of matches.

3.2 Layerwise relevance propagation

Naive feature matching results in a large set of matches, some of which are unimportant to the model prediction. To filter this large set of candidate matches, we first use LRP to determine relevance values for each neuron in the selected intermediate layer l. The resulting matrix takes shape $w \times h \times c$, and represents the estimated relevances of the values in the intermediate feature map at l. Given a keypoint match between (i_1, j_1) and (i_2, j_2) in the first and second images, respectively, we compute a relevance score for the match:

$$Rel((i_1, j_1), (i_2, j_2)) = \left(\sum_k R^{1,l}_{i_1,j_1,k} \right) \times \left(\sum_k R^{2,l}_{i_2,j_2,k} \right) \tag{1}$$

where $R^{1,l}$ and $R^{2,l}$ are the intermediate relevance matrices for the two images, and we keep the n highest-scoring matches ($n = 20$ in our figures).

Neurons typically draw information from a wider surrounding region, or receptive field, which is not captured by a visualization with lines connecting approximated keypoints. To precisely visualize the pixel-space

contributions of matched features, we use LRP to backpropagate from the selected intermediate layer l to the original image for a set of top matches (ranked according to the relevance metric presented in Equation 1). Given a feature match between keypoints (i_1, j_1) and (i_2, j_2), we backpropagate for each image in the pair separately. We first mask the intermediate activation matrix, A^l, to the values included in the matched keypoint descriptor. This takes the form:

$$\mathcal{M}(A^l_{i,j}) = \begin{cases} A^l_{i,j}, & \text{if } (i,j) = (i_1, j_1) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

We then backpropagate relevance from $\mathcal{M}(A^l_{i,j})$, using the rules defined by LRP. The result of this backpropagation takes the same shape as the original input image, which can be summed channel-wise for RGB images to produce a 2D heatmap. We visualize the pixel-wise relevances by assigning a different color map to each match, then combining all matches into a single color-coded visualization.

3.3 Quantitative explainability metrics

Because explainability is inherently qualitative, it is difficult to define metrics that can quantify explainability performance. We propose two quantitative metrics which seek to capture the "plausibility" of PAIR-X explanations.

Inverted residual mean. Our first metric, the inverted residual mean, aims to capture whether the feature matches follow a "ground truth" homography $\mathcal{H}$, as a proxy to understand whether the matches correctly align the image subject. Matches that do not follow a homography will typically appear less plausible. $\mathcal{H}$ is calculated using classical techniques; we use a SuperPoint extractor to extract local features and a LightGlue matcher to find feature matches, then estimate $\mathcal{H}$ from the matches (DeTone et al., 2017; Lindenberger et al., 2023). Each keypoint $p_1 = (i_1, j_1)$ is projected from the first image using this "ground truth" homography as:

$$\mathcal{H}\left(\begin{bmatrix} i_1 \\ j_1 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} i'_1 \\ j'_1 \\ w' \end{bmatrix} \rightarrow p'_1 = \frac{1}{w'} \begin{bmatrix} i'_1 \\ j'_1 \end{bmatrix}. \tag{3}$$

For the final score S_1, we take the reciprocal of the average of the residuals on the second image (between the projected points and the matched points) across all feature matches M (before LRP filtering):

$$S_1 = \frac{|M|}{\sum_{(p_1,p_2)\in M} ||p'_1 - p_2||}. \tag{4}$$

We take the reciprocal because residual means are typically concentrated around small values, with a long tail of high-value outliers. Taking the reciprocal allows for improved separability of the smaller values.

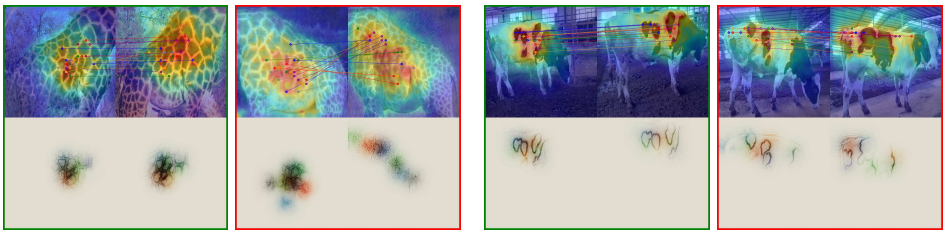
Relevance-weighted match coverage. The second metric we propose, relevance-weighted match coverage, aims to understand what proportion of relevant regions are successfully matched by PAIR-X. Visualizations that fail to show matches between the most important regions of an image will also appear less informative. Each feature that has been matched by PAIR-X is weighted by its relevance score, summed, and then divided by the sum of all relevance scores:

$$\frac{\sum_{(i,j)\in\mathcal{K}_1} R^1_{i,j} + \sum_{(i,j)\in\mathcal{K}_2} R^2_{i,j}}{\sum_{i,j} R^1_{i,j} + \sum_{i,j} R^2_{i,j}}, \tag{5}$$

where $\mathcal{K}_1$ and $\mathcal{K}_2$ denote the sets of matched keypoints of the two images, before being filtered by relevance. For each dataset evaluated, we compute these metrics across top-ranked correct and incorrect pairs following the procedure described in Suppl. Sec. B.4.

