Supplementary Material for "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

1 Overview

In this supplementary material we present more information about the dataset (including a datasheet for the dataset) and extensive results that could not fit in the main paper. In Sec. 2 we include a datasheet for our dataset, author statement, and hosting, licensing, and maintenance plan. In Sec. 3 we present more details about our dataset such as dataset specifications. In Sec. 4 we present full quantitative and qualitative benchmarking results on previous SoTA models trained across different datasets and ablation studies on the modalities.

The data is publicly available at `https://allclear.cs.cornell.edu`. Our code for accessing the dataset and benchmark result reproduction can be found at `https://github.com/Zhou-Hangyu/allclear`.

The Croissant metadata tool was not used because it does not support the metadata format we used in our dataset. Specifically, we use a hierarchical structure with dictionaries of lists to store the file path and corresponding timestamp for each image within each sample. The Croissant framework currently does not support parsing such a format. We will provide Croissant metadata file once support for this format is available in the future.

*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.

2 Datasheet

We include a datasheet for our dataset following the methodology from "Datasheets for Datasets" Gebru et al. [2021]. In this section, we include the prompts from Gebru et al. [2021] in blue, and in black are our answers.

2.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to facilitate research development on cloud removal in satellite imagery. The task we include allows a trained model to output a clear image given three (or more) cloudy satellite images. Specifically, our task is more temporally aligned than previous benchmarks.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Hangyu Zhou, Chia-Hsiang Kao, Cheng Perng Phoo, Utkarsh Mall, Bharath Hariharan, and Kavita Bala at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

This work was funded by the National Science Foundation (IIS-2144117).

Any other comments?

We specify the bands we collect for Sentinel-1, Sentinel-2, and Landsat-8/9. All images are sampled at 10-meter spatial resolution.

2.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

An individual instance in the benchmark dataset is a set of input images, target (clear) images, cloud and shadow masks, land use and land cover maps, and metadata. The input images primarily consist of Sentinel-2 images, while auxiliary sensor information such as Sentinel-1 and Landsat 8/9 may be included if specified in the arguments. Additionally, the number of timestamps for the input images can be 3, 6, or 12, indicating that the inputs contain images from different time frames, typically covering approximately 30 days of image collection, given the average revisit time for Sentinel-2 is 5 days. The cloud and shadow masks are binary spatial maps for each input and target Sentinel-2 image. The land use and land cover maps correspond to the target images. The metadata includes geolocation information such as latitude and longitude, as well as timestamps, sun elevation, sun azimuth, and precomputed cloud coverage.

How many instances are there in total (of each type, if appropriate)?

There are 278,613 training instances, 14,215 validation instances, and 55,317 benchmarking instances.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all instances from 23,742 ROIs (Regions of Interest) for the year 2022. It does not include all ROIs around the world, but it is a representative subset. We believe the samples are representative of the larger geographic coverage, as the ROI selection was balanced using land use and land cover maps.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

We describe an instance using an ordered pair $<I_1, I_2, I_3, T, M_1, M_2, M_3, M_T, DW, metadata>$. Specifically, there are three input cloudy images I_1, I_2, I_3 and a single target image T, each of spatial size $\mathbf{R}^{256\times256}$. The number of channels is 13 for Sentinel-2, 2 for Sentinel-1, and 11 for Landsat-8/9. The cloud and shadow masks for input M_1, M_2, M_3 and target M_T are all the same size as the inputs, with the number of channels being 5. These channels represent the cloud probability, binary cloud mask, and binary shadow mask with dark pixel thresholds of 0.2, 0.25, and 0.3. The DW indicates the land cover and land use maps, which have the same spatial size and resolution, with nine classes representing water, trees, grass, flooded vegetation, crops, shrub and scrub, built-up areas, bare land, and snow and ice. The $metadata$ includes geolocation (latitude and longitude), sun elevation and azimuth, and timestamps.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance is paired with a target clear image as ground truth. The target clear images are selected as images with cloud coverage less than 10%.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

All the information is included in the instances.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Relationships between instances are made explicit in the temporal and spatial domains. Specifically, the metadata for each instance includes information on their corresponding geolocations and timestamps, thereby establishing the relationships between instances based on their location and time of capture.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

We provide a train-validation-test split for our benchmark. The number of instances in train, validation, and test split are 278,613, and 14,215, and 55,317, respectively.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There are no redundancies in the dataset, as each instance is constructed to be non-overlapping with others in the spatiotemporal domain. However, errors in the dataset may arise from the cloud and shadow masks, since the cloud detection module is not yet perfect or 100% accurate, and similarly, the shadow mask may not be entirely accurate as it is derived from the cloud masks.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained as we provide all images with associated masks and metadata. This dataset is free for non-commercial usage and available to the public. For example, using our download code allows for collecting more metadata or other satellite imagery.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' nonpublic communications)? If so, please provide a description.

