Deep Occlusion Framework for Multimodal Earth Observation Data

Burak Ekim and Michael Schmitt, *Senior Member, IEEE*

***Abstract*— Advancements in Earth observation (EO) have led to an increase in the volume of and easier access to multimodal geospatial data, making environmental monitoring and analysis more accessible. However, understanding the influence of each input modality on decision-making within deep learning models remains an open challenge. This letter proposes a deep occlusion framework to enhance the interpretability of a multimodal model for land naturalness assessment, using a supervised pixelwise regression task for naturalness mapping with the input modalities Sentinel-2 and Sentinel-1 imagery, land cover maps, and nighttime lights intensity data. The proposed framework systematically occludes individual input modalities to create modality-level influence scores. Influence scores are attributed to input modalities by measuring the distance between the embedding of the nonoccluded input and the embedding of the input with a single modality occluded, revealing how each modality influences predictions and clarifying their contributions (and, thus, importance) in the model's decision-making process. The results provide further insights into how input modalities influence the model's decision-making at both the sample level, enabling regional case studies, and the dataset level, allowing for data pruning and improving training and inference times. The code is available at https://github.com/burakekim/embedding_occlusion.**

***Index Terms*— Earth observation (EO), environmental conservation, feature attribution, interpretable machine learning, naturalness, occlusion sensitivity maps (OSMs).**

I. Introduction

THROUGHOUT history, humans have continuously modified the landscape to accommodate their needs. From cutting down forests to building urban metropolises, these transformations have been instrumental to human advancement but have also significantly altered our planet. As ecosystems face ever-increasing pressures from modern human activities, Earth observation (EO) has become a valuable tool [1]. Serving as our eyes from above, EO systems provide vital insights into our interactions with the natural environment, playing a crucial role in monitoring and mitigating the effects of modern human-induced environmental changes [2].

The ability of EO to map human influence at multiple scales offers a consistent, repeatable, and scalable method to capture various indicators of human activity and its impact, making it especially valuable for monitoring remote areas without further environmental strain. However, this era of data abundance presents a complex paradox: while the vast volume of data enables the advancement of data-driven deep learning methods, it also significantly complicates our analyses. This complexity underscores the need for developing algorithms specifically designed to address the challenges of EO data [3].

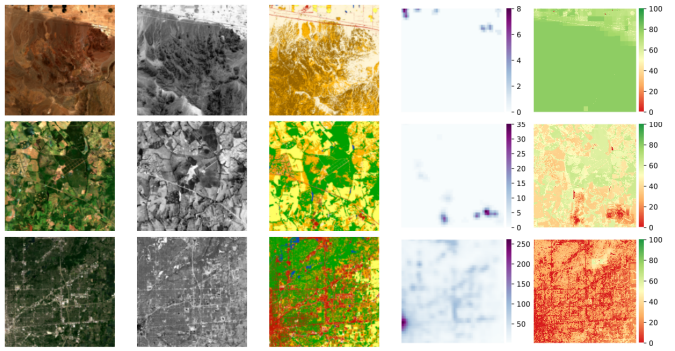

Fig. 1. Sample images (dataset IDs 372358, 19442, and 900000093). From (Left) to (Right): Sentinel-2 RGB bands, Sentinel-1 VH band, ESA's WorldCover landcover map, VIIRS nighttime lights data, and land naturalness annotation.

Despite the benefits, the massive amount of data from EO can be overwhelming and confusing. EO data come in various modalities, each acquired or produced through different methods, offering unique pieces of information. This diversity means that for different tasks, some modalities might be more useful than others. However, the complex and often opaque nature of modern deep learning algorithms adds another layer of difficulty, making it harder to interpret which modality the model relies on most when learning to solve a task. Understanding which data modalities are most informative for specific tasks helps in retaining the predictive performance while removing the necessity of additional data modalities, which simplifies model complexity, optimizes resource use, and enhances interpretability. It also allows for more targeted data collection and reduces noise and redundancy, contributing to more robust and efficient models.

II. Interpretable Machine Learning in EO

Earth observation generates an abundance of data across various modalities. These diverse data sources can be incredibly rich in information, capturing different aspects of the Earth's surface. Given this complexity and the sheer volume of data, traditional methods of feature extraction and

Received 9 August 2024; revised 7 September 2024; accepted 10 September 2024. Date of publication 16 September 2024; date of current version 26 September 2024. This work was supported by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) through the 'MapInWild' Project under Grant SCHM 3322/4-1. *(Corresponding author: Burak Ekim.)*

The authors are with the Department of Aerospace Engineering, University of the Bundeswehr Munich, 85577 Neubiberg, Germany (e-mail: burak.ekim@unibw.de; michael.schmitt@unibw.de).

