Characterization of out-of-distribution samples from uncertainty maps using supervised machine learning

Lina E. Budde[1 [0000−0001−9545−3018]], Dimitri Bulatov[2 [0000−0002−0560−2591]], Eva Strauss[2 [0000−0003−4834−8092]], Kevin Qiu[2 [0000−0003−1512−4260]], and Dorota Iwaszczuk[1 [0000−0002−5969−8533]]

[1] Technical University of Darmstadt, Civil and Environmental Engineering Sciences Remote Sensing and Image Analysis, 64287 Darmstadt, Germany {lina.budde, dorota.iwaszczuk}@tu-darmstadt.de

[2] Scene Analysis Division, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, 76275 Ettlingen, Germany {dimitri.bulatov, eva.strauss, kevin.qiu}@iosb.fraunhofer.de

Abstract. The quality of land use maps often refers to the data quality, but distributional uncertainty between training and test data must also be considered. In order to address this uncertainty, we follow the strategy to detect out-of-distribution samples using uncertainty maps. Then, we use supervised machine learning to identify those samples. For the investigations, we use an uncertainty metric adapted from depth maps fusion and Monte-Carlo dropout based predicted probabilities. The results show a correlation between out-of-distribution samples, misclassifications and uncertainty. Thus, on the one hand, out-of-distribution samples are identifiable through uncertainty, on the other hand it is difficult to distinguish between misclassification, anomalies and out-of-distribution.

Keywords: Potsdam dataset · error detection · semantic segmentation.

1 Introduction

1.1 Motivation

A major use case of semantic segmentation in remote sensing is the generation of land cover maps. Since these maps serve as a basis for derived products and also political decisions, they must be up-to-date, complete, and trustworthy. While up-to-dateness of the map can be satisfied by the temporal availability of the image data, completeness and trustworthiness mean that the object-related classes correspond to the actual object types in the real world. Hereby, both data and model quality influence the completeness and trustworthiness.

Moreover, completeness in the use of semantic segmentation requires that the world of classes is closed. However, especially while working with high-resolution data from urban scenes, there are out-of-distribution (OOD) samples and anomalies. These samples are usually unified into a particular class, which is called

"clutter", "void", "urban asset", or similar [21,23]. Due to limited reference data, which also may include mislabelings, OOD samples may occur in the test data, but do not belong to any object-specific class in the training samples. Anomalous pixels, in contrast, are present in both, training and test data, but the appearance of anomalous pixels differs from class representing training samples, e.g. image defects.

1.2 Problem statement

Due to the mixing of OOD and anomalies in a joint heterogeneous "clutter" class, this class is often misclassified. However, anomalies and OOD samples are not the only cause of misclassifications. Thus, reasons for misclassifications are among others: (i) incorrect ground truth labels, (ii) anomalies in the images, (iii) OOD samples or (iv) model overfitting. This mix of different sources of misclassifications makes detection and understanding of OOD samples more complex and without specific labeling, the separation of OOD samples is difficult.

Recently, a publication establishing correlation of OOD, anomalous samples and misclassifications based on uncertainty came out [19]. The next logical step would be establishing a workflow allowing for prediction on whether an uncertainty corresponds to an OOD sample, to a misclassification, or to a false alarm.

To do this, it is crucial to understand the two different types of errors. First, not all uncertain predictions are incorrect nor anomalous (false positives, error type 1) and second, not all misclassifications are necessarily uncertain (false negatives, error type 2). To address both types of errors, error detection strategies depend on the task [19]. When undetected misclassification is costly (as in the case of credit card fraud), one must choose a threshold for the false-positive error generously and put more effort into interactive detection of the false-negative errors. Contrarily, if addressing possible misclassifications is more costly, as in surveillance tasks, then too high false alarm rates are undesirable [25]. In addition to uncertainty, other attributes can be helpful in characterizing OOD samples and supporting their identification.

1.3 Our solution in a nutshell

The aim of this work is to characterize OOD samples for identification with different machine learning methods. In addition, two semantic segmentation models are compared in terms of their ability to identify OOD samples. We assume, the OOD samples are part of a "clutter" class in the data. After inference with Monte-Carlo dropout [8], a uncertainty map is computed and thresholded. The threshold value is derived automatically to segment the pixels exhibiting higher uncertainties into connected components. These components have radiometric and geometric features, which, in turn, represent an input for conventional classifiers such as Random Forest [3] or Import Vector Machines [30]. The classifier outputs binary labels for regions to belong to the possible OOD samples and anomalies or to be misclassified.