No, Sentinel-1, Sentinel-2, and Landsat-8/9 imageries are free to use for non-commercial usage and publicly accessible.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

The satellite images have a medium spatial resolution of 10 meters. We do not believe it includes content that is offensive, insulting, or threatening.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset

No, it does not identify any subpopulations.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No, the images are of medium resolution, making it impractical to identify or track individuals.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No, it does not contain sensitive information.

Any other comments?

None.

2.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The dataset is built upon the publicly available Sentinel-2, Sentinel-1, and Landsat-8/9 satellite imagery.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

The raw satellite images were collected using Google Earth Engine APIs [2].

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset is not a sample of a larger dataset.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The first authors are involved in the data collection process.

[2] https://developers.google.com/earth-engine

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataset is built with satellite imagery in the year 2022. The image captured time stamps for each image in each instance are explicitly labeled.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

The study was exempted from IRB as we do not collect any individual/personal information from users.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Our dataset does not contain information about individuals.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Our dataset does not contain information about individuals.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Our dataset does not contain information about individuals.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Our dataset does not contain information about individuals.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Our dataset does not contain information about individuals.

Any other comments?

None.

2.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

We preprocessed the Sentinel-2 and Landsat-8/9 images with value clipping and normalization. Detailed steps are depicted in Section 3.2.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

We do not do extra pre-processing of the downloaded image dataset. The preprocessing steps are done on the fly.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

Not applicable.

Any other comments?

None.

2.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

The dataset presented a novel task and has not been used for any tasks yet.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

N/A.

What (other) tasks could the dataset be used for?

Our datasets can be used to create benchmarks for sequence-to-sequence cloud removal as well. For example, the input images are a sequence of images where the clear ones are masked, and the target is the original sequence. The provided metadata contains sun position information and capture timestamps, which may be applied for more generative purposes. Our datasets provide a large corpus of cloudy satellite images, which can potentially facilitate developing cloud and shadow detection models.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

Our dataset does not contain information about individuals, so it should not result in unfair treatment of individuals or groups.

Are there tasks for which the dataset should not be used? If so, please provide a description.

None.

Any other comments?

None.

2.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset is publicly available on the internet.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset can be downloaded from Cornell's server at `https://allclear.cs.cornell.edu`. The dataset currently does not have a DOI, but we are planning to get one.

When will the dataset be distributed?

The dataset is available (since June 2024).

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is available under Creative Commons Attribution-NonCommercial 4.0 International License.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

Since our dataset is derived from Sentinel-2, Sentinel-1, and Landsat-8/9 images. Please also refer to Sentinel terms of service[3] and Landsat terms of service[4].

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No, there are no restrictions on the dataset.

Any other comments?

None.

2.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset is hosted and supported by web servers at Cornell. The CS department at Cornell will be maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Hangyu and Chia-Hsiang can be contacted via email (hz477@cornell.edu, and ck696@cornell.edu). More updated information can be found on the dataset webpage.

Is there an erratum? If so, please provide a link or other access point.

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

The updates to the dataset will be posted on the dataset webpage.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Our dataset does not contain information about individuals.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers

In case of updates, we plan to keep the older version of the dataset on the webpage.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If

[3] https://scihub.copernicus.eu/twiki/do/view/SciHubWebPortal/TermsConditions

[4] https://www.usgs.gov/emergency-operations-portal/data-policy

so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

We also provide the script downloading code in our codebase, which details our downloading configuration to ensure the dataset can be extended and augmented freely without inconsistency. Others may also do so by contacting the original authors about incorporating more fixes/extensions.

Any other comments?

None.

2.8 Author Statement

The authors assume full responsibility for any potential rights violations and the verification of data licensing.

2.9 Hosting, Licensing, and Maintenance Plan

The benchmarking dataset is hosted on a Cornell server and is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License. The first authors are responsible for maintaining the dataset.

