Machine Learning with Applications 20 (2025) 100653

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Building consistency in explanations: Harmonizing CNN attributions for satellite-based land cover classification

Timo T. Stomberg [a],*, Lennart A. Reißner [a], Martin G. Schultz [b,c], Ribana Roscher [a,d]

[a] Institute of Geodesy and Geoinformation, University of Bonn, Niebuhrstraße 1a, Bonn, 53113, Germany
[b] Jülich Supercomputing Centre, Forschungszentrum Jülich, Wilhelm-Johnen-Straße, Jülich, 52428, Germany
[c] Department of Computer Science, University of Cologne, Weyertal 86-90, Cologne, 50931, Germany
[d] Institute of Bio- and Geosciences, Forschungszentrum Jülich, Wilhelm-Johnen-Straße, Jülich, 52428, Germany

ARTICLE INFO

Keywords:
Explainable machine learning
Attribution methods
Feature representations
Feature space
Land cover
Satellite imagery

ABSTRACT

Explainable machine learning has gained substantial attention for its role in enhancing transparency and trust in computer vision applications. Attribution methods like Grad-CAM and occlusion sensitivity analysis are frequently used to identify how features contribute to predictions of neural networks. However, a key challenge is that different attribution methods often produce different outcomes undermining trust in their results. Furthermore, the unique characteristics of remote sensing imagery pose additional challenges for attribution interpretation: it primarily comprises continuous "stuff" classes rather than objects, exhibits fine-grained spatial variability, contains mixed pixels, is often multispectral, and exhibits spatially heterogeneity. To tackle this challenge, we present a novel methodology that harmonizes attributions, resulting in: 1. greater consistency across different attribution methods; 2. more meaningful explanations when validated against known segmentation ground truth; and 3. enhanced transparency and traceability. This is achieved by coherently linking feature representations to attributions derived from analyzing the training data, enabling direct attribution assignment to features in (unseen) images. We evaluate our methodology using two satellite-based land cover classification datasets, three convolutional neural network architectures, and nine attribution methods. Harmonizing attributions increases the Pearson correlation coefficient between different attribution methods by an average of 0.18 across all datasets, models, and methods; and improves the micro F1-score — a measure of accuracy — by 12%. We demonstrate that Grad-CAM attributions are inherently well-aligned with the features, whereas other gradient-based attribution methods exhibit significant noise, mitigated through harmonization. It further enhances the resolution of occlusion-based attribution maps and adjusts misleading explanations.

1. Introduction

Convolutional neural networks (CNNs) have revolutionized computer vision by effectively learning complex patterns from images, with significant applications in satellite imagery analysis for environmental monitoring. However, their lack of transparency in decision-making raises concerns, particularly when understanding predictions is crucial for building trust and ensuring responsible use. To address this, multiple explainable machine learning methods have been developed, such as Gradient-weighted Class Activation Mapping (Grad-CAM) by Selvaraju et al. (2020) and occlusion sensitivity analysis by Zeiler and Fergus (2014), which help to interpret CNN decisions by highlighting important features and pixels. Despite their usefulness, a significant challenge with current attribution methods is their inconsistency. Different techniques often yield differing explanations for the same prediction, undermining trust in their results and raising questions about which one is best suited for a given task. This creates a pressing need for harmonization techniques that ensure meaningful and reliable explanations across diverse attribution methods.

Unlike conventional computer vision tasks that focus on object-centric images with distinct entities such as cars, animals, or people ("things"), satellite-based land cover classification primarily deals with "stuff" classes — continuous, amorphous regions characterized by fine-grained textures such as vegetation, soil, or water. This is further complicated by mixed pixels, where multiple land cover types contribute to a single pixel's spectral signature. Additionally, remote sensing images exhibit high spectral dimensionality, with multiple bands beyond the

* Corresponding author.
E-mail addresses: timo.stomberg@uni-bonn.de (T.T. Stomberg), reissner@uni-bonn.de (L.A. Reißner), m.schultz@fz-juelich.de (M.G. Schultz), r.roscher@fz-juelich.de (R. Roscher).

https://doi.org/10.1016/j.mlwa.2025.100653
Received 11 February 2025; Received in revised form 20 March 2025; Accepted 3 April 2025
Available online 6 May 2025

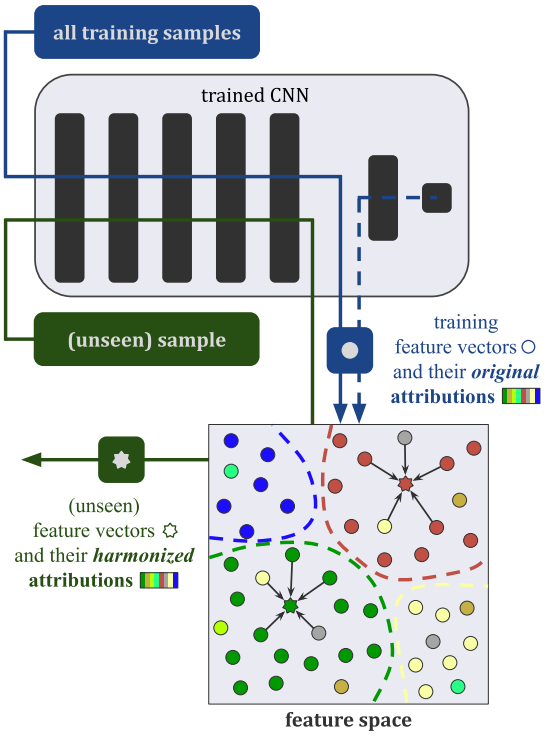
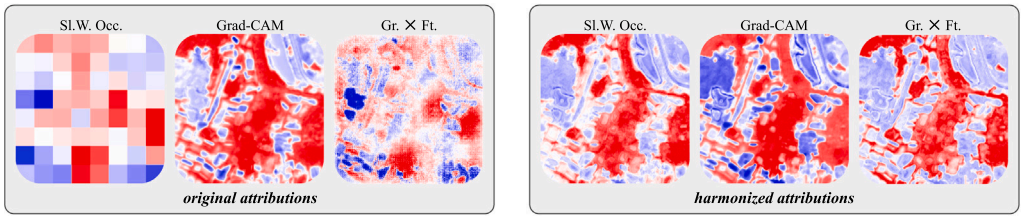
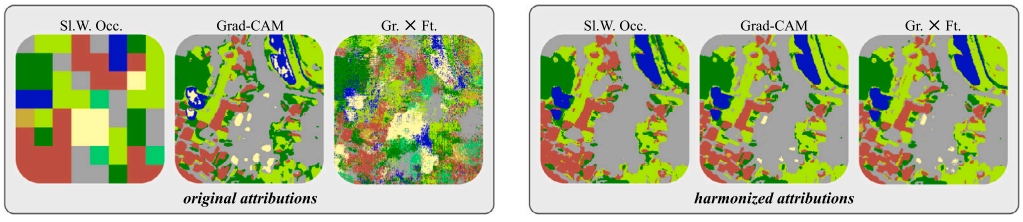


Fig. 1. Methodology. The feature vectors of all training samples' feature maps build the feature space. As similar feature vectors represent similar concepts, regions with comparable attributions emerge, as indicated by the dashed lines. The color represents the predominant class of an attribution vector (argmax). Using k-nearest neighbors, (unseen) feature vectors are assigned the average value of the surrounding attributions, resulting in more coherent and improved explanations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Attribution maps for the Urban/Built-up class ■ , with red indicating positive and blue indicating negative values. The color range for each attribution method is determined based on the 95th percentile of attribution values.



(b) Attribution maps showing the predominant classes of the attribution vectors (argmax). Table 1 provides the land cover legend.

Fig. 2. Original and harmonized attribution maps of the deep UH-Net layer for the DFC2020 dataset are compared across three attribution methods. The original attribution maps (left-hand side) show notable dissimilarities: Sliding Window Occlusions produce low-resolution maps due to the window size, while Gradients×Features introduces noise. Additionally, some explanations differ among the attribution methods. Harmonization (right-hand side) enhances resolution, reduces noise, and aligns the explanations. As a result, attribution maps become more consistent and similar across attribution methods. The corresponding satellite image and its ground truth are shown in Figure Fig. 3. Attribution maps for all models, attribution methods, and datasets are shown in Appendix B, Figures Fig. B.9, B.10, B.11. Abbreviations are provided at the end of the manuscript.

visible spectrum, such as near-infrared and shortwave infrared, providing rich yet complex feature information. Spatial heterogeneity further adds to the challenge, as land cover varies due to seasonal changes, diverse landscape structures, and atmospheric influences. The nature of remote sensing data makes it more challenging to derive meaningful explanations compared to conventional object-centric tasks. As features cannot rely on well-defined objects or edges, remote sensing features often form spatially extensive gradients with unstable and fragmented explanations. However, interpretability is crucial in this field, as it involves critical applications such as environmental monitoring, where decisions can significantly impact society and ecosystems (Alotaibi & Nassif, 2024). Further challenges of explainable machine learning and geospatial data are discussed by Xing and Sieber (2023).

To address the challenges of inconsistent and potentially uninformative attributions, we propose a novel methodology that harmonizes attributions by utilizing the model's learned feature space. Typically, attributions are computed for the features of *individual* samples. However, we argue that most feature representations and attributions generalize across *all* data and can therefore be aggregated using the training dataset. The harmonized attribution of an (unseen) feature is computed by the local average of attributions within the feature space, as illustrated in Fig. 1. Our approach is simple yet effective. It reduces noise in gradient-based attribution maps and enhances the resolution of occlusion-based ones. It further adjusts misleading explanations, defined as cases in which significant attributions are assigned to classes with low prediction scores. Overall, harmonization provides the following key contributions:

1. Greater consistency across different attribution methods making the choice of attribution method less critical (see Fig. 2).
2. More meaningful explanations when validated against known segmentation ground truth of unseen test data.
3. Enhanced transparency and traceability through a mechanistic and intuitive analysis of the feature space.

We evaluate our methodology using:

- Two datasets for land cover classification:
  - DFC2020 by Schmitt et al. (2019),
  - Ben-ge by Mommert et al. (2023);
- Three CNN architectures; each architecture is trained 10 times, and the results are averaged:
  - VGG-16 (Simonyan & Zisserman, 2015),
  - ResNet-18 (He et al., 2016),
  - UH-Net, an interpretable-by-design CNN based on Stomberg et al. (2021);
- Nine attribution methods (explained in Appendix A):
  - Sliding Window Occlusions (Zeiler & Fergus, 2014),
  - k-means Occlusions, based on Stomberg et al. (2023),
  - Grad-CAM (Selvaraju et al., 2020),
  - Gradients×Features (Shrikumar et al., 2017),
  - Layer-wise Relevance Propagation (Bach et al., 2015),
  - Integrated Gradients (Sundararajan et al., 2017),
  - DeepLift (Shrikumar et al., 2017),
  - Gradient-SHAP (Lundberg & Lee, 2017),
  - DeepLift-SHAP (Lundberg & Lee, 2017).

2. Related work and theoretical foundations

2.1. Feature representations

Research in mechanistic interpretability has contributed to a general understanding of how CNNs process image data (Carter et al., 2019; Olah et al., 2017, 2018). Each layer outputs a feature map in which the feature vectors ("pixels") encode increasingly complex properties of the input data. Starting with pixel colors, convolutions and poolings progressively expand the receptive field and allow the network to capture edges, textures, and, with deeper layers, more task-specific patterns.

In mechanistic interpretability, there is significant interest in examining the internal mechanisms of neural networks, focusing on two aspects: the features that represent learned concepts and the circuits — groups of neurons responsible for specific computations (Saphra & Wiegreffe, 2024). In contrast, attribution maps primarily aim to visualize individual input–output relationships without exploring the internal workings of neural networks. Our approach bridges these two fields.

2.2. Attribution methods

Prior research on explainable machine learning for remote sensing has primarily focused on feature attribution (Höhl et al., 2024). Attribution methods assign significance scores to features based on their impact on the model's prediction. The principles of attribution methods used in this work are explained in Appendix A. They can be divided into two fundamental groups: Perturbation-based attribution methods modify the features and evaluate resulting changes in the prediction. Gradient-based attribution methods involve the gradients of the model's prediction with respect to the features. However, these methods often exhibit significant variability, making it unclear which explanations are most reliable: Kakogeorgiou and Karantzalos (2021) evaluate their appearance and robustness on multi-label remote sensing benchmark datasets; Hsu and Li (2023) compare several qualitative abilities for geospatial datasets; Mohan and Peeples (2023) assay the robustness, faithfulness, randomization, complexity, localization, and axiomatic; and Nieradzik et al. (2024) assess their similarities with imagery from unmanned aerial vehicles. Some studies combine attribution methods to unite their respective advantages (Dhore et al., 2024; Gulum et al., 2021; Selvaraju et al., 2020).

Similar to our approach, Stomberg et al. (2023) harmonize attributions by employing a CNN layer's feature space. They are motivated by the issue that attribution values are incomparable between image patches, limiting large-scale mapping in remote sensing imagery. Their methodology is limited to low-dimensional feature spaces and is thus constrained to specific layers or architectures. We generalize their approach and evaluate how harmonization enhances the consistency among attribution methods and improves their explanations.

2.3. High-level semantics with input-level resolution.

Attribution maps can typically be computed at any layer. Some methods are originally designed for input-level analysis (e.g., Sliding Window Occlusions), while others target deeper layers (e.g., Grad-CAM). Each approach has distinct advantages: input-level analysis offers the highest resolution, whereas deep-layer analysis yields more meaningful results by capturing high-level semantics. Stomberg et al. (2021) present an interpretable-by-design architecture combining a U-Net (Ronneberger et al., 2015) and a classifier head, originally named "jUngle-Net" due to its application in a wilderness classification task. At the intermediate layer, attributions exhibit both high resolution and high-level semantics. Skip connections establish a strong link between deep and input features, ensuring that attributions remain closely tied to the input and are thus more interpretable. Our UH-Net architecture is based on their idea, visualized in Fig. 5, and described in Section 4.2.

3. Methodology

Our approach for harmonizing attributions is illustrated in Fig. 1. Specifically, we compute the feature maps for all training samples in a pre-selected layer. The feature map for layer l and training sample n, denoted by $F_{l,n}$, has dimensions $D_l \times H_l \times W_l$ representing the number of channels, height, and width, respectively. Across the training dataset with N samples, this yields N feature maps that form a set of feature vectors, expressed as $\{\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_{N\times H_l \times W_l}\}, \vec{f}_i \in \mathbb{R}^{D_l}$ and define the learned feature space. In the feature space, similar feature vectors — and thus similar feature representations — are positioned close together, while dissimilar ones are spaced apart. This proximity enables us to average the attributions from nearby feature vectors, forming the foundation of our harmonization approach.

