Cloud removal using the DSen2-CR model, trained with Twin Mask Adaptive Regularized Loss

Koen Oppenhuis
Leiden University
s1692836

Kay Gijzen
Leiden University
s3724808

Abstract

Cloud cover poses significant challenges in applications that use satellite observations. In this study, we introduce the Twin Mask Adaptive Regularized Loss (TMARL) function, an adaptation of the Cloud Adaptive Regularized Loss (CARL) function. In contrast to CARL, the TMARL function separately handles cloud and cloud-shadow masks. Furthermore, the emphasis TMARL gives to either the cloud mask or the cloud-shadow mask is parameterized; therefore it allows for a more nuanced optimization of the cloud removal network DSen2-CR. Our results indicate that TMARL improves reproduction accuracy, with $\alpha = 0.8$ yielding the best overall performance. However, in cases of heavy cloud cover, CARL remains superior by minimizing hallucinated artifacts.

Keywords

Cloud removal, Optical imagery, SAR-optical, Deep learning, Data fusion, Residual network

1 Introduction

On average, approximately 55% of the Earth's land surface is covered by clouds [5]. Despite improvements in satellite imaging technologies, cloud cover remains a persistent issue. Thick clouds block reflectance signals, obscuring the surface and creating significant spatial and temporal data gaps. These gaps are particularly problematic for applications that require consistent and continuous observations, such as disaster or agricultural monitoring. Cloud removal techniques for satellite imagery offer a promising solution by generating cloud-free data, which can improve the accuracy and reliability of such applications.

Meraner et al. [8] address this problem by proposing the DSen2-CR model, a deep residual neural network designed to remove clouds from Sentinel-2 observations. The model uses SAR-optical data fusion to enhance and guide image reconstruction, with SAR data coming from Sentinel-1 observations. Furthermore, the authors introduce a special cloud-adaptive loss aimed at optimizing the retention of the original image information, allowing for improved cloud removal and enhanced model performance.

We propose an improved approach to cloud removal that better accounts for the distinct characteristic of clouded and shadowed pixels in satellite imagery. Our method builds upon the Cloud Adaptive Regularized Loss (CARL) function by Meraner et al., which has been shown to be effective in cloud removal tasks. However, the original CARL formulation treats the errors of clouded and shadowed pixels equally, which may not be ideal, as the error associated with clouded pixels is typically higher than that of shadowed pixels [11]. In this study, we introduce the Twin Mask Adaptive Regularized Loss (TMARL) function, an enhanced version of CARL, which parameterizes the losses for both cloud and shadow regions separately. This approach allows for a more nuanced approach, where the errors in clouded and shadowed regions are weighted differently. By assigning distinct weights to the contributions of cloud and shadow pixels, the model can optimize each region according to its specific characteristics, improving the overall cloud removal performance.

The effectiveness of the proposed TMARL function is evaluated using the DSen2-CR model. The results are analyzed by comparing the performance of TMARL with the original CARL function, highlighting improvements in the reconstruction of cloud-free images. Additionally, the research explores the trade-offs between different values of the TMARL parameters and asses the impact of separately handling cloud and shadow regions on model performance.

This report is written as part of the 2024 Urban Computing course at Leiden University. The paper by Meraner et al. is selected as a starting point. Any mention of "the original paper", without any explicit other mention, refers to the paper by Meraner et al. [8]. Our contributions to the original paper are:

- Introduction of the Twin Mask Adaptive Regularized Loss (TMARL): A novel loss function that extends the Cloud Adaptive Regularized Loss (CARL), but treats clouded and shadowed pixels separately.
- Evaluation with the DSen2-CR model: The TMARL function is tested on the DSen2-CR deep residual network to assess its effectiveness in cloud removal.
- Parameter trade-off analysis of TMARL: Different TMARL parameter settings are investigated, in order to evaluate their impact on model performance.

2 Related work

In recent years, effort have been made to improve the quality of satellite observations. Several studies have been conducted on the reconstruction of missing information in remote sensing data. Numerous approaches have been developed to address the specific task of cloud removal in optical imagery. Traditional methods can be broadly grouped into three categories: inpainting, multitemporal, and multispectral techniques. Many approaches combine elements from these categories to create hybrid solutions.

Inpainting techniques fill cloud-affected regions by utilizing surface information from the unobscured areas of the same image [7]. These methods do not rely on additional images, but their effectiveness is limited to scenarios with small clouds.

Multitemporal methods reconstruct cloudy scenes by incorporating data from reference images captured under clear sky conditions. These approaches are the most widely used, as they replace cloud-covered pixels with real cloud-free observations. However, they

face challenges when dealing with rapidly changing surface conditions due to temporal discrepancies between the target scene and the reference image [3].

