JUNGLE-NET: USING EXPLAINABLE MACHINE LEARNING TO GAIN NEW INSIGHTS INTO THE APPEARANCE OF WILDERNESS IN SATELLITE IMAGERY

T. Stomberg[1]*, I. Weber[2]†, M. Schmitt[3], R. Roscher[1]

[1] Institute of Geodesy and Geoinformation, University of Bonn, Germany - (timo.stomberg, ribana.roscher)@uni-bonn.de
[2] AMLS, University of Applied Sciences Koblenz, Germany - immanuel.weber@hs-koblenz.de
[3] Department of Geoinformatics, Munich University of Applied Sciences - michael.schmitt@hm.edu

KEY WORDS: Scene classification, Explainability, Interpretability, Deep neural networks

ABSTRACT:

Explainable machine learning has recently gained attention due to its contribution to understanding how a model works and why certain decisions are made. A so far less targeted goal, especially in remote sensing, is the derivation of new knowledge and scientific insights from observational data. In our paper, we propose an explainable machine learning approach to address the challenge that certain land cover classes such as wilderness are not well-defined in satellite imagery and can only be used with vague labels for mapping. Our approach consists of a combined U-Net and ResNet-18 that can perform scene classification while providing at the same time interpretable information with which we can derive new insights about classes. We show that our methodology allows us to deepen our understanding of what makes nature wild by automatically identifying simple concepts such as wasteland that semantically describes wilderness. It further quantifies a class's sensitivity with respect to a concept and uses it as an indicator for how well a concept describes the class.

1. INTRODUCTION

Machine learning (ML) methods are successfully used in remote sensing for various tasks such as classification, detection, or parameter prediction. In general, the main goal of these tasks is high accuracy and high efficiency. However, especially for scientific applications also other characteristics such as the comprehensibility and reliability of the results are considered important in order to ensure the scientific value of the outcome and to increase trust in the learned models. Besides the actual solving of the application task and the mere learning of relationships between observed data and the desired output, a recent but not yet widespread use of ML is the derivation of new scientific knowledge (Roscher et al., 2020a). In order to get closer to such goals, explainable ML has been strongly promoted in research in recent years.

Explainable ML aims at the understanding of the underlying reasons for the produced decisions and in which way a particular model works (Samek et al., 2020). Little work has been done so far in remote sensing, particularly with satellite images, to understand better what has been learned. (Roscher et al., 2020b) discuss first works in this direction and show that explainability is often used to align the models with existing knowledge, for example, to improve models and to correct obvious flaws in case of wrong decisions. To this point, explainable ML has been used less to uncover previously unknown patterns and to derive novel scientific insights.

One possible application of explainable ML to uncover unknown patterns is the mapping of only weakly defined phenomena such as *wilderness*. Although there is a plethora of scientific work discussing this topic from a philosophical perspective (e.g., (Bastmeijer, 2016)), there is no clear physical definition of what makes nature wild. However, we see a high relevance for mapping wilderness areas using remote sensing observations, as this can be an important source of information for stakeholders in the context of establishing new protected areas.

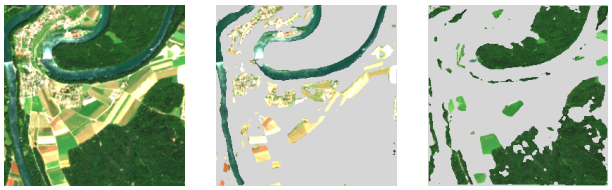


Figure 1. Sample satellite image with two derived concepts that semantically describe water and man-made structures on the one hand and forests on the other hand.

Here, we propose an explainable ML approach to derive novel scientific insights. We analyze and evaluate our approach for satellite-based classification of wilderness, which is only vaguely defined from a technical perspective. This uncertain definition hinders an automatic mapping of *wilderness* with an ML approach. For example, the choice of reference data is not clear, and thus the development of a suitable model cannot be ensured. As a weak label proxy, we use the structural classification as a protected wilderness area, which can be derived from administrative data provided by the International Union for the Conservation of Nature (IUCN). However, since such data is not all-encompassing and actual wilderness may also exist outside such protected areas, this annotation may be incomplete and noise-prone.

The contributions of our work are the following:

- Common interpretation methods are tailored to RGB images and objects characterized by strong image gradi-

*Corresponding author
†T. Stomberg and I. Weber equally contributed to this work.

ents compared to the background. In contrast, the scenes of interest in this work represent natural areas and are therefore not characterized by distinctive image gradients. Moreover, they do not represent a typical foreground-background division of the scene, which is commonly used for explanatory analysis (Adebayo et al., 2018). Also, the use of multi-spectral satellite imagery provides us with more bands, which we use for our analyses.

- Due to the uncertain definition of *wilderness*, existing labels only vaguely describe the scenes. This is taken into account when designing the explainable ML method by addressing wilderness identification through scene classification rather than semantic segmentation, which removes the burden of requiring accurate location-specific information.
- We generate potential concepts in a human-understandable space, that are visually interpretable and explainable with domain knowledge (see Fig. 1). By concept, we mean something that can be described semantically. In our case, two concepts cover surface types as water, man-made structures and forests, which are sensitive towards the complex land cover phenomenon *non-wilderness*. Unlike common methods that gain explainability for novel concepts by using a relation to existing concepts, we exploit the high interpretive power of our models' internal representation, which can be explained.

Overall, to make a step in the direction of better land cover mapping, we propose an explainable ML approach to deepen our understanding of what makes nature *wild* and derive novel insights about this land cover class.