Digital Object Identifier 10.1109/LGRS.2024.3460812

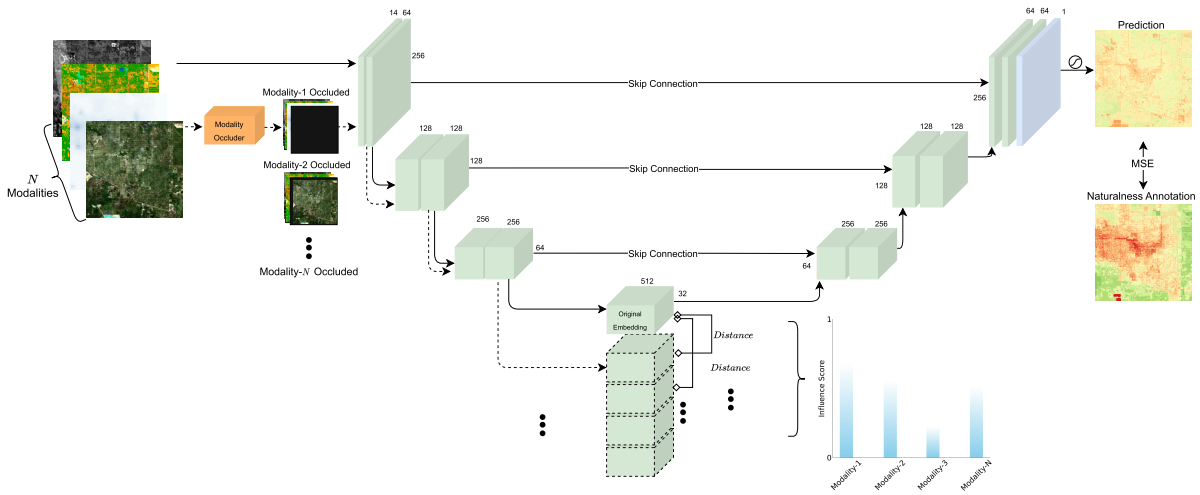

Fig. 2. Framework for assessing land naturalness using multimodal EO data. The iterative occlusion process, shown by dashed lines, repeated "N" times for the N input modalities. In each iteration, one modality is zeroed out to create an occluded embedding. Comparing these occluded embeddings with the baseline from the nonoccluded input reveals the impact of each modality on the model's output.

analysis often fall short. Feature attribution aims to shed light on the contributions of individual features in the decision-making process of machine learning models. By assigning importance scores to each feature, these methods offer a quantitative measure of how each variable influences a model's prediction, thereby enhancing the model's transparency and interpretability.

Various methods have been developed to achieve this, each with its unique advantages and computational considerations [4]. For instance, SHapley Additive exPlanations (SHAPs) [5] use concepts from cooperative game theory to fairly distribute the "contribution" of each feature to a given prediction. On the other hand, local interpretable model-agnostic explanations (LIMEs) [6] operate by perturbing the input data and fitting a simpler, interpretable model to approximate the complex model's behavior for individual predictions. Gradient-weighted class activation mapping (Grad-CAM) [7] provides another avenue by generating heatmaps to highlight regions in the input image that are most influential for prediction. The paper [8] shows that interpretable ML methods, such as LIME and SHAP, assist in clarifying the decision-making mechanisms of multilabel deep learning models in remote sensing tasks, improving both interpretability and trustworthiness. Channel attention networks (CANs) [9] are deep learning models optimized for multispectral imagery. The authors evaluate them on the SpaceNet semantic segmentation dataset. The CAN approach not only outperforms the existing models in segmentation accuracy but also improves the interpretability of the network function through its soft attention mechanism, which effectively allocates attention away from noisy channels. The paper [10] finds that attention mechanisms improve the performance of specific CNN backbones, such as SegNet and U-Net, in building segmentation tasks. Using DeepLIFT, the study also further shows that these mechanisms improve model interpretability and detection accuracy without adding much computational complexity. The study [11] introduces a new multiscale spectral–spatial attention network for hyperspectral image classification. The network combines 2-D octave convolution and 3-D DenseNet to extract complex spatial and spectral features. As for feature extraction and network performance, the authors use two attention mechanisms bottleneck attention module (BAM) and efficient channel attention (ECA). BAM is used to assign proper weight values to each spectral band, effectively suppressing insignificant spectral bands and reducing redundancy. ECA is applied in both the spatial and spectral feature extraction subnetworks to the interactions among feature channels, thereby boosting the network's feature extraction capability.