The remainder of the paper begins with the related work in Section 2, followed by the methodology description in Section 3. Section 4 describes the experimental setup that provides the basis for the results in Section 5. Finally, Section 6 contains conclusions and future work.

2 Related Work

The deep learning methods are very powerful, but they have the dubious reputation to be black boxes taking rather non-transparent decisions. Therefore, in recent years, deep learning models have been increasingly evaluated regarding uncertain predictions. This trend can also be observed in the remote sensing domain [15,4,6]. To identify errors and especially find their sources, observing uncertainty is not enough. According to the current literature, this leads to anomaly detection and detection of OOD samples.

Anomaly detection is more of an unsupervised task because we do not know a-priori what the anomaly is [25]. Therefore, architectures specialized in the anomaly detection task are frequently applied [29]. Examples of unsupervised strategies for retrieving anomalies include mixture-model-based [11], density-based and reconstruction-based approaches [29]. However, semi-supervised or even fully supervised methods are not completely uncommon. For example, [24] investigate a component based on so-called meta-classification in the autonomous driving domain and [27] use Random Forest (RF) to classify thermal anomalies versus shortcomings of a thermal simulator.

The survey of [10] points out that in addition to model and data uncertainties, there is also a distributional uncertainty. This distributional uncertainty is related to OOD samples. To solve the difficulty of distinguishing between in-domain and OOD samples, several authors cited in [10] identify these samples by perturbing the input data [16], or analyzing the softmax probabilities [13], possibly with relaxations of the neural network model, for example using dropouts [4]. To overcome the issues the limited amount of ground truth information and the closed world assumption, the OOD detection becomes important [26,9].

Information about errors and causes, such as OOD samples, are important to improve the quality of land cover maps. However, the results of the semantic segmentation are desired as well [1]. To our knowledge, there is no approach that provides semantic segmentation and at the same time methods to identify the OOD samples in remote sensing. Thus, our contribution contains:

1. dual task pipeline with semantic segmentation and uncertainty mapping,
2. quantitative analysis of the correlation between uncertainty and possible errors and
3. characterization and identification of OOD samples using machine learning.

3 Methodology

By using Monte Carlo dropout, the average softmax outputs of an input image over multiple inferences can be interpreted as probabilities [8]. The resulting

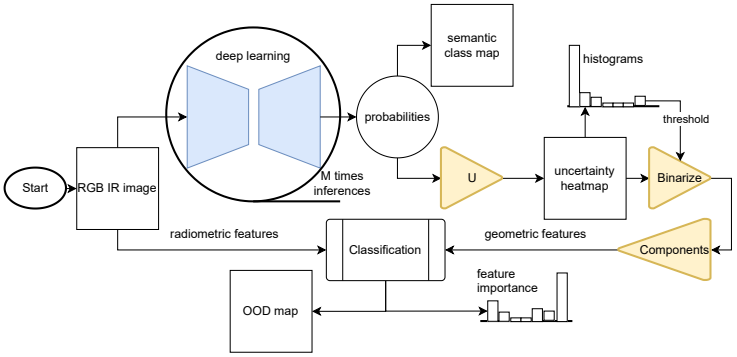

Fig. 1: The flowchart visualizes the processing steps. Based on the probabilities from the deep learning models, the semantic map, and the uncertainty heatmap are generated. By analyzing the histogram distribution, the heatmap is binarized, and the connected components are selected and assigned to a class OOD.

varying inferences can be evaluated in two ways: the variance between the inferences and, from the mean of the inferences, the ambiguities of the class probabilities. For the identification of the OOD regions, we use such ambiguities. Ideally, the maximum probability is near one while all other classes' probabilities are near zero. The class with the maximum probability is selected for the semantic segmentation task. In contrast, if the class probabilities are uniformly distributed, the class decision is subject to the greatest uncertainty. According to Figure 1, we determine the deviations from the ideal case by calculating the related uncertainty (Section 3.1). To characterize and identify the OOD samples, a meta-classification on the component level is performed using Random Forest with geometric and radiometric features (Section 3.3). To use such components for classification, the uncertainties are binarized by the histogram-based threshold (Section 3.2). For evaluation, the labeling described in Section 3.4 allows to detect the errors of the uncertainty and OOD maps.