3 Dataset Curation

We define a sample (i.e., an instance) from the AllClear dataset using an ordered pair $<I_1, I_2, I_3, T, M_1, M_2, M_3, M_T, DW, metadata>$. Specifically, there are three input cloudy images I_1, I_2, I_3 and a single target image T, each of spatial size $\mathbf{R}^{256\times256}$. The number of channels is 13 for Sentinel-2, 2 for Sentinel-1, and 11 for Landsat-8/9. We set Sentinel-2 to be the main sensor (i.e., we evaluate models' performance on reconstructing Sentinel-2 images) and use the other satellites as auxiliary ones. The cloud and shadow masks for input M_1, M_2, M_3 and target M_T are all the same size as the inputs, with the number of channels being 5. These channels represent the cloud probability, binary cloud mask, and binary shadow mask with dark pixel thresholds of 0.2, 0.25, and 0.3. Notably, the cloud and shadow masks are paired with and derived from Sentinel-2 input images only. The DW indicates the land cover and land use maps derived from Dynamic World (DW) V1 algorithm [Brown et al., 2022], which have the same spatial size and resolution, with nine classes representing water, trees, grass, flooded vegetation, crops, shrub and scrub, built-up areas, bare land, and snow and ice. The $metadata$ includes geolocation (latitude and longitude), sun elevation and azimuth, and timestamps.

For the benchmark dataset, we ensured that every target image have a corresponding land cover map generated by Dynamic World to enable stratified evaluation. After removing instances without corresponding land cover maps, we found that 98 out of 3,796 original test ROIs were disqualified, so we moved them to the training split to maintain benchmark dataset quantity and quality. For benchmark evaluation, we notice that some ROIs can provide over 30 test instances while some ROIs only have single test instance as shown in Figure 1, and thus we decide to sample one instance for each ROI to avoid oversampling, resulting in 3,698 benchmark instances. Future works can include more test instances as an alternative to gain a more comprehensive evaluation on model performance. The statistics of our dataset are based on the final version after these adjustments.

Table 1: *AllClear* Specifications

Specification	Description
Satellites	Sentinel-1/2, Landsat-8/9
ROIs	23708 (train, validation, test: 19013, 997, 3698)
Periods	2022.01.01 - 2022.12.31
Spectrum	Covering all useful bands with raw values
Cloud	Covering all cloud coverages without filtering
Metadata	Latitude, longitude, time-stamp, sun elevation, sun azimuth
File Format	Cloud Optimized GeoTIFF (COG) with ZSTD compression
# of images	4354652
# Sentinel-2 images	2185076 (train, validation, test: 1755206, 90590, 339280)
# Sentinel-1 images	897239 (train, validation, test: 721991, 38500, 136748)
# Landsat-8 images	637341 (train, validation, test: 510876, 26611, 99854)
# Landsat-9 images	634996 (train, validation, test: 508818, 26535, 99643)

We also provide the dataset assets in Table 2, specifying the bands we collected for each satellite sensors, the cloud and shadow masks, and the metadata. For Landsat-8/9, we use the Tier 1 TOA (top-of-atmosphere) Reflectance collection from the Google Earth Engine. For cloud and shadow masks, we use the binary cloud mask from Channel 2 and the binary shadow mask from Channel 5 by default for all our experiments.

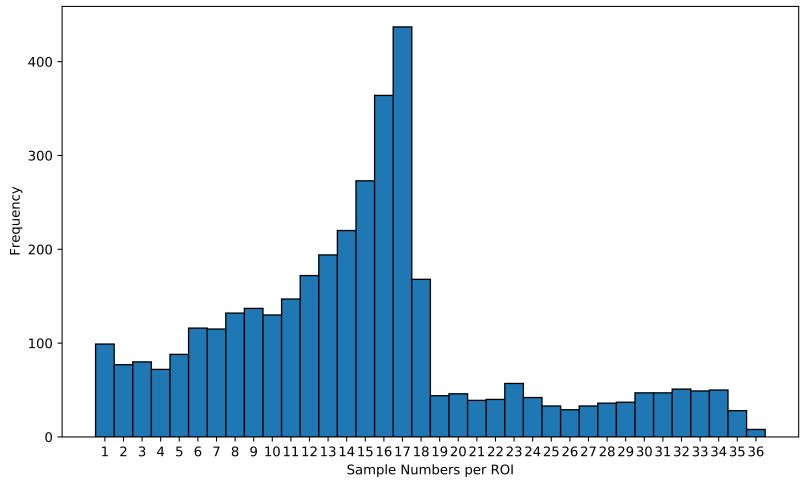

Figure 1: Histogram of the number of instance per test ROI.