2. RELATED WORK

Among the most notable remote sensing applications is the automatic mapping of land cover from satellite images (Ma et al., 2018, Zhang et al., 2016). While in most cases land cover mapping is addressed generically, i.e., aiming at comprehensive land cover class schemes, many works target the identification of specific classes, e.g., water (Isikdogan et al., 2017), croplands (Kussul et al., 2017), or urban areas (Qiu et al., 2020). However, unlike these commonly used land cover classes or other well-defined classes (e.g., defined by Copernicus Global Land Service (Buchhorn et al., 2020)), the land cover class `wilderness` is only vaguely defined from a technical perspective, and there is no clear physical delineation of what is a wilderness area. In such cases, accurate mapping is limited, and more sophisticated ML methods must be applied.

A promising direction to overcome this challenge is explainable ML, which can be used to derive novel insights, for example, into the characteristics of classes (Tuia et al., 2021). According to (Roscher et al., 2020a) three properties are beneficial to learn an explainable ML model: transparency, interpretability, and explainability. Transparency is the accessibility to properties such as the model's structure or the motivation why specific components were chosen. Interpretability is the property of representing certain processes in an ML model in a human-understandable space. At the same time, explainability is achieved by combining these interpretations with domain knowledge in the context of a particular application. These three properties are found in different approaches with different levels. An overview in the context of remote sensing can be found, for example, in (Roscher et al., 2020b). Although a model may have been learned using various ML techniques, the goals of explainable ML are mostly mentioned in the context of deep neural networks, whose decision-making is often difficult for humans to comprehend due to their complexity.

There are mainly two groups of approaches used to increase interpretability and explainability when combined with domain knowledge: post-hoc interpretation methods and interpretation-by-design methods. Post-hoc interpretation methods analyze the outcomes and decisions of an already learned model utilizing the input. (Samek et al., 2020) discuss several local and global post-hoc interpretation tools that visualize processes in a neural network in different ways. Especially in the field of image analysis, heatmaps are often used to mark regions relevant for the decision-making process (Kierdorf et al., 2020, Lapuschkin et al., 2019). Among the most common methods for producing heatmaps are layer-wise relevance propagation (Montavon et al., 2019) and occlusion sensitivity maps (Zeiler, Fergus, 2014). In contrast to post-hoc interpretation methods, interpretation-by-design methods produce interpretations and explanations by an imposed representation of model components or latent variables that we can link to domain knowledge. Here, one of the most prominent works in the field of neural networks is network dissection, where units in neural networks are linked to human-understandable concepts (Zhou et al., 2018a). The identification and disentanglement of such concepts has long been the subject of research (Arendsen et al., 2020, Marcos et al., 2019, Zhou et al., 2018b, Kim et al., 2018, Wigness et al., 2014). However, the approaches generally use existing concepts from other annotated image datasets such as Broden (Bau et al., 2017) or embed the concepts with other databases into a common space.

3. DATA

In the following, we present the data we use for finding concepts for wilderness with our presented framework.

Dataset preparation For the investigations in this paper, we have created a dataset connecting optical multi-spectral imagery acquired by the Sentinel-2 satellites with annotations indicating `wilderness`. For that purpose, we developed an automated data processing chain based on Google Earth Engine (Gorelick et al., 2017), which is depicted in Fig. 2.

While the Sentinel-2 data preparation mainly consisted of temporal mosaicking to ensure cloud-free observations, the more challenging aspect in this context was to annotate wilderness areas with *wilderness* being a term that is only vaguely defined from a technical perspective. To circumvent this problem, we relied on the definition of IUCN, which provides a classification of protected areas. Their Category Ib is called *wilderness area* and described as follows: "These areas are a protected domain in which biodiversity and ecosystem processes (including evolution) are allowed to flourish or experience restoration if previously disturbed by human activity." To emphasize the difference to our understanding of wilderness, we denote this category *IUCN wilderness area* in the following.

For our dataset, we created a spatial buffer around IUCN wilderness areas represented as spatial polygons in the World Database on Protected Areas (WDPA)[1] for Europe. We then rasterized the resulting rectangle to a pixel spacing of 10 m (i.e., corresponding to the maximum spatial resolution of Sentinel-2),

[1] `https://www.protectedplanet.net/en/thematic-areas/wdpa`

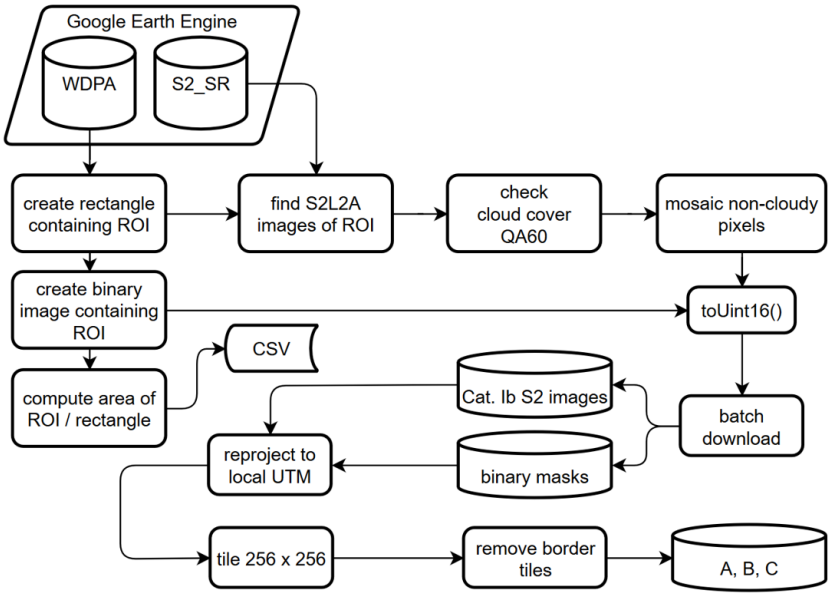
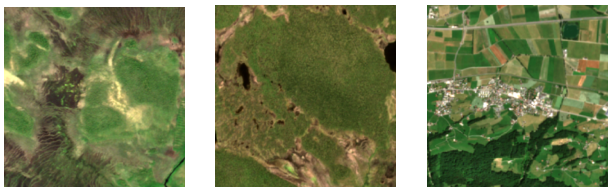