Attributions. To study the relationship between feature maps and model predictions, we compute attribution maps for each feature map and class utilizing a desired attribution method. The attribution map for layer l and training sample n, denoted as $A_{l,n}$, has dimensions $C \times H_l \times W_l$ where C is the number of land cover classes. This approach diverges from the conventional procedure of attribution map computation in three ways: 1. Attributions are calculated for all classes, not just the predicted ones. 2. Negative attribution values are retained instead of being thresholded using ReLU, preserving their original state. 3 Attributions are left unnormalized instead of transforming them to the $[0, 1]$ range, preserving their original state. These choices ensure that the raw attribution values reflect the model's internal logic without introducing biases from post-processing. As feature maps and attribution maps share identical spatial dimensions, each feature vector $\vec{f}_i$ can be assigned directly to an attribution, resulting in a set of training attributions $\{\vec{a}_1, \vec{a}_2, \dots, \vec{a}_{N \times H_l \times W_l}\}, \vec{a}_i \in \mathbb{R}^C$.

Harmonized attributions. To predict the harmonized attributions for the mth (unseen) test sample, we first compute its feature map $F'_{l,m}$ with feature vectors $\{\vec{f}'_1, \vec{f}'_2, \dots, \vec{f}'_{H_l \times W_l}\}, \vec{f}'_i \in \mathbb{R}^{D_l}$. For each feature vector $\vec{f}'_i$, we identify the k nearest neighbors in the training feature space, $\{\vec{f}_1, \vec{f}_2, \dots, \vec{f}_{N \times H_l \times W_l}\}$ and the corresponding attributions, $\{\vec{a}_j, \dots\}$, are averaged. Averaging the attributions of nearby training feature vectors provides relevant information about the feature representations in that region of the feature space leading to more meaningful attributions.

4. Experiments

In this section, we provide a detailed explanation of the experimental procedures. The results are presented in Section 5. The methodology is tested on two datasets for land cover classification (Section 4.1), three CNN architectures (Section 4.2), and nine attribution methods (explained in Appendix A). Attribution computations are detailed in Sections 4.3 and 4.4; Section 4.5 outlines the procedure for evaluation.

4.1. Datasets

Both datasets used in our study — DFC2020 and Ben-ge — present unique challenges typical of remote sensing applications. DFC2020 includes imagery sampled from diverse geographic regions, meaning that models trained on this dataset must generalize across different climatic zones, vegetation types, and seasonal variability. Ben-ge, though detailed, suffers from label noise due to the reliance on automated classification products.

DFC2020. The DFC2020 dataset has been created for the 2020 IEEE GRSS Data Fusion Contest (Schmitt et al., 2019) and land cover classes have been semiautomatically annotated with a 10-meter resolution for 6114 patches with a size of 256 × 256 pixels each. The Sentinel-2 images have 13 spectral bands, with values ranging from 0 to 10,000, and are sampled from seven globally distributed regions. We run experiments for a random split among all samples and divide the image values by 10,000 so that they range from 0 to 1. The land cover classification scheme contains 10 classes from which 8 are present in the dataset, listed in Table 1. A sample is shown in Fig. 3.

Ben-ge (BigEarthNet). Ben-ge by Mommert et al. (2023) extends the BigEarthNet dataset by Sumbul et al. (2021) with geographical and environmental data. BigEarthNet contains about 590,000 Sentinel-2 images with 12 spectral bands and a size of 120 × 120 pixels each, with values ranging from 0 to 10,000. We reduce the dataset size by randomly selecting 20% of the images. Using the provided data split, this results in approximately 54,000 training samples, and 25,000 samples each for validation and testing. We divide the image values by 10,000 so that they range from 0 to 1. Ben-ge provides ESA WorldCover (Zanaga et al., 2022) for each Sentinel scene which is a global land cover product with a resolution of 10 meters and 11 land cover classes. Three land cover classes are nearly not covered by BigEarthNet (< 0.01%) and are ignored in our experiments. The remaining 8 classes and their frequencies are listed in Table 2. A sample is shown in Fig. 4.

Table 1
DFC2020 land cover classes and their relative areas within the DFC2020 dataset.

	Forest	23%
	Shrubland	6%
	Grassland	10%
	Wetland	5%
	Cropland	18%
	Urban/Built-up	10%
	Barren	3%
	Water	25%

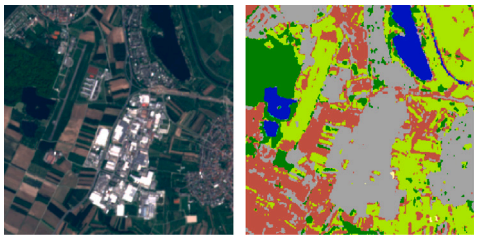


Fig. 3. DFC2020 sample. Sentinel-2 image and segmentation ground truth.

Table 2
ESA WorldCover classes and their relative areas within the Ben-ge dataset.

	Tree cover	41%
	Shrubland	1%
	Grassland	22%
	Cropland	15%
	Built-up	2%
	Bare/sparse vegetation	0.1%
	Permanent water bodies	17%
	Herbaceous wetland	0.5%

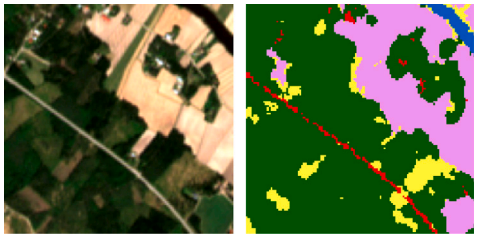


Fig. 4. Ben-ge sample. Sentinel-2 image and segmentation ground truth.

Image-wise, multi-class labels. We define image-wise multi-labels $\vec{y} \in \{0, 1\}^C$ for both datasets to train our CNNs. For this, we label a land cover class 1, if it covers more than 10% of the segmentation ground truth; and 0, if it covers less.

4.2. CNN architectures and training

We train the CNNs end-to-end on the given datasets with image-wise multi-labels $\vec{y} \in \{0, 1\}^C$ as described in Section 4.1. Each model uses sigmoid activation, binary cross-entropy loss, the AdamW optimizer, and a batch size of 32. The learning rate is linearly warmed up for 5 epochs and reduced by a factor of 10 if the validation loss stagnates for 5 epochs. The model state with the lowest validation loss is selected. The learning rates, weight decays, and epochs with the lowest validation loss are listed in Table 3 Each architecture is trained 10 times with different initializations and randomizations. The results are averaged. Training is done with PyTorch on an NVIDIA A100 40 GB PCIe GPU.

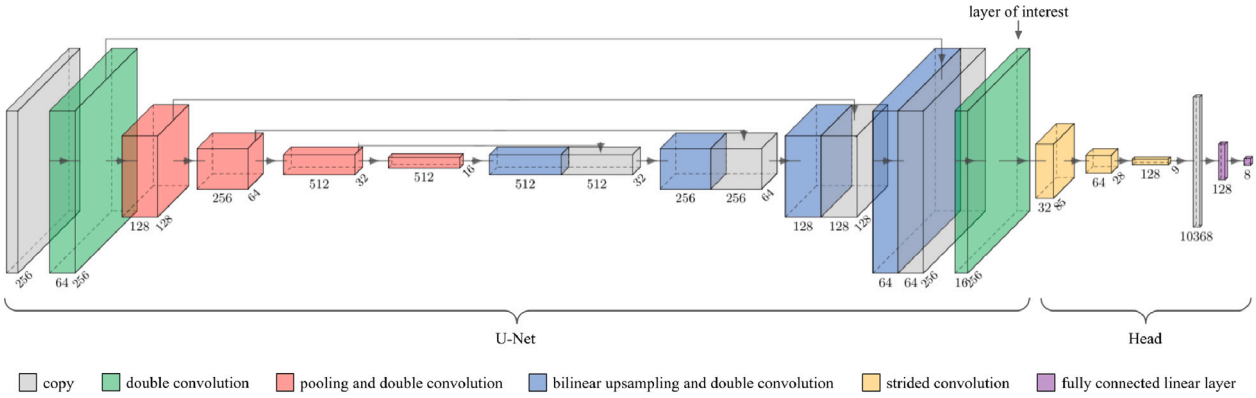


Fig. 5. UH-Net. The interpretable-by-design model combines a U-Net with a shallow classifier Head. The intermediate feature map matches the resolution of the input image and is of interest for computing the attribution maps. Shown is the architecture for the DFC2020 dataset, illustrated with PlotNeuralNet by Iqbal (2018).

Table 3
CNN training parameters, average number of epochs, and the metrics for the test datasets (in %). Params: Parameters; LR: Learning Rate; WD: Weight Decay; M: Million.

Dataset	Model	#Params	LR	WD	#Epochs	Accuracy	F1 *micro*	F1 *macro*
DFC2020	VGG-16	134.3 M	1e−5	1e−4	41 ± 7	95 ± 0	90 ± 0	87 ± 0
	ResNet-18	11.2 M	1e−4	1e−4	47 ± 8	95 ± 0	91 ± 0	87 ± 1
	UH-Net	19.0 M	1e−4	1e−4	35 ± 14	95 ± 0	89 ± 1	83 ± 1
Ben-ge	VGG-16	134.3 M	1e−6	1e−4	57 ± 16	96 ± 0	91 ± 0	77 ± 1
	ResNet-18	11.2 M	1e−6	1e−4	58 ± 9	96 ± 0	90 ± 0	71 ± 1
	UH-Net	17.7 M	1e−4	1e−4	36 ± 7	97 ± 0	94 ± 0	76 ± 1

VGG-16. We use a standard VGG-16, modified for the multi-channel input images and the number of targets, totaling 134.3 million parameters.

ResNet-18. We use a standard ResNet-18, modified for the multi-channel input images and the number of targets, totaling 11.2 million parameters.

UH-Net. We combine a modified U-Net with a shallow classifier Head resulting in a deep, hidden layer matching the resolution of the input image (see Fig. 5). The U-Net is adjusted with padding for size preservation, batch normalization, and bilinear upsampling as proposed by Odena et al. (2016). For Ben-ge, pooling is omitted in the first and third encoding steps due to the small image size. The U-Net's last layer is set to 16 channels with batch normalization but no activation function. The classifier head includes three strided convolutional layers and two linear layers. The UH-Net has 19.0 million parameters for DFC2020 and 17.7 million for Ben-ge.

Table 3 shows the performances of the trained CNNs by listing the accuracies and F1-scores of the test data. The F1-score is defined as the harmonic mean of precision and recall. When a metric is described as *micro*, it is computed globally, while *macro* indicates that the metric is calculated independently for each class and then averaged.

4.3. Attributions

We compute attributions utilizing nine attribution methods, namely Sliding Window Occlusions, k-means Occlusions, Grad-CAM, Gradients×Features, Integrated Gradients, Layer-wise Relevance Propagation, DeepLift, Gradient-SHAP, and DeepLift-SHAP. For each CNN, we independently compute the attributions for the input layer and a deep layer using the Captum library by Kokhlikyan et al. (2020).

Parameters. For Sliding Window occlusions, we cover a total of $8 \times 8 = 64$ non-overlapping, equally-sized squares. Similarly, in k-means Occlusions, we cover 64 clusters. For both occlusion methods, we set the occlusion value to zero. For Integrated Gradients, we use zeros as the baseline and approximate the integral over 10 steps using the Gauss–Legendre method. For Layer-wise Relevance Propagation, we apply the Epsilon Rule. For DeepLift, we use zeros as the baseline. For Gradient-SHAP, we involve 100 random baseline samples, and each run is repeated five times with a noise of the standard deviation scaled by 0.01 added to the input sample. For DeepLift-SHAP, we use 10 random baseline samples. As constituted in Section 3, we do *not* threshold attributions to positive values using ReLU and we do *not* normalize attributions.

VGG-16. VGG-16 has 13 convolutional layers from which we select the last convolutional layer with 512 channels. For DFC2020 with an input size of 256×256 pixels, its resolution is 16×16 feature vectors; for Ben-ge (120×120), it is 7×7. The layer is followed by a Max Pooling operation with a kernel size of 2×2. For the odd Ben-ge feature maps (7×7), this causes all attribution methods — except Grad-CAM — to assign zero attributions to the right column and bottom row. Additionally, due to the low resolution, Sliding Window and k-means Occlusions occlude each feature vector independently, yielding identical attribution maps.

ResNet-18. ResNet-18 consists of four basic blocks. We select the last layer of the second block due to its higher resolution compared to the layers in blocks three and four. The layer has a resolution of 32×32 for DFC2020 and 15×15 for Ben-ge, and 128 channels. We do not apply Layer-wise Relevance Propagation as originally no rules are defined for skip connections.

UH-Net. For the UH-Net, we select the last layer of the U-Net with 16 channels and the same resolution as the input image.

4.4. Harmonized attributions

We perform the following steps independently for each model, dataset, layer of interest (input and deep layer), and attribution method. First, we compute feature vectors and their attributions from the training dataset and select a representative random subset of approximately 100,000 training feature vectors for performance reasons. Next, we use the k-nearest neighbor regressor of the RAPIDS cuML Python library by Raschka et al. (2020) to find the $k = 100$ nearest neighbors of a vector. Since Euclidean distances become less meaningful in high-dimensional spaces (Aggarwal et al., 2001), we employ cosine similarity as our distance metric. (The selection of the number of training feature vectors is quite robust. Our findings suggest that it should be at least 10,000, but beyond this threshold, the results remain largely unchanged. Similarly, the choice of the number of nearest neighbors, k, is highly robust, though it should be at least 20. Additionally, we observe similar performance when using the Euclidean metric instead of cosine similarity.)

4.5. Evaluation

Similarity between attribution methods. To compare two attribution methods, we calculate the Pearson correlation coefficient for their attributions across the test dataset. The Pearson correlation coefficient is the covariance of two random variables A and B normed by the product of their standard deviations:

$$\mathrm{sim}_P = \frac{\mathrm{cov}(A, B)}{\sigma_A \sigma_B} \tag{1}$$

In our case, the random variables A and B represent attribution values from two sets of attribution vectors $\{\vec{a}_1, \ldots, \vec{a}_n\}$ and $\{\vec{b}_1, \ldots, \vec{b}_n\}$. The vectors $\vec{a}_i$ correspond to one attribution method, while the vectors $\vec{b}_i$ correspond to the other. The Pearson correlation coefficient can have values between -1 (negative correlation) and 1 (positive correlation), where zero indicates no correlation.

Segmentation ground truth. We compare the predominant class of the attribution vectors with the segmentation ground truth of the test data. We apply bilinear interpolation to the attribution maps if their sizes differ from the ground truth maps, which is required for VGG-16 and ResNet-18. As the segmentation classes of both datasets are imbalanced (see Tables 1, 2), we consider the F1-score in addition to the accuracy.

5. Results

This section presents the results obtained from the *deep* CNN layers. Results related to the inputs are presented separately in Appendix C. In the following, *original* attributions refer to the unmodified outputs of the attribution methods, whereas *harmonized* attributions are generated using our proposed approach.

5.1. Visual appearance

A selection of original and harmonized attribution maps is shown in Fig. 2. Attribution maps of all architectures, datasets, and attribution methods are shown in Appendix B, Figs. B.9, B.10, B.11. For visualizing the maps, a random model is selected from the set of ten.

