Effectiveness of Satellite Imagery Cloud Removal using Multi-Modal Deep Learning in Downstream Applications

Nick Hauptvogel (1010801624)

Index Terms—Satellite Imagery, Cloud Removal, Deep Learning, Land Cover Classification, Remote Sensing

Abstract—Recent advances in Deep Learning enable explicit, multi-modal cloud removal from satellite imagery using synthetic aperture radar (SAR) inputs as auxiliary data source. However, to fully assess the usability of those predictions, they must prove to be applicable in downstream applications as well. Using land cover classification as an example, we explore the efficacy of an explicit cloud removal step in a satellite image processing pipeline by testing the predictive performance on a classification network, which was trained on corresponding cloud-free data. Our findings confirm that cloud-removed data is performing better than using out-of-distribution cloudy images during inference, but still demonstrates a significant performance gap to ideal, cloud-free patches. Further, a certain bias in mispredictions can be observed, specifically decreasing performance on classes that include high spatial frequency details (e.g. Urban, Forest). A qualitative input feature attribution reveals that even non-cloudy, detailed areas of images get blurred in the cloud-removal step, such that the classifier confuses them for less detail-reliant classes (e.g. Croplands, Grasslands).

1 Introduction

The availability of multi-spectral satellite imagery has significantly increased with new and more earth observation satellites, creating numerous important applications, among others change monitoring, weather forecasting, land cover classification, and disaster monitoring. However, a major and unavoidable challenge in satellite imagery is cloud cover: approximately 67% of the Earth's surface is covered with clouds at any given time, with land masses experiencing cloud covering of 55% on average [1]. Generally, clouds obstruct visibility and complicate the aforementioned applications, essentially creating an image reconstruction task, in which parts of the image's information is missing and has to be restored.

To tackle this issue, various approaches for cloud removal from satellite images have been devised. Traditional methods have proven to be limited in their effectiveness, especially with opaque cloud cover. Deep Learning models, trained with datasets that include cloudy imagery, implicitly condition the network to ignore the cloudy regions. Aside from that, explicit cloud removal has gained popularity, especially with advanced architectures like Generative Adversarial Networks (GANs) and Diffusion models. Moreover, the integration of synthetic aperture radar (SAR) data has proven to be an effective auxiliary data source to assist cloud removal models, further enhancing the performance of cloud removal methods.

Using land cover classification as an example, we explore the efficacy of an explicit cloud removal step in a satellite image processing pipeline rather than disregarding those images in the case where a model has not seen cloudy images during training and therefore not learned to ignore cloudy regions. For such a separate cloud removal step to become standard, the generated image distribution must align as closely as possible with cloud-free data, ensuring the highest quality results. In this work, we compare the performance of cloud-removed data from a state-of-the-art Deep Neural Network with corresponding cloud-free satellite imagery on the task of land cover classification, specifically using a model that is trained on cloud-free data. With this, we want to address whether researchers and industry using models that have been trained on cloud-free data could benefit from using cloud-removed data additionally (instead of disregarding it), e.g. to get better real-time monitoring. Further, we provide a data separability study to estimate the proximity of data distributions between cloud-free and cloud-removed satellite images. Lastly, to explore the qualitative impact of the reconstructed image regions to the final output, an input feature attribution study with Gradient-weighted Class Activation Mapping (Grad-CAM) [2] is performed and reported.

2 Related Work

In general, cloud removal via Deep Learning can be categorized two-fold: Mono- vs. multi-modal and mono- vs. multi-temporal approaches. In multi-modal approaches, auxiliary data (most often SAR data) is used to improve the reconstruction results, which has been proven effective in various studies. Grohnfeldt et al. [3] proposed a conditional GAN which includes SAR data from the ESA Copernicus Sentinel-1 mission. Meraner et al. [4] presented paired Sentinel-1 SAR and Sentinel-2 cloudy and cloud-free data in a data set called SEN12MS-CR and, alongside, a novel Convolutional Neural Network (CNN) architecture for cloud removal. Other multi-modal Deep Neural Network architectures for cloud removal are models utilizing other GAN variants (cycle-consistent GANs in [5]), self-attention mechanisms [6] [7] or diffusion models [8]. In multi-temporal architectures, matched images from the same patch across time are used to complement clouded areas and ideally allow for at least some cloud-free information in every part of the image. Examples include again GAN-based methods [9], CNNs [10], self-attention models [11] and diffusion models [12]. Often, multi-temporal approaches have difficulty with

rapidly changing landscapes (e.g. agricultural regions) due to the limited revisit time of earth observation satellites to the same region. Also, the matching process of images itself is additional overhead that has to be performed before cloud removal can be applied. Ebel et al. [10] therefore presented an extension to the SEN12MS-CR data set called SEN12MS-CR-TS, which includes multiple cloudy images per cloud-free ground truth, enabling multi-temporal support. Both SEN12MS-CR and SEN12MS-CR-TS are de-facto standard datasets in this area of research [4] [5] [12] [8] [10] [11] [6] [13] [14] [15] [7].

For downstream task evaluation, Gu et al. [14] demonstrate with the use case of land cover classification that explicit cloud removal results in higher precision and recall compared to a network that is trained on cloudy data and assumed to learn to ignore cloudy regions in the input implicitly. Further, Gawlikowski et al. [13] demonstrate the issues that arise from out-of-distribution test data, i.e. cloudy data, when training on cloud-free images, as most satellite imaging data sets are published with cloudy patches removed. With land cover classification as their downstream task, they employ a detailed analysis of the data distribution of images per land cover class and cloud coverage, classification results per class as well as network confidence predictions, finding that models make overconfident and wrong predictions with increasing cloud cover. Additionally, they investigate the input feature attribution via so-called saliency maps, indicating the model's attention to certain pixels conditioned on a particular prediction class and show that cloudy and cloud-free data can be effectively separated by analysing the network logits (activations before applying the last softmax layer), demonstrating a clear difference in data distributions.

What both downstream task evaluations lack however is a detailed analysis of cloud-removed images compared to cloud-free data. By achieving very good results expressed in metrics like structural similarity (SSIM), peak-signal-to-noise ratio, etc., it is unclear whether those results are enough to perform well in downstream tasks. Ideally, the proximity of cloud-free and cloud-removed data distributions is so high that downstream tasks cannot distinguish between them and work as well on cloud-removed data as on real cloud-free images.

3 Methodology

3.1 Data Set

Due to the need of both land cover classification labels and cloudy images for given cloud-free satellite observations, this work utilizes both the SEN12MS [16] and SEN12MS-CR [4] datasets.

