

# UnCRtainTS: Uncertainty Quantification for Cloud Removal in Optical Satellite Time Series

Patrick Ebel\*

patrick.ebel@tum.de

Vivien Sainte Fare Garnot<sup>†</sup>

vsaint@ics.uzh.ch

Michael Schmitt<sup>‡</sup>

michael.schmitt@unibw.de

Jan Dirk Wegner<sup>†</sup>

jandirk.wegner@uzh.ch

Xiao Xiang Zhu\*

xiaoxiang.zhu@tum.de

\* Technical University of Munich <sup>†</sup> University of Zurich <sup>‡</sup> University of the Bundeswehr Munich

## Abstract

Clouds and haze often occlude optical satellite images, hindering continuous, dense monitoring of the Earth's surface. Although modern deep learning methods can implicitly learn to ignore such occlusions, explicit cloud removal as pre-processing enables manual interpretation and allows training models when only few annotations are available. Cloud removal is challenging due to the wide range of occlusion scenarios—from scenes partially visible through haze, to completely opaque cloud coverage. Furthermore, integrating reconstructed images in downstream applications would greatly benefit from trustworthy quality assessment. In this paper, we introduce UnCRtainTS, a method for multi-temporal cloud removal combining a novel attention-based architecture, and a formulation for multivariate uncertainty prediction. These two components combined set a new state-of-the-art performance in terms of image reconstruction on two public cloud removal datasets. Additionally, we show how the well-calibrated predicted uncertainties enable a precise control of the reconstruction quality.

## 1. Introduction

Multispectral, optical satellite imagery allows for large-scale assessments of the environment like crop monitoring [58, 71] and global vegetation height estimation [45, 46]. Clouds, haze and other atmospheric disturbances, however, often occlude large parts of optical satellite images, particularly during meteorological winter season [40] and over landcover such as rainforests [4]. Neural networks trained on extensive amounts of annotated data may implicitly learn to ignore task-irrelevant cloudy observations [55, 58, 59]. Yet, explicit cloud removal as a pre-processing step can further improve model performance and is valuable if ground

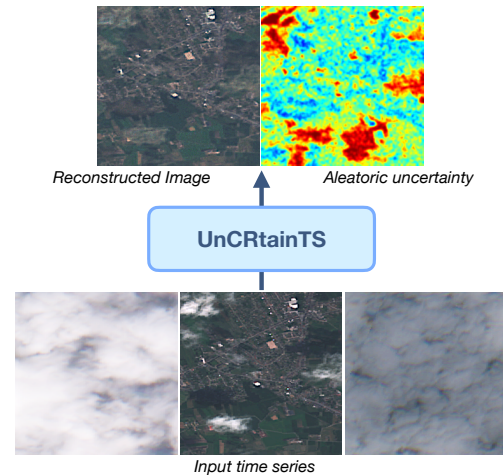


Figure 1. **Overview:** Our attention-based UnCRtainTS architecture predicts a single cloud-free image from a sequence of cloudy observations. For each reconstructed pixel, our method also estimates the aleatoric uncertainty of the prediction. Note how higher uncertainties (in red) are associated with persistent occlusion, cloud shadow, or with specific land cover types.

truth annotations for supervised training are scarce [30]. Cloud removal prior to training or applying a pre-trained task-specific model also permits a seamless analysis using traditional non-learning methods or visualisation [51].

Hence, cloud removal is an active field of research boasting a large body of literature on image reconstruction methods to recover cloud-free observations [4, 12, 17, 20, 29, 54, 61, 62]. Such methods are typically evaluated in terms of image restoration metrics, e.g. mean squared error or structural similarity (SSIM), providing an aggregated measure of reconstruction quality. These metrics, however, provide little insight into how reliable a given reconstruction is on a pixel-wise or image-by-image basis. To address this shortcoming, we introduce uncertainty estimation to satellite im-

age reconstruction, specifically to the task of multi-temporal cloud-removal in optical satellite images. Predicting uncertainties that correlate with the empirical errors of a neural net is at the core of the growing field of probabilistic deep learning [39, 65, 68]. By modelling the uncertainty and training for a negative log likelihood (NLL) objective, such approaches allow to jointly learn a model for making a prediction and estimate the prediction’s variances. If well-calibrated, the predicted uncertainties can be very valuable for downstream usage by providing a measure of a reconstruction’s confidence. Uncertainty quantification has been successfully applied in univariate remote sensing regression problems such as canopy height regression [46] or flood risk estimation [8]. Here, we extend uncertainty quantification to multivariate regression for satellite image reconstruction. We obtain experimentally well-calibrated uncertainties that enable flagging poorly reconstructed images. We also show that multivariate uncertainty prediction requires a multivariate uncertainty model for better calibration.

Aleatoric uncertainty prediction implies training with a pixel-based Negative Log Likelihood (NLL) loss. On the other hand, image reconstruction losses like SSIM or perceptual loss are typically used in existing cloud removal methods to better retrieve high-frequency details [10, 12, 74]. Here, we introduce a novel neural architecture that operates on feature maps at full resolution. It leverages attention-based temporal encoding, allowing it to outperform previous state-of-the-art approaches even when trained via a pixel-based loss. In sum, our contributions are:

- We introduce multivariate uncertainty quantification to the task of multispectral satellite image reconstruction, to obtain both reconstructions and variance estimates.
- We propose a novel neural network architecture achieving state-of-the-art results on two challenging benchmark datasets for optical satellite cloud removal.
- We obtain well-calibrated uncertainties that allow to measure and control the quality of reconstructed images for risk-mitigation in downstream applications.

## 2. Related Work

### 2.1. Cloud Removal in Satellite Image Time Series

Optical satellite image reconstruction [64], and specifically cloud removal, pose a long-standing challenge in remote sensing [15, 33, 35, 49, 50]. Contemporary deep learning approaches can be categorised into mono-temporal [4, 17, 20, 56, 75], mono-temporal & multi-modal [12, 29, 54], multi-temporal [61] and multi-temporal & multi-modal methods [14, 62]. Here, we consider the reconstruction task in a multi-temporal & multi-modal setting.

Spatial encoding of image reconstruction is either done with UNet-like encoder-decoder backbones [37, 57, 76] that

spatially down-sample the intermediate representations [12, 17, 29], or with architectures preserving the full resolution of the images [44, 54]. While the first are computationally more efficient especially in the multi-temporal setting, the latter tend to better preserve the spatial structure in the reconstructed images. In fact, downsampling architectures often necessitate auxiliary perceptual [12, 13, 36, 38] or structural similarity losses [72, 73] to recover high-frequency information. The combination of such cost functions with a probabilistic training objective for uncertainty prediction is not straightforward. Therefore, we design an architecture that operates on full resolution feature maps and make design choices to reduce its computational complexity. For temporal encoding, we draw inspiration from recent work in satellite time series encoding [21, 22, 59] and rely on self-attention to integrate the temporal information.

