# AllClear: A Comprehensive Dataset and Benchmark for Cloud Removal in Satellite Imagery

**Hangyu Zhou**[1,*] **Chia-Hsiang Kao**[1*], **Cheng Perng Phoo**[1], **Utkarsh Mall**[2], **Bharath Hariharan**[1], **Kavita Bala**[1]

[1]Computer Science, Cornell University
[2]Computer Science, Columbia University

## Abstract

Clouds in satellite imagery pose a significant challenge for downstream applications. A major challenge in current cloud removal research is the absence of a comprehensive benchmark and a sufficiently large and diverse training dataset. To address this problem, we introduce the largest public dataset — *AllClear* for cloud removal, featuring 23,742 globally distributed regions of interest (ROIs) with diverse land-use patterns, comprising 4 million images in total. Each ROI includes complete temporal captures from the year 2022, with (1) multi-spectral optical imagery from Sentinel-2 and Landsat 8/9, (2) synthetic aperture radar (SAR) imagery from Sentinel-1, and (3) auxiliary remote sensing products such as cloud masks and land cover maps. We validate the effectiveness of our dataset by benchmarking performance, demonstrating the scaling law — the PSNR rises from 28.47 to 33.87 with $30\times$ more data, and conducting ablation studies on the temporal length and the importance of individual modalities. This dataset aims to provide comprehensive coverage of the Earth's surface and promote better cloud removal results.

## 1 Introduction

Satellite image recognition enables environmental monitoring, disaster response, urban planning [Pham et al., 2011, Wellmann et al., 2020], crop-yield prediction [Doraiswamy et al., 2003], and many more applications, but is held back significantly due to occlusion by clouds. Roughly 67% of the Earth's surface is covered by clouds at any given moment [King et al., 2013]. The limited availability of cloud-free captures is especially problematic for time-sensitive events like wildfire control [Kyzirakos et al., 2014, Thangavel et al., 2023] and flood damage assessment [Rahman and Di, 2020]. Consequently, developing effective cloud removal techniques is crucial for maximizing the utility of remote sensing data in various domains.

A major challenge holding back research into cloud removal is the lack of comprehensive datasets and benchmarks. A survey of publicly available datasets for cloud removal (Table 1) reveals several problems. First, most existing datasets are sampled from a small set of locations and thus have limited geographical diversity [Ebel et al., 2020, Huang and Wu, 2022, Ebel et al., 2022], impacting both the effectiveness of training and the rigor of evaluation. Second, many existing datasets filter out very cloudy images (e.g., more than 30% cloud coverage), thus preventing trained models from tackling practical situations with extensive cloud cover [Sarukkai et al., 2020, Requena-Mesa et al.,

*Lead authors. Correspondence to : Hangyu Zhou hz477@cornell.edu, Chia-Hsiang Kao ck696@cornell.edu

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

| Dataset | Regions | # ROIs | # Images | Satellites |
|---|---|---|---|---|
| STGAN [Sarukkai et al., 2020] | Worldwide | 945 | 3,101 | Sentinel-2 |
| Sen2_MTC [Huang and Wu, 2022] | Worldwide | 50 | 13,669 | Sentinel-2 |
| EarthNet2021 [Requena-Mesa et al., 2021] | Europe | 32,000 | 960,000 | Sentinel-2 |
| SEN12MS-CR [Ebel et al., 2020] | Worldwide | 169 | 366,654 | Sentinel-1/2 |
| SEN12MS-CR-TS [Ebel et al., 2022] | Worldwide | 53 | 917,580 | Sentinel-1/2 |
| AllClear | Worldwide | 23,742 | 4,354,652 | Sentinel-1/2, LandSat-8/9 |

**Table 1:** Summary of publicly available cloud removal datasets.

2021] (Figure 1). Third, some existing benchmarks use ground-truth cloud-free images captured at a very different time point from the time the input images are captured [Sarukkai et al., 2020, Ebel et al., 2022]. This means that many changes may have occurred on the ground between the capture of the input and the target images, introducing noise in the evaluation. Finally, existing datasets incorporate a very limited set of sensors/modalities (i.e., Sentinel-2), limiting the information available to models for faithful cloud removal.

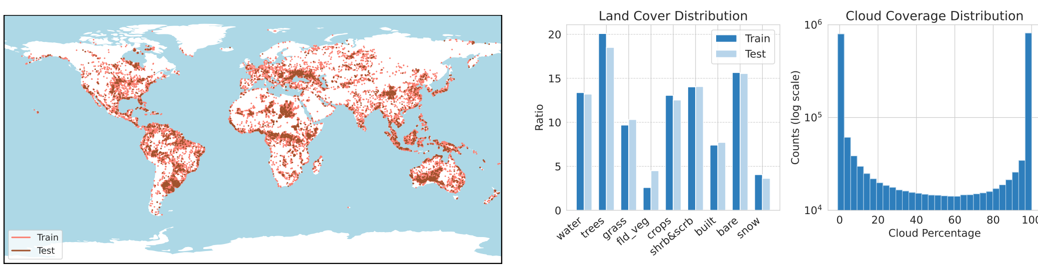


**Figure 1:** Left: Geographical distribution of *AllClear* ROIs; middle: land cover distribution of *AllClear* for training and testing set; right: cloud coverage distribution of the entire *AllClear* dataset.