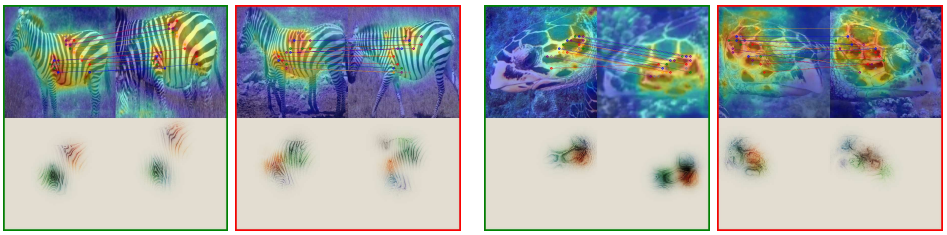
Patterned species

PAIR-X performs best on fine-grained tasks with highly-localized and structured details, e.g. re-ID for patterned species such as giraffes, cows, zebras, and sea turtles. These patterns capture unique spatial arrangements for individuals, thus the visualizations for correct and incorrect matches are interpretably different. Green outlines indicate correct image matches (same individual), and red outlines indicate incorrect matches (different individuals).



Giraffes - Correct and Incorrect Matches

Cows - Correct and Incorrect Matches

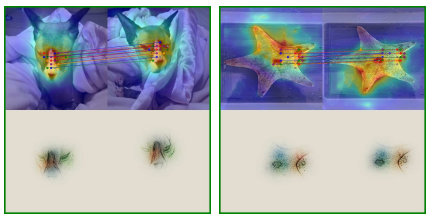


Plains Zebras - Correct and Incorrect Matches

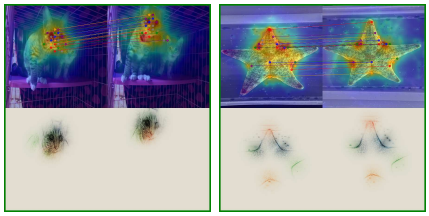
Sea Turtles - Correct and Incorrect Matches

Unpatterned species

PAIR-X is designed to optimally explain fine-grained, localized features that follow unique spatial arrangements for different categories. **For species without structured biometric patterns such as stripes, we find PAIR-X explanations are more useful for individuals with unique markings.** Without these markings, PAIR-X can sometimes highlight features that are invariant between individuals, producing misleading explanations for incorrect matches.



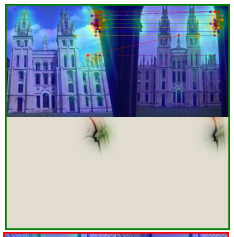
Cat and Sea Star Images with Distinctive Markings

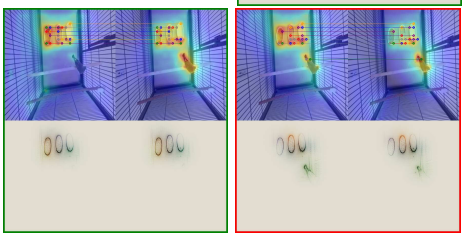


Cat and Sea Star Images without Distinctive Markings

Identifying spurious correpondences

PAIR-X can help identify when model decisions are based on irrelevant information. To the right, it shows the model's focus on foreground information.

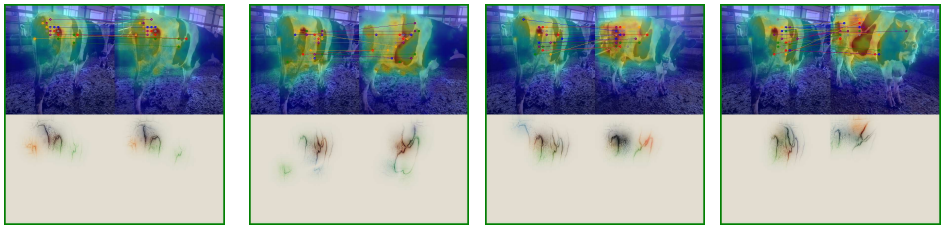




For these bird images, PAIR-X captures the background information contributing to the model prediction.

Failure mode: extreme pose variation

As with many classical feature-matching techniques, PAIR-X performance degrades as pose variation becomes increasingly extreme.



Minor pose difference | Moderate pose difference | Major pose difference | Extreme pose difference

Figure 4: Qualitative analysis of trends in PAIR-X outputs.

