

Evaluating Explainable Artificial Intelligence Methods for Multi-label Deep Learning Classification Tasks in Remote Sensing

Ioannis Kakogeorgiou* and Konstantinos Karantzas

Remote Sensing Laboratory, National Technical University of Athens, Zographou, 15780, Greece; karank@central.ntua.gr

* Correspondence: gkakogeorgiou@central.ntua.gr; Tel.: +302107721673

Abstract: Although deep neural networks hold the state-of-the-art in several remote sensing tasks, their black-box operation hinders the understanding of their decisions, concealing any bias and other shortcomings in datasets and model performance. To this end, we have applied explainable artificial intelligence (XAI) methods in remote sensing multi-label classification tasks towards producing human-interpretable explanations and improve transparency. In particular, we developed deep learning models with state-of-the-art performance in the benchmark BigEarthNet and SEN12MS datasets. Ten XAI methods were employed towards understanding and interpreting models' predictions, along with quantitative metrics to assess and compare their performance. Numerous experiments were performed to assess the overall performance of XAI methods for straightforward prediction cases, competing multiple labels, as well as misclassification cases. According to our findings, *Occlusion*, *Grad-CAM* and *Lime* were the most interpretable and reliable XAI methods. However, none delivers high-resolution outputs, while apart from *Grad-CAM*, both *Lime* and *Occlusion* are computationally expensive. We also highlight different aspects of XAI performance and elaborate with insights on black-box decisions in order to improve transparency, understand their behavior and reveal, as well, datasets' particularities.

Keywords: interpretability; explainability; multi-class; reliability; multilabel image classification; deep neural networks; XAI; black-box models; BigEarthNet; SEN12MS;

1. Introduction

Deep neural networks have achieved remarkable success in real-world applications in various engineering fields [1] as well as in remote sensing [2][3]. However, from a scientific standpoint, black-box artificial intelligence (AI) solutions with non or questionable transparency, interpretability, and explainability are still barriers. Contrary to more simple and self-explaining models (e.g., linear regression), deep neural networks lack interpretability due to their non-linear and complex design. Even though deeper models can identify and model complex patterns as well as enable significantly higher performance, the deployment of black-box solutions in remote sensing and other disciplines is not straightforward for critical decision making. In particular, there is a tendency for a trade-off between model performance and its interpretability [4].

To tackle this challenge, explainable AI (XAI) methods could provide human interpretable explanations to better understand machine learning black-box decisions. XAI methods could help users/ practitioners further evaluate their models beyond standard performance metrics (e.g., accuracy metric) by analyzing and inspecting individual predictions through examining their explanations [5]. Moreover, these methods could potentially reveal biases in the trained dataset, classes, multiple labels, and other spurious or artifactual correlations learned by a model [6]. Moreover, further insights could be gained, in cases, e.g., that a model surpasses human performance; it may have encompassed scientific knowledge that can be extracted via an XAI method providing insights to the domain experts and scientific community [7]. Additionally, the comprehensive study of the delivered explanations can be beneficial for further designing a deep architecture and the debugging process of a model [8].

In Remote Sensing (RS) and Earth Observation (EO), deep complex, black-box models are broadly being employed due to their successful performance [2]. Indeed, deep learning architectures are currently holding state-of-the-art performance in several RS tasks like indicatively: scene classification, multi-label classification, semantic

segmentation, weakly supervised semantic segmentation and object detection [9][10][11][12]. However, challenges such as model interpretability need to be overcome and further investigated [13].

In particular, the majority of RS studies that have considered XAI methods have been conducted for applications in the area of bio- and geosciences [14]. For instance, [15] trained a Deep Convolutional Neural Network (DCNN) to identify Plant diseases on hyperspectral images. The authors employed the *Saliency* method [16] to increase the confidence in their model's predictive capability. Moreover, in a similar manner, for the prediction of meteorological phenomena using radar data, explanation methods such as *Saliency* and *Grad-CAM* [17] were exploited for model interpretation and visualization [18]. *Class Activation Maps (CAM)* [19] of a trained ResNet-50 on aerial images were visualized to gain a better understanding of its internal representations [20]. Additionally, [21] integrated the *Grad-CAM++* [22] approach in their models to generate adversarial examples during the training process to improve performance for Land-use classification and object detection. [23] trained Convolutional Neural Networks (CNN) that incorporated multiple *CAMs* to improve aircraft type recognition in aerial images. A theoretical study conducted by [24] investigated the derivatives of model functions learned by several supervised and unsupervised kernel methods in order to analyze and understand them.

Regarding multispectral and radar satellite data, *Regression Activation Maps (RAM)* were integrated into the pipeline of [25] methodology providing insight into the relevant conditions leading to crop yield variability using MODIS imagery. Additionally, based on the benchmark multi-label BigEarthNet dataset, [26] utilized *Layer-wise Relevance Propagation (LRP)* [27] heatmaps to explain the characteristics of the different loss functions that have been examined. Recurrent Neural Networks for land use classification using Sentinel-2 data [28] and crop yield estimation along with remote-sensing variables [29] were scrutinized, looking at the hidden units distribution. Self-explaining models, such as Gaussian Processes were also utilized on data acquired by optical and microwave sensors, and further examined by sensitivity analysis [30]. [31] developed an interpretable by-design CNN model in order to understand the connections between landscape scenicness and the presence of landcover classes. So far, and to the best of our knowledge, none comparison and analysis of different XAI methods have been performed on satellite multispectral imagery.

Towards this direction, in this study, we have developed deep learning models that deliver high classification accuracy results in multi-label RS benchmark datasets. To better understand their decision-making process, we have utilized various XAI methods, gaining valuable insights regarding their overall performance. In particular, we have trained Densely Connected Convolutional Networks (DenseNet) [32] in challenging RS benchmark datasets, i.e., BigEarthNet [33] and SEN12MS [34], and achieved state-of-the-art results by integrating learning rate scheduler, early stopping, and data augmentation techniques. We further contribute with the benchmarking of ten XAI methods by evaluating their explainability and their performance towards understanding the decisions/predictions of the developed deep neural networks. Specifically, we have utilized the following XAI methods: *Saliency*, *Input \times Gradient*, *Integrated Gradients*, *Guided Backpropagation*, *Grad-CAM*, *Guided Grad-CAM*, *Lime*, *Occlusion*, *DeepLift* as well as we have integrated the *SmoothGrad* approach. *Max-Sensitivity*, *Area Under the Most Relevant First* perturbation curve, *File Size* and *Computational Time* were employed in order to quantitatively assess the performance of the studied XAI methods. Moreover, we qualitatively assessed XAI methods for cases with single or competing multiple labels and misclassification cases. Finally, a detailed analysis of XAI methods performance is presented to evaluate different aspects of their explainability and applicability. Our results stress that certain XAI methods can lead to the extraction of significant insights regarding black-box models' decisions for various prediction cases as well as datasets' composition.

2. Materials and Methods

In this section, we describe the employed XAI methods as well as the utilized evaluation metrics in order to quantitatively assess and compare their performance. Moreover, the two benchmark remote sensing datasets, i.e., BigEarthNet and SEN12MS, are presented for multi-label classification tasks. The developed deep learning models based on Densely Connected Convolutional Networks (DenseNet) are also demonstrated while their performance is evaluated and discussed. Specific implementation details are also included in the last part of this section.

2.1. Explainable AI Methods

For the rest of the paper, let a model $f \in \mathcal{F}$ be a function $f : \mathbb{R}^D \rightarrow \mathbb{R}^C$ with input $x = (x_1, \dots, x_D) \in \mathbb{R}^D$ and output $f(x) = (f_1, \dots, f_C)(x) \in \mathbb{R}^C$ which is the amount of evidence for predicting a number of C classes in a multiclass problem. An explanation method $\Phi : \mathcal{F} \times \mathbb{R}^D \rightarrow \mathbb{R}^D$ attributes relevance, contribution, or importance scores of the prediction $f_c(x)$ of a class to each feature x_d .

The **Saliency (Sal)** method [16] is estimated by the gradient of the output $f_c(x)$ with respect to (w.r.t.) the input x :

$$\Phi_{Sal}(f_c, x) = \nabla f_c(x) \quad (1)$$

Briefly, the *Saliency* approach indicates which features need to be changed the least to affect the score of a particular class the most.

The **Input \times Gradient (InputXGrad)** method [35] is an extension of the *Saliency* method, which uses the element-wise product of the input and the gradient:

$$\Phi_{InputXGrad}(f_c, x) = x \odot \nabla f_c(x) \quad (2)$$

Intuitively, the *Input \times Gradient* product is the total contribution of each feature to the approximately linearized model's output.

The **Integrated Gradients (IntGrad)** method [36] is defined as the integral of the gradients along the straight-line path from a baseline $x' = (x'_1, \dots, x'_D)$ to the input $x = (x_1, \dots, x_D)$:

$$\Phi_{IG}^d(f_c, x) = (x_d - x'_d) \times \int_0^1 \frac{\partial f_c(\tilde{x})}{\partial \tilde{x}_d} \bigg|_{\tilde{x} = x'_d + a(x_d - x'_d)} da \quad \forall d \in \{1, \dots, D\} \quad (3)$$

The baseline x' input can be a root point of the desired function f_c and can be obtained using an optimization technique [7] or more simply an input that represents the non-appearance of a feature in the original x . For instance, it can be the black/ all zero image or even random noise [37].

The **Guided Backpropagation (GuidedBackprop)** method [38] combines both *Saliency* and *Deconvnet* [39] methods. *Guided Backpropagation* varies from the other two methods in the way they handle backpropagation through ReLU functions. Specifically, it computes the gradient of the output $f_c(x)$ w.r.t. the input, except that when propagating through ReLU functions, only non-negative gradients are backpropagated.

The **Grad-CAM** method [17] computes the gradients of the output $f_c(x)$ w.r.t. feature map activations A^k of a given layer. At that point, the gradients are averaged for each channel k (along width W and height H dimensions) to obtain the importance weights:

$$a_k^c = \frac{1}{H \cdot W} \sum_i^W \sum_j^H \frac{\partial f_c}{\partial A_{ij}^k}(x) \quad (4)$$

Consequently, the average gradient for each channel a_k are multiplied by the layer activations A^k and the results are summed over all channels. Finally, a ReLU is applied to the output, returning only non-negative attributions:

$$L_{Grad-CAM}^c = ReLU(\sum_k a_k^c A^k) \quad (5)$$

The *Grad-CAM* explanation has the same spatial dimension as the feature maps A^k . Thus, bilinear interpolation can be used in order to upsample the *Grad-CAM* output to the input image resolution:

$$\Phi_{Grad-CAM}(f_c, x) = UPSAMPLE_{Bilinear}(L_{Grad-CAM}^c) \quad (6)$$

Furthermore, [17] suggested utilizing the last convolutional layer's feature maps since they encode high-level features and retain spatial information. Intuitively, *Grad-CAM* highlights coarse regions of the image that have a positive contribution to $f_c(x)$.

The **Guided Grad-CAM** method [17] is the element-wise product of *Guided Backpropagation* with the upsampled *Grad-CAM* attributions:

$$\Phi_{Guided\ Grad-CAM}(f_c, x) = \Phi_{Grad-CAM}(f_c, x) \odot \Phi_{GuidedBackprop}(f_c, x) \quad (7)$$

This method combines the fine-grained details of *Guided Backpropagation* with the course localization advantages of *Grad-CAM*.

The **Occlusion** [39] systematically replaces different contiguous rectangular patches $p_i \in P$ of the input image with a given baseline and monitors the decrease of the prediction function $f_c(x)$:

$$\Phi_{Occlusion}^{p_i}(f_c, x) = f_c(x) - f_c(x'_{p_i}) \quad (8)$$

For overlapping patches, the corresponding output differences are averaged to compute the final contribution of that region. In the case that the baseline is the all-zero patch, then the reference point is $x'_{p_i} = x \odot (1 - \mathbb{1}_{p_i})$.

The **DeepLift** method [35] is a recursive backpropagation-based method that attributes the difference in the outputs $f_c(x) - f_c(x')$ on the differences between the input $x = (x_1, \dots, x_D)$ and a baseline $x' = (x'_1, \dots, x'_D)$. We present *DeepLift* similar to the formulation by [40]. Specifically, *DeepLift* runs forward propagation for both original x and baseline x' inputs. Then, it stores the difference of the weighted activation $\Delta z_{ij} = w_{ij}^{(l, l+1)}(x_i^l - x'^l_i)$, where x_i^l and x'^l_i are the activations of the neuron i in the l layer and $w_{ij}^{(l, l+1)}$ is the weight between neuron i and neuron j in the next $l + 1$ layer. Finally, the backpropagation of the difference in the outputs to the final input contributions $\Phi_{DeepLift}^d(f_c, x)$ is achieved with the following rules:

$$r_i^{(l)} = \begin{cases} f_c(x) - f_c(x'), & i = c \\ 0, & i \neq c \end{cases} \quad (9)$$

$$r_i^{(l)} = \sum_j \frac{\Delta z_{ij}}{\sum_{i'} \Delta z_{i'j}} r_j^{(l+1)} \quad (10)$$

$$\Phi_{DeepLift}^d(f_c, x) = r_d^{(1)}, \quad (11)$$

where L is the output layer and $r_d^{(1)}$ is the contribution of each feature on the input layer. The equation (10) is called the *Rescale* rule in [35]. Also, [35] proposed the *RevealCancel* rule, which is not examined in this study (for further details see [35]).

The **Lime** [5] explanations can be achieved by approximating locally the predictions $f_c(x)$ around a specified input x using a simpler self-explanatory surrogate model (e.g., linear regression). *Lime*'s core idea is that the surrogate model is trained on different *interpretable* features from the original model inputs. For instance, in our case, an *interpretable* representation \tilde{z} is a binary vector indicating the presence or absence of a contiguous patch of the image, such that $h(\tilde{z}) = \tilde{x}$ where h is a mapping between the binary vectors and the corresponding perturbed image. The patches can be super-pixels that might have been derived from an image segmentation algorithm. Consequently, in order to train the surrogate model, randomly binary vectors \tilde{z} samples are drawn and then evaluated by the original model. Another interesting aspect of the *Lime* method is that each generated sample \tilde{z} is weighted by the corresponding similarity of \tilde{x} to the instance of interest x . In our case, we used linear regression for the surrogate model $g(\tilde{z}) = w \cdot \tilde{z}$ and an exponential kernel defined on the Frobenius norm between the original and the perturbed input $\pi_x(\tilde{z}) = \exp(-\|x - h(\tilde{z})\|_F^2)$ as a similarity function. The surrogate model is obtained by the minimization of:

$$g = \underset{g'}{\operatorname{argmin}} \sum_{\tilde{z}} \pi_x(\tilde{z}) \cdot ((f_c \circ h)(\tilde{z}) - g'(\tilde{z})) \quad (12)$$

The positive $\Phi_{LIME}^i(f_c, x) = w_i$ explanations (i.e., weights of the linear model g) highlight the regions that contribute to the prediction $f_c(x)$. Negative and close to zero $\Phi_{LIME}^i(f_c, x)$ explanations indicate areas with a negative and neutral contribution, respectively.

The **SmoothGrad** (SG) [41] is a method that can be used on top of other attribution methods. *SmoothGrad* generates multiple samples by adding Gaussian noise to the original input x and averages the calculated attributions:

$$\Phi_{SG}(f_c, x) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} [\Phi(f_c, x + \varepsilon)] \quad (13)$$

This approach can be understood as an averaging process that makes the initial explanation $\Phi(f_c, x)$ smoother [7]. While *SG* was mainly introduced as an averaging process of the vanilla *Saliency* method, the same paper [41] highlights that this smoothing procedure can be used to augment any gradient-based method. Thus, we applied *SG* combined with the *Saliency*, *Input \times Gradient* and *Integrated Gradients* methods and evaluated the results.

2.2. Evaluating Metrics

In order to evaluate quantitatively the performance of the aforementioned XAI methods, we employed the *Max-Sensitivity* metric, *Area Under the Most Relevant First* perturbation curve, the produced *File Size* and the *Computation Time*.

The **Max-Sensitivity** (MS) [42] metric measures the reliability in terms of the maximum change in an explanation $\Phi(f_c, x)$ with small input perturbations x' and it is estimated using Monte Carlo sampling:

$$SENS_{MAX}(\Phi, f_c, x, r) = \max_{\|x' - x\|_\infty \leq r} \|\Phi(f_c, x') - \Phi(f_c, x)\|_F, \quad (14)$$

where $\|\cdot\|_\infty$ is the maximum norm, $\|\cdot\|_F$ is the Frobenius norm and r is the input neighborhood radius. Naturally, we would not prefer an explanation to have high *Max-Sensitivity*, since that would entail differing explanations with minor variations in the input. This fact might lead us to distrust the explanations.

Another way to assess the performance of XAI methods quantitatively is the **Most Relevant First (MoRF)** perturbation curve [43]. This procedure measures the reliability of an explanation by testing how fast the $f_c(x)$

decreases, while we progressively remove information (e.g., perturb pixels) from the input x (e.g., image), that appears as the most relevant by the explanation $\Phi(f_c, x)$.

Specifically, let $x = (x_1, \dots, x_D) \in \mathbb{R}^D$ be the input. We denote by $\Phi(f_c, x)^\downarrow \in \mathbb{R}^D$ a vector with the same components as $\Phi(f_c, x) = (\varphi_1, \dots, \varphi_D) \in \mathbb{R}^D$, but sorted in non-increasing order. There is a permutation $\sigma: \{1, 2, \dots, D\} \rightarrow \{1, 2, \dots, D\}$, such that:

$$\Phi(f_c, x)^\downarrow = (\varphi_{\sigma(1)}, \dots, \varphi_{\sigma(D)}) \in \mathbb{R}^D \quad (15)$$

We call the permutation σ as *argsort* and for each $j < k \Rightarrow \varphi_{\sigma(j)} \geq \varphi_{\sigma(k)}$. Also, we define the sets $S_j := \{\sigma(i) : i \leq j\}$ for $j = 1, 2, \dots, D$. Thus, the k *Most Relevant* features of x are the components positioned at $S_k = \{\sigma(1), \sigma(2), \dots, \sigma(k)\}$. Finally, if we reformulate the input as a finite sequence $x = (x_d)_{d=0}^D$, the perturbed input based on the k *Most Relevant* features is:

$$x_{MoRF}^{(k)} = \left(\begin{cases} x'_d, & \text{if } d \in S_k \\ x_d, & \text{if } d \notin S_k \end{cases} \right)_{d=0}^D, \quad (16)$$

where x'_d is a perturbed feature. For instance, in the context of an image, this perturbation might be a zero, a random or an interpolated pixel value. The faster the curve $f_c(x_{MoRF}^{(k)})$ for $k = 1, 2, \dots, D$ decreases, the more reliable the explanation is.

In order to quantify and compare the degree of this decrease, we use the **Area Under the Most Relevant First (AUC - MoRF)** perturbation curve. Based on the *trapezoidal rule*, this area is:

$$AUC_{MoRF}(\Phi, f_c, x, r) = \sum_{k=2}^D \frac{f_c(x_{MoRF}^{(k-1)}) + f_c(x_{MoRF}^{(k)})}{2} \quad (17)$$

Thus, we would like to minimize the *AUC - MoRF* score for an explanation.

Moreover, the **File Size** of an XAI visualization was proposed by [7] as a quantification metric of the included amount of information. The smaller the file size, the less complex information in the results, thus potentially more concise and interpretable for a human.

The **Computation Time** is an essential metric to quantify the performance of any given algorithm. High computation time is a barrier to XAI methods' applicability as well as to their integration in complex pipelines with potential requirements for real-time performance.

2.3. Multi-label Remote Sensing Datasets

In our study, we employed two remote sensing benchmark datasets, i.e., BigEarthNet [33] and SEN12MS [34], which have recently gained significant attention from the research community towards developing and assessing deep learning architectures for remote sensing image understanding tasks.

In particular, BigEarthNet consists of 590,326 non-overlapping Sentinel-2 image patches, which were constructed by 125 different tiles with less than 1% cloud cover. The tiles were acquired between June 2017 and May 2018 and are distributed over ten different European countries. Each image patch is a section of i) 120×120 pixels for 10m bands (B2, B3, B4 and B8); ii) 60×60 pixels for 20m bands (B5, B6, B7, B8A, B11 and B12); and iii) 20×20 pixels for 60m bands (B1 and B9). The B10 Cirrus band of Sentinel-2 has been excluded as it does not provide information about the land surface. BigEarthNet is annotated with labels provided by the CORINE Land Cover (CLC) map of 2018 [44]. However, in this paper, the alternative nomenclature with 19 classes proposed by [45] for BigEarthNet images was used, as it better expresses single-date Sentinel-2 images. For the train, validation and test sets, we used the suggested split of 269,695 images in training, 123,723 images in the validation and 125,866 images

in the testing set. Additionally, as [33] recommended, ~71,000 images that are fully covered by seasonal snow, cloud and cloud shadows are already eliminated in these sets.

SEN12MS consists of 180,662 triples of Sentinel-1 synthetic aperture radar (SAR) data, Sentinel-2 full multispectral imagery and MODIS-derived land cover image patches. The dataset was constructed by 252 different scenes, which were acquired between December 2016 and November 2017 and are globally distributed. SEN12MS was developed based on a sophisticated workflow to avoid cloud-affected images. Each image patch consists of 256×256 pixels in size, and there is a 50% overlap between adjacent patches. This overlapping should be taken into account during the training and testing split.