To address these limitations and facilitate future research in cloud removal, we introduce the largest and most comprehensive dataset to date, *AllClear*. To ensure sufficient coverage of the planet's diversity, *AllClear* includes 23,742 regions of interest (ROIs) scattered across the globe with diverse land cover patterns, resulting in four million multi-spectral images. *AllClear* includes data from three different satellites (i.e., Sentinel-1, Sentinel-2, and LandSat-8/9) captured over a year (2022) at each ROI, allowing models to better interpolate missing information. We use this dataset to create a more rigorous sequence-to-point benchmark with more temporally aligned ground truth. Finally, besides the enormous amount of raw satellite images, we also curated a rich set of metadata for each individual image (e.g., geolocation, timestamp, land cover map, cloud masks, etc.) to support building future models for the cloud removal challenge as well as to enable stratified evaluation.

We evaluate existing state-of-the-art on AllClear and find that existing models are undertrained; training on our larger and more diverse training set significantly improves performance. We also find that models that use the full suite of available sensors as well as a longer temporal sequence of captures perform much better. Taken together, our contributions are:

- We introduce to-date the largest dataset for cloud removal, as well as a comprehensive and stratified evaluation benchmark,
- We demonstrate that our significantly larger and more diverse training set improves model performance, and
- We show empirically the importance of leveraging multiple sensors and longer time spans.

# 2 Background

## 2.1 Existing Cloud Removal Datasets

Advances in cloud removal research for satellite imagery have led to the development of several datasets with unique characteristics and limitations. STGAN introduced two cloud removal datasets and established the multi-temporal task format of using three images as input [Sarukkai et al., 2020]. However, the dataset discards all image crops with more than 30% cloud cover, leading to only 3K images. Following STGAN, Huang and Wu [2022] find that the annotations in STGAN can be incorrect and propose Sen2_MTC with four times more images. The Sen_MTC dataset first samples 50 tiles globally and proceeds to divide the large tile into pieces, restricting the sampling regional diversity. STGAN and Sen_MTC also do not describe their *data processing levels* (e.g., level-1C Top-of-Atmosphere or level-2A Surface Reflectance imagery), making it hard to compare models trained on different datasets. Different from the STGAN and Sen2_MTC datasets, the SEN12MS-CR dataset features synthetic-aperture radar (SAR) images to augment the optics imagery. However, it has a single image pair per data point. The successor is SEN12MS-CR-TS [Ebel et al., 2022], featuring multi-temporal (multiple images per location) multi-modality paired images. For each location, 30 Sentinel-1 and Sentinel-2 images from 2018 are temporally aligned and paired to form spatiotemporal patches. However, the temporal differences between the two modalities can be as large as 14 days, and the temporal difference between the input and the target can be as large as a year, resulting in noise in the evaluation. In addition, the authors construct a sequence-to-point cloud removal task in which images from this dataset with more than 50% cloud coverage are excluded. EarthNet2021 [Requena-Mesa et al., 2021] also provides sequences of carefully curated Sentinel-2 images with a spatial resolution of 20m and bands of RGB and Infrared. However, the dataset excluded spatiotemporal patches with high cloud coverage and is thus not an ideal dataset for cloud removal.

## 2.2 Cloud Removal Methodology

Early work on cloud removal used a conditional GAN to map a single image to its cloudless version conditioning on the NIR channel [Enomoto et al., 2017] or SAR images [Grohnfeldt et al., 2018]. These early attempts fall short of generalizing to real cloudy images [Ebel et al., 2020, Stucker et al., 2023]. Singh and Komodakis [2018] and Ebel et al. [2020] improve this setup by using a cycle-consistency loss. Other approaches learn the mapping from SAR images to their corresponding multi-spectral bands [Bermudez et al., 2018, 2019, Wang et al., 2019, Fuentes Reyes et al., 2019]. More recently, with the advent and rise of transformers, multi-head attention modules have been introduced for cloud removal tasks. Yu et al. [2022] casts the cloud as image distortion and designs a distortion-aware module to restore the cloud-free images. Zou et al. [2023a] utilized multi-temporal inputs along with a multi-scale attention autoencoder to exploit the global and local context for reconstruction. Ebel et al. [2023] also adopts a multi-temporal inputs and attention autoencoder but also proposes to estimate the aleatoric uncertainty of the prediction, which controls the quality of the reconstruction for risk-mitigation applications. Jing et al. [2023], Zou et al. [2023b] proposed to utilize diffusion training objective for cloud-free image generation where the inputs only rely on the optimal images and SAR imagery is not taken into consideration. Similarly but more generally, Khanna et al. [2023] proposed a generative foundation model for satellite imagery, but is not tailored for the cloud removal task.

# 3 Dataset

## 3.1 Regions-of-Interest Selection

We choose our ROIs to satisfy two objectives: (a) coverage of most of the land surface and (b) a balanced sampling of land cover types. This balanced sampling in particular ensures that smaller but more popular locations like cities are as well represented as the large swathes of wilderness. To get these ROIs, we follow a two-step procedure: curating a pool of ROI candidates and then building train/benchmark subgroups balanced across land cover types, as shown in Figure 1. This ensures both the benchmark and the training sets contain a sufficient amount of data representing various land cover types.

