UAB
Universitat Autònoma
de Barcelona

Interpretability Analysis in the Classification of Nature-Related Images in Social Media Videos

Bachelor's Thesis (TFG)

Author: Paula Feliu Criado 1630423

Profesors: Johannes Langemeyer & Ramin Solyemani

Date: May 18, 2025

Contents

Abstract

Abstract

1 Introduction and Context

Urbanization and digitalization have weakened direct interactions with nature, leading to a disconnect from the environment and community values . This "extinction of experience" describes a cycle where reduced exposure to nature across generations erodes familiarity, appreciation, and engagement. Soga and Gaston, 2016.

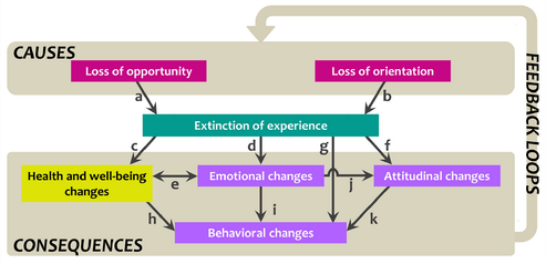


Figure 1: "Extinction of Experience". Soga and Gaston, 2016

Figure 1 illustrates how stressors like climate change, urban living, and increased screen time drive a feedback loop that diminishes opportunities and motivation to engage with nature, further reducing environmental concern and stewardship.

Psychological studies confirm that reduced contact with nature not only diminishes environmental concern but also shifts individual values toward self-centered aspirations, consequently impacting generosity, community involvement, and overall social connectedness Weinstein et al., 2009. As we can see in Figure 2, (data from Soga and Gaston, 2016), the first and second charts compare rates of children's outdoor activities at different points in time and highlights a marked decline in participation and time spent outside. In a related vein, the last chart demonstrates the downward trend in per capita visits to natural areas, underscoring the broader societal shift away from meaningful nature experiences.

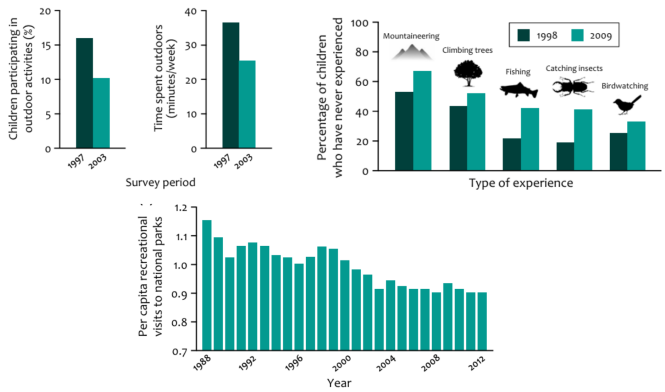


Figure 2: Statistics about the exposure to nature. Soga and Gaston, 2016

The BIG-5 project explores how digital platforms can foster human–nature connections by offering virtual encounters with the natural world. Its core concept of Digital Relational Values (DRV) defined as the nature-related values and attachments that arise within online communities through these indirect experiences Barcelona Supercomputing Center, 2021. This thesis contributes to that framework by deepening the understanding of how DRVs form and evolve in digital contexts. Recent work shows that such virtual interactions can not only compensate for lost physical contact with nature but also strengthen environmental appreciation, care, and stewardship—and may even motivate real-world conservation behaviors. Langemeyer and Calcagni, 2022.

To measure DRVs, the project use artificial intelligence (AI) to analyze social-media images for nature elements, human–nature interactions, and emotional context (Väisänen et al., 2021). Because trust in these insights hinges on transparency, technqiues as Grad-CAM, SHAP, and LIME will be applied to explain which visual features drive the model's decisions, aligning algorithmic outputs with human perception and strengthening the bridge between digital and real-world nature experiences.

2 Approach

This project is divided into two main phases, each playing a crucial role in achieving the overall goal of understanding and explaining the decision-making process of deep learning models in the context of nature-related image classification.

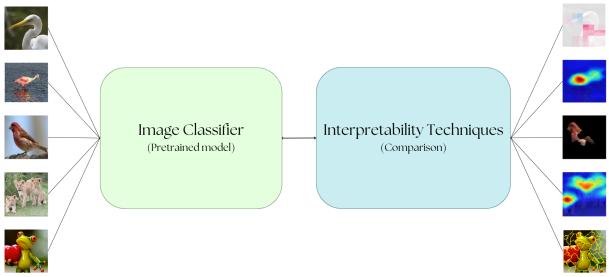


Figure 3: Approach

The first phase focuses on the development of an image classifier based on deep learning techniques. This classifier will categorize nature-related images into predefined classes, and its outputs will serve as the foundation for the second phase of the project: the comparison of interpretability techniques. The

goal of this phase is to assess how well different methods can explain the decisions made by the model.

Image Classification System

The image classifier will categorize images into four distinct classes, with each class structured to ensure mutual exclusivity within its category. The four categories are:

1. **Nature vs. Non-Nature**
2. **Biotic vs. Abiotic**
3. **Material vs. Immaterial**
4. **Landscape Type**: Classifying the image into one of the predefined landscape types.

Each image will be assigned one and only one label per class, ensuring exclusivity within each category. This approach follows a multiclass classification structure, where the classes are mutually exclusive, but independent of one another.

To develop the image classifier, deep learning techniques will be employed, particularly Convolutional Neural Networks (CNNs), which are highly effective for image classification tasks. The model will be trained to predict the most likely label for each of the four categories based on the visual content of the images.

The development of the classifier, is crucial for answering the following key questions:

1. What regions of the image does the model consider most relevant for classification?
2. How do these regions vary across different interpretability techniques?
3. How does the model's classification criteria compare to human classification criteria?

Interpretability Techniques

The second phase of the project will focus on addressing these questions. To achieve this, the trained classifier will be used as a foundation, and various interpretability techniques will be applied to analyze the model's decision-making process. Techniques such as Grad-CAM and LIME will be used to highlight the regions of the images that are most influential in the classifier's predictions.

The goal is to evaluate how effectively these interpretability methods can explain the model's outputs and how well they align with human classification criteria. By comparing the results across these techniques, we aim to identify which methods provide the most transparent insights into the model's behavior.