Multispectral techniques are effective for haze and thin cirrus clouds, where optical signals are only partially obstructed by wavelength-dependent absorption and reflection. These methods leverage information directly from the original scene without requiring external data, but their application is limited to semi-transparent clouds [4].

In addition to traditional methods for cloud removal, data-driven approaches leveraging deep learning have recently gained significant attention. Deep neural networks (DNNs) offer the potential to address many challenges associated with traditional algorithms. Ebel et al. [1] propose a novel attention-based neural architecture that maps a cloudy input time series to a single cloud-free optical image. The authors demonstrate how well-calibrated uncertainties can enhance reconstruction quality. Meraner et al. [8] propose a deep residual neural network for cloud-removal in single-temporal Sentinel-2 satellite imagery. Their approach incorporates three key features:

- A data fusion strategy that integrates Sentinel-1 SAR imagery.
- A cloud-adaptive loss function that incorporates a binary cloud and cloud-shadow mask to guide the learning process, prioritizing the retention of input information.
- A globally diverse training dataset sampled across various meteorological seasons.

Cloud and cloud-shadow detection methods, used to generate binary cloud and cloud-shadow masks, can be categorized into two categories [6]: a classical algorithm-based approach and a machine learning approach. Classical algorithm-based approaches primarily rely on threshold-based methods. These methods struggle with generalization across varying environments and conditions. Machine learning approaches train predictive models using data, employing techniques like support vector machines and convolutional neural networks. While these approaches are flexible, their performance heavily depends on the quality and diversity of the training data. The cloud-shadow mask implementation by Meraner et al. uses a combination of cloud detection [9] and cloud-shadow [12] detection methods. Both methods follow an algorithm-based approach.

3 Methods

3.1 Cloud Adaptive Regularized Loss

The Cloud Adaptive Regularized Loss (CARL) function, proposed by Meraner et al. [8], consists of a regularization term and a cloud adaptive term. The regularization term $\mathcal{L}_{reg}$ is given by Equation 1, it is a standard mean absolute error between the predicted output image (denoted by $\boldsymbol{P}$) and the cloud-free target image (denoted by $\boldsymbol{T}$). N_{tot} is the number of pixels in all channels of the optical images.

$$\mathcal{L}_{\text{reg}} = \frac{|| \boldsymbol{P} - \boldsymbol{T} ||_1}{N_{\text{tot}}} \tag{1}$$

The cloud adaptive term incorporates a binary cloud and cloud-shadow mask ($\boldsymbol{CSM}$). This mask has the same size as input images. An entry in the $\boldsymbol{CSM}$ corresponds to a pixel in the input image, and has a 1 for pixels that are either obstructed by clouds or shadows of clouds, and 0 otherwise. The CARL function is given by Equation 2.

$$\mathcal{L}_{\text{CARL}} = \overbrace{\frac{|| \boldsymbol{CSM} \odot (\boldsymbol{P} - \boldsymbol{T}) + (\boldsymbol{1} - \boldsymbol{CSM}) \odot (\boldsymbol{P} - \boldsymbol{I}) ||_1}{N_{\text{tot}}}}^{\text{cloud adaptive part}} + \lambda \overbrace{\frac{|| \boldsymbol{P} - \boldsymbol{T} ||_1}{N_{\text{tot}}}}^{\text{target reg. part}} \tag{2}$$

$\boldsymbol{P}$, $\boldsymbol{T}$ and $\boldsymbol{I}$ denote the predicted output, the cloud-free target, and cloudy input images respectively. $\boldsymbol{1}$ denotes a matrix of ones with the same spatial dimensions as the images and the $\boldsymbol{CSM}$. The operator $\odot$ is an element wise multiplication, it is applied over all channels.

Each pixel in the input image can fall into one of three categories: obscured by clouds, obscured by cloud shadows, or non-obscured. Accordingly, the cloud-adaptive term computes the mean absolute error loss as follows:

- For obscured pixels, the loss is calculated relative to the target image $\boldsymbol{T}$.
- For non-obscured pixels, the loss is calculated relative to the input image $\boldsymbol{I}$.

This approach trains the network to preserve the cloud-free regions of the input image while leveraging multi-temporal information exclusively for reconstructing clouded and shadowed areas.