Although there has been considerable effort in using interpretation and feature attribution methods to better understand models in a post hoc manner, equipping models intrinsically with such methods has received relatively less attention. Exploring the potential of models with inherent feature attribution capabilities could pave the way for better utilization of vast amounts of EO data by focusing efforts on the most important modalities for the task at hand. In this letter, we propose a deep learning framework equipped with interpretability capabilities. The proposed framework enables input feature attribution by uncovering the influence of input features (i.e., formulated as modalities) through systematic occlusion, shedding light on how each modality contributes to the deep learning model's predictions.

III. Methodology

We use a vanilla U-Net architecture that starts with a DoubleConv block, expanding the input to 64 channels. This is followed by four downsampling layers and four upsampling layers. The downsampling process employs max pooling to increase the number of channels from 64 to 512 while reducing the spatial dimensions. Conversely, upsampling uses bilinear interpolation to restore the original dimensions, decreasing the number of channels back to 64. The final layer produces a single-channel output, matching the input size, with LeakyReLU activation applied. The first convolutional layer is designed to accept four input modalities, as shown in Fig. 1,

with a total of fourteen channels across these modalities. The entire architecture comprises approximately 17 million trainable parameters.

As for the interpretability component, as illustrated in Fig. 2, we systematically analyze the impact of different modalities on the model's representations by using an occlusion strategy. This involves generating two sets of embeddings: one from the original input and another from the input with specific modalities occluded, effectively reducing those modalities to a zero state. The purpose is to discern how the absence of a given input modality affects the model's embedding space and, by extension, its predictions.

To quantitatively assess the influence of each modality, we compute the Euclidean distance between the embeddings from the original and occluded inputs. A larger distance indicates a significant alteration in the embedding due to the occlusion, suggesting that the occluded modality plays a crucial role in the model's decision-making process. Conversely, smaller distances imply minimal change, indicating the occluded modality's lesser importance.

It is crucial to note that the distances provide relative rather than absolute information about modality importance. The comparison of distance metrics across different modalities offers insights into their relative contributions to the model's predictions. This approach not only helps in understanding the model's sensitivity to specific input features but also assists in feature selection and model interpretability enhancements. The choice of Euclidean distance is driven by its straightforward interpretation and ability to measure the magnitude of change in the model's representation, providing a clearer understanding of each modality's impact.

IV. Implementation Details and Experiments

A. Dataset and Training

In this study, we build on the MapInWild dataset [12], a comprehensive dataset originally curated for the task of wilderness mapping, by expanding it with a newly curated annotation source called the naturalness index (NI) [13].

MapInWild consists of data from diverse modalities, including Sentinel-2, Sentinel-1, VIIRS nighttime lights, and ESA WorldCover. The dataset comprises 8144 images, each with a size of 1920 × 1920 pixels, amounting to approximately 350 GB in total. Beyond its diverse set of modalities, MapInWild serves as an ideal test bed for our framework due to its representativeness and the diversity of its samples. To ensure, MapInWild's AOIs are sampled from the World Database of Protected Areas using a climate map and a land cover type-aware semiautomated approach, guided by weights calculated from the Köppen–Geiger climate classification map and the ESA WorldCover map. These weights are inversely normalized to ensure that underrepresented polygons received adequate sampling, thereby enhancing the dataset's spatial coverage and representability. Seasonal variations and hemispheres are also accounted for, making the dataset versatile and robust.

As for the newly introduced annotation source, the NI maps are calculated at a 10 × 10 m spatial resolution and are based on four proxies indirectly measuring human absence: population density, land transformation, accessibility, and electrical power. We refer the readers to [13] for a detailed explanation of proxies. The NI maps accompanying MapInWild dataset range between 0 and 100 representing the level of naturalness, with the higher score representing a higher degree of naturalness. Sample geodata from MapInWild expanded with newly created naturalness annotations are shown in Fig. 1.