In this paper, as input into the model, we used only Sentinel-2 image patches, excluding the B10 Cirrus band. For the labeling, we used the 1st of the four provided bands of the MODIS-derived land cover patches, which contains land cover following the International Geosphere-Biosphere Programme (IGBP) classification scheme [46]. Moreover, SEN12MS was modified in order to become more appropriate for multi-label classification tasks. Thus, we used all classes that appeared from each patch as multi-labels. Also, we used the simplified International Geosphere-Biosphere Programme (IGBP) classification scheme as presented by [47] following the IEEE GRSS Data Fusion Contest 2020 (DFC2020). This simplified scheme consists of 10 aggregated classes compared to the original IGBP with 17 classes. For the train, validation and test sets, we propose a new balanced split* appropriate for the multi-label classification task. Table 1 and Table 2 show that this proposed split preserves the same meteorological seasonal and class distribution on the different sets. Additional details for the proposed split are presented in Figures S1, S2 and S3 in the supplementary material.

Table 1. The proposed balanced seasonal distribution (%) for each split in SEN12MS.

Split	Fall	Spring	Summer	Winter
Train	34.2	23.2	24.1	18.5
Val	35.6	22.4	24.3	17.7
Test	33.0	22.2	25.6	19.2

Table 2. The proposed balanced class distribution (%) for each split in SEN12MS.

Split	Forest	Shrubland	Savanna	Grassland	Wetlands	Croplands	Urban/ Built-up	Snow/ Ice (*100)	Barren	Water
Train	27.0	12.5	53.6	42.6	6.8	32.1	23.9	0.4	7.8	11.5
Val	29.1	13.1	51.8	45.1	7.9	34.3	24.3	3.3	7.9	12.7
Test	26.4	12.1	51.1	43.6	8.9	33.5	26.3	2.7	7.8	11.7

2.4 Developed Deep Learning Models and Performance Results

In this section, the developed machine learning models are described in detail. Specifically, we adopted DenseNet-121 architecture, in which each convolutional layer receives feature maps from all previous layers and transmits its feature maps to all subsequent layers. We modified the first layer of the DenseNet-121 to adapt to the 12 input bands (B10 Cirrus band is excluded) for both datasets, as well as the final classification layer was changed to output 19 and 10 classes for BigEarthNet and SEN12MS, respectively. Contrary to SEN12MS, in which the Sentinel-2 bands are already stacked, for BigEarthNet, we applied cubic interpolation to upsample the 20m and 60m bands to match the 10m resolution bands. During training, we employed the Adam algorithm [48] to minimize the Binary Cross Entropy loss with an initial learning rate of 10^{-3} . We also employed a scheduler to reduce the learning rate when the validation set's loss has stopped decreasing. Moreover, we utilized early stopping based on the loss of the validation set. The batch size for the SEN12MS was 32 samples and for the BigEarthNet was 64. Additionally, we employed random rotations of the input images by -90° , 0° , 90° , or 180° and horizontal flips in order to augment the datasets. It is worth mentioning that for both datasets the developed models were trained from scratch. We also selected 0.4 as a threshold for the multi-label classification for BigEarthNet and 0.3 for SEN12MS. These thresholds maximized the F_1 scores on the validation sets.

* New SEN12MS balanced split available here:
<https://pithos.oceanos.grnet.gr/public/23vXFjRgMquw6dZcwF6D71>

In order to evaluate the developed models, we assessed different overall and per class metrics, including F₁-score, Recall, Precision, Hamming Loss, Rank Loss, Coverage Error and compared our results with the literature (Tables 3,4,S1,S2,S3). For BigEarthNet, the developed model resulted in state-of-the-art performance across all overall metrics compared with recently reported efforts [45] (Table 3). Also, improved F₁ scores were derived for each class as well (Table S1). Although our average Precision for each class is higher than the corresponding ones from [45], our Precision results per class were also comparable (Table S2). Our developed model led to a higher than 80% F₁ score for the classes: *Arable Land*, *Coniferous Forest*, *Mixed Forest*, *Inland Waters*, and *Marine Waters*. The lowest F₁ scores (i.e., < 60%) were reported for: *Industrial or Commercial Units* and *Natural Grassland and Sparsely Vegetated Areas* (Table S1).

For the SEN12MS dataset, which is a more recent one and was employed in the recent 2020 IEEE GRSS Data Fusion contest [49], there was not, as for BigEarthNet, an available multi-label classification model in the literature to compare our results. Our model achieved (Table 4) an overall F₁ score of 74.4%. The highest F₁ scores (i.e., > 75%) were achieved for the classes: *Forest*, *Savanna*, *Urban/ Built-up*, and *Water* (Table S3). Conversely, the model resulted in the lowest F₁ scores (i.e., < 60%) for the classes: *Shrubland* and *Barren*. More quantitative results, reporting on the per-class accuracy and performance of the developed models are presented in the supplementary material (Tables S1,S2,S3).

Table 3. Comparing the performance of the developed model against the state-of-the-art in BigEarthNet [45]. With bold the higher score per evaluation metric.

Metric	K-Branch CNN	VGG16	VGG19	ResNet50	ResNet101	ResNet152	DenseNet121 (Ours)
F₁-score (%)	72.73	76.01	75.96	77.11	76.49	76.53	82.22
Recall (%)	78.96	75.85	76.71	77.44	77.45	76.24	84.70
Precision (%)	71.61	81.05	79.87	81.39	80.18	81.72	83.60
Hamming Loss	0.093	0.077	0.079	0.075	0.077	0.075	0.061
One Error	0.103	0.073	0.071	0.072	0.082	0.072	0.029
Rank Loss	0.056	0.048	0.048	0.047	0.049	0.046	0.026
Coverage Error	4.730	4.603	4.606	4.613	4.628	4.552	3.893

Table 4. The overall performance of the developed model in the SEN12MS Dataset.

	F₁-score (%)	Recall (%)	Precision (%)	Hamming Loss	One Error	Rank Loss	Coverage Error
DenseNet121 (Ours)	74.35	82.46	73.80	0.122	0.138	0.057	2.928

2.5. Implementation Details

For each XAI method, various parameters were tested in order to assess their overall sensitivity and performance. Based on a trial-and-error investigation, the configuration that provided the most stable and robust results was selected. In particular, for the visualization of *Saliency*, *Input \times Gradient* and *Integrated Gradients*, the absolute values of attribution were utilized. In the case of *Integrated Gradients*, we employed the black/ all zero image as a baseline reference input, following the proposed initial approach. To approximate the integral, we used 50 samples along the path. Regarding *Grad-CAM*, we used the feature maps of the last convolutional layer, as proposed by [17]. Furthermore, for the baseline x' requirement of *DeepLift*, we used a blurred version of the original image as [35] proposed for CIFAR10 data [50]. Considering *Occlusion*, we applied a 15×15 sliding window with 5×5 strides on BigEarthNet and a 32×32 window with 6×6 strides on SEN12MS. Also, we used the black/ all zero patches for the replacements of the baseline x'_{p_i} inputs. For *Lime*, we used super-pixels that have been derived from the SLIC image

segmentation algorithm [51]. Additionally, 64 segments were exploited for BigEarthNet, as well as 100 segments for SEN12MS. For each case, 6000 samples were generated in order to train the linear surrogate models. We selected to present results for the super-pixels with the highest positive weights of the linear model. When *SmoothGrad* was integrated on top of other methods, 30 samples were generated to obtain the results. Regarding the presented visualization of XAI methods, we employ colored gray-scale images that indicate the sum of attributions along the channels/ bands. Last but not least, we ran our experiments on a single RTX 3070 (8GB memory) GPU, Ryzen 7 3700X CPU, 32GB RAM and employed implementations from the Captum library [52]. All presented results regarding the performance evaluation are based on the aforementioned hardware configuration.

3. Experimental Results and Evaluation

In this section, we present the experimental results and evaluation outcomes from the application of XAI for remote sensing multi-label classification tasks. Firstly, subsection 3.1 presents a quantitative evaluation and comparison of the considered XAI methods (Table 5, Figure 1). Followingly, subsection 3.2 describes the qualitative evaluation of XAI explanations for: (i) single class correct predictions (Figures 2 and 3) (ii) multiple competing classes predictions (Figures 4,5,6 and 7) (iii) cases that the developed deep learning model failed to predict the underlying classes (Figures 8 and 9) and (iv) failures due to dataset/ inexact labeling (Figures 10 and 11).

3.1. Quantitative Evaluation

In order to quantify the reliability of the studied XAI methods we utilized *Max-Sensitivity* and *AUC-MoRF* metrics. In particular, the assessment of XAI performance in terms of methods' sensitivity against slightly different noisy input was achieved using *Max-Sensitivity* metric. *AUC-MoRF* was also performed to examine how the progressive removal of those pixels that contributed the most to model's decision according to XAI, affected model prediction.

Additionally, the produced *File Size* in Kilobytes (KB) after JPEG compression for every XAI output was assessed in order to quantify the interpretability regarding the amount of information included in the extracted explanations. The *Computation Time*, as an aspect of methods' applicability, is also evaluated and recorded in seconds (hardware, software configuration details at Sect. 2.5). Table 5 demonstrates all aforementioned quantitative metrics for both datasets. We mention that the lower the scores the better the performance for all metrics.

Regarding *Max-Sensitivity* metric, *Occlusion*, *Lime* and *Grad-CAM* achieved the lowest scores, i.e., less than 0.20 for BigEarthNet and less than 0.11 for SEN12MS. On the other hand, $Input \times Gradient$ presented the highest scores for both datasets (i.e., 0.59 for BigEarthNet and 0.30 for SEN12MS). For all cases, the integration of *SmoothGrad* approach led to lower *Max-Sensitivity* scores than the uncustomized XAI methods (i.e., *Saliency*, $Input \times Gradient$ and *Integrated Gradients*).

Moreover, Figure 1 demonstrates *MoRF* perturbation curve for all XAI methods for both datasets. At each iteration, 20% of remained pixels are removed based on XAI explanations until the whole information/ pixels of the image are removed. The removed pixels are imputed based on nearest interpolation. It is expected that XAI methods that indicate more precisely those image pixels/ regions that bind model's decision will be the ones that will rapidly lose their prediction performance (mean output score) as those pixels get more and more perturbed (iterations). Random explanation (pink dashed curve) is also presented as a baseline since it is expected that by perturbing randomly image pixels, the performance will be decreased at a slower pace.

According to Figure 1a,b *Occlusion* (purple dashed curve), *Lime* (yellow dashed curve) and *Grad-CAM* (black dashed curve) achieved the fastest decrease in their curves. This observation is also quantitatively demonstrated in Table 5. In particular, *Occlusion*, *Lime* and *Grad-CAM* resulted in less than 16.38 for BigEarthNet and less than 28.87 for SEN12MS regarding the *AUC-MoRF* metric. On the contrary, $Input \times Gradient$ presented the nearest curve to the random baseline curve (Figure 1a,b), as well as achieved the highest *AUC-MoRF* scores (i.e., 26.15 for BigEarthNet and 38.8 for SEN12MS) (Table 1). Based on *AUC-MoRF*, the integration of *SmoothGrad*, for the most cases, led to slightly more reliable explanations.

Table 5. Quantitative Metrics (lower scores indicate higher performance for all metrics)

Method	Max-Sensitivity		AUC-MoRF		File Size (KB)		Computation Time (Sec)	
	BigEarthNet (120 × 120)	SEN12MS (256 × 256)	BigEarthNet (120 × 120)	SEN12MS (256 × 256)	BigEarthNet (120 × 120)	SEN12MS (256 × 256)	BigEarthNet (120 × 120)	SEN12MS (256 × 256)
Sal	0.55	0.27	21.47	36.92	6.22	23.35	0.07	0.08
Sal w. SG	0.24	0.14	19.57	36.32	5.71	22.54	1.75	1.94
InputXGrad	0.59	0.30	26.15	38.80	5.70	18.07	0.05	0.06
InputXGrad w. SG	0.29	0.16	25.90	38.67	5.43	17.68	0.14	0.24
IntGrad	0.38	0.26	25.64	37.40	5.07	17.91	0.23	0.36
IntGrad w. SG	0.19	0.13	25.50	37.64	4.86	16.72	4.47	7.96
Guided Backprop	0.36	0.23	23.14	34.67	5.50	22.68	0.06	0.07
Grad-CAM	0.14	0.03	16.38	23.62	1.89	6.31	0.03	0.03
Guided Grad-CAM	0.50	0.23	23.66	35.71	4.57	16.55	0.08	0.1
DeepLift	0.42	0.25	23.41	34.65	5.65	19.33	0.11	0.13
Occlusion	0.14	0.03	16.32	28.87	3.02	3.84	8.73	29.71
Lime	0.20	0.11	14.39	27.65	3.24	7.91	5.80	18.67
Random								

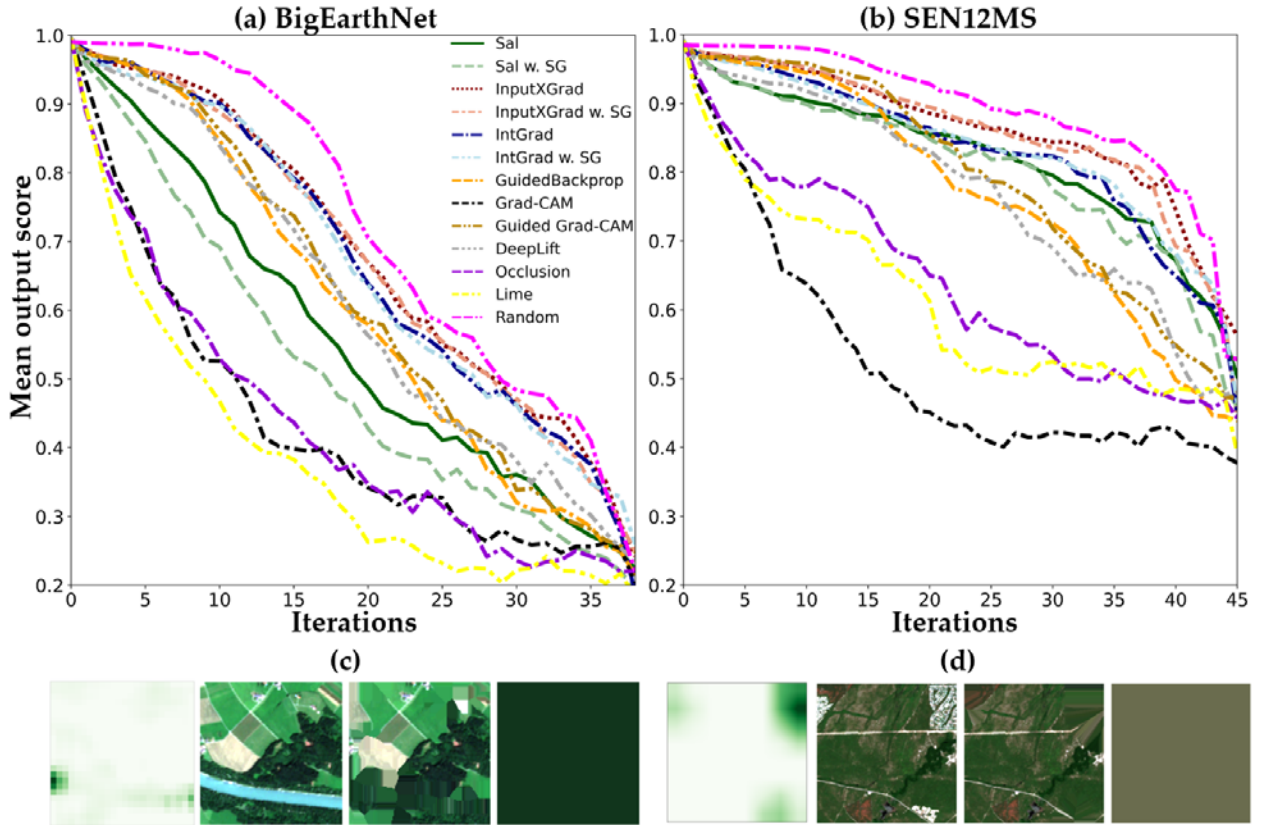


Figure 1. *MoRF* perturbation curves for all XAI methods for (a) BigEarthNet and (b) SEN12MS datasets. Curves indicate the drop in model’s performance (mean output sigmoid probability scores) for each iteration that an additional 20% of image pixels are perturbed based on XAI explanations. (c) Representative example for *Inland Waters* class prediction (BigEarthNet); from left to right: *Occlusion* explanations, S2 RGB image, intermediate iteration, final iteration. (d) Representative example for *Urban/ Built-up* class prediction (SEN12MS); from left to right: *Grad-CAM* explanations, S2 RGB image, intermediate iteration, final iteration.

Concerning *File Size*, the results were positively correlated with the *Max-Sensitivity* results ($r=0.69$, $p\text{-value} < 0.05$ for BigEarthNet and $r=0.82$, $p\text{-value} < 0.005$ for SEN12MS). Specifically, *Occlusion*, *Lime* and *Grad-CAM* presented the smallest *File Size* (i.e., < 3.24 KB for BigEarthNet and < 7.91 KB for SEN12MS). Therefore, these methods can provide rough localization information to the user. On the other hand, all other methods provide higher resolution and more detailed explanations than *Occlusion*, *Lime* and *Grad-CAM*, leading to larger file sizes.

According to Table 5, except for *Saliency w. SG*, *Integrated Gradients w. SG*, *Occlusion* and *Lime*, all the other methods required less than 0.36 sec for a single execution. The selection of parameters in the aforementioned methods affects their efficiency, and consequently the *Computation Time*. More specifically, in the case of *Integrated Gradients* with *SmoothGrad*, the bottleneck was the large number of samples that proceeded in multiple GPU batches (i.e., 4.47 sec for BigEarthNet and 7.96 sec for SEN12MS). Moreover, the *Computation Time* of *Lime* and *Occlusion* was affected by the number of the generated samples and occluded images, as well as by the required time to compute $f_c(x')$ for each sample. By using a small stride length towards high-resolution explanations in the case of *Occlusion* method, *Computation Time* further increased (i.e., 8.73 sec for BigEarthNet and 29.71 sec for SEN12MS).

Additionally, only *Occlusion* and *Lime* presented a significant increase in the required computation time from the 120×120 pixels (BigEarthNet) to the 256×256 pixels (SEN12MS). This fact indicates a lower capacity against scalability. Regarding *Occlusion*, this increase is mainly due to the additional generated occluded samples based on different stride size selection. On the other hand, for *Lime*, we generated 6000 samples for both datasets; hence the only overhead was the additional required time for the evaluation of $f_c(x')$ for each sample x' .

It is worth mentioning that although the lowest *Max-Sensitivity* scores as well as the smallest *File Sizes* were derived from the *Occlusion* and *Grad-CAM* methods, only *Grad-CAM* required low *Computation Time* regardless of the image size.

3.2 Qualitative Evaluation

For qualitative assessment of XAI methods, visual examination was performed as well as evaluation of sensitivity w.r.t. the input. Also, sensitivity regarding different labels for the same inputs was examined in order to assess the degree of explanations' dependence on the label.

3.2.1. Explaining Single Class Correct Predictions

Initially, we studied numerous cases in both datasets that the developed models managed to predict correctly the underlying classes based on the ground truth. It should be noted that we, also, performed our own verification process, which was based on an intensive photo-interpretation supported by high-resolution satellite images derived from Microsoft Bing.

Two indicative cases that the models managed to accurately predict the class are presented in Figures 2, 3 and corresponding Figures S4 and S5 in the supplementary material. In Figure 2 (and S4), the resulted explanations for predicting the class *Urban Fabric* are presented for the BigEarthNet dataset. The model managed to accurately predict *Urban Fabric* with a 0.99 sigmoid probability score.

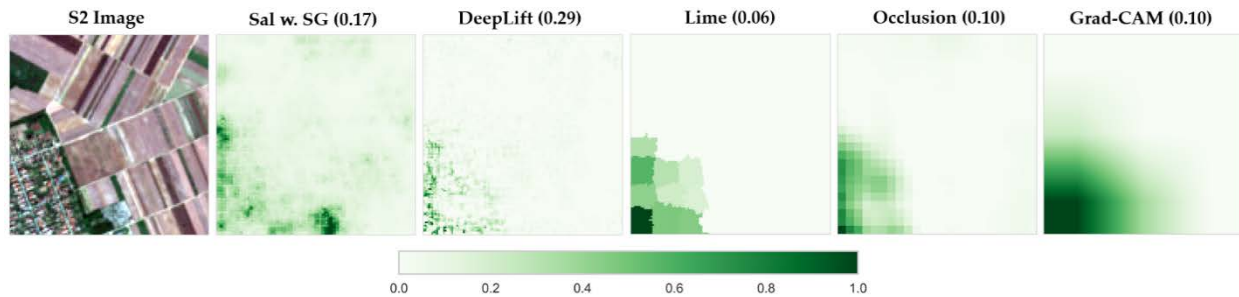


Figure 2. Explaining the Predictions of DenseNet for the class of *Urban Fabric* in the BigEarthNet dataset (Image ID: S2A_MSIL2A_20171002T094031_53_67). *Max-Sensitivity* score in parenthesis after method's name.

In particular, after a close look, one can observe that all XAI methods managed to explain and locate the image-regions (i.e., South-West) that correspond to the *Urban Fabric* class. Overall, the studied methods *Lime*, *Occlusion*, and *Grad-CAM* were the most insensitive w.r.t. the input with *Max-Sensitivity* score less than 0.10. This fact indicates that these methods are reliable for the specific case. On the opposite side, *Saliency* and *Input \times Gradient* were the most sensitive ones (with a *Max-Sensitivity* score of more than 0.65). Moreover, when the *SmoothGrad* approach was integrated, all methods resulted in lower *Max-Sensitivity* scores indicating that *SmoothGrad* contributes to a more robust and insensitive outcome. After a close look and visual examination, no significant differences were observed after applying *SG*, apart from the fact that a certain amount of noise in regions irrelevant to the label was eliminated. However, for the *Integrated Gradients* method, adding *SG* did not result in any difference since the method already delivers smooth results due to its inherent computations (Figure 2, Figure S4).