3.1 Uncertainty Determination

In contrast to using common uncertainty metrics such as entropy [6], [4] propose a confidence metric which was originally developed in the context of depth maps fusion [18] and adapted by us to measure uncertainty [19]. The confidence has a value range from zero to $\frac{1}{C}$ where C denotes the number of classes. By subtracting the confidence values from one, the confidence is interpreted as uncertainty

$$U = 1 - \left(\sum_{c=1}^{C} \exp \left[\frac{-(s - \min\limits_{c \in C}(s))^2}{Q_{0.75} \left[0.5 \cdot (s - \min\limits_{c \in C}(s)) \right]^2} \right] \right)^{-1}, \tag{1}$$

where $Q_{0.75}$ is the 75%-quantile and adjusts the uncertainty values. Further,

$$s = s_{c,h,w} = \min(-\log_2(\overline{P}_{c,h,w}), 2048) \tag{2}$$

is used in (1) using the truncated negative logarithm of the averaged predicted probabilities for each class c, pixel position h and w. The upper bound 2048 avoids numerical problems when using 16-bit values.

3.2 Histogram Analysis

For analyzing the correlation between the uncertainty values from (1) with OOD samples and misclassifications, the corresponding distribution is determined. To calculate the distribution, we represent the uncertainty values as a discrete set $X = \{x_0, x_1, \ldots, x_N\}$, whereby $x_0 = \min(U)$ and $x_N = \max(U)$ with U from 1. For $n \in \{1, \ldots, N\}$, we have $x_n = x_0 + nh$ with bin width h and the number of bins $N + 1$. Let $f, g : x_n \in X \mapsto \mathbb{R}$ be two functions describing a histogram with bin width h, chosen in the way that $x_N = \max(U)$, and the number of bins $N + 1$. Hereby, f represents the distribution of mask pixels which are related to OOD samples or misclassifications and g the complement mask pixels. The bin height is defined from the frequency m as density

$$f(x_n) = \frac{m_n}{\sum_i m_n \cdot h} \tag{3}$$

and for g analogously. We call x_k an intersection of f and g, and it is defined by the change of the sign

$$\begin{gathered} \Leftrightarrow \\ \exists k: \ f(x_{k-1}) \geq g(x_{k-1}) \wedge f(x_k) < g(x_k) \\ \vee \\ \exists k: \ f(x_{k-1}) \leq g(x_{k-1}) \wedge f(x_k) > g(x_k). \end{gathered} \tag{4}$$

Based on [19], the first intersection x_{k^*} is used as a threshold for binarization of the uncertainty.

3.3 Connected Component Analysis

Based on the binary classification with the threshold from (4), median filter and morphological opening is applied to avoid isolated pixels. The resulting uncertainty map is segmented based on connected components. For each component, features are computed and used to characterize the different error sources in a machine-learning framework.

The problem is non-linear and the features correlate strongly. In order to cope with the non-linearity, the RF classifier [3] is known to be quite suitable; it can also cope with correlated features to a certain extent. For example, if one uses a few features for one decision tree and increases the number of decision trees, then the trees are supposed to be less correlated and the overall accuracy

Table 1: If the reference mask and the uncertainty map is unequal this correspond with error type 1 and 2.

	$\neg M$	M
$\neg B$	background	error type 2 (false negative)
B	error type 1 (false positive)	identified

increases. This strategy, theoretically well-founded in [14] and applied, among others, in [27], has been adopted in our work as well. The classification of the components leads to a feature importance, used to characterize the OOD map, which shows the overlap between uncertainty and "clutter" class.

We are motivated to compare RF with a classifier that is able to cope with correlated features. To this end, the Import Vector Machines classifier (IVM) was applied. This classifier was developed by [30] and represents an alternative to Support Vector Machines in the sense that the output is probabilistic and is therefore comparable with that of RF. We used the implementation of [22] with radial-bases functions for kernel trick. The results are shown in Supplementaries.

3.4 Evaluation

For evaluating the distributions similarity, the Kullback-Leibler (KL) divergence is used at the maximum density of the mask class $x_{\max}$ from the histograms in (3) [2]

$$\mathrm{KL} = g(x_{\max}) \cdot \log\left(\frac{g(x_{\max})}{f(x_{\max})}\right) - g(x_{\max}) + f(x_{\max}) \tag{5}$$

$$x_{\max} = \underset{x \in X}{\arg\max}\ f(x). \tag{6}$$

Furthermore, the evaluation of the resulting maps, that means the semantic segmentation, the uncertainty and the OOD maps, use the confusion matrix. This confusion matrix leads to F1-scores and overall accuracy, but also to the amount of occurrences of error type 1 and 2. Let B be a binary uncertainty map based on the threshold from (4) and M a mask used as reference. The results of the confusion matrix are labeled accordingly with Table 1. Furthermore, the Cohen's kappa coefficient is used to measure the correlation between detected OOD samples and clutter.