Original attributions. Among the CNNs studied, VGG-16 has the lowest resolution, followed by ResNet-18, with UH-Net preserving the input resolution. The gradient-based attribution methods produce more noisy attribution maps, especially when applied to VGG-16 and ResNet-18 (Figs. B.9, B.10). For the UH-Net, cluster-like areas with reduced noise occur (Fig. B.11). Fig. 2(a) (Gradients×Features) reveals an uneven distribution of gradients across the Urban/Built-up class, likely due to the texture-based nature of remote sensing data. The extreme noise for ResNet-18 stems from strided convolutions (3×3 kernel with 2×2 stride) which include every second feature vector twice. This creates a checkerboard pattern within gradient maps and transfers to the gradient-based attribution maps. Grad-CAM, though gradient-based, is less noisy as gradients are averaged across channels. Sliding Window Occlusions result in a low resolution which is particularly noticeable for the UH-Net.

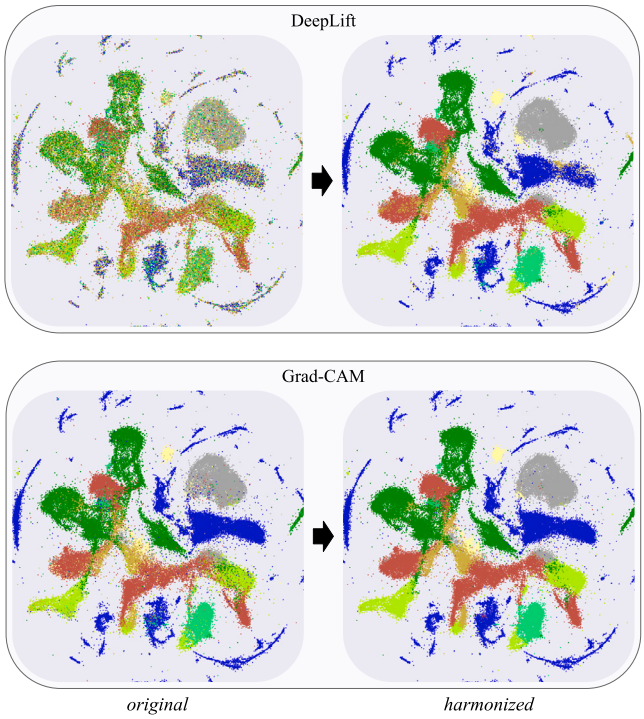


Fig. 6. Feature space of the deep UH-Net layer for the DFC2020 training dataset. The four illustrations differ only in the colorization of the feature vectors. On the left-hand sides, colors indicate the predominant classes (argmax) within the *original* attributions. The *harmonized* attributions (right) result in a smoother alignment. Grad-CAM attributions are already well-aligned before harmonization. For visualization purposes, the 16-dimensional feature vectors are reduced to two dimensions using UMAP (McInnes et al., 2020; Raschka et al., 2020). Table 1 provides the land cover legend.

Harmonized attributions. Harmonization mainly compensates for two aspects, both of which frequently appear in remote sensing: It reduces attribution noise, particularly for gradient-based methods; and it can adjust misleading explanation — significant attributions to classes with low prediction scores. For example, in Grad-CAM, Fig. B.11(b), initial attributions to *Shrubland* are reassigned to *Grassland*, aligning with the model's predictions and ground truth. For Sliding Window Occlusions, harmonization significantly improves the resolution, which is discussed in Section 5.2. Overall, harmonized attributions show greater similarity across attribution methods compared to the original attributions.

5.2. Feature space

Fig. 6 compares a feature space for various attribution methods. The original DeepLift shows noisy attribution-feature alignment. Harmonization smoothens this mapping and aligns DeepLift's attributions closer to Grad-CAM, which already exhibits strong attribution-feature alignment. For gradient-based methods like DeepLift, the noisy alignment is caused by noisy gradients. On the other hand, Sliding Window Occlusions result in maps with low resolution introducing noisy attributions within the feature space. Harmonization mitigates this, improving resolution within the image space (Fig. 2).

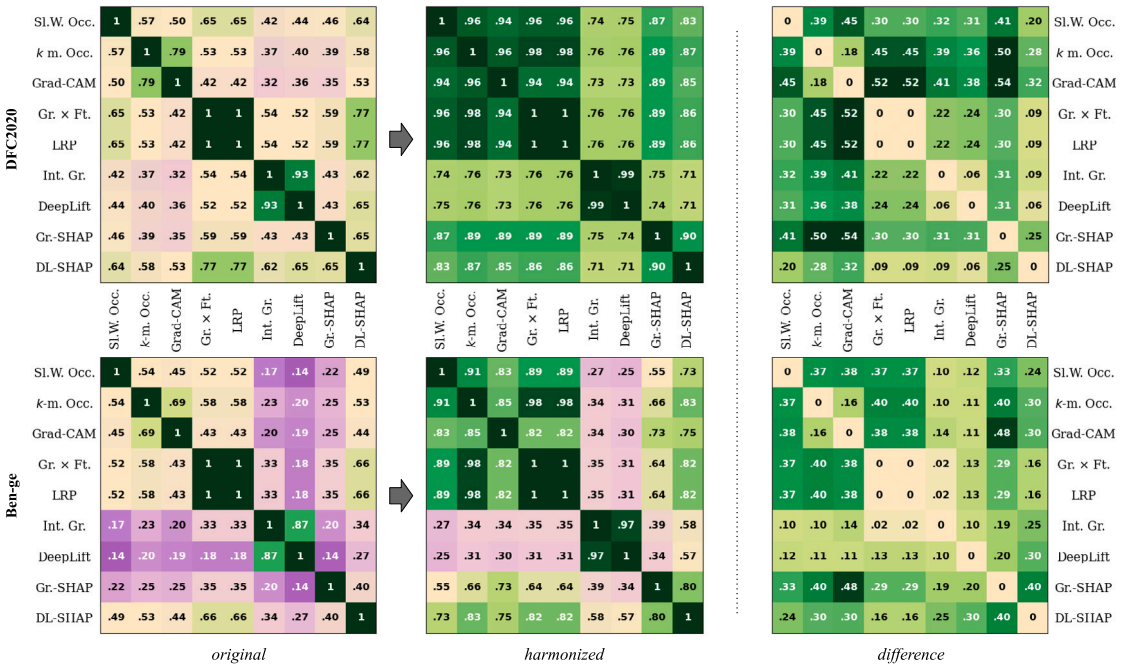


Fig. 7. Similarities between attribution methods for the deep UH-Net layer. Shown are the Pearson correlation coefficients for DFC2020 (top) and Ben-ge (bottom), averaged across ten models. From left to right: the similarities of the *original* attributions, the *harmonized* attributions, and the *difference* in similarities. Abbreviations are provided at the end of the manuscript.

5.3. Similarity between attribution methods

The Pearson correlation coefficients between attribution methods are shown in Fig. 7 (UH-Net), and Appendix B, Figs. B.12, B.13 (VGG-16, ResNet-18). Layer-wise Relevance Propagation is equivalent to Gradients×Features which is in line with Shrikumar et al. (2017) who state that these attribution methods are equivalent under the basic rule (z-Rule), provided all activations are piecewise linear. We further observe a high correlation between Integrated Gradients and DeepLift as stated by Ancona et al. (2018). They correlate with each other but have the weakest correlation with the other attribution methods.

Harmonization significantly increases the similarity between attribution methods, with only few exceptions. For VGG-16, we achieve an average increase in Pearson correlation of 0.12 across both datasets; for ResNet-18, the increase is 0.16; and for UH-Net, it is 0.27. Sliding Window Occlusions, k-means Occlusions, Grad-CAM, Gradients×Features, and Layer-wise Relevance Propagation share high similarities. Integrated Gradients and DeepLift exhibit the lowest similarity with other methods even after harmonization.

5.4. Segmentation ground truth

The evaluation metrics comparing the predominant attributions with the segmentation ground truth are listed in Table 4 (UH-Net), and Appendix B, Tables B.5, B.6 (VGG-16, ResNet-18). Through harmonization, all metrics improve (for all architectures and attribution methods), with a few exceptions where the metrics remain unchanged. The UH-Net shows a particularly strong improvement in the F1-scores. Assuming the models' predictions are rational, attribution methods aligning more closely with the segmentation ground truth are more likely to offer meaningful explanations. We thus conclude that harmonized attributions offer better explanations than the original attribution methods.

Integrated Gradients and DeepLift show poor agreement with the segmentation ground truth across all CNNs. For VGG-16 and UH-Net, Grad-CAM performs best comparing the original attribution methods. After harmonization, Grad-CAM, Gradients×Features, Layer-wise Relevance Propagation, Sliding Window Occlusions, and k-mean Occlusions show good results. For ResNet-18, DeepLift-SHAP performs best among the original attribution methods. DeepLift-SHAP and Gradient-SHAP perform best among the harmonized attributions.

5.5. Similarities between original and harmonized attributions

The Pearson correlation coefficients between the original and harmonized attributions are presented in Fig. 8, where a higher coefficient indicates greater similarity. Overall, Grad-CAM exhibits the highest correlation, suggesting that harmonizing its attributions has a relatively low effect. This implies that Grad-CAM attributions are already well-aligned with the feature vectors, as also evident by the feature space visualization in Fig. 6. In contrast, Gradient-SHAP shows the lowest correlation between original and harmonized attributions, revealing that harmonization has a stronger effect.

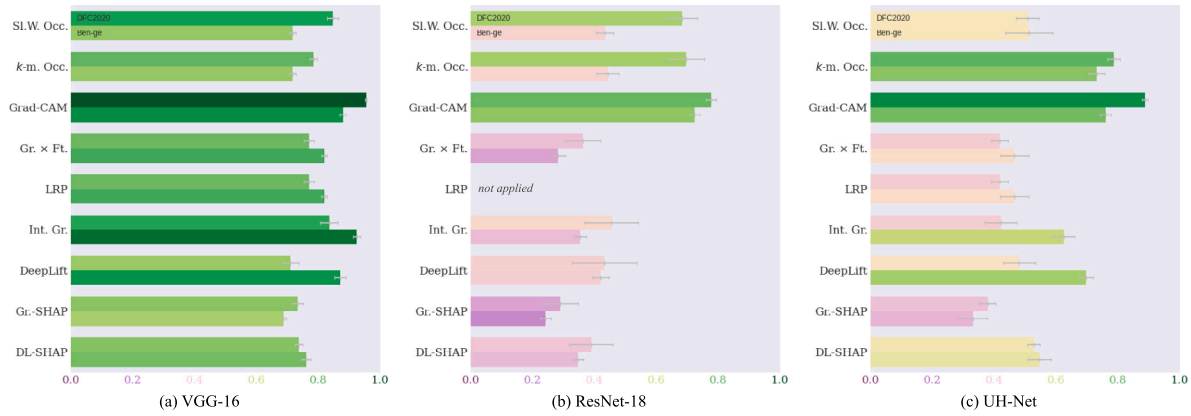
Comparing the CNNs, VGG-16 exhibits the highest correlations due to a relatively simple relationship between the analyzed layer and the prediction. The layer is not followed by further convolutions, but Max Pooling, Global Average Pooling, and three fully connected linear layers, resulting in a more straightforward alignment between feature vectors and original attributions.

6. Discussion

Our results demonstrate that harmonizing attributions across the learned feature space enhances the interpretation of CNN predictions. By averaging attributions inside the learned feature space, harmonization compensates for the inherent inconsistencies of individual attribution methods leading to more coherent and reliable explanations. This is evident by the examination of attribution maps, visualization of learned feature spaces, evaluation of similarities, and comparison of attributions with segmentation ground truth.

Table 4
Metrics comparing the predominant attribution class (original and harmonized) with the segmentation ground truth for the deep UH-Net layer (in %). Deviations in the differences may result from rounding. The best values in a dataset column, along with those up to 2% lower, are highlighted in bold.

Dataset	Attribution method	Accuracy			F1 (micro)			F1 (macro)		
		orig.	harm.	diff.	orig.	harm.	diff.	orig.	harm.	diff.
DFC2020	Sl.W. Occlusions	88 ± 1	**95** ± 0	8 ± 1	50 ± 3	**81** ± 0	30 ± 3	42 ± 3	**69** ± 1	27 ± 3
	k-means Occlusions	**94** ± 0	**95** ± 0	1 ± 0	**77** ± 1	**80** ± 0	4 ± 1	**65** ± 2	**69** ± 1	4 ± 1
	Grad-CAM	**95** ± 0	**95** ± 0	1 ± 0	**78** ± 1	**80** ± 1	2 ± 1	**67** ± 2	**69** ± 1	2 ± 1
	Gradients×Features	86 ± 1	**95** ± 0	9 ± 1	42 ± 3	**80** ± 0	37 ± 3	35 ± 2	**69** ± 1	33 ± 2
	LRP	86 ± 1	**95** ± 0	9 ± 1	42 ± 3	**80** ± 0	37 ± 3	35 ± 2	**69** ± 1	33 ± 2
	Integrated Gradients	84 ± 1	**93** ± 2	9 ± 1	35 ± 6	72 ± 7	37 ± 4	30 ± 4	62 ± 5	32 ± 3
	DeepLift	85 ± 2	**93** ± 2	8 ± 1	40 ± 6	73 ± 7	33 ± 4	34 ± 5	63 ± 5	29 ± 3
	Gradient-SHAP	84 ± 1	**94** ± 0	10 ± 0	36 ± 2	77 ± 1	42 ± 2	30 ± 2	66 ± 1	36 ± 2
	DeepLift-SHAP	89 ± 0	**93** ± 0	4 ± 1	56 ± 2	73 ± 2	17 ± 2	46 ± 2	64 ± 1	18 ± 2
Ben-ge	Sl.W. Occlusions	90 ± 2	**95** ± 0	5 ± 1	60 ± 6	**80** ± 2	20 ± 5	36 ± 4	**55** ± 1	19 ± 4
	k-means Occlusions	**93** ± 1	**95** ± 0	2 ± 0	**73** ± 2	**79** ± 1	6 ± 2	**46** ± 2	**54** ± 1	8 ± 1
	Grad-CAM	**94** ± 0	**95** ± 0	1 ± 0	**74** ± 1	**79** ± 1	5 ± 1	**47** ± 1	**54** ± 1	7 ± 0
	Gradients×Features	89 ± 1	**95** ± 0	6 ± 1	55 ± 4	**78** ± 1	23 ± 3	32 ± 2	**53** ± 1	21 ± 2
	LRP	89 ± 1	**95** ± 0	6 ± 1	55 ± 4	**78** ± 1	23 ± 3	32 ± 2	**53** ± 1	21 ± 2
	Integrated Gradients	85 ± 1	86 ± 2	2 ± 2	38 ± 4	46 ± 9	7 ± 6	26 ± 2	28 ± 5	2 ± 3
	DeepLift	85 ± 1	86 ± 3	1 ± 2	40 ± 5	44 ± 12	4 ± 7	26 ± 3	26 ± 6	0 ± 4
	Gradient-SHAP	83 ± 1	92 ± 1	9 ± 0	30 ± 3	67 ± 3	37 ± 2	20 ± 2	45 ± 2	25 ± 1
	DeepLift-SHAP	89 ± 1	**93** ± 0	4 ± 1	56 ± 2	72 ± 1	16 ± 2	38 ± 1	49 ± 1	11 ± 1



(a) VGG-16 (b) ResNet-18 (c) UH-Net

Fig. 8. Similarities between original and harmonized attributions. The upper bar of an attribution shows the Pearson correlation coefficient obtained for the DFC2020 dataset; the lower bar the one for Ben-ge. The colorization highlights the values also represented by the bars as shown on the *x*-axis. The values are averaged across ten models; the gray lines indicate the standard deviations. Abbreviations are provided at the end of the manuscript. We do not apply Layer-wise Relevance Propagation for the ResNet-18 as originally no rules are defined for skip connections. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Input vs. deep layer. Few attribution methods are typically applied to deeper layers, originating from their foundational articles. The inventors of Grad-CAM (Selvaraju et al., 2020) explicitly suggest an application to deeper layers, likely because Grad-CAM originates from Class Activation Mapping (CAM, Zhou et al., 2016), which can be applied to the last convolutional layer only. However, doing so seems generally reasonable from findings in mechanistic interpretability (Carter et al., 2019; Olah et al., 2017, 2018) which conclude that representations in deeper layers are more refined and closely aligned with the output (see also Section 2.1). In contrast, the input feature vectors (pixels) represent only the concept of (multispectral) color — not even structure. However, structure often plays a crucial role in differentiating land cover classes. For example, cropland and grassland may share similar spectral signatures but differ in spatial structure, requiring explanations to be both spatially and spectrally aware. Another challenge for applying attributions to the input is the high number of channels in multispectral remote sensing images, which can introduce increased noise. Based on our results, we conclude that applying attribution methods is generally more effective in deeper layers (Section 5 vs. Appendix C). Further below (*Limitations*), we also discuss why harmonization is not working effectively in the input.