In a similar manner, in Figure 3 (and corresponding Figure S5), the extracted explanations for the class *Water* are presented for the SEN12MS dataset. The model accurately predicted *Water* with a 0.99 sigmoid probability score. In particular, experimental results indicate that *Occlusion*, *Lime*, *Grad-CAM*, *Saliency*, *Input \times Gradient*, *Integrated Gradients* and the corresponding methods with *SmoothGrad* presented interpretable explanations, as they successfully identified the water region in the image (i.e., the West and North-West region). Instead, *DeepLift*, *Guided Backpropagation*, *Guided Grad-CAM* were not able to interpret DenseNet's decision adequately. More specifically, *Guided* methods focused on irrelevant regions of the image that contained striking image features like edges (i.e., urban area). Regarding *Max-Sensitivity*, the lowest scores were achieved by *Occlusion* (0.01), *Grad-CAM* (0.01), and *Lime* (0.08) methods. Additionally, Figure S5 shows that all methods resulted in lower *Max-Sensitivity* scores and visually smoother explanations when the *SmoothGrad* approach was integrated.

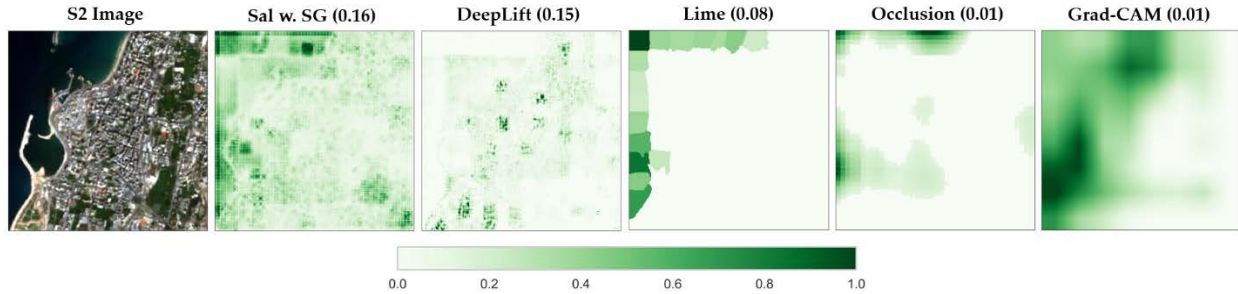


Figure 3. Explaining the Predictions of DenseNet for the class of *Water* in the SEN12MS dataset (Image ID: ROIs1970_fall_s2_112_p727). *Max-Sensitivity* score in parenthesis after method's name.

3.2.2. Explaining Correct Predictions for Multiple Competing Classes

Furthermore, we studied several cases that the developed models successfully predicted the underlying classes for cases that multiple competing classes were labeled in the same image. Two indicative cases for BigEarthNet are presented in Figures 4, 5 and corresponding Figures S6a, S6b, S7a, S7b in the supplementary material.

More specifically, in Figure 4 (and corresponding S6a and S6b), the derived visual explanations after the classification of two semantically-diverse classes, i.e., *Urban Fabric* and *Broad-leaved Forest* are presented. The model managed to accurately predict both *Urban Fabric* and *Broad-leaved Forest* with a higher than 0.99 sigmoid probability score. Overall, we observed that *Occlusion*, *Lime* and *Grad-CAM* methods were the most interpretable and sensitive w.r.t. different classes, as they presented accurate localization information to the user. *Guided Backpropagation* failed to explain DenseNet's decision against the two labels since it focused on the same image regions for both classes, indicating that it is less reliable. *Guided Grad-CAM* was slightly more sensitive (as it utilizes *Grad-CAM*) but still underperformed. *DeepLift*, *Saliency*, *Input \times Gradient*, and *Integrated Gradients* were sensitive to each different label and provided different explanations. Nevertheless, in the *Urban Fabric* class, they provided more informative results than *Broad-leaved Forest*.

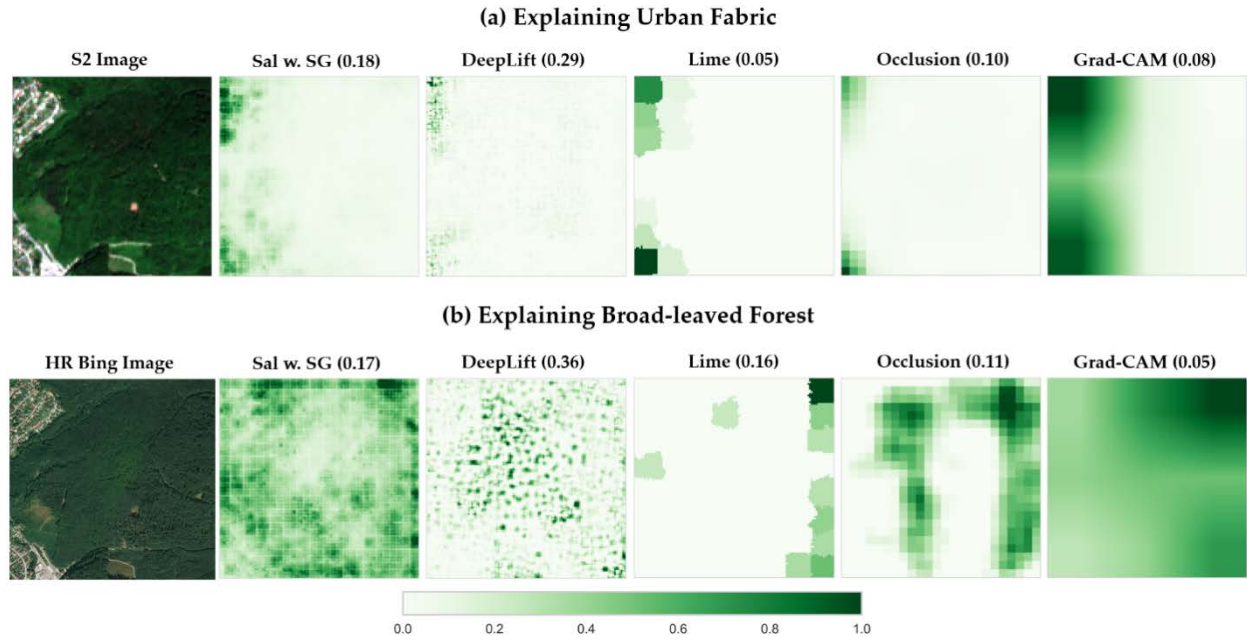


Figure 4. Explaining the Predictions of DenseNet for the class of (a) *Urban Fabric* and (b) *Broad-leaved Forest* in BigEarthNet dataset (Image ID: S2A_MSIL2A_20180506T100031_72_48). *Max-Sensitivity* score in parenthesis after method's name.

Moreover, in order to evaluate the performance of the studied XAI methods, we examined cases with competing classes that belong semantically to the same family and exist concurrently in the same image. Therefore, in Figures 5 (and corresponding Figures S7a, S7b), an indicative case from BigEarthNet is presented with the classes *Urban Fabric* and *Industrial Units* that belong to the same Artificial Surfaces super-class. The model accurately predicted both *Urban Fabric* and *Industrial Units* with 0.90 and 0.64 sigmoid probability scores, respectively. Overall, derived explanations revealed that the model correctly focused on the North image region for the *Urban Fabric* and on the North-West area for the *Industrial Units*. In particular, all methods managed to focus on the actual image regions that correspond to considered labels. However, *Grad-CAM* and *Integrated Gradients*, when explaining decisions related to the *Urban Fabric* class, focused additionally on pixels/sub-regions of the *Industrial Units* class. Compared to previous examples, *Guided Backpropagation* was slightly more sensitive in this case (Figure S7a, S7b).

Additionally, another indicative case from the SEN12MS dataset with two classes (i.e., *Urban/ Built-up* and *Croplands*) is demonstrated in Figure 6 (and corresponding S8a and S8b). The model managed to accurately predict both *Urban/ Built-up* and *Croplands* with 0.94 and 1.00 sigmoid probability scores, respectively. Overall, all methods managed to grab the regions that correspond to *Urban* class (i.e., North-West and East/ South-East regions). However, for *Croplands* class prediction, only *Occlusion* and *Grad-CAM* methods were interpretable. The rest of the methods failed to deliver valuable information regarding the model decision for *Croplands* class to the user. With respect to different classes, almost all methods were sensitive. *Occlusion* and *Grad-CAM* were the most class-discriminative methods though, and thus, the most reliable. On the opposite side, *Guided Backpropagation* focused on points where image intensity changes sharply and its result was almost the same when explained the two different class predictions. Similar to *Broad-leaved* class from BigEarthNet (Figures 4, S6b), *Saliency*, *Input \times Gradient*, and *Integrated Gradients* methods did not manage to localize *Croplands* class, which is not spatially centralized in a specific image region, as well.

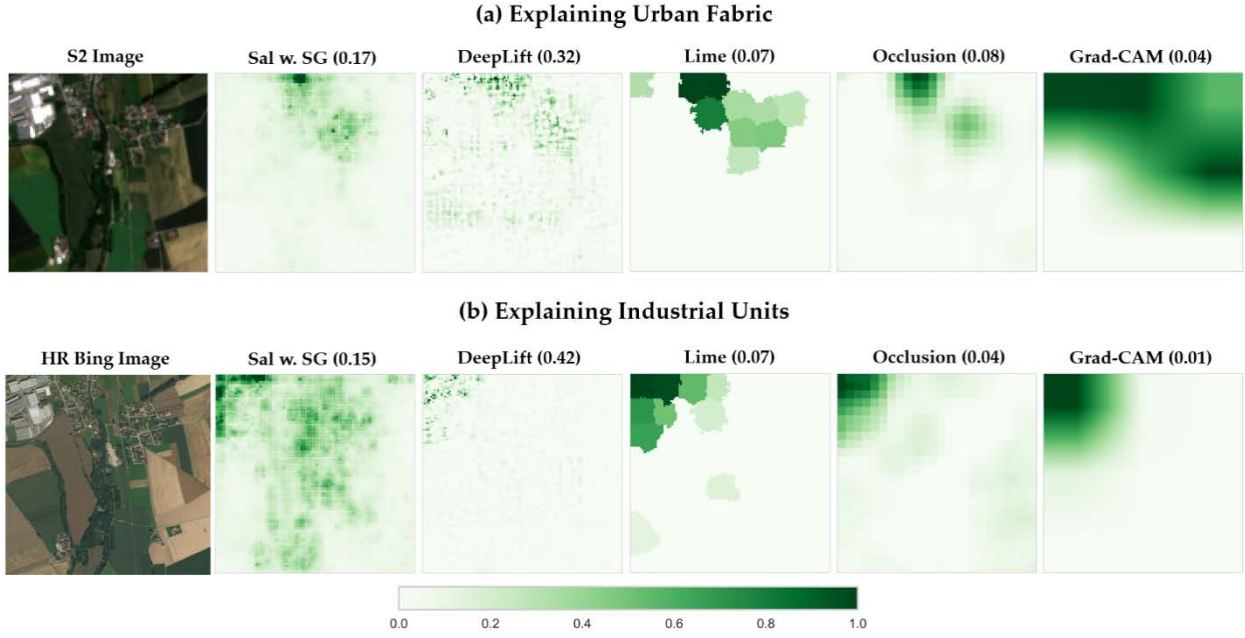


Figure 5. Explaining the Predictions of DenseNet for the class of (a) *Urban Fabric* and (b) *Industrial Units* in BigEarthNet dataset (Image ID: S2A_MSIL2A_20170613T101031_47_45). *Max-Sensitivity* score in parenthesis after method's name.

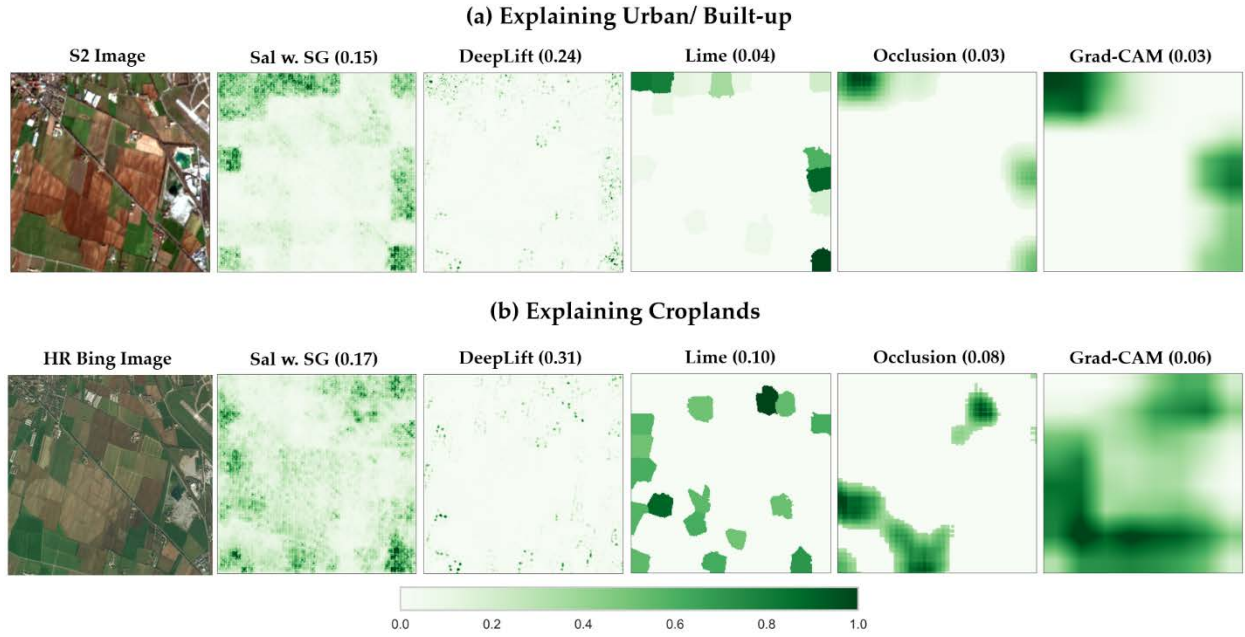


Figure 6. Explaining the Predictions of DenseNet for the class of (a) *Urban/ Built-up* and (b) *Croplands* in SEN12MS dataset (Image ID: ROIs2017_winter_s2_144_p760). *Max-Sensitivity* score in parenthesis after method's name.

Moreover, another indicative case for the explanations of *Forest* and *Savanna* class decisions on SEN12MS dataset is demonstrated in Figures 7 (and corresponding Figures S9a, S9b). The model accurately predicted both *Forest* and *Savanna* with 0.97 and 1.00 sigmoid probability scores, respectively. After visual inspection, we observed that all methods provided valuable information for *Forest* class prediction to the user. *Occlusion* and *Grad-CAM* were the most interpretable. However, only *Grad-CAM* managed to highlight the *Savanna* class region, indicating that this method is the most class-discriminative one. It is worth mentioning that *Savanna* class identification is challenging

also for an expert, as *Savanna* consists of a mixed ecosystem and may span spatially the entire image. Additionally, it is highlighted that *Guided Backpropagation* focused on image regions that are not necessarily relevant to the examined class.

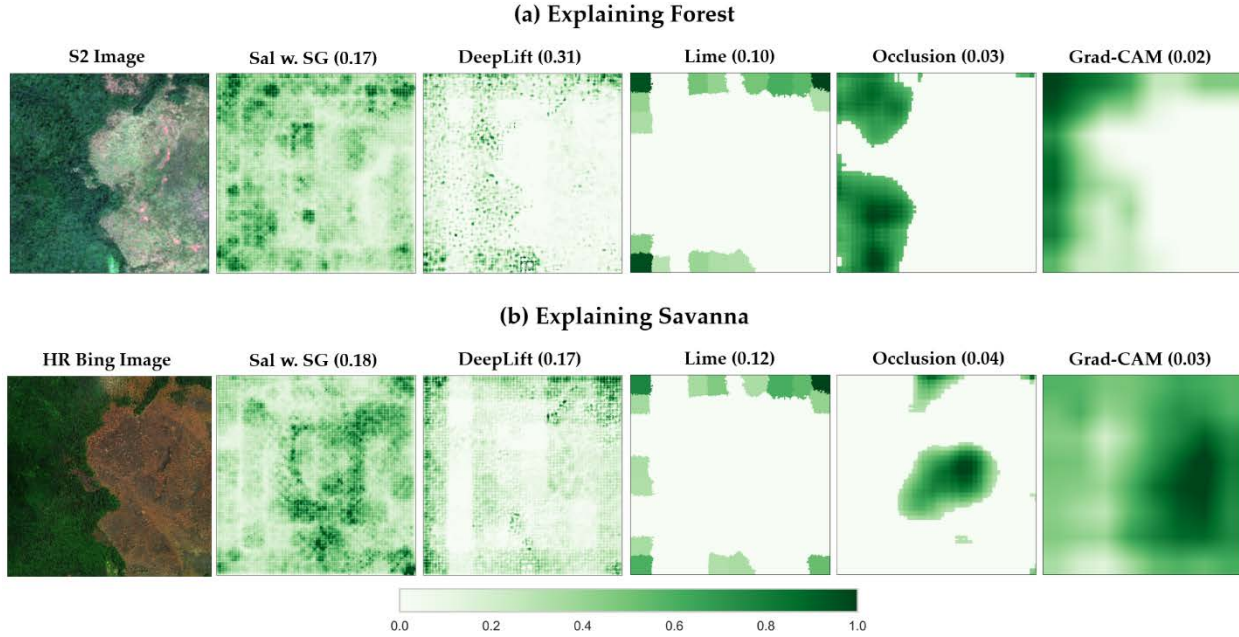


Figure 7. Explaining the Predictions of DenseNet for the class of (a) *Forest* and (b) *Savanna* in SEN12MS dataset (Image ID: ROIs1970_fall_s2_35_p297). *Max-Sensitivity* score in parenthesis after method's name.

3.2.3. Explaining Model Failure

Furthermore, we examined several cases that the models failed to predict the correct classes, which were included in the ground truth/ testing set and verified by us. In such model failure cases, explanations were used to obtain important insights regarding the performance of the considered XAI methods.

In particular, as Figure 8 demonstrates (as well as the corresponding Figure S10), the developed model incorrectly predicted *Coniferous Forest* as a label in this particular image. Indeed, the model focused on the dark green area (i.e., South-East region), which is probably a recently irrigated crop field and misclassified it as a *Coniferous Forest*, based on the relatively darker intensity values. All XAI methods agreed with each other and focused on this particular image region.

Another indicative case is demonstrated in Figures 9 and S11. The model predicted the class *Water* with high confidence (0.95 sigmoid probability). Indeed, although it seems that a stream is crossing this highly dense urban area, *Water* is correctly not part of the ground truth due to its size and width. This misclassification case was due to the extended shadows that are covering the image. In particular, explanation methods indicated that the model was mainly focused on this darker cover with cloud shadows and not on any detected stream, river or water area. Except for *DeepLift* and *Guided Backpropagation*, the rest of the methods are explaining this decision quite successfully. Similar images with cloud shadows from the same region (i.e., ROIs1970_fall_s2_116) were also investigated to confirm our results. We have to mention that during data preparation for SEN12MS, [34] developed a sophisticated workflow to generate cloud-free S2 mosaics. However, only a few images with cloud shadows were included in the dataset; thus, the model could not generalize adequately in cases with cloud shadows.

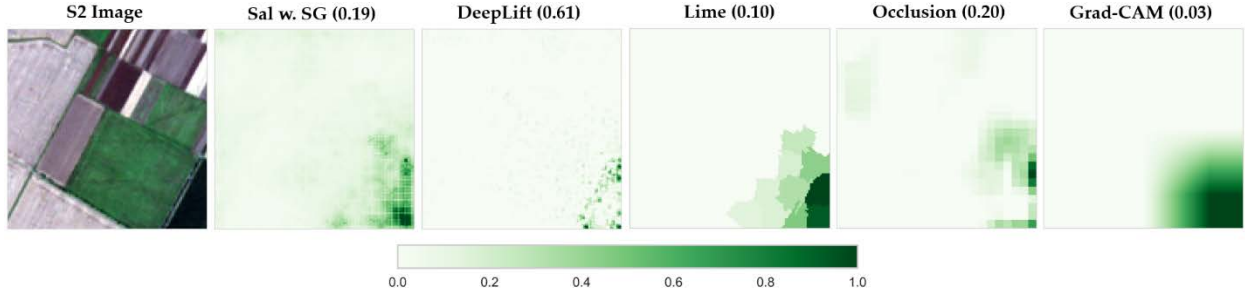


Figure 8. Explaining the Predictions of DenseNet for *Coniferous Forest* in BigEarthNet (Image ID: S2A_MSIL2A_20171002T094031_64_36). *Max-Sensitivity* score in parenthesis after method's name.