### 2.2. Uncertainty Quantification

Uncertainty can be partitioned into *epistemic* or model uncertainty, and *aleatoric* or data uncertainty. Epistemic uncertainty accounts for the uncertainty on the model’s weights, and can be estimated for instance with ensemble methods [43, 70], or monte-carlo dropout [19] in deep nets. Aleatoric uncertainty captures the randomness inherent to the data. In the case of optical satellite image reconstruction, aleatoric uncertainty may thus help flagging restorations based on too little evidence. In the recent deep learning literature, aleatoric uncertainty estimation is achieved via likelihood maximization with a parametric model of the noise distribution [1, 63, 65, 67, 68]. This is a common technique in safety-critical applications, such as solving inverse problems in biomedical imaging [2, 5, 9, 16, 27, 47, 48, 69]. Uncertainty quantification is of growing interest in remote sensing [26], with applications to forest assessments, flood hazard monitoring, geophysical modeling, landcover classification and out-of-distribution detection [8, 24, 25, 45, 46, 52]. As prior remote sensing work covers uncertainty quantification for univariate regression problems, the multivariate extension has yet to be explored. To our knowledge, the aforementioned contributions are either on image reconstruction in the biomedical domain or target specific remote sensing downstream tasks, such that ours is the first work to investigate uncertainty quantification for multispectral satellite image reconstruction. The current lack of uncertainty quantification in the cloud removal literature is a significant research gap because reconstructed satellite images may guide safety-critical downstream applications or human judgement alike, such that pixel-wise measures of confidence would be beneficial.

## 3. Methods

We follow the problem statement of the public cloud removal benchmark SEN12MS-CR-TS [14]. Each sam-

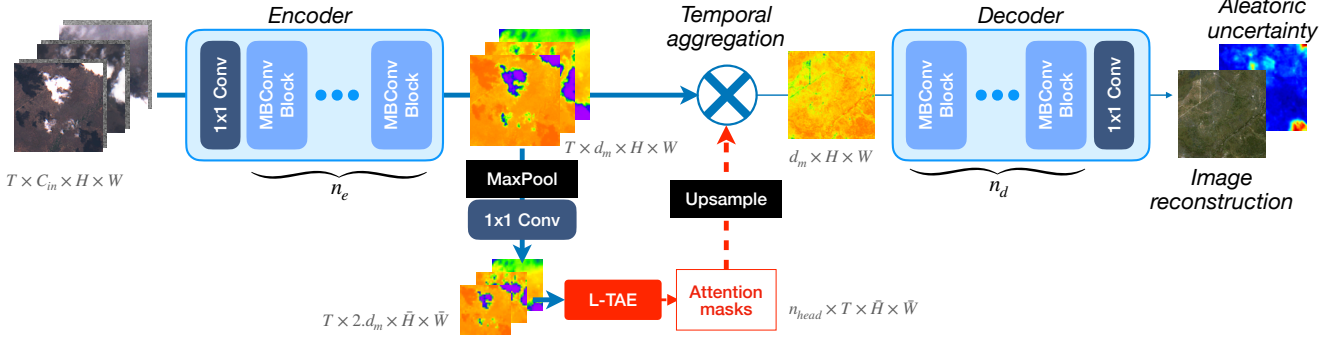


Figure 2. **UnCRtainTS**. The network consists of three main parts, applied along a main branch of MBConv blocks [60] that is processing feature maps at full input resolution: First, an *encoder* is applied in parallel to the  $T$  time points. Then, an *attention-based temporal aggregator* computes attention mask by applying an L-TAE to downsampled feature maps, used to aggregate the sequence of observations. Finally, the temporally integrated feature map is processed by a *decoding block*, yielding the image reconstruction and aleatoric uncertainty.

ple  $i$  of the  $N$ -sized dataset consists of a pair  $(\mathbf{X}^i, Y^i)$ , where  $\mathbf{X}^i = [X_1^i, \dots, X_T^i]$  is the input time series of size  $[T \times C_{in} \times H \times W]$  containing cloudy pixels, and  $Y^i$  is the target cloud-free image of shape  $[K \times H \times W]$ .  $T$  denotes the number of dates in the input sequence,  $C_{in}$  and  $K$  the number of input and output channels, and  $H \times W$  the two spatial dimensions of the images. As in [14], we set  $T = 3$ ,  $C_{in} = 15$ ,  $K = 13$ ,  $H = W = 256$ . Note that  $C_{in} \neq K$  because Sentinel-1 radar observations are utilized as additional input. Furthermore, aleatoric uncertainty quantification introduces additional output channels to describe the modeled noise distribution. For convenience, we drop the  $i$  superscript in the rest of this section.

### 3.1. Network Architecture

Our proposed UnCRtainTS network architecture maps a cloudy input time series to a single cloud-free optical image. As explained in Sec. 2.1, we make the explicit choice to perform spatial encoding only on full-resolution feature maps to allow for good performance when training with a pixel-based loss. To ease the impact of this choice on the computational load of the architecture, we rely on efficient MBConv blocks [60]. They combine depthwise convolution and regular pointwise convolutions for computationally efficient spatial encoding. We perform temporal encoding on downsampled feature maps via the attention-based L-TAE [21], which is designed for satellite image time series and computationally more efficient than transformers. The network architecture is illustrated in Fig. 2 and further described in the following paragraphs.

**Pre-aggregation shared encoder** The  $T$  different input images are processed in parallel by a shared spatial encoding branch. This encoder is composed of a pointwise convolution  $C_{in} \rightarrow d_m$ , followed by a specifiable number  $n_e$  of MBConv blocks. Following [22] we use group normal-

isation in the encoding branch. All MBConv blocks map to  $d_m \rightarrow 2 \times d_m \rightarrow d_m$  channels and contain Squeeze-Excitation layers [34]. Ultimately, each input image  $X_t$  is mapped to a feature map  $f_t$  of the same resolution.

**Attention-based temporal aggregation** Following recent literature, we employ self-attention to aggregate a sequence of feature maps  $[f_1, \dots, f_T]$  into a single one. We first down-sample features  $f_t$  with a single max-pooling operation to low resolution feature maps  $\hat{f}_t$  of size  $[d_m \times \bar{H} \times \bar{W}]$ . We set  $\bar{H} = \bar{W} = 32$ , to limit computation while providing sufficient resolution to group cloudy pixels, which typically cluster in space. We re-project the downsampled features via a linear layer  $d_m \rightarrow 2 \times d_m$ . Next, as in [22], the low-resolution features  $\hat{f}_t$  are processed pixel-wise with an L-TAE [21, 23]: we obtain attention masks over the  $T$  observations for each pixel position of the low resolution feature maps. Contrary to previous work, we only use the L-TAE’s attention masks, and omit attention-weighting of the sequence of low resolution feature maps. We upsample the attention masks to the full resolution via bilinear interpolation, and apply them to the sequence of high resolution feature maps  $[f_1, \dots, f_T]$ . This results in a single feature map  $\hat{f}$  of shape  $[d_m \times H \times W]$ . We use a dropout rate of 0.1 on the attention masks after upsampling, and the temporal aggregation is done with L-TAE’s channel grouping strategy [21].