SEN12MS was published in 2019 and contains 180,662 triplets of $256px \times 256px$ patches, originating from the ESA Copernicus Sentinel 1 and Sentinel 2 missions. Each triplet contains a Sentinel-1 (S1) SAR image with two polarimetric channels (VV, VH), a Sentinel-2 (S2) multi-spectral optical image with 13 channels and a Moderate Resolution Imaging Spectroradiometer (MODIS) land cover map using four difference classification schemes. The triplets are derived from so-called scenes, which are sampled from all inhabited

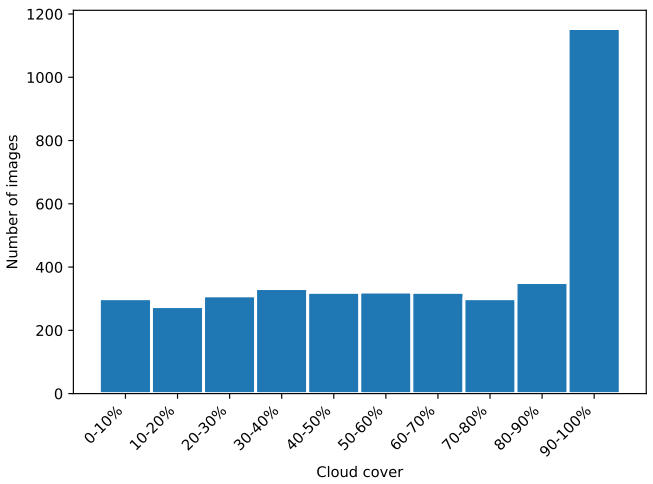


Fig. 1: Cloud coverage histogram on test data set (3982 images).

continents and across all seasons. For this work, in order to confine computational load, only approximately 1/4 of scenes are utilized (selection seed 1158). They were taken in spring as defined on the northern hemisphere (1 March 2017 to 30 May 2017), but still cover all continents like the full data set. For the classification labels, the International Geosphere Biosphere Programme (IGBP) scheme in its simplified version is used, containing 10 different classes. For each patch, a single label is defined by the mode of all single pixel classes.

In order to pair the patches from SEN12MS with cloudy data, the compatible SEN12MS-CR is utilized, which was created with the intention of exploring cloud-removal in Deep Learning. Next to the S1 and S2 patches, it includes a cloudy S2 image from the same meteorological season in order to limit surface changes, amounting to 157,521 patches in total. All patches in the data set are intersected with SEN12MS to have quadruplets of S1, S2, S2 cloudy and land cover data, resulting in 28,396 patches. Using the train, validation and test splits defined for SEN12MS-CR, set sizes of 22,124, 2,290 and 3,982 are obtained respectively. The cloud cover distribution in the test set is shown in Figure 1 using standalone cloud detector *s2cloudless* [17]. It is apparent that the cloud coverage seems to be distributed uniformly except the very high cloud coverage between 90-100%, which occurs more frequently compared to the rest. As this distribution will tend to challenge the cloud removal model more, it is expected to see more pronounced differences between cloud-free and cloud-removed data due to the bias towards near-full cloud coverage.

Another statistic on the data set is the class distribution on the pre-defined data set splits, visualized in Figure 2. It is apparent that there are pronounced imbalances between the classes, as e.g. Shrub Land, Wetlands, Snow & Ice and Barren are hardly represented at all, while Savannas make up approx. 30% of the training set. The validation split composition of classes also differs from the train and test sets, especially for Forest, Croplands and Urban & Built-up. Nonetheless, the distributions (especially train and test) are comparably similar, which is why no adaptation of the

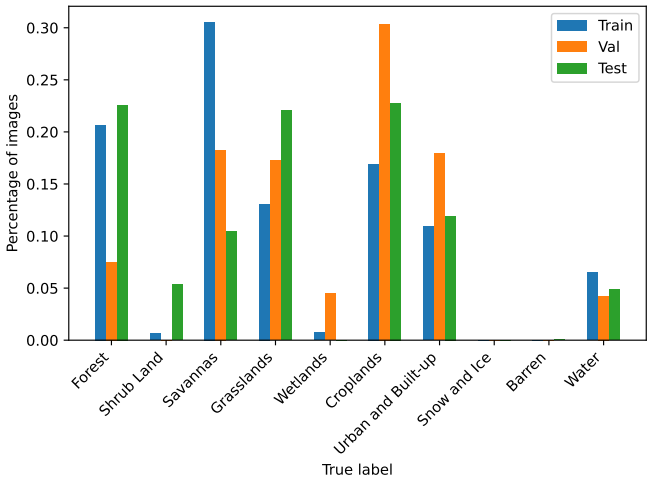


Fig. 2: Class distribution histogram on data set.

data set splits is carried out. Further, an adaptation of splits would mean the exclusion of pre-trained models for the cloud removal step, which have been trained on those data splits.

3.2 Cloud Removal

In order to compare the cloud-free patch performance against a cloud-removed counterpart, the paired cloudy images are handed to a pre-trained state-of-the-art cloud removal Deep Neural Network to obtain cloud-free predictions. This is done with all patches from the test set only, as the other splits were used in the training process. As cloud removal model, the mono-temporal and multi-modal version of UnCRtainTS by Ebel et al. [11] is utilized, as it is the best performing model to the best of our knowledge at the time of writing. Primarily intended to work on another, more recent evolution of SEN12MS-CR containing time-series of cloudy images (in order to ideally see all areas in a patch cloud-free at least once), it outperforms all other mentioned related work in the mono-temporal setting as well except for SSIM, in which it is the second best model after [6]. Architecturally, it features an encoder to spatially encode the cloudy input and an attention mechanism operating on downsampled feature maps to compute attentions masks. Applying those masks to the spatially encoded feature maps, a decoder produces the cloud-removed output. Next to its performance, the model architecture is chosen due to the availability of open-source code and pre-trained model weights. The output of this step are cloud-removed predictions for all test set patches.

3.3 Land Cover Classification

Having obtained the cloud-free predictions from the cloud removal model, the downstream task of land cover classification can be explored. For that, a Convolutional Neural Network is trained on the cloud-free train and validation splits, before exposing it to the cloud-free and cloud-removed patches during inference. The outputs for both classes of patches are analysed in a subsequent step. For the model architecture, a common and widely used classification architecture is taken, namely a ResNet50 [18]. In a study of Schmitt et al. [19], the authors benchmarked different architectures for land cover classification (including a ResNet50 and DenseNet121), and reported adequate performances for these architectures. They also utilize the SEN12MS dataset, however with different splits, such that the pre-trained models for land cover classification cannot be employed. Instead, their open-source code is slightly adapted and used as a basis for training. In order to match the expected input size of the model, all images are cropped to $224px \times 224px$. An overview of the full methodology is shown in Figure 3

4 Evaluation

The evaluation of the land cover classification task features a comparative performance analysis of cloud-removed and cloud-free data, as well as a separability study to estimate the proximity of data distributions. Moreover, a qualitative analysis shall be carried out, highlighting differences in the input patch areas between both classes.

To bridge the gap to the metrics used in the cloud-removal step and examine their correlation with downstream task performance, we will introduce the peak-signal-to-noise ratio, the structural similarity index (SSIM), and spectral angle mapper (SAM) [20].