Table 1: Quantitative Metrics Across Datasets

Dataset	ρ_{res}	$\Delta_{res} \uparrow$	ρ_{mc}	$\Delta_{mc} \uparrow$
AAUZebraFish (Haurum et al., 2020)	0.73	0.56	0.75	0.78
ATRW (Li et al., 2019a)	0.72	0.16	0.79	0.18
BelugaIDv2 (Me, 2022a)	0.27	-0.01	0.43	-0.00
BirdIndividualID (Ferreira et al., 2020)	0.62	0.07	0.71	0.03
CTai (Freytag et al., 2016)	0.24	0.09	0.57	0.08
CZoo (Freytag et al., 2016)	0.27	0.14	0.36	0.24
CatIndividualImages (Lin, 2020)	0.88	0.08	0.92	0.13
CowDataset (Fu & Gong, 2021)	0.70	0.39	0.85	0.78
Cows2021v2 (Gao et al., 2021)	0.77	0.74	0.80	1.32
DogFaceNet (Mougeot et al., 2019)	**0.48**	**-0.01**	**0.70**	**0.02**
ELPephants (Korschens & Denzler, 2019)	0.21	0.09	0.32	0.16
FriesianCattle2015v2 (Andrew et al., 2016)	0.74	0.22	0.81	0.24
FriesianCattle2017 (Andrew et al., 2017)	0.74	0.08	0.89	0.06
GiraffeZebraID (Parham et al., 2017)	0.38	0.15	0.41	0.40
Giraffes (Miele et al., 2021)	**0.74**	**0.57**	**0.75**	**0.85**
HyenaID2022 (Me, 2022b)	0.46	-0.03	0.63	0.02
IPanda50 (Wang et al., 2021)	0.38	0.24	0.44	0.21
LeopardID2022 (Me, 2022c)	0.50	0.04	0.54	0.01
LionData (Dlamini & van Zyl, 2020a)	0.18	-0.05	0.28	0.03
MacaqueFaces (Witham, 2018)	0.24	0.05	0.43	0.09
NDD20v2 (Trotter et al., 2020)	0.44	0.04	0.56	0.10
NOAARightWhale (Khan et al., 2015)	0.45	0.16	0.62	0.22
NyalaData (Dlamini & van Zyl, 2020b)	0.30	0.23	0.62	0.14
OpenCows2020 (Andrew et al., 2021)	0.78	0.31	0.80	0.27
ReunionTurtles (Jean, 2024)	0.50	0.18	0.74	0.13
SMALST (Zuffi et al., 2019)	0.46	-0.48	0.67	-0.19
SeaStarReID2023 (Wahltinez & Wahltinez, 2024)	0.64	0.38	0.68	0.29
SeaTurtleID2022 (Adam et al., 2024)	0.26	0.17	0.28	0.14
SeaTurtleIDHeads (Adam et al., 2024)	0.26	0.22	0.31	0.16
SealID (Nepovinnykh et al., 2022)	0.56	0.50	0.88	0.13
SouthernProvinceTurtles (Sou, 2024)	0.69	0.14	0.75	0.43
WhaleSharkID (Holmberg et al., 2009)	0.19	0.31	0.73	0.05
ZakynthosTurtles (Papafitsoros, 2024)	0.48	0.00	0.87	0.00
ZindiTurtleRecall (Zin, 2022)	0.46	0.03	0.71	0.10
Oxford5k (Philbin et al., 2007)	0.64	0.37	0.69	0.58

Inverted residual mean (res) and relevance-weighted match coverage (mc) across datasets, aggregated within each dataset using both Spearman's rank correlation coefficient (ρ) between each metric and model match score, and binned Bhattacharyya distance (Δ) of each metric between correct and incorrect matches. Positive Δ values indicate that PAIR-X improves separation for similarly scored matches. Metrics for **bolded** datasets are shown in more detail in Figure 5.

4 Results

Using the multispecies re-ID model Miew-ID[2] as our deep metric model (Otarashvili et al., 2024), we evaluate PAIR-X across 34 public datasets for animal re-ID from WildlifeDatasets (Čermák et al., 2023), as well as the Oxford5k building dataset (Philbin et al., 2007). See Suppl. Sec. D for results on an additional model, and Suppl. Sec. F for a preliminary expansion to transformers. The metrics defined in Sec. 3.3 (which we refer to as the PAIR-X metrics) are aggregated across each dataset using two additional values described

[2]weights publicly available at `https://huggingface.co/conservationxlabs/miewid-msv2`

below. Results are presented in Table 1. In Figure 5, we visualize our metrics across individual image pairs in specific datasets.

Dataset Metrics. Both metrics are computed for a fixed set of image pairs for each dataset (see Suppl. Sec. B for details on pair selection). To aggregate these metrics across pairs within each dataset, we computed two additional values for each PAIR-X metric. First, we measure the rank correlation between the PAIR-X metric and the model similarity scores, to ascertain whether PAIR-X visualizations appear more plausible for pairs that the model scores more highly. This is done using Spearman's rank correlation coefficient ρ (see Table 1). Second, we measure the separability of correct and incorrect pairs using the PAIR-X metric, while controlling for the model similarity score. The goal of this separability test is to ascertain whether, for correct and incorrect pairs that the model cannot distinguish, PAIR-X visualizations appear more plausible for correct matches than incorrect ones. This value is denoted as Δ in Table 1, and it is measured by binning over cosine similarity, then taking a weighted average of bin-wise Bhattacharyya distances (see Suppl. Sec. B.3 for exact details).
In Figure 5, we visualize and discuss possible distributions of our metrics for correct and incorrect pairs, to develop intuition for the meaning of these metrics.