UH-Net. Due to the drawback that deeper CNN layers usually have lower resolution, using the UH-Net architecture is advantageous in terms of interpretability. It enables meaningful attributions of high-level features at high resolution. The UH-Net demonstrates the highest concordance with the segmentation ground truth. It achieves similarly high correlations among the harmonized attribution methods as VGG-16, with both significantly outperforming ResNet-18.

Weakly-supervised segmentation. Attribution methods are regularly used for weakly-supervised segmentation, where image-wise but not pixel-wise labels are available (Ahn & Kwak, 2018; Chong et al., 2021; Kwak et al., 2017; Wang et al., 2020). Combining the UH-Net with our harmonization technique, we yield promising results with high accuracies that are competitive with current research. For instance, Hanna et al. (2023) use sparse multimodal vision transformers, obtaining an Intersection over Union (IoU) of 34% on the DFC2020 dataset. By employing the UH-Net and harmonized attributions, we achieve an IoU of 56% (based on a different data split).

In weakly-supervised segmentation, feature vectors are often actively disentangled to enhance feature representations for the task at hand. Approaches typically assume that pixels with similar colors and spatial connectivity belong to the same class. This improves various aspects, including the delineation of objects when applying attribution methods. Building on this, Jonnarth and Felsberg (2022) introduce a feature similarity loss; Kwak et al. (2017) propose a superpixel pooling layer that enforces uniform features within each superpixel determined utilizing the input image; and Zeng et al. (2023) present a Global Superpixel Consistency Module, ensuring that similar superpixels in the input image transfer to similar features in the penultimate feature map. Similar approaches could potentially be integrated with the UH-Net

and our harmonization framework. Additionally, addressing the class imbalance could improve macro metrics. Possible strategies include applying a weighted loss function or augmenting data samples from underrepresented classes.

Vision transformers and graph convolutional networks. We have harmonized the attributions within the learned feature space constructed from the feature vectors of the feature maps generated by the convolutional layers of CNNs. In a Vision Transformer (Dosovitskiy et al., 2021), an image is partitioned into patches, which are then processed by multi-head attention. Attention maps can be derived from the attention weights of a multi-head attention layer. Analogous to the *channels* in a convolutional layer, the *heads* in this context could serve to define attention vectors (instead of feature vectors). Harmonization would then be carried out in the attention space, following a similar logic. Alternatively, the features of a normalization layer in an encoder block could be used for harmonization. In a Graph Convolutional Network (Kipf & Welling, 2017), the input image is transformed into graph-structured data, such as by creating superpixels where each superpixel serves as a node, and the relationships between neighboring nodes are characterized by edges. The nodes are represented by feature vectors which pass through the network. Attributions and harmonized attributions can be computed similarly to the approach used for pixel-based CNNs in this study. A potential research question, which is however beyond the scope of this paper, is whether harmonization in Vision Transformers or Graph Convolutional Networks leads to similar effects as it does in CNNs.

Limitations. Harmonization is constrained by the degree of entanglement within the learned feature space as the separation of distinct feature representations is crucial. Similar feature vectors that represent different concepts or classes cannot be well harmonized. Entanglement is more frequent in the initial layers because they process low-level features, such as color or texture. When different land cover classes share similar colors, their attributions become mixed in the input space. In our land cover classification tasks, the high agreement between attributions and segmentation ground truth (Section 5.4) suggests that the deep feature vectors are disentangled with respect to the eight classes. However, as the number of classes increases and interactions between them grow, the likelihood of entanglement rises.

A domain shift between training and test data poses limitations not only for model predictions but also for the reliability of attributions. If a CNN encounters out-of-distribution input features, they may either be mapped to out-of-distribution deep features or mistakenly to in-distribution deep features. In both cases, the resulting features may be misattributed both by the original and the harmonized attributions.

Grad-CAM. Grad-CAM attributions are most aligned with the deep feature vectors among all attribution methods we evaluated in Section 5.5. Notably, Grad-CAM is one of the few gradient-based attribution methods passing the sanity checks by Adebayo et al. (2018). It further performs best under various evaluations by Yang and Kim (2019) who examine how the attributions of the same method vary under different conditions: two different models given the same input (model contrast score), two different inputs given the same model (input dependence rate), and two similar inputs given the same model (input independence rate). Grad-CAM is also among the best-performing attribution methods utilizing Most Relevant First (Samek et al., 2017) on CNNs trained for land cover classification (Kakogeorgiou & Karantzalos, 2021). Most Relevant First evaluates an explanation by observing how quickly the prediction score decreases as information is progressively removed. Further, Grad-CAM performs best when testing the robustness of explanations by perturbating the input in a semantically meaningful way (Yang & Kim, 2019).

A possible explanation for Grad-CAM's strong performance in such tests is its high alignment with feature vectors. These vectors are critical in a CNN's computation, as they encapsulate the learned representations driving the final decision. Another explanation is that, unlike most attribution methods, Grad-CAM is typically applied to deeper layers rather than the input level — also in tests described above. This approach leads to better explanations, as discussed in the following.

7. Conclusion

Interpretable results are crucial in remote sensing, as this field involves critical applications like environmental monitoring, where decisions can have significant societal and ecological impacts. Harmonizing attributions across the learned feature space enhances the interpretability of CNN predictions for land cover classification. By averaging attributions inside the learned feature space, our approach compensates for the inherent inconsistencies of individual attribution methods, leading to more coherent and reliable explanations. When combined with the interpretable-by-design UH-Net, harmonization opens up opportunities for transparent classifiers and significant enhancement of accuracies in weakly-supervised segmentation, while also contributing to advances in mechanistic interpretability.

CRediT authorship contribution statement

Timo T. Stomberg: Conceptualization, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Lennart A. Reißner:** Investigation. **Martin G. Schultz:** Supervision, Writing – review & editing. **Ribana Roscher:** Resources, Supervision, Writing – review & editing.

Abbreviations

CNN	Convolutional Neural Network
DL	DeepLift
Ft.	Features
Gr.	Gradients
Int.	Integrated
IoU	Intersection over Union
k-m.	*k*-means
LR	Learning Rate
LRP	Layer-wise Relevance Propagation
M	Million
Occ.	Occlusions
Params	Parameters
Sl.W.	Sliding Window
WD	Weight Decay

Code availability

The code for the presented methodology is available at https://gitlab.jsc.fz-juelich.de/kiste/harmon.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge funding from the Deutsche Forschungsgemeinschaft, Germany (DFG, German Research Foundation; RO 4839/7-1 | STO 1087/2-1, and EXC-2070—390732324—PhenoRob); the German Federal Ministry for the Environment, Nature Conservation, and Nuclear Safety (67KI2043, KISTE); and the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE).

Appendix A. Attribution methods

In the following terminology, the term *feature map* encompasses the term *input image*. Unless specified otherwise, *gradients* refer to the partial derivatives of the non-activated output with respect to the features. A *feature vector* includes the features at a certain position of a feature map across all channels; simply put, it is a "pixel" of a feature map.

Sliding Window Occlusions. By systematically covering sections of a feature map with a sliding window, one can observe how the model's prediction changes, revealing the attributions of the occluded regions (Zeiler & Fergus, 2014).

k-means Occlusions. We introduce a modified form of Activation Space Occlusion Sensitivities (ASOS) by Stomberg et al. (2023) which is more computationally efficient. We apply k-means (Lloyd, 1982) to cluster the feature map's feature vectors into N clusters. Each cluster is occluded separately and the attributions are divided by the number of occluded pixels. The fundamental idea is to simultaneously occlude similar features, assuming that similar features also share similar attributions.

Gradient-weighted Class Activation Mapping (Grad-CAM). The average gradients across each channel are multiplied with the corresponding channels of the feature map (Selvaraju et al., 2020). The gradients are computed applying Guided Backpropagation (Springenberg et al., 2015).

Gradients×Features. The gradients of each feature vector are multiplied with the corresponding feature vectors (Shrikumar et al., 2017).

Integrated Gradients. A baseline map is chosen, often set to zeros. The integrated gradients are computed by averaging the gradients as the features transition linearly from the baseline map to the original feature map. The integrated gradients are multiplied by the difference between the feature map and the baseline map (Sundararajan et al., 2017).

Layer-wise Relevance Propagation (LRP). The model's prediction is propagated back through the layers. The propagation is performed according to specific propagation rules that ensure the preservation of relevance within each layer throughout the network (Bach et al., 2015).

Deep Learning Important FeaTures (DeepLift). A baseline map is chosen, often set to zeros. The difference in all features in all layers is computed by forward passing the baseline map and the original feature map. Subsequently, consistent contribution scores are computed for each feature by backpropagating these differences via a specific rule (Shrikumar et al., 2017).

Gradient-SHAP. Gradient-SHAP combines concepts from SHapley Additive exPlanations (SHAP, Lundberg and Lee, 2017), Integrated Gradients (Sundararajan et al., 2017), and SmoothGrad (Smilkov et al., 2017). As baselines, a set of training samples is chosen. Noise is added to the baseline maps and the original feature map. The attribution is averaged over multiple iterations.

DeepLift-SHAP. DeepLift-SHAP combines concepts from SHAP (Lundberg & Lee, 2017) and DeepLift (Shrikumar et al., 2017). Similar to Gradient-SHAP, a set of training samples is chosen as baselines and the attribution is averaged.

Appendix B. Results of the deep layers

Visual appearance. In Section 5.1, Fig. 2, we show a selection of attribution maps for the UH-Net. Attribution maps for all models, attribution methods, and datasets are shown in Figs. B.9, B.10, B.11.

Similarity between attribution methods. In Section 5.3, Fig. 7, we show the Pearson correlation coefficients between attribution methods for the UH-Net. The Pearson correlation coefficients for VGG-16 and ResNet-18 are shown in Figs. B.12, B.13.

In the case of Ben-ge and VGG-16, two effects arise due to the low resolution of the analyzed layer (7×7). First, Sliding Window Occlusions and k-means Occlusions produce identical results because both methods occlude each feature vector individually. Second, Grad-CAM exhibits low similarity with the other methods because the feature maps have an odd size and are followed by a Max Pooling operation with a 2×2 kernel. This operation causes all attribution methods — except Grad-CAM — to assign zero attributions to the rightmost column and bottom row of the attribution maps, affecting 27 % of the feature vectors.

Segmentation ground truth. In Section 5.4, Table 4, we show the metrics comparing the predominant attribution class with the segmentation ground truth for the UH-Net. For VGG-16 and ResNet-18, this is shown in Tables B.5, B.6.

Appendix C. Results of the inputs

In Section 5, we focus on the results of the deep CNN layers. In the following, we also present our results for the inputs of the CNNs. We do not compute Layer-wise Relevance Propagation for the UH-Net's input as there are no rules defined for skip connections.

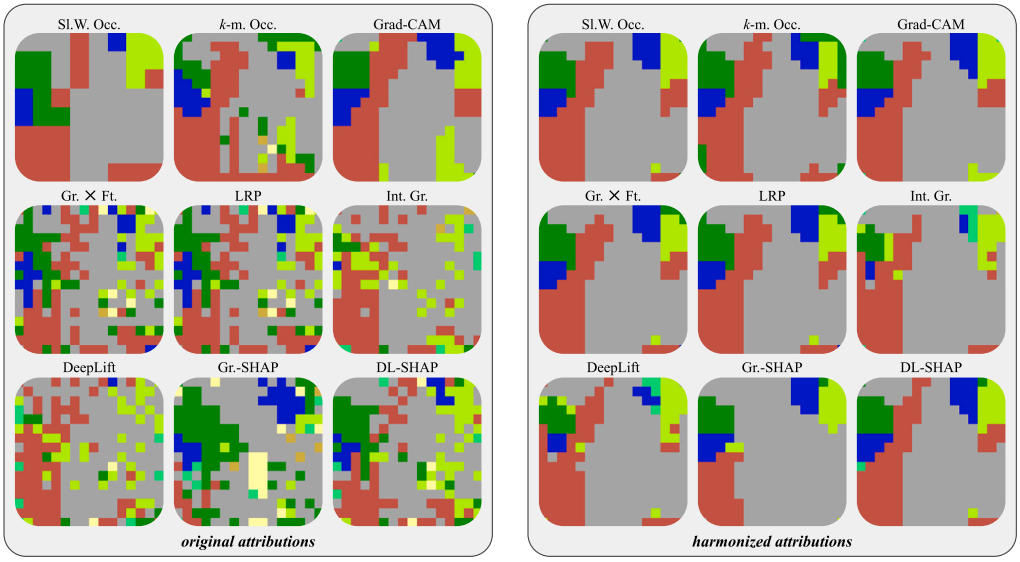
Visual appearance. A selection of original and harmonized attribution maps is shown in Fig. C.14. The gradient-based standard attribution maps (DeepLift-SHAP is shown as an example) are significantly more noisy than they are for the deep layers. Harmonization can compensate for this only to a limited extent. Large areas of the forest are misleadingly explained by Sliding-Window Occlusions and Grad-CAM. Harmonization does not compensate for this in most cases. For the UH-Net, the gradient-based attribution maps appear less noisy than for VGG-16 and ResNet-18. This is likely due to the skip connections within the U-Net, allowing the gradients to backpropagate more directly to the input.

Similarity between attribution methods. The Pearson correlation coefficients among all attribution methods are shown in Figs. C.15, C.16, C.17. Harmonizing the attributions often increases their similarity across the attribution methods, though the effect is typically minor compared to the overall high dissimilarity of attributions. Significant increases in similarity are observed only in a few cases, most notably with the UH-Net.