Figure 2. Flowchart of the dataset preparation in Google Earth Engine. A, B, and C refer to training, validation, and test sets, respectively. The abbreviations are defined as follows: WDPA – World Database of Protected Areas, S2_SR/S2L2A – Sentinel-2 Level 2A imagery, QA60 – 60m resolution quality band of Sentinel-2, ROI – region of interest, CSV – comma-separated value file, UTM – Universal Transverse Mercator coordinate system.



(a) (b) (c)

Figure 3. Three example images from the dataset: (a) positive example indicating a wilderness area; (b) hard negative example extracted in the vicinity of a wilderness area; (c) easy negative extracted from a populated area in Europe.

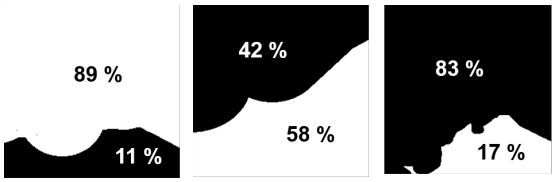


Figure 4. Conversion of pixel-level binary annotations into single-label scene annotations. Patches with more than 80% of wilderness pixels receive label `wilderness` (left). Patches with relatively similar shares of wilderness and non-wilderness pixels are removed (center). Patches with less than 20% of wilderness pixels received the scene label `non-wilderness`.

transforming it to a binary target layer. Since areas in the surroundings of IUCN wilderness areas can be expected to be visually at least similar to the protected areas, those non-wilderness areas can be considered hard negatives samples. To ease the network's training, especially during the first iterations, we added a few simple examples sampled from randomly selected European cities. Last, the data was split into tiles with a size of 256×256 pixels, which corresponds to an area of about 6.6 km^2. Positive, hard negative, and easy negative examples are shown in Fig. 3.

Since this work uses scene classification as a backbone task rather than semantic segmentation, we further reduced the pixel-level binary annotations to scene labels `wilderness` and `non-wilderness`. This was done by keeping only samples showing less than 20 % or more than 80 % of IUCN wilderness areas for negative and positive examples, respectively (cf. Fig. 4). In the end, this results in a dataset of 5300 images.

Data Split Most of our collected `wilderness` data comes from Sweden, Finland, and Estonia. `Non-wilderness` samples mainly come from large European cities as Berlin, Zurich, or Vienna. We divide our data into three independent and spatial consistent subsets for training, validation, and testing. To do so, we spatially cluster all data samples as described in the following: First, data samples with a distance of more than 5 km are separated. This leads to a few large clusters so that in a second step, clusters with more than 100 samples are split into several clusters using the k-means algorithm. Small clusters with less than ten samples are added to the training set to encourage versatile training. The remaining clusters are assigned randomly but uniformly in size until the final split fractions of the datasets are about 60/20/20 %. Thus, we get three disjunct datasets of 3055 samples for training (1066 of them labeled as `wilderness`), 1162 (274) samples for validation, and 1082 (345) for testing.

4. METHODOLOGY

Fig. 5 schematically shows our framework. One part of our framework, illustrated with a red box, is a deep neural network for scene classification. The input is a $(H \times W \times B)$-dimensional image, and the output is a K-dimensional vector indicating the estimated class of the scene. In our case, the output is a two-dimensional confidence score for the classes `non-wilderness` and `wilderness`. The second part of our framework, illustrated with a blue box, identifies concepts in a dataset. In the following, we will explain the single components in more detail.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2021-317-2021 | © Author(s) 2021. CC BY 4.0 License.