**Post-aggregation decoding** The temporally aggregated feature map  $\hat{f}$  is processed by a decoding branch, which consists of a specifiable number  $n_d$  of batch-normalized MBConv blocks and a final  $d_m \rightarrow C_{out}$  pointwise convolution followed by a non-linearity. For every channel predicting image reconstruction, we use a sigmoidal function to squash the outputs into the data’s valid range. For channels predicting aleatoric uncertainty (see next section), we use a

softplus activation to ensure positivity, as in [32, 63, 67].

### 3.2. Aleatoric uncertainty prediction

Here, we explain how our UnCRtainTS method predicts an aleatoric uncertainty value for each reconstructed pixel. As UnCRtainTS is trained with pixel-wise losses, we henceforth adopt a pixel-based notation. We consider the set of pixels of cardinal  $n$  contained in the dataset. We denote each pixel reconstruction by  $\hat{\mathbf{y}}_j$  and the corresponding ground truth by  $\mathbf{y}_j$ , both vectors of dimension  $K$ .

**Image reconstruction** In the default setting of satellite image reconstruction, the network only regresses the target pixel values. Hence, in this setting,  $C_{out} = K$  and the predictions are typically supervised with L2 loss [3, 11]:

$$\mathcal{L}_2(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{n} \sum_{j=1}^n \|\hat{\mathbf{y}}_j - \mathbf{y}_j\|_2^2. \quad (1)$$

**Multivariate negative log-likelihood loss** Predicting aleatoric uncertainty assumes a parametric noise distribution with a likelihood function. We then optimise the likelihood of the observed data as a function of the input and the distribution’s parameters, using a negative log-likelihood (NLL) cost function [6]. Following the literature [39], we model aleatoric uncertainty on the reconstructed pixel with a  $K$ -variate Normal distribution centered at the predicted value  $\hat{\mathbf{y}}_j$  and with positive definite covariance matrix  $\Sigma$ :

$$\mathcal{N}(\mathbf{y}_j | \hat{\mathbf{y}}_j, \Sigma) = \frac{1}{\sqrt{|\Sigma|} (2\pi)^{\frac{K}{2}}} \exp\left(-\frac{1}{2} \|\hat{\mathbf{y}}_j - \mathbf{y}_j\|_M\right), \quad (2)$$

with  $\|\cdot\|_M$  the Mahalanobis distance, defined as:

$$\|\hat{\mathbf{y}}_j - \mathbf{y}_j\|_M = (\hat{\mathbf{y}}_j - \mathbf{y}_j)^T \Sigma^{-1} (\hat{\mathbf{y}}_j - \mathbf{y}_j). \quad (3)$$

Subsequently, the negative log likelihood loss writes as:

$$\mathcal{L}_{NLL}(\mathbf{y}_j | \hat{\mathbf{y}}_j, \Sigma) \propto \sum_{j=1}^n \log(|\Sigma_j|) + \|\hat{\mathbf{y}}_j - \mathbf{y}_j\|_M. \quad (4)$$

Fitting a multivariate distribution raises the question of whether a full description of the covariance matrix should be pursued or if any structural constraints on  $\Sigma$  are preferable. NLL optimization does become notoriously difficult when involving full covariance matrices [63, 65].

**Diagonal covariance matrix** We define  $\Sigma$  as a diagonal matrix with diagonal elements  $\sigma^2 = (\sigma_1^2, \dots, \sigma_K^2)$ . This greatly simplifies the inverse and determinant computations in Eq. 4. The diagonal model allows for different variance predictions per channel, which we experimentally find

to be beneficial. However, cross-channel interactions in aleatoric predictions are not captured under this assumption, and such modelling is left for further research. To predict the variances, we set  $C_{out}$  to  $2 \times K = 26$ . The diagonal entries of  $\Sigma$  serve as aleatoric uncertainty prediction for the corresponding output channel:

$$\mathbf{u}_j = [u_j^1, \dots, u_j^K] = [\sigma_1^2, \dots, \sigma_K^2]. \quad (5)$$

## 4. Experiments

### 4.1. Data

We conduct our experiments on the SEN12MS-CR [12] and SEN12MS-CR-TS [14] datasets for mono-temporal and multi-temporal cloud removal. Both are challenging image reconstruction benchmark datasets with about 50% cloud coverage over regions distributed across the whole planet and all seasons. The datasets contain ground range detected dual-polarization C-band  $S1$  measurements as well as co-registered level-1C top-of-atmosphere reflectance  $S2$  products, curated from Google Earth Engine [28] and subsequently handled as documented in the two associated publications. The mono-temporal dataset contains 169 regions, whereas SEN12MS-CR-TS focuses on a global subset of 53 large areas. All regions of the datasets are utilized for training, validation and testing, with the respective splits as originally defined. Unless specified otherwise, experiments on SEN12MS-CR-TS are run on  $T = 3$  time points, which is a reasonable number of revisits for the cloud removal task and has been a prevalent choice in prior work [14, 61, 62]. All data are of spatial dimensions  $H = W = 256$  px and we use the full spectrum of all 13 optical bands. Analogous to preceding studies combining information of SAR and optical imagery [14, 15, 35, 54, 75] we use both Sentinel-1 and Sentinel-2 data to reconstruct images of the latter (i.e.,  $C_{S1} = 2$ ,  $C_{S2} = C_{out} = 13$ , and  $C_{in} = C_{S1} + C_{S2} = 15$ ).  $S1$  data are preprocessed as in [12, 14] and  $S2$  pixel-values are divided by 1000. Finally, binary cloud masks are calculated via s2cloudless [77]—a lightweight and commonly deployed cloud detector [7, 66]. The cloud masks are used for sampling cloud-free target images at train time, statistical evaluations of results, and in prior work for losses that are cloud-sensitive [54].

### 4.2. Implementation details

**Architectures** We train the proposed UnCRtainTS in its default setting with  $n_e = 1$  pre- and  $n_d = 5$  post-aggregation MBConv blocks. The input convolution maps to  $d_m = 128$  channels, so that MBConv blocks map to  $128 \rightarrow 256 \rightarrow 128$  channels with the default expansion factor 0.25 in their Squeeze-Excitation layers. The L-TAE’s parameters are kept to their default values  $n_{head} = 16$ , and key dimension  $d_k = 4$ . For mono-temporal considera-



tions, we use the same architecture and simply discard the unnecessary L-TAE-based aggregation. We compare our architecture against the baselines already evaluated on the SEN12MS-CR [12] and SEN12MS-CR-TS [14] datasets. We also evaluate the performance of U-TAE [22] a state-of-the-art satellite image time series encoder, using the official implementation with minor adaptations to our task <sup>1</sup>.

**Training** To assess the contribution of uncertainty modelling we train two variants: *UnCRtainTS - no  $\sigma$* , trained with L2 loss only, i.e., without uncertainty prediction, and *UnCRtainTS* trained with the NLL loss of Eq. 4 predicting uncertainties together with the reconstructed image. We use the ADAM optimizer [41] with an initial learning rate of 0.001, at a batch size of 4 as in [22]. All models are trained for 20 epochs with an exponential learning rate decay of 0.8, such that the rate decays by roughly one order of magnitude every 10 epochs. Models are evaluated on the validation split each epoch and the checkpoint with best validation loss is used for testing.