For curating the ROI pool, unlike previous work that followed random ROI selection [Sarukkai et al., 2020, Huang and Wu, 2022, Ebel et al., 2020, 2022, Xu et al., 2023], we use grid sampling to select an ROI every 0.1°latitude and every $0.1°\cos(\theta)$ longitude, where $\theta$ is the latitude, from 90°S to 90°N. The intuition behind this approach is that the same 0.1°longitude can represent 11.1 km at the equator and 4.35 km at 67°latitude. This weighting provides a simple yet effective method for not over-sampling high-latitude areas. By excluding ocean areas using the `GeoPandas` package, we select a total of 1,087,947 ROIs.

Next, we select ROIs from the pool to achieve a more balanced dataset over land-cover use while considering the natural imbalance of land cover distribution on the earth's surface. We leverage the land cover data from the Dynamic World product [Brown et al., 2022] from Google Earth Engine, which is a 10-meter resolution Land Use / Land Cover (LULC) dataset containing class probabilities and label information for nine classes: water, tree, grass, flooded vegetation, crops, shrub and scrub, built, bare, and snow and ice. Specifically, we calculate the all-year median of the LULC in 2022 as an estimate for the land use and land cover for each ROI. We iteratively select ROIs from the candidate pool such that the average land cover for all classes (except snow and ice) is greater than 10 percent in the benchmark set and 5 percent in the train set.

Finally, for a fairer comparison with models trained on previous datasets, we take an additional measure to exclude the ROIs that are close to the SEN12MS-CR-TS dataset [Ebel et al., 2022]. Specifically, the size of tiles in the SEN12MS-TR-CS dataset is $40 \times 40$ km$^2$. So we exclude the ROIs in AllClear that are within a 50 km radius of the ROIs in SEN12MS-CR-TS.

### 3.2 Data Preparation

*AllClear* contains three different types of open-access satellite imagery made available by the Google Earth Engine (GEE) platform [Gorelick et al., 2017]: Sentinel-2A/B [Drusch et al., 2012], Sentinel-1A/B [Torres et al., 2012], and Landsat 8/9 [Williams et al., 2006]. For Sentinel-2, we collected all thirteen bands of Level-1C orthorectified top-of-atmosphere (TOA) reflectance product. For Sentinel-1, we acquired the S1 Ground Range Detected (GRD) product with two polarization channels (VV and VH). All the raw images in *AllClear* were resampled to 10-meter resolution. We follow the default GEE preprocessing steps during all the downloading process. In addition, we include the Dynamic World Land Cover Map for all the Sentinel-2 imagery [Brown et al., 2022]. For each selected ROI, our goal is to collect all $2.56 \times 2.56$ km$^2$ patches in 2022 with a spatial resolution of 10 meters. We adopt the Universal Transverse Mercator (UTM) coordinate reference system (CRS), following Ebel et al. [2020, 2022], Zhao et al. [2023], which divides the Earth into 60 zones, each spanning 6 degrees of longitude, to ensure minimal distortion, especially along the longitude axis. Since satellite imagery is often captured in large tiles that do not necessarily conform to the boundaries of UTM zones, gaps (NaN values) can occur where the tile data does not cover the entire ROI. In such cases, we exclude all images containing NaN values to maintain data quality.

**Data Preprocessing.** For Sentinel-1, following Ebel et al. [2022], we clip the values in the VV channel of S1 to $[-25; 0]$ and those of the VH channels to $[-32.5, 0]$. For Sentinel-2 and Landsat 8/9, we clip the raw values to $[0, 10000]$ [Ebel et al., 2022, Huang and Wu, 2022]. The values are then normalized to the range of $[0, 1]$.

**Cloud and Shadow Mask Computation.** The cloud and shadow masks are indispensable to this dataset as they are used for guiding evaluation metric computation by masking out regions where there are clouds and shadows in the target images. To obtain the cloud mask, we use the S2 Cloud Probability dataset available on Google Earth Engine. This dataset is built by using S2cloudless [Zupanc, 2017], an automated cloud-detection algorithm for Sentinel-2 imagery based on a gradient boosting algorithm, which shows the best overall cloud detection accuracy on opaque clouds and semi-transparent clouds in the Hollstein reference dataset [Hollstein et al., 2016, Skakun et al., 2022] and the LCD PixBox dataset [Paperin et al., 2021, Skakun et al., 2022].

As for the shadow mask, ideally the cloud shadows can be estimated using the sun azimuth and cloud height but the latter information cannot be obtained. We therefore proceed with curating the shadow mask following documentation in Google Earth Engine [jdbcode, 2023]. The shadow is estimated by computing dark pixels and projecting cloud regions. For the dark pixels, we use the Scene Classification Map (SCL) band values from Sentinel-2 to remove water pixels, as water pixels can resemble shadows. We then threshold the NIR pixel values with a threshold of 1e-4 to create a map of dark pixels. Finally, we take the intersection of the dark pixel map and the projected cloud regions to obtain the cloud shadow masks.