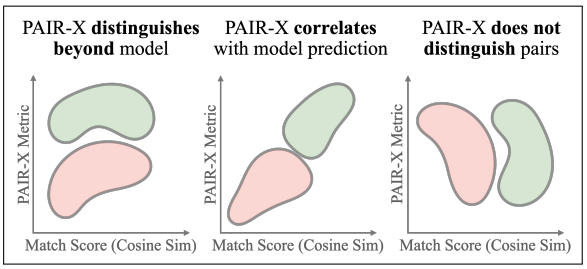
5 Discussion

5.1 Performance across applications

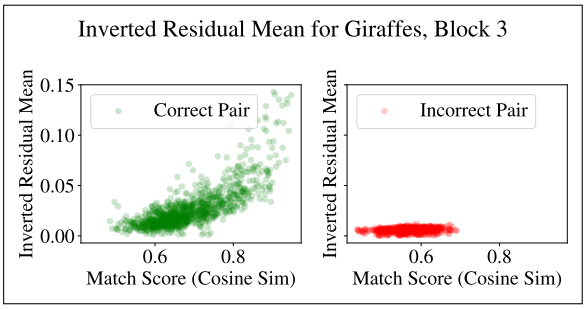
PAIR-X performs best on fine-grained tasks with highly-patterned or highly-localized features. As an example, the giraffe data in WildlifeDatasets consists of high-quality, well-cropped images of Reticulated Giraffes, a species with dense, uniquely oriented patterns of highly-localized features. As shown in Fig. 5, the PAIR-X scores for this dataset follow two relevant trends. First, we see a strong positive correlation between match score and PAIR-X score, indicating that PAIR-X visualizations are more plausible for higher-scoring image pairs. It also suggests that PAIR-X is unlikely to produce highly plausible but misleading visualizations for image pairs with low model match scores. Second, we see that the PAIR-X scores show an additional dimension of separability between correct and incorrect pairs. For image pairs that the model assigns equivalent match scores, the PAIR-X metrics suggest that visualizations appear, on average, more plausible for correct than incorrect pairs. This is a promising result, and if additional separability between correct and incorrect pairs can be achieved through this type of method, it could perhaps be directly utilized to improve model accuracy.

Our method shows potential for fine-grained explainability beyond animals, particularly for other tasks with highly-structured and localized features such as building facades. We conduct a detailed analysis of performance on the Oxford5k dataset in Suppl. Sec. C.1.

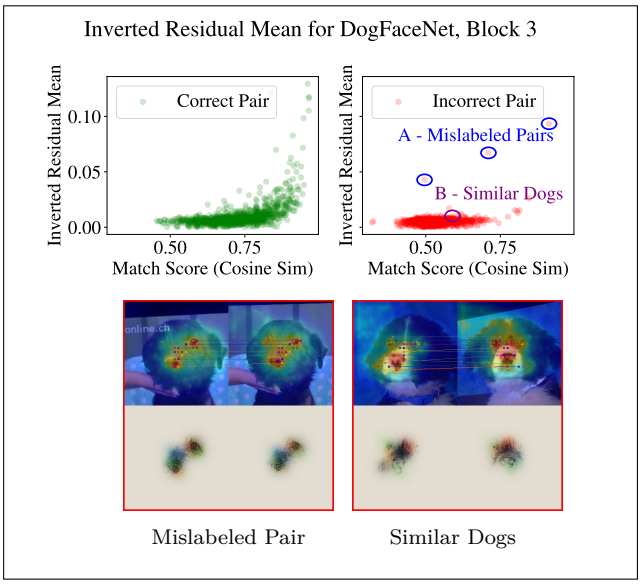
PAIR-X is less well-suited for fine-grained tasks with less-localized or less-structured distinguishing features. As an example, for images in DogFaceNet, we see a much lower degree of separability between correct and incorrect pairs. Qualitatively, we see that this is largely due to structural similarities between individuals: PAIR-X is, for instance, likely to find matches between eyes and noses of incorrectly-matched dog pairs, especially when comparing different individuals from the same breed. This raises the question of what optimal explainability would look like in this case, where identification may be more gestalt than localizeable (*i.e.* subtle variations in relative spacing between eyes and nose, as opposed to unique patterns of stripes). As we see in Figure 5, many of the incorrect pairs that receive high model similarity scores are of very similar-looking dogs. In these cases, PAIR-X provides informative visualizations of the features that contribute most to similarity (*e.g.* eyes, nose). However, as these facial features contribute to similarity for both correct and incorrect matches, PAIR-X may produce misleading visualizations for incorrect matches, and is thus not as useful for manual review of model predictions. This is in contrast to highly patterned species, where highly-matched but uniquely-structured patterns for each individual are easily distinguished in our visualizations, and thus lead to higher PAIR-X metric scores. Examples on additional less-patterned species (cats and starfish) are shown in Figure 4.