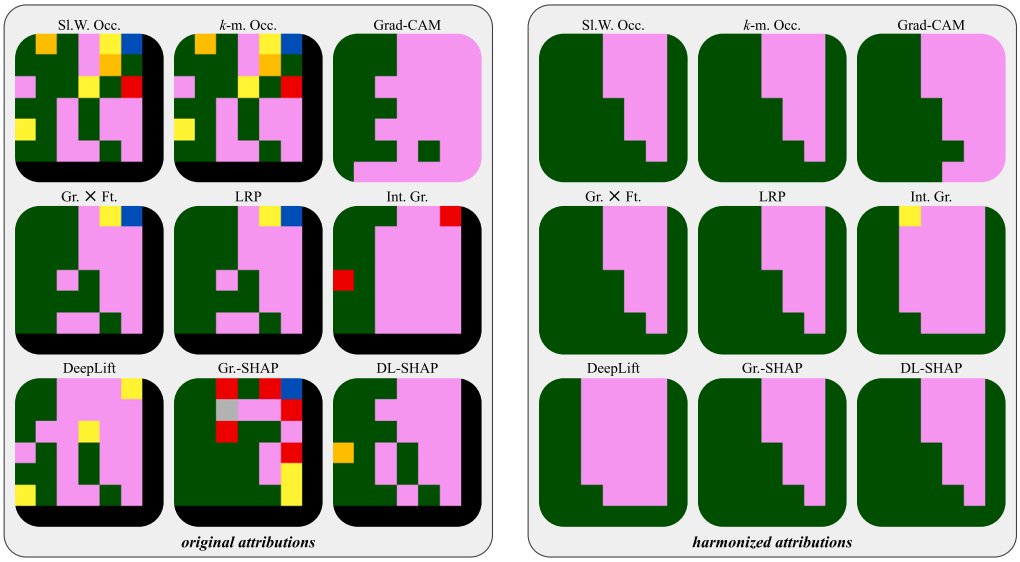
Segmentation ground truth. Tables C.7, C.8, C.9 list the evaluation metrics comparing the predominant attributions with the segmentation ground truth. Some metrics show a significant upward trend after harmonization, with the highest impact observed for UH-Net using DFC2020. However, there are also cases where harmonization leads to a decline in metrics. For Ben-ge, Sliding Window Occlusions (or k-means occlusions) perform best across all CNNs. For DFC2020, DeepLift-SHAP yields the best performance across all CNNs.

Similarities between original and harmonized attributions. The similarities between original and harmonized attributions for the inputs are shown in Fig. C.18. For all models, correlation is present for Sliding Window Occlusions, k-means Occlusions, and Grad-CAM, but hardly for the other methods. The likely reason is the high level of noise for the gradient-based methods, but its low level (or absence) for the occlusion-based methods and Grad-CAM. For the UH-Net, noise is generally lower due to the presence of skip connections. It thus exhibits higher similarity values also for the gradient-based methods.

Discussion. Section 6 discusses why attributions are more meaningful in deeper layers than in the input and the reasons for the low performance of harmonization when applied to the input.

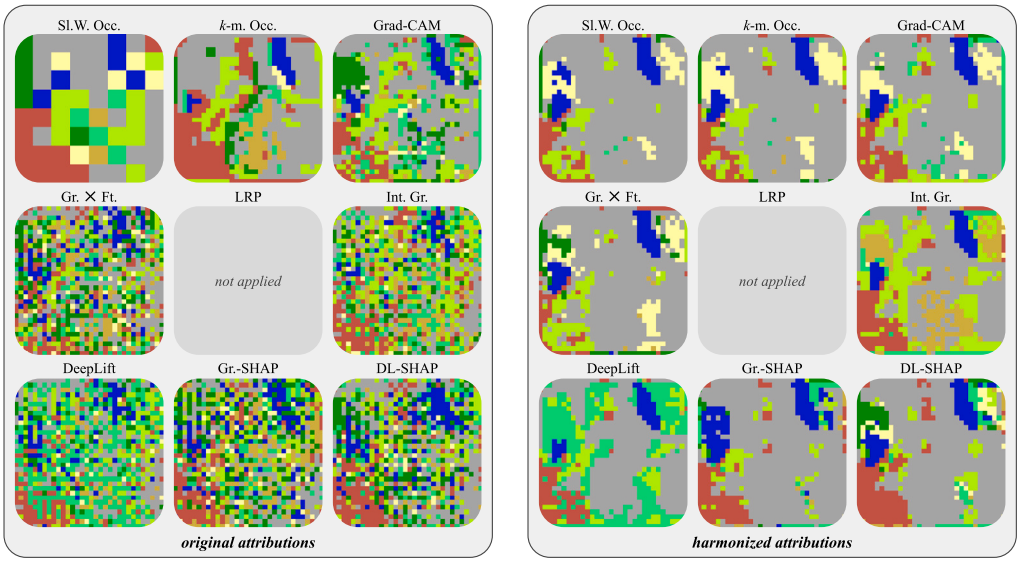


(a) DFC2020. The model correctly predicts the classes ■ Forest, ■ Grassland, ■ Cropland, and ■ Urban/Built-up as these classes occur with a relative area >10 %. Satellite image and ground truth are shown in Figure 3; Table 1 provides the land cover legend.

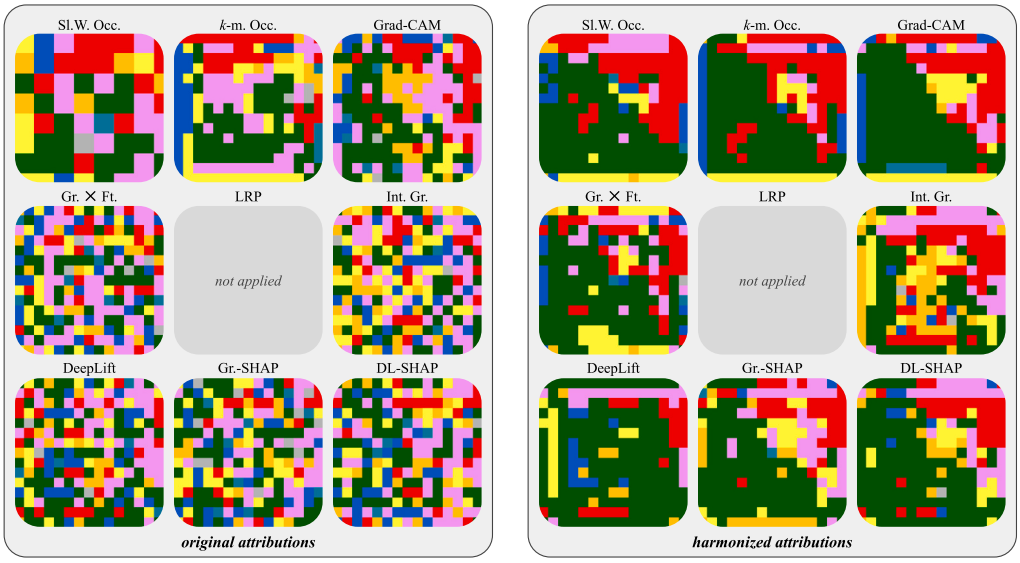


(b) Ben-ge. The model correctly predicts the classes ■ Tree cover and ■ Cropland as these classes occur with a relative area >10 %. ■ Black indicates that all attributions are zero. This occurs due to the odd feature map size of 7 × 7, followed by Max Pooling with a 2 × 2 kernel. Satellite image and ground truth are shown in Figure 4; Table 2 provides the land cover legend.

Fig. B.9. Original and harmonized attribution maps of the deep **VGG-16** layer. Harmonization (right-hand side) enhances the similarity among the attribution methods compared to the original attributions (left). The color represents the predominant class of an attribution vector (argmax).

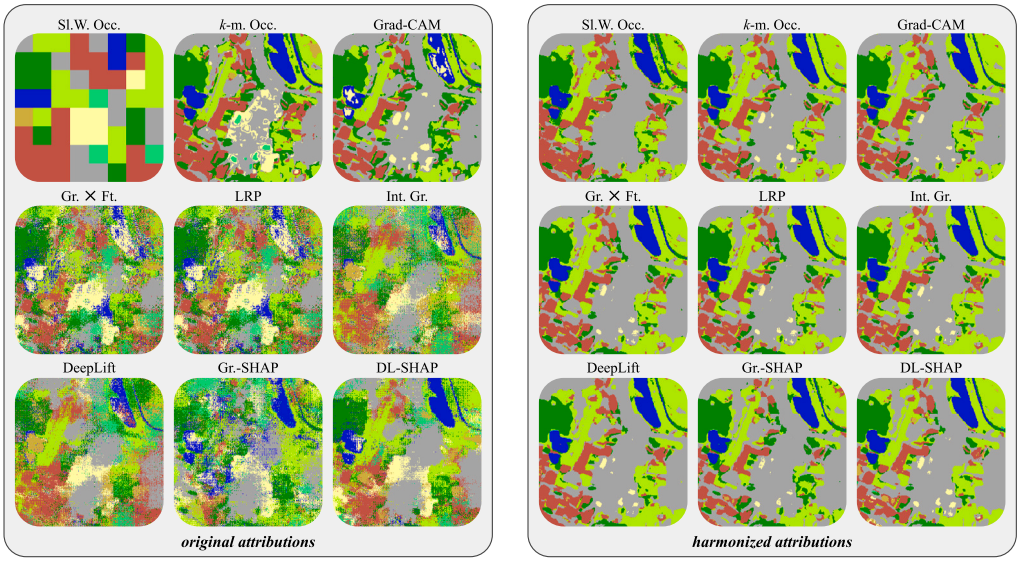


(a) DFC2020. The model correctly predicts the classes Forest, Grassland, Cropland, and Urban/Built-up as these classes occur with a relative area >10 %. Satellite image and ground truth are shown in Figure 3; Table 1 provides the land cover legend..

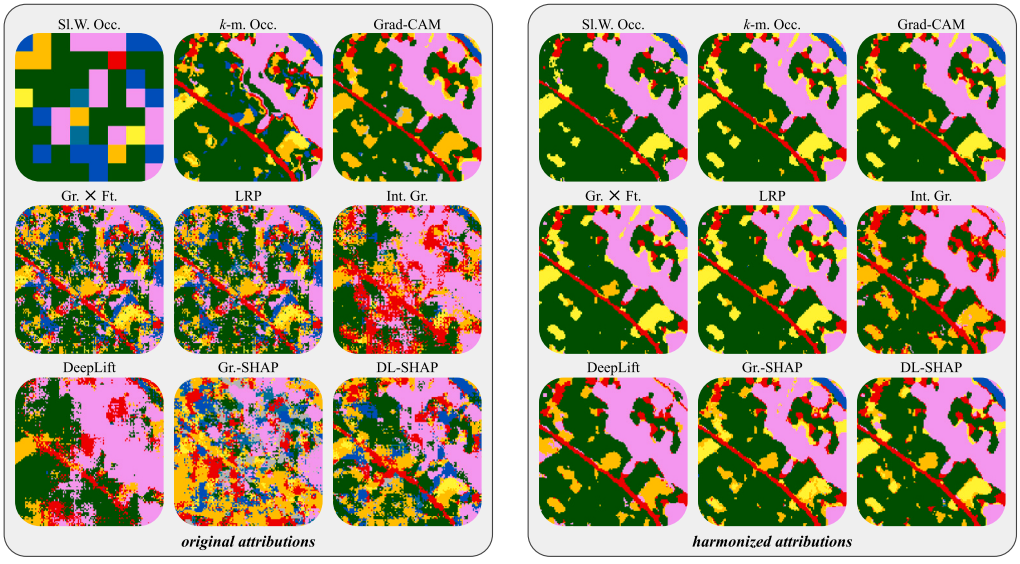


(b) Ben-ge. The model predicts the classes Tree cover, Cropland, and Grassland. Grassland is not included in the label as the relative area threshold is >10 %. Satellite image and ground truth are shown in Figure 4; Table 2 provides the land cover legend.

Fig. B.10. Original and harmonized attribution maps of the deep **ResNet-18** layer. Harmonization (right-hand side) enhances the similarity among the attribution methods compared to the original attributions (left). The color represents the predominant class of an attribution vector (argmax). We do not apply Layer-wise Relevance Propagation as originally no rules are defined for skip connections.



(a) DFC2020. The model correctly predicts the classes Forest, Grassland, Cropland, and Urban/Built-up as these classes occur with a relative area >10 %. Satellite image and ground truth are shown in Figure 3; Table 1 provides the land cover legend.



(b) Ben-ge. The model predicts the classes Tree cover, Cropland, and Grassland. Grassland is not included in the label as the relative area threshold is >10 %. Satellite image and ground truth are shown in Figure 4; Table 2 provides the land cover legend.

Fig. B.11. Original and harmonized attribution maps of the deep **UH-Net** layer. Harmonization (right-hand side) enhances the similarity among the attribution methods compared to the original attributions (left). The color represents the predominant class of an attribution vector (argmax).

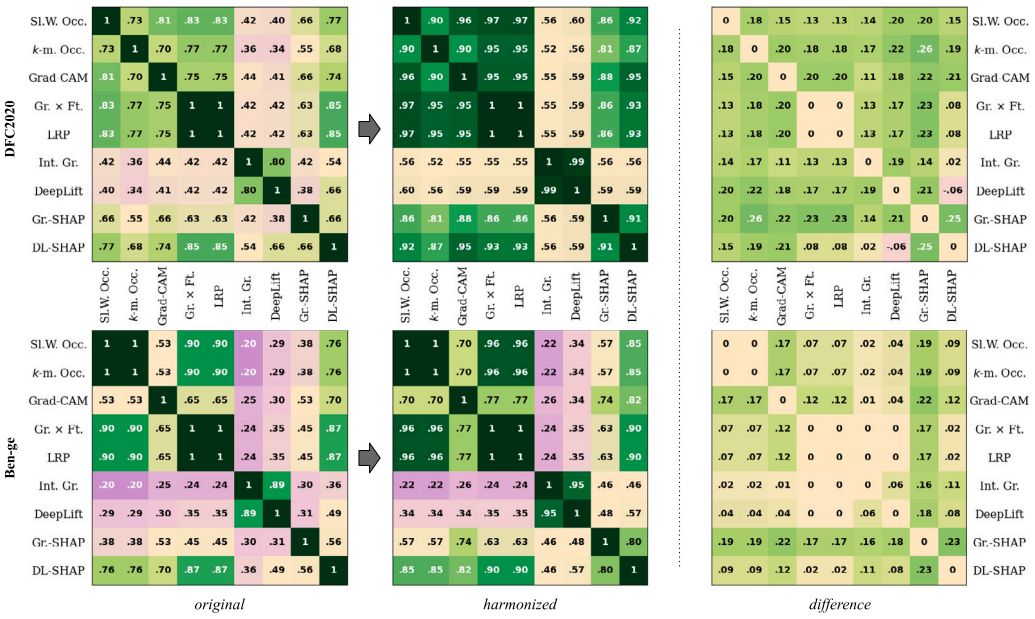


Fig. B.12. Similarities between attribution methods for the deep **VGG-16** layer. Shown are the Pearson correlation coefficients for DFC2020 (top) and Ben-ge (bottom), averaged across ten models. From left to right: the similarities of the *original* attributions, the *harmonized* attributions, and the *difference* in similarities. *The low similarities of Grad-CAM attributions for Ben-ge are caused by the odd-sized feature maps (7 × 7), followed by a Max Pooling operation with a 2 × 2 kernel. This causes all attribution methods — except Grad-CAM — to assign zero attributions to the right column and bottom row of the attribution maps, affecting 27 % of the feature vectors.*.