We concatenate all of the inputs and form a data cube of 14 channels (10 × Sentinel-2, 2 × Sentinel-1, single-band ESA WorldCover, and single-band VIIRS DNB). Sentinel-2 bands are normalized by dividing with 10 000, and the other sources are normalized with their mean and standard deviation values calculated on the train set. We use the dataset in its original shape and apply on-the-fly random cropping with shape 256 × 256 pixels and apply a series of augmentation methods to improve the model's robustness, including random horizontal and vertical flips, sharpness adjustments, random erasing, and Gaussian blurring. To mitigate the class imbalance in the annotations, we feed the inverse mean value of each patch as sample weights to the model to balance the representation of the NI values.

We use the mean-squared error as the criterion for the loss function and the Adam optimizer with an initial learning rate of 10^{-4}. The training uses a batch size of 16 and leverages early stopping based on validation error, with a patience of 40 epochs to address overfitting. We implement a cosine annealing learning rate with warm restarts, setting the initial cycle to three epochs and maintaining the same length for each subsequent cycle. The scheduler adjusts the learning rate based on epoch intervals, aiming to fine-tune the model's training process for optimal performance. We exclude no-data values from loss and error calculations, as well as from influencing the sampling weights. We use an 80:10:10 split ratio for training, validation, and test sets.

B. Experiments

The experimental setup consists of three main steps. First, we collect sample-level and test set-level modality influence values during the test phase to study the deep occlusion method in a post hoc way. To validate these influence values, we conduct an ablation study by passing each modality to the model individually and observing the level of agreement with the test set-level modality influence values. Finally, to explore the versatility of the *modality occluder*, we investigate alternative methods for transitioning input features from an information state to a zero state, such as using zeros and random noise. It is important to note that the proposed approach performs the embedding-level distance calculation and subsequent evaluations during test time.

V. Results and Discussion

A. Post Hoc Investigation: Sample-Level Influence Values

We calculate the influence values for each test set sample and investigate the individual influence values to enable sample-level interpretations. These results provide an understanding of individual input modalities and their assigned

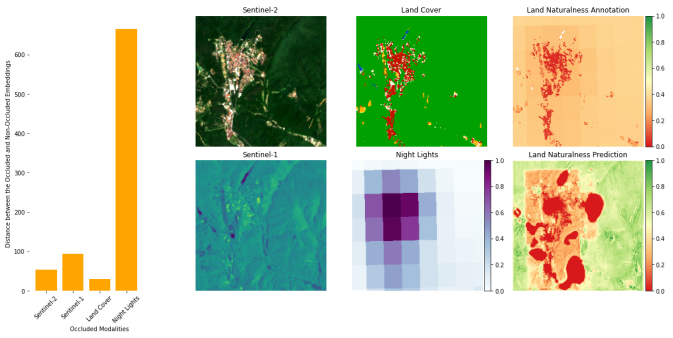

Fig. 3. Sample-level influence scores, highlighting the impact of occluding the nighttime lights modality on the model's output. The dominant influence of nighttime lights is reflected in the significantly larger distance on the embedding level.

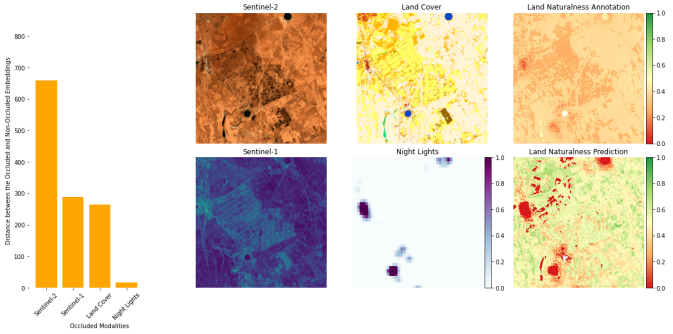

Fig. 4. Sample-level influence scores, highlighting the impact of occluding the Sentinel-2 modality on the model's output, with other modalities influencing the prediction to certain extents. The relatively higher impact of Sentinel-2 is reflected in the larger distance on the embedding level.

attribution values during test time. Fig. 3 reveals that in the context of residential areas, the *nighttime lights* modality exhibits a strong influence on the model's predictions, which is depicted by the significant embedding distances when this modality is occluded. The sparsity of nighttime lights may be related to the modality providing easily discernible cues when present—likely due to its distinctive and sparse signal that provides clear indications of human activity. Furthermore, Fig. 4 illustrates a scenario where Sentinel-2 has the greatest influence on the model's prediction, followed by Sentinel-1 and *land cover*. While this focuses on modality attribution values for a single sample, we extend this analysis to the set level by averaging the attribution values across all samples. The results of this aggregated analysis are presented in Fig. 5, as described in Section V-B.