**Evaluation** For image reconstruction performance, we report the Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) as well as Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM) [73] and the Spectral Angle Mapper (SAM) metric [42]. We assess the quality of the uncertainty predictions via Uncertainty Calibration Error (UCE) [31]

$$UCE(e, u) = \sum_{p=1}^P \frac{N_p}{N} |e(B_p) - u(B_p)|, \quad (6)$$

where  $e(B_p)$  denotes the RMSE of  $N_p$  pixel predictions in bin  $B_p$ ,  $P = 20$  is the bin count and a bin's uncertainty  $u(B_p)$  is given in terms of Root Mean Variance (RMV):

$$u(B_p) = \sqrt{\frac{1}{N_p} \sum_{j \in B_p} \frac{1}{K} \sum_{k=1}^K u_j^k}. \quad (7)$$

UCE quantifies the deviation between the predicted uncertainty and the empirical reconstruction error. Low UCE corresponds to well-calibrated uncertainties. We also report a patch-wise calibration metric termed  $UCE_{im}$ , where RMSE and RMV are spatio-spectrally averaged across all pixels of a given image before calculating calibration.

### 4.3. UnCRtainTS

In this section we show the experimental performance of our approach, both in terms of image reconstruction and aleatoric uncertainty prediction.

Table 1. **Multi-temporal image reconstruction experiment.** We evaluate models for  $T = 3$  inputs on SEN12MS-CR-TS benchmark. UnCRtainTS outperforms all learnable approaches on every metric, and performs best on all measures while predicting well calibrated uncertainties (bottom table).

Model	↓ RMSE	↑ PSNR	↑ SSIM	↓ SAM
least cloudy	0.079	—	0.815	12.204
DSen2-CR [54]	0.060	26.04	0.810	12.147
STGAN [61]	0.057	25.42	0.818	12.548
CR-TS Net [14]	0.051	26.68	0.836	10.657
U-TAE [22]	0.051	27.05	0.849	11.649
UnCRtainTS - no $\sigma$ (ours)	<b>0.049</b>	27.23	0.859	10.168
<b>UnCRtainTS</b> (ours)	0.051	<b>27.84</b>	<b>0.866</b>	<b>10.160</b>

	$UCE_{im}$	UCE
UnCRtainTS (ours)	0.010	0.007

**Multi-temporal image reconstruction** We benchmark our method against established heuristics and baselines of [14, 22, 54, 61]. We report the performance of these methods in Table 1. UnCRtainTS sets a new state-of-the-art performance in terms of PSNR, SSIM, and SAM. Our architecture trained without uncertainty prediction (UnCRtainTS - no  $\sigma$ ) scores second best on all those metrics and first in RMSE. This shows that our neural architecture alone outperforms existing approaches, and uncertainty prediction further improves the reconstruction performance. Compared to U-TAE, the architecture improves by 1pt SSIM while the uncertainty prediction increases the performance by another 0.7pt. Note that uncertainty prediction has a slightly detrimental impact on RMSE performance ( $-0.002$ ). This is in line with recent evidence that NLL optimization involves a trade-off between mean and variance estimate optimization that may hinder regression performance [63, 65]. However this does not impact the image similarity metrics. Lastly, in terms of parameter efficiency, our model counts 0.5M parameters. For comparison, the competitive U-TAE baseline [22] which performs third-best consists of 1.2M trainable weights, such that UnCRtainTS is relatively lightweight.

**Aleatoric uncertainty prediction** We show the uncertainty calibration metrics of our method at image and pixel level in Table 1. Those values should be compared to the test RMSE: at the pixel (resp. image) level the average error made on the reconstruction uncertainty is around 7 (resp. 5) times smaller than the average reconstruction error, showing satisfactory calibration. In other words, our method predicts uncertainty values that correlate well with the empirical reconstruction error. To demonstrate how uncertainty predictions can be useful in practice, we show how they allow filtering bad predictions. We rank all reconstructed images of the test set sorted by increasing  $UCE_{im}$  and accumulate squared errors from the least to the most uncertain samples. The monotonous curve in Fig. 3 displays a linear relation

<sup>1</sup>github.com/VSainteuf/utae-paps

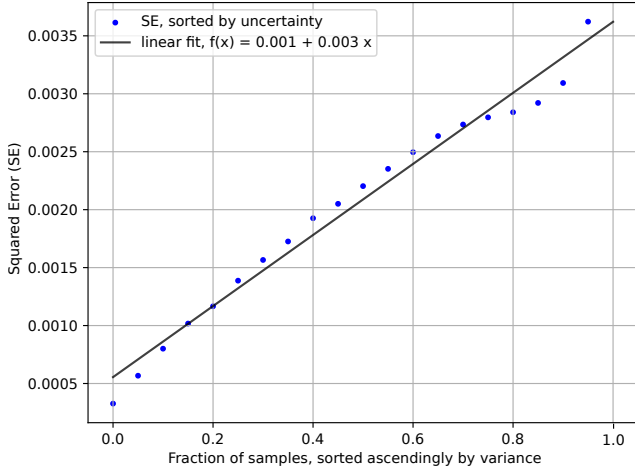


Figure 3. **Controlling error** on the test split by discarding top uncertain samples. Discarding the top 50% of uncertain reconstructions almost halves prediction error, enabling risk management.

between error and uncertainty, such that error can be step-wise decreased by uncertainty-based filtering. In practice, this enables controlling risk in downstream applications on the restored satellite images.

#### 4.4. Architecture design

To support the previous results and our architecture design choices, we systematically investigate UnCRtainTS’ hyper-parameter sensitivity. Here, all model instances are trained with L2 loss only. Because UnCRtainTS operates on feature maps at full resolution, computational complexity is an important design criterion. In addition to its image reconstruction metrics, we report each model’s number of trainable parameters and Floating Point Operations Per Second (GFLOPS), estimated via FAIR’s *fvcore* package [18].

Table 2. **Block setup.** Evaluation of the UnCRtainTS backbone for varying numbers of pre- and post-aggregation MBCConv blocks.

MBCConv		params (k)	GFLOPS	↓ RMSE	↑ PSNR	↑ SSIM	↓ SAM
$n_e$	$n_d$						
1	3	400	29.3	0.052	27.03	0.859	11.614
1	4	483	34.0	0.050	27.00	0.851	11.771
1	5	568	38.7	<i>0.049</i>	27.23	0.859	<b>10.168</b>
1	6	654	43.4	0.050	<b>27.55</b>	<i>0.860</i>	10.471
1	7	740	48.1	0.049	27.21	0.859	<i>10.300</i>
0	5	483	24.6	0.052	26.97	0.853	11.002
1	5	568	38.7	<i>0.049</i>	27.23	0.859	<b>10.168</b>
2	5	654	52.9	<b>0.048</b>	<b>27.55</b>	<b>0.864</b>	10.641

**Spatial processing** We explore the influence of the number of MBCConv blocks before ( $n_e$ ) and after ( $n_d$ ) temporal aggregation in Table 2. Using  $n_e = 2$  blocks in the encoder instead of one, brings a 0.5pt increase in SSIM, while

the performance gain is marginal on the three other metrics. More pressing, due to the parallel processing of the input sequence of feature maps, this setup incurs the highest computational complexity of 52.9 GFLOPS. In terms of post-aggregation blocks, performance peaks around 5 – 6 modules, with 5 modules being best on one metric and a close second on two more. For these reasons we choose  $n_e = 1$  pre and  $n_d = 5$  post aggregation blocks as default configuration. We also note that the ( $n_e = 0$ ) model performs competitively while being very lightweight and directly aggregating the input features. Indeed, it performs comparable to the U-TAE baseline. This secondary result shows that competitive performance can be obtained with very light architectures.