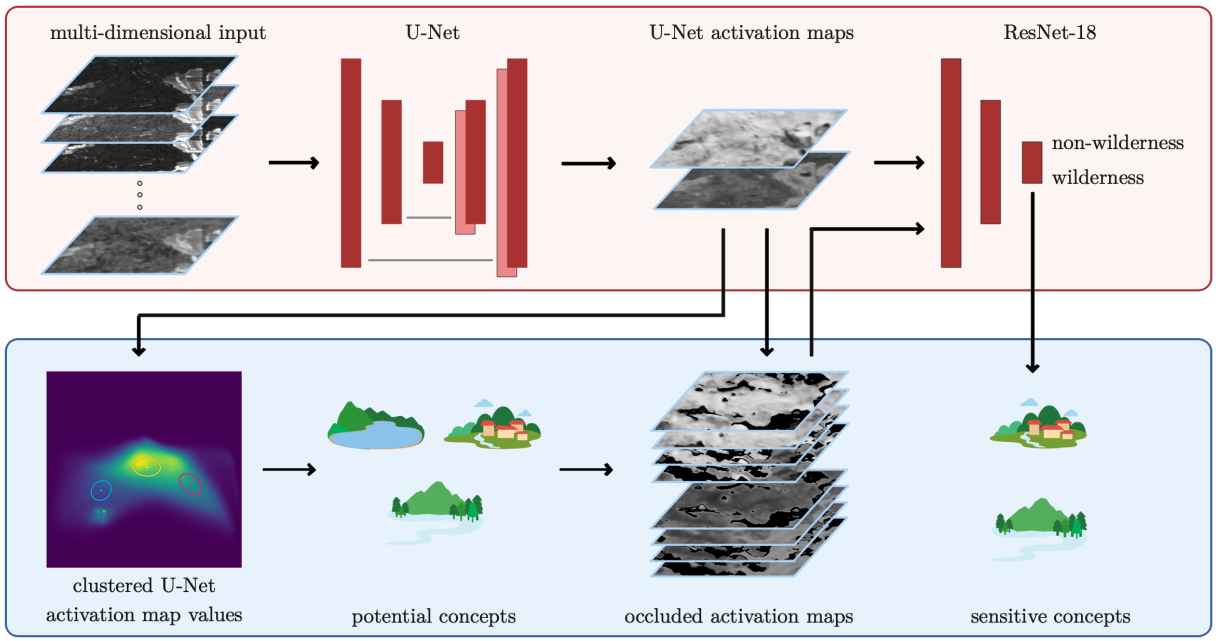


Figure 5. Overview of the jUngle-Net and our methodology. The scene classification pipeline (red box) classifies multi-dimensional images into the classes `wilderness` or `non-wilderness`. The interpretation pipeline (blue box) derives potential concepts by clustering U-Net activation maps of correctly classified images. Sensitive concepts belonging to the classes are identified by comparing the confidence scores of the ResNet-18 obtained with non-occluded U-Net activation maps and concept-occluded activation maps.

jUngle-Net architecture Due to the vague definition of the target phenomenon *wilderness* and the use of proxy labels, we do not aim for a pixel-wise semantic segmentation of the input image data but reduce the problem of scene classification into a proxy `wilderness` class. To gain scientific knowledge about *wilderness*, we are building an interpretation-by-design network that will classify it while providing introspection into the network's internal representation. The first part of jUngle-Net is an encoder-decoder-based U-Net (Ronneberger et al., 2015) that transforms input data of the form $(H \times W \times B)$ into a new representation of the form $(H \times W \times L)$. This intermediate representation is then passed to the second part, a ResNet-18 (He et al., 2016), which classifies it into K classes. We choose $L = K$ so that the U-Net's feature dimension is equal to the number of classes. Since $L < B$, the interface between the two networks is a bottleneck, which enforces a feature reduction while maintaining the spatial extent.

In detail, we use a U-Net structure with four downward steps and replace the transpose convolution layers of the upward steps with bi-linear up-sampling operations, as described in (Odena et al., 2016). This upsampling prevents checkerboard artifacts in the outputs, which we observed with the standard transposed convolutions. To account for the larger number of input feature dimensions compared to RGB images, we modify the first convolutional layer of the U-Net to put out 128 rather than 64 feature dimensions. We continue this doubling of dimensions over the rest of the network. The U-Net's final layer is activated with the tanh function, to get activation map values in the range of -1 and 1. We use the a ResNet-18 with sigmoid activation to output class scores. In contrast to softmax activation, sigmoid allows outputs in which both classes - `wilderness` and `non-wilderness` - are not forced to sum to one and can be seen as a more independent class uncertainty estimate. For the model's training, we use the binary-cross-entropy loss. We train both networks end-to-end, and therefore both are optimized together. The representation that emerges during training at the interface between the two networks is the base for our subsequent concept identification.

Identifying potential concepts For the identification of specific concepts, we map all vectors $\boldsymbol{u}_{h,w}$ from each position in the $(H \times W \times L)$-stack of U-Net activation maps to an L-dimensional space to analyze which combinations of activations are most common. In doing so, we double the number of `wilderness` samples to get a more balanced amount of `wilderness` and `non-wilderness` samples. This way we assume the following clustering algorithm not to prefer `non-wilderness` concepts. Assuming a Gaussian mixture distribution of the mapped activations, we determine the regions of maximum density using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). We compute C clusters using EM, which we consider as potential concepts. Each cluster is described by a mean value, a covariance matrix, and a mixing coefficient. This allows an assignment of each activation map point to one of the clusters by the argument of the maximum of the likelihood values. Since image and activation map have the same size, we can assign a concept to each image pixel as well.

Quantifying concept-occlusion sensitivity We quantify the sensitivity of each potential concept by occluding their areas in the U-Net activation maps. Our method is based on the idea of occlusion sensitivity maps (OSMs) developed by (Zeiler, Fergus, 2014). OSMs are heatmaps indicating the sensitivity of a trained neural network to partial occlusions in the input image. To produce an OSM, a patch is moved over the image with a certain stride and occludes pixels by replacing their values with a fill value. Deviations in the output scores resulting from the occlusions indicate the influence of the image regions on the result. Both positive and negative contributions to the output score can occur.

Compared to standard OSM, there are mainly three modifications in our approach: First, instead of occluding parts in the input image, we occlude parts of the U-Net activation maps before passing them to the ResNet-18. Second, instead of occluding with fixed-size patches, we occlude with areas representing a potential concept, i.e., as indicated by the concept map. Third, instead of presenting the sensitivity of patch-wise occlusion in a heatmap, we evaluate the impact of covering single potential concepts per image. We then evaluate these results for multiple images in histograms to analyze whether the concepts are sensitive and class-specific.

We occlude the U-Net activation maps with zero and pass them to the ResNet-18, which provides K predicted scores in a K-dimensional vector $\mathbf{y}_i$ for each occlusion i. We compare these values with vector $\mathbf{y}_{\text{orig}}$, which is obtained by the non-occluded activation maps, and consider the absolute deviations $\Delta_i = \mathbf{y}_{\text{orig}} - \mathbf{y}_i$ as a measure for concept-sensitivity.