3 State of the Art

3.1 Image Classification and Transfer Learning

Image classification with deep learning has advanced through pretrained models and transfer learning. Pretrained networks—trained on large datasets like ImageNet—learn versatile features, from simple edges to complex object parts. Repurposing these models for new tasks provides a strong initialization, cutting down on task-specific data needs and speeding up convergence Bengio, 2012.

Transfer learning adapts a pretrained model to a new domain by using it as a fixed feature extractor and retraining or fine-tuning its higher layers. Low-level filters (e.g., edge detectors) remain mostly unchanged, while deeper layers adjust to the target data. This method excels when labeled data are scarce, delivering high performance in specialized applications such as nature image classification Pan and Yang, 2010.

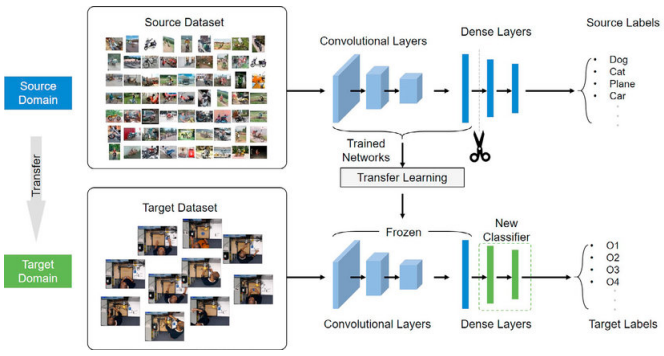


Figure 4: Transfer Learning Schema (Kumar, 2024)

Several landmark models underpin modern image classification through transfer learning such as **AlexNet** Krizhevsky et al., 2012, **VGGNet** Simonyan and Zisserman, 2015, **ResNet** He et al., 2016 and **GoogLeNet** Szegedy et al., 2015.

By leveraging these pretrained backbones, researchers can quickly fine-tune models for specialized tasks, cutting down on both computational cost and training time—especially valuable when labeled data are scarce.

3.2 Interpretability Techniques in Classification Models

Despite impressive accuracy, CNNs often operate as "black boxes," prompting the need for interpretability methods to reveal decision-making processes.

Class Activation Mapping (CAM) and Grad-CAM

CAM and Grad-CAM Selvaraju et al., 2017 are widely used for visually interpreting CNN decisions by generating heatmaps highlighting the regions influential for predictions. Grad-CAM uses gradients from the final convolutional layer to identify discriminative regions, effectively aligning the model's focus with human attention. Grad-CAM's effectiveness in revealing biases or dataset artifacts has made it essential for transparent AI evaluations.

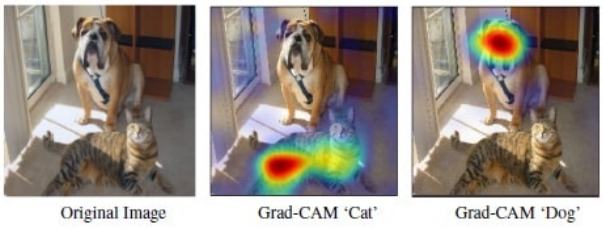


Figure 5: Grad-CAM. Adapted from Selvaraju et al., 2017.

Local Interpretable Model-Agnostic Explanations (LIME)

LIME Ribeiro et al., 2016 provides local, detailed explanations by segmenting images into superpixels and evaluating their influence on model predictions through perturbations. Despite its granularity, LIME faces stability and variability challenges, though it remains valuable for individual prediction explanations.

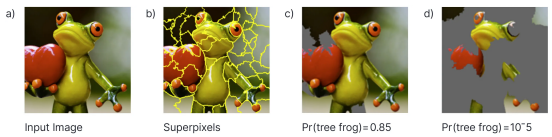


Figure 6: LIME. Adapted from Ribeiro et al., 2016.

SHapley Additive exPlanations (SHAP)

SHAP Lundberg and Lee, 2017 uses game theory to quantify pixel-level feature contributions to predictions, ensuring consistency and robustness in interpretations. Its unified approach integrates local and global interpretability, making it effective for validating model behavior.

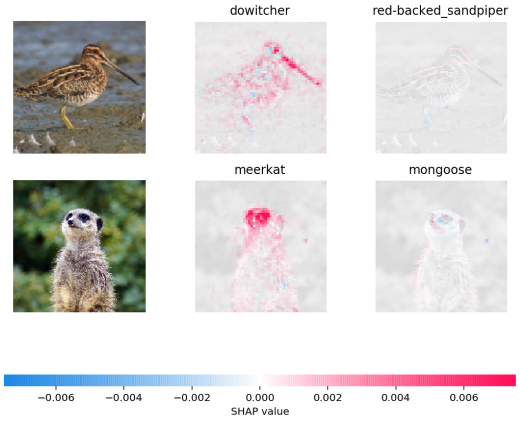


Figure 7: SHAP. Adapted from Lundberg and Lee, 2017.

Occlusion Sensitivity

Occlusion Sensitivity is a perturbation-based interpretability technique for CNN image classification, originally introduced by Zeiler and Fergus, 2014. The approach systematically occludes (masks or replaces) different portions of an input image to measure changes in the model's prediction confidence. If covering a particular region causes a significant drop in the target class probability, that region is inferred to be important for the model's decision. They used this method to verify that their ImageNet CNN was focusing on the object itself rather than spurious background context.

Figure 8: Occlusion Sensitivity. Zeiler and Fergus, 2014.

By sliding a gray patch across the image and recording the classifier's output, one can create an "importance" heatmap highlighting key object parts. This method is model-agnostic and intuitive, though it requires many forward passes (one per occluded position) and its resolution is limited by the occlusion window size. For example, in their results occluding a dog's face drastically reduced the "Pomeranian" class score (while occluding background had little effect), indicating that the face region was crucial for

the prediction. An adapted figure from Zeiler and Fergus (2014) could be captioned to illustrate this sensitivity map.

XRAI

XRAI Kapishnikov et al., 2019 is a region-based attribution method that produces more coherent visual explanations for CNN image classification. Unlike pixel-level saliency approaches, XRAI builds upon Integrated Gradients by aggregating attributions over meaningful image segments. The algorithm first oversegments the input image (e.g., into superpixels) and computes an attribution score (via integrated gradients) for each segment. It then iteratively adds the most salient segments, merging them into larger highlighted regions that strongly influence the prediction. This yields saliency maps that emphasize entire objects or important parts of the image, rather than scattered individual pixels.