3.2 Cloud and Cloud Shadow detection

For the generation of the $\boldsymbol{CSM}$, a combination of methods proposed by Schmitt et al. [9] (cloud detection) and Zhai et al. [12] (cloud-shadow detection) is used. These methods yield a binary cloud mask ($\boldsymbol{CM}$) and a binary cloud-shadow mask ($\boldsymbol{SM}$) respectively. The masks have the same dimensions as $\boldsymbol{CSM}$. The $\boldsymbol{CSM}$, as used in the original CARL function is given by Equation 3, it is comprised of the logical OR of $\boldsymbol{CM}$ and $\boldsymbol{SM}$.

$$\boldsymbol{CSM} = \boldsymbol{CM} \vee \boldsymbol{SM} \tag{3}$$

3.3 Twin Mask Adaptive Regularized Loss

In the CARL function, the errors of clouded and shadowed pixels are treated equally. However, this approach has a potential limitation: the error between clouded pixels and the target image is typically higher than that of shadowed pixels. A more refined approach involves weighting the contributions of cloud and shadow regions, reducing the risk of one dominating the optimization process. This reweighing enables a more balanced optimization of the network.

Building on this idea, we propose a novel adaptation of the original CARL function, called the Twin Mask Adaptive Regularized Loss (TMARL) function. The TMARL function, shown in Equation 4, is designed to account for the distinct characteristics of clouds and shadows in satellite imagery. Shadows, being generally darker and exhibiting different spatial features compared to clouds, require separate treatment in the loss function. By parameterizing the losses for both cloud and cloud-shadow regions, the TMARL function

provides the flexibility to tune the relative influence of each, which can lead to improved performance in cloud removal tasks.

$$\mathcal{L}_{\text{TMARL}} = \alpha\ \mathcal{L}_{\text{cloud}} + (1-\alpha)\ \mathcal{L}_{\text{shadow}} + \lambda \frac{||\ \boldsymbol{P}-\boldsymbol{T}\ ||_1}{N_{\text{tot}}} \tag{4}$$

In this formulation, the cloud loss $\mathcal{L}_{\text{cloud}}$ (Equation 5) and the shadow loss $\mathcal{L}_{\text{shadow}}$ (Equation 6) are computed using their respective cloud mask $\boldsymbol{CM}$ and cloud-shadow mask $\boldsymbol{SM}$. The parameter α controls the relative weight of $\mathcal{L}_{\text{cloud}}$ compared to $\mathcal{L}_{\text{shadow}}$. Parameter α should be a value in the interval $[0, 1]$. When $\alpha = 0.5$, the errors in the shadowed regions and errors in the clouded regions are given equal weight. For $\alpha < 0.5$, the shadowed regions are prioritized, assigning greater weight to their errors. Conversely, for $\alpha > 0.5$, the focus shift to the clouded regions, with their errors receiving greater weight. The parameterized losses ensure that the network can prioritize cloud and shadow regions independently, optimizing for each feature's unique characteristics.

$$\mathcal{L}_{\text{cloud}} = \frac{||\ \boldsymbol{CM} \odot (\boldsymbol{P}-\boldsymbol{T}) + (\boldsymbol{1}-\boldsymbol{CM}) \odot (\boldsymbol{P}-\boldsymbol{I})\ ||_1}{N_{\text{tot}}} \tag{5}$$

$$\mathcal{L}_{\text{shadow}} = \frac{||\ \boldsymbol{SM} \odot (\boldsymbol{P}-\boldsymbol{T}) + (\boldsymbol{1}-\boldsymbol{SM}) \odot (\boldsymbol{P}-\boldsymbol{I})\ ||_1}{N_{\text{tot}}} \tag{6}$$

3.4 DSen2-CR

The model used to evaluate the proposed loss function TMARL is DSen2-CR model as proposed by Meraner et al.[8]. This model follows the ResNet architecture design. A Keras[1] implementation with Tensorflow[2] backend of the DSen2-CR model in Python is available[3].

Both the DSen2-CR model structure and its residual block architecture are depicted in Figure 1. For each section of the network, the number of layers and the two spatial dimensions are specified in parentheses. As the network is fully convolutional, it can process input images with arbitrary spatial dimensions m during both trainig and inference. F denotes the chosen feature dimension, while B denotes the number of residual blocks integrated into the network.

In order to reconstruct regions that are obscured by thick clouds, with no available earth surface information, DSen2-CR incorporates a SAR image as a prior. Thus the model takes a Sentinel 1 SAR image and a Sentinel 2 optical image as inputs. The SAR channels are concatenated to the channels of the input optical image.