5. EXPERIMENTS

5.1 Experimental Setup

Our code to train and validate our model is based on pytorch and pytorch-lightning. We standard-normalize the input data using band-specific mean and standard deviation values based on the 95 % quantile of the complete dataset in the training procedure. To increase data variability during training, we apply random horizontal and vertical flipping as image augmentation. We train the model for 40 epochs with a batch size of 32 samples, which takes about 60 minutes with an NVIDIA Tesla V100. The SGD optimizer is used with a learning rate of 10^{-3}, a momentum of 0.95, and a weight decay of 10^{-3}. For warm-up and annealing of the learning rate and momentum, we employ the one-cycle scheduler.

For our investigations, we classify all test data samples with the trained model and save the confidence scores from ResNet-18 and the U-Net activation maps. We treat the number of Gaussian mixture components C as a hyperparameter and tune it to find the best concepts with regard to interpretabilty and sensitivity. We choose the number of potential concepts to be $C = 3$.

5.2 Results and Discussion

5.2.1 Scene classification The confusion matrix for the scene classification results on the test dataset is shown in Table 1. The resulting accuracies for the two classes `non-wilderness` and `wilderness` are 0.76 and 0.84. The F1-scores are 0.71 and 0.83, respectively.

Table 1. Confusion matrix of the scene classification results of classes `non-wilderness` (`nw`) and `wilderness` (`w`).

		target		
		`nw`	`w`	Total
prediction	`nw`	558	55	613
	`w`	179	290	469
	total	737	345	1082

5.2.2 Finding potential concepts To find potential concepts, we use the described procedure in Sec. 4 and cluster the vectors $\boldsymbol{u}_{h,w}$ in the stack of U-Net activation maps. We only use correctly predicted test samples, which comprise 558 correctly predicted samples for `non-wilderness` and 290 correctly predicted samples for `wilderness`. The empirical densities in Fig. 6 show that nearly all activation map values concentrate around a few hot spots. It is also apparent that densities are more distributed across the value range of the first U-Net activation map than the range of the second U-Net activation map, which results in a larger spread of the clusters along the first range. We identify clusters characterized by a high density of activation values with the EM-algorithm, but also alternative approaches such as mean-shift can be used. We obtain the best results with regard to interpretability and sensitivity with $C = 3$ clusters. Choosing a higher number of clusters, e.g. $C = 4$, leads to concepts being similar in large parts and therefore less sensitive. Choosing a lower number, e.g. $C = 2$, leads to a less interpretable mixture of concepts.

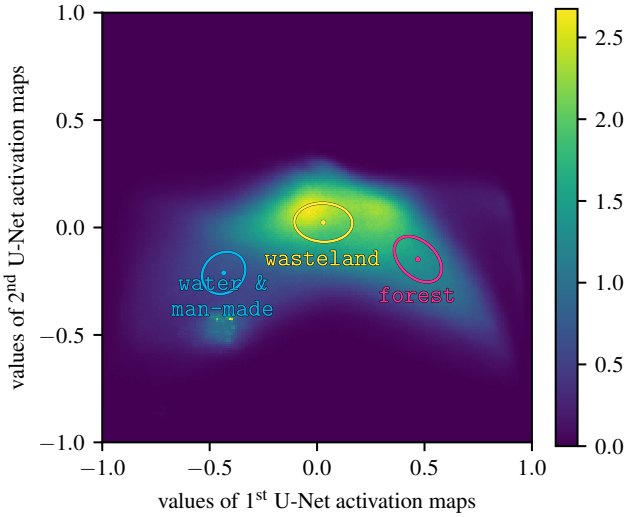
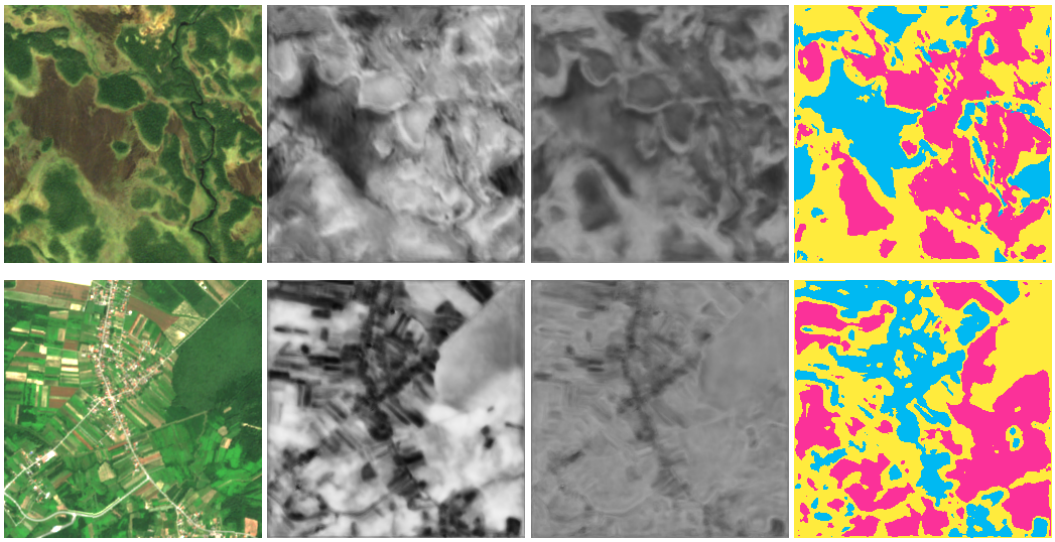


Figure 6. Empirical densities of the two U-Net activation maps aggregated from all samples. The ellipses represent the 1-sigma isolines of the three Gaussian mixture distribution components. The points mark their mean values.