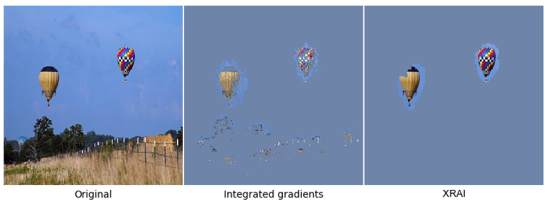


Figure 9: XRAI. Kapishnikov et al., 2019.

Kapishnikov et al. (2019) show that XRAI's region-focused explanations align better with true object locations and improve localization performance over standard saliency maps. For instance, they compare an image of a balloon where XRAI clearly highlights the balloon itself with minimal noise, while the pixel-wise heatmap appears diffuse. An adapted figure from their paper could be captioned to illustrate how XRAI produces a cleaner, region-based attribution concentrated on the object.

All figures presented in this section are adapted from the official papers of each method.

4 Objectives

The specific objectives of the project are:

1. **Data acquisition and processing:** Collect and preprocess images from different social media platforms to create a well-structured and labeled dataset.
2. **Classifier development:** Implement and train a neural network-based classifier to categorize nature-related visual content.
3. **Interpretability techniques:** Apply and compare methods such as CAM, Grad-CAM, SHAP, and LIME to visualize and analyze model decision-making.
4. **Comparison and evaluation:** Assess the effectiveness of interpretability techniques to determine which is most effective in this context.
5. **Conclusions and implications:** Evaluate how AI models contribute to the understanding of nature-related content and their potential impact on environmental awareness.

5 Methodology and Planning

The project is structured into clearly defined phases over approximately five months, each designed to ensure a comprehensive approach from research through analysis. The timeline and activities for each phase are summarized below and visually represented in Figure 10.

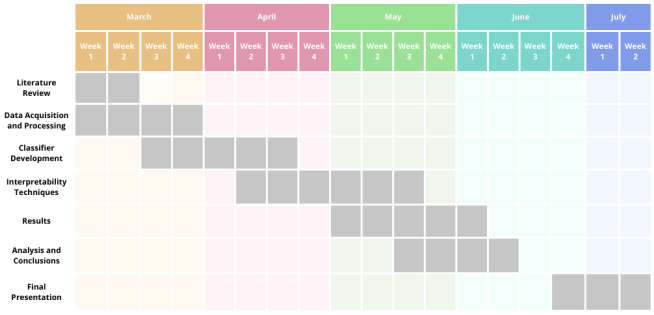


Figure 10: Gantt Chart

Literature Review (Weeks 1-3, March): Conduct a targeted literature review on image classification and interpretability to build a solid theoretical base.

Data Acquisition and Processing (Week 2 of March - Mid-April): Collect, label, and clean a diverse nature-themed image dataset from Platform X to ensure high quality and representativeness.

Classifier Development (End of March - April): Fine-tune a pretrained CNN (e.g., ResNet, EfficientNet), including hyperparameter tuning and overfitting safeguards, to create an accurate classifier.

Interpretability Techniques (Mid-April - Mid-May): Apply different techniques to visualize and dissect the model's decisions.

Results Analysis (May - June): Analyse and compare the outputs, draw conclusions, and prepare a final report and presentation summarizing findings and future directions.

This integrated approach ensures a logical progression through each stage, optimizing both methodological rigor and effective time allocation.

6 Data Collection and Dataset Creation

Creating a comprehensive and high-quality dataset tailored specifically to our project's objectives involved meticulous planning, execution, and rigorous standardization. In contrast to state-of-the-art projects that rely on publicly available datasets, our dataset is manually curated and built entirely by our team. This in-house approach ensures that our data precisely captures the diversity and authenticity of social media content related to nature interactions. The entire process was organized into three key stages: data collection, data labeling, and protocol implementation.

6.1 Data Collection

Our dataset consists exclusively of visual content sourced from social media, specifically from the platform X. To achieve a diverse and representative dataset, we systematically collected approximately 1700 high-quality images. These images were gathered using targeted searches, focusing on nature-related hashtags and verified accounts known for posting content aligned with the project's thematic focus on human-nature interactions.

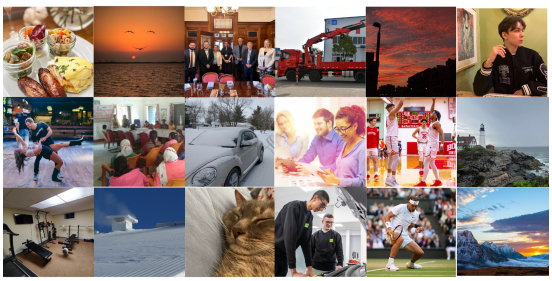


Figure 11: Examples of images in the dataset

6.2 Data Labeling

Once the images were collected, the next critical step involved labeling them according to our defined classification tasks. To achieve consistency and accuracy in this process, the labeling was performed using LabelStudio, a specialized tool designed for structured and reproducible labeling of image data. Each image was carefully examined and assigned labels across four specific, mutually exclusive categories:

1. **Nature vs. Non-Nature (Nature visual)**: Identifying whether the content represents nature or not.
2. **Biotic vs. Abiotic (Nep_materiality visual)**: Distinguishing between images containing living organisms (biotic) and non-living elements (abiotic).
3. **Material vs. Immaterial (Nep_biological visual)**: Classifying the content as either physically tangible (material) or abstract/intangible representations (immaterial).
4. **Landscape Type (Landscape-type visual)**: Categorizing images into predefined landscape types—artificial surfaces, agricultural areas, forests and semi-natural areas, wetlands, water bodies, other or none of these.

To maintain a high standard and reliability, all coders were required to reach an agreement on the assigned labels. This step was critical in ensuring data consistency, minimizing ambiguity, and enabling accurate model training.

6.3 Protocol Implementation

To ensure clear and consistent labeling, the BIG-5 group established a detailed coding protocol to standardize image classification:

- **Object of Analysis**: Relational values (RVs) conveyed explicitly or implicitly in visual and textual social media content.
- **Analytical Flow**: Coders first confirm relevant natural elements or activities, then resolve any visual–textual contradictions.
- **Categorization**: Defined criteria and examples for each dimension (Material vs. Immaterial, Biotic vs. Abiotic, Landscape Type).