Three possible distributions of correct and incorrect matches using our PAIR-X metrics.



The PAIR-X metric separates correctly and incorrectly matched pairs well for highly-patterned species such as giraffes, suggesting that PAIR-X proves highly explainable in these cases.



The PAIR-X metric struggles to separate correct from incorrect matches for less-highly-patterned species like dogs. However, this analysis identified several outliers in the DogFaceNet dataset which were found to be mislabeled copies of the same images, with distortions such as rotations and cropping. This ability to flag mislabeled pairs is potentially useful. However, some incorrect pairs still receive moderately high scores from both the model and PAIR-X simply because they are similar-looking.

Figure 5: We provide visualizations to build intuition for interpreting plots of PAIR-X metrics (inverted residual mean vs. match score), as well as real examples of these plots for a higher-performing (giraffe) and a lower-performing species (dogs).

5.2 Quantitative comparisons between methods

Ideally it would be possible to not only quantify the performance of our method across datasets, but also quantitatively compare our method to other explainability techniques. General-purpose quantifiable metrics for explainability are still out of reach, and our PAIR-X metrics assume explicit pairwise feature matching to calculate, and are thus not directly applicable to, e.g., CAM-based methods. Thus, we rely on qualitative comparisons and expert interviews to measure the relative interpretability and usefulness of explainations from different methods. That said, we want to highlight the value that our method-specific metrics provide. The ability to quantify explainability in PAIR-X allows a user to efficiently determine whether PAIR-X is a good fit for their task of interest.

5.3 Applicability to real-world use cases

In an envisioned use case for animal re-ID, PAIR-X visualizations could be used to more efficiently manually validate model predictions, especially in cases where the closest correct match and closest incorrect match are scored similarly. Qualitatively, PAIR-X visualizations help to isolate important information and to visually align images, reducing the manual labor required for match verification. As discussed in Section 4, the difference in PAIR-X metrics between correct and incorrect pairs with similar match scores suggests that visualizations for correct pairs would, on average, appear more visually plausible. This is especially true for datasets with a high degree of separability, as measured by Δ in Table 1. We further analyze the applicability of PAIR-X to real-world use cases via expert interviews and a brief analysis of computational costs.

5.3.1 Expert interviews

Because explainability is highly subjective, we found it important to consider perspectives from downstream model users about the usability of PAIR-X. Animal re-ID is a niche topic, with a very limited number of experts capable of manual re-ID on these datasets, which limited the pool of people from whom to collect feedback. However, we interviewed three experts in giraffe re-ID, and discuss a few key insights gained from those conversations.

In real-world deployments of giraffe re-ID models, experts are tasked with manually verifying large batches of image labels, which frequently requires reviewing between five and twenty top-ranked database matches per query image. In this setting, efficiency is critical. The experts we interviewed agreed that explainability visualizations that highlight relevant image regions are helpful for directing user attention and speeding up verification.

We asked experts about their preferences between PAIR-X, SIFT feature matching such as HotSpotter, and Grad-CAM++. While experts found Grad-CAM++ to be more useful than *no* explainability, they found the fine-grained information provided by PAIR-X to be more helpful. Between classical SIFT feature matching and PAIR-X, experts were split. One expert had used HotSpotter extensively before switching to deep models, and thus had a preference for the classical feature-matching visualization, but recognized that its inability to scale to their current database rendered it no longer usable. Another expert, who had not previously used HotSpotter, found those visualizations to contain too many matches to be interpretable, and preferred PAIR-X for its filtered set of feature matches.

5.3.2 Computational efficiency

Since experts are interactively reviewing large numbers of images (one of the experts we interviewed had reviewed more than 100,000 during their time in the field), low latency, and therefore computational efficiency, is essential. On a single A100 GPU, we find that creating a typical explanation for 10 backpropagated matches requires 5 seconds, which demonstrates the feasibility of using PAIR-X at scale. We expand upon the factors influencing computational efficiency in Suppl. Sec. E.

6 Conclusion

We present PAIR-X, a novel fine-grained explainability technique based on a combination of deep feature matching and layer-wise relevance propagation (LRP), which provides explanations of pairwise similarity based on pretrained deep metric learning models. We demonstrate promising results on a diverse collection of animal re-ID datasets, as well as the Oxford-5k building dataset. Qualitatively, the explanations produced by PAIR-X are finer-grained than existing CAM-based techniques, as well as easier to interpret thanks to explicit matching of relevant image features and the color-coded propagation of those features back into image space. We furthermore propose a set of quantitative metrics which show that PAIR-X is in many cases able to distinguish correct from similarly scoring incorrect (*i.e.* confusing) matches. While PAIR-X may produce misleading explanations for species with a high degree of structural similarity, we show that there are many patterned species PAIR-X is applicable to. The experts we interviewed unanimously agreed that PAIR-X explanations are useful and informative, and WildMe[3], a cross-species animal re-identification platform, has expressed intent to deploy our method for all of its patterned species, emphasizing its applicability and usefulness in real-world settings.