We cluster the U-Net activation maps based on the found cluster parameters and use them to segment the input images. As an example, we show the segmentation of two samples in Fig. 7 labeled as `non-wildness` and `wildness` that show an urban and a wild area. Besides the input's RGB bands, we display both U-Net activation maps and the segmentation derived using the three clusters found in the activation maps.

We identify surface types covered by these clusters to associate each cluster with a concept. Although each concept can be associated with several surface types, for simplicity, we name each concept only by its most frequent one(s). Typical examples for each of the three clusters are shown in Fig. 8, where we masked areas gray that are not part of the respective cluster. The first concept, denoted as `water & man-made`, mainly represents water bodies, man-made structures, and rocky and sandy surfaces. The second concept, denoted as `forest`, comprises different forest and other high vegetation types. The last concept is diverse, but the majority contains wastelands and rough surfaces. We refer to it as `wasteland`.

5.2.3 Concept-occlusion sensitivities In this experiment, we verify whether the identified concepts are connected to the `non-wilderness` or `wilderness` class by analyzing concept-occlusion sensitivities. The histograms in Fig. 9 show the confidence score deviations due to occlusions in the U-Net activation maps: those on the left hand side show the deviations for



(a) RGB (b) U-Net map 1 (c) U-Net map 2 (d) Concepts

Figure 7. RGB bands, U-Net activation maps, and derived concept segmentations (cyan: `water & man-made`, magenta: `forests`, yellow: `wasteland`) of two samples (top: `wilderness`, bottom: `non-wilderness`).

`non-wilderness` samples (Fig. 9a) and those on the right hand side show the deviations for `wilderness` samples (Fig. 9b).

The sensitivity of the `water & man-made` concept in `non-wilderness` samples can be reflected in the top histogram of Fig. 9a. The deviations of the `non-wilderness` score (blue bars) are negative, which means that the occlusions affected the confidence score towards the `wilderness` class. The histogram corresponding to deviations of the `wilderness` confidence score (green bars) shows the opposite behavior, which also means a change towards `wilderness`. The wide spread of the histograms shows that large changes to the classification score and even complete switches of the classification result occur. Consequently, we assign the concept `water & man-made` to be connected to the `non-wilderness` class. The corresponding histogram on the right-hand site in Fig. 9b shows the deviations when occluding the `water & man-made` concept in `wilderness` samples. Here, the shift towards the `wilderness` class is significantly less distinct. We assume that the classifier's decision is less impacted by that when it is already confident, that a sample is of class `wilderness` and that it can rely on more information than contained just by that concept. Compared to the `water & man-made` concept, we see similar distributions for the `forest` concept (middle row in Fig. 9), which therefore is also connected to the `non-wilderness` class. For the concept `wasteland` (bottom row), we see the opposite behavior. Here, the occlusion leads to deviations towards the `non-wilderness` class for samples classified as `wilderness` (Fig. 9b). However, it is notable, that the occlusion of the `wasteland` concept has a smaller effect on `wilderness` samples, than the occlusions of the `water & man-made` and `forest` concepts have on `non-wilderness` samples (Fig.9a). We therefore regard the `wasteland` concept to be slightly sensitive towards the `wilderness` class.

5.2.4 Interpret sensitive concepts We found two concepts, `water & man-made` and `forest`, that are clearly connected to the `non-wilderness` class, whereas we found one concept that is slightly sensitive to `wilderness`. It seems, that it is easier for the model to find `non-wilderness` than `wilderness` indicators. We assume that this is because 1.) there are easy-detectable `non-wilderness` indicators such as cities or streets, but there are more difficult ones for `wilderness`, and 2.) IUCN wilderness areas do generally not show easy `non-wilderness` indicators like streets, whereas non-IUCN wilderness areas might show many indicators for `wilderness`.

However, we observe, that the `wasteland` concept is more sensitive in samples labeled as `wilderness`, than it is in samples labeled as `non-wilderness`. As shown in Fig. 7, the `wasteland` concept often contains fields in `non-wilderness` samples, whereas it often contains wastelands in `wilderness` samples. It appears, that the model uses information from the `wasteland` concept more in samples labeled as `wilderness` and hard `non-wilderness` samples.

On a broader view, we consider `wilderness` to be more characterized by the absence of some concepts rather than others' existence. This is especially true for the concept `water & man-made`, where the latter aspect means that less human impact is observable in a scene.

6. CONCLUSION AND FUTURE DIRECTION

We have presented an interpretation-by-design network for the derivation of novel scientific insights, in particular, for the refinement of the description of ill-defined classes. We investigated jUngle-Net in the context of remote sensing for the vaguely defined target phenomenon *wilderness*. JUngle-Net consists of two parts that are trained end-to-end, namely a ResNet-18 that is primarily responsible for scene classification and a U-Net that provides an interpretable representation. The interpretable representation is visually human-understandable and can be explained with domain knowledge leading to potential concepts for wilderness. Our results show that potential concepts can be found and verified with interpretation tools like occlusion sensitivity. Overall, with our work, we see a starting point in using ML methods that go beyond the classical goal of maximizing accuracy but lead us to new scientific insights. We believe that our approach is as well applicable to data in other study areas to find sensitive concepts that explain the decision of jUngle-Net. We hope to encourage other researchers to try similar approaches for their applications.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2021-317-2021 | © Author(s) 2021. CC BY 4.0 License.