- **RVs Identification**: Recognition of explicit or implicit personal, cultural, and social connections to nature, guided by documented examples.
- **Contradiction Handling**: Procedures for documenting and collaboratively discussing discrepancies between text and images.
- **Consistency**: Regular calibration meetings to refine criteria, resolve ambiguities, and maintain reliability.

Importance and Potential Issues: Thorough, unambiguous guidelines minimize labeling errors and ensure consistency across coders. Without them, inconsistent labels degrade classifier accuracy, compromise interpretability analyses, and undermine the project's validity.

6.4 Final Dataset Preparation

Following successful image collection and labeling according to the established protocol, we structured the dataset systematically:

- All images were consolidated into a centralized directory structure for easy accessibility.
- A CSV file was created, clearly linking each image (via a unique image ID) with its respective labels for each classification task, as assigned by the coders.

  This later file included details such as the **image ID**, an unique identifier for each collected image and its **assigned labels**, the specific classes allocated across all four predefined categories.

This meticulous and structured approach in data handling and labeling ensures that the final dataset is robust, consistent, and reliable, forming a solid foundation for accurate model training and meaningful analysis in subsequent stages of our project.

7 Image Classifier Development

This section describes the design, customization, and fine-tuning of our multitask image classification system. Given the nuances of our classification tasks and the limited amount of labeled data available, we worked on a flexible framework that incorporates several state-of-the-art pretrained models (Convolutional neural network, CNN). This section outlines the key design choices and justifies the selection of these models as robust feature extractors for our project.

7.1 Pretrained Model Selection

For this approach, a custom backbone module has been implemented that allows the selection among four widely recognized pre-trained architectures:

DenseNet121: Characterized by dense connectivity between layers, DenseNet121 helps in better gradient flow and feature reuse. This architecture is especially advantageous when training data is limited because its design facilitates learning more diversified features with fewer parameters.

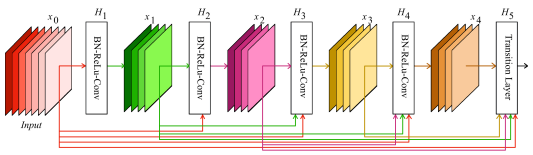


Figure 12: DenseNet Architecture

ResNet18: A lighter variant of the ResNet family, ResNet18 offers a balance between computational efficiency and representational power. Its reduced complexity makes it suitable for projects with scarce data, as it mitigates the risk of overfitting while still capturing essential visual patterns.

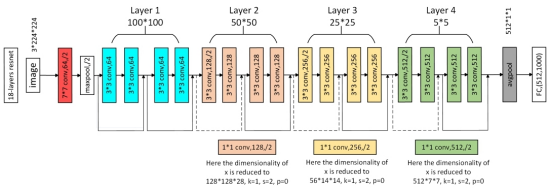


Figure 13: ResNet18 Architecture

EfficientNetB0: EfficientNetB0 is known for its compound scaling method, which optimizes depth, width, and resolution in a balanced manner. This results in a model that is both resource-efficient and effective at extracting salient features, making it a strong candidate when leveraging limited datasets.

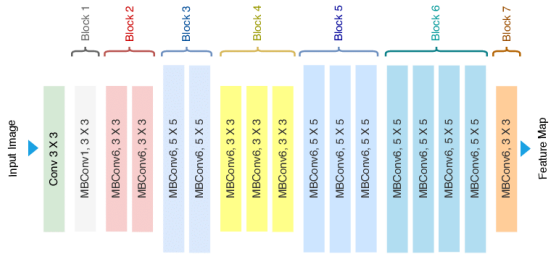


Figure 14: EfficientNetB0 Architecture

ResNet50 (Default): Although deeper and more parameter-intensive, ResNet50's robust architecture—utilizing residual connections—has proven successful for large-scale image recognition tasks. Its high-dimensional feature representations can be beneficial in capturing subtle nuances; however, care must be taken to avoid overfitting with scarce data.

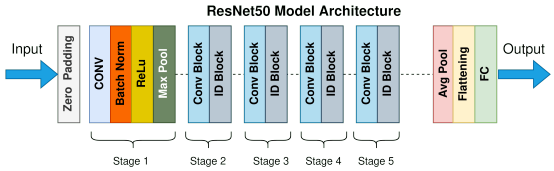


Figure 15: ResNet50 Architecture

All these models come pretrained on ImageNet dataset, ensuring that the learned features are general enough to transfer effectively to our target tasks. By comparing these options, our workflow facilitates empirical investigation of how the choice of backbone affects not only the classification performance but also downstream interpretability analyses.

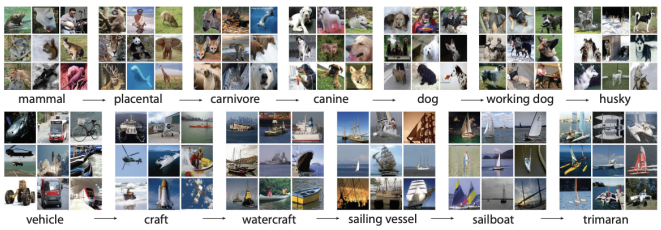


Figure 16: Imagenet

7.2 Dataset Preparation

As we have seen previously, the dataset for training and evaluating our model was created from high-quality images collected from the social media platform X.

All labels were standardized and mapped into numerical format to be compatible with neural network training. A structured data loader was created to efficiently handle and preprocess images, applying standard transformations such as resizing to 224x224 pixels, normalization, and conversion to tensor format.

7.3 Multitask Classifier Architecture

The network separates feature extraction from task-specific decisions to handle four simultaneous classification tasks (nature vs. non-nature, materiality, biological content, landscape type):

1. **Feature Extraction:** Remove the final layer of a chosen backbone, letting it output a high-dimensional feature vector (512, 1024, 1280, or 2048 dims, respectively).

2. **Task-Specific Classifier Heads:** From this shared feature vector, we branch into four fully connected heads: *Nature Visual Head*, *Materiality Head*, *Biological Content Head*, *Landscape Type Head*.

By decoupling these tasks, the model can learn task-specific decision boundaries from a shared feature space, boosting performance in data-scarce settings.