4 Experimental Setup

The presented methodology is tested on the ISPRS Potsdam dataset [23]. Before using this data for deep learning, the sample images are prepared as in [19].

4.1 Dataset Preparation

First, the dataset is split into training, validation, and test data. The multi-spectral images with channels red, green, blue, and near-infrared are used. However, to process images of size 6000 × 6000 pixels, they are divided into small patches of 512 × 512 pixels. The test data are tiled by an overlap of 50 pixels. In the reference, six classes are defined: impervious surface, building, low vegetation, tree, car, and clutter. In this contribution we are mainly interested in the latter class which contains the OOD samples. In total, 818 training and 243 validation patches are randomly selected. For testing, a total number of 1350 patches is used.

4.2 Deep Learning Models

For the experiments, we use a U-Net [20] and DeepLab V3+ [5] model, which is referred to as DeepLab in this work. For both models, a ResNet 101 encoder [12] with ImageNet [7] initialization is used. To be able to use multi-spectral images with more than three channels, the first convolution layer is changed accordingly. For training, the AdamW optimizer [17] and the cross entropy loss function are used. The hyperparameter settings are a learning rate of 0.001, a weight decay of 0.01, a batch size of 9, a dropout rate of 0.5, and 300 epochs. The SMP Toolbox [28] is used for the model construction. Data augmentation is applied with 90°, 180° and 270° rotations. During testing, we evaluate the models with 100 Monte-Carlo inferences.

4.3 Evaluation Strategy

For each model, two different masks are used to generate the histogram distributions: the clutter pixels extracted from the ground truth and the model predictions compared with the ground truth, representing misclassifications. In addition, we select 75% of our connected components as training data and the rest as test data. For feature extraction, we use the multi-spectral image channels and entropy features with four different kernel sizes as radiometric features. The mean value and variance over all pixels lying in this component are calculated. To these radiometric features, we add two geometric properties of the connected component, namely area and eccentricity, yielding 18 features in total. Training data is further balanced in order for the learner not to overfit towards the most frequent class. The ratio between the number of training examples of the most and least frequent class was set to 1, 1.5, 2.5, and 4. The components are evaluated with RF using 10, 25, and 50 trees.

5 Results and Discussion

5.1 Semantic Segmentation

For the semantic segmentation task, both models provide similar quantitative results (Table 2). Looking closely at the differences between U-Net and DeepLab

Table 2: The quality of the semantic segmentation is evaluated by the overall accuracy and the F1 values of the test dataset using the U-Net and DeepLab model. In addition, the class frequency from the ground truth is presented. The results are given as a percentage value.

model	overall accuracy	impervious surfaces	building	low vegetation	tree	car	clutter
ground truth frequency		38.4	24.9	16.5	12.3	2.2	5.6
U-Net	86.2	90.4	**92.9**	80.6	**81.1**	**89.8**	48.8
DeepLab	**86.5**	**90.5**	92.8	**81.5**	80.9	88.2	**53.1**

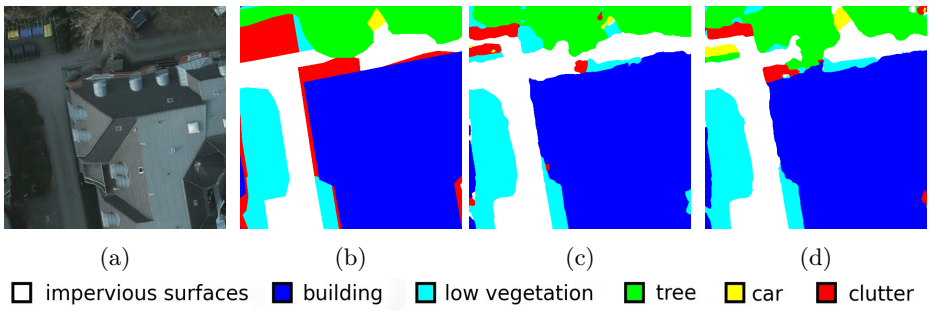

Fig. 2: Visualization of the semantic segmentation: (a) RGB input, (b) ground truth, (c) prediction from U-Net, and (d) prediction from DeepLab.

for the different classes, U-Net performs better (+1.6 %) for the car class, which is the least frequent class and can be considered as a small object class with high-contrast and high in-class variation. In contrast, in the clutter class (red pixels), DeepLab outperforms U-Net by 4 %. The differences in the other classes are below 1 %. Looking at the clutter class in the example displayed in Figure 2, anomalies at the building edges represented by facade pixels could not well predicted (Figure 2c, 2d) compared to the the ground truth (Figure 2b). Although the F1 score for clutter is better for DeepLab, U-Net predicts the garbage cans better in this example. This shows the difficulty of prediction of OOD samples and anomalies. For this reason, the following analyses focus on this clutter class which contains the OOD samples.