### 3.3 Benchmarking Task Setup and Evaluation

For evaluation, we construct a sequence-to-point task using our AllClear dataset with train, validation, and test splits of 278,613, 14,215, and 55,317 samples, respectively. Each instance contains three input images $(u_1, u_2, u_3)$, a target clear image $(v)$, input cloud and shadow masks, target cloud and shadow masks, timestamps, and metadata such as latitude, longitude, sun elevation angle, and sun azimuth. Sentinel-2 images are considered the main sensor modality, while sensors such as Sentinel-1 and LandSat-8/9 are auxiliary. Unlike previous datasets, we do not threshold the cloud coverage in the input images Sarukkai et al. [2020], Requena-Mesa et al. [2021], Ebel et al. [2022]. We also provide multiple options for cloud and shadow masks with different thresholds for users to use.

We address two temporal misalignment problems found in previous datasets: misalignment between source and target images (where the difference can be months apart) and misalignment when pairing main sensors with auxiliary sensors (where the difference can be at most two weeks) [Ebel et al., 2022]. To avoid temporal misalignment issues, the target clear images are chosen from four consecutive spatial-temporal patches. In particular, the time stamps of the input and target images are either in the order $[u_1, v, u_2, u_3]$ or in the order $[u_1, u_2, v, u_3]$. This ensures that the target image does not include any novel or unseen changes that occurred after the capture of the cloudy images. For auxiliary sensors, we select the auxiliary satellite images within a two-day difference from the respective Sentinel-2 images. We fill the corresponding channels with ones if no auxiliary sensor images match are available. More details about the construction of these inputs and targets is in the supplementary.

Note that our target images may still have some clouds (since it is difficult to get a cloud-free image within each time span). To reach a balance between having diverse scenarios and limit metric inaccuracy, we set target images to have less than 10% cloud and shadow (combined) coverage and exclude the cloudy pixels when calculating the metrics. We modified various pixel-based metrics to compute only over the cloud-free areas. We adopt the following metrics common in cloud removal literature: mean absolute error (MAE), root mean square error (RMSE), peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM), and structural similarity index measure (SSIM).

## 4 Experiments

We next evaluate the usefulness of our dataset for both evaluation and training.

### 4.1 Benchmarking prior methods on the AllClear test set

**Selection of SoTA model architecture.** For a fair comparison between datasets, we choose among the SoTA models for comparison. Specifically, we choose prior state-of-the-art models that are pre-trained on SEN12-MS-CR-TS for the benchmark because AllClear and SEN12-MS-CR-TS are both Top-of-Atmosphere imagery and contain all the bands of Sentinel-2. Notably, other previous datasets such as STGAN and Sen2_MTC are excluded because the pre-processing methodology and imagery production type are not explicitly mentioned, making direct deployment of previous models on the AllClear dataset unfair and not comparable. Therefore, we exclude models such as CTGAN [Huang and Wu, 2022], PMAA [Zou et al., 2023a], and DiffCR [Zou et al., 2023b] which use these datasets to train. Instead, we choose UnCRtainTS model [Ebel et al., 2023], a sequence-to-point model, and U-TILISE [Stucker et al., 2023], a sequence-to-sequence model, both pre-trained on the SEN12MS-TR-CS dataset and public available, for our experiments. For this evaluation, all models receive three images as input. Specifically, they receive both Sentinel-2 and Sentinel-1 images concatenated along the channel dimension.

**Results.** The benchmark results are shown in Table 2. We first notice that simple baselines *least cloudy* and *mosaicing* perform well on the dataset. UnCRtainTS performs slightly better than these simple baselines in terms of SSIM and SAM. On the other hand, the U-TILISE model falls short of reaching the performance of the simple baselines. Since U-TILISE is a sequence-to-sequence model, we adopt it for sequence-to-point evaluation by choosing the image from the output sequence with the lowest MAE score as the model output. Notably, the training of U-TILISE involves adding sampled cloud masks to the cloud-free images as inputs, and it is trained to recover the original cloud-free sequence. The model is evaluated in a similar manner. The distribution disparity between the sampled cloud masks and the real clouds may contribute to the low score of U-TILISE in the real scenario. The good performance of *least cloudy* and *mosaicing* is intriguing. We conjecture that part of the reason may be that in AllClear, the temporal gap between input images and target images is smaller, so simply averaging or choosing from the input images is likely to yield good results.

**Table 2:** Benchmark performance of previous SoTA models evaluated on our AllClear benchmark dataset. The best performing values are in **bold** and the second best is underlined.

| Model | Training Dataset | PSNR (↑) | SSIM (↑) | SAM (↓) | MAE (↓) |
|---|---|---|---|---|---|
| Least Cloudy | - | 28.864 | <u>0.836</u> | <u>6.982</u> | 0.078 |
| Mosaicing | - | **29.824** | 0.754 | 23.58 | **0.045** |
| UnCRtainTS [Ebel et al., 2023] | SEN12MS-CR-TS | <u>29.009</u> | **0.898** | **5.972** | <u>0.039</u> |
| U-TILISE Stucker et al. [2023] | SEN12MS-CR-TS | 24.660 | 0.807 | 7.765 | 0.083 |