7.4 Model Training, Fine-Tuning, and Comparative Evaluation

To make the most of limited data, we fine-tune each multitask backbone end-to-end using the Adam optimizer with an adaptive learning rate and a composite, task-balanced cross-entropy loss. Extensive data augmentation (random cropping, flips, color jitter) combats overfitting, while the weighted loss ensures no single task dominates training.

We split the data 90/10 into training and test sets and we track task-specific accuracy, loss curves, and confusion matrices to compare architectures quantitatively.

8 Results of Image Classifier

In this section, we present the quantitative outcomes obtained from fine-tuning four pretrained backbones (ResNet18, ResNet50, EfficientNetB0 and DenseNet121) on our multitask classification problems. We report the final test loss as well as task-specific accuracies for each model.

Model	Test Loss	Nature (%)	Materiality (%)	Biological (%)	Landscape (%)
ResNet18	5.2291	69.46	67.07	68.86	61.08
ResNet50	4.2544	79.04	74.85	77.25	64.07
EfficientNetB0	3.9973	77.25	76.65	75.45	66.47
DenseNet121	3.8308	80.84	78.44	79.04	56.89

Table 1: Test Loss and Task Accuracies for Different Pretrained Backbones

Table 1 highlights a clear efficiency–accuracy spectrum across our four backbones. ResNet18, the lightest and fastest, achieves moderate performance, making it suitable for quick prototyping or edge deployment but less reliable on fine-grained tasks.

ResNet50 consistently outperforms ResNet18 by leveraging a richer feature space—especially boosting binary classification by 10 pp.

EfficientNetB0 strikes the best balance: it matches or exceeds ResNet50 on binary tasks, and leads on the multiclass landscape task thanks to its compound scaling with fewer parameters.

DenseNet121 achieves the lowest loss and highest accuracies on three of four tasks due to its dense connectivity and feature reuse, though it underperforms on the landscape task—suggesting that even high-capacity models may require targeted augmentation for fine-grained categories.

Overall, for purely binary distinctions (nature, materiality, biological) the ranking is *DenseNet121* > *ResNet50* > *EfficientNetB0* > *ResNet18*, whereas for the multiclass landscape task it is *EfficientNetB0* < *ResNet50* < *ResNet18* < *DenseNet121*.

8.1 Backbone Selection

To choose the most suitable model for applying interpretability methods, we evaluate not only average performance but also stability across tasks and efficiency (inference time and complexity). Below are the mean accuracy and weighted F1 scores for each model:

Model	Mean Acc. (%)	Mean F1
ResNet18	66.12	0.660
ResNet50	73.30	0.7375
EfficientNetB0	73.46	0.735
DenseNet121	73.80	0.725

Table 2: Mean accuracy and weighted F1 by model

As it can be seen in Table 2, ResNet18 was discarded for poor performance, while ResNet50 and EfficientNetB0 achieved similar mean accuracies (73.3 % vs. 73.5 %) and F1 scores. However, EfficientNetB0 offers faster inference, lower computational cost, and the highest accuracy on the challenging landscape-type task (66.5 %), whereas DenseNet121 despite its top mean accuracy (73.8 %), falls to 56.9 % on that task.

After this analysis, it has been chosed **EfficientNetB0** as the base model for interpretability techniques, as it balances overall accuracy, task stability, and computational efficiency, enabling agile and consistent results.

9 Interpretability Techniques

In this chapter, we present and discuss only those techniques that yielded meaningful and interpretable results. While several additional methods were implemented during the course of this study, their outputs did not meet the criteria for meaningful analysis and are therefore excluded from the main text. Detailed descriptions of these supplementary techniques and observed outcomes, can be found in Appendix A.

9.1 GradCAM

GradCAM (Gradient-weighted Class Activation Mapping) is a post hoc interpretability method designed to produce coarse localization maps, highlighting which regions of an input image are most influential for a particular class prediction. By combining the spatial information contained in deep convolutional feature maps with the gradient signal flowing back from the class score, GradCAM produces a heatmap that can be overlaid on the original image to visually explain model decisions.

Implementation

We integrate Grad-CAM into our multi-task classifier by wrapping the full network in a lightweight `SingleOutputWrapper` exposing only the logits for the target task, then registering two hooks on the chosen convolutional layer (`target_layer`): a forward hook to save its activations A, and a backward hook to capture the gradients $\partial y^c/\partial A$. To generate a Grad-CAM map for class c, we do a forward pass to get the logit y^c, call `loss.backward()` on it, and from the backward hook compute channel weights

$$\alpha_k = \frac{1}{Z}\sum_{i,j}\frac{\partial y^c}{\partial A_k^{ij}}.$$

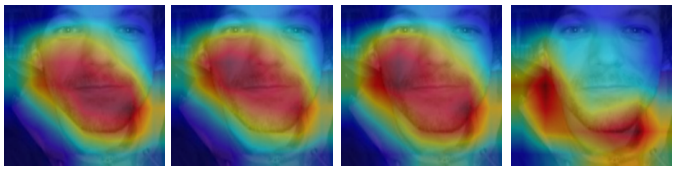
We then form

$$L_{\text{GradCAM}} = \text{ReLU}\Big(\sum_k \alpha_k \, A_k\Big),$$

normalize to $[0, 1]$, resize to the input resolution, and blend with the de-normalized image via a colormap and alpha blending. This hook-based design leaves the core training/inference loop untouched and works for any convolutional layer without modifying the original model code.

Results on GradCAM

Figure 17: Original image



(a) Nature (b) Biological (c) Materiality (d) Landscape

Figure 18: GradCAM results

The four Grad-CAM maps reveal distinct attention patterns by task: for the three binary classifiers (Nature, Biological, Materiality), all high-intensity regions align on the person's face—skin texture, hair contours and other facial features—showing the model associates a human subject with "nature" and with living versus material elements. By contrast, the Landscape Type head shifts to background structures and sky, using contextual spatial patterns.

First, the overlap across binary tasks suggests feature entanglement—converging on the same facial pattern—which may limit finer distinctions. Second, the landscape heatmap, though correctly oriented, is diffuse: key elements like trees or water lack crisp localization. Incorporating hierarchical attention or dedicated focus layers could both disentangle overlapping features in the binary heads and sharpen spatial precision in the landscape head.

Grad-CAM effectively explains each head's semantic focus and points to two improvement paths: better feature separation for binary concepts and enhanced background localization for landscape classification, boosting interpretability and performance.