5.2 Uncertainty Analysis

As mentioned in Section 3, the mean value can be calculated from the standard deviations per class from several inferences. In our results, this mean is 0.011 for DeepLab and 0.005 for U-Net. Therefore, the predicted probabilities can be considered reliable.

The class prediction, in contrast, has ambiguities represented by the uncertainties from (1). The relationship between uncertainty and clutter is displayed by the density distributions in Figure 3. The highest uncertainty values corre-

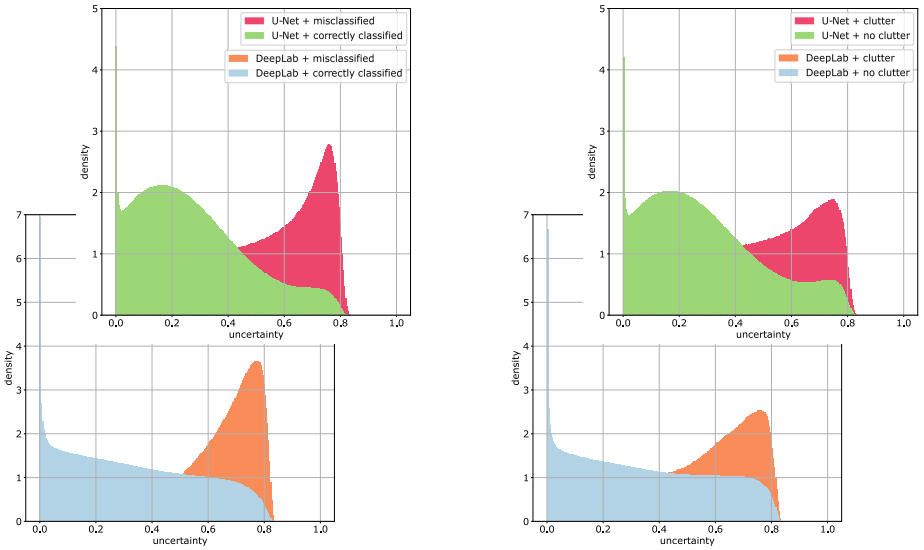

Fig. 3: The distributions of the U-Net and DeepLab uncertainties with the misclassification mask are presented on the left and with the clutter mask as in [19] on the right. The imbalance of the data is considered by calculating density.

Table 3: The KL divergence measures the similarity of the uncertainty distributions for U-Net and DeepLab. The intersection point at the misclassification mask is used as a threshold for the connected component formation.

model	KL clutter mask	KL misclassification mask	threshold
U-Net	0.95	3.19	0.434
DeepLab	1.01	3.62	0.508

late with clutter. However, the same correlation occurs between uncertainty and misclassifications. Thus, clutter and misclassification are also correlated. Even if the peak density of uncertainty for misclassified pixels is higher compared with the clutter. For both, clutter and misclassification masks, the maximum density increases when using DeepLab. The KL divergences between the distributions are given in Table 3. Although the density increases, the lower divergence indicates a more similar distribution of clutter versus no-clutter and correctly versus misclassified pixels for U-Net. Independent of the model, data related misclassifications are often characterized by high uncertainty. However, the source of the misclassification, i.e. wrong ground truth label, OOD sample, or anomaly, can hardly be distinguished based on the distribution.

However, in the example in Figure 4a and 4c the highest uncertainties occur in the predicted clutter area and the object borders. By the threshold values from Table 3 and the labeling principle from Table 1 the uncertainty maps are evaluated. Thus, Figure 4b and 4d visualizes the errors by using only the uncertainty