PSNR is a pixel-based metric comparing the mean squared error per pixel between images and scaling by the maximum possible signal (therefore, the higher, the better).

$$PSNR(x,y) = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE(x,y)}\right)$$

In contrast, SSIM focuses on large-scale structures and perceived visual similarity instead of per-pixel comparisons. It computes the luminance, contrast and structure between the images and outputs the weighted sum of those. It is limited in $[0,1]$, where higher means better. Lastly, SAM computes the angle between two image spectra (lower is better). By normalizing both, it is insensitive to gain factors and rather focuses on the similarity of spectrum composition. It is computed as

$$SAM(x,y) = \arccos\left(\frac{x \cdot y}{\|x\|_2 \cdot \|y\|_2}\right)$$

Ideally, by scoring well in those metrics, the reconstructed images should be suited for downstream tasks, which is subject of this study. For the classification, well-established metrics will be used to compare the predictions. For performance, the precision, recall and F1-score are computed from the confusion matrix. Macro averaging is used as averaging method across classes, because the data set is imbalanced, still all classes receive the same weight. Accuracy is not a suitable metric due to the imbalance in the distribution.

By using cloud-removed images for inference on a classifier trained on cloud-free data and observing the network's outputs, this can be seen as an application of Out-of-Distribution detection to determine whether cloud-free and cloud-removed patches can be easily separated by the

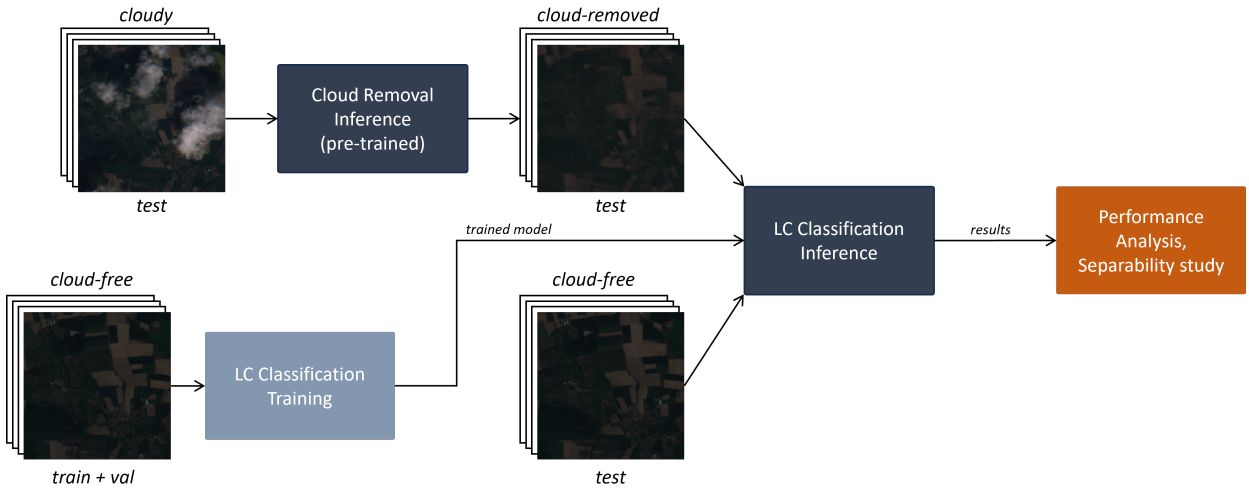


Fig. 3: Methodology approach for comparing cloud-removed and cloud-free performance in land cover classification.

model's predictions, indicating a dissimilarity in data distributions. Different metrics exist to use for the separation, based on the model's logits (outputs before applying the softmax function to obtain probabilities) or the probabilities itself. Gawlikowski et al. [13] provide several examples of metrics, which are adopted in this study. We will investigate separability based on the maximum probability (or confidence), entropy of the output probability vector, sum of the logit values (logit-sum), and so-called *precision*, i.e. sum of exponential logit values. For all metrics, the area-under-the-curve (AUC) for precision/recall (PR) and receiver-operating characteristic (ROC) curves will be measured, whereas an area of 1 corresponds to perfect separability.

To understand what might drive mispredictions on cloud-removed data and which scenarios most problems are caused by, a qualitative input feature attribution method called Gradient-weighted Class Activation Mapping (Grad-CAM) [2] is performed. This method computes the gradient $\frac{\partial y_c}{\partial A_k}$ of a logit y_c for an output class c with respect to k feature maps A_k in the last convolutional layer of a CNN (just before the fully-connected layers), as those layers promise the most correspondence between semantic meaning and spatial structure in the network. Then, those gradients are averaged in each feature map (called global average pooling) to obtain a weight α_k^c indicating the influence of this particular feature map on the output class.

$$\alpha_k^c = \frac{1}{H \times W} \sum_i \sum_j \frac{\partial y_c}{\partial A_{k,i,j}}$$

A saliency map M_c for this class (in the dimensions of the last convolutional layer) is then computed as linear combination of all feature maps with the alpha values as weights, fed into a rectified linear unit.

$$M_c = \text{ReLU}\left(\sum_k \alpha_k^c A_k\right)$$

In a final step, the saliency map is upsampled bilinearly to the dimensions of the input, indicating which parts of the input are salient in the classifiers prediction for the particular class. Despite recent work [21] uncovering weaknesses in Grad-CAM due to the global averaging of gradients, which can cause highlighting of regions that have not contributed to the prediction, the method is still used due to the qualitative-only nature of this analysis step, as well as time constraints in the project.

5 Results

Exemplary predictions of the pre-trained UnCRtainTS on the test set are shown in Figure 4. It is apparent that the model is able to reconstruct the image well when there are only thin or partial clouds. The thicker the clouds and the more high-frequency details are present on the ground cover (e.g. urban areas), the worse the model performs. In case of full cloud cover, the model has to rely solely on the SAR data which is insensitive to clouds and has to infer the optical reconstruction without any reference, recognizable by similar structures, but different colors in the predicted image. The pre-trained model's performance according to the presented metrics is shown in Table 1, alongside the reported values from the original study [11]. It is apparent that the quality of the reconstruction in terms of the image reconstruction metrics is comparable to the original work, making the application of the cloud-removed images suitable to be evaluated in the downstream task.

TABLE 1: UnCRtainTS performance on selected test set

Test set	PSNR	SSIM	SAM
Ours	28.79	0.884	8.10
Ebel et al. [11]	28.90	0.880	8.32

5.1 Classifier Training

The ResNet50 is trained based on the open source code of Schmitt et al. [16], with all hyperparameter settings as proposed in their work. For the entirety of the training and tests in the downstream task, only the multi-spectral S2