4 Experiments

4.1 Baseline evaluation on models pre-trained on STGAN dataset and Sen2_MTC dataset

We provide the qualitative and quantitative evaluation of the baseline models on AllClear. In Table 3, the full table for all the available baseline models from other papers are shown. The results reveal that models trained on STGAN and Sen2_MTC all give worse performance on AllClear. In Figure 2, we show the corresponding visualization for some test samples in AllClear.

4.2 Ablation studies on multi-modality

In this subsection, we explore the integration of multiple sensors into the input data. As described in the main manuscript, we concatenate multi-spectral Sentinel-2 images with Sentinel-1 and Landsat images to create an input with multiple channels. However, due to the differing revisit intervals of these satellites, there can be gaps in the input sequences, meaning that some Sentinel-2 images may not have corresponding Sentinel-1 or Landsat-8/9 images.

To address these gaps, we experimented with different preprocessing strategies, as shown in Table 4. We discovered that filling the gaps with different constant values significantly impacts the results. Specifically, filling with zeros yielded better performance compared to filling with ones. Also, we provided additional experiments adding an extra input dimension called the "availability mask," which is filled with zeros if there is no paired Sentinel-1 image and ones otherwise, but this approach did not improve results.

Additionally, while outcomes regarding using extra Landsat images were inconsistent, filling gaps with zeros for Landsat produced the best results, albeit still lower than using only Sentinel-1 and Sentinel-2 alone. This might be due to the low-resolution of Landsat imagery; we suggest a model redeisgn to fully exploit Landsat images.

We also revisited the results with the scaling law using the new preprocessing method for Sentinel-1 gaps. As shown in Table 5, the scaling law holds for both preprocessing methods. Additionally, when it comes to full dataset training, the preprocessing methods do not cause significant differences in the results. Interestingly, the overall results improve when the Sentinel-1 gaps are filled with constant zeros during the small and medium dataset regimes, indicating a potential inductive bias of filling with constant zeros.