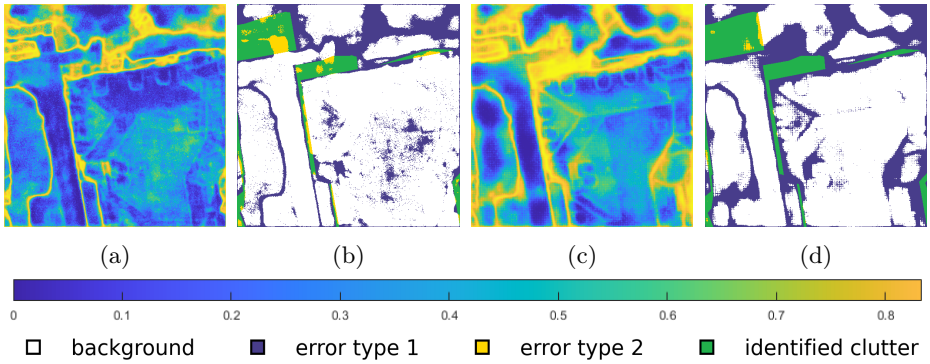

Fig. 4: The uncertainty values for U-Net (a) and DeepLab (c) are evaluated with clutter mask (b) and (d), respectively. Type 1 (blue) and type 2 (yellow) errors occur in both models. The green regions represent the area where clutter appears together with high uncertainty.

Table 4: Proportion of errors represented as percentage of the uncertainty maps. Differences to the sum of 100 is caused by rounding.

mask	model	background	error type 1	error type 2	identified
misclassification	U-Net	71	19	4	7
misclassification	DeepLab	66	24	3	8
clutter	U-Net	72	22	3	3
clutter	DeepLab	66	28	2	3

values for the OOD identification. Applying the clutter mask, all other classes are assigned to the background. Due to high uncertainty, areas with misleading clutter regions (error type 1, blue) exist. In contrast, low uncertainties leads to missing clutter identification (error type 2, yellow). The correctly identified clutter regions (green) have higher completeness in DeepLab in this example. However, with DeepLab, larger contiguous error type 1 regions form.

Evaluated quantitatively, 3 % of the clutter pixels cannot be identified via their uncertainty with U-Net (error type 2 in Table 4). With DeepLab, the proportion of error type 2 pixels decreases at the expense of the amount of error type 1 and correct background class. In contrast to error type 2, error type 1 is very high with a proportion of up to 28 %. Thus, to reduce the type 1 error and analyze the characteristics of the misclassified and clutter pixels in more detail, the connected component analysis is used.

5.3 Connected Component Analysis

The results from the component-based classification in Table 5 show that U-Net and DeepLab models correlate moderately with the reference masks in the assessment of clutter pixels and quite low in the assessment of the incorrect pix-

Table 5: Selection of results from the meta-classification. We specify the number of trees in Random Forest classifier. By OA and κ, we denote Overall Accuracy and Cohen's kappa coefficient, respectively. All numbers are given in percentages. For the results of all configurations, including import vector machines, see Supplementary Material (Table S1).

mask	model	metric	balancing factor 1			balancing 1.5			balancing 2.5		
			10 Trees	25 T	50 T	10 T	25 T	50 T	10 T	25 T	50 T
clutter	DeepLab	OA	66.4	66.1	66.1	73.6	74.3	74.8	81.8	83.1	83.8
		κ	14.9	16.0	16.7	17.9	20.8	**21.2**	16.8	18.9	20.4
clutter	U-Net	OA	67.1	66.8	67.1	74.9	75.5	75.9	82.4	82.8	83.3
		κ	18.5	20.6	21.2	19.7	22.0	22.8	22.1	23.2	**23.8**
misclassification	DeepLab	OA	59.3	60.5	62.0	65.5	66.3	67.6	67.8	69.5	70.0
		κ	8.5	**11.3**	13.4	8.9	8.6	9.2	5.5	6.4	5.3
misclassification	U-Net	OA	62.7	63.9	64.5	68.1	69.9	70.5	69.2	70.5	71.1
		κ	12.6	16.6	**17.0**	11.3	14.4	14.3	8.8	9.3	9.0

els. The results of U-Net are always slightly higher than that of DeepLab. One problem of lower performance of DeepLab is related to the connected components. Due to the larger size of the components, they include both clutter and no-clutter pixels, which makes a clear assignment difficult.

In addition, we are dealing with extremely unbalanced data: the highest overall accuracies can be achieved by simply assigning all connected components to the "normal" class, which sets κ to zero. Not mentioned in Table 5 is the dependency on the minimum leaf size parameter of RF. Smaller minimum leaf size tends to overfit for the small number of trees while a larger value makes RF more independent of the tree number.

The features used are strongly correlated, which not only exacerbates the overfitting, but also makes it more difficult to assess their importance. Nevertheless, the feature importance values of RF in Figure 5 show more detailed characteristics of the uncertain regions compared to deep learning results. The variance inside each component is less important than the mean values. The top three features using 50 Trees and the U-Net results are the mean blue, red and near infrared compared to area, mean red and the first mean entropy feature using DeepLab. However, the importance values for all features except eccentricity vary in a range of 0.9 to 2.0 and are thus similar in importance.