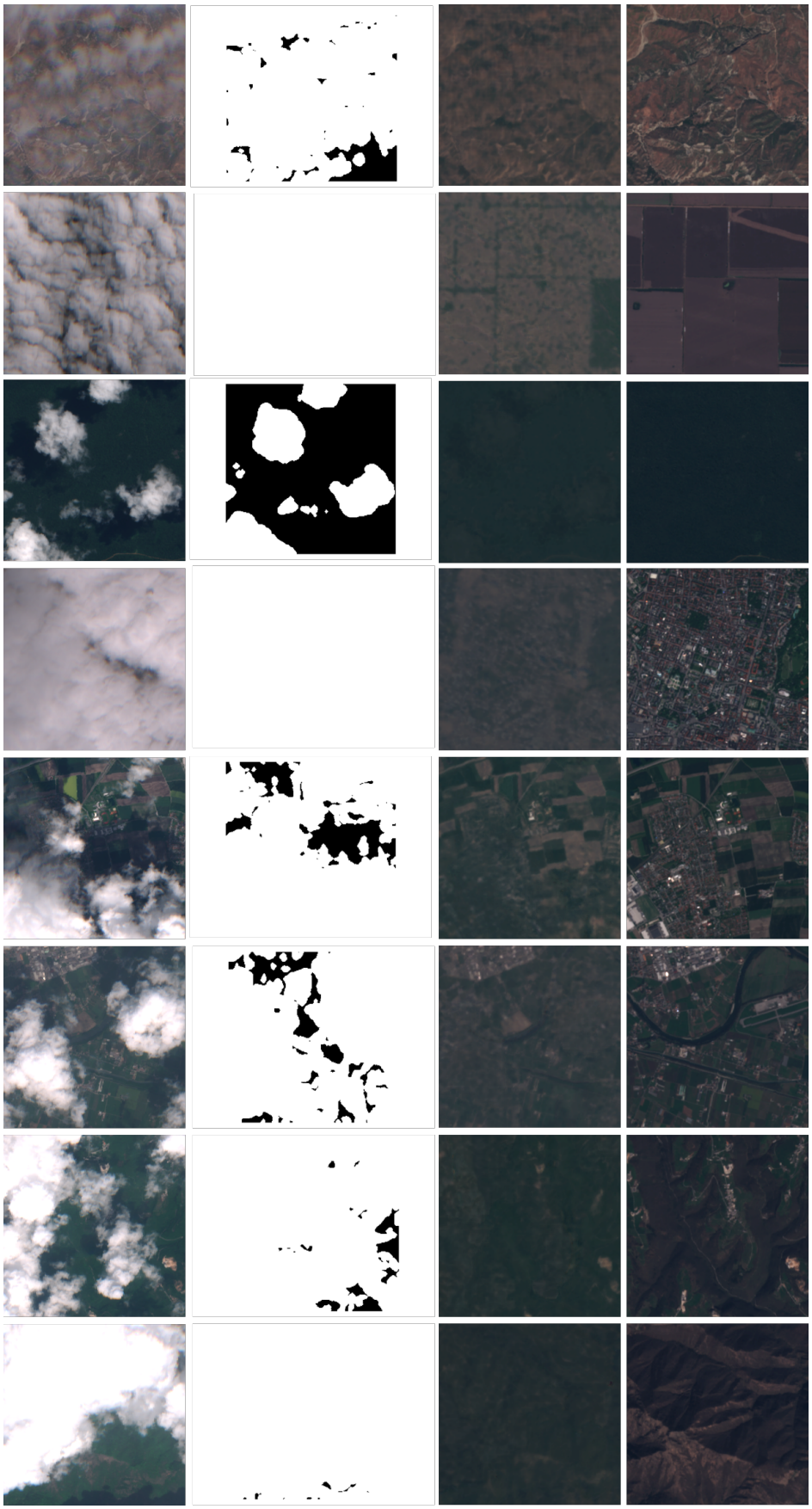


Fig. 4: Example predictions of UnCRtainTS. Different patches in rows. Per row: Cloudy image, computed cloud mask, prediction, target image.

imagery is considered. Due to time and computational constraints using Google Colab, no hyperparameter tuning was conducted. However, this means mostly that the absolute performance of the classifier will not be at its achievable maximum, while the relative performance between cloud-free and cloud-removed images is still assumed to be representative. The best validation F1-score is achieved after 10 epochs with 0.711 (micro-averaging) / 0.471 (macro-averaging). As the training loss keeps decreasing in the following epochs, while the F1-scores drop again, there might be overfitting of the model occurring, presumably due to the reduced amount of training data with the same model complexity as intended for the full data set.

5.2 Classification Performance

Testing the trained network on the test split with cloud-free and cloud-removed images yields the confusion matrices shown in Figure 5. Additionally, the network is tested with the original cloudy data to obtain a comparable baseline, i.e. expected performance on Out-of-Distribution samples. The cloud-free matrix indicates adequate generalization behaviour for the classes Forest, Savannas, Croplands, Urban and Water. Wetlands, Snow & Ice and Barren have too little support to be evaluable. For Shrub Land and Grasslands, the performance of the network is poor even on cloud-free data, which could presumably be caused by the difference in data distributions, namely significantly less samples in the training split. For the cloud-removed images, the predictions seem to be biased towards the class "Savannas", which is visualized in a difference of the confusion matrices in Figure 6. For all classes except Grasslands, the fraction of correctly labeled samples drops significantly. Only 8% of urban patches are classified correctly, meaning a decrease of 59%. However, the classes Water, Savannas, Croplands and Grasslands are not affected substantially by using their cloud-removed counterparts. The test of cloudy images reveals a bias of the network towards the prediction "Urban". For all remaining classes except Water, classification performance drops even more than for the cloud-removed equivalents.

Next to predictions grouped by class, another consideration is the difference in classification performance grouped by cloud cover of the original cloudy image (before removal). Figure 7 shows the F1-score, precision, recall and average confidence (maximum probability) per 10% cloud coverage. Note that the cloud-free predictions should not be impacted by the grouping, therefore there should not be considered to be a trend (as no cloudy images have been used). It is apparent how cloudy images worsen the performance with increasing cloud cover. In the case of cloud cover $> 90\%$, only a fraction of images can be classified correctly. Interestingly, the confidence of the model does not seem to suffer and even increases with higher cloud cover, indicating overconfident mispredictions as analysed in [13]. The cloud-removed images, on the other side, demonstrate a middle ground between cloud-free and cloudy performance. While the precision does not decrease much, the recall is deteriorating with increasing cloud cover (before removal), but still significantly better than the cloudy comparison. Interestingly, the confidence in predictions remains high, but levels off slightly below cloud-free and cloudy baselines, indicating less confident predictions.

Grouping the presented classification performance metrics by their SSIM score on the test set (as calculated during cloud removal inference), we can examine correlations between numerical quality of the reconstruction (in terms of SSIM) and downstream performance. The results are shown in Figure 8. Other than expected, a higher reconstruction score does not result in overall better classification performance. While the precision increases, recall even drops slightly, making the F1-score mostly constant across SSIM scores. The network becomes slightly more confident with images that were reconstructed better according to SSIM.

5.3 Separability Study

The previous results have demonstrated a noticeable difference in classification performance between cloud-free and cloud-removed data. Therefore, it is not expected that the distribution of network predictions will be congruent, which would result in no performance difference. However, some insights can be gained from comparing the distributions of different Out-of-Distribution metrics as presented in Section 4. The distributions are visualized in Figure 9. In the maximum probability, the cloudy predictions seem to be more frequently represented in the higher confidences, which is in line with the confidence analysis before, confirming the finding of overconfident mispredictions on cloudy images. The high cloud-free prediction confidences are similar to the cloud-removed patches, while in the middle range (approx. 0.7-0.95), more cloud-removed samples are located. Again, this is in line with the observation that the confidence is indeed lower than on the cloud-free test set, which is advantageous considering the associated dip in performance. Here, a good separability of distributions might be appropriate given the differing predictive power on the cloud-removed set. A similar image can be drawn from the cross entropy distributions, as the cross entropy in the output probability vector is strongly correlated with the confidence (e.g. a perfect prediction has cross-entropy 0). It is observable that the cloud-free predictions have cross-entropies close to 0 by a multiplicative factor (hence the logarithmic scale in the graph), while cloud-removed patches are more frequent in higher entropies, again signaling uncertainty.