4 Data

The dataset we will use is the SEN12MS-CR dataset [2] (publicly available at[4]) is used, this dataset is an evolution of the SEN12MS dataset [10]. The SEN12MS-CR dataset includes observations from 175 globally distributed Regions of Interest (ROI), captured across all seasons in the year 2018. Each ROI exists out of a cloudy and cloud-free optical multi-spectral Sentinel 2 observations, as well as the corresponding synthetic aperture radar (SAR) Sentinel-1

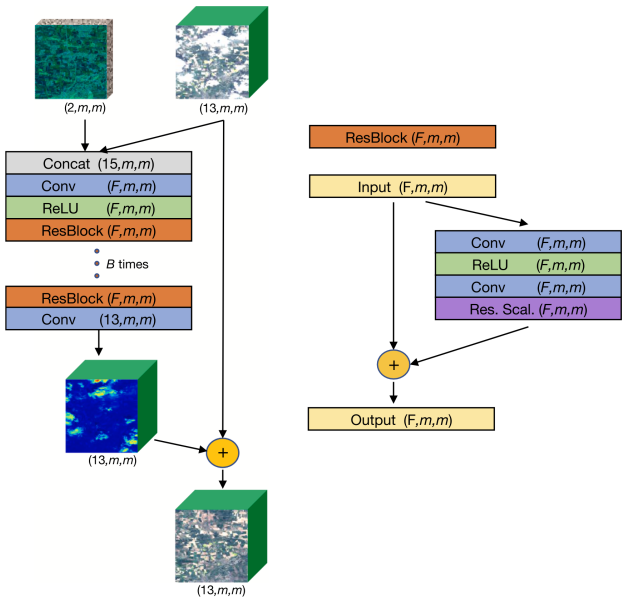


Figure 1: Reproduced from [8]. Left: DSen2-CR model structure. Right: Residual block structure. m denotes the spatial dimensions of input images, F denotes the feature dimension, and B denotes the number of residual blocks.

observation. The average cloud coverage of all data is approximately 48%.

The ROIs are cut up into a total of 122,218 patch triplets with a size of 256 × 256 pixels with a 128 pixel stride. This means that neighboring patches have a 50% overlap. These settings were chosen to maximize the number of patches, but still have the patches be independent from each other. Figure 2 shows an example of a patch triplet from the dataset. All patches are provided in the form of 16-bit GeoTIFFs containing the following specific information:

- Sentinel-1 SAR: 2 channels representing sigma nought backscatter values in dB scale for VV and VH polarization.
- Sentinel-2 Multi-Spectral: 13 channels representing the 13 spectral bands (B1, B2, B3, B4, B5, B6, B7, B8, B8a, B9, B10, B11, B12).
- MODIS Land Cover: 4 channels representing to IGBP, LCCS Land Cover, LCCS Land Use, and LCCS Surface Hydrology layers.

Because the SEN12MS-CR dataset spans diverse geospatial regions and seasons, this enables the development and benchmarking of globally generalizable cloud removal methods. This eliminates the need to retrain models for specific locations or seasonal variations. Therefore, it addresses the limitations of existing cloud removal research, which often evaluates methods on narrowly defined and geographically distinct regions.

[1] https://keras.io/
[2] https://www.tensorflow.org/
[3] https://github.com/ameraner/dsen2-cr
[4] https://mediatum.ub.tum.de/1474000

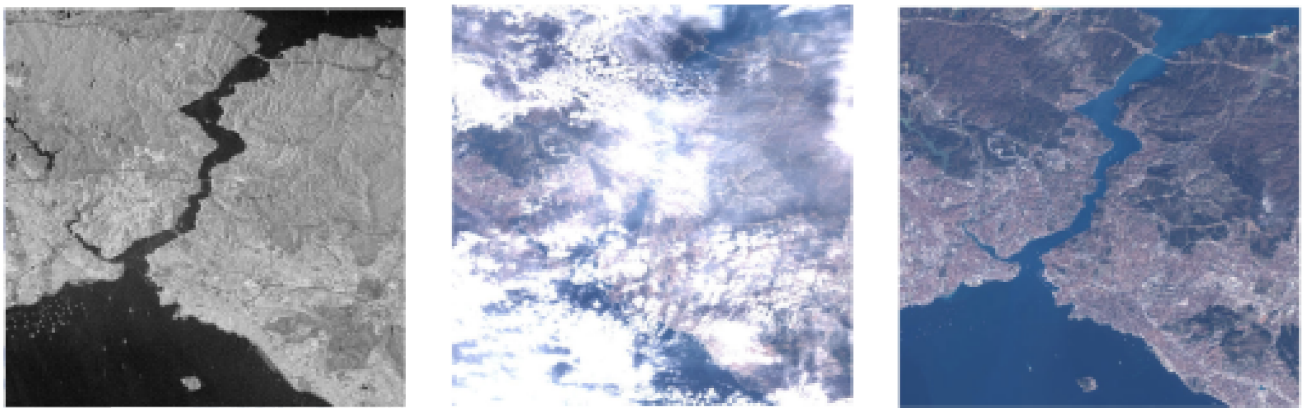


Figure 2: An example 256 × 256 patch triplet from the SEN12MS-CR dataset. Left: Sentinel-1 SAR image, Middle: Cloudy Sentinel-2 optical observation, Right: Cloud-free Sentinel-2 optical observation.