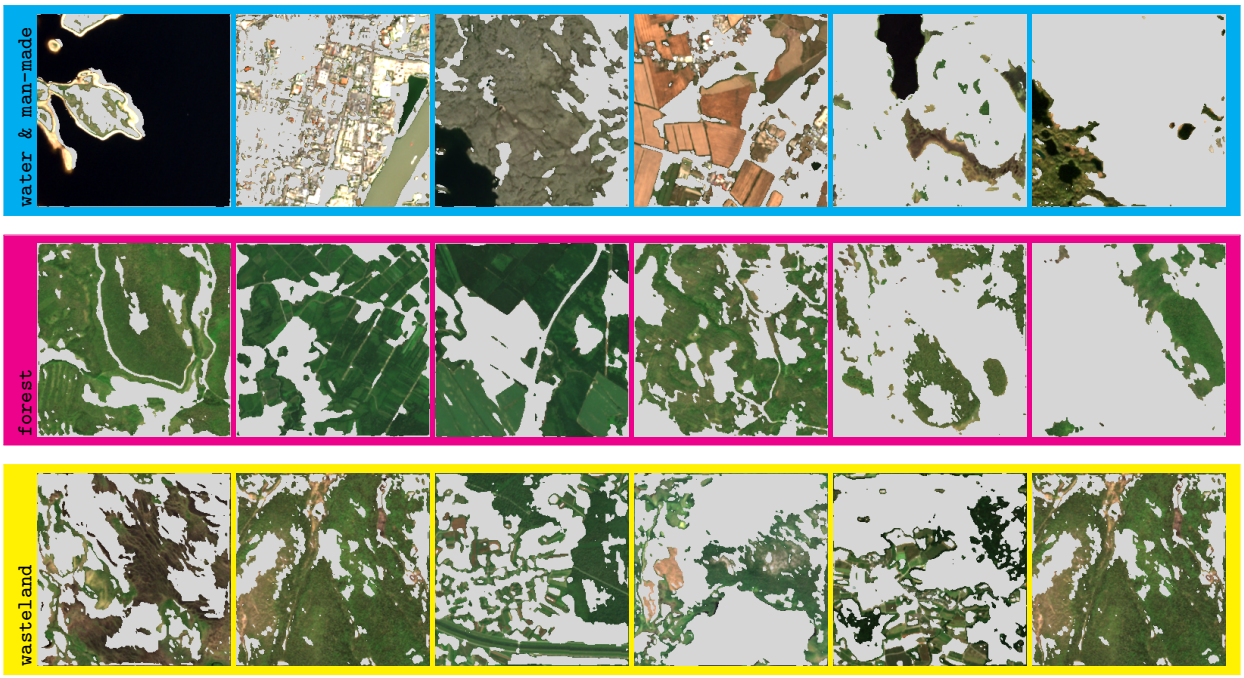


Figure 8. Correctly classified samples for each of the potential concepts. Colored areas correspond to given concepts, and gray areas are part of other concepts. `Water & man-made`: The first four samples show highly sensitive areas, which means areas that are highly connected to the `non-wilderness` class. The last two samples show non-sensitive areas of the same concept. `Forest`: Again, the first four areas are highly connected to `non-wilderness`, while the last two areas are non-sensitive. `Wasteland`: The first three samples show areas that are slightly connected to `wilderness`, whereas the last three areas are slightly connected to `non-wilderness`.

In this paper, we used protected areas as a weak label proxy for `wilderness`, but also other classes can be used as a proxy. One challenge that needs to be considered in future work is that actual non-labeled wilderness exists outside of labeled protected areas causing, for example, false positives during evaluation. Furthermore, no distilling approaches have been performed so far or tested whether light-weight models lead to similar results. Therefore, another promising research direction would be to reduce parameters and optimize convergence so that the model could be better applied on small datasets. Moreover, we consider a more detailed analysis of the likelihood values and posterior probabilities to make better statements about the assignment of a concept. Both likelihood values and posterior probabilities derived from the Gaussian mixture components could be used to describe how well a sample can be assigned to a concept. Another interesting future research question is whether this approach can be used to select and evaluate training data for non-well-defined classes.

ACKNOWLEDGEMENTS

We want to thank Max Helleis, a former student of the ESPACE master's program at the Technical University of Munich, for his support in the creation of the dataset used in this paper. This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 - 390732324 - PhenoRob. In addition, we acknowledge funding from the German Technical and Scientific Association for Gas and Water (DVGW) as part of G201819 - Antonia, from the German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety under grant no 67KI2043 - KISTE, and DFG as part of the project RO 4839/5-1 / SCHM 3322/4-1 - MapInWild.

REFERENCES

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B., 2018. Sanity checks for saliency maps. *NeurIPS*, 9505–9515.

Arendsen, P., Marcos, D., Tuia, D., 2020. Concept Discovery for The Interpretation of Landscape Scenicness. *MAKE*, 2(4), 397–413.

Bastmeijer, K., 2016. *Wilderness protection in Europe: the role of international, European and national law*. Cambridge University Press.

Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A., 2017. Network dissection: Quantifying interpretability of deep visual representations. *Proceedings of the CVPR*, 6541–6549.