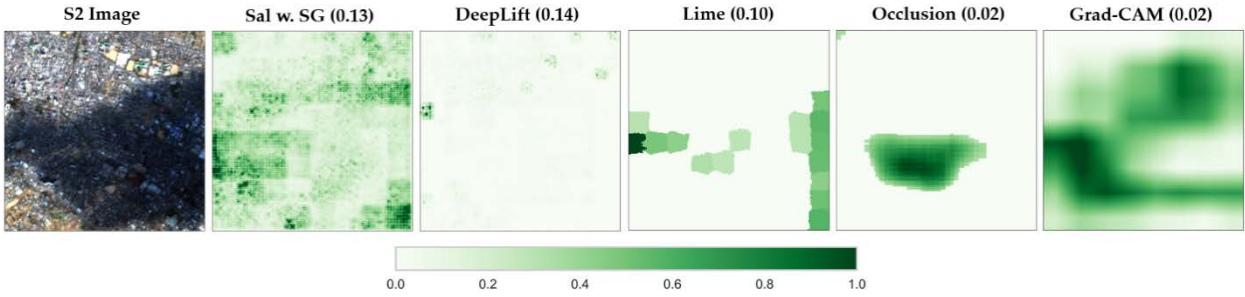


Figure 9. Explaining the Predictions of DenseNet for *Water* in SEN12MS (Image ID: ROIs1970_fall_s2_116_p703). *Max-Sensitivity* score in parenthesis after method's name.

3.2.4. Explaining Failure due to Dataset/ Inexact Labeling

During the intensive examination of numerous model prediction cases and in particular, cases that model prediction and labeling were not consistent, we also focused on cases where inexact labeling in the testing dataset occurred. For instance, seaport installations (which according to CORINE belong to the *Industrial Units* class) were not labeled as *Industrial Units* in an image of BigEarthNet dataset (Figures 10, S12). Nevertheless, the model managed to predict *Industrial Units* with a 0.78 sigmoid probability score. For this particular case, the majority of XAI methods focused on the actual industrial buildings that exist in the North area of the image explaining sufficiently the correct decision of the developed model. However, *Saliency* and *Guided Backpropagation* focused additionally on irrelevant regions (i.e., South and South-West regions), delivering less clear explanations.

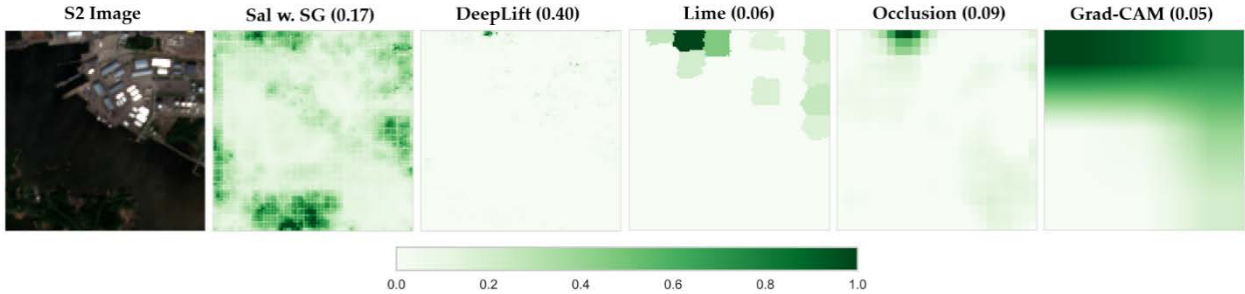


Figure 10. Explaining the Predictions of DenseNet for *Industrial Units* in BigEarthNet (Image ID: S2A_MSIL2A_20170613T101031_23_87). *Max-Sensitivity* score in parenthesis after method's name.

Moreover, another indicative example of inexact labeling in the SEN12MS dataset is demonstrated in Figures 11 (and S13), where *Urban/ Built-up* class is not included in labeling. Similarly, the model managed to predict *Urban/*

Built-up with a 0.97 sigmoid probability score. In particular, none of the MODIS land cover pixels were labeled with *Urban*. However, all XAI methods agreed with each other and correctly indicated the actual urban areas in the image (North-West and South-East regions) that the model managed to detect accurately.

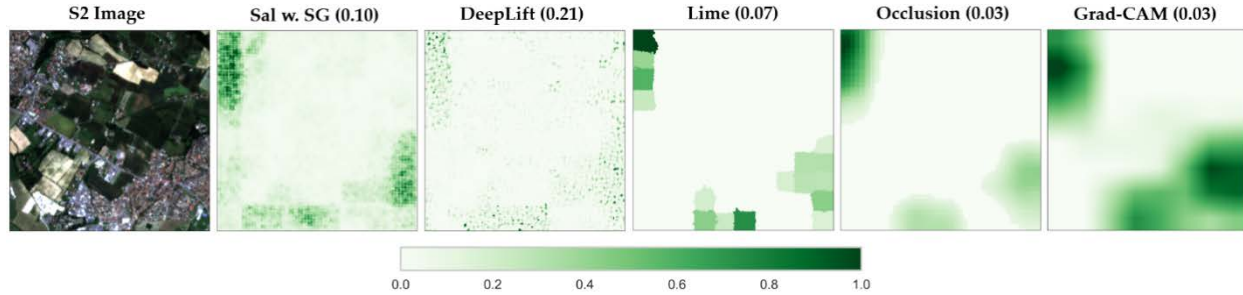


Figure 11. Explaining the Predictions of DenseNet for *Urban/ Built-up* in SEN12MS (Image ID: ROIs1158_spring_s2_148_p197). *Max-Sensitivity* score in parenthesis after method's name.

4. Discussion

Based on the aforementioned outcomes, we discuss the gained insights during the evaluation of XAI methods for multi-label classification tasks. Moreover, we summarize the explainability and applicability aspects of the studied methods derived from both quantitative and qualitative results (Table 6). Finally, further extracted insights regarding the black-box model decisions and the studied datasets composition are described.

4.1. Performance of Explainable AI methods for Multi-label Classifications Tasks

According to our findings, *Occlusion*, *Grad-CAM* and *Lime* were the most interpretable XAI methods, as well as the ones that were able to explain competing multi-class decisions (class-discriminative) by also locating the corresponding image regions successfully. Quantitatively, these three methods achieved relatively low sensitivity (*Max-Sensitivity*) scores. This fact indicates that there are no significantly different explanations when input slightly varies [42], which leads to trust their explanations (high reliability). *AUC-MoRF* metric results also confirmed that these methods are reliable. Additionally, these methods produced the smallest output file sizes and thus were likely more interpretable, which is in accordance with the literature [7].

On the other hand, *Guided Backpropagation* was the least reliable method as it was insensitive concerning different classes. For several cases, it focused on image regions with striking primitives (e.g., edges) independently of the predicted class, leading to outputs almost identical to the ones of an edge detector. This behavior was also observed, to a lesser extent, from the *Saliency*, *Input \times Gradient* and *Integrated Gradients* methods. Similar observations and shortcomings have also been reported by [53] through their saliency checks, consisting of model parameters and data randomization tests. They concluded that saliency methods mostly employ the edges of an image. For classes that contain sharp edges, all methods may seem visually reliable highlighting these image features; however, this can be misleading as, e.g., *Guided Backpropagation* keeps highlighting the same image regions independently of the considered thematic classes.

Additional insights extracted through the qualitative evaluation are presented below. *Saliency*, *Input \times Gradient*, *Integrated Gradients* and *DeepLift* in certain cases failed to provide accurate information regarding the location of the class; this was mainly for classes that were spatially distributed in the image (e.g., *Savanna*). Moreover, *Guided Grad-CAM* was more sensitive than *Guided Backpropagation* w.r.t. different labels. We also confirm the study by [17] that indicated that the *Guided Grad-CAM* is able to highlight more fine-grained details than *Grad-CAM* in the explanations. Furthermore, experiments made with *SG* indicated that methods that integrated this approach achieved lower sensitivity scores than default methods. Visually there were no significant differences in the explanations; however,

noise in the default methods’ outputs from irrelevant image regions was reduced. This finding is in accordance also with [41].

Table 6 summarizes XAI methods explainability regarding reliability, interpretability and output resolution as well as applicability in terms of computation time, scalability against different input sizes, model agnostic and reference input requirement. Aforementioned XAI aspects are demonstrated based on both quantitative and qualitative outcomes. Reliability reflects the quantitative results from *Max-Sensitivity* and *AUC-MoRF* metrics (Sect. 3.1) as well as our own qualitative visual examination (Sect. 3.2). Similarly, interpretability describes the *File Size* results (Sect. 3.1) and the qualitative evaluation (Sect. 3.2).

Table 6. An overall evaluation for the studied XAI methods.

The assigned scores correspond to Yes/ No (Y, N) and Low/ Moderate/ High (L, M, H) performance.

Explainability				Applicability			
Method	Reliability	Interpretability	Resolution	Model Agnostic	Computation Time	Scalability	Required Reference
Sal	M	M	H	N	L	H	N
InputXGrad	M	M	H	N	L	H	N
IntGrad	M	M	H	N	M	H	Y
Guided Backprop	L	L	H	N	L	H	N
Grad-CAM	H	H	L	N	L	H	N
Guided Grad-CAM	L	M	H	N	L	H	N
DeepLift	M	L	H	N	L	H	Y
Occlusion	H	H	M	Y	H	L	Y
Lime	H	H	M	Y	H	L	Y

As presented in Table 6, although *Occlusion*, *Grad-CAM* and *Lime* are the most reliable and interpretable, they are the lowest resolution methods. Instead, *Saliency*, *Input \times Gradient*, *Integrated Gradients*, *Guided Backpropagation*, *Guided-grad CAM* and *DeepLift* methods are high-resolution; thus, they deliver fine-grained details in their output. Regarding applicability aspects, only *Lime* and *Occlusion* are model agnostic methods (i.e., they are independent of the employed model/ prediction method). This fact reveals that the specific methods are applicable beyond the existence of model function partial derivatives w.r.t. the inputs, or neural networks architecture. Also, in general, the fewer the prerequisites for the execution of an XAI method, the more straightforward applicability for the user. Specifically, in our study, the determination of a satisfactory reference point was time-consuming (e.g., the scale of the blurring for the *DeepLift* reference input), leading to the corresponding scores in Table 6.

Overall, we observe that there is not a single method that stands out as the best one. *Grad-CAM* is highly reliable, interpretable, scalable, and requires less computational time but does not provide high-resolution outputs. Both *Lime* and *Occlusion* are highly reliable, interpretable and model agnostic with moderate resolution outputs but with relatively high computational time and low scalability.

4.2. XAI for further insights in black-box models and benchmark datasets

Through our study, the developed models resulted in state-of-the-art performance in all evaluated metrics for both datasets. We have to mention though, that classes with the highest F_1 scores or high probability predictions for the examined cases did not necessarily lead to straightforward interpretable explanations (e.g., localize the corresponding image regions for a given class). For instance, in Figure S14 the model correctly predicted *Coniferous Forest* class with 1.00 sigmoid probability. Additionally, the overall F_1 score for this class was 87.79% (Table S1). However, all XAI methods did not achieve quite informative results, especially regarding the dominant location in the image. On the other hand, in the case presented in Figures 5 and S7b, all methods resulted in quite interpretable explanations for the *Industrial or Commercial Units* class with 0.64 sigmoid probability. For the considered class, F_1 score was 52.79% (Table S1), as well as Precision score was 71.18% (Table S2) which was lower than the average Precision score. Similarly, in Figure S9b from SEN12MS dataset, despite the high 1.00 sigmoid probability score and the high overall F_1 score 84.65% for *Savanna* class (Table S3), XAI models resulted in relatively poor explanations regarding the location of the particular class in the image.

XAI methods contributed with further insights towards understanding models' predictions due to datasets particularities like training set class distribution. More specifically, by studying results and explanations for BigEarthNet, we observed that, when the model correctly predicted *Marine Water* class, occasionally, it was also focused on *Beaches, Dunes, Sands*. At the same time, for several cases, the model could not successfully predict *Beaches, Dunes, Sands* area (e.g., Figure S15). Additionally, F_1 score for this class was the third lowest (i.e., 63.30%) (Table S1). This fact is probably attributed to two highly correlated reasons. Firstly, a high percentage (i.e., 79%) of images labeled with *Beaches, Dunes, Sands* also include the *Marine Water* label (Figure S16). Secondly, a significant number of images that both *Marine Water* and *Beaches, Dunes, Sands* classes were depicted, were not properly labeled with *Beaches, Dunes, Sands*. These findings are also in-line with a recent study [54], which indicates that small beaches, dunes and sand plains were omitted from the *Beaches, Dunes, Sands* class of CLC2018 in Norway, while a considerable area with ocean/water was assigned in the *Beaches, Dunes, Sands* class.

In a similar way, in SEN12MS, a high percentage (i.e., ~65%) of images with *Wetlands* includes *Water* label, while a high percentage (i.e., ~39%) of the cases with *Water* contains *Wetlands* as well (Figure S1). Moreover, by studying intensively numerous cases, we found several inexact cases that the class *Wetlands* was missing from the ground truth labels. Indeed, [55] reported that the largest omission errors for IGBP classification were recorded for certain classes, including *Wetlands*. Thus, XAI methods indicated that the correct *Water* class predictions were focused on image regions/ features of both *Water* and *Wetlands* (Figure S17). At the same time, the model did not manage to predict the existence of *Wetlands* in images into which *Water* was dominating, resulting in False Negative cases (Figure S17).

5. Conclusions

To sum up, we developed deep learning models with state-of-the-art performance in multi-label RS benchmark datasets towards various XAI methods evaluation. In order to quantitatively evaluate XAI performance, *Max-Sensitivity*, *Area Under the Most Relevant First* perturbation curve, *File Size* and *Computational Time* metrics were utilized. Extensive experiments were performed to qualitatively examine and assess the function of the considered methods. We further investigated different aspects of XAI methods regarding their applicability and explainability. Through our evaluation procedure, we found that none of the XAI methods stands out as the best one. *Occlusion*, *Grad-CAM* and *Lime* were the most interpretable and reliable XAI methods presenting the lowest *Max-Sensitivity* and *AUC-MoRF* scores. However, none of them provides high-resolution outputs and apart from *Grad-CAM*, both *Lime* and *Occlusion* are not computationally efficient. Overall, our findings indicate that XAI provides valuable insights for deep black-box models' performance and decisions as well as benchmark datasets' composition and shortcomings.

Acknowledgments: Part of this work was funded by the Operational Program "Competitiveness, Entrepreneurship and Innovation 2014-2020" (co-funded by the European Regional Development Fund).

References

1. Deng, L.; Yu, D. Deep Learning: Methods and Applications. *SIG* **2014**, *7*, 197–387, doi:[10.1561/20000000039](https://doi.org/10.1561/20000000039).
2. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep Learning in Remote Sensing Applications: A Meta-Analysis and Review. *ISPRS Journal of Photogrammetry and Remote Sensing* **2019**, *152*, 166–177, doi:[10.1016/j.isprsjprs.2019.04.015](https://doi.org/10.1016/j.isprsjprs.2019.04.015).
3. Qiangqiang Yuan, Huanfeng Shen, Tongwen Li, Zhiwei Li, Shuwen Li, Yun Jiang, Hongzhang Xu, Weiwei Tan, Qianqian Yang, Jiwen Wang, Jianhao Gao, Liangpei Zhang, **2020**. Deep learning in environmental remote sensing: Achievements and challenges, *Remote Sensing of Environment*, Volume 241, <https://doi.org/10.1016/j.rse.2020.111716>.
4. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, 2013; ISBN 978-1-4614-6848-6.
5. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery: New York, NY, USA, August 13 2016; pp. 1135–1144.
6. Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; Müller, K.-R. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nature Communications* **2019**, *10*, 1096, doi:[10.1038/s41467-019-08987-4](https://doi.org/10.1038/s41467-019-08987-4).
7. Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C.J.; Müller, K.-R. Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond. *arXiv:2003.07631 [cs, stat]* **2020**.
8. Cadamuro, G.; Gilad-Bachrach, R.; Zhu, X. Debugging Machine Learning Models. In Proceedings of the ICML Workshop on Reliable Machine Learning in the Wild; 2016.
9. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs. *IEEE Transactions on Geoscience and Remote Sensing* **2018**, *56*, 2811–2821, doi:[10.1109/TGRS.2017.2783902](https://doi.org/10.1109/TGRS.2017.2783902).
10. Sumbul, G.; Demir, B. A Deep Multi-Attention Driven Approach for Multi-Label Remote Sensing Image Classification. *IEEE Access* **2020**, *8*, 95934–95946, doi:[10.1109/ACCESS.2020.2995805](https://doi.org/10.1109/ACCESS.2020.2995805).
11. P. Li, P. Chen, Y. Xie and D. Zhang, **2020**. Bi-Modal Learning With Channel-Wise Attention for Multi-Label Image Classification," in IEEE Access, vol. 8, pp. 9965-9977, doi: 10.1109/ACCESS.2020.2964599
12. Wang, S.; Chen, W.; Xie, S.M.; Azzari, G.; Lobell, D.B. Weakly Supervised Deep Learning for Segmentation of Remote Sensing Imagery. *Remote Sensing* **2020**, *12*, 207, doi:[10.3390/rs12020207](https://doi.org/10.3390/rs12020207).
13. Camps-Valls, G.; Reichstein, M.; Zhu, X.; Tuia, D. ADVANCING DEEP LEARNING FOR EARTH SCIENCES: FROM HYBRID MODELING TO INTERPRETABILITY. In Proceedings of the IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium; September 2020; pp. 3979–3982.
14. Roscher, R.; Bohn, B.; Duarte, M.F.; Garcke, J. EXPLAIN IT TO ME – FACING REMOTE SENSING CHALLENGES IN THE BIO- AND GEOSCIENCES WITH EXPLAINABLE MACHINE LEARNING. In Proceedings of the ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences; Copernicus GmbH, August 3 2020; Vol. V-3–2020, pp. 817–824.
15. Nagasubramanian, K.; Jones, S.; Singh, A.K.; Sarkar, S.; Singh, A.; Ganapathysubramanian, B. Plant Disease Identification Using Explainable 3D Deep Learning on Hyperspectral Images. *Plant Methods* **2019**, *15*, 98, doi:[10.1186/s13007-019-0479-8](https://doi.org/10.1186/s13007-019-0479-8).
16. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In Proceedings of the Workshop at International Conference on Learning Representations; 2014.
17. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); October 2017; pp. 618–626.

18. McGovern, A.; Lagerquist, R.; Gagne, D.J.; Jergensen, G.E.; Elmore, K.L.; Homeyer, C.R.; Smith, T. Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bulletin of the American Meteorological Society* **2019**, *100*, 2175–2199, doi:[10.1175/BAMS-D-18-0195.1](https://doi.org/10.1175/BAMS-D-18-0195.1).
19. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 2016; pp. 2921–2929.
20. Vasu, B.; Rahman, F.U.; Savakis, A. Aerial-CAM: Salient Structures and Textures in Network Class Activation Maps of Aerial Imagery. In Proceedings of the 2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP); June 2018; pp. 1–5.
21. Yang, R.; Xu, X.; Xu, Z.; Ding, C.; Pu, F. A Class Activation Mapping Guided Adversarial Training Method for Land-Use Classification and Object Detection. In Proceedings of the IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium; July 2019; pp. 9474–9477.
22. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV); March 2018; pp. 839–847.
23. Fu, K.; Dai, W.; Zhang, Y.; Wang, Z.; Yan, M.; Sun, X. MultiCAM: Multiple Class Activation Mapping for Aircraft Recognition in Remote Sensing Images. *Remote Sensing* **2019**, *11*, 544, doi:[10.3390/rs11050544](https://doi.org/10.3390/rs11050544).
24. Johnson, J.E.; Laparra, V.; Pérez-Suay, A.; Mahecha, M.D.; Camps-Valls, G. Kernel Methods and Their Derivatives: Concept and Perspectives for the Earth System Sciences. *PLOS ONE* **2020**, *15*, e0235885, doi:[10.1371/journal.pone.0235885](https://doi.org/10.1371/journal.pone.0235885).
25. Wolanin, A.; Mateo-García, G.; Camps-Valls, G.; Gómez-Chova, L.; Meroni, M.; Duveiller, G.; Liangzhi, Y.; Guanter, L. Estimating and Understanding Crop Yields with Explainable Deep Learning in the Indian Wheat Belt. *Environ. Res. Lett.* **2020**, *15*, 024019, doi:[10.1088/1748-9326/ab68ac](https://doi.org/10.1088/1748-9326/ab68ac).
26. Yessou, H.; Sumbul, G.; Demir, B. A Comparative Study of Deep Learning Loss Functions for Multi-Label Remote Sensing Image Classification. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS); Hawaii, USA, 2020.
27. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE* **2015**, *10*, e0130140, doi:[10.1371/journal.pone.0130140](https://doi.org/10.1371/journal.pone.0130140).
28. Campos-Taberner, M.; García-Haro, F.J.; Martínez, B.; Izquierdo-Verdiguier, E.; Atzberger, C.; Camps-Valls, G.; Gilabert, M.A. Understanding Deep Learning in Land Use Classification Based on Sentinel-2 Time Series. *Scientific Reports* **2020**, *10*, 17188, doi:[10.1038/s41598-020-74215-5](https://doi.org/10.1038/s41598-020-74215-5).
29. Pérez-Suay, A.; Adsua, J.E.; Piles, M.; Martínez-Ferrer, L.; Díaz, E.; Moreno-Martínez, A.; Camps-Valls, G. Interpretability of Recurrent Neural Networks in Remote Sensing. In Proceedings of the IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium; September 2020; pp. 3991–3994.
30. Martínez-Ferrer, L.; Piles, M.; Camps-Valls, G. Crop Yield Estimation and Interpretability With Gaussian Processes. *IEEE Geoscience and Remote Sensing Letters* **2020**, 1–5, doi:[10.1109/LGRS.2020.3016140](https://doi.org/10.1109/LGRS.2020.3016140).
31. Levering, A.; Marcos, D.; Lobry, S.; Tuia, D. Interpretable Scenicness from Sentinel-2 Imagery. In Proceedings of the IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium; September 2020; pp. 3983–3986.
32. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 2017; pp. 2261–2269.
33. Sumbul, G.; Charfuelan, M.; Demir, B.; Markl, V. Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. In Proceedings of the IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium; July 2019; pp. 5901–5904.
34. Schmitt, M.; Hughes, L.H.; Qiu, C.; Zhu, X.X. SEN12MS - A CURATED DATASET OF GEOREFERENCED MULTI-SPECTRAL SENTINEL-1/2 IMAGERY FOR DEEP LEARNING AND