9.2 LIME

Local Interpretable Model-agnostic Explanations (LIME) is a perturbation-based method that provides localized, human-interpretable explanations for any black-box classifier. For images, LIME segments an input into superpixels and learns a sparse linear model that approximates the classifier's behavior in the vicinity of a particular instance. The resulting explanation highlights which superpixels most positively influence the predicted class.

Implementation

We encapsulate our multitask network (or a single-output head via `SingleOutputWrapper`) in a `LIMEExplainer`, which exposes a `predict_fn` that takes NumPy arrays of perturbed images, applies our standard preprocessing (resize, normalize) and returns class probabilities for the target head. Given an input image, we first segment it into M superpixels (via SLIC or Quickshift), then generate N perturbed samples by randomly "hiding" subsets of these superpixels (replacing them with a constant hide-color) and recording the model's predictions.

We fit a weighted linear regression—weighting each sample by its ℓ_2 proximity to the original image—on the binary indicators of superpixel presence versus the predicted probability of the target class. The top k superpixels (those with the largest positive coefficients) are combined into a binary mask highlighting the regions most supportive of the prediction. Finally, we optionally refine the mask with morphological opening/closing, map it to a colormap (e.g. `JET`) and save the result.

Results on LIME

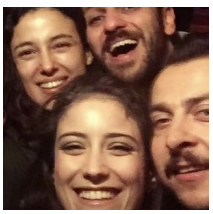


Figure 19: Original image

(a) Nature (b) Biological (c) Materiality (d) Landscape

Figure 20: LIME results

The LIME explanations for all four heads show that localized regions drive each prediction. For "Nature", highlighted superpixels fall on faces—skin, contours and hair—indicating the model equates human subjects with "nature" in a social-media context. The "Biological" head selects nearly the same facial regions and in "Materiality", the mask again centers on the face but spreads slightly to clothing and hair highlights, showing that textural cues also inform the decision. By contrast, "Landscape Type" shifts emphasis to broader patches in the upper corners (hints of structures or ambient light) so the model uses scene context for its decision.

In sum, LIME offers intuitive, human-readable explanations—confirming reliance on faces for binary tasks and on environmental cues for landscape classification. However, its superpixel granularity can be noisy or overlapping, pointing to improvements like adaptive segmentation or region-merging to sharpen interpretability.

9.3 Occlusion Sensitivity

Occlusion Sensitivity is a straightforward perturbation-based technique that measures how the classifier's output changes when parts of the input image are systematically "occluded" (replaced with a baseline value). By sliding a fixed-size patch over the image and recording the drop in the predicted class score, we obtain an importance map indicating which regions are most critical to the model's decision.

Implementation

We encapsulate our occlusion-sensitivity method in an `OcclusionSensitivity` class by first wrapping the multitask network in `SingleOutputWrapper` for the target task, then iterating over a grid of patches defined by `patch_size` and `stride`. We compute the original class logit y^c once; for each patch location (x, y), we create an occluded image $x^{\setminus p}$ by filling the `patch_size` $\times$ `patch_size` block at (x, y) with a constant baseline (e.g. 0), re-apply the normalization transform, and forward it to obtain $y^c_{\setminus p}$. The local importance score for patch p is

$$\Delta_p \;=\; y^c \;-\; y^c_{\setminus p}.$$

We then average overlapping Δ_p contributions for each pixel to produce a smooth heatmap, clamp negatives to zero, normalize to $[0, 1]$, resize to the original image resolution, apply a JET colormap overlay, and save the blended result.

Results on Occlusion Sensitivity

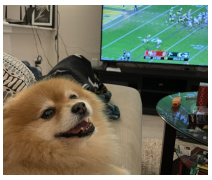
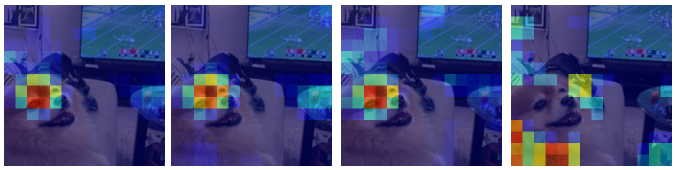


Figure 21: Original image



(a) Nature (b) Biological (c) Materiality (d) Landscape

Figure 22: Occlusion Sensitivity results

The occlusion maps in Figure 22 reveal that masking the dog's face causes the largest confidence drop for the Nature and Biological heads—confirming they rely on the central animal. In Materiality, the most important patch spans the fur and muzzle, but a secondary hotspot on the tabletop edge indicates the model also uses material context (soft fur vs. hard surface) to distinguish materials. Finally, Landscape Type shifts attention to the background TV and surroundings, showing dependence on environmental context.

These results validate occlusion sensitivity as a perturbation-based check where binary heads focus sharply on the animal's face, the landscape head on scene cues and suggest that smaller or adaptive patches could produce crisper importance maps for finer structures.

9.4 XRAI

XRAI (eXplanation with Ranked Area Integrals) builds on Integrated Gradients to produce region-based attributions. Instead of attributing importance pixel by pixel, XRAI first segments the input

into meaningful regions (e.g. via Felzenszwalb's algorithm), computes integrated gradients for each pixel, and then aggregates and ranks regions by their average attribution. This yields a smooth, region-focused heatmap that highlights cohesive areas most responsible for the model's decision.

Implementation

For each task, we wrap the multi-task network in `SingleOutputWrapper` to isolate the target head and instantiate `IntegratedGradients` on this wrapper. Given an input x, we compute attributions

$$\texttt{atts} = \text{IG.}\texttt{attribute}\big(x,\ \texttt{baseline} = 0,\ \texttt{target} = c,\ n\_steps = S\big)\,. \tag{1}$$

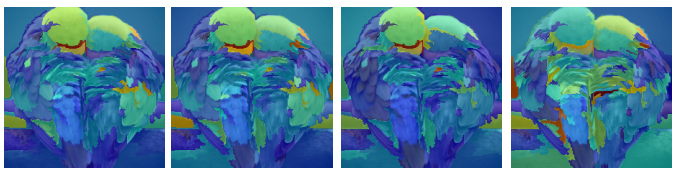
sum and clamp to positive values across channels to form a raw pixel map, and normalize it to $[0, 1]$. We then segment the de-normalized RGB image using Felzenszwalb's method into regions R, compute each region's mean attribution

$$A_R = \frac{1}{|R|} \sum_{(i,j)\in R} \max\big(0, \text{IG}(i, j)\big),$$

fill all pixels in R with A_R, re-normalize to $[0, 1]$, resize to the original resolution, apply a colormap (e.g. JET), and blend with the original image at opacity α. This combines the path-integral approximation of gradients from a zero baseline to the input, region segmentation, aggregation, and final visualization in one streamlined workflow.