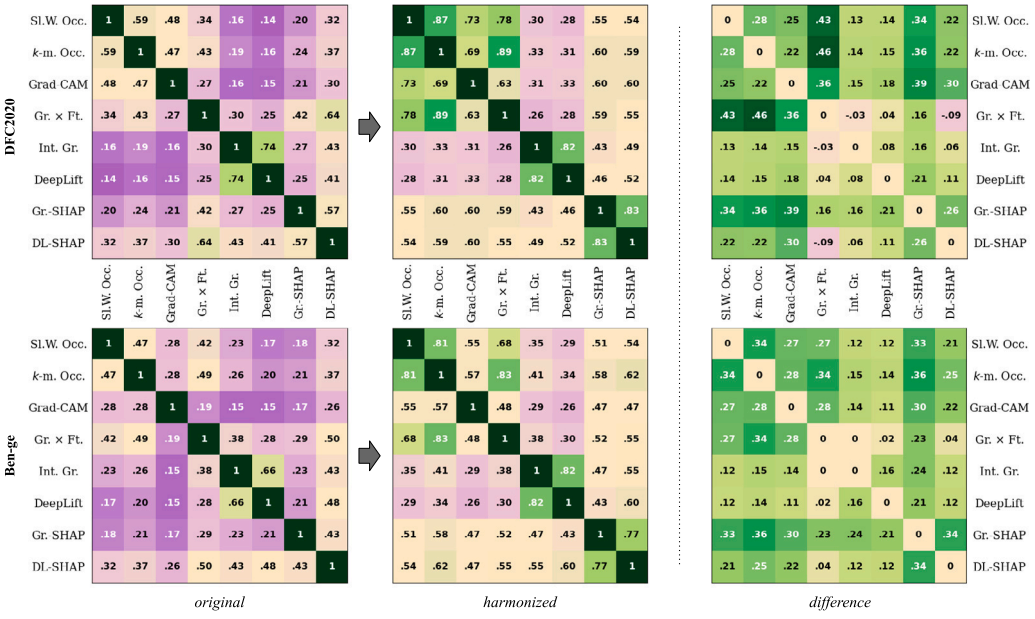


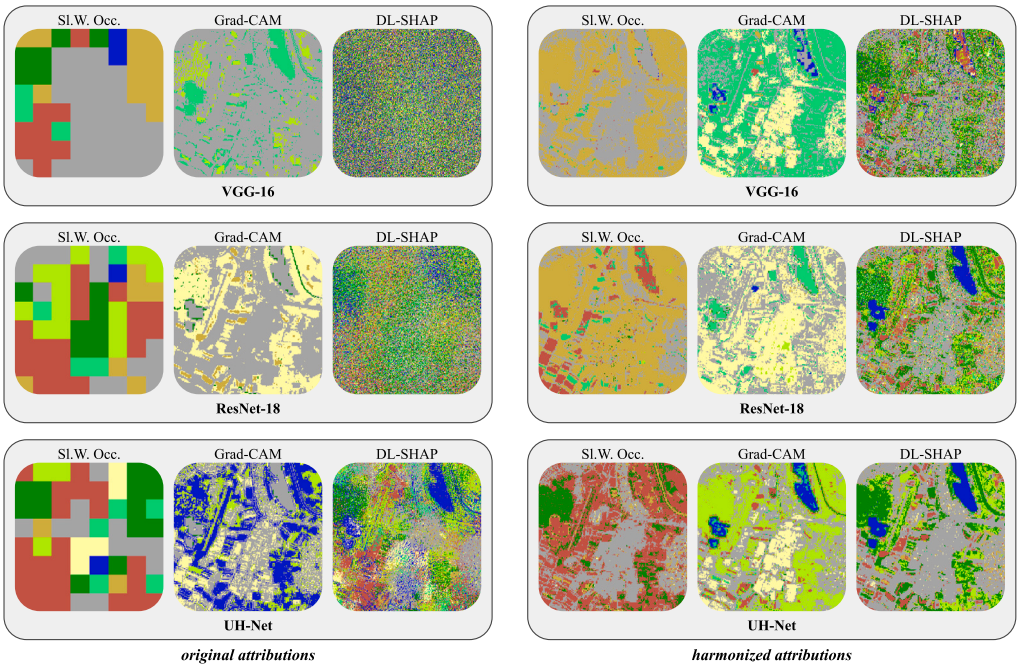
Fig. B.13. Similarities between attribution methods for the deep **ResNet-18** layer. Shown are the Pearson correlation coefficients for DFC2020 (top) and Ben-ge (bottom), averaged across ten models. From left to right: the similarities of the *original* attributions, the *harmonized* attributions, and the *difference* in similarities. We do not apply Layer-wise Relevance Propagation as originally no rules are defined for skip connections.

Table B.5
Metrics comparing the predominant attribution class (original and harmonized) with the segmentation ground truth for the deep **VGG-16** layer (in %). Deviations in the differences may result from rounding. The best values in a dataset column, along with those up to 2% lower, are highlighted in bold.

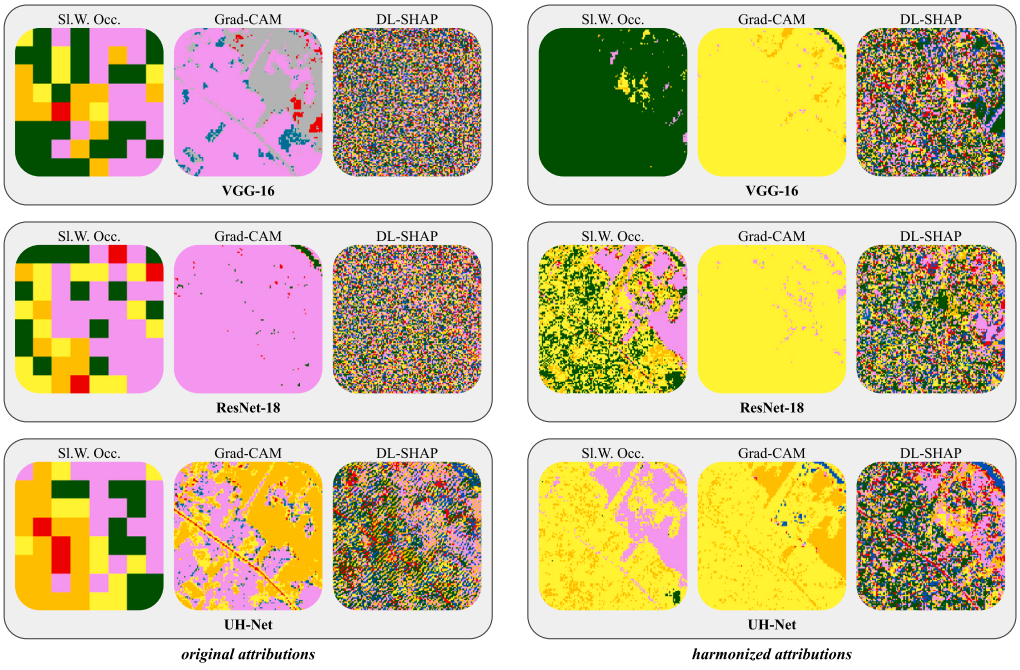
Dataset	Attribution method	Accuracy			F1 (micro)			F1 (macro)		
		orig.	harm.	diff.	orig.	harm.	diff.	orig.	harm.	diff.
DFC2020	Sl.W. Occlusions	**93** ± 0	**94** ± 0	1 ± 0	72 ± 1	**76** ± 0	4 ± 1	61 ± 1	**65** ± 1	4 ± 1
	k-means Occlusions	**92** ± 0	**94** ± 0	2 ± 0	68 ± 1	**74** ± 0	7 ± 1	56 ± 1	**63** ± 1	7 ± 1
	Grad-CAM	**94** ± 0	**94** ± 0	0 ± 0	**75** ± 0	**75** ± 0	0 ± 0	**65** ± 0	**65** ± 0	0 ± 0
	Gradients×Features	**92** ± 0	**94** ± 0	2 ± 0	69 ± 1	**75** ± 0	7 ± 1	57 ± 1	**65** ± 1	8 ± 1
	LRP	**92** ± 0	**94** ± 0	2 ± 0	69 ± 1	**75** ± 0	7 ± 1	57 ± 1	**65** ± 1	8 ± 1
	Integrated Gradients	86 ± 3	89 ± 3	2 ± 1	45 ± 11	55 ± 12	10 ± 2	38 ± 8	47 ± 9	9 ± 2
	DeepLift	87 ± 3	89 ± 3	2 ± 1	47 ± 12	57 ± 13	10 ± 2	40 ± 9	49 ± 10	9 ± 2
	Gradient-SHAP	91 ± 0	**93** ± 0	2 ± 0	63 ± 1	71 ± 1	9 ± 1	51 ± 1	60 ± 1	9 ± 1
	DeepLift-SHAP	**93** ± 0	**93** ± 0	0 ± 0	71 ± 1	**74** ± 0	3 ± 1	61 ± 1	**65** ± 0	4 ± 1
Ben-ge	Sl.W. Occlusions	88 ± 0	**91** ± 0	3 ± 0	53 ± 1	65 ± 1	12 ± 1	32 ± 1	40 ± 1	8 ± 1
	k-means Occlusions	88 ± 0	**91** ± 0	3 ± 0	53 ± 1	65 ± 1	12 ± 1	32 ± 1	40 ± 1	8 ± 1
	Grad-CAM	**91** ± 0	**92** ± 0	1 ± 0	**63** ± 2	**69** ± 1	5 ± 1	**43** ± 1	**45** ± 1	2 ± 1
	Gradients×Features	**91** ± 0	**92** ± 0	2 ± 0	**63** ± 1	**69** ± 0	6 ± 1	40 ± 1	**43** ± 0	3 ± 1
	LRP	**91** ± 0	**92** ± 0	1 ± 0	**63** ± 1	**69** ± 0	6 ± 1	40 ± 1	**43** ± 0	3 ± 1
	Integrated Gradients	87 ± 1	88 ± 1	1 ± 0	48 ± 3	54 ± 3	5 ± 1	27 ± 3	29 ± 4	2 ± 1
	DeepLift	87 ± 1	89 ± 1	1 ± 0	49 ± 3	55 ± 3	6 ± 1	27 ± 2	30 ± 3	3 ± 1
	Gradient-SHAP	88 ± 0	**91** ± 0	3 ± 0	51 ± 2	64 ± 1	13 ± 1	32 ± 1	41 ± 1	10 ± 1
	DeepLift-SHAP	**91** ± 0	**92** ± 0	1 ± 0	**62** ± 1	**68** ± 0	6 ± 1	**41** ± 1	**45** ± 0	4 ± 1

Table B.6
Metrics comparing the predominant attribution class (original and harmonized) with the segmentation ground truth for the deep **ResNet-18** layer (in %). Deviations in the differences may result from rounding. The best values in a dataset column, along with those up to 2% lower, are highlighted in bold.

Dataset	Attribution method	Accuracy			F1 (micro)			F1 (macro)		
		orig.	harm.	diff.	orig.	harm.	diff.	orig.	harm.	diff.
DFC2020	Sl.W. Occlusions	**85** ± 1	**86** ± 1	1 ± 1	40 ± 5	43 ± 6	3 ± 2	33 ± 3	34 ± 4	1 ± 1
	k-means Occlusions	**85** ± 1	85 ± 1	1 ± 1	39 ± 4	41 ± 5	2 ± 2	33 ± 3	33 ± 4	1 ± 1
	Grad-CAM	**85** ± 2	**86** ± 2	1 ± 0	39 ± 7	42 ± 8	3 ± 2	34 ± 5	35 ± 6	2 ± 1
	Gradients×Features	**84** ± 1	**87** ± 2	3 ± 1	34 ± 4	46 ± 6	12 ± 3	27 ± 3	35 ± 5	7 ± 2
	Integrated Gradients	83 ± 1	84 ± 1	1 ± 1	32 ± 3	37 ± 6	5 ± 3	28 ± 3	33 ± 5	4 ± 3
	DeepLift	83 ± 1	85 ± 2	2 ± 1	34 ± 4	42 ± 6	8 ± 3	30 ± 3	37 ± 5	7 ± 4
	Gradient-SHAP	83 ± 1	**88** ± 1	5 ± 1	31 ± 3	**51** ± 5	21 ± 4	27 ± 3	**45** ± 3	18 ± 3
	DeepLift-SHAP	**86** ± 1	**88** ± 1	2 ± 1	**43** ± 5	**51** ± 4	8 ± 4	**38** ± 5	**46** ± 3	8 ± 3
Ben-ge	Sl.W. Occlusions	**82** ± 1	83 ± 2	1 ± 1	27 ± 4	32 ± 8	5 ± 4	17 ± 2	19 ± 4	2 ± 2
	k-means Occlusions	**82** ± 1	83 ± 2	1 ± 1	27 ± 5	30 ± 7	3 ± 3	17 ± 3	18 ± 4	1 ± 1
	Grad-CAM	**83** ± 2	83 ± 2	1 ± 1	**31** ± 7	33 ± 9	2 ± 3	**20** ± 4	20 ± 5	0 ± 1
	Gradients×Features	80 ± 1	82 ± 2	1 ± 1	21 ± 3	26 ± 7	5 ± 5	13 ± 1	15 ± 3	2 ± 2
	Integrated Gradients	**82** ± 1	83 ± 2	2 ± 1	26 ± 2	34 ± 7	8 ± 5	16 ± 1	19 ± 3	3 ± 2
	DeepLift	**82** ± 1	**84** ± 1	2 ± 1	**30** ± 2	37 ± 5	7 ± 4	16 ± 1	19 ± 3	3 ± 2
	Gradient-SHAP	**82** ± 1	**86** ± 2	5 ± 2	26 ± 2	**44** ± 8	18 ± 6	17 ± 1	**27** ± 4	10 ± 3
	DeepLift-SHAP	**83** ± 1	**86** ± 2	3 ± 1	**31** ± 3	**43** ± 7	12 ± 5	**20** ± 2	**26** ± 4	6 ± 2



(a) DFC2020. Satellite image and ground truth are shown in Figure 3; Table 1 provides the land cover legend.



(b) Ben-ge. Satellite image and ground truth are shown in Figure 4; Table 2 provides the land cover legend.

Fig. C.14. Original and harmonized **input** attribution maps for the three models (VGG-16 at the top, ResNet-18 in the middle, and UH-Net at the bottom) and three attribution methods. The color represents the predominant class of an attribution vector (argmax).