References

Turtle recall: Conservation challenge, 2022. URL `https://zindi.africa/competitions/turtle-recall-conservation-challenge`.

Southern province turtles. `https://www.kaggle.com/datasets/wildlifedatasets/southernprovinceturtles`, 2024. Kaggle.

[3]https://www.wildme.org/

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. doi: 10.1109/TPAMI.2012.120.

Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, Sep 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00711-8. URL `https://doi.org/10.1038/s42256-023-00711-8`.

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attnlrp: Attention-aware layer-wise relevance propagation for transformers, 2024. URL `https://arxiv.org/abs/2402.05602`.

Lukáš Adam, Vojtěch Čermák, Kostas Papafitsoros, and Lukas Picek. Seaturtleid2022: A long-span dataset for reliable sea turtle re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7146–7156, 2024.

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors, 2022. URL `https://arxiv.org/abs/2112.05814`.

Christopher J. Anders, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. Software for dataset-wide xai: From local explanations to global insights with Zennit, CoRelAy, and ViRelAy. *CoRR*, abs/2106.13200, 2021.

William Andrew, Sion Hannuna, Neill Campbell, and Tilo Burghardt. Automatic individual holstein friesian cattle identification via selective local coat pattern matching in rgb-d imagery. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 484–488. IEEE, 2016.

William Andrew, Colin Greatwood, and Tilo Burghardt. Visual localisation and individual identification of holstein friesian cattle via deep learning. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 2850–2859, 2017.

William Andrew, Jing Gao, Siobhan Mullan, Neill Campbell, Andrew W. Dowsey, and Tilo Burghardt. Visual identification of individual holstein-friesian cattle via deep metric learning. *Computers and Electronics in Agriculture*, 185:106133, June 2021. ISSN 0168-1699. doi: 10.1016/j.compag.2021.106133. URL `http://dx.doi.org/10.1016/j.compag.2021.106133`.

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, July 2015.

Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35, 99-109., 1943.

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *CoRR*, abs/1710.11063, 2017. URL `http://arxiv.org/abs/1710.11063`.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization, 2021. URL `https://arxiv.org/abs/2012.09838`.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf.

Jonathan P. Crall, Charles V. Stewart, Tanya Y. Berger-Wolf, Daniel I. Rubenstein, and Siva R. Sundaresan. Hotspotter — patterned species instance recognition. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 230–237, 2013. doi: 10.1109/WACV.2013.6475023.

Debayan Deb, Susan Wiper, Alexandra Russo, Sixue Gong, Yichun Shi, Cori Tymoszek, and Anil Jain. Face recognition: Primates in the wild, 2018. URL https://arxiv.org/abs/1804.08790.

Saurabh Desai and Harish G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 972–980, 2020. doi: 10.1109/WACV45572.2020.9093360.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. *CoRR*, abs/1712.07629, 2017. URL http://arxiv.org/abs/1712.07629.

Nkosikhona Dlamini and Terence L van Zyl. Automated identification of individuals in wildlife population using siamese neural networks. In *2020 7th international conference on soft computing & machine intelligence (ISCMI)*, pp. 224–228. IEEE, 2020a.

Nkosikhona Dlamini and Terence L van Zyl. Automated identification of individuals in wildlife population using siamese neural networks. In *2020 7th international conference on soft computing & machine intelligence (ISCMI)*, pp. 224–228. IEEE, 2020b.

Nkosikhona Dlamini and Terence L van Zyl. Automated identification of individuals in wildlife population using siamese neural networks. In *2020 7th International Conference on Soft Computing Machine Intelligence (ISCMI)*, pp. 224–228, 2020. doi: 10.1109/ISCMI51676.2020.9311574.

Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks, 2021. URL https://arxiv.org/abs/2011.08891.

André C Ferreira, Liliana R Silva, Francesco Renna, Hanja B Brandl, Julien P Renoult, Damien R Farine, Rita Covas, and Claire Doutrelant. Deep learning-based methods for individual recognition in small birds. *Methods in Ecology and Evolution*, 11(9):1072–1085, 2020.

Philipp Fischer, Alexey Dosovitskiy, and Thomas Brox. Descriptor matching with convolutional neural networks: a comparison to sift, 2015. URL https://arxiv.org/abs/1405.5769.

Alexander Freytag, Erik Rodner, Marcel Simon, Alexander Loos, Hjalmar S Kühl, and Joachim Denzler. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *Pattern Recognition: 38th German Conference, GCPR 2016, Hannover, Germany, September 12-15, 2016, Proceedings 38*, pp. 51–63. Springer, 2016.