B. Agreement Between Individual Modality Performance and Test Set-Level Influence Values

We benchmark the test performance when each modality is individually input to the model. The mean absolute error (MAE) and root-mean-squared error (RMSE) of the test set are presented in Table I. Sentinel-2 and Sentinel-1 stand out with the lowest MAE and RMSE, followed by the nighttime lights and land cover modalities, when each modality is used as the sole input. In addition, we conduct experiments where each modality is occluded one at a time, with the model being trained using the remaining modalities.

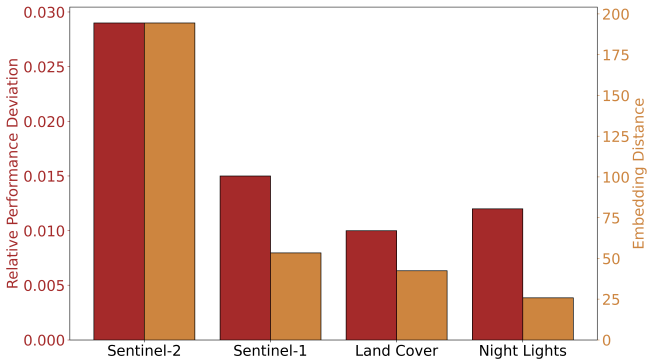

Fig. 5. Comparison of relative performance deviations and embedding distances for each modality. Red bars represent performance deviations from an all-modality scenario, while yellow bars show the embedding distances between occluded and nonoccluded embeddings, measuring their influence.

A noticeable drop in predictive performance is observed in these cases, which we hypothesize is due to the increased data volume without any additional useful information. Furthermore, feeding only the Sentinel-1 modality yields the highest predictive performance, and occluding only the Sentinel-2 modality (*Sentinel-2 occluded*) results in the lowest predictive performance. We hypothesize that the results in this table serve as a proxy for the *importance* of each modality, which aligns with what the proposed method suggests.

We first compared the test MAE values derived from using each individual modality as the sole input to the model against the test MAE from using all modalities combined (*all modalities* in Table I). This comparison involved calculating the absolute difference between the MAE of the all-modalities-present scenario and the MAE of each individual modality. These absolute differences, which we refer to as relative performance deviations, illustrate how the performance of each single modality deviates from the performance achieved when all modalities are combined. This analysis highlights the impact of using individual modalities compared with a scenario where all modalities are utilized.

Then, we calculate the modality influence values over test set with the proposed occlusion-based occlusion embedding distance method. This allows us to assess the relative performance deviation of individual modalities from a scenario where all these modalities are present.

Finally, we compare the test set-level influence values with the relative performance deviations in Fig. 5. The high degree of agreement in the distributions suggests that inferring relative influence score information from latent space could be interpreted as input feature attribution values. Furthermore, in the test set, land naturalness prediction heavily depends on the *Sentinel-2* bands, followed by *Sentinel-1*, while the other modalities have relatively weak influence.

C. Occlusion Strategy Investigation

In the last experiment, we apply four occlusion strategies—filling with zeros, ones, random noise, and Gaussian noise—to the input modalities to observe their impact on the model's representations. Each occlusion method alters input modality a unique way, simulating the absence of information or

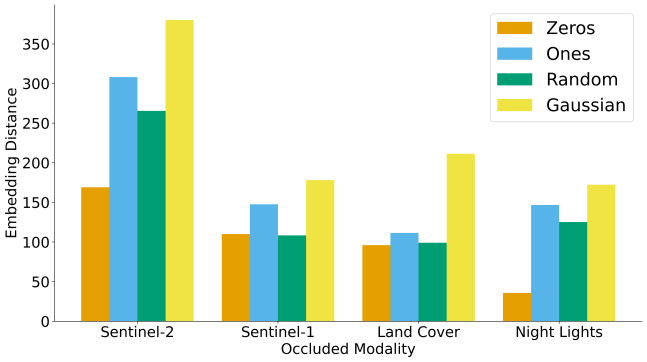

Fig. 6. Comparison of embedding distances for different occlusion strategies across modalities. Each bar represents the embedding space perturbation when a specific modality is occluded with zeros, ones, random noise, or Gaussian noise.