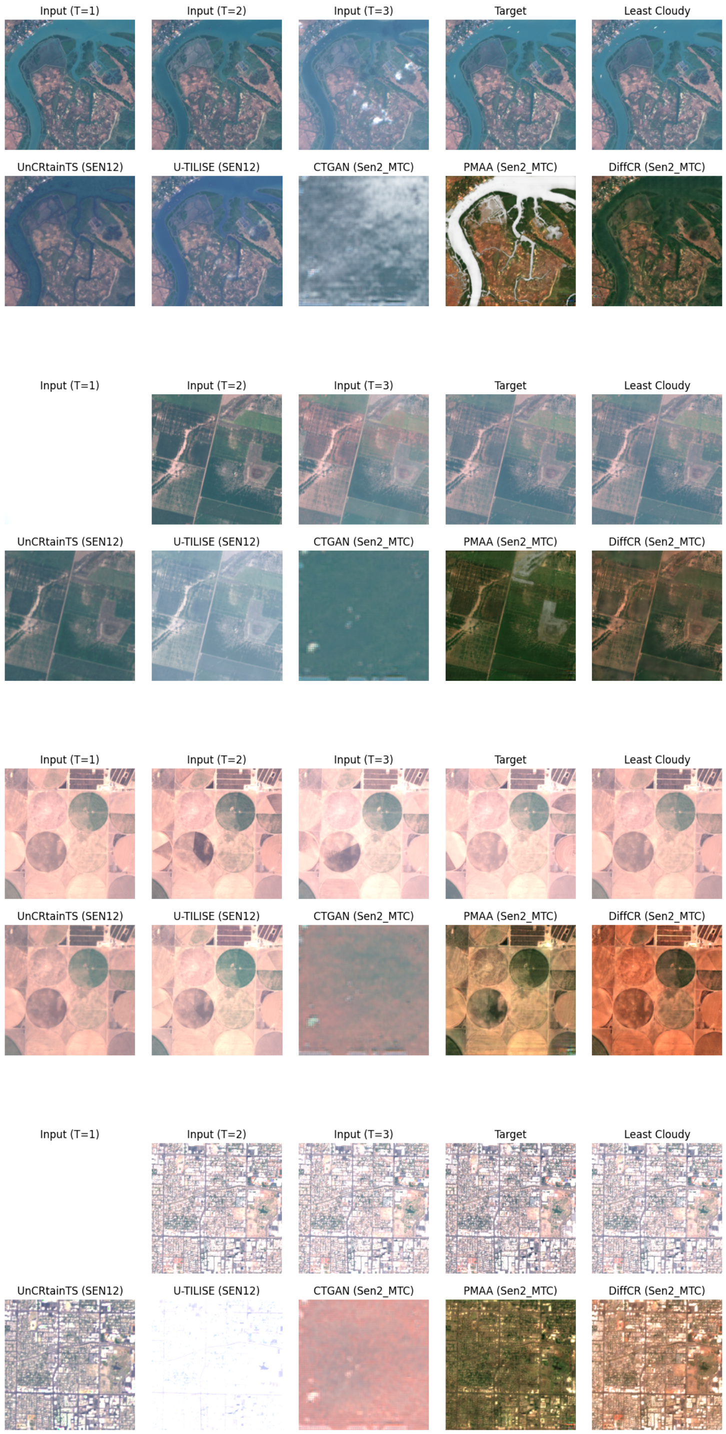

Figure 2: Qualitative comparison of the results from different baseline models. The results from four ROIs are shown, including three input images, the target image, the simple baseline result (i.e., Least Cloudy), and the outputs from previous pre-trained models. Specifically, we added the dataset that the model is pre-trained on in the bracket. The results show that the pre-trained UnCRtainTS attains the best qualitative results among all the pre-trained models, while U-TILISE performs well when the input images are mostly clear. On the contrary, CTGAN, PMAA, and DiffCR, pre-trained on a smaller dataset [Huang and Wu, 2022], show several color shifts.

Table 2: List of assets available for each instance.

Data Type	Channels	Wavelength	Description
Sentinel-2	B1	443.9 nm (S2A) / 442.3 nm (S2B)	Aerosols.
	B2	496.6 nm (S2A) / 492.1 nm (S2B)	Blue.
	B3	560 nm (S2A) / 559 nm (S2B)	Green.
	B4	664.5 nm (S2A) / 665 nm (S2B)	Red.
	B5	703.9 nm (S2A) / 703.8 nm (S2B)	Red Edge 1.
	B6	740.2 nm (S2A) / 739.1 nm (S2B)	Red Edge 2.
	B7	782.5 nm (S2A) / 779.7 nm (S2B)	Red Edge 3.
	B8	835.1 nm (S2A) / 833 nm (S2B)	NIR.
	B8A	864.8 nm (S2A) / 864 nm (S2B)	Red Edge 4.
	B9	945 nm (S2A) / 943.2 nm (S2B)	Water vapor.
	B10	1373.5 nm (S2A) / 1376.9 nm (S2B)	Cirrus.
	B11	1613.7 nm (S2A) / 1610.4 nm (S2B)	SWIR 1.
	B12	2202.4 nm (S2A) / 2185.7 nm (S2B)	SWIR 2.
Sentinel-1	VV	5.405 GHz	Dual-band cross-polarization, vertical transmit/horizontal receive.
	VH	5.405 GHz	Single co-polarization, vertical transmit/vertical receive.
Landsat-8/9	B1	0.43 - 0.45 µm	Coastal aerosol.
	B2	0.45 - 0.51 µm	Blue.
	B3	0.53 - 0.59 µm	Green.
	B4	0.64 - 0.67 µm	Red.
	B5	0.85 - 0.88 µm	Near infrared.
	B6	1.57 - 1.65 µm	Shortwave infrared 1.
	B7	2.11 - 2.29 µm	Shortwave infrared 2.
	B8	0.52 - 0.90 µm	Band 8 Panchromatic.
	B9	1.36 - 1.38 µm	Cirrus.
	B10	10.60 - 11.19 µm	Thermal infrared 1, resampled from 100m to 30m.
	B11	11.50 - 12.51 µm	Thermal infrared 2, resampled from 100m to 30m.
Land use	Label	-	Pixel-wise land cover labels.
Cloud and shadow masks	Channel 1	Cloud probability (%)	Derived from s2cloudless product.
	Channel 2	Binary cloud mask	Derived from thresholding cloud probability at 30.
	Channel 3	Binary shadow mask	Threshold for dark pixel set to 0.20.
	Channel 4	Binary shadow mask	Threshold for dark pixel set to 0.25.
	Channel 5	Binary shadow mask	Threshold for dark pixel set to 0.30.
Metadata	List of attributes	-	Latitude, longitude, sun elevation, sun azimuth, capture timestamp.