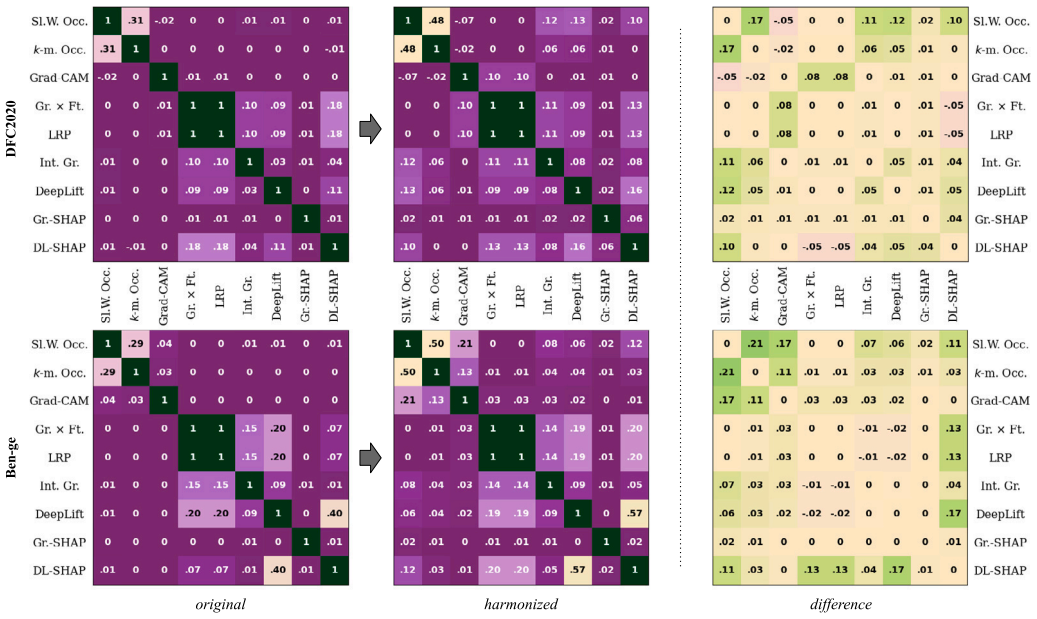


Fig. C.15. Similarities between attribution methods for the **VGG-16 input**. Shown are the Pearson correlation coefficients for DFC2020 (top) and Ben-ge (bottom), averaged across ten models. From left to right: the similarities of the *original* attributions, the *harmonized* attributions, and the *difference* in similarities.

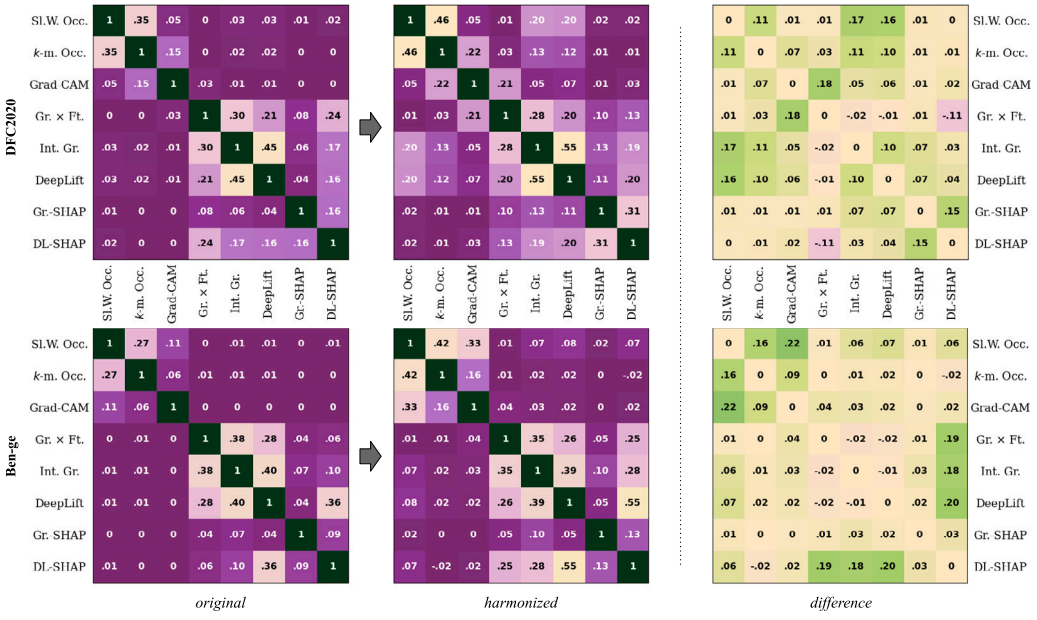


Fig. C.16. Similarities between attribution methods for the **ResNet-18 input**. Shown are the Pearson correlation coefficients for DFC2020 (top) and Ben-ge (bottom), averaged across ten models. From left to right: the similarities of the *original* attributions, the *harmonized* attributions, and the *difference* in similarities.

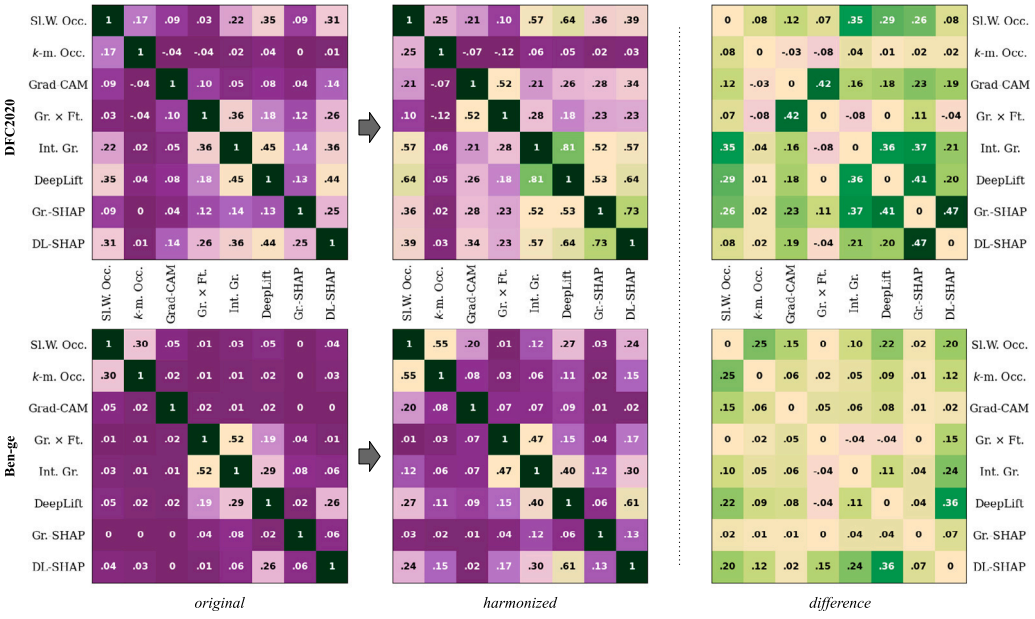


Fig. C.17. Similarities between attribution methods for the **UH-Net input**. Shown are the Pearson correlation coefficients for DFC2020 (top) and Ben-ge (bottom), averaged across ten models. From left to right: the similarities of the *original* attributions, the *harmonized* attributions, and the *difference* in similarities.

Table C.7
Metrics comparing the predominant attribution class (original and harmonized) with the segmentation ground truth for the **VGG-16 input** (in %). Deviations in the differences may result from rounding. The best values in a dataset column, along with those up to 2% lower, are highlighted in bold.

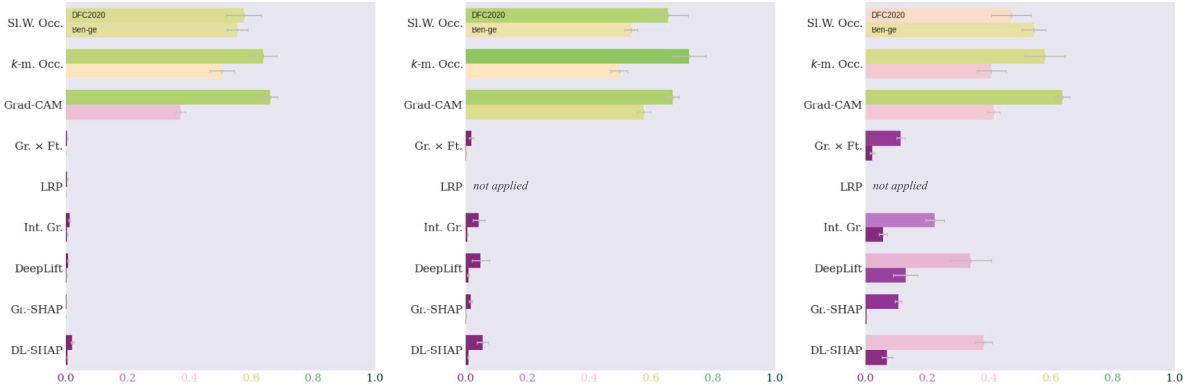
Dataset	Attribution method	Accuracy			F1 (micro)			F1 (macro)		
		orig.	harm.	diff.	orig.	harm.	diff.	orig.	harm.	diff.
DFC2020	Sl.W. Occlusions	**81** ± 1	81 ± 1	0 ± 1	23 ± 3	23 ± 5	0 ± 5	20 ± 3	17 ± 6	-4 ± 5
	k-means Occlusions	**82** ± 1	80 ± 2	-1 ± 1	**26** ± 5	21 ± 6	-5 ± 3	**23** ± 3	18 ± 4	-5 ± 2
	Grad-CAM	**81** ± 1	81 ± 1	0 ± 0	23 ± 2	22 ± 2	-1 ± 1	19 ± 2	18 ± 2	-1 ± 1
	Gradients×Features	**80** ± 0	80 ± 0	0 ± 0	18 ± 0	19 ± 1	1 ± 1	15 ± 0	16 ± 1	1 ± 1
	LRP	**80** ± 0	80 ± 0	0 ± 0	18 ± 0	19 ± 1	1 ± 1	15 ± 0	16 ± 1	1 ± 1
	Integrated Gradients	79 ± 0	80 ± 1	0 ± 1	17 ± 0	19 ± 2	2 ± 2	14 ± 0	16 ± 2	2 ± 1
	DeepLift	**80** ± 0	80 ± 1	1 ± 1	18 ± 0	22 ± 2	4 ± 2	15 ± 0	18 ± 1	3 ± 1
	Gradient-SHAP	78 ± 0	80 ± 0	1 ± 0	14 ± 0	19 ± 1	5 ± 1	12 ± 0	16 ± 1	4 ± 1
	DeepLift-SHAP	**80** ± 0	**84** ± 1	4 ± 0	19 ± 1	**35** ± 2	16 ± 2	16 ± 0	**29** ± 2	13 ± 1
Ben-ge	Sl.W. Occlusions	**84** ± 1	**88** ± 2	4 ± 1	**35** ± 3	**51** ± 8	16 ± 5	**23** ± 2	**27** ± 4	5 ± 3
	k-means Occlusions	**83** ± 1	82 ± 3	-1 ± 2	31 ± 6	26 ± 13	-5 ± 8	19 ± 3	15 ± 5	-4 ± 2
	Grad-CAM	80 ± 0	81 ± 0	1 ± 0	19 ± 1	24 ± 1	5 ± 1	11 ± 1	8 ± 1	-3 ± 1
	Gradients×Features	80 ± 0	80 ± 0	0 ± 0	19 ± 0	18 ± 0	-1 ± 0	12 ± 0	11 ± 0	-1 ± 0
	LRP	80 ± 0	80 ± 0	0 ± 0	19 ± 0	18 ± 1	-1 ± 0	12 ± 0	11 ± 0	-1 ± 0
	Integrated Gradients	80 ± 0	80 ± 0	0 ± 0	20 ± 0	20 ± 1	0 ± 1	12 ± 0	12 ± 1	0 ± 1
	DeepLift	80 ± 0	80 ± 0	0 ± 0	21 ± 0	21 ± 1	1 ± 1	12 ± 0	12 ± 1	0 ± 0
	Gradient-SHAP	79 ± 0	79 ± 0	0 ± 0	14 ± 0	16 ± 1	1 ± 1	10 ± 0	10 ± 1	1 ± 1
	DeepLift-SHAP	80 ± 0	81 ± 0	1 ± 0	20 ± 0	26 ± 1	6 ± 1	12 ± 0	15 ± 1	3 ± 1

Table C.8
Metrics comparing the predominant attribution class (original and harmonized) with the segmentation ground truth for the **ResNet-18 input** (in %). Deviations in the differences may result from rounding. The best values in a dataset column, along with those up to 2% lower, are highlighted in bold.

Dataset	Attribution method	Accuracy			F1 (micro)			F1 (macro)		
		orig.	harm.	diff.	orig.	harm.	diff.	orig.	harm.	diff.
DFC2020	Sl.W. Occlusions	**80** ± 1	79 ± 1	-1 ± 1	19 ± 3	15 ± 4	-4 ± 3	18 ± 2	11 ± 2	-7 ± 2
	k-means Occlusions	**81** ± 1	80 ± 1	-1 ± 0	**24** ± 4	20 ± 4	-4 ± 2	**21** ± 3	16 ± 4	-5 ± 2
	Grad-CAM	**80** ± 1	79 ± 2	-1 ± 1	19 ± 4	17 ± 7	-2 ± 3	17 ± 3	15 ± 5	-3 ± 3
	Gradients×Features	**81** ± 0	80 ± 1	0 ± 0	**23** ± 1	21 ± 2	-2 ± 2	**20** ± 1	18 ± 1	-2 ± 1
	Integrated Gradients	**80** ± 0	81 ± 1	1 ± 1	21 ± 1	25 ± 5	4 ± 4	17 ± 1	21 ± 3	4 ± 2
	DeepLift	**80** ± 0	82 ± 2	1 ± 1	**22** ± 1	27 ± 6	5 ± 5	18 ± 1	23 ± 4	5 ± 3
	Gradient-SHAP	**79** ± 0	83 ± 1	4 ± 1	17 ± 0	31 ± 5	14 ± 4	15 ± 0	26 ± 4	12 ± 3
	DeepLift-SHAP	**81** ± 0	**87** ± 1	7 ± 1	**24** ± 1	**50** ± 6	26 ± 5	**20** ± 1	**41** ± 5	21 ± 4
Ben-ge	Sl.W. Occlusions	**83** ± 1	**83** ± 2	1 ± 2	**30** ± 2	**33** ± 8	3 ± 7	**19** ± 1	**17** ± 4	-3 ± 3
	k-means Occlusions	**81** ± 2	79 ± 2	-2 ± 1	25 ± 7	15 ± 9	-10 ± 3	16 ± 4	10 ± 5	-6 ± 1
	Grad-CAM	**81** ± 0	**81** ± 0	0 ± 0	23 ± 2	24 ± 1	1 ± 1	12 ± 1	8 ± 1	-4 ± 0
	Gradients×Features	80 ± 0	80 ± 0	0 ± 0	19 ± 1	18 ± 1	0 ± 1	11 ± 0	11 ± 1	0 ± 0
	Integrated Gradients	80 ± 0	80 ± 0	0 ± 0	19 ± 0	20 ± 1	1 ± 1	12 ± 0	12 ± 1	0 ± 1
	DeepLift	80 ± 0	**81** ± 0	0 ± 0	21 ± 0	23 ± 1	1 ± 1	12 ± 0	13 ± 1	1 ± 1
	Gradient-SHAP	79 ± 0	80 ± 0	1 ± 0	16 ± 0	18 ± 1	2 ± 1	10 ± 0	12 ± 1	2 ± 0
	DeepLift-SHAP	80 ± 0	**82** ± 0	2 ± 0	20 ± 0	27 ± 1	7 ± 1	12 ± 0	**16** ± 1	4 ± 0

Table C.9
Metrics comparing the predominant attribution class (original and harmonized) with the segmentation ground truth for the **UH-Net input** (in %). Deviations in the differences may result from rounding. The best values in a dataset column, along with those up to 2% lower, are highlighted in bold.