The precision, i.e. sum of exponential logits, only displays major differences in the small range $[0 - 50]$, in which the cloudy patches are predominant, followed by cloud-free and cloud-removed samples.

Lastly, the logit-sum is a metric that is applied before obtaining actual probabilities, indicating the raw output of the neural network. In this case, it can be seen that the distributions of cloud-free and cloud-removed images are in fact very similar, and the distribution of cloudy images is noticeably different. As this metric considers the raw output before converting to probabilities, it shows that the model is not behaving significantly differently on cloud-removed data, speaking in favor of a proximity of distributions. The results in 2 confirm the qualitative findings for the most part. The best separability between cloud-free and cloud-removed images can be achieved with maximum probability, followed by precision. Logit-sum and Cross-Entropy

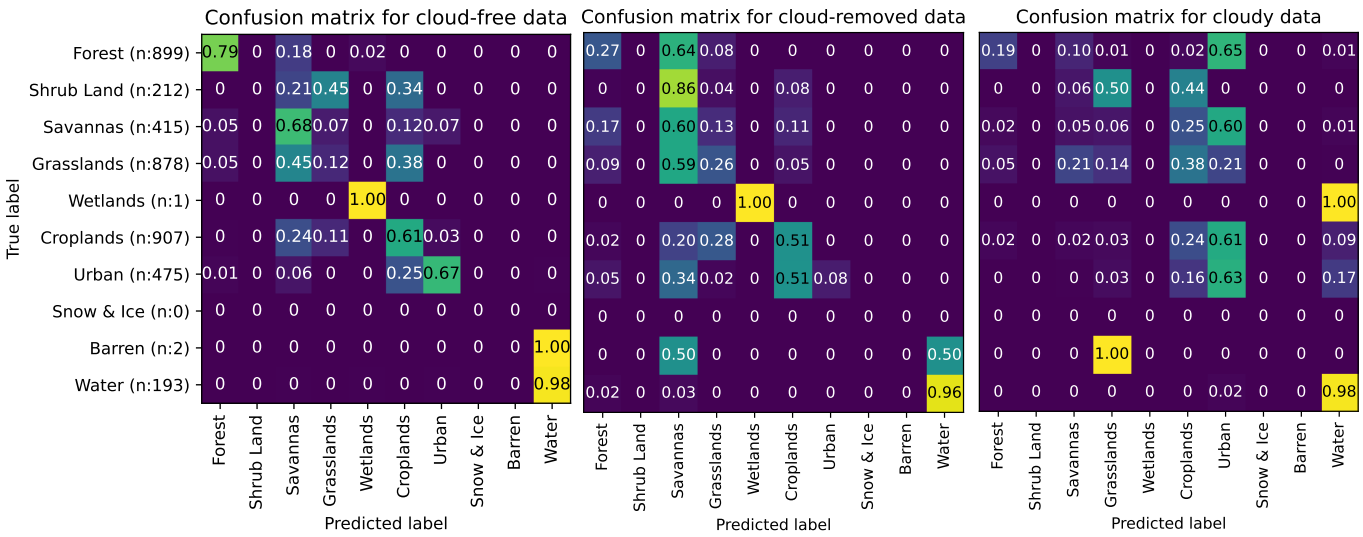


Fig. 5: Confusion matrices of test data in cloud-free, cloud-removed and cloudy splits.

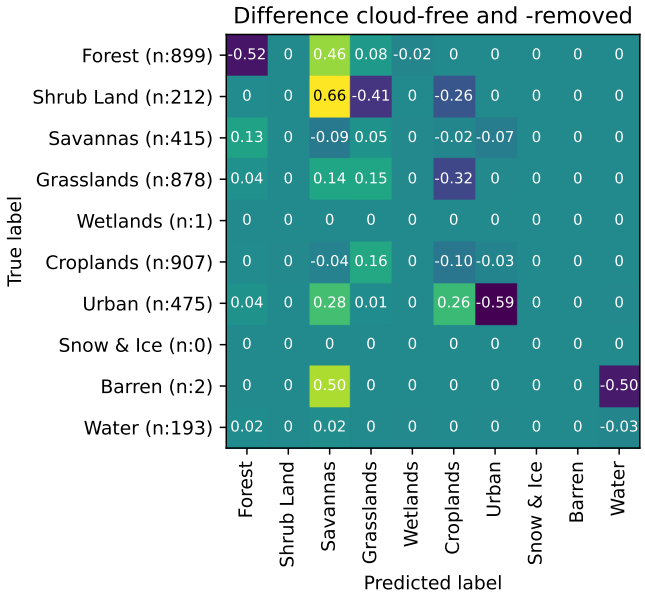


Fig. 6: Difference in confusion matrices for cloud-free and cloud-removed data.

can be worse discriminated than taking random guesses (corresponding to 0.5), indicating that the distributions are comparably close.

5.4 Input Feature Attribution via Grad-CAM

The trained classifier is used with the input feature attribution mechanism Grad-CAM to obtain saliency maps of the model's attention in the input for a certain output class. Analysing the outputs of Grad-CAM, in the following, common misperceptions, confusions and reoccurring patterns are described. For each example, the Figures show the cloudy, cloud-free and cloud-removed image, as well as the saliency map of the cloud-free image for the correct target

TABLE 2: Separability study: ROC/PR AUC for cloudy and cloud-removed data, all compared to cloud-free distribution

Method	Clear vs. Cloud-removed		Clear vs. Cloudy	
	PR	ROC	PR	AUC
Max. prob.	0.642	0.609	0.597	0.564
Cross-Entropy	0.425	0.372	0.381	0.290
Precision	0.584	0.527	0.583	0.595
Logit-sum	0.464	0.430	0.383	0.329

class A_k^t and the saliency maps of the cloud-removed patch for the predicted (false) and true output class, $A_k^{'f}$ and $A_k^{'t}$.

Figure 10 demonstrates one common misperception using the cloud-removed data. Examining the activation heatmap for the predicted class on the cloud-removed image (e) shows that mostly the areas of previous cloud cover are considered in the prediction. With respect to the true class, the model only considers the bottom edge, which is not sufficient for predicting the patch correctly as such.

In Figure 11, another common observation is demonstrated. Even though the cloud cover, is not thick nor fully-covering (84% according to binary cloud mask), the cloud-removed image gets blurred in most regions, changing the prediction to croplands. This especially impacts urban scenes, as high-frequency details are not sufficiently transferred to the cloud-removed image.

The consequence of a certain in-painting pattern is shown in Figure 12. The forest is almost completely covered by clouds, removing the clouds introduces a landscape that gets mistakenly predicted as grasslands. Similar to this, Figure 13 shows a mixed scene, in which the mode of pixel-wise classes corresponds to croplands. Due to the imperfect transfer of high-frequency details during cloud-removal, the cloud-removed image gets predicted (correctly) as croplands, but only due to the structure of the in-painted urban area (note the attention on the previously urban area to predict croplands in (e)), not due to the actual croplands around the urban parts.