5 Experiments

Code for the implementation of TMARL and the following experiments is made publicly available on GitHub[5].

5.1 Research question

This study aims to explore the potential benefits of treating cloud and shadow masks separately within the loss function for cloud removal in satellite imagery. By distinguishing between the two, the approach seeks to better capture their unique characteristics, which may lead to improved model performance in reconstructing cloud-free images. thus posing the research questions:

- "How does separating the cloud mask and the shadow mask during the training affect the performance of the DSen2-CR model for cloud removal?"
- "What is the optimal value α for optimizing DSen2-CR using the $\mathcal{L}_{TMARL}$ as loss function?"

5.2 DSen2-CR configuration

The DSen2-CR model will be used to assess the proposed loss function TMARL. The model will be optimized using the TMARL function as its loss function, with parameter tuning following the process outlined in the next subsection. Additionally, the DSen2-CR model will be trained using the CARL loss function as a baseline for comparison. The hyperparameters for training the DSen2-CR models, along with their corresponding values, are presented in Table 1.

Parameter	Value
B (Residual blocks)	16
F (Feature dimension)	256
Epochs	8
Batch size	16
Learning rate	7e-5

Table 1: Hyperparameters used for training the DSen2-CR model.

[5]https://github.com/Kaygijzen/dsen2-cr-tmarl

5.3 TMARL Tuning

Since TMARL is parameterized by α, we can either put an emphasis on $\mathcal{L}_{\text{cloud}}$, the error of the clouded regions; or an emphasis on $\mathcal{L}_{\text{shadow}}$ the error of the shadowed regions. We can emphasize $\mathcal{L}_{\text{cloud}}$ by setting $\alpha > 0.5$). Conversely, we can emphasize $\mathcal{L}_{\text{shadow}}$ by settings $\alpha < 0.5$. Thus, in order to find suitable values for the parameterized $\mathcal{L}_{\text{TMARL}}$ we consider the configurations as shown in Table 2.

Situation	L_{cloud}	L_{shadow}
$\alpha = 0.5$	0.5	0.5
$\alpha < 0.5$	0.2	0.8
$\alpha > 0.5$	0.8	0.2

Table 2: Configurations of α and and their effect on the loss in

The λ value, which manages the strength of the regularization term, is held at 1 for these experiments.

5.4 Evaluation metrics

To evaluate the final performance of a trained configuration of the DSen2-CR network, we utilize the pixel-wise comparison metric mean average error (MAE). The MAE of the predicted image is computed in three ways:

- *Reproduction*, the error between the predicted image and the non-obscured regions of the input image:

$$\text{MAE}_{\text{reprod.}} = \frac{|| \boldsymbol{P} - ((\boldsymbol{1} - \boldsymbol{CSM}) \odot \boldsymbol{I}) ||_1}{N_{tot}}$$

- *Reconstruction*, the error between the predicted image and the obscured regions of the target image:

$$\text{MAE}_{\text{recon.}} = \frac{|| \boldsymbol{P} - (\boldsymbol{CSM} \odot \boldsymbol{T}) ||_1}{N_{tot}}$$

- *Total*, the cumulative error of the reproduction error and the reconstruction error. The combination of the non-obscured regions of the input image and the reconstructed regions