Dataset	Attribution method	Accuracy			F1 (micro)			F1 (macro)		
		orig.	harm.	diff.	orig.	harm.	diff.	orig.	harm.	diff.
DFC2020	Sl.W. Occlusions	83 ± 1	87 ± 3	4 ± 2	33 ± 5	48 ± 11	15 ± 8	28 ± 3	38 ± 8	10 ± 6
	k-means Occlusions	84 ± 2	82 ± 2	-1 ± 1	35 ± 7	30 ± 8	-5 ± 3	29 ± 6	24 ± 7	-5 ± 2
	Grad-CAM	84 ± 1	85 ± 2	1 ± 0	37 ± 5	40 ± 6	3 ± 2	30 ± 4	32 ± 4	2 ± 1
	Gradients×Features	81 ± 1	84 ± 1	3 ± 1	23 ± 2	37 ± 6	14 ± 5	19 ± 2	29 ± 4	9 ± 3
	Integrated Gradients	82 ± 1	89 ± 2	7 ± 2	28 ± 3	56 ± 9	29 ± 6	23 ± 2	46 ± 5	23 ± 3
	DeepLift	84 ± 1	**92** ± 2	8 ± 2	36 ± 5	67 ± 9	31 ± 6	30 ± 4	56 ± 6	26 ± 4
	Gradient-SHAP	80 ± 0	90 ± 1	10 ± 0	20 ± 1	62 ± 2	42 ± 2	17 ± 1	51 ± 2	34 ± 2
	DeepLift-SHAP	**87** ± 1	**93** ± 0	6 ± 1	**49** ± 3	**72** ± 1	23 ± 3	**40** ± 2	**61** ± 1	21 ± 3
Ben-ge	Sl.W. Occlusions	**85** ± 1	85 ± 1	0 ± 1	41 ± 4	41 ± 4	-1 ± 3	**28** ± 3	**25** ± 2	-3 ± 2
	k-means Occlusions	**87** ± 1	**88** ± 2	1 ± 1	**48** ± 6	**51** ± 7	2 ± 4	**28** ± 3	**24** ± 3	-5 ± 2
	Grad-CAM	82 ± 1	82 ± 1	0 ± 0	28 ± 3	27 ± 3	-1 ± 2	16 ± 2	12 ± 3	-4 ± 2
	Gradients×Features	82 ± 0	80 ± 0	-2 ± 0	29 ± 2	22 ± 2	-7 ± 1	17 ± 1	13 ± 1	-4 ± 1
	Integrated Gradients	82 ± 0	81 ± 1	-1 ± 0	27 ± 2	25 ± 3	-2 ± 2	16 ± 1	15 ± 2	-2 ± 1
	DeepLift	81 ± 1	81 ± 1	0 ± 1	25 ± 2	24 ± 4	-1 ± 2	16 ± 1	14 ± 2	-2 ± 1
	Gradient-SHAP	78 ± 0	79 ± 0	1 ± 0	13 ± 0	16 ± 1	2 ± 1	9 ± 0	11 ± 1	2 ± 0
	DeepLift-SHAP	83 ± 0	**86** ± 0	2 ± 0	33 ± 1	42 ± 2	10 ± 1	21 ± 1	**25** ± 1	4 ± 1



(a) VGG-16 (b) ResNet-18 (c) UH-Net

Fig. C.18. Similarities between original and harmonized attributions for the CNNs' **inputs**. The upper bar of an attribution shows the Pearson correlation coefficient obtained for the DFC2020 dataset; the lower bar the one for Ben-ge. The colorization highlights the values also represented by the bars as shown on the *x*-axis. The values are averaged across ten models; the gray lines indicate the standard deviations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Data availability

The data that has been used is publically available. The code that has been used has been published by the authors.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *31*, In *Advances in neural information processing systems*. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In G. Goos, J. Hartmanis, J. Van Leeuwen, J. Van Den Bussche, & V. Vianu (Eds.), *Lecture Notes in Computer Science*: *vol. 1973*, *Database theory — ICDT 2001* (pp. 420–434). Springer Berlin Heidelberg: http://dx.doi.org/10.1007/3-540-44503-X_27, URL http://link.springer.com/10.1007/3-540-44503-X_27.

Ahn, J., & Kwak, S. (2018). Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 4981–4990). Salt Lake City, UT: IEEE, http://dx.doi.org/10.1109/CVPR.2018.00523, URL https://ieeexplore.ieee.org/document/8578621/.

Alotaibi, E., & Nassif, N. (2024). Artificial intelligence in environmental monitoring: in-depth analysis. *Discover Artificial Intelligence*, *4*(1), 84. http://dx.doi.org/10.1007/s44163-024-00198-1, URL https://link.springer.com/10.1007/s44163-024-00198-1.

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. URL http://arxiv.org/abs/1711.06104. arXiv:1711.06104 [cs].

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, *10*(7), Article e0130140. http://dx.doi.org/10.1371/journal.pone.0130140, URL https://dx.plos.org/10.1371/journal.pone.0130140.

Carter, S., Armstrong, Z., Schubert, L., Johnson, I., & Olah, C. (2019). Activation Atlas. *Distill*, *4*(3), Article e15. http://dx.doi.org/10.23915/distill.00015, URL https://distill.pub/2019/activation-atlas.

Chong, Y., Chen, X., Tao, Y., & Pan, S. (2021). Erase then grow: Generating correct class activation maps for weakly-supervised semantic segmentation. *Neurocomputing*, *453*, 97–108. http://dx.doi.org/10.1016/j.neucom.2021.04.103, URL https://linkinghub.elsevier.com/retrieve/pii/S0925231221006615.

Dhore, V., Bhat, A., Nerlekar, V., Chavhan, K., & Umare, A. (2024). Enhancing Explainable AI: A hybrid approach combining GradCAM and LRP for CNN interpretability. http://dx.doi.org/10.48550/arXiv.2405.12175, URL http://arxiv.org/abs/2405.12175. arXiv:2405.12175 [cs] version: 1.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 [Cs]. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929.

Gulum, M. A., Trombley, C. M., & Kantardzic, M. (2021). Improved deep learning explanations for prostate lesion classification through grad-CAM and saliency map fusion. In *2021 IEEE 34th international symposium on computer-based medical systems (CBMS)* (pp. 498–502). Aveiro, Portugal: IEEE, http://dx.doi.org/10.1109/CBMS52027.2021.00099, URL https://ieeexplore.ieee.org/document/9474664/.

Hanna, J., Mommert, M., & Borth, D. (2023). Sparse multimodal vision transformer for weakly supervised semantic segmentation. In *2023 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 2145–2154). Vancouver, BC, Canada: IEEE, http://dx.doi.org/10.1109/CVPRW59228.2023.00208, URL https://ieeexplore.ieee.org/document/10208918/.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778). Las Vegas, NV, USA: IEEE, http://dx.doi.org/10.1109/CVPR.2016.90, URL http://ieeexplore.ieee.org/document/7780459/.

Höhl, A., Obadic, I., Fernández-Torres, M.-A., Najjar, H., Oliveira, D. A. B., Akata, Z., Dengel, A., & Zhu, X. X. (2024). Opening the Black Box: A systematic review on explainable artificial intelligence in remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, *12*(4), 261–304. http://dx.doi.org/10.1109/MGRS.2024.3467001, URL https://ieeexplore.ieee.org/abstract/document/10742949. Conference Name: IEEE Geoscience and Remote Sensing Magazine.

Hsu, C. Y., & Li, W. (2023). Explainable GeoAI: can saliency maps help interpret artificial intelligence's learning process? An empirical study on natural feature detection. *International Journal of Geographical Information Science*, *37*(5), 963–987. http://dx.doi.org/10.1080/13658816.2023.2191256, URL https://www.tandfonline.com/doi/full/10.1080/13658816.2023.2191256.

Iqbal, H. (2018). HarisIqbal88/PlotNeuralNet v1.0.0. http://dx.doi.org/10.5281/zenodo.2526396.

Jonnarth, A., & Felsberg, M. (2022). Importance sampling cams for weakly-supervised segmentation. In *ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2639–2643). [ISSN: 2379-190X] http://dx.doi.org/10.1109/ICASSP43922.2022.9746641, URL https://ieeexplore.ieee.org/abstract/document/9746641.

Kakogeorgiou, I., & Karantzalos, K. (2021). Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, *103*, Article 102520. http://dx.doi.org/10.1016/j.jag.2021.102520.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. http://dx.doi.org/10.48550/arXiv.1609.02907, URL http://arxiv.org/abs/1609.02907. arXiv:1609.02907 [cs].

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for PyTorch. _eprint: 2009.07896.

Kwak, S., Hong, S., & Han, B. (2017). Weakly supervised semantic segmentation using superpixel pooling network. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1), http://dx.doi.org/10.1609/aaai.v31i1.11213, URL https://ojs.aaai.org/index.php/AAAI/article/view/11213.

Lloyd, S. (1982). Least squares quantization in PCM. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, *28*(2), 129–137. http://dx.doi.org/10.1109/TIT.1982.1056489.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *vol. 30*, *Advances in neural information processing systems*. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform manifold approximation and projection for dimension reduction. URL http://arxiv.org/abs/1802.03426. arXiv:1802.03426 [cs, stat].

Mohan, A., & Peeples, J. (2023). Quantitative analysis of primary attribution explainable artificial intelligence methods for remote sensing image classification. In *IGARSS 2023 - 2023 IEEE international geoscience and remote sensing symposium* (pp. 950–953). http://dx.doi.org/10.1109/IGARSS52108.2023.10281981, URL http://arxiv.org/abs/2306.04037. arXiv:2306.04037 [cs].

Mommert, M., Kesseli, N., Hanna, J., Scheibenreif, L., Borth, D., & Demir, B. (2023). Ben-Ge: Extending bigearthnet with geographical and environmental data. In *IGARSS 2023 - 2023 IEEE international geoscience and remote sensing symposium* (pp. 1016–1019). Pasadena, CA, USA: IEEE, http://dx.doi.org/10.1109/IGARSS52108.2023.10282767, URL https://ieeexplore.ieee.org/document/10282767/.

Nieradzik, L., Stephani, H., Sieburg-Rockel, J., Helmling, S., Olbrich, A., & Keuper, J. (2024). Challenging the Black Box: A comprehensive evaluation of attribution maps of CNN applications in agriculture and forestry. URL http://arxiv.org/abs/2402.11670. arXiv:2402.11670.

Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts. *Distill*, http://dx.doi.org/10.23915/distill.00003.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), Article e7. http://dx.doi.org/10.23915/distill.00007, URL https://distill.pub/2017/feature-visualization.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. *Distill*, *3*(3), Article e10. http://dx.doi.org/10.23915/distill.00010, URL https://distill.pub/2018/building-blocks.

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. URL http://arxiv.org/abs/2002.04803. arXiv:2002.04803 [cs, stat].

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical image computing and computer-assisted intervention – MICCAI 2015* (pp. 234–241). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K.-R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(11), 2660–2673. http://dx.doi.org/10.1109/TNNLS.2016.2599820, URL https://ieeexplore.ieee.org/abstract/document/7552539. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

Saphra, N., & Wiegreffe, S. (2024). Mechanistic?. http://dx.doi.org/10.48550/arXiv.2410.09087, URL http://arxiv.org/abs/2410.09087. arXiv:2410.09087 [cs].

Schmitt, M., Hughes, L., Ghamisi, P., Yokoya, N., & Hänsch, R. (2019). 2020 IEEE GRSS data fusion contest. http://dx.doi.org/10.21227/rha7-m332.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, *128*(2), 336–359. http://dx.doi.org/10.1007/s11263-019-01228-7.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th international conference on machine learning* (pp. 3145–3153). PMLR, [ISSN: 2640-3498] URL https://proceedings.mlr.press/v70/shrikumar17a.html.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. http://dx.doi.org/10.48550/arXiv.1409.1556, URL http://arxiv.org/abs/1409.1556. arXiv:1409.1556.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). SmoothGrad: removing noise by adding noise. URL http://arxiv.org/abs/1706.03825. arXiv:1706.03825 [cs, stat].

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015). Striving for simplicity: The all convolutional net. URL http://arxiv.org/abs/1412.6806. arXiv:1412.6806 [cs].

Stomberg, T. T., Leonhardt, J., Weber, I., & Roscher, R. (2023). Recognizing protected and anthropogenic patterns in landscapes using interpretable machine learning and satellite imagery. *Frontiers in Artificial Intelligence*, *6*, Article 1278118. http://dx.doi.org/10.3389/frai.2023.1278118, URL https://www.frontiersin.org/articles/10.3389/frai.2023.1278118/full.

Stomberg, T., Weber, I., Schmitt, M., & Roscher, R. (2021). Jungle-Net: Using explainable machine learning to gain new insights into the appearance of wilderness in satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *V-3-2021*, 317–324. http://dx.doi.org/10.5194/isprs-annals-V-3-2021-317-2021.

Sumbul, G., de Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., & Markl, V. (2021). BigEarthNet-MM: A large scale multi-modal multi-label benchmark archive for remote sensing image classification and retrieval. *IEEE Geoscience and Remote Sensing Magazine*, *9*(3), 174–180. http://dx.doi.org/10.1109/MGRS.2021.3089174, URL http://arxiv.org/abs/2105.07921. arXiv:2105.07921 [cs].

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In D. Precup, & Y. W. Teh (Eds.), *Proceedings of machine learning research*: *vol. 70*, *Proceedings of the 34th international conference on machine learning* (pp. 3319–3328). PMLR, URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Wang, Y., Zhang, J., Kan, M., Shan, S., & Chen, X. (2020). Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 12272–12281). Seattle, WA, USA: IEEE, http://dx.doi.org/10.1109/CVPR42600.2020.01229, URL https://ieeexplore.ieee.org/document/9157474/.

Xing, J., & Sieber, R. (2023). The challenges of integrating explainable artificial intelligence into GeoAI. *Transactions in GIS*, *27*(3), 626–645. http://dx.doi.org/10.1111/tgis.13045, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.13045. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.13045.

Yang, M., & Kim, B. (2019). Benchmarking attribution methods with relative feature importance. URL http://arxiv.org/abs/1907.09701. arXiv:1907.09701.

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.-E., Xu, P., Ramoino, F., & Arino, O. (2022). ESA WorldCover 10 m 2021 v200. http://dx.doi.org/10.5281/zenodo.7254221, URL https://zenodo.org/records/7254221.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Lecture notes in computer science*, *Computer vision – ECCV 2014* (pp. 818–833). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-10590-1_53.

Zeng, X., Wang, T., Dong, Z., Zhang, X., & Gu, Y. (2023). Superpixel consistency saliency map generation for weakly supervised semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–16. http://dx.doi.org/10.1109/TGRS.2023.3264232, URL https://ieeexplore.ieee.org/document/10097682/.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. (pp. 2921–2929). URL https://openaccess.thecvf.com/content_cvpr_2016/html/Zhou_Learning_Deep_Features_CVPR_2016_paper.html.