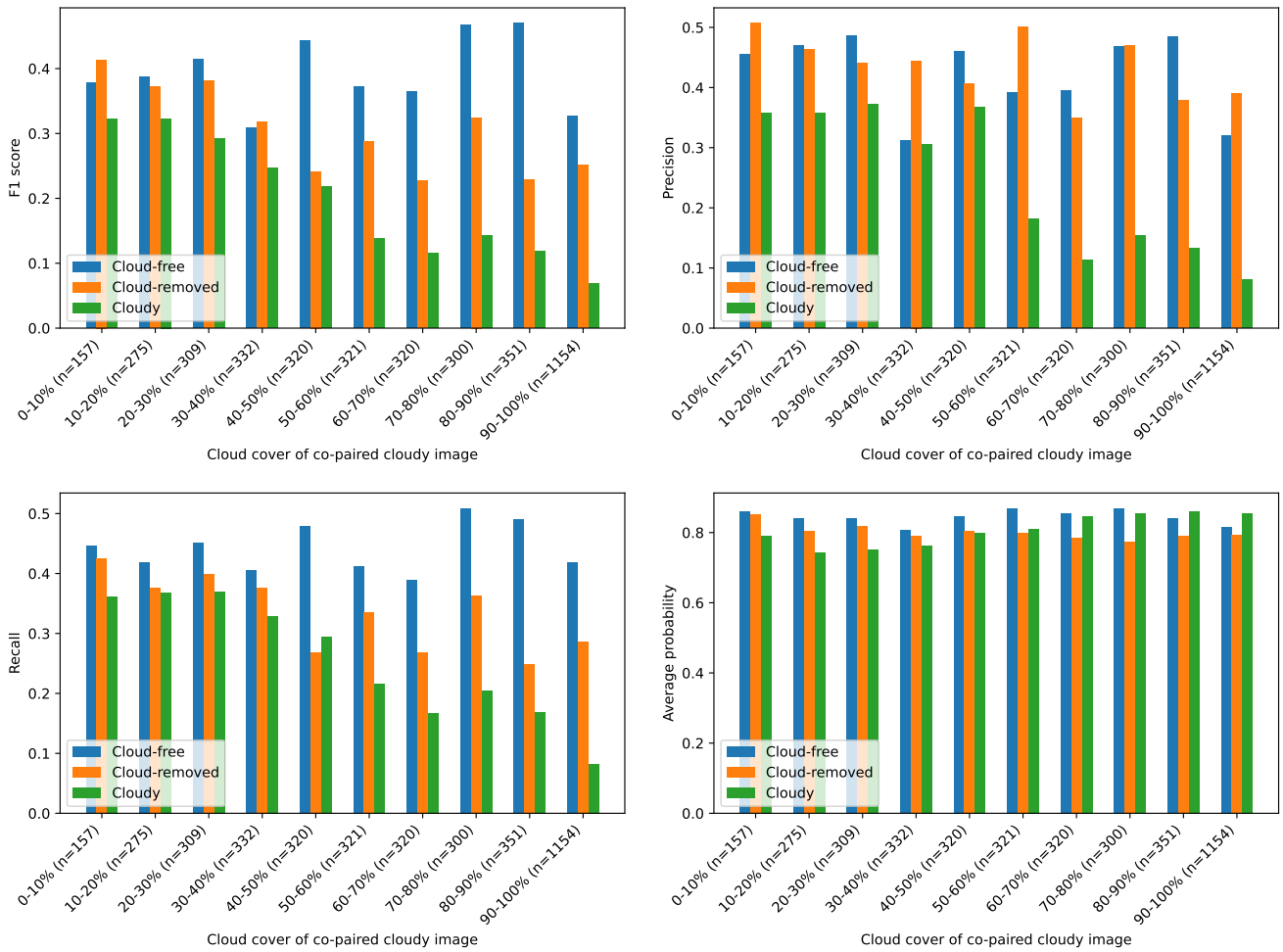


Fig. 7: Classification performance and average confidence, grouped by cloud coverage. Top Left: F1-score. Top Right: Precision. Bottom Left: Recall. Bottom Right: Average Confidence

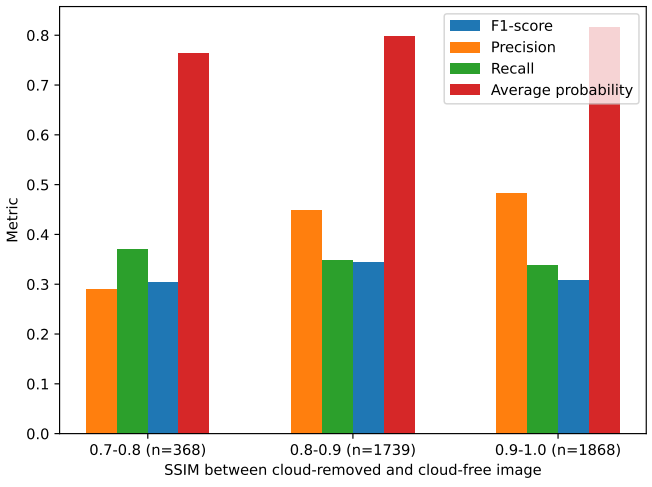


Fig. 8: Classification performance per SSIM score in cloud removal.

6 Discussion

The aforementioned results pose the answer to the initially asked research question, which will be discussed in the following. First, it is reasonable to assume that the quality of the cloud-removed images cannot be significantly improved by any other method at this time, as the obtained metric results equal the results of Ebel et al. [11], who show that their approach UnCRtainTS beats other state-of-the-art methods on a variety of metrics. Having confirmed the quality of the cloud-removed data, the training of the classifier on the cloud-free training set is remains limited due to the project's time and available computational constraints. However, as we are interested in a comparative performance result, this circumstance is acknowledged and accepted. With more hyperparameter tuning and other model architectures, better results will be possible to obtain almost certainly.