- DATA FUSION. In Proceedings of the ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences; Copernicus GmbH, September 16 2019; Vol. IV-2-W7, pp. 153–160.
35. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features Through Propagating Activation Differences. In Proceedings of the International Conference on Machine Learning; PMLR, July 17 2017; pp. 3145–3153.
 36. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the Proceedings of the 34th International Conference on Machine Learning - Volume 70; JMLR.org: Sydney, NSW, Australia, August 6 2017; pp. 3319–3328.
 37. Goh, G.S.W.; Lapuschkin, S.; Weber, L.; Samek, W.; Binder, A. Understanding Integrated Gradients with SmoothTaylor for Deep Neural Network Attribution. *arXiv:2004.10484 [cs, stat]* **2020**.
 38. Springenberg, J.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for Simplicity: The All Convolutional Net.; 2015.
 39. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the Computer Vision – ECCV 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, 2014; pp. 818–833.
 40. Ancona, M.; Ceolini, E.; Öztireli, C.; Gross, M. Towards Better Understanding of Gradient-Based Attribution Methods for Deep Neural Networks. In Proceedings of the International Conference on Learning Representations; February 2018.
 41. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. Smoothgrad: Removing Noise by Adding Noise. In Proceedings of the Workshop on Visualization for Deep Learning, ICML; Sydney, Australia, 2017.
 42. Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A.; Inouye, D.I.; Ravikumar, P.K. On the (In)Fidelity and Sensitivity of Explanations. In Proceedings of the Advances in Neural Information Processing Systems; 2019; pp. 10967–10978.
 43. Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; Müller, K. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems* **2017**, *28*, 2660–2673, doi:10.1109/TNNLS.2016.2599820.
 44. Feranec, J.; Soukup, T.; Hazeu, G.; Jaffrain, G. *European Landscape Dynamics : CORINE Land Cover Data*; CRC Press, 2016; ISBN 978-1-315-37286-0.
 45. Sumbul, G.; Kang, J.; Kreuziger, T.; Marcelino, F.; Costa, H.; Benevides, P.; Caetano, M.; Demir, B. BigEarthNet Dataset with A New Class-Nomenclature for Remote Sensing Image Understanding. *arXiv:2001.06372 [cs]* **2020**.
 46. Loveland, T.R.; Belward, A.S. The International Geosphere Biosphere Programme Data and Information System Global Land Cover Data Set (DISCover). *Acta Astronautica* **1997**, *41*, 681–689, doi:[10.1016/S0094-5765\(98\)00050-2](https://doi.org/10.1016/S0094-5765(98)00050-2).
 47. Schmitt, M.; Prexl, J.; Ebel, P.; Liebel, L.; Zhu, X.X. WEAKLY SUPERVISED SEMANTIC SEGMENTATION OF SATELLITE IMAGES FOR LAND COVER MAPPING – CHALLENGES AND OPPORTUNITIES. In Proceedings of the ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences; Copernicus GmbH, August 3 2020; Vol. V-3–2020, pp. 795–802.
 48. Kingma, DP; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference for Learning Representations; San Diego, 2015.
 49. Robinson, C.; Malkin, K.; Jojic, N.; Chen, H.; Qin, R.; Xiao, C.; Schmitt, M.; Ghamisi, P.; Hansch, R.; Yokoya, N. Global Land Cover Mapping with Weak Supervision: Outcome of the 2020 IEEE GRSS Data Fusion Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2021**, *1*–1, doi:[10.1109/JSTARS.2021.3063849](https://doi.org/10.1109/JSTARS.2021.3063849).
 50. Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. *University of Toronto* **2012**.
 51. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2012**, *34*, 2274–2282, doi:[10.1109/TPAMI.2012.120](https://doi.org/10.1109/TPAMI.2012.120).

52. Kokhlikyan, N. Captum: A Unified and Generic Model Interpretability Library for PyTorch. *CoRR* **2020**, *abs/2009.07896*.
53. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity Checks for Saliency Maps. In Proceedings of the Proceedings of the 32nd International Conference on Neural Information Processing Systems; Curran Associates Inc.: Red Hook, NY, USA, December 3 2018; pp. 9525–9536.
54. Aune-Lundberg, L.; Strand, G.-H. The Content and Accuracy of the CORINE Land Cover Dataset for Norway. *International Journal of Applied Earth Observation and Geoinformation* 2021, 96, 102266, doi:10.1016/j.jag.2020.102266.
55. Sulla-Menashe, D.; Gray, J.M.; Abercrombie, S.P.; Friedl, M.A. Hierarchical Mapping of Annual Global Land Cover 2001 to Present: The MODIS Collection 6 Land Cover Product. *Remote Sensing of Environment* 2019, 222, 183–194, doi:10.1016/j.rse.2018.12.013.

Supplementary Material

Evaluating Explainable Artificial Intelligence Methods for Multi-label Deep Learning Classification Tasks in Remote Sensing

Ioannis Kakogeorgiou* and Konstantinos Karantzas

Remote Sensing Laboratory, National Technical University of Athens, Zographou, 15780, Greece; karank@central.ntua.gr

* Correspondence: gkakogeorgiou@central.ntua.gr; Tel.: +302107721673

Table S1. Per Class F₁ Scores on the BigEarthNet Dataset.

Class	K-Branch CNN	VGG16	VGG19	ResNet50	ResNet101	ResNet152	DenseNet121 (Ours)
Urban Fabric	71.52	74.49	74.45	74.84	74.50	74.32	79.51
Industrial or Commercial Units	38.66	44.41	42.68	48.55	49.11	50.15	52.79
Arable Land	80.55	82.40	82.04	83.85	82.96	82.92	87.02
Permanent Crops	47.47	51.53	48.59	51.91	42.35	56.46	68.68
Pastures	70.42	70.52	70.46	72.38	71.47	72.35	77.65
Complex Cultivation Patterns	62.68	62.68	61.98	66.03	65.91	64.07	72.93
Land Principally Occupied by Agriculture, with significant Areas of Natural Vegetation	59.68	59.61	61.08	60.94	63.41	60.95	70.18
Agro-forestry Areas	71.30	72.88	72.42	70.49	60.08	74.29	79.97
Broad-leaved Forest	73.03	73.85	73.14	74.05	73.98	75.36	79.82
Coniferous Forest	82.73	85.18	84.66	85.41	85.67	85.11	87.79
Mixed Forest	78.27	78.84	78.77	79.44	80.00	79.64	83.79
Natural Grassland and Sparsely Vegetated Areas	39.14	40.12	38.58	47.55	49.00	50.92	55.22
Moors, Heathland and Sclerophyllous Vegetation	43.79	52.23	54.03	59.41	55.70	46.69	68.57
Transitional Woodland-shrub	62.46	59.86	60.91	53.47	51.99	60.38	69.97
Beaches, Dunes, Sands	38.71	46.04	42.13	61.46	59.39	58.18	63.30
Inland Wetlands	48.07	54.65	53.59	60.64	59.10	60.44	66.79
Coastal Wetlands	19.85	21.15	17.88	47.71	27.26	45.88	63.52
Inland Waters	74.31	80.39	82.06	83.69	83.40	80.81	87.49
Marine Waters	88.28	96.55	96.76	97.53	97.77	95.16	98.70
Average	60.58	63.55	62.96	67.33	64.90	67.06	74.40

Table S2. Per Class Precision Scores on the BigEarthNet Dataset.

Class	K-Branch CNN	VGG16	VGG19	ResNet50	ResNet101	ResNet152	DenseNet121 (Ours)
Urban Fabric	70.69	76.80	75.90	72.82	78.11	78.80	80.06
Industrial or Commercial Units	47.23	56.25	48.99	53.86	54.70	52.88	71.18
Arable Land	85.67	86.68	86.54	82.26	77.76	85.37	86.38
Permanent Crops	43.53	52.49	45.02	76.93	76.73	69.16	72.25
Pastures	70.48	75.58	77.38	66.83	68.23	79.56	78.92
Complex Cultivation Patterns	51.14	66.01	64.21	68.23	61.97	70.75	69.74
Land Principally Occupied by Agriculture, with significant Areas of Natural Vegetation	56.80	66.36	63.59	64.04	61.45	66.42	66.56
Agro-forestry Areas	61.54	68.16	66.32	82.64	84.11	71.40	74.47
Broad-leaved Forest	70.45	78.77	75.06	81.03	80.09	73.80	80.76
Coniferous Forest	80.70	84.64	85.46	83.37	84.48	84.93	86.22
Mixed Forest	71.03	79.62	77.31	76.90	79.11	78.57	80.14
Natural Grassland and Sparsely Vegetated Areas	44.25	45.28	48.72	68.53	65.96	56.03	72.52
Moors, Heathland and Sclerophyllous Vegetation	52.32	57.67	50.51	68.36	70.11	74.73	70.17
Transitional Woodland- shrub	50.71	63.65	63.43	70.88	70.81	64.48	68.25
Beaches, Dunes, Sands	93.10	38.82	36.27	66.67	57.38	58.45	61.54
Inland Wetlands	60.00	62.88	66.93	73.67	68.70	71.04	75.87
Coastal Wetlands	39.81	41.51	20.59	82.54	27.76	60.96	61.96
Inland Waters	72.79	89.09	83.28	84.49	82.75	86.63	88.18
Marine Waters	98.93	95.99	97.01	96.80	97.21	98.84	98.59
Average	64.27	67.70	64.87	74.78	70.92	72.78	75.99

Table S3. Per Class Scores on the SEN12MS Dataset.

Class	Precision (%)	Recall (%)	F ₁ -score (%)	Support
Forest	78.21	75.65	76.90	9748
Shrubland	45.12	47.21	46.14	4482
Savanna	77.34	93.47	84.65	18860
Grassland	62.01	80.89	70.20	16100
Wetlands	56.26	75.51	64.48	3295
Croplands	71.69	78.39	74.89	12359
Urban/ Built-up	78.20	80.06	79.12	9697
Snow/Ice	0.00	0.00	0.00	10
Barren	47.51	77.30	58.85	2890
Water	72.09	88.65	79.52	4327
Average	58.84	69.71	63.48	81768