Our example (Figure 6) shows our final predicted clutter map, which contains OOD samples and anomalies. Comparing this predicted map with the RGB input and the corresponding ground truth, this map detected more anomalies correctly as labeled in the ground truth as clutter, for example the distorted tree pixels. However, with the used features the separation between OOD samples and anomalies could not reached.

There are some errors of type 2 that we leave unchecked in this work, but mostly they are close to the correctly identified components. Overall, DeepLab

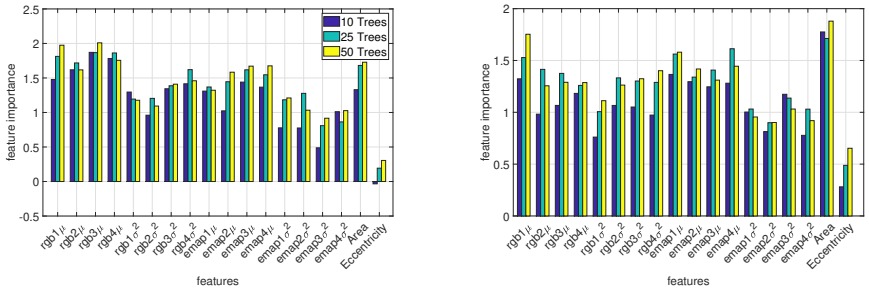

Fig. 5: Feature importance of the 18 different component-based features using RF for U-Net (left) and DeepLab (right) for identification of clutter.

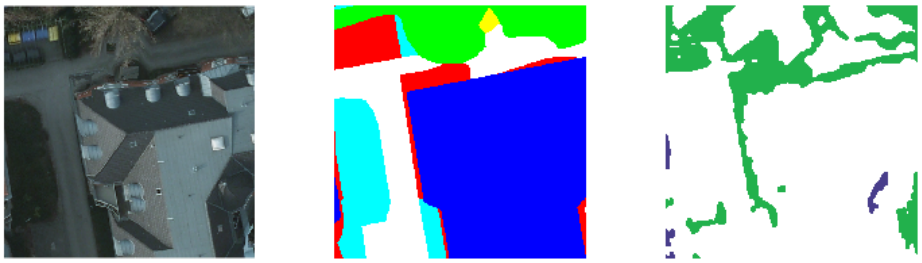

Fig. 6: The RGB, the ground truth, and the predicted clutter map are presented from left to right. The clutter map contains the desired OOD samples as well as anomalies. The green color represents the correctly identified components, the blue components shows the errors of the RF classifier. The true negative and all not used image parts are white background.

results tend to produce quite large components, which are not of much help while the results of U-Net seem more concordant with the anomalous regions.

6 Conclusion and Future Work

We presented a contribution on detection and analysis of OOD samples. Two goals pursued here were to establish a correlation between the uncertain and OOD samples within a workflow for semantic segmentation as well as their identification using connected component analysis and shallow learning. The first goal could be achieved with the applied deep learning models and the generated uncertainty maps. Significant values for KL divergence in the distributions confirm the correlation between the classification uncertainty and OOD samples and anomalies, represented by clutter pixels. For the second goal, we analyzed the uncertain regions with methods of machine learning. We considered low-level image-based features and Random Forest as conventional classifier. At the current stage of our research, moderate correlations could be established since the

values of Cohen's kappa have not exceeded 0.25 for clutter pixels and 0.2 for misclassifications.

To improve the detection in future work, the dependency on very big and inhomogeneously filled with clutter or misclassified pixels should addressed. Thus, applying clustering methods such as super-pixel could generate more categorical components. Further, evaluation of multi-modal data is becoming increasingly popular in remote sensing; thus extending the classifier with additional features derived from 3D data may increase the accuracy. In this paper, we concentrated on the type 1 errors. However, the type 2 errors, that is, possible misclassification despite allegedly certain regions, are equally important and should be tackled in the future. Finally, evaluation on further dataset would provide some clues on the generalizability of the proposed methods.

Acknowledgment

We would like to thank Timo Kullmann for his help during the implementations of the neural networks.