Lili Fu and He Gong. Cow dataset. 10 2021. doi: 10.6084/m9.figshare.16879780.v1. URL https://figshare.com/articles/dataset/data_set_zip/16879780.

Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns, 2020. URL https://arxiv.org/abs/2008.02312.

Jing Gao, Tilo Burghardt, William Andrew, Andrew W Dowsey, and Neill W Campbell. Towards self-supervision for video identification of individual holstein-friesian cattle: The cows2021 dataset. *arXiv preprint arXiv:2105.01938*, 2021.

Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.

Joakim Bruslund Haurum, Anastasija Karpova, Malte Pedersen, Stefan Hein Bengtson, and Thomas B. Moeslund. Re-identification of zebrafish using metric learning. In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 1–11, 2020. doi: 10.1109/WACVW50321.2020.9096922.

Jason Holmberg, Bradley Norman, and Zaven Arzoumanian. Estimating population size, structure, and residency time for whale sharks rhincodon typus through collaborative photo-identification. *Endangered Species Research*, 7(1):39–53, 2009.

Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Benjamin Hughes and Tilo Burghardt. Automated visual fin identification of individual great white sharks. *International Journal of Computer Vision*, 122:542–557, 2017.

Claire Jean. Reunionturtles. `https://www.kaggle.com/datasets/wildlifedatasets/reunionturtles`, 2024. Kaggle.

Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps. *IEEE Transactions on Image Processing*, PP:1–1, 06 2021. doi: 10.1109/TIP.2021.3089943.

Sachin Karmani, Thanushon Sivakaran, Gaurav Prasad, Mehmet Ali, Wenbo Yang, and Sheyang Tang. Kpca-cam: Visual explainability of deep computer vision models using kernel pca, 2024. URL `https://arxiv.org/abs/2410.00267`.

Mahmut Kaya and Hasan Şakir Bilge. Deep metric learning: A survey. *Symmetry*, 11(9):1066, 2019.

Marcella J Kelly. Computer-aided photograph matching in studies using individual identification: an example from serengeti cheetahs. *Journal of Mammalogy*, 82(2):440–449, 2001.

Christin B. Khan, Shashank, and Wendy Kan. Right whale recognition. `https://kaggle.com/competitions/noaa-right-whale-recognition`, 2015. Kaggle.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020.

Matthias Korschens and Joachim Denzler. Elpephants: A fine-grained dataset for elephant re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

Mayank Lahiri, Chayant Tantipathananandh, Rosemary Warungu, Daniel I. Rubenstein, and Tanya Y. Berger-Wolf. Biometric animal databases from field photographs: identification of individual zebra in the wild. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450303361. doi: 10.1145/1991996.1992002. URL `https://doi.org/10.1145/1991996.1992002`.

Shuyuan Li, Jianguo Li, Hanlin Tang, Rui Qian, and Weiyao Lin. Atrw: a benchmark for amur tiger re-identification in the wild. *arXiv preprint arXiv:1906.05586*, 2019a.

Weihao Li, Omid Hosseini Jafari, and Carsten Rother. Deep object co-segmentation. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler (eds.), *Computer Vision – ACCV 2018*, pp. 638–653, Cham, 2019b. Springer International Publishing. ISBN 978-3-030-20893-6.

Xingjian Li, Qiming Zhao, Neelesh Bisht, Mostofa Rafid Uddin, Jin Yu Kim, Bryan Zhang, and Min Xu. Diffcam: Data-driven saliency maps by capturing feature differences. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10327–10337, 2025.

Tzu-Yuan Lin. Cat individual images, 2020. URL https://www.kaggle.com/datasets/timost1234/cat-individuals.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed, 2023. URL https://arxiv.org/abs/2306.13643.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60 (2):91–110, November 2004. ISSN 0920-5691. doi: 10.1023/B:VISI.0000029664.99615.94. URL http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

Wild Me. Beluga id 2022, 2022a. URL https://lila.science/datasets/beluga-id-2022/.

Wild Me. Hyena id 2022, 2022b. URL https://lila.science/datasets/hyena-id-2022/.

Wild Me. Leopard id 2022, 2022c. URL https://lila.science/datasets/leopard-id-2022/.

Vincent Miele, Gaspard Dussert, Bruno Spataro, Simon Chamaillé-Jammes, Dominique Allainé, and Christophe Bonenfant. Revisiting animal photo-identification using deep metric learning and network analysis. *Methods in Ecology and Evolution*, 12(5):863–873, 2021.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. *Layer-Wise Relevance Propagation: An Overview*, pp. 193–209. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_10. URL https://doi.org/10.1007/978-3-030-28954-6_10.

Guillaume Mougeot, Dewei Li, and Shuai Jia. A deep learning approach for dog face verification and recognition. In Abhaya C. Nayak and Alok Sharma (eds.), *PRICAI 2019: Trends in Artificial Intelligence*, pp. 418–430, Cham, 2019. Springer International Publishing. ISBN 978-3-030-29894-4.

Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. *CoRR*, abs/2008.00299, 2020. URL https://arxiv.org/abs/2008.00299.