Results on XRAI

Figure 23: Original image



(a) Nature (b) Biological (c) Materiality (d) Landscape

Figure 24: XRAI results

Figure 24 shows XRAI's region-focused saliency maps. For Nature, the entire bird silhouette is highlighted as one cohesive region—unlike pixel-level methods—linking animate form to "natural" content; the Biological mask is nearly identical, confirming both heads use the same cues. In Materiality, XRAI emphasizes the bird's plumage plus nearby branches and perch, indicating that texture (feathers) and context (wood grain) inform material decisions.

For Landscape Type, attention shifts to the corners of the image, showing reliance on environmental patterns. However, small spurious edge regions and a noisy landscape boundary suggest overfragmentation, and the near-identical Nature and Biological maps point to a need for task-specific parameter tuning to disentangle correlated concepts.

10 Comparison on Interpretability Techniques

Future work: Apply each explanation method to the same (or to several identical) image(s) and compare their outputs side by side, in order to determine for which types of images each technique performs best.

11 Conclusions

APPENDIX

1 Relevant Projects in Nature and Social Media Imagery

ProtoPNet for Fine-Grained Classification

The Prototypical Part Network (ProtoPNet) (Chen et al., 2019) introduced interpretability into fine-grained image classification (e.g., bird species) by learning prototypical image patches. ProtoPNet provides explanations such as "this image resembles this prototypical part," maintaining accuracy comparable to traditional CNNs while providing transparent, case-based reasoning essential for ecological research.

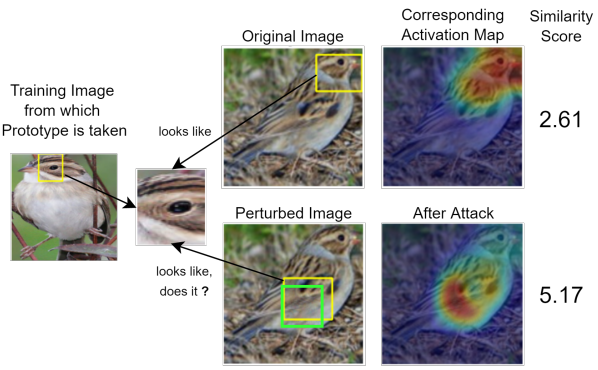


Figure 25: ProtoPNet. Adapted from (Chen et al., 2019).

Bird Species Classification with XAI

Kumar and Kondaveeti (2024) developed a pipeline for bird species classification using transfer learning and LIME. Their study emphasized that accuracy alone doesn't guarantee interpretability, as high-performing models could rely on background features. By quantitatively evaluating explanations with Intersection-over-Union (IoU), they showed that EfficientNet-B0 had superior interpretability alongside high accuracy, demonstrating the critical role of interpretability in ecological research.

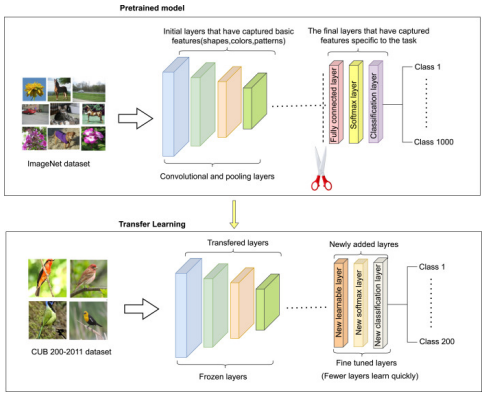


Figure 26: Model scheme. Adapted from Kumar and Kondaveeti (2024).

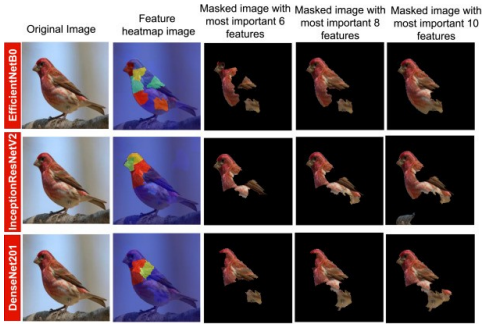


Figure 27: IoU results. Adapted from Kumar and Kondaveeti (2024).

Explainable Face Recognition (XFR)

The XFR project (Williford, May, & Byrne, 2020) provided a rigorous evaluation framework for face recognition interpretability. Through controlled face-triplet experiments, they benchmarked interpretability techniques like subtree Excitation Back-Propagation and DISE, significantly improving accuracy in highlighting distinguishing facial features.

They tried to answer the following question: *"What image region of the probe example is most similar to the mate example and least similar to the non-mate one?*. However, their approach has critical implications for ethical and unbiased deployment of AI in social contexts.

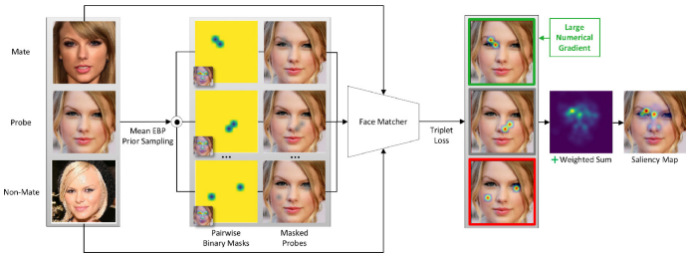


Figure 28: XFR. Adapted from (Williford, May, & Byrne, 2020).

Integrating interpretability into CNN-based image classification significantly enhances transparency and trustworthiness, especially critical for domains like ecological monitoring and social media analysis.

Techniques such as Grad-CAM, LIME, and SHAP provide complementary insights into model behavior, highlighting the importance of combining multiple interpretability methods for robust explanations. Projects like ProtoPNet and Concept Bottlenecks underline that interpretability need not compromise accuracy, while XFR emphasize interpretability's role in identifying biases and ensuring ethical AI deployments. Collectively, these studies affirm that combining accurate models with effective interpretability is essential for deploying AI systems responsibly in sensitive domains.