One of the first observations is the performance decrease in correctly labeled samples. Overall, the classes "Forest" and "Urban", suffer the most, while "Savannas" and "Croplands" can keep or even improve performance. This leads to the assumption that the cloud-removal process induces a bias towards those classes, which generally tend to comprise less high-frequency spatial features, such as trees, houses, streets etc. Another important fact is the imbalance of some

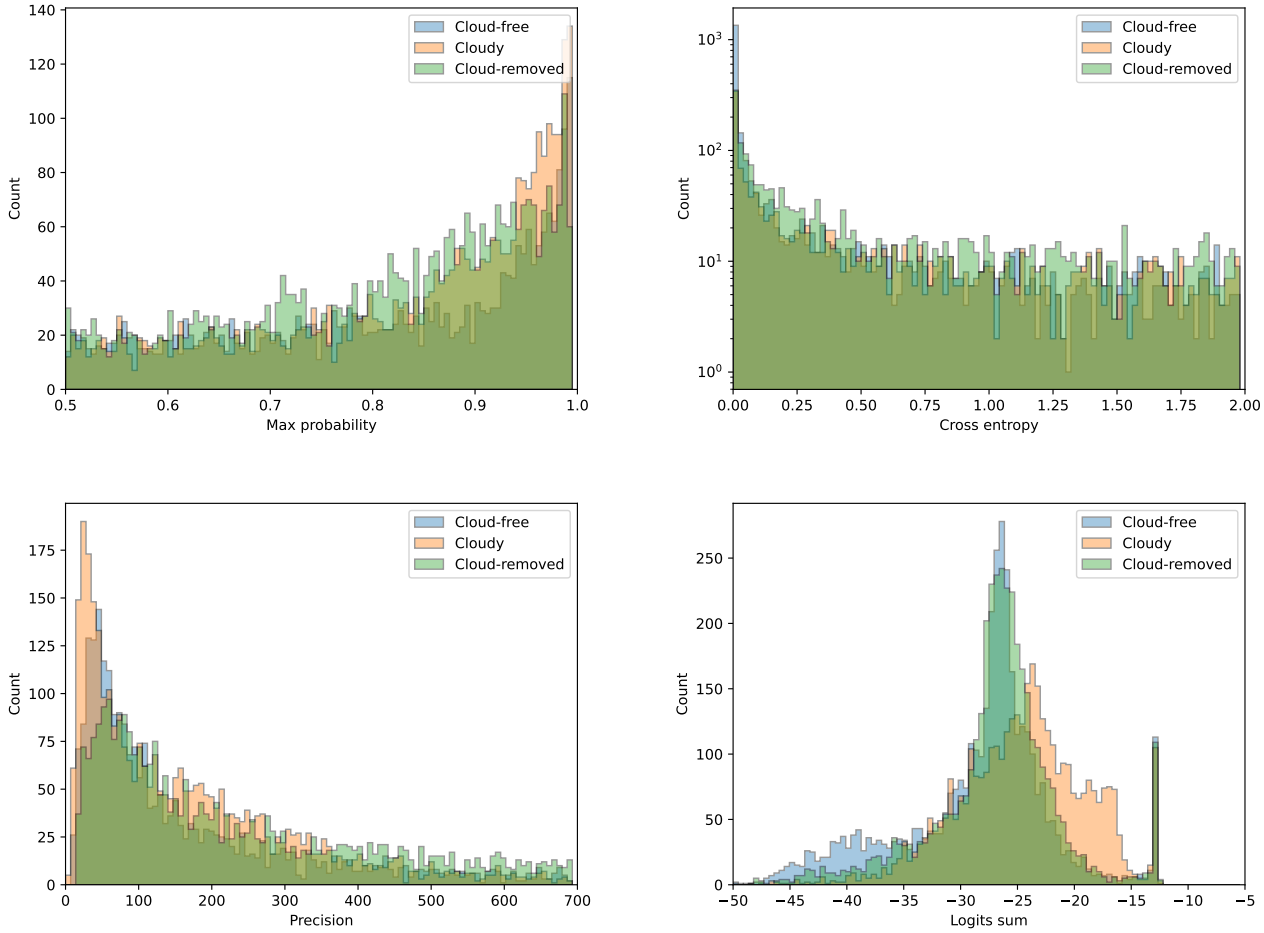
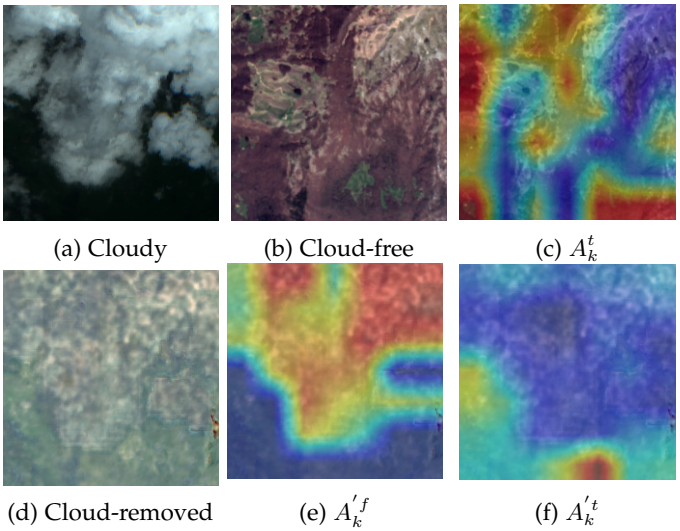


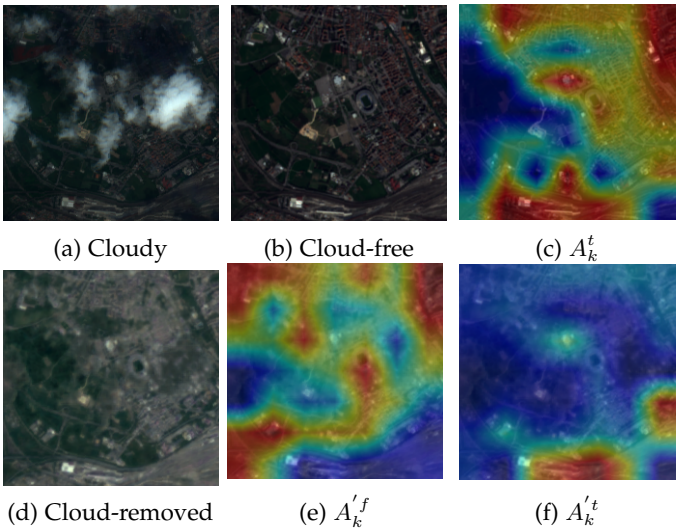
Fig. 9: Distributions of different separability features. Top Left: Max. probability. Top Right: Cross Entropy. Bottom Left: Precision. Bottom Right: Logit Sum



(a) Cloudy (b) Cloud-free (c) A_k^t

(d) Cloud-removed (e) $A_k^{'f}$ (f) $A_k^{'t}$

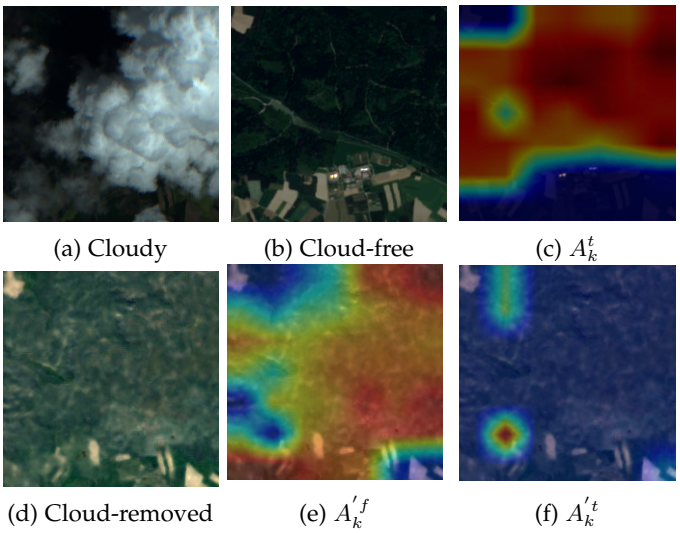
Fig. 10: Cloud-removed false prediction: Savannas (true) vs. Grasslands (pred)