**Failure cases.** To understand the performance of the state-of-the-art better, we visualize the output images generated using the state-of-the-art model UnCRtainTS [Ebel et al., 2023], which was trained on the SEN12MS-CR-TS dataset [Ebel et al., 2022]. In Figure 2, we evaluate the pre-trained model on AllClear testing cases where it receives three cloudy images as input. Overall, we observe three primary failure modes in the model's performance: (1) The model fails to draw from clear input images, particularly when the other two images are cloudy. This issue may arise because the model was trained exclusively on images with less than 50% cloud coverage, as noted by the authors [Ebel et al., 2023]. (2) The model often struggles to recover the correct color spectrum, even when the input images are mostly clear. We hypothesize that this is due to the relatively small dataset size, leading to a lack of generalization ability. (3) The model frequently fails to generalize to snow-covered land. We speculate that this is due to insufficient sampling of diverse snowy regions during training.

### 4.2 Training on AllClear

**Table 3:** Benchmark Performance for UnCRtainTs models retrained on AllClear.

| Evaluation Dataset | Training Dataset (fraction used) | PSNR (↑) | SSIM (↑) | SAM (↓) | MAE (↓) |
|---|---|---|---|---|---|
| SEN12MS-CR-TS | SEN12MS-CR-TS | **27.838** | **0.866** | **9.455** | **0.036** |
| | AllClear (3.4%) | 26.256 | 0.847 | 10.411 | 0.041 |
| AllClear | SEN12MS-CR-TS | **29.009** | 0.898 | **5.972** | 0.039 |
| | AllClear (3.4%) | 28.474 | **0.906** | 6.373 | **0.036** |

We next evaluate the benefits of training on AllClear. For this purpose, we use UnCRtainTS given its good performance on prior benchmarks. To evaluate if there is any domain difference between AllClear and the previous SEN12MS-TR-CS dataset, we first run an equal-training-set-size comparison. We train UnCRtainTS on a *subset* of AllClear that is of the same size as the the training

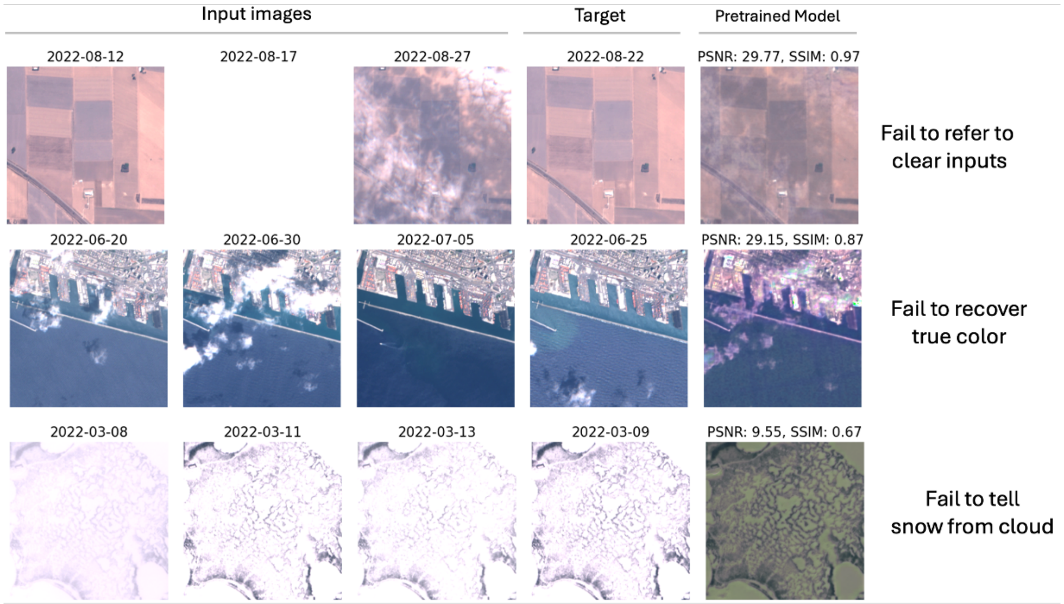


**Figure 2:** Failure case from UnCRtainTS [Ebel et al., 2023], a previous SOTA model trained on the SEN12MS-CR-TS [Ebel et al., 2022] cloud removal dataset.

set size used in UnCRtainTS training, which is 10,167 data points. We also follow the training hyperparameters as in the original paper to avoid extra tuning. As shown in Table 3, when both models are evaluated on AllClear (i.e., the bottom two rows in Table 3), we observe that UnCRtainTS models pre-trained on both datasets have comparable results across the four metrics. This suggests that there is no noticeable domain difference between the two datasets.

**Scaling with AllClear.** We next evaluate how much we can scale UnCRtainTS using the large training set available with AllClear. Specifically, we curate a dataset of various scale using random sampling from the training dataset while evaluating on the same validation set. Table 4 shows the results. We find that more training data clearly improves accuracy significantly across all metrics, resulting in a more than 10% improvement in PSNR. Figure 5 shows that with a larger dataset the model is able to better remove clouds and better preserve the color. This suggests that cloud removal models trained on past datasets are in general *undertrained* and AllClear's large training set is extremely useful to help the models fit the data better.