4.3 Correlation between Cloud Removal Quality and Cloud and Shadow Coverage

We illustrate the relationship between qualitative performance and cloud and shadow coverage in Figure 3. From the left to the right columns, we quantify the cloud and shadow mask using (1) average cloud coverage, (2) average shadow mask coverage, (3) consistent cloud coverage, and (4) consistent shadow coverage. Specifically, consistent cloud (shadow) coverage refers to the percentage of pixels in the input images that are always covered by clouds (shadows). This shows a consistent trend where higher cloud coverage correlates with decreased quality of the target images, consistent with previous observations. The strips in the subplots, especially in the left column at x-axis values of 0.33, 0.67, and 1.0, are due to the fact that some images are fully clouded, resulting in more data points in particular positions in those subplots. During shadow mask synthesis, we discard regions of shadow masks that overlap with cloud masks. Thus images with low shadow percentage may have extremely high or extremely low cloud coverage. This explains the high variance of model performance in the low shadow percentage region.

Table 3: Benchmark performance of previous SoTA models evaluated on our AllClear benchmark dataset.

Model	Training Dataset	PSNR (↑)	SSIM (↑)	SAM (↓)	MAE (↓)
Least Cloudy	-	28.864	0.836	6.982	0.078
Mosaicing	-	29.824	0.754	23.58	0.045
UnCRtainTS	SEN12MS-CR-TS	29.009	0.898	5.972	0.039
U-TILISE	SEN12MS-CR-TS	24.660	0.807	7.765	0.083
CTGAN	Sen2_MTC	27.783	0.840	8.800	0.041
PMAA	STGAN	12.455	0.460	8.072	0.240
PMAA	Sen2_MTC	24.328	0.768	8.680	0.078
DiffCR	STGAN	17.998	0.642	9.512	0.117
DiffCR	Sen2_MTC	25.220	0.744	9.382	0.060

Table 4: Multi-modality ablation studies. UnCRtainTS models are trained on a 10K subset of samples from our datasets with various setups. *S1* and *LS* denote Sentinel-1 and Landsat images, respectively. Preprocessing methods: *FZ* - Fill zeros, *FO* - Fill ones, *AM* - Availability mask. *FZ/FO* indicates filling gaps with constant zeros/ones when no nearby S1 images are available. The best-performing results are **bolded** and the second best are <u>underlined</u>.

Sentinel-2	Sentinel-1	Landsat-8/9	Preproc.	PSNR (↑)	SSIM (↑)	SAM (↓)	MAE (↓)
✓	✓		S1: FO	28.474	0.906	6.373	0.036
✓			-	31.725	0.920	<u>6.084</u>	0.026
✓	✓		S1: AM	30.506	0.922	6.258	0.027
✓	✓		S1: FZ	**33.107**	**0.930**	**5.719**	**0.022**
✓	✓	✓	S1: FO, LS: FO	30.040	0.898	6.989	0.033
✓	✓	✓	S1: FZ, LS: FO	31.416	0.914	6.622	0.026
✓	✓	✓	S1: FZ, LS: FZ	<u>32.522</u>	<u>0.923</u>	6.233	<u>0.024</u>

Table 5: Scaling law of our model on our AllClear datasets with UnCRtainTS as backbone architecture, with gaps being zeros. Preprocessing methods: *FZ* - Fill zeros, *FO* - Fill ones. *FZ/FO* indicates filling gaps with constant zeros/ones when no nearby S1 images are available. The best-performing results are **bolded** and the second best are <u>underlined</u>.