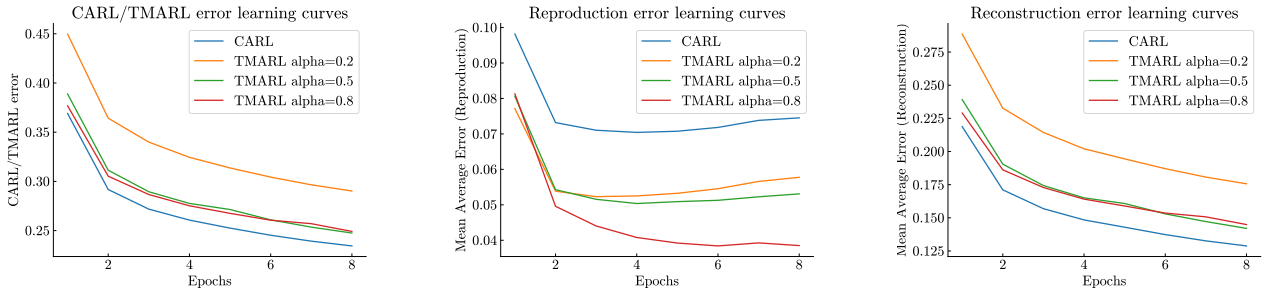


Figure 3: The learning curves of the four models with their different loss functions. The left-side Figure shows the total $\mathcal{L}_{\text{CARL}}$ or $\mathcal{L}_{\text{TMARL}}$ during training, the middle Figure shows the Reproduction Error of the models during training, and the right-side Figure shows the Reconstruction Error during training. All curves show similar behavior, which is understandable because of the similarity in loss functions. $\mathcal{L}_{\text{TMARL}}$ shows a clear improvement for the reproduction error for all values of α. In general $\alpha = 0.8$ outperforms the other 2 tested values.

of the target image gives the best estimate of the output image:

$$\text{MAE}_{\text{total}} = \text{MAE}_{\text{reprod.}} + \text{MAE}_{\text{recon.}}$$

5.5 Training, validation and testing split

We use a subset of the SEN12MS-CR dataset due to the large size of the full dataset, which would make training the model impractical. A uniform random sample of 50% of the available data is selected. This subset is then divided into training (80%), testing (10%) and validation (10%) sets to evaluate and optimize the DSen2-CR models.

6 Results

Training of the various configurations has been performed using the Data Science Lab provided by the Leiden Institute of Computer Science. The training was conducted on 4 NVIDIA GeForce GTX TITAN X GPUs.

6.1 Training results

After training the 4 models with the different loss functions, the concluding errors found, can be seen in Table 3. In comparison to $\mathcal{L}_{\text{CARL}}$, all $\mathcal{L}_{\text{TMARL}}$ variants have a lower $\text{MAE}_{\text{reprod.}}$. Seeing an improvement in the $\text{MAE}_{\text{reprod.}}$ for all $\mathcal{L}_{\text{TMARL}}$ models means that separating the cloud mask and the shadow mask in the loss function helps the model to learn the clear parts of the input images better. In Meraner et al. [8] $\mathcal{L}_{\text{CARL}}$ is proposed as a method to optimize this error, so it is interesting to find an improvement on this.
Contrary to this, the $\text{MAE}_{\text{recon}}$ is lower for the model that uses the $\mathcal{L}_{\text{CARL}}$ loss function. Separating the cloud and shadow masks does not seem to improve the $\text{MAE}_{\text{recon}}$. This would suggest that, instead of getting extra information from see-through clouds and dark shadows, the model is maybe getting too much information which it cannot handle, for instance large sections of thick clouds. In the cumulative error, $\text{MAE}_{\text{total}}$, you still see an improvement for $\mathcal{L}_{\text{TMARL}}$ with $\alpha = 0.5$ and $\alpha = 0.8$, indicating that these loss functions outperform the standard $\mathcal{L}_{\text{CARL}}$ loss function.

In Figure 3, the learning curves during training are shown. Because of the similarity in loss functions, the learning curves show similar behavior too each other. Regarding the optimal value of α, $\alpha = 0.8$ outperforms the other α values. A higher value of α means that the loss with the cloud masks is weighted more than the loss with the shadow masks. One possible explanation for this is that there is in general a higher amount of cloud regions in comparison to shadow regions. The higher α-value would balance the weighing of these regions better.

	MAE (ρ_{TOA})		
Method	**Reprod**	**Recon**	**Total**
DSen2-CR on $\mathcal{L}_{\text{CARL}}$	0.079	**0.123**	0.202
DSen2-CR on $\mathcal{L}_{\text{TMARL}}$ ($\alpha = 0.2$)	0.067	0.150	0.217
DSen2-CR on $\mathcal{L}_{\text{TMARL}}$ ($\alpha = 0.5$)	0.060	0.139	0.199
DSen2-CR on $\mathcal{L}_{\text{TMARL}}$ ($\alpha = 0.8$)	**0.040**	0.139	**0.179**