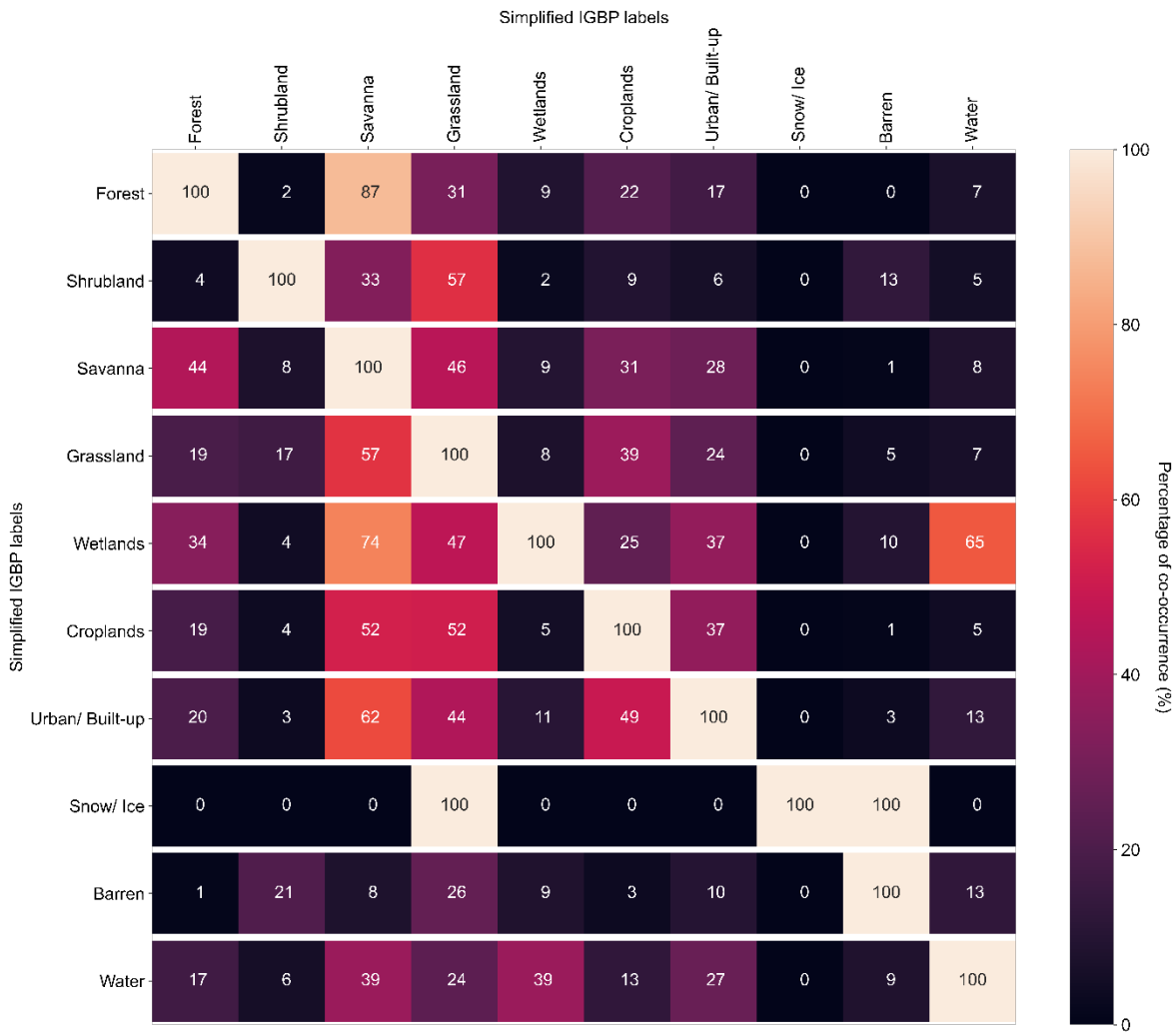


Figure S1. SEN12MS training set co-occurrence

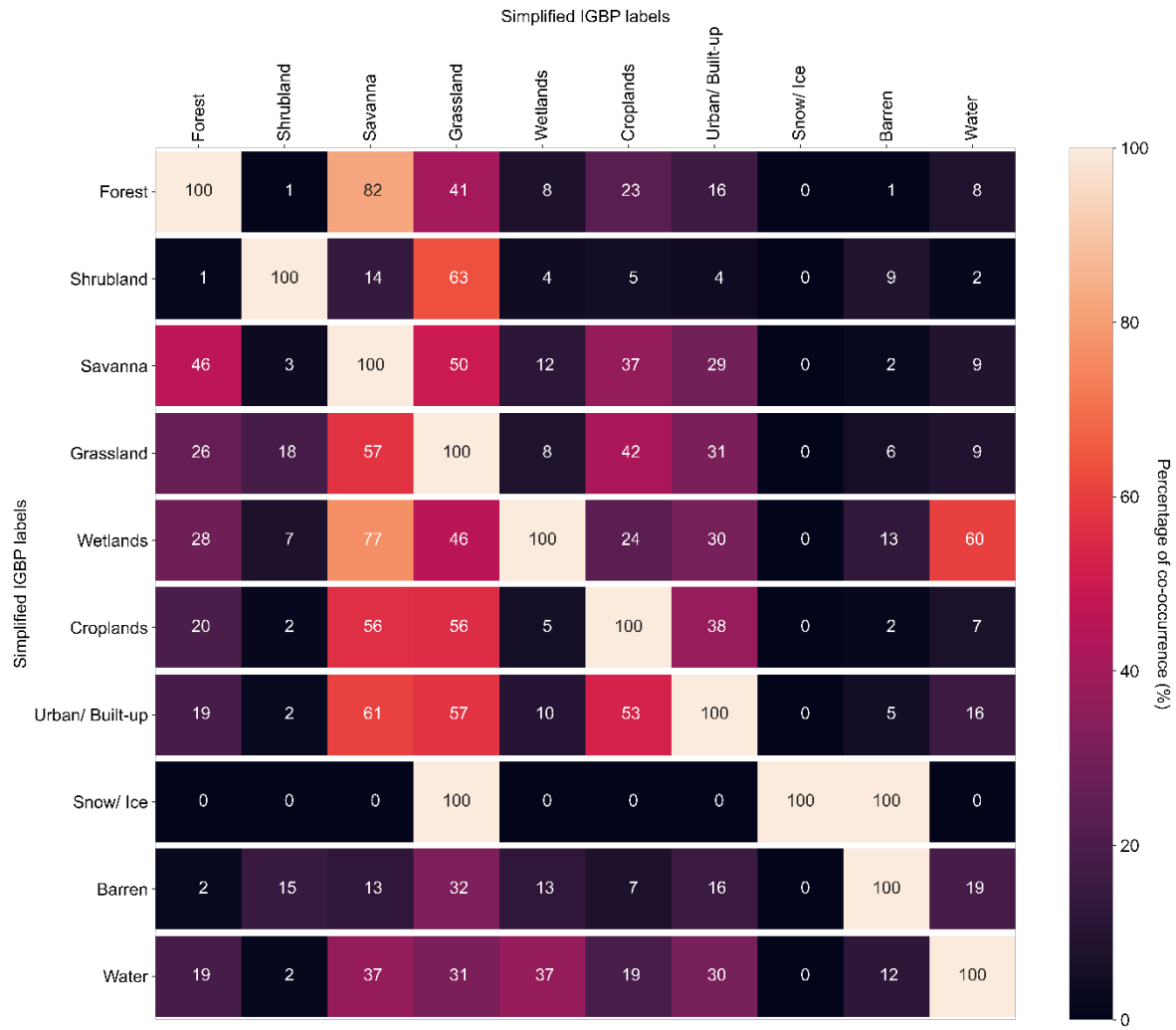


Figure S2. SEN12MS validation set co-occurrence

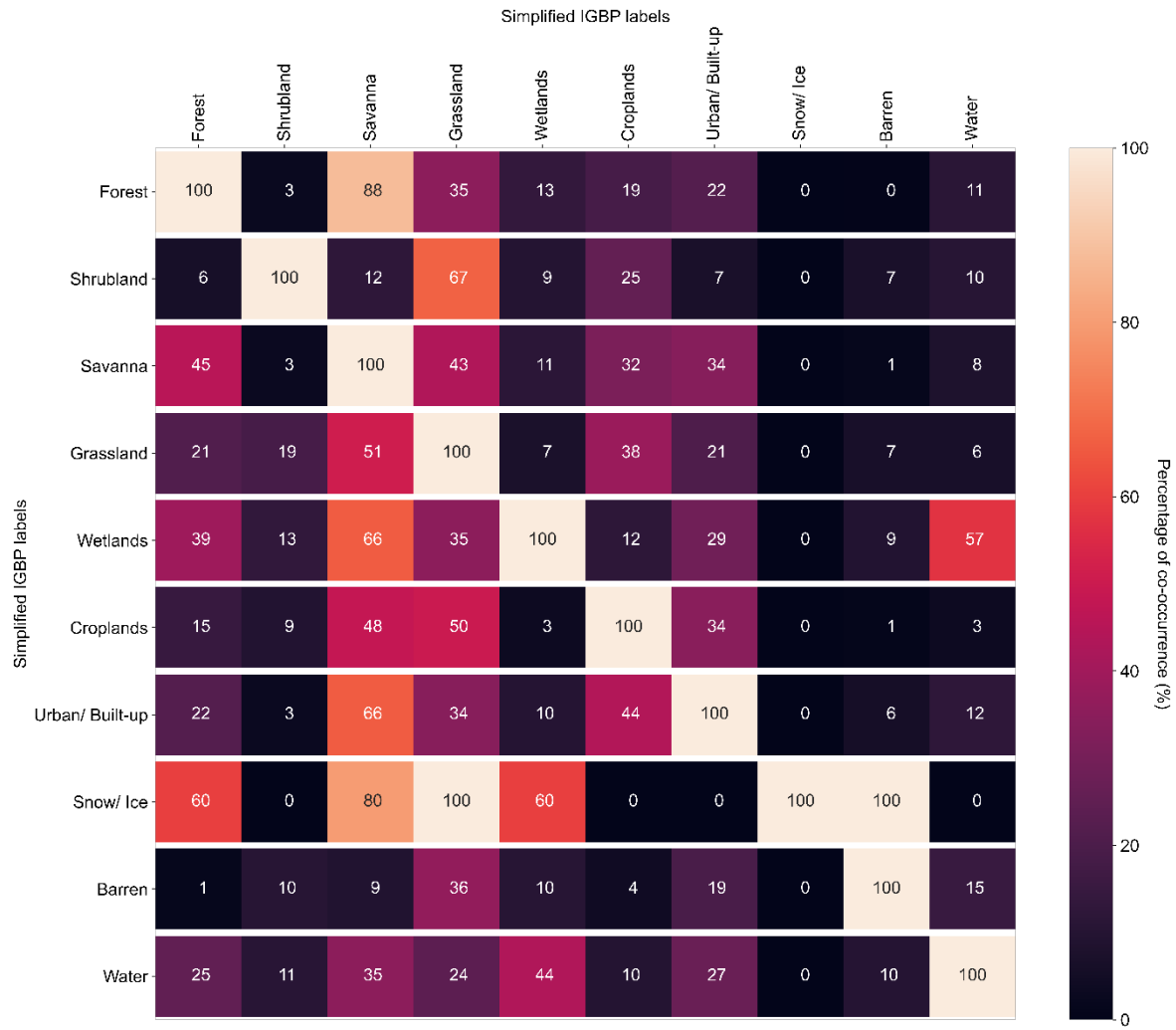


Figure S3. SEN12MS testing set co-occurrence

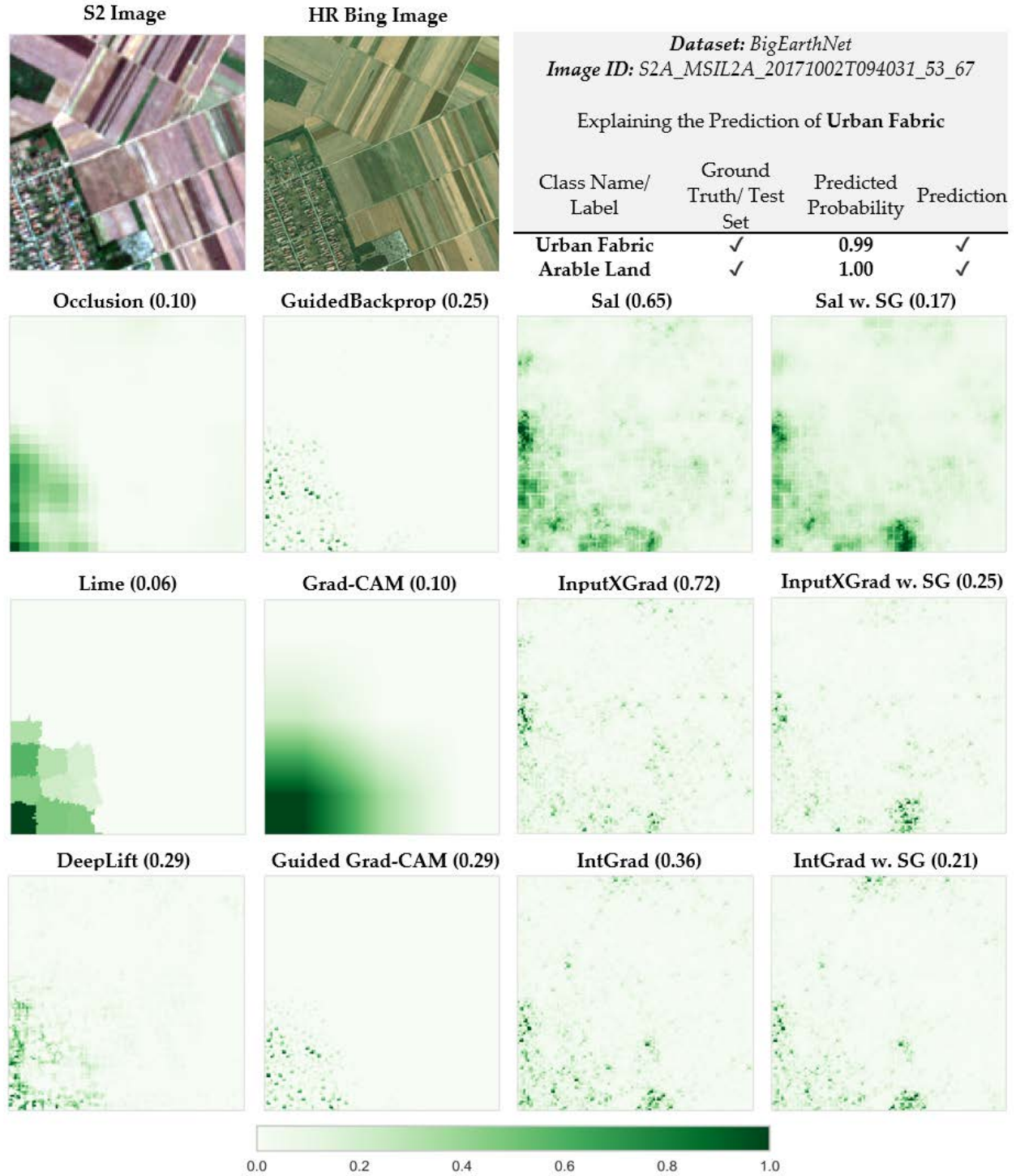


Figure S4. Explaining the Predictions of DenseNet for the class of *Urban Fabric* in BigEarthNet dataset (Image ID: S2A_MSIL2A_20171002T094031_53_67). *Max-Sensitivity* score in parenthesis after method's name.

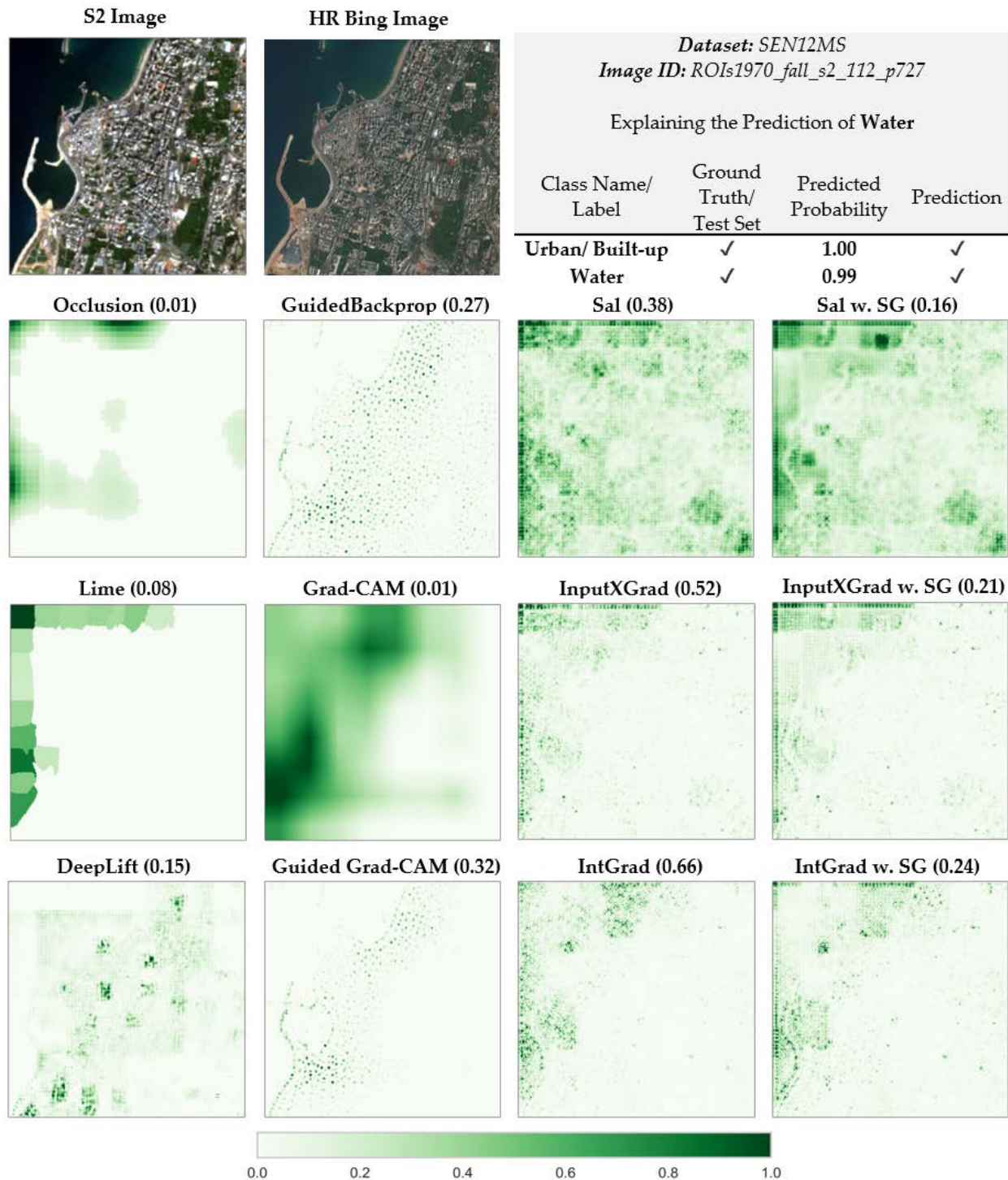


Figure S5. Explaining the Predictions of DenseNet for *Water* in SEN12MS (Image ID: ROIs1158_spring_s2_17_P110). *Max-Sensitivity* score in parenthesis after method's name.

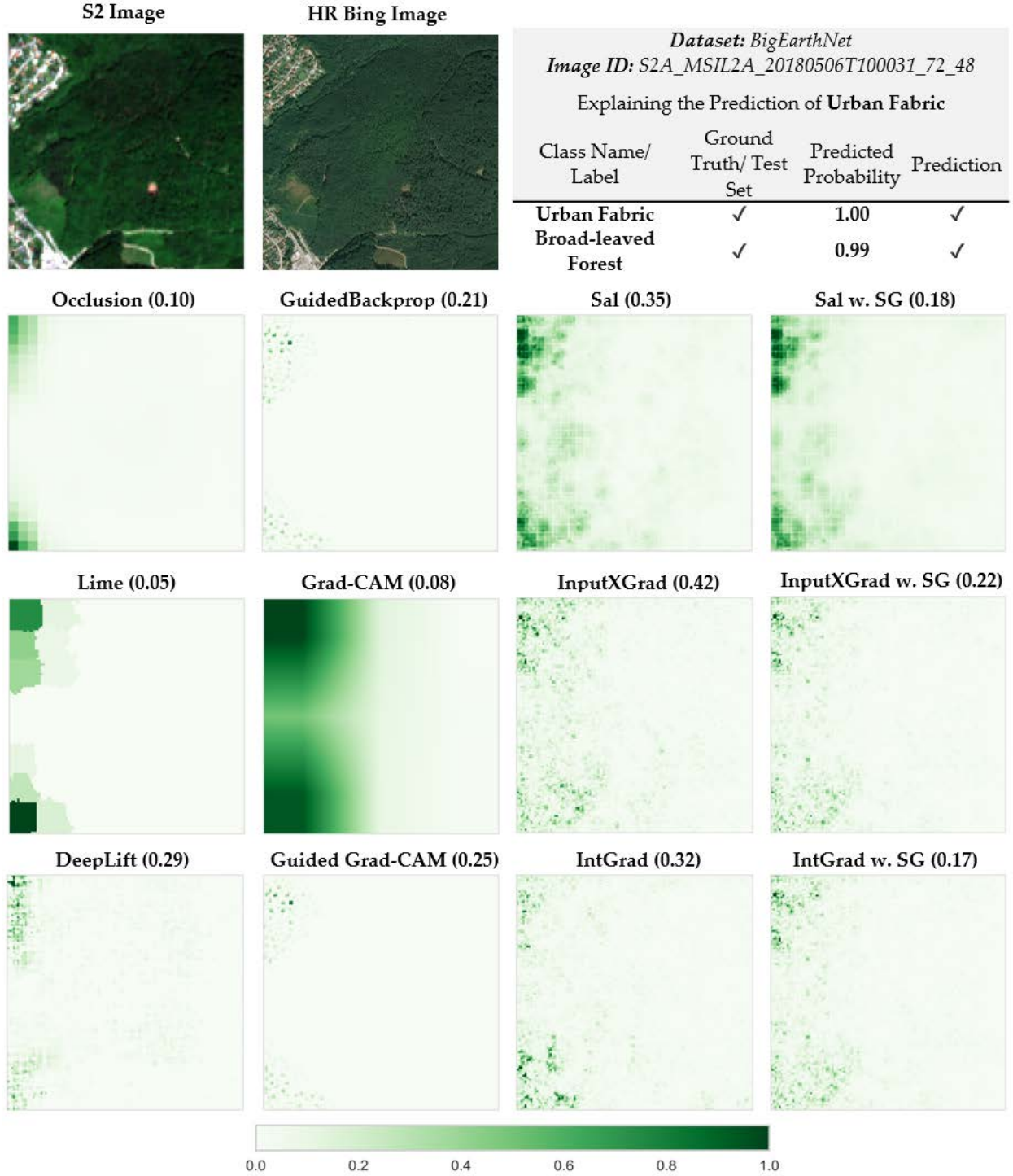


Figure S6a. Explaining the Predictions of DenseNet for *Urban Fabric* class in BigEarthNet dataset (Image ID: S2A_MSIL2A_20180506T100031_72_48). *Max-Sensitivity* score in parenthesis after method's name.

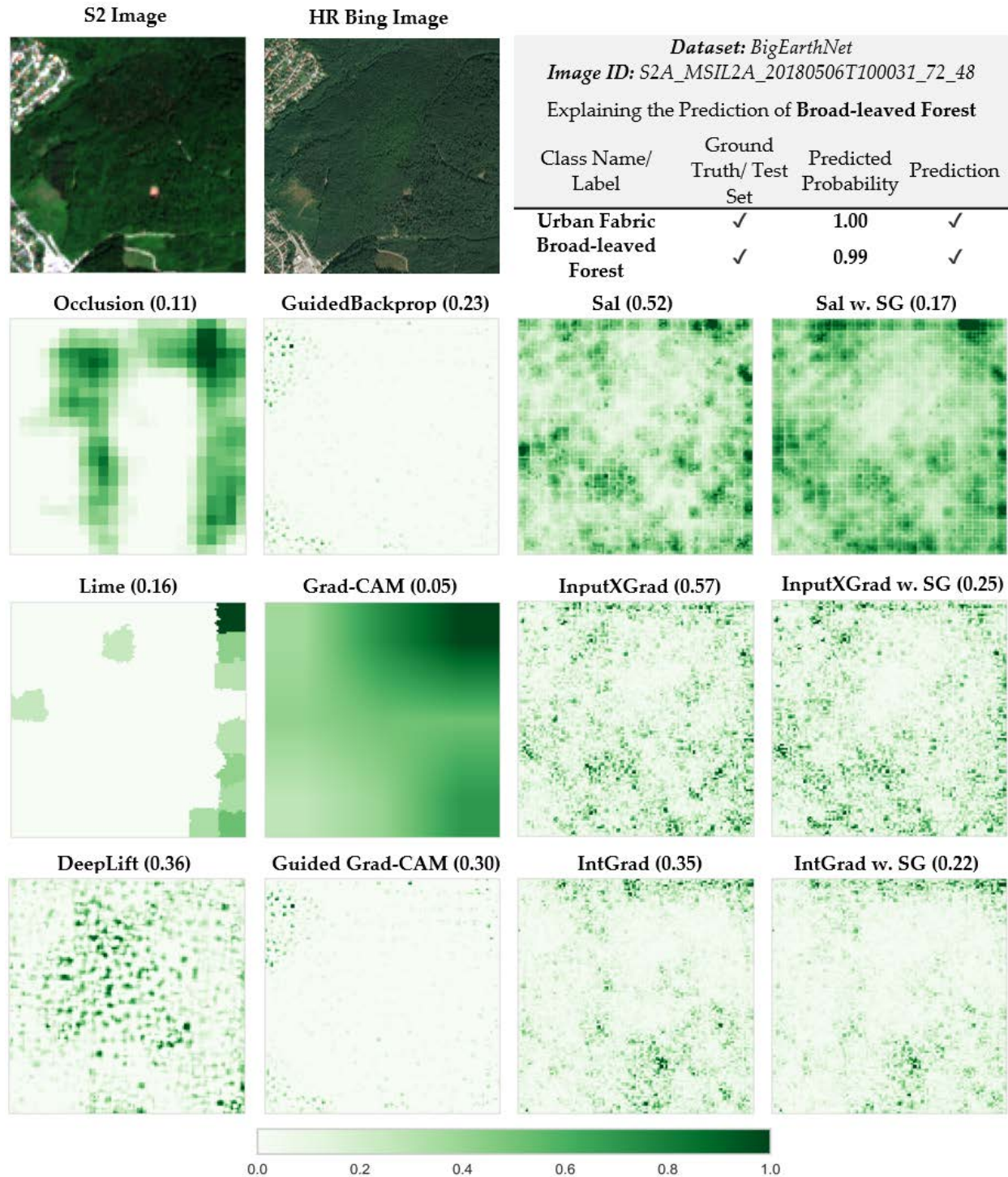


Figure S6b. Explaining the Predictions of DenseNet for *Broad-leaved Forest* class in BigEarthNet dataset (Image ID: S2A_MSIL2A_20180506T100031_72_48). *Max-Sensitivity* score in parenthesis after method's name.



Figure S7a. Explaining the Predictions of DenseNet for *Urban Fabric* in BigEarthNet (Image ID: S2A_MSIL2A_20170613T101031_47_45). *Max-Sensitivity* score in parenthesis after method's name.

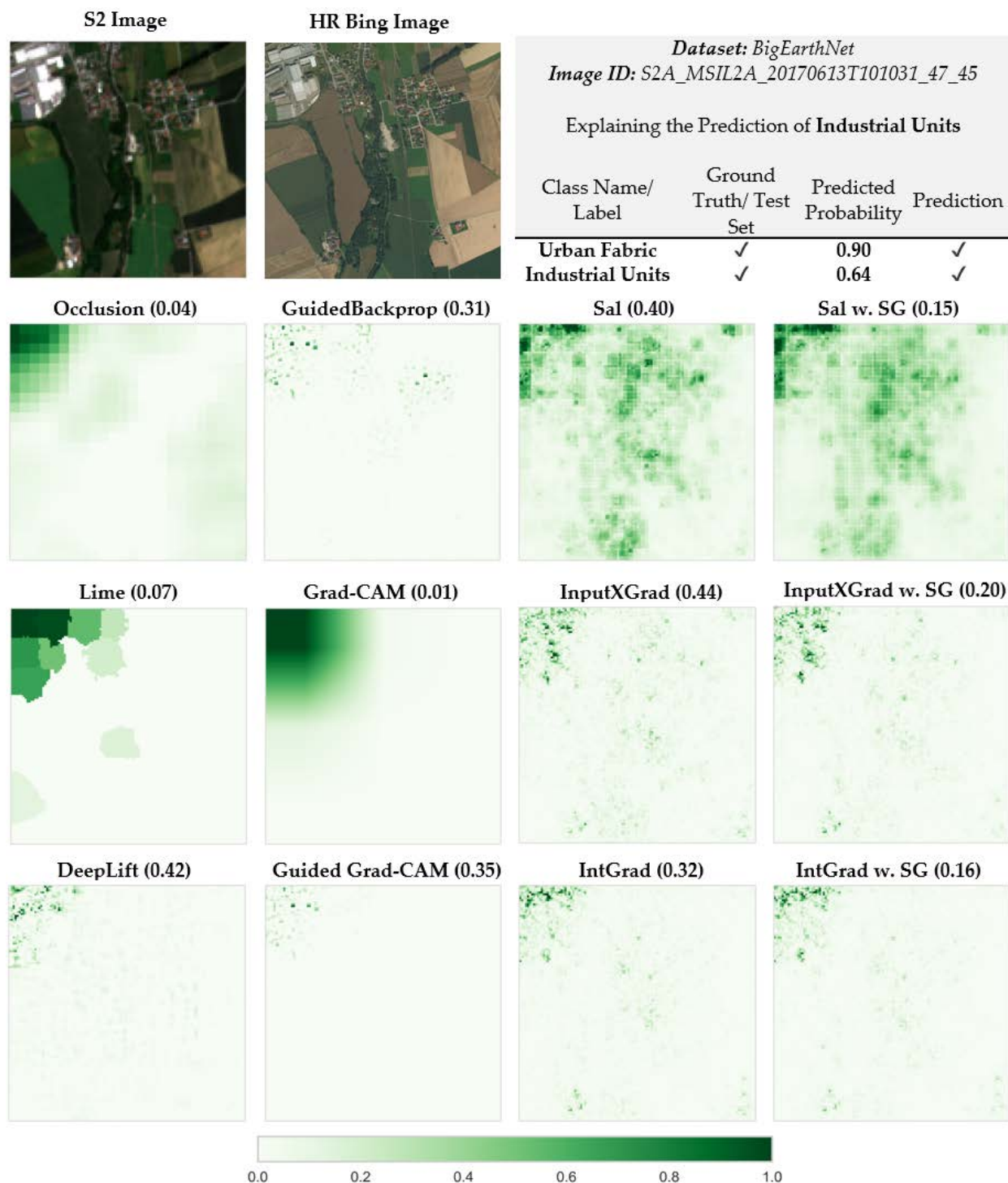


Figure S7b. Explaining the Predictions of DenseNet for *Industrial Units* in BigEarthNet (Image ID: S2A_MSIL2A_20170613T101031_47_45). *Max-Sensitivity* score in parenthesis after method's name.