1.1 Guided GradCAM

Guided GradCAM fuses GradCAM's coarse "where"—localizing important regions at the feature-map level—with guided backpropagation's fine "what"—propagating only positive gradients through ReLUs to preserve edges and textures—yielding class-specific saliency maps that are both spatially precise and richly detailed.

Implementation

We extend our single-output wrapper to expose task-specific logits and register two sets of hooks: (1) on the target convolutional layer (e.g. `backbone.backbone.layer4`) to cache forward activations A_k and backward gradients $\partial y^c/\partial A_k$, and (2) on every ReLU to save its forward output and, in the backward pass, clamp negative gradients and mask them by the cached positive activations. At inference, for each image and its predicted class c, we:

1. **Grad-CAM mask:** backpropagate the target class score, average the spatial gradients to get channel weights α_k, then weight and sum the activations A_k with a ReLU:

$$M_{\text{GradCAM}} = \text{ReLU}\Big(\sum_k \alpha_k\, A_k\Big),$$

   followed by normalization and upsampling.

2. **Guided backpropagation:** run a second backward pass through the ReLU-hooked model to extract per-pixel gradients $G(x)$, zeroing out any negative contributions.

3. **Fusion:** element-wise multiply the upsampled M_{GradCAM} with each channel of $G(x)$ and sum across channels to get a single saliency map, then normalize.

4. **Overlay:** map the fused saliency to a colormap and blend it with the original image at adjustable opacity.

This hook-based design requires no changes to the core model and works with any pretrained backbone.

Results on Guided GradCAM

Figure 29: Original image

(a) Nature (b) Biological (c) Materiality (d) Landscape

Figure 30: Guided GradCAM results

1.2 SHAP

SHapley Additive exPlanations (SHAP) is a unified framework based on cooperative game theory that attributes model predictions to input features. In the image domain, SHAP computes per-pixel (or per-patch) contribution scores indicating how much each region "pushes" the predicted class score up or down relative to a baseline. This provides both local explanations for individual predictions and a consistent measure of feature importance.

Implementation

We use `shap.GradientExplainer` to approximate Shapley values for deep networks by first wrapping the multitask model in `SingleOutputWrapper` to isolate the logits of the target task and sampling a background set of B representative images (e.g. 50) from the training loader as our baseline. We initialize the explainer on the GPU (optionally enabling `local_smoothing`) and, for each test image x, call `explainer.shap_values` to obtain an array $\phi(c, i, j)$ of shape $C \times h \times w$. These values are computed by averaging output–input gradients over small perturbations around the background, yielding an approximate Shapley map for each channel c and spatial location (i, j). We then aggregate across channels via

$$M_{\text{SHAP}}(i, j) \;=\; \sum_{c} \big|\phi(c, i, j)\big|,$$

normalize $M_{\text{SHAP}}(i, j)$ to $[0, 1]$, resize it to the original image resolution, and map it to a divergent or sequential colormap (e.g. JET or PLASMA), where warm colors denote positive contributions and cool/dark colors denote neutral or negative influence. Finally, we overlay this heatmap on the denormalized RGB image at a chosen opacity α. In our refined workflow, we further apply morphological opening/closing to remove small artifacts and extract contours around the top-percentile regions for clearer visualization.

Results on SHAP

Figure 31: Original image

(a) Nature (b) Biological (c) Materiality (d) Landscape

Figure 32: SHAP results

At the time of writing, this interpretability technique remain under evaluation. The initial experiments did not yield sufficiently clear or expected outcomes (as it can be seen in the previous images) to include as definitive results. Consequently, I am currently exploring alternative configurations and additional post-processing strategies. Once more robust and interpretable outputs are obtained, these sections will be populated with detailed analyses and visualizations to complement the techniques already presented.

References

Barcelona Supercomputing Center. (2021). Fostering internet-based values of the environment [Defines Digital Relational Values (DRV) on social media].

Bengio, Y. (2012, July). Deep learning of representations for unsupervised and transfer learning. In I. Guyon, G. Dror, V. Lemaire, G. Taylor, & D. Silver (Eds.), *Proceedings of icml workshop on unsupervised and transfer learning* (pp. 17–36, Vol. 27). PMLR. https://doi.org/10.5555/3044805.3044807

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Kapishnikov, A., Sipper, M., Ishay, M., & Keller, Y. (2019). Xrai: Better attributions through regions. *Proceedings of the IEEE International Conference on Computer Vision*, 2955–2964. https://doi.org/10.1109/ICCV.2019.00305

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105. https://doi.org/10.5555/2999134.2999257

Kumar, S. (2024). *Transfer learning and data augmentation in deep learning* [Accessed: 2025-05-18]. https://www.tredence.com/blog/transfer-learning-and-data-augmentation-in-deep-learning

Langemeyer, J., & Calcagni, F. (2022). Virtual spillover effects: What social media has to do with relational values and global environmental stewardship [Suggests social media can foster relational values and stewardship]. *Ecosystem Services*, *53*, 101400. https://doi.org/10.1016/j.ecoser.2021.101400

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774. https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, June). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In J. DeNero, M. Finlayson, & S. Reddy (Eds.), *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Demonstrations* (pp. 97–101). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-3020

Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *ICCV*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*. https://arxiv.org/abs/1409.1556

Soga, M., & Gaston, K. J. (2016). Extinction of experience: The loss of human–nature interactions [Introduces and reviews the "extinction of experience" concept]. *Frontiers in Ecology and the Environment*, *14*(2), 94–101. https://doi.org/10.1002/fee.1225

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Väisänen, T., Heikinheimo, V., Hiippala, T., & Toivonen, T. (2021). Exploring human–nature interactions in national parks with social media photographs and computer vision [Applied AI to Flickr photos to monitor public engagement with nature]. *Conservation Biology*, *35*(2), 424–436. https://doi.org/10.1111/cobi.13704

Weinstein, N., Przybylski, A. K., & Ryan, R. M. (2009). Can nature make us more caring? effects of immersion in nature on intrinsic as-

pirations and generosity [Experimental evidence that nature exposure shifts values toward prosocial behavior]. *Personality and Social Psychology Bulletin*, *35*(10), 1315–1329. https://doi.org/10.1177/0146167209341649

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, *8689*, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53