Table 3. **Head count.** Quantitative evaluation of the UnCRtainTS backbone with varying number of self-attention heads.

$n_{head}$	params (k)	GFLOPS	↓ RMSE	↑ PSNR	↑ SSIM	↓ SAM
1	556	38.7	<b>0.049</b>	<b>27.56</b>	0.856	10.497
4	559	38.7	0.052	27.40	0.856	10.825
8	563	38.7	0.051	27.00	0.851	11.131
16	568	38.7	<b>0.049</b>	27.23	<i>0.859</i>	<b>10.168</b>
32	588	38.8	0.051	27.12	<b>0.861</b>	<i>10.245</i>
64	621	38.9	0.051	27.24	0.858	11.054

**Temporal aggregation** Second, we explore the effect of the number of attention heads on the reconstruction quality. Table 3 shows that performances are closeby and differences in computational costs are negligible. We opt for 16 heads, in line with the literature [22].

**Mono-temporal image reconstruction** To validate our resolution-preserving network design, we re-train and evaluate UnCRtainTS on the mono-temporal SEN12MS-CR dataset for cloud removal. That is, we consider the special case of  $T = 1$  to investigate the model’s spatio-spectral restoration qualities and benchmark against the competitive baselines of [4, 17, 20, 29, 54, 56, 75]. Albeit being primarily designed for time series cloud removal, UnCRtainTS achieves best performances on all metrics except for SSIM, where it ranks second best following the recently published mono-temporal vision transformer architecture of [75]. The competitive performance achieved by the spatial encoding part of our architecture supports our choice of relying on MBCConv blocks operating on full resolution feature maps.

#### 4.5. Uncertainty Modelling

In this section, we provide additional experiments and ablations on the uncertainty prediction part of our method.

**Comparison of covariance models** UnCRtainTS predicts aleatoric uncertainties using a diagonal covariance

Table 4. **Mono-temporal image reconstruction experiment.** Evaluation of models for  $T = 1$  inputs on the SEN12MS-CR benchmark. UnCRtainTS is best on all metrics except SSIM, where it is second following the recent vision transformer of [75].

Method	↓ MAE	↑ PSNR	↑ SSIM	↓ SAM
McGAN [17]	0.048	25.14	0.744	15.676
SAR-Opt-cGAN [29]	0.043	25.59	0.764	15.494
SAR2OPT [4]	0.042	25.87	0.793	14.788
SpA GAN [56]	0.045	24.78	0.754	18.085
Simulation-Fusion GAN [20]	0.045	24.73	0.701	16.633
DSen2-CR [54]	0.031	27.76	0.874	9.472
GLF-CR [75]	<i>0.028</i>	<i>28.64</i>	<b>0.885</b>	<i>8.981</i>
UnCRtainTS (ours)	<b>0.027</b>	<b>28.90</b>	<i>0.880</i>	<b>8.320</b>

Table 5. **Uncertainty models.** Evaluation of different uncertainty models and of two ensembles of 5 UnCRtainTS instances (bottom), with and without SAR measurements as auxiliary input data.

model	↓ RMSE	↑ PSNR	↑ SSIM	↓ SAM	↓ UCE <sub>im</sub>	↓ UCE
isotropic $\Sigma$	0.053	26.74	0.842	11.77	0.029	0.023
UnCRtainTS	0.051	27.84	0.866	<b>10.16</b>	<b>0.010</b>	<i>0.007</i>
ensemble	<i>0.049</i>	<b>28.19</b>	<b>0.872</b>	<i>10.18</i>	<i>0.012</i>	<b>0.002</b>
ensemble <sub>noSAR</sub>	<b>0.048</b>	27.97	0.869	10.76	0.018	0.014

model, enabling different uncertainty predictions across channels. Here, this choice is compared to the simpler option of an isotropic covariance model. In the isotropic setting, we model the covariance matrix as  $\Sigma = \sigma^2 \mathbf{I}_K$  where  $\sigma^2$  is scalar and  $\mathbf{I}_K$  the  $K$ -dimensional identity matrix. This model assumes that the aleatoric uncertainty across channels can be described with a single value. We compare the performance of those two methods in Table 5. The diagonal matrix model is best overall, outperforming on all metrics. These results clearly demonstrate that uncertainty prediction for satellite image reconstruction requires channel-specific uncertainty predictions. Indeed, modeling a diagonal covariance matrix over a simplistic isotropic description entails a three-fold reduction of the final uncertainty calibration error.

**Combined epistemic and aleatoric modelling** To give a full picture of uncertainty, we complement aleatoric uncertainty modelling with epistemic uncertainty estimation. We re-train the diagonal model with different weight initializations and samples of training batches to obtain a deep ensemble of  $M = 5$  member networks [43]. The members’ reconstructions and uncertainty predictions are averaged via:

$$\hat{\mathbf{y}}^M = \frac{1}{M} \sum_{m=1}^M \hat{\mathbf{y}}^m \quad (8)$$

$$(\sigma^M)^2 = \frac{1}{M} \sum_{m=1}^M (\sigma^m)^2 + \frac{1}{M} \sum_{m=1}^M (\hat{\mathbf{y}}^m)^2 - (\hat{\mathbf{y}}^M)^2 \quad (9)$$

to obtain the ensemble reconstruction  $\hat{\mathbf{y}}^M$  and total uncertainty  $(\sigma^M)^2$ . As shown on Table 5, the 5-member ensemble achieves the best reconstruction performances overall. The full ensemble also achieves the best pixel-based calibration at 0.002 UCE, Deep ensembles come at a computational cost both at training and inference time, but can prove valuable for the integration in downstream applications.

Table 6. **Repeated Measures.** Evaluation of our ensemble of UnCRtainTS models with varying numbers of input time points.

input length T	↓ RMSE	↑ PSNR	↑ SSIM	↓ SAM	↓ UCE <sub>im</sub>	↓ UCE
2	0.051	27.78	0.861	10.86	0.012	0.004
3	<i>0.049</i>	<i>28.19</i>	<i>0.872</i>	<i>10.18</i>	0.012	<i>0.002</i>
4	<b>0.047</b>	<b>28.41</b>	<b>0.875</b>	<b>9.99</b>	0.013	<b>0.001</b>

**Uncertainty vs. sequence length** To evaluate the effect of the number of input time points  $T$  on performances, we perform inference with the UnCRtainTS ensemble on input time series of lengths  $T = 2, 3, 4$ . Table 6 shows that longer sequences help achieve both better image reconstruction quality and uncertainty calibration. This confirms the intuition that longer sequences, where additional samples are likely cloud-free, facilitate the restoration task and provide growing evidence for better calibration. Table 6 also underlines that the  $T = 3$  case considered in the main experiments makes for a challenging setting.