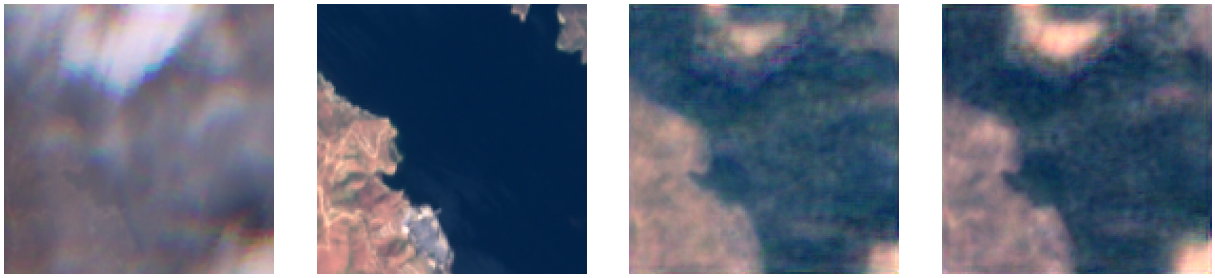
Table 3: MAE values for different methods evaluated on the validation set.

6.2 Prediction examples

Numerical values do not always paint the complete picture. Therefore we also include some examples wherein the differences between DSen2-CR on $\mathcal{L}_{\text{CARL}}$ and DSen2-CR on $\mathcal{L}_{\text{TMARL}}$ are visible in the predictions made by these models. For these predictions DSen2-CR on $\mathcal{L}_{\text{TMARL}}$ uses an α of 0.8.
In Figure 4, we show an example in which we think DSen2-CR on $\mathcal{L}_{\text{TMARL}}$ made a better prediction, and in Figure 5 we show an example in which we think DSen2-CR on $\mathcal{L}_{\text{CARL}}$ made a better prediction.
In pictures labeled with (a), you can see the cloudy images. The pictures labeled with (b) are the clear target images. Finally, the pictures labeled with (c) and (d) are the predictions made by the models. The predictions in (c) are made by DSen2-CR on $\mathcal{L}_{\text{CARL}}$ and the prediction in (d) is made by DSen2-CR on $\mathcal{L}_{\text{TMARL}}$.
The differences between these cloud removal tasks is the amount of cloud coverage. In Figure 4a, there is very light cloud coverage, with some cloud shadows. In Figure 5a, almost the entire picture

(a) (b) (c) (d)

Figure 4: Comparison of DSen2-CR on $\mathcal{L}_{\text{CARL}}$ with DSen2-CR on $\mathcal{L}_{\text{TMARL}}$ ($\alpha = 0.8$). In (a) and (b) the input and target image are shown. In (c) the (d) the predictions made by DSen2-CR on $\mathcal{L}_{\text{CARL}}$ with DSen2-CR on $\mathcal{L}_{\text{TMARL}}$ are shown. For this example, the reproduction of the input image is important. This is due to the very light cloud and cloud shadow coverage. In (c) the model uses too much of the colors from the target image, and therefore this is a worse prediction than (d), which uses the colors from the input image.



(a) (b) (c) (d)

Figure 5: Comparison of DSen2-CR on $\mathcal{L}_{\text{CARL}}$ with DSen2-CR on $\mathcal{L}_{\text{TMARL}}$ ($\alpha = 0.8$). In (a) and (b) the input and target image are shown. In (c) the (d) the predictions made by DSen2-CR on $\mathcal{L}_{\text{CARL}}$ with DSen2-CR on $\mathcal{L}_{\text{TMARL}}$ are shown. For this example, the reconstruction of the input image is important. This is due to the heavy cloud coverage. In (d) the model mistakenly creates islands where the cloud coverage is at its thickest. You can also see this in (c), but here these islands are smaller and less bright. Therefore, the prediction in (c) is a bit better.

is covered by clouds, and you can only see little through them. As seen in Table 3 DSen2-CR on $\mathcal{L}_{\text{TMARL}}$, performs slightly better in reproduction and DSen2-CR on $\mathcal{L}_{\text{CARL}}$ slightly better on reconstruction. You can see this in these two examples.
In Figure 4c, DSen2-CR on $\mathcal{L}_{\text{CARL}}$ uses too much of the colors from the target image, while, in Figure 4d, DSen2-CR on $\mathcal{L}_{\text{TMARL}}$ correctly uses the colors from the input image.
In Figure 5c and d, DSen2-CR on $\mathcal{L}_{\text{CARL}}$ and DSen2-CR on $\mathcal{L}_{\text{TMARL}}$ both correctly make the shape of land in the bottom right. However they both create some extra islands in the places where the cloud coverage is the thickest. DSen2-CR on $\mathcal{L}_{\text{TMARL}}$ does this more than DSen2-CR on $\mathcal{L}_{\text{CARL}}$ and therefore, the latter performs a bit better in the reconstruction loss.