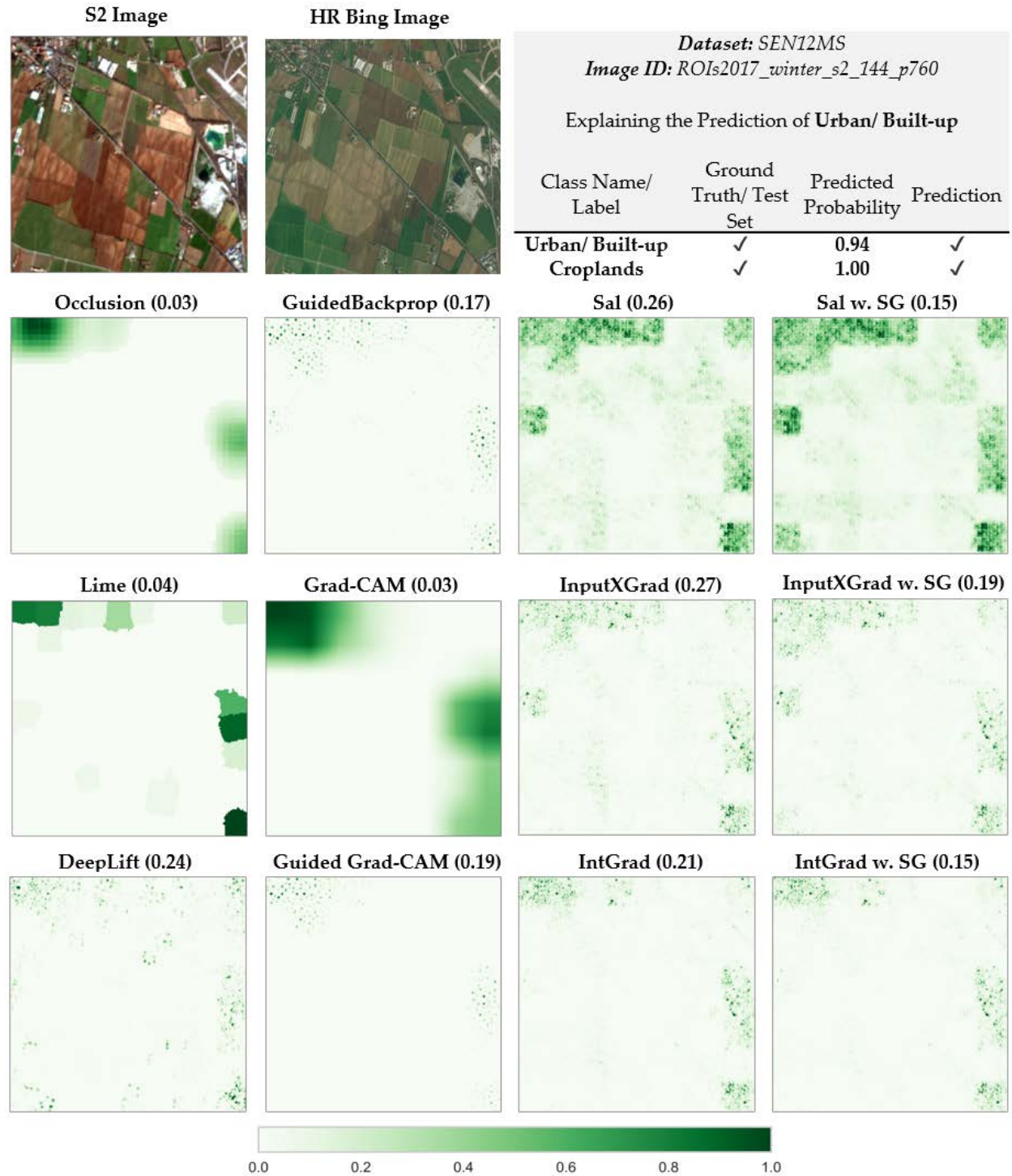


Figure S8a. Explaining the Predictions of DenseNet for *Urban/ Built-up* in SEN12MS (Image ID: ROIs2017_winter_s2_144_p760). *Max-Sensitivity* score in parenthesis after method's name.

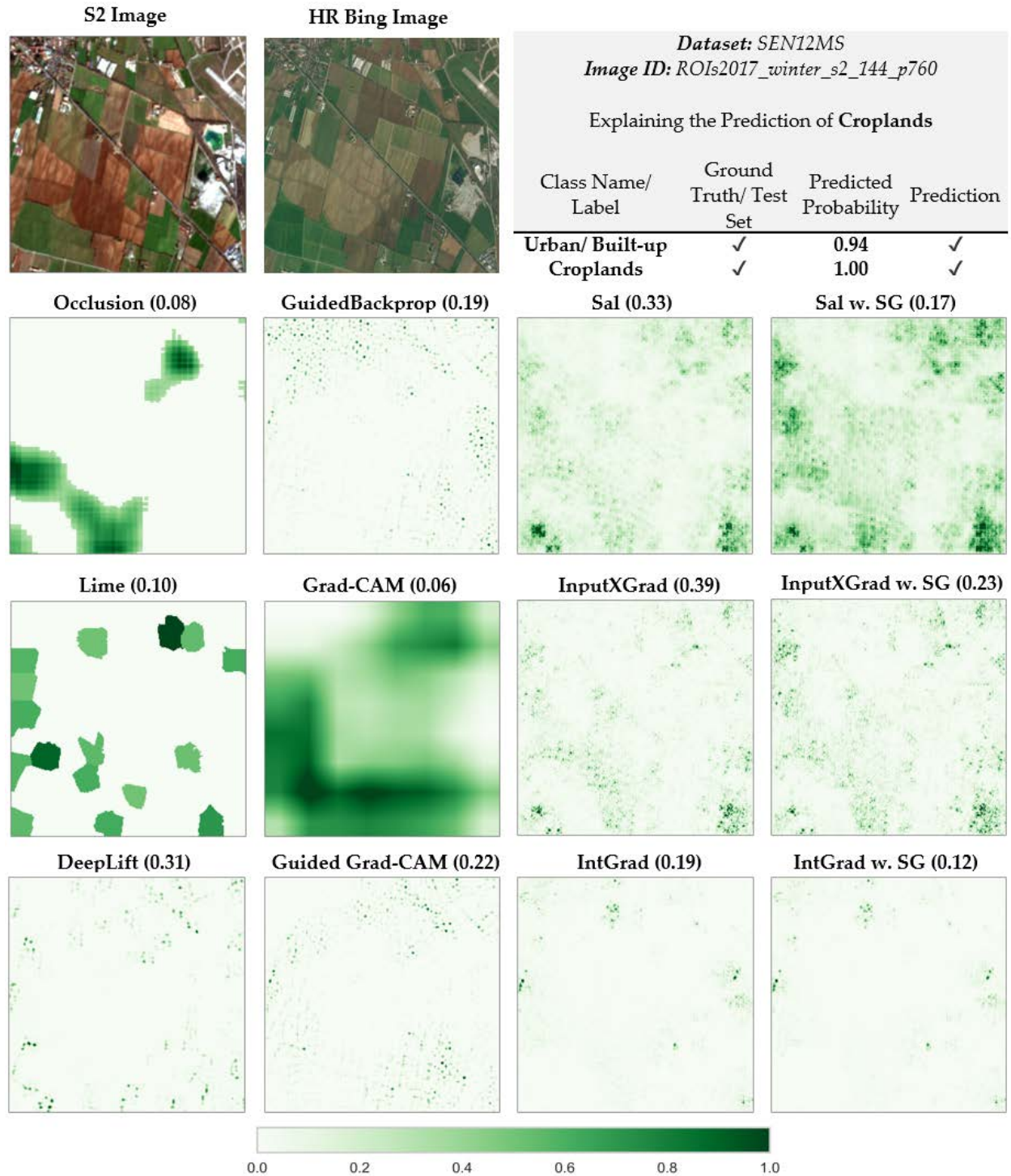


Figure S8b. Explaining the Predictions of DenseNet for *Croplands* in SEN12MS (Image ID: ROIs2017_winter_s2_144_p760). *Max-Sensitivity* score in parenthesis after method's name.

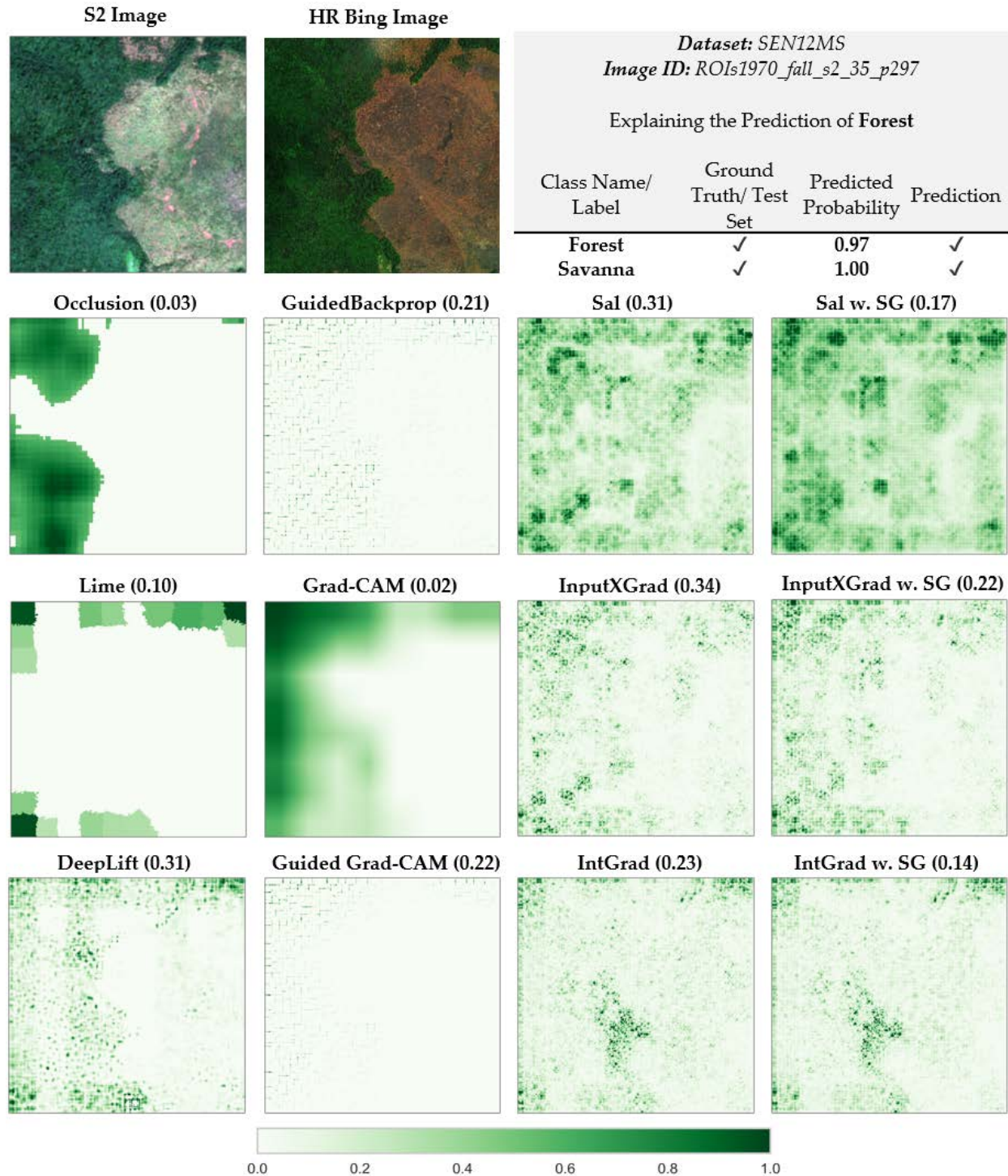


Figure S9a. Explaining the Predictions of DenseNet for *Forest* in SEN12MS (Image ID: ROIs1970_fall_s2_35_p297). *Max-Sensitivity* score in parenthesis after method's name.

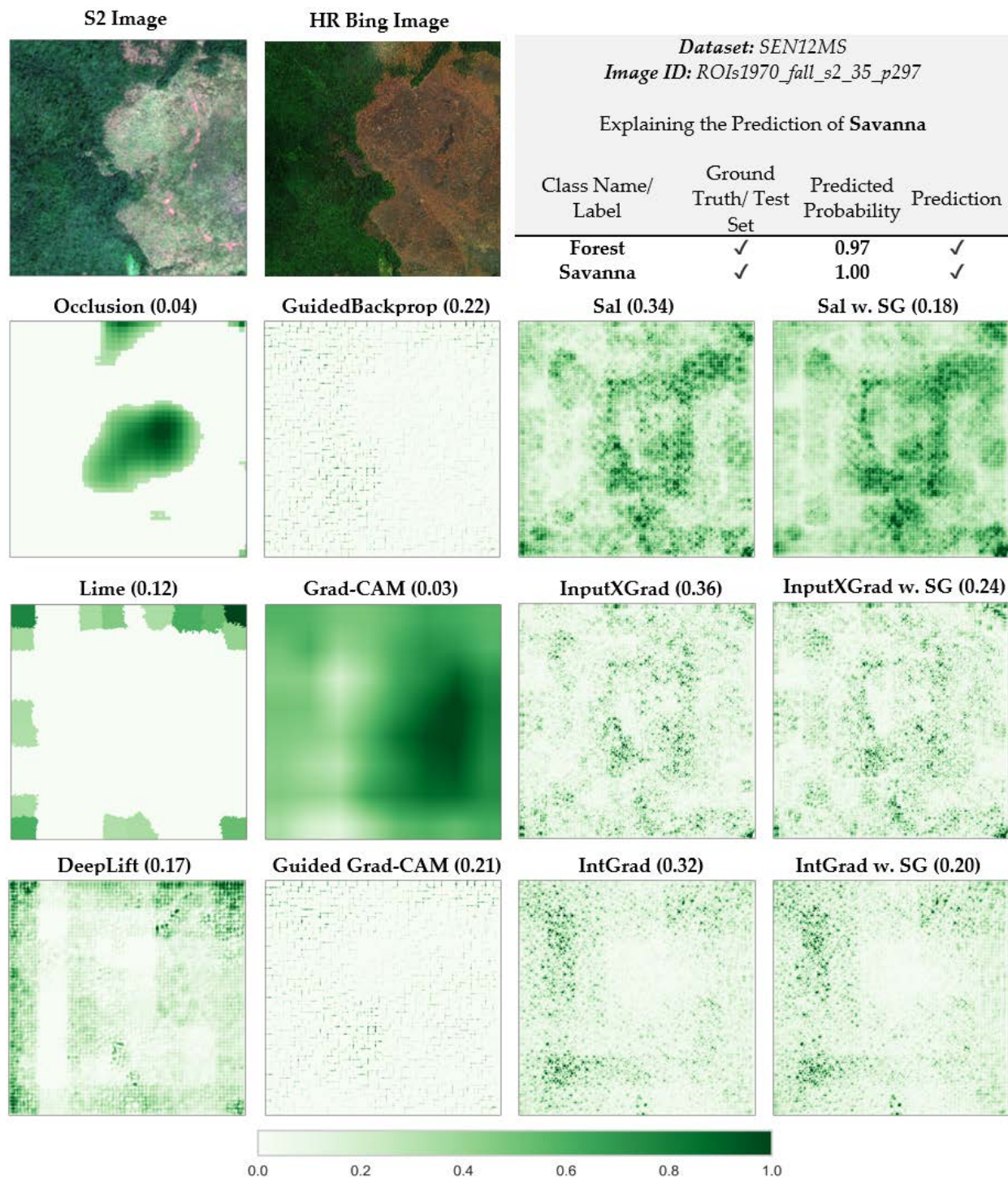


Figure S9b. Explaining the Predictions of DenseNet for *Savanna* in SEN12MS (Image ID: ROIs1970_fall_s2_35_p297). *Max-Sensitivity* score in parenthesis after method's name.

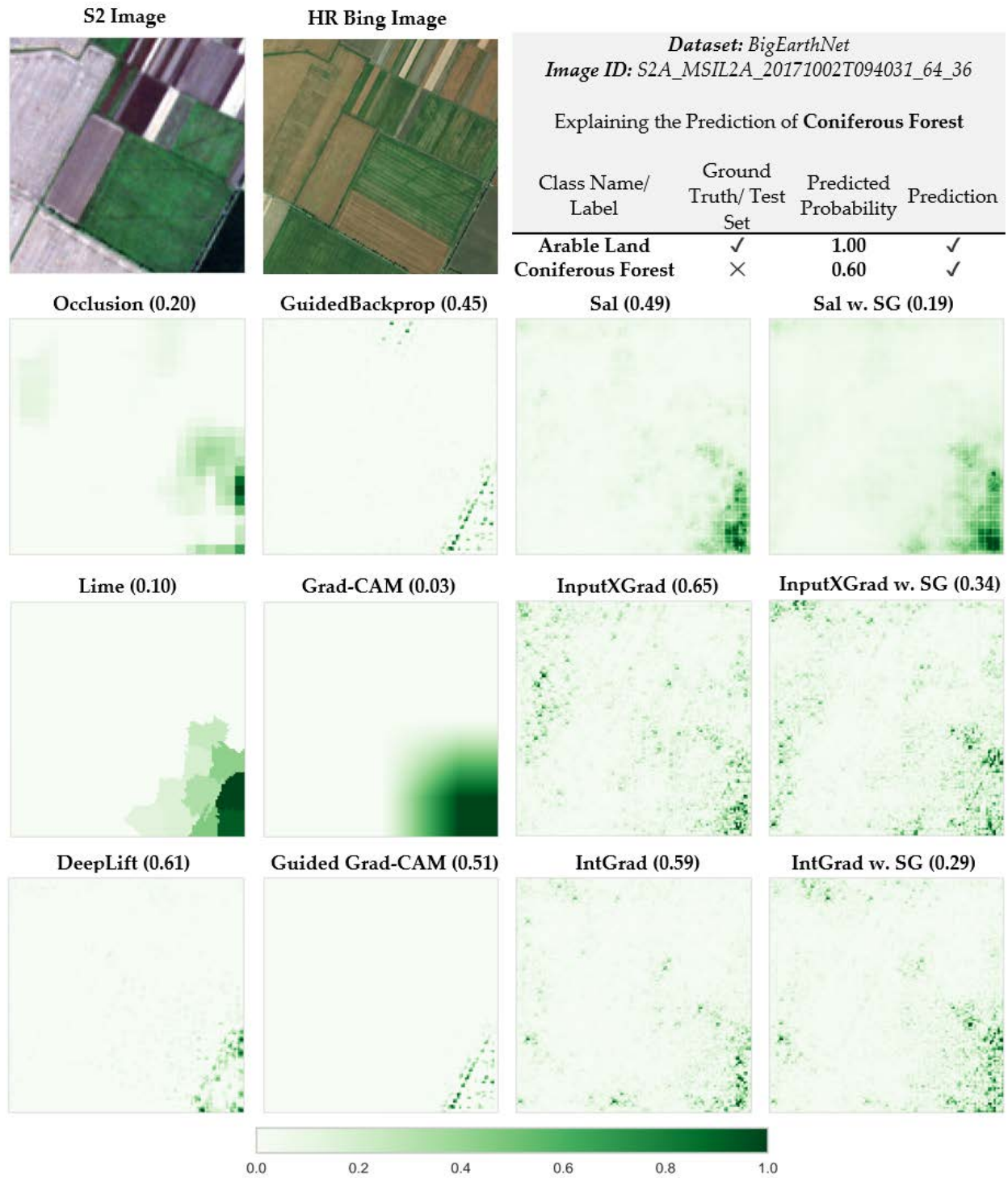


Figure S10. Explaining the Predictions of DenseNet for *Coniferous Forest* in BigEarthNet (Image ID: S2A_MSIL2A_20171002T094031_64_36). *Max-Sensitivity* score in parenthesis after method's name.

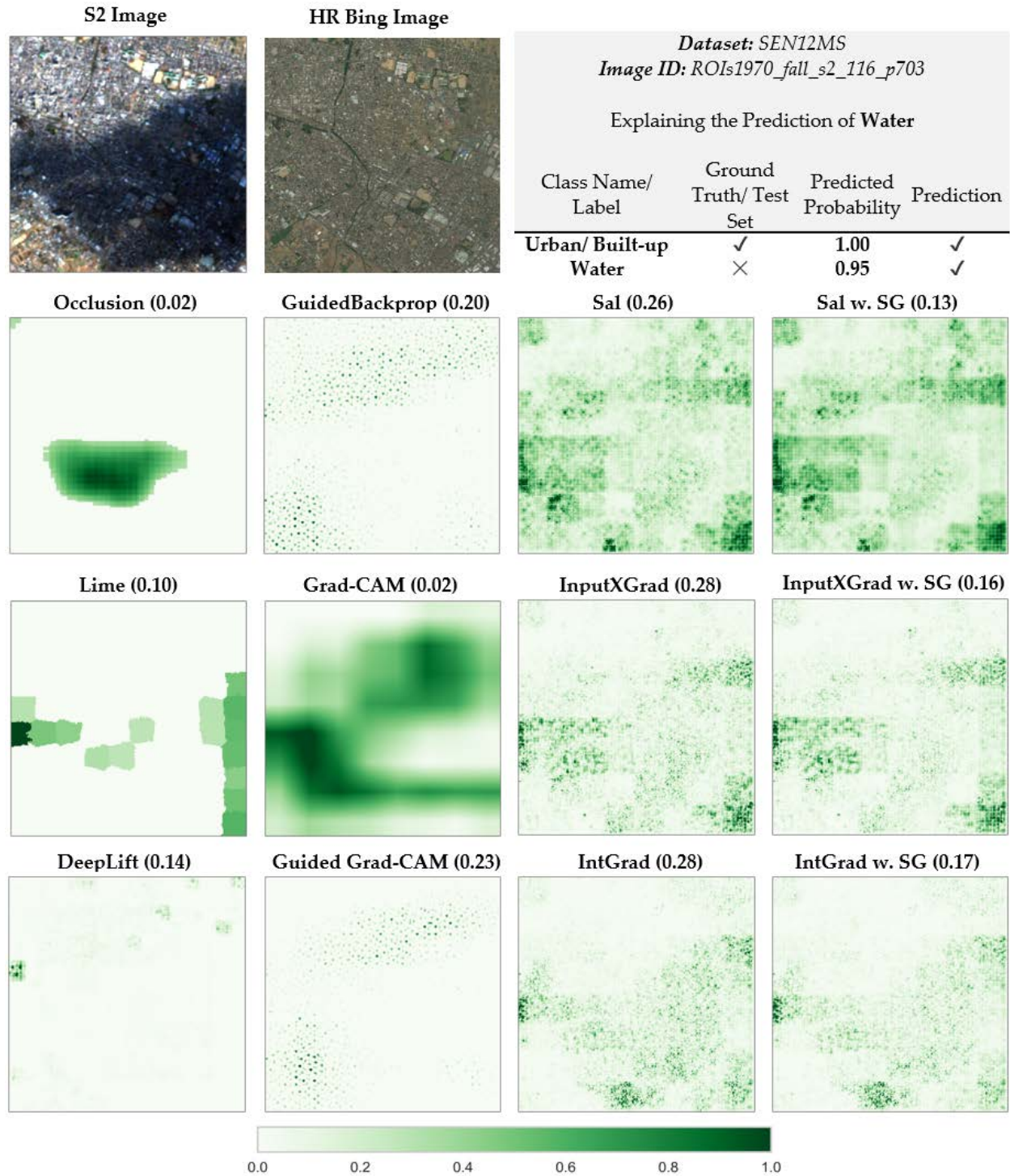


Figure S11. Explaining the Predictions of DenseNet for *Water* in SEN12MS (Image ID: ROIs1970_fall_s2_116_p703). *Max-Sensitivity* score in parenthesis after method's name.