**SAR reduces uncertainty** We obtain a second ensemble trained without using SAR as auxiliary inputs, to explore the benefits of radar data. We show its performance on the bottom row of Table 5. The single-sensor ensemble achieves a considerably higher UCE at both image and pixel level. This suggests that the additional information contained in the SAR inputs is beneficial to improve the trustworthiness of the reconstructions.

**Qualitative results** Complementary to the quantitative measures, Fig. 4 shows UnCRtainTS’ image restorations and uncertainty maps across varying levels of cloud coverage. Of particular interest is the uncertainty predictions not only being sensitive to clouds and cloud shadows, but also capturing other dynamics such waves breaking on a shore or the coloring of maturing crops. UnCRtainTS attends to differences in the input time series—not entirely unlike sequence-based cloud detectors explicitly designed for spotting transients across repeated measures [53]—and then, due to their temporary nature, attributes them an elevated aleatoric uncertainty.

## 5. Conclusion

We introduced UnCRtainTS, a novel method for combining uncertainty quantification with cloud removal from



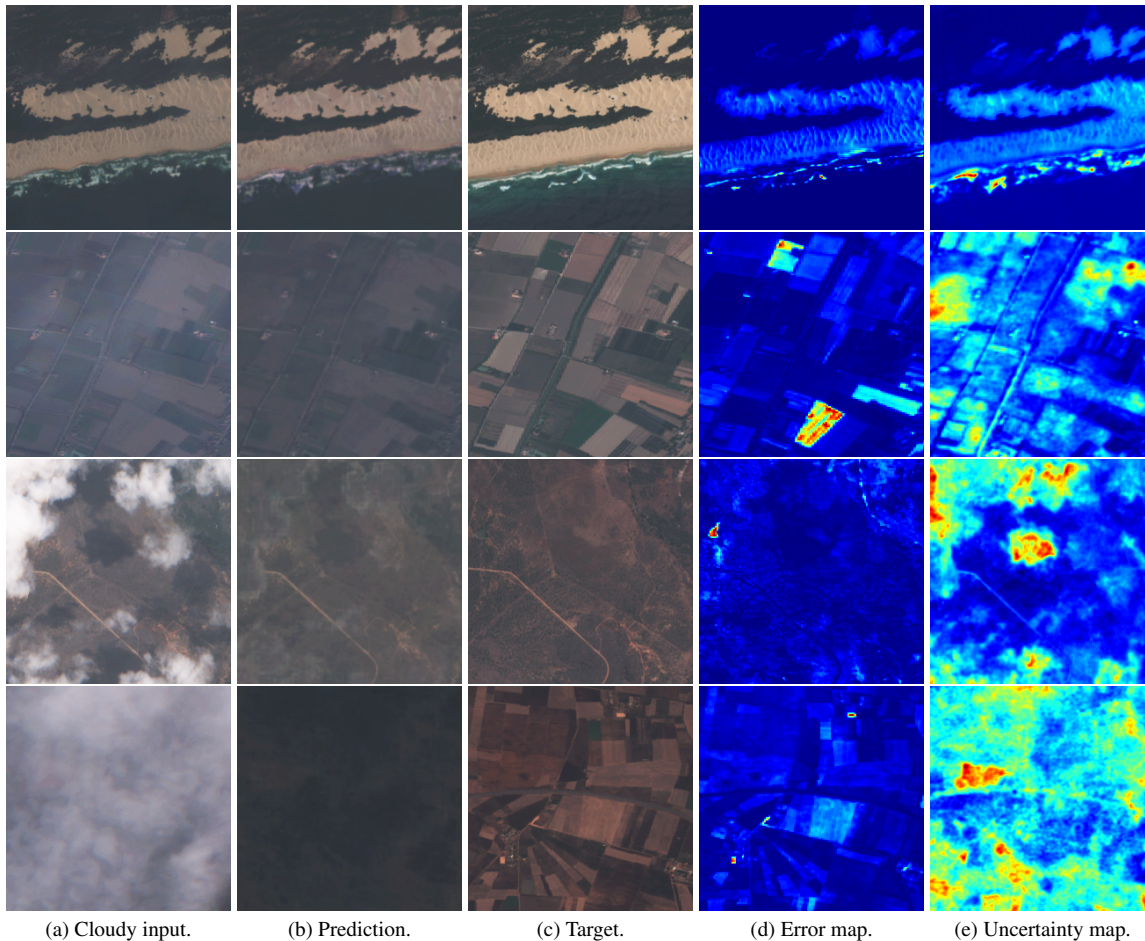


Figure 4. **Exemplary images.** Detail view on exemplary satellite images and predictions by UnCRtainTS with a diagonal covariance matrix model. Rows: Four different samples from the test split. The illustrated cases show mild atmospheric distortions, semi-transparent haze, partly dense cloud coverage and cloud coverage with no visibility at all. Columns: The input sequence’s least-cloudy image ( $T = 3$ ), UnCRtainTS’ image reconstruction, the clear-view target image, the map of squared error residuals as well as the map of UnCRtainTS’ variance predictions. Note the model’s sensitivity to transients captured in the input time series, such as the ocean’s white wash, changing crops as well as clouds and cloud shadow. UnCRtainTS captures these changing circumstances as data-inherent, aleatoric uncertainty.

optical satellite image time series. While prior contributions applied uncertainty prediction in biomedical imaging or to univariate remote sensing downstream applications, our work is the first to investigate multivariate uncertainty quantification for multispectral satellite image reconstruction. UnCRtainTS features an attention-based neural architecture that outperforms all competitors benchmarked on the satellite image reconstruction task. Our proposed method includes a formulation of aleatoric uncertainty prediction for image reconstruction based on diagonal covariance matrices, as well as an estimation of epistemic uncertainty via deep ensembles. The conducted experiments show that both of our contributions, the new architecture combined with uncertainty quantification, set a new state-of-the-art image reconstruction performance on SEN12MS-CR-TS. Finally, the outcomes highlight how our well-calibrated uncertainties can effectively serve as a measure to control re-

construction quality and help integration in risk-sensitive downstream applications. Our results encourage further explorations of more complex multivariate uncertainty models for image reconstructions. Our code is provided at [https://patrickTUM.github.io/cloud\\_removal/](https://patrickTUM.github.io/cloud_removal/).

**Acknowledgements** This work is jointly supported by the Federal Ministry for Economic Affairs and Energy of Germany in the project “AI4Sentinels– Deep Learning for the Enrichment of Sentinel Satellite Imagery” (FKZ50EE1910), by the German Federal Ministry of Education and Research (BMBF) in the framework “AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond” (01DD20001) and by the German Federal Ministry of Economics and Technology in the framework of the “national center of excellence ML4Earth” (50EE2201C).