**Table 4:** Scaling law of our model on our AllClear datasets with UnCRtainTS as backbone architecture.

| Fraction of Data | # data point | PSNR (↑) | SSIM (↑) | SAM (↓) | MAE (↓) |
|---|---|---|---|---|---|
| 1% | 2,786 | 27.035 | 0.898 | 5.972 | 0.039 |
| 3.4% | 10,167 | 28.474 | 0.906 | 6.373 | 0.036 |
| 10% | 27,861 | 32.997 | 0.923 | 6.038 | **0.023** |
| 100% | 278,613 | **33.868** | **0.936** | **5.232** | **0.021** |

### 4.3 Stratified evaluations

We use the available land-cover type labels in *AllClear* to conduct a stratified evaluation across land-cover types (Figure 3). We generally find that both PSNR and SSIM metrics are much worse for both water bodies and snow cover. Water bodies have transient wave patterns, and snow cover is also often transient, which may explain the difficulty of predicting these classes. Snow may also be confused with cloud.

Following past work [Ebel et al., 2022], we also perform a stratified evaluation of accuracy relative to the extent of cloud cover and shadows (Figure 5). For cloud cover, generally performance decreases

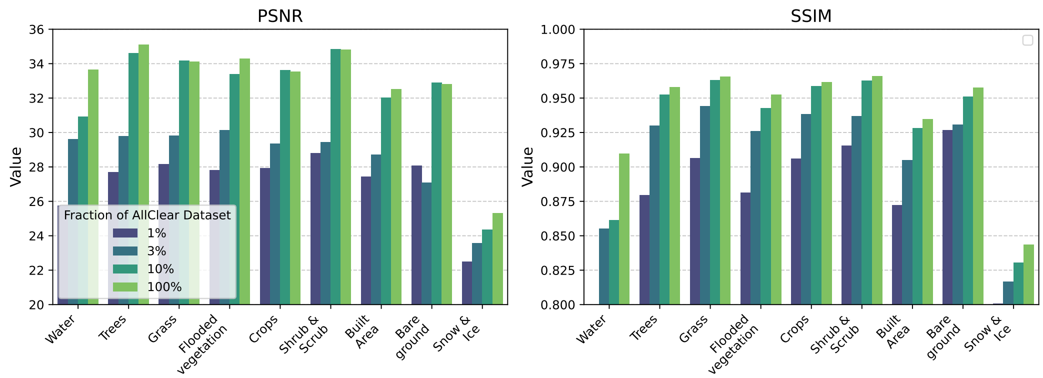


**Figure 3:** Land cover stratified evaluation of models trained with different fractions of the AllClear dataset: $1\%$, $3.4\%$, $10\%$, and $100\%$.

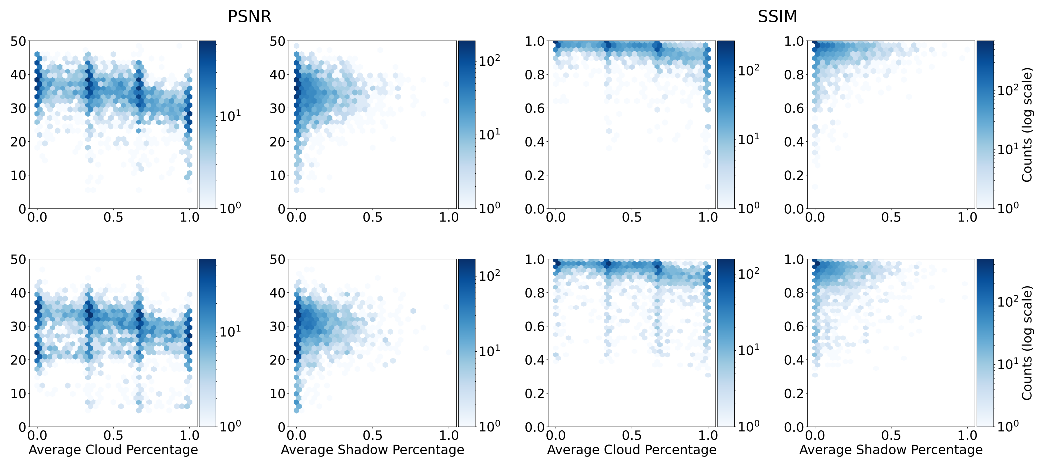


**Figure 4:** Cloud removal quality measured by PSNR (left column) and SSIM (right column) at different cloud and shadow coverage levels. The top row represents models trained on the full AllClear dataset, and the bottom row represents models trained on the SEN12MS-CR-TS dataset.

with cloud percentage, which is expected. Training on a larger dataset (AllClear) substantially improves accuracy for low and medium cloud cover, but not for fully clouded regions. Note that the striped pattern is because of fully cloudy images as explained in the Appendix. Shadows are generally less of a problem, and shadow percentage seems to be uncorrelated with performance.

### 4.4 Effect of various temporal spans

We next use our benchmark to see whether the common practice of using 3 input images is sufficient. We compare two models, one using 3 images and the other using all 12 images captured at that location. Both models are trained on a 10k subset of *AllClear*. The results, shown in Table 5, suggest that in fact a longer timespan significantly improves accuracy. Future cloud removal techniques should therefore consider longer timespans.