7 Conclusion

In this study, we evaluated different loss function for training the DSen2-CR model in the context of cloud removal in satellite imagery. Specifically, we compared the standard $\mathcal{L}_{\text{CARL}}$ loss function with the proposed $\mathcal{L}_{\text{TMARL}}$ variants, which introduce separate handling of cloud and cloud-shadow masks. Our findings indicate that $\mathcal{L}_{\text{TMARL}}$ improves the reproduction error $\text{MAE}_{\text{reprod.}}$, suggesting that explicitly distinguishing between cloud and shadow-cloud regions helps the model better reconstruct clear parts of the input image.

Among the tested α values in $\mathcal{L}_{\text{TMARL}}$, $\alpha = 0.8$ outperformed other configurations in terms of cumulative error $\text{MAE}_{\text{total}}$, with a value of $\text{MAE}_{\text{total}} = 0.179$. Since $\alpha > 0.5$, we observe a benefit in weighting the cloud mask more heavily than the cloud-shadow mask. This aligns with the fact that clouded regions are more prevalent than shadowed regions in the dataset.

However, the reconstruction error $\text{MAE}_{\text{recon.}}$ was lower for the model trained with $\mathcal{L}_{\text{CARL}}$, this suggests that treating the cloud and cloud-shadow mask separately may introduce challenges in handling highly obstructed areas with thick clouds. This trade-off was further illustrated in qualitative prediction examples, where DSen20CR trained with $\mathcal{L}_{\text{TMARL}}$ better preserved input colors in lightly clouded regions, while DSen2-CR trained with $\mathcal{L}_{\text{CARL}}$ performed better in heavily clouded scenarios by reducing hallucinated artifacts.

Overall, our results suggests that incorporating separate cloud and cloud-shadow masks in the loss function improves the reconstruction of the clear parts of the input image. $\mathcal{L}_{\text{TMARL}}$ performed

best with parameter $\alpha = 0.8$. However, the traditional $\mathcal{L}_{\text{CARL}}$ function performs best, especially in cases of extreme cloud cover. Future work could explore adaptive weighting strategies to dynamically adjust α based on cloud density.

References

[1] Patrick Ebel, Vivien Sainte Fare Garnot, Michael Schmitt, Jan Dirk Wegner, and Xiao Xiang Zhu. 2023. UnCRtainTS: Uncertainty quantification for cloud removal in optical satellite time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2086–2096.

[2] Patrick Ebel, Andrea Meraner, Michael Schmitt, and Xiao Xiang Zhu. 2020. Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery. *IEEE Transactions on Geoscience and Remote Sensing* 59, 7 (2020), 5866–5878.

[3] Patrick Ebel, Yajin Xu, Michael Schmitt, and Xiao Xiang Zhu. 2022. SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–14.

[4] Kenji Enomoto, Ken Sakurada, Weimin Wang, Hiroshi Fukui, Masashi Matsuoka, Ryosuke Nakamura, and Nobuo Kawaguchi. 2017. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 48–56.

[5] Michael D King, Steven Platnick, W Paul Menzel, Steven A Ackerman, and Paul A Hubanks. 2013. Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites. *IEEE transactions on geoscience and remote sensing* 51, 7 (2013), 3826–3852.

[6] Seema Mahajan and Bhavin Fataniya. 2020. Cloud detection methodologies: Variants and development—A review. *Complex & Intelligent Systems* 6, 2 (2020), 251–261.

[7] Fan Meng, Xiaomei Yang, Chenghu Zhou, and Zhi Li. 2017. A sparse dictionary learning-based adaptive patch inpainting method for thick clouds removal from high-spatial resolution remote sensing imagery. *Sensors* 17, 9 (2017), 2130.

[8] Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael Schmitt. 2020. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing* 166 (2020), 333–346.

[9] Michael Schmitt, Lloyd H Hughes, Chunping Qiu, and Xiao Xiang Zhu. 2019. Aggregating cloud-free sentinel-2 images with google earth engine. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 4 (2019), 145–152.

[10] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. 2019. SEN12MS–A curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789* (2019).

[11] Katelyn Tarrio, Xiaojing Tang, Jeffrey G Masek, Martin Claverie, Junchang Ju, Shi Qiu, Zhe Zhu, and Curtis E Woodcock. 2020. Comparison of cloud detection algorithms for Sentinel-2 imagery. *Science of Remote Sensing* 2 (2020), 100010.

[12] Han Zhai, Hongyan Zhang, Liangpei Zhang, and Pingxiang Li. 2018. Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery. *ISPRS journal of photogrammetry and remote sensing* 144 (2018), 235–253.

Received 31 Januari 2025