## References

- [1] Navid Ansari, Hans-peter Seidel, Nima Vahidi Ferdowsi, and Vahid Babaei. Autoinverse: Uncertainty aware inversion of neural networks. In *Advances in Neural Information Processing Systems*, 2022. 2
- [2] Javier Antorán, Riccardo Barbano, Johannes Leuschner, José Miguel Hernández-Lobato, and Bangti Jin. Uncertainty estimation for computed tomography with a linearised deep image prior. *arXiv preprint arXiv:2203.00479*, 2022. 2
- [3] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *ACM Computing Surveys*, 2020. 4
- [4] J. D. Bermudez, P. N. Happ, D. A. B. Oliveira, and R. Q. Feitosa. SAR to optical image synthesis for cloud removal with generative adversarial networks. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2018. 1, 2, 6, 7
- [5] Sayantan Bhadra, Varun A Kelkar, Frank J Brooks, and Mark A Anastasio. On hallucinations in tomographic image reconstruction. *IEEE Transactions on Medical Imaging*, 2021. 2
- [6] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006. 4
- [7] Justin Braaten, Kurt Schwehr, and Simon Ilyushchenko. More accurate and flexible cloud masking for Sentinel-2 images. [Medium](https://medium.com/google-earth/more-accurate-and-flexible-cloud-masking-for-sentinel-2-images-766897a9ba5f). <https://medium.com/google-earth/more-accurate-and-flexible-cloud-masking-for-sentinel-2-images-766897a9ba5f>, 2020. Accessed: 2022-10-16. 4
- [8] Priyanka Chaudhary, João P Leitão, Tabea Donauer, Stefano D’Aronco, Nathanaël Perraudin, Guillaume Obozinski, Fernando Perez-Cruz, Konrad Schindler, Jan Dirk Wegner, and Stefania Russo. Flood uncertainty estimation using deep ensembles. *Water*, 2022. 2
- [9] Hyungjin Chung, Eun Sun Lee, and Jong Chul Ye. Mr image denoising and super-resolution using regularized reverse diffusion. *arXiv preprint arXiv:2203.12621*, 2022. 2
- [10] Faramarz Naderi Darbaghshahi, Mohammad Reza Mohammadi, and Mohsen Soryani. Cloud removal in remote sensing images using generative adversarial networks and sar-to-optical image translation. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 2
- [11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. 4
- [12] Patrick Ebel, Andrea Meraner, Michael Schmitt, and Xiao Xiang Zhu. Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 1, 2, 4, 5
- [13] Patrick Ebel, Michael Schmitt, and Xiao Xiang Zhu. Internal learning for sequence-to-sequence cloud removal via synthetic aperture radar prior information. In *International Geoscience and Remote Sensing Symposium*, 2021. 2
- [14] Patrick Ebel, Yajin Xu, Michael Schmitt, and Xiao Xiang Zhu. SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 2, 3, 4, 5
- [15] Robert Eckardt, Christian Berger, Christian Thiel, and Christiane Schmillius. Removal of optically thick clouds from multi-spectral satellite images using multi-frequency SAR data. *Remote Sensing*, 2013. 2, 4
- [16] Vineet Edupuganti, Morteza Mardani, Shreyas Vasanawala, and John Pauly. Uncertainty quantification in deep mri reconstruction. *IEEE Transactions on Medical Imaging*, 40(1):239–250, 2020. 2
- [17] Kenji Enomoto, Ken Sakurada, Weimin Wang, Hiroshi Fukui, Masashi Matsuoka, Ryosuke Nakamura, and Nobuo Kawaguchi. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 1, 2, 6, 7
- [18] FAIR. Flop Counter for PyTorch Models. [Github](https://github.com/facebookresearch/fvcore). <https://github.com/facebookresearch/fvcore>, 2019. Accessed: 2023-01-05. 6
- [19] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*. PMLR, 2016. 2
- [20] Jianhao Gao, Qiangqiang Yuan, Jie Li, Hai Zhang, and Xin Su. Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks. *Remote Sensing*, 2020. 1, 2, 6, 7
- [21] Vivien Sainte Fare Garnot and Loic Landrieu. Lightweight temporal self-attention for classifying satellite images time series. In *International Workshop on Advanced Analytics and Learning on Temporal Data*. Springer, 2020. 2, 3
- [22] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2, 3, 5, 6
- [23] Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. Satellite image time series classification with pixel-set encoders and temporal self-attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3
- [24] Jakob Gawlikowski, Sudipan Saha, Anna Kruspe, and Xiao Xiang Zhu. An advanced dirichlet prior network for out-of-distribution detection in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 2
- [25] Jakob Gawlikowski, Sudipan Saha, Julia Niebling, and Xiao Xiang Zhu. Robust distribution-shift aware sar-optical data fusion for multi-label scene classification. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 911–914. IEEE, 2022. 2
- [26] Jakob Gawlikowski, Cedrique Rovile Njiteucheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021. 2
- [27] Jan Glaubitz, Anne Gelb, and Guohui Song. Generalized sparse bayesian learning and application to image reconstruction. *SIAM/ASA Journal on Uncertainty Quantification*, 2023. 2