(a) Cloudy (b) Cloud-free (c) A_k^t

(d) Cloud-removed (e) $A_k^{'f}$ (f) $A_k^{'t}$

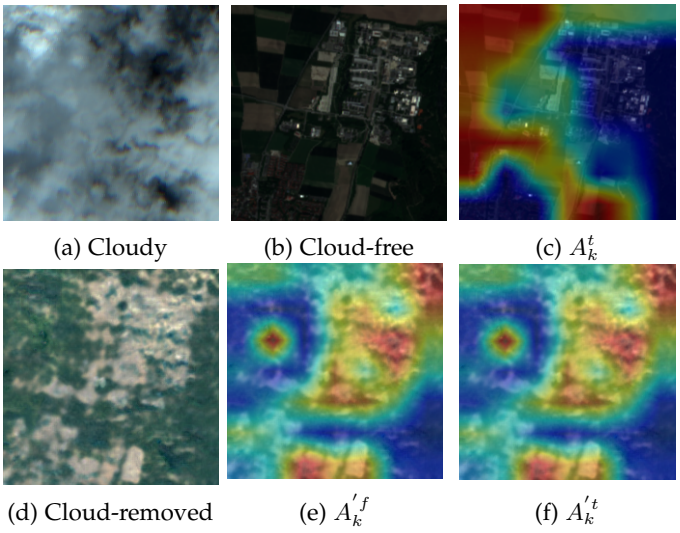
Fig. 11: Cloud-removed false prediction: Urban (true) vs. Croplands (pred)



(a) Cloudy (b) Cloud-free (c) A_k^t

(d) Cloud-removed (e) $A_k^{'f}$ (f) $A_k^{'t}$

Fig. 12: Cloud-removed false prediction: Forest (true) vs. Grasslands (pred)



(a) Cloudy (b) Cloud-free (c) A_k^t

(d) Cloud-removed (e) $A_k^{'f}$ (f) $A_k^{'t}$

Fig. 13: Correct prediction: Croplands

classes, which are underrepresented in the training and test set and therefore poorly classified. For the sake of this study, those classes are simply ignored.

By grouping the performance per cloud cover, we could show that higher cloud cover before removal worsens the performance in the downstream task. This is in line with the bias assumption, which is stronger fulfilled, the more cloudy areas have to be filled in. Even though the performance does not drop as significantly as for cloudy images, it is not en par with cloud-free counterparts. Interestingly, this performance gap is visible in the confidence of the network, whose distributions for cloud-removed and cloud-free patches differ in terms of entropy and confidence score. This is advantageous, as the model expresses its uncertainty, rather than confidently mispredicting images, seen in the case of cloudy samples. Trying to distinguish distributions according to the raw model output (logit-sum) is harder, indicating that the style of in-painting is close to the cloud-free data (even though biased to other classes, as seen before). The model is therefore only less confident, but not confused by the generated data.

Lastly, the qualitative analysis reveals blurring of high-frequency details (e.g. in urban and forest scenes), leading to high activations in different classes than the true label. Unfortunately, the model seems to make use of the generated parts of the patch to output a prediction, which can be different in style than the rest of the patch.

7 Conclusion

In conclusion, the analysis shows that using cloud-removed data is performing better than using cloudy images on a cloud-free trained model. However, by analysing per-class performance, we demonstrate a certain bias in the cloud-removed images, leading to network preference of certain classes. Therefore, we conclude that the introduced bias can affect downstream tasks negatively and should be handled with care. One of the main issues is performance on patches that rely of high-frequency spatial patterns, such as urban scenes. Moreover, unusual events or small hidden objects (e.g. wildfires, ships), occluded by clouds will not be in-painted, therefore this method is generally not applicable to downstream tasks specializing on those events. In order to use cloud-removed images in a downstream task, further analysis of the cloud-removed images is necessary. For example, a metric to favor strong gradients or high spatial frequencies would allow the tuning of cloud-removal models to keep those details instead of blurring in-painted areas.

REFERENCES

[1] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 7, pp. 3826–3852, 2013.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01228-7

[3] C. Grohnfeldt, M. Schmitt, and X. Zhu, "A conditional generative adversarial network to fuse sar and multispectral optical data for cloud removal from sentinel-2 images," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 1726–1729.

[4] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 333–346, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271620301398

[5] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu, "Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5866–5878, 2021.

[6] F. Xu, Y. Shi, P. Ebel, L. Yu, G.-S. Xia, W. Yang, and X. X. Zhu, "Glf-cr: Sar-enhanced cloud removal with global–local fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 192, pp. 268–278, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271622002064

[7] S. Han, J. Wang, and S. Zhang, "Former-cr: A transformer-based thick cloud removal method with optical and sar imagery," *Remote Sensing*, vol. 15, no. 5, 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/5/1196

[8] R. Jing, F. Duan, F. Lu, M. Zhang, and W. Zhao, "Denoising diffusion probabilistic feature-based network for cloud removal in sentinel-2 imagery," *Remote Sensing*, vol. 15, no. 9, 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/9/2217

[9] V. Sarukkai, A. Jain, B. Uzkent, and S. Ermon, "Cloud removal in satellite images using spatiotemporal generative networks," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 1785–1794.

[10] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu, "SEN12ms-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022. [Online]. Available: https://doi.org/10.1109%2Ftgrs.2022.3146246

[11] P. Ebel, V. S. F. Garnot, M. Schmitt, J. D. Wegner, and X. X. Zhu, "Uncrtaints: Uncertainty quantification for cloud removal in optical satellite time series," 2023.

[12] X. Zhao and K. Jia, "Cloud removal in remote sensing using sequential-based diffusion models," *Remote Sensing*, vol. 15, no. 11, 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/11/2861

[13] J. Gawlikowski, P. Ebel, M. Schmitt, and X. X. Zhu, "Explaining the effects of clouds on remote sensing scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9976–9986, 2022.

[14] Z. Gu, P. Ebel, Q. Yuan, M. Schmitt, and X. Zhu, "Explicit haze & cloud removal for global land cover classification," 07 2022.

[15] R. Mao, H. Li, G. Ren, and Z. Yin, "Cloud removal based on sar-optical remote sensing data fusion via a two-flow network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 7677–7686, 2022.

[16] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "Sen12ms – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W7, pp. 153–160, 2019. [Online]. Available: https://isprs-annals.copernicus.org/articles/IV-2-W7/153/2019/

[17] A. Zupanc, "Improving cloud detection with machine learning," Jul 2020. [Online]. Available: https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[19] M. Schmitt and Y.-L. Wu, "Remote sensing image classification with the sen12ms dataset," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-2-2021, 2021, pp. 101–106.

[20] F. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, and A. Goetz, "The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data," *Remote Sensing of Environment*, vol. 44, no. 2, pp. 145–163, 1993, airbone Imaging Spectrometry. [Online]. Available: https://www.sciencedirect.com/science/article/pii/003442579390013N

[21] R. L. Draelos and L. Carin, "Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks," 2021.