Ekaterina Nepovinnykh, Tuomas Eerola, Vincent Biard, Piia Mutka, Marja Niemi, Mervi Kunnasranta, and Heikki Kälviäinen. Sealid: Saimaa ringed seal re-identification dataset. *Sensors*, 22(19):7602, 2022.

Giang Nguyen, Mohammad Reza Taesiri, and Anh Nguyen. Visual correspondence-based explanations improve ai robustness and human-ai team accuracy, 2023. URL https://arxiv.org/abs/2208.00780.

Lasha Otarashvili, Tamilselvan Subramanian, Jason Holmberg, J. J. Levenson, and Charles V. Stewart. Multispecies animal re-id using a large community-curated dataset, 2024. URL https://arxiv.org/abs/2412.05602.

Frederik Pahde, Galip Ümit Yolcu, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Optimizing explanations by network canonization and hyperparameter search, 2023. URL https://arxiv.org/abs/2211.17174.

Kostas Papafitsoros. Zakynthos turtles. https://www.kaggle.com/datasets/wildlifedatasets/zakynthosturtles, 2024. Kaggle.

Jason Parham, Jonathan Crall, Charles Stewart, Tanya Berger-Wolf, and Daniel I Rubenstein. Animal population censusing at scale with citizen science and photographic identification. In *AAAI spring symposium-technical report*, 2017.

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007. doi: 10.1109/CVPR.2007.383172.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL https://arxiv.org/abs/1602.04938.

Stefan Schneider, Graham W. Taylor, Stefan S. Linquist, and Stefan C. Kremer. Past, present, and future approaches using computer vision for animal re-identification from camera trap data. *CoRR*, abs/1811.07749, 2018. URL http://arxiv.org/abs/1811.07749.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. URL http://arxiv.org/abs/1610.02391.

Suraj Srinivas and Francois Fleuret. Full-gradient representation for neural network visualization, 2019. URL https://arxiv.org/abs/1905.00780.

Peter T Stevick, Per J Palsbøll, Tim D Smith, Mark V Bravington, and Philip S Hammond. Errors in identification using natural markings: rates, sources, and effects on capturerecapture estimates of abundance. *Canadian Journal of Fisheries and Aquatic Sciences*, 58(9):1861–1870, 2001. doi: 10.1139/f01-131. URL https://doi.org/10.1139/f01-131.

Mengyu Tan, Wentao Chao, Jo-Ku Cheng, Mo Zhou, Yiwen Ma, Xinyi Jiang, Jianping Ge, Lian Yu, and Limin Feng. Animal detection and classification from camera trap images using different mainstream object detection architectures. *Animals (Basel)*, 12(15), August 2022.

Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pp. 10096–10106. PMLR, 2021.

Cameron Trotter, Georgia Atkinson, Matt Sharpe, Kirsten Richardson, A Stephen McGough, Nick Wright, Ben Burville, and Per Berggren. Ndd20: A large-scale few-shot dolphin dataset for coarse and fine-grained categorisation. *arXiv preprint arXiv:2005.13359*, 2020.

Nikolai Ufer and Bjorn Ommer. Deep semantic feature matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6 2014. ISSN 2167-8359. doi: 10.7717/peerj.453. URL https://doi.org/10.7717/peerj.453.

Oscar Wahltinez and Sarah J Wahltinez. An open-source general purpose machine learning framework for individual animal re-identification using few-shot learning. *Methods in Ecology and Evolution*, 15(2): 373–387, 2024.

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, 2020. URL https://arxiv.org/abs/1910.01279.

Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1386–1393, 2014.

Le Wang, Rizhi Ding, Yuanhao Zhai, Qilin Zhang, Wei Tang, Nanning Zheng, and Gang Hua. Giant panda identification. *IEEE Transactions on Image Processing*, 30:2837–2849, 2021.

Hendrik Weideman, Chuck Stewart, Jason Parham, Jason Holmberg, Kiirsten Flynn, John Calambokidis, D. Barry Paul, Anka Bedetti, Michelle Henley, Frank Pope, and Jerenimo Lepirei. Extracting identifying contours for african elephants and humpback whales using a learned appearance model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

Claire L Witham. Automated face recognition of rhesus macaques. *Journal of neuroscience methods*, 300: 157–165, 2018.

Meng Zheng, Srikrishna Karanam, Terrence Chen, Richard J Radke, and Ziyan Wu. Towards visually explaining similarity models. *arXiv preprint arXiv:2008.06035*, 2020.

Sijie Zhu, Taojiannan Yang, and Chen Chen. Visual explanation for deep metric learning, 2021. URL `https://arxiv.org/abs/1909.12977`.

Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images" in the wild". In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5359–5368, 2019.

Vojtěch Čermák, Lukas Picek, Lukáš Adam, and Kostas Papafitsoros. Wildlifedatasets: An open-source toolkit for animal re-identification, 2023. URL `https://arxiv.org/abs/2311.09118`.