Fraction of Data	# data point	Preproc.	PSNR (↑)	SSIM (↑)	SAM (↓)	MAE (↓)
1%	2,786	S1: FO	27.035	0.898	5.972	0.039
3.4%	10,167	S1: FO	28.474	0.906	6.373	0.036
10%	27,861	S1: FO	32.997	0.923	6.038	0.023
100%	278,613	S1: FO	<u>33.868</u>	**0.936**	**5.232**	**0.021**
1%	2,786	S1: FZ	32.039	0.922	6.469	0.024
3.4%	10,167	S1: FZ	33.107	0.930	5.719	<u>0.022</u>
10%	27,861	S1: FZ	33.163	0.929	5.606	0.023
100%	278,613	S1: FZ	**34.148**	<u>0.935</u>	<u>5.338</u>	**0.021**

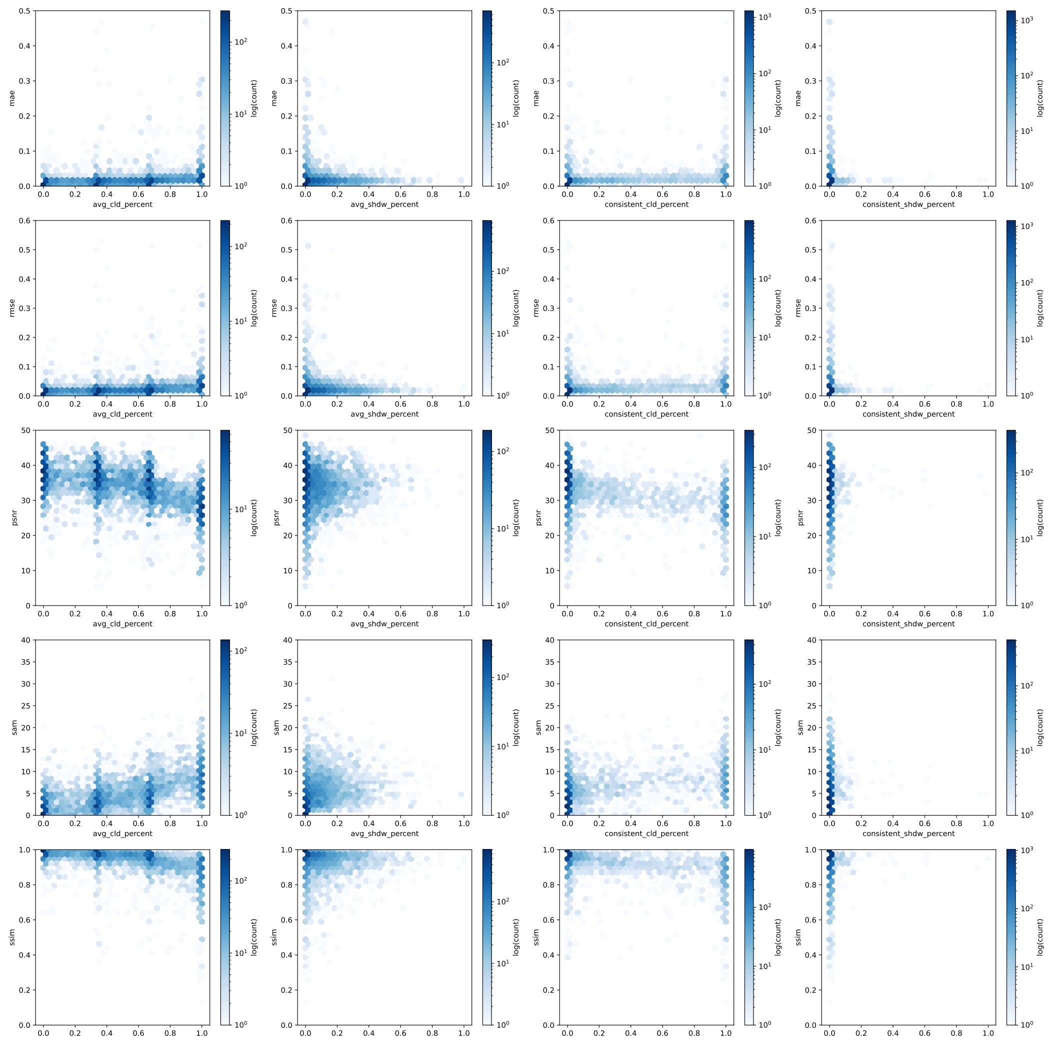

Figure 3: Correlation between cloud removal quality and cloud and shadow coverage of UnCRtainTS trained on full AllClear train set, evaluated on the AllClear test set. From left to right, the columns indicate average cloud coverage, average shadow mask coverage, consistent cloud coverage, and consistent shadow coverage. From top to bottom, the rows indicate the metrics MAE, RMSE, PSNR, SAM, and SSIM. The subplots show a consistent trend that a higher cloud coverage rate correlates with lower image reconstruction quality.

References

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.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021.

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.