References

1. Bevandić, P., Krešo, I., Oršić, M., Šegvić, S.: Simultaneous semantic segmentation and outlier detection in presence of domain shift. In: Fink, G.A., Frintrop, S., Jiang, X. (eds.) Pattern Recognition, Lecture Notes in Computer Science, vol. 11824, pp. 33–47. Springer International Publishing, Cham (2019)
2. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004). https://doi.org/10.1017/CBO9780511804441
3. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
4. Budde, L.E., Bulatov, D., Iwaszczuk, D.: Identification of misclassified pixels in semantic segmentation with uncertainty evaluation. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XLIII-B2-2021**, 441–448 (2021). https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-441-2021, https://isprs-archives.copernicus.org/articles/XLIII-B2-2021/441/2021/
5. Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. CoRR **abs/1802.02611** (2018), http://arxiv.org/abs/1802.02611
6. Dechesne, C., Lassalle, P., Lefèvre, S.: Bayesian deep learning with monte carlo dropout for qualification of semantic segmentation. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. pp. 2536–2539 (2021). https://doi.org/10.1109/IGARSS47720.2021.9555043
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer vision and Pattern Recognition. pp. 248–255. IEEE (2009)
8. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. ArXiv **abs/1506.02142** (2015)
9. Gawlikowski, J., Saha, S., Kruspe, A., , Zhu, X.X.: Towards out-of-distribution detection for remote sensing. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. pp. 8676–8679 (2021). https://doi.org/10.1109/IGARSS47720.2021.9553266

10. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A.M., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X.: A survey of uncertainty in deep neural networks. CoRR **abs/2107.03342** (2021), https://arxiv.org/abs/2107.03342
11. Hazel, G.G.: Multivariate gaussian mrf for multispectral scene segmentation and anomaly detection. IEEE transactions on geoscience and remote sensing **38**(3), 1199–1211 (2000)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
13. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
14. James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning, vol. 112. Springer (2013)
15. Kampffmeyer, M., Salberg, A.B., Jenssen, R.: Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 680–688 (2016). https://doi.org/10.1109/CVPRW.2016.90
16. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690 (2017)
17. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. CoRR **abs/1711.05101** (2017), http://arxiv.org/abs/1711.05101
18. Pollefeys, M., Nistér, D., Frahm, J.M., Akbarzadeh, A., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Kim, S.J., Merrell, P., Salmi, C., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewénius, H., Yang, R., Welch, G., Towles, H.: Detailed Real-Time Urban 3D Reconstruction from Video. International Journal of Computer Vision **78**(2-3), 143–167 (2008). https://doi.org/10.1007/s11263-007-0086-4, dx.doi.org/10.1007/s11263-007-0086-4
19. Qiu, K., Bulatov, D., Budde, L.E., Kullmann, T., Iwaszczuk, D.: Influence of out-of-distribution examples on the quality of semantic segmentation in remote sensing. In: IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium. Pasadena (July 2023)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28, https://arxiv.org/abs/1505.04597v1
21. Roscher, R., Volpi, M., Mallet, C., Drees, L., Wegner, J.D.: Semcity toulouse: A benchmark for building instance segmentation in satellite images. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **V-5-2020**, 109–116 (2020)
22. Roscher, R., Waske, B., Forstner, W.: Incremental import vector machines for classifying hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing **50**(9), 3463–3473 (2012)
23. Rottensteiner, F., Sohn, G., Gerke, M., , Wegner, J., Breitkopf, U., Jung, J.: Results of the ISPRS benchmark on urban object detection and 3D building reconstruction. ISPRS Journal of Photogrammetry and Remote Sensing **93**, 256–271 (07 2014). https://doi.org/10.1016/j.isprsjprs.2013.10.004
24. Rottmann, M., Maag, K., Chan, R., Hüger, F., Schlicht, P., Gottschalk, H.: Detection of false positive and false negative samples in semantic segmentation (08122019), http://arxiv.org/pdf/1912.03673v1

25. Ruff, L., Kauffmann, J.R., Vandermeulen, R.A., Montavon, G., Samek, W., Kloft, M., Dietterich, T.G., Müller, K.R.: A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE **109**(5), 756–795 (2021)
26. da Silva, C.C.V., Nogueira, K., Oliveira, H.N., dos Santos, J.A.: Towards open-set semantic segmentation of aerial images. CoRR **abs/2001.10063** (2020), https://arxiv.org/abs/2001.10063
27. Strauß, E., Bulatov, D.: A region-based machine learning approach for self-diagnosis of a 4d digital thermal twin. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences **10** (2022)
28. Yakubovskiy, P.: Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch (2020)
29. Yuan, S., Wu, X.: Trustworthy anomaly detection: A survey (16022022), http://arxiv.org/pdf/2202.07787v1
30. Zhu, J., Hastie, T.: Kernel logistic regression and the import vector machine. Journal of Computational and Graphical Statistics **14**(1), 185–205 (2005)