- [28] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017. 4
- [29] Claas Grohnfeldt, Michael Schmitt, and Xiaoxiang Zhu. A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images. In *IEEE International Geoscience and Remote Sensing Symposium*, 2018. 1, 2, 6, 7
- [30] Ziqi Gu, Patrick Ebel, Qiangqiang Yuan, Michael Schmitt, and Xiao Xiang Zhu. Explicit haze & cloud removal for global land cover classification. *CVPR 2022 Workshop on Multimodal Learning for Earth and Environment*, 2022. 1
- [31] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*. PMLR, 2017. 5
- [32] Ali Harakeh, Jordan Hu, Naiqing Guan, Steven L Waslander, and Liam Paull. Estimating regression predictive distributions with sample networks. *arXiv preprint arXiv:2211.13724*, 2022. 4
- [33] Gensheng Hu, Xiaoyi Li, and Dong Liang. Thin cloud removal from remote sensing images using multidirectional dual tree complex wavelet transform and transfer least square support vector regression. *Journal of Applied Remote Sensing*, 2015. 2
- [34] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3
- [35] Bo Huang, Ying Li, Xiaoyu Han, Yuanzheng Cui, Wenbo Li, and Rongrong Li. Cloud removal from optical satellite imagery with SAR imagery using sparse representation. *IEEE Geoscience and Remote Sensing Letters*, 2015. 2, 4
- [36] Jieon Hwang, Chushi Yu, and Yoan Shin. SAR-to-optical image translation using ssim and perceptual loss based cycle-consistent gan. In *International Conference on Information and Communication Technology Convergence*. IEEE, 2020. 2
- [37] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2
- [38] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*. Springer, 2016. 2
- [39] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 2017. 2, 4
- [40] Michael D. King, Steven Platnick, W. Paul Menzel, Steven A. Ackerman, and Paul A. Hubanks. Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites. *IEEE Transactions on Geoscience and Remote Sensing*, 2013. 1
- [41] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 5
- [42] Fred A Kruse, AB Lefkoff, JW Boardman, KB Heidebrecht, AT Shapiro, PJ Barloon, and AFH Goetz. The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data. In *AIP Conference Proceedings*. American Institute of Physics, 1993. 5
- [43] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017. 2, 7
- [44] Charis Lanaras, José Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018. 2
- [45] Nico Lang, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. A high-resolution canopy height model of the earth. *arXiv preprint arXiv:2204.08322*, 2022. 1, 2
- [46] Nico Lang, Nikolai Kalischek, John Armston, Konrad Schindler, Ralph Dubayah, and Jan Dirk Wegner. Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles. *Remote Sensing of Environment*, 2022. 1, 2
- [47] Max-Heinrich Laves, Sontje Ihler, Jacob F Fast, Lüder A Kahrs, and Tobias Ortmaier. Well-calibrated regression uncertainty in medical imaging with deep learning. In *Medical Imaging with Deep Learning*. PMLR, 2020. 2
- [48] Max-Heinrich Laves, Malte Tölle, and Tobias Ortmaier. Uncertainty estimation in medical image denoising with bayesian deep image prior. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*. Springer, 2020. 2
- [49] Xinghua Li, Huanfeng Shen, Liangpei Zhang, Hongyan Zhang, Qiangqiang Yuan, and Gang Yang. Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning. *IEEE Transactions on Geoscience and Remote Sensing*, 2014. 2
- [50] Chao-Hung Lin, Po-Hung Tsai, Kang-Hua Lai, and Jyun-Yuan Chen. Cloud removal from multitemporal satellite images using information cloning. *IEEE Transactions on Geoscience and Remote Sensing*, 2012. 2
- [51] Han Liu, Peng Gong, Jie Wang, Xi Wang, Grant Ning, and Bing Xu. Production of global daily seamless data cubes and quantification of global land cover change from 1985 to 2020 - imap world 1.0. *Remote Sensing of Environment*, 2021. 1
- [52] Mingliang Liu, Dario Grana, and Leandro Passos de Figueiredo. Uncertainty quantification in stochastic inversion with dimensionality reduction using variational autoencoder. *Geophysics*, 87(2):M43–M58, 2022. 2
- [53] Vincent Lonjou, Camille Desjardins, Olivier Hagolle, Beatrice Petrucci, Thierry Tremas, Michel Dejus, Aliaksei Makarau, and Stefan Auer. MACCS-ATCOR joint algorithm (MAJA). In Adolfo Comerón, Evgueni I. Kassianov, and Klaus Schäfer, editors, *Remote Sensing of Clouds and the Atmosphere XXI*. International Society for Optics and Photonics, SPIE, 2016. 7
- [54] Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael 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*, 2020. 1, 2, 4, 5, 6, 7
- [55] Nando Metzger, Mehmet Ozgur Turkoglu, Stefano D’Aronco, Jan Dirk Wegner, and Konrad Schindler. Crop classification under varying cloud cover with neural ordinary differential equations. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 1
- [56] Heng Pan. Cloud removal for remote sensing imagery via spatial attention generative adversarial network. *arXiv preprint arXiv:2009.13015*, 2020. 2, 6, 7
- [57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015. 2
- [58] Marc Rußwurm and Marco Körner. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 1
- [59] Marc Rußwurm and Marco Körner. Self-attention for raw optical satellite time series classification. *ISPRS journal of photogrammetry and remote sensing*, 2020. 1, 2
- [60] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3
- [61] Vishnu Sarukkai, Anirudh Jain, Burak Uzkent, and Stefano Ermon. Cloud removal from satellite images using spatiotemporal generator networks. In *The IEEE Winter Conference on Applications of Computer Vision*, 2020. 1, 2, 4, 5
- [62] Alessandro Sebastianelli, Erika Puglisi, Maria Pia Del Rosso, Jamila Mifdal, Artur Nowakowski, Pierre Philippe Mathieu, Flora Pirri, and Silvia Liberata Ullo. PLFM: Pixel-level merging of intermediate feature maps by disentangling and fusing spatial and temporal data for cloud removal. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 1, 2, 4
- [63] Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2021. 2, 4, 5
- [64] Huanfeng Shen, Xinghua Li, Qing Cheng, Chao Zeng, Gang Yang, Huifang Li, and Liangpei Zhang. Missing information reconstruction of remote sensing data: A technical review. *IEEE Geoscience and Remote Sensing Magazine*, 2015. 2
- [65] Nicki Skafte, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. *Advances in Neural Information Processing Systems*, 2019. 2, 4, 5
- [66] Sergii Skakun, Jan Wevers, Carsten Brockmann, Georgia Doxani, Matej Aleksandrov, Matej Batič, David Frantz, Ferran Gascon, Luis Gómez-Chova, Olivier Hagolle, et al. Cloud mask intercomparison exercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. *Remote Sensing of Environment*, 2022. 4
- [67] Andrew Stirn, Hans-Hermann Wessels, Megan Schertzer, Laura Pereira, Neville E Sanjana, and David A Knowles. Faithful heteroscedastic regression with neural networks. *arXiv preprint arXiv:2212.09184*, 2022. 2, 4
- [68] Hiroshi Takahashi, Tomoharu Iwata, Yuki Yamanaka, Masanori Yamada, and Satoshi Yagi. Student-t variational autoencoder for robust density estimation. In *International Joint Conference on Artificial Intelligence*, 2018. 2
- [69] Malte Tölle, Max-Heinrich Laves, and Alexander Schlaefer. A mean-field variational inference approach to deep image prior for inverse problems in medical imaging. In *Medical Imaging with Deep Learning*. PMLR, 2021. 2
- [70] Mehmet Ozgur Turkoglu, Alexander Becker, Hüseyin Anil Gündüz, Mina Rezaei, Bernd Bischl, Rodrigo Caye Daudt, Stefano D’Aronco, Jan Dirk Wegner, and Konrad Schindler. FiLM-ensemble: Probabilistic deep learning via feature-wise linear modulation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2
- [71] Mehmet Ozgur Turkoglu, Stefano D’Aronco, Gregor Perich, Frank Liebisch, Constantin Streit, Konrad Schindler, and Jan Dirk Wegner. Crop mapping from image time series: Deep learning with multi-scale label hierarchies. *Remote Sensing of Environment*, 2021. 1
- [72] Xiaoke Wang, Guangluan Xu, Yang Wang, Daoyu Lin, Peiguang Li, and Xiujing Lin. Thin and thick cloud removal on remote sensing image by conditional generative adversarial network. In *International Geoscience and Remote Sensing Symposium*, 2019. 2
- [73] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 2, 5
- [74] Zhaobin Wang, Yikun Ma, and Yaonan Zhang. Hybrid cgan: Coupling global and local features for sar-to-optical image translation. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 2
- [75] Fang Xu, Yilei Shi, Patrick Ebel, Lei Yu, Gui-Song Xia, Wen Yang, and Xiao Xiang Zhu. GLF-CR: SAR-enhanced cloud removal with global-local fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022. 2, 4, 6, 7
- [76] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2
- [77] Anze Zupanc. Improving cloud detection with machine learning. [Sentinel-Hub](https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13). <https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13>, 2017. Accessed: 2019-10-10. 4