**Table 5:** Effect of different temporal length.

| # Consecutive Frame as Input | PSNR (↑) | SSIM (↑) | SAM (↓) | MAE (↓) |
|---|---|---|---|---|
| 3 | 28.474 | 0.906 | 6.373 | 0.036 |
| 12 | 30.399 | 0.919 | 5.920 | 0.028 |

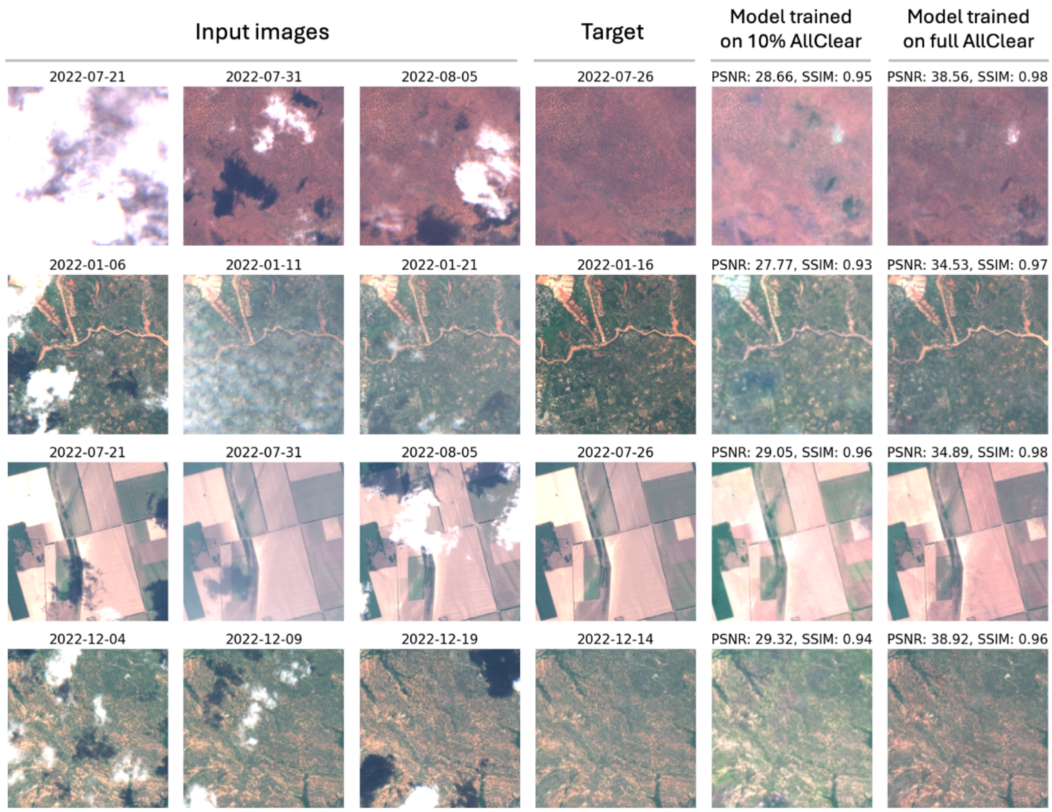


**Figure 5:** Scaling the training dataset by ten-folds gives better qualitative results.

# 5 Conclusion

This paper has introduced *AllClear*, the most extensive and diverse dataset available for cloud removal research. The larger training set significantly advances state-of-the-art performance. Our dataset also enables stratified evaluation on cloud coverage and land cover, and ablations of the sequence length and sensor type. We hope that future research can build on this benchmark to advance cloud removal, for instance by exploring the dynamics between SAR and multispectral images.

## References

JD Bermudez, PN Happ, DAB Oliveira, and RQ Feitosa. Sar to optical image synthesis for cloud removal with generative adversarial networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:5–11, 2018.

Jose D Bermudez, Patrick N Happ, Raul Q Feitosa, and Dario AB Oliveira. Synthesis of multispectral optical images from sar/optical multitemporal data using conditional generative adversarial networks. *IEEE Geoscience and Remote Sensing Letters*, 16(8):1220–1224, 2019.

Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, Simon Ilyushchenko, et al. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific Data*, 9(1):251, 2022.

Paul C Doraiswamy, Sophie Moulin, Paul W Cook, and Alan Stern. Crop yield assessment from remote sensing. *Photogrammetric engineering & remote sensing*, 69(6):665–674, 2003.

Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012.

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*, 59(7):5866–5878, 2020.

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*, 60:1–14, 2022.

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

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 48–56, 2017.

Mario Fuentes Reyes, Stefan Auer, Nina Merkle, Corentin Henry, and Michael Schmitt. Sar-to-optical image translation based on conditional generative adversarial networks—optimization, opportunities and limits. *Remote Sensing*, 11(17):2067, 2019.

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.

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 *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1726–1729. IEEE, 2018.

André Hollstein, Karl Segl, Luis Guanter, Maximilian Brell, and Marta Enesco. Ready-to-use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in sentinel-2 msi images. *Remote Sensing*, 8(8):666, 2016.