TABLE I
PERFORMANCE METRICS PER INDIVIDUAL MODALITY

Modality	MAE (↓)	RMSE (↓)
All Modalities	0.109	0.136
Sentinel-2	0.125	0.159
Sentinel-1	0.123	0.157
Land Cover	0.128	0.163
Night Lights	0.126	0.157
Sentinel-2 Occluded	0.17	0.194
Sentinel-1 Occluded	0.145	0.166
Land Cover Occluded	0.13	0.161
Night Lights Occluded	0.132	0.173

introducing variability to gauge the model's dependency on the given modality.

The resulting pattern, as shown in Fig. 6, suggests a consistent trend across all occlusion types regarding the relative influence of each modality. Sentinel-2 and Sentinel-1 modalities appear to have more significant impact on the embedding space, as indicated by larger distances, suggesting that these modalities are highly influential in the model's prediction process. Conversely, Sentinel-1 and nighttime lights modalities resulted in smaller embedding space perturbations, indicating a lesser degree of influence. The coherence of these findings across different occlusion strategies reinforces the reliability of the occlusion sensitivity analysis as a method for interpreting model behavior through its manipulated embedding space. The agreement between the occlusion types supports the conclusion that despite the difference in how information is withheld or distorted, the model consistently identifies the same modalities as more or less influential, which is a promising result for the robustness of feature attribution in EO data analysis.

VI. CONCLUSION

This letter proposes a deep occlusion framework embedded within a multimodal learning context, enhancing interpretability through feature attribution on the embedding level. By systematically occluding individual input modalities, our method allows for a nuanced analysis of how each modality influences the model's predictions. Applied to the task of land naturalness mapping, the proposed method reveals that the Sentinel-2 modality has the most significant impact on the model output, with other modalities contributing to varying extents. The proposed framework does not make assumptions about the task or model, enabling it to be applied across different architectures and applications. The consistency between changes observed on the embedding level due to modality occlusion and the corresponding shifts in the prediction output validates the hypothesis about the importance of specific modalities. The results provide deeper insights into how input modalities influence the model decision-making process, both at the sample level—enabling regional case studies—and at the dataset level—facilitating data pruning and improving training and inference times. As future work, we intend to perform this analysis on a band level and explore different occlusion and distance strategies to unlock further insights.

ACKNOWLEDGMENT

The authors acknowledge the University of the Bundeswehr Munich for providing open access funding.

REFERENCES

[1] D. Tuia et al., "Artificial intelligence to advance earth observation: A perspective," 2023, *arXiv:2305.08413*.
[2] D. Tuia et al., "Perspectives in machine learning for wildlife conservation," *Nature Commun.*, vol. 13, no. 1, pp. 1–15, 2022.
[3] E. Rolf, K. Klemmer, C. Robinson, and H. Kerner, "Mission critical—Satellite data is a distinct modality in machine learning," 2024, *arXiv:2402.01444*.
[4] B. Ekim and M. Schmitt, "Explaining multimodal data fusion: Occlusion analysis for wilderness mapping," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2023, pp. 962–965.
[5] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arXiv:1705.07874*.
[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, 2016, pp. 1135–1144.
[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[8] I. Kakogeorgiou and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 103, Dec. 2021, Art. no. 102520.
[9] A. A. Bastidas and H. Tang, "Channel attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 881–888.
[10] E. H. Zaryabi, L. Moradi, B. Kalantar, N. Ueda, and A. A. Halin, "Unboxing the black box of attention mechanisms in remote sensing big data using XAI," *Remote Sens.*, vol. 14, no. 24, p. 6254, 2022.
[11] L. Liang, S. Zhang, J. Li, A. Plaza, and Z. Cui, "Multi-scale spectral–spatial attention network for hyperspectral image classification combining 2D octave and 3D convolutional neural networks," *Remote Sens.*, vol. 15, no. 7, p. 1758, Mar. 2023.
[12] B. Ekim, T. T. Stomberg, R. Roscher, and M. Schmitt, "MapInWild: A remote sensing dataset to address the question of what makes nature wild [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 1, pp. 103–114, Mar. 2023.
[13] B. Ekim, Z. Dong, D. Rashkovetsky, and M. Schmitt, "The naturalness index for the identification of natural areas on regional scale," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, Dec. 2021, Art. no. 102622.