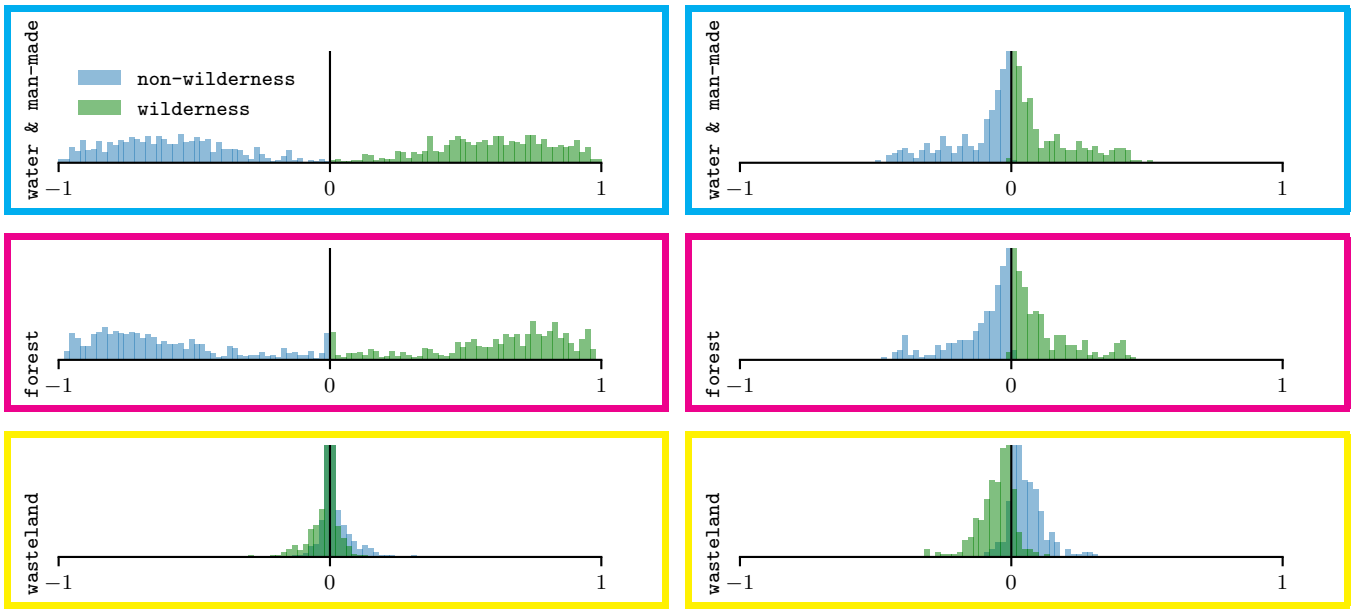
Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., Smets, B., 2020. Copernicus global land cover layers—collection 2. *Remote Sensing*, 12(6), 1044.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the RSS: Series B (Methodological)*, 39(1), 1–22.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *RSE*, 202, 18–27.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the CVPR*, 770–778.

Isikdogan, F., Bovik, A. C., Passalacqua, P., 2017. Surface water mapping by deep learning. *JSTARS*, 10(11), 4909–4918.



(a) `non-wilderness` samples

(b) `wilderness` samples

Figure 9. Confidence score deviations when covering image areas of different concepts. a) and b) show the effects for samples that are labeled `non-wilderness` and `wilderness`. Blue and green histograms show the changes to the `non-wilderness` and `wilderness` scores. The scores are more sensitive to the concepts of `water & man-made` and `forest`, and less sensitive to `wasteland`.

Kierdorf, J., Garcke, J., Behley, J., Cheeseman, T., Roscher, R., 2020. What identifies a whale by its fluke? on the benefit of interpretable machine learning for whale identification. *ISPRS Annals*, V-2-2020, 1005–1012.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. et al., 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *ICML*, PMLR, 2668–2677.

Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A., 2017. Deep learning classification of land cover and crop types using remote sensing data. *IEEE GRSL*, 14(5), 778–782.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.-R., 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), 1096.

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., Ideker, T., 2018. Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, 15(4), 290–298.

Marcos, D., Lobry, S., Tuia, D., 2019. Semantically interpretable activation maps: What-where-how explanations within CNNs. *ICCV Workshop*, 4207–4215.

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.-R., 2019. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, 193–209.

Odena, A., Dumoulin, V., Olah, C., 2016. Deconvolution and Checkerboard Artifacts. *Distill*. http://distill.pub/2016/deconv-checkerboard.

Qiu, C., Schmitt, M., Geiß, C., Chen, T.-H. K., Zhu, X. X., 2020. A framework for large-scale mapping of human settlement extent from Sentinel-2 images via fully convolutional neural networks. *ISPRS P&RS*, 163, 152–170.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, Springer, 234–241.

Roscher, R., Bohn, B., Duarte, M. F., Garcke, J., 2020a. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8, 42200–42216.

Roscher, R., Bohn, B., Duarte, M., Garcke, J., 2020b. Explain it to Me-Facing Remote Sensing Challenges in the-and Geosciences with Explainable Machine Learning. *ISPRS Annals*, 3, 817–824.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., Múller, K.-R., 2020. Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond. *arXiv:2003.07631*.

Tuia, D., Roscher, R., Wegner, J. D., Jacobs, N., Zhu, X. X., Camps-Valls, G., 2021. Towards a Collective Agenda on AI for Earth Science Data Analysis. *IEEE GRSM*.

Wigness, M., Draper, B. A., Beveridge, J. R., 2014. Selectively guiding visual concept discovery. *IEEE WACV*, IEEE, 247–254.

Zeiler, M. D., Fergus, R., 2014. Visualizing and understanding convolutional networks. *ECCV*, Springer, 818–833.

Zhang, L., Zhang, L., Du, B., 2016. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE GRSM*, 4(2), 22–40.

Zhou, B., Bau, D., Oliva, A., Torralba, A., 2018a. Interpreting deep visual representations via network dissection. *IEEE TPAMI*, 41(9), 2131–2145.

Zhou, B., Sun, Y., Bau, D., Torralba, A., 2018b. Interpretable Basis Decomposition for Visual Explanation. *Lecture Notes in Computer Science*, 11212 LNCS, 122–138.