Gi-Luen Huang and Pei-Yuan Wu. Ctgan: Cloud transformer generative adversarial network. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 511–515. IEEE, 2022.

jdbcode. Sentinel-2 cloud masking with s2cloudless. https://developers.google.com/earth-engine/tutorials/community/sentinel-2-s2cloudless, 2023. Accessed: 2023-06-05.

Ran Jing, Fuzhou Duan, Fengxian Lu, Miao Zhang, and Wenji Zhao. Denoising diffusion probabilistic feature-based network for cloud removal in sentinel-2 imagery. *Remote Sensing*, 15(9):2217, 2023.

Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. *arXiv preprint arXiv:2312.03606*, 2023.

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*, 51(7):3826–3852, 2013.

Kostis Kyzirakos, Manos Karpathiotakis, George Garbis, Charalampos Nikolaou, Konstantina Bereta, Ioannis Papoutsis, Themos Herekakis, Dimitrios Michail, Manolis Koubarakis, and Charalambos Kontoes. Wildfire monitoring using satellite images, ontologies and linked geospatial data. *Journal of web semantics*, 24:18–26, 2014.

M. Paperin, J. Wevers, K. Stelzer, and C. Brockmann. PixBox Sentinel-2 pixel collection for CMIX (Version 1.0), 2021. URL https://doi.org/10.5281/zenodo.5036991.

Hai Minh Pham, Yasushi Yamaguchi, and Thanh Quang Bui. A case study on the relation between city planning and urban growth using remote sensing and spatial metrics. *Landscape and Urban Planning*, 100(3):223–230, 2011.

Md Shahinoor Rahman and Liping Di. A systematic review on case studies of remote-sensing-based flood crop loss assessment. *Agriculture*, 10(4):131, 2020.

Christian Requena-Mesa, Vitus Benson, Markus Reichstein, Jakob Runge, and Joachim Denzler. Earthnet2021: A large-scale dataset and challenge for earth surface forecasting as a guided video prediction task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1132–1142, 2021.

Vishnu Sarukkai, Anirudh Jain, Burak Uzkent, and Stefano Ermon. Cloud removal from satellite images using spatiotemporal generator networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1796–1805, 2020.

Praveer Singh and Nikos Komodakis. Cloud-gan: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1772–1775. IEEE, 2018.

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*, 274:112990, 2022.

Corinne Stucker, Vivien Sainte Fare Garnot, and Konrad Schindler. U-tilise: A sequence-to-sequence model for cloud removal in optical satellite time series. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

Kathiravan Thangavel, Dario Spiller, Roberto Sabatini, Stefania Amici, Sarathchandrakumar Thottuchirayil Sasidharan, Haytham Fayek, and Pier Marzocca. Autonomous satellite wildfire detection using hyperspectral imagery and neural networks: A case study on australian wildfire. *Remote Sensing*, 15(3):720, 2023.

Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, BjÖrn Rommen, Nicolas Floury, Mike Brown, et al. Gmes sentinel-1 mission. *Remote sensing of environment*, 120:9–24, 2012.

Lei Wang, Xin Xu, Yue Yu, Rui Yang, Rong Gui, Zhaozhuo Xu, and Fangling Pu. Sar-to-optical image translation using supervised cycle-consistent adversarial networks. *Ieee Access*, 7:129136–129149, 2019.

Thilo Wellmann, Angela Lausch, Erik Andersson, Sonja Knapp, Chiara Cortinovis, Jessica Jache, Sebastian Scheuer, Peleg Kremer, André Mascarenhas, Roland Kraemer, et al. Remote sensing in urban planning: Contributions towards ecologically sound policies? *Landscape and urban planning*, 204:103921, 2020.

Darrel L Williams, Samuel Goward, and Terry Arvidson. Landsat. *Photogrammetric Engineering & Remote Sensing*, 72(10):1171–1178, 2006.

Fang Xu, Yilei Shi, Patrick Ebel, Wen Yang, and Xiao Xiang Zhu. Multimodal and multiresolution data fusion for high-resolution cloud removal: A novel baseline and benchmark. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2023.

Weikang Yu, Xiaokang Zhang, and Man-On Pun. Cloud removal in optical remote sensing imagery using multiscale distortion-aware networks. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5, 2022.

Mingmin Zhao, Peder Olsen, and Ranveer Chandra. Seeing through clouds in satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

Xuechao Zou, Kai Li, Junliang Xing, Pin Tao, and Yachao Cui. Pmaa: A progressive multi-scale attention autoencoder model for high-performance cloud removal from multi-temporal satellite imagery. *arXiv preprint arXiv:2303.16565*, 2023a.

Xuechao Zou, Kai Li, Junliang Xing, Yu Zhang, Shiying Wang, Lei Jin, and Pin Tao. Diffcr: A fast conditional diffusion framework for cloud removal from optical satellite images. *arXiv preprint arXiv:2308.04417*, 2023b.

Anze Zupanc. Improving cloud detection with machine learning. *Accessed: Oct*, 10:2019, 2017.

## Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 4
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Yes we do
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We did not tune hyper-parameters
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include our codebase as supplementary
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Our data is downloaded from Google Earth Engine, with full open-access.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] Models are all publicly available.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]