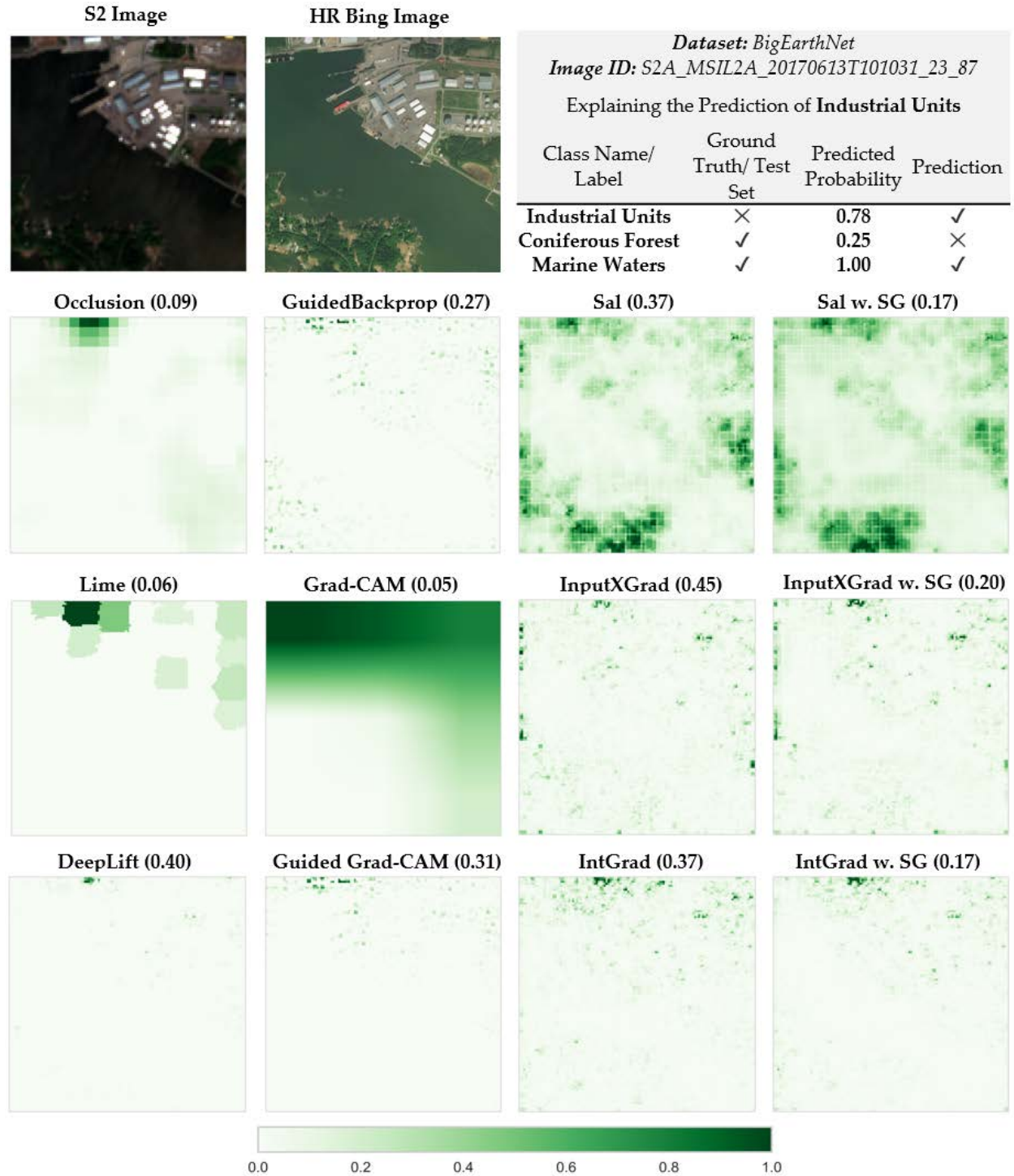


Figure S12. Explaining the Predictions of DenseNet for *Industrial Units* in BigEarthNet (Image ID: S2A_MSIL2A_20170613T101031_23_87). *Max-Sensitivity* score in parenthesis after method's name.

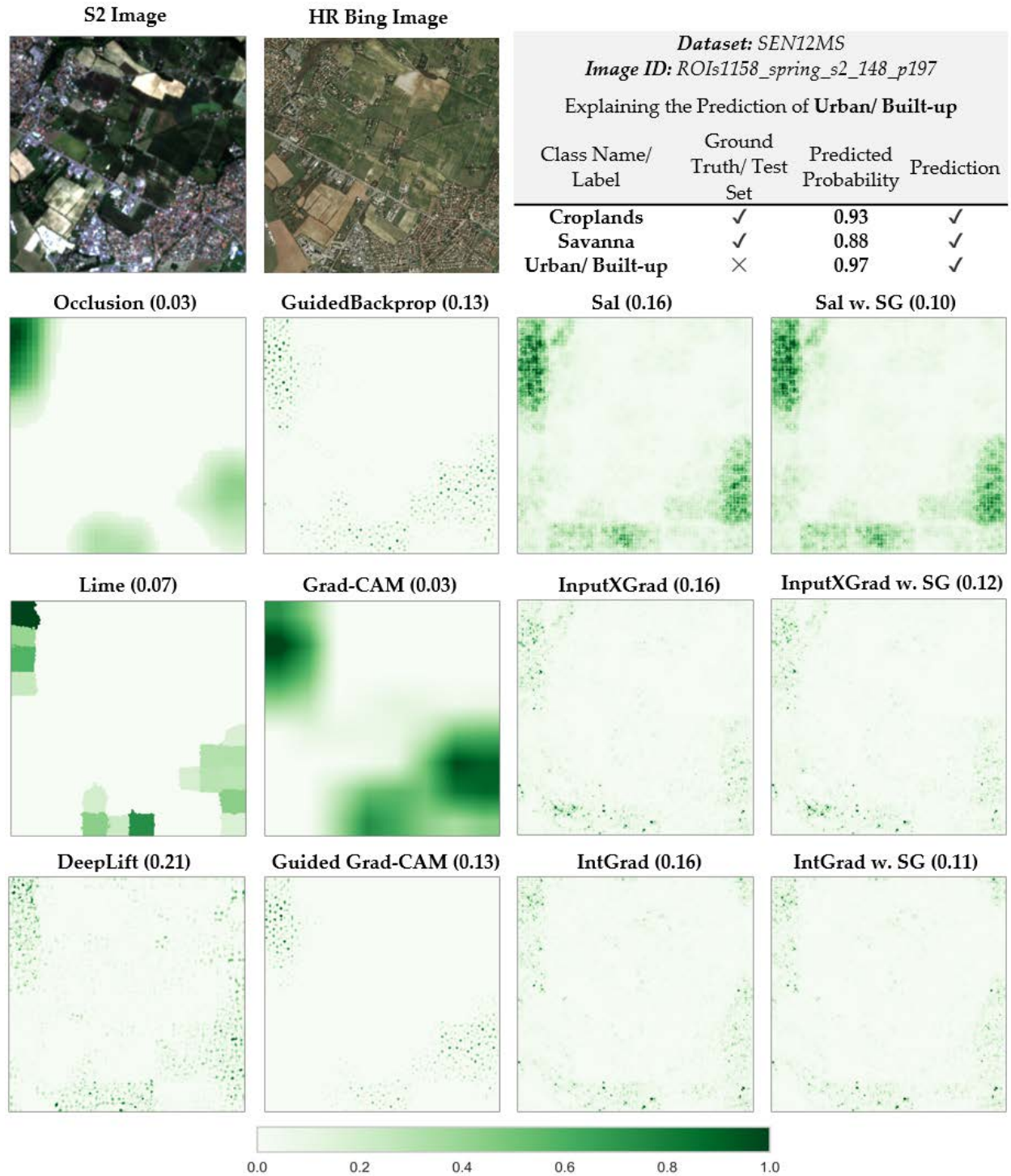


Figure S13. Explaining the Predictions of DenseNet for *Urban/ Built-up* in SEN12MS (Image ID: ROIs1158_spring_s2_148_p197). *Max-Sensitivity* score in parenthesis after method's name.

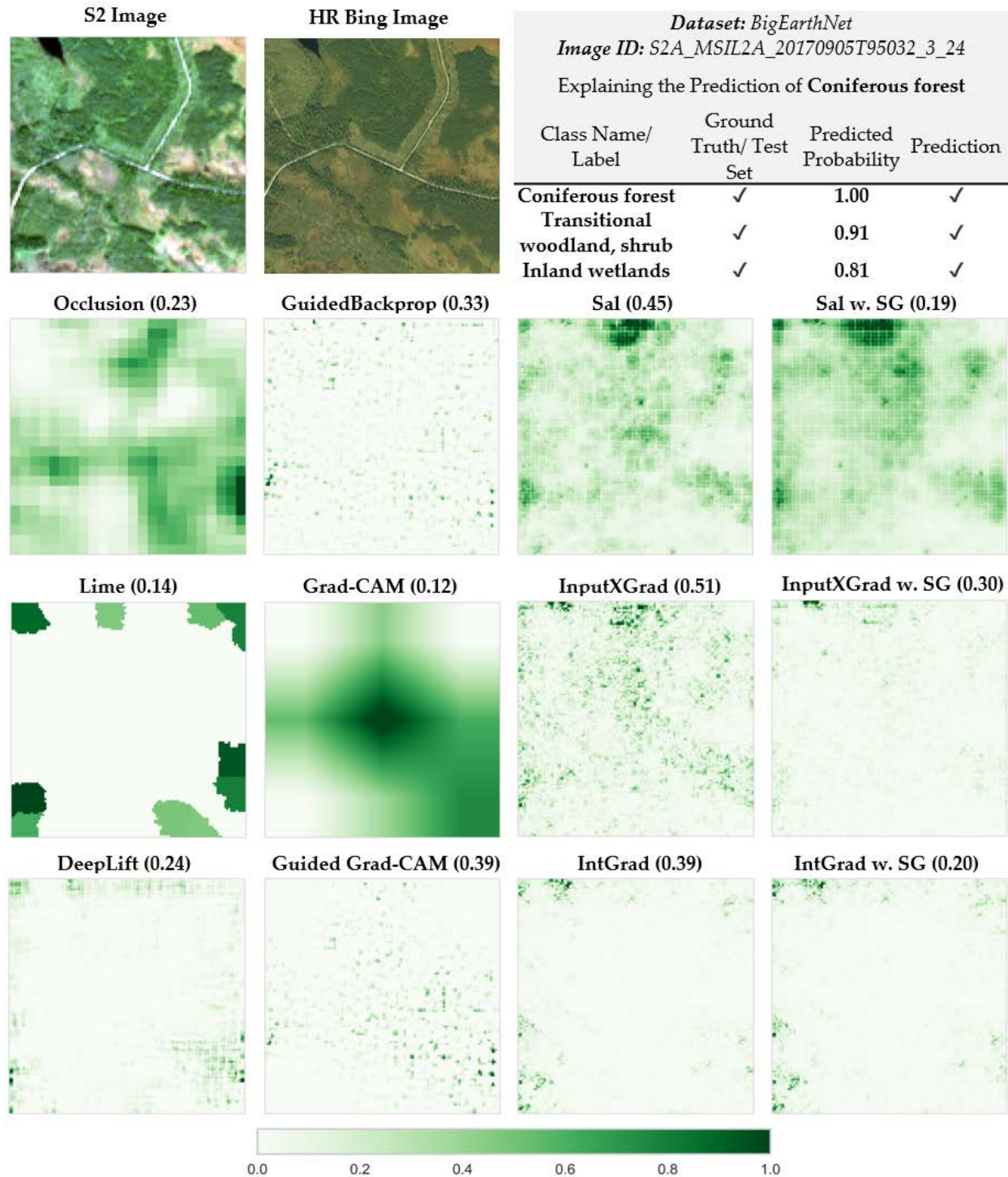


Figure S14. Explaining the Predictions of DenseNet for *Coniferous Forest* in BigEarthNet (Image ID: S2A_MSIL2A_20170905T95032_3_24). *Max-Sensitivity* score in parenthesis after method's name.

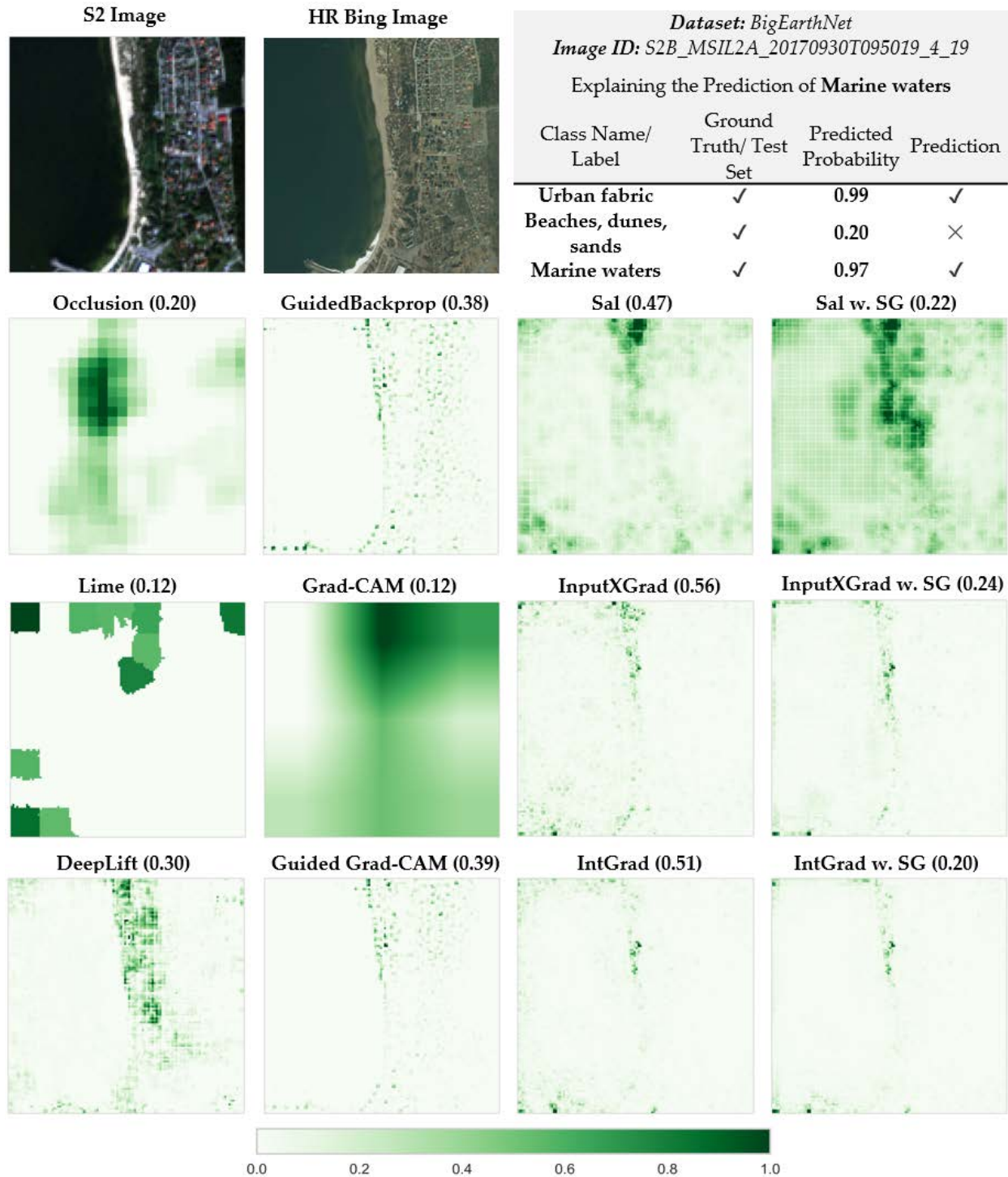


Figure S15. Explaining the Predictions of DenseNet for *Marine Waters* in BigEarthNet (Image ID: S2B_MSIL2A_20170930T095019_4_19). *Max-Sensitivity* score in parenthesis after method's name.



Figure S16. BigEarthNet training set co-occurrence

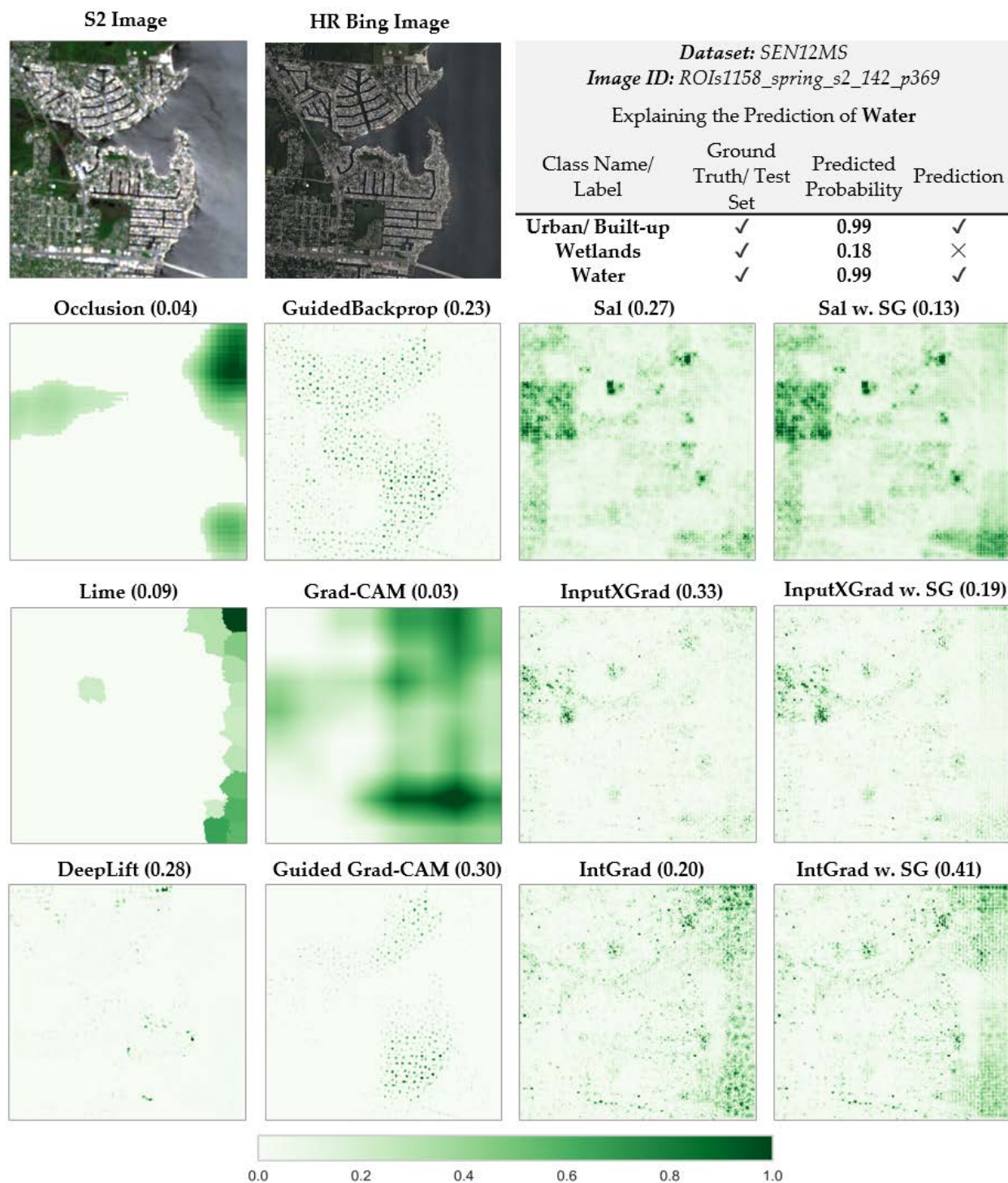


Figure S17. Explaining the Predictions of DenseNet for *Water* in SEN12MS (Image ID: ROIs1158_spring_s2_142_p369). *Max-Sensitivity* score in parenthesis after method's name.