
Supplementary Material for “AllClear: A Comprehensive Dataset and Benchmark for Cloud Removal in Satellite Imagery”

Hangyu Zhou^{1*}, Chia-Hsiang Kao^{1*}, Cheng Perng Phoo¹,
Utkarsh Mall², Bharath Hariharan¹, Kavita Bala¹

¹Computer Science, Cornell University

²Computer Science, Columbia University

1 Overview

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

8 The data is publicly available at <https://allclear.cs.cornell.edu>. Our code for accessing the
9 dataset and benchmark result reproduction can be found at [https://github.com/Zhou-Hangyu/
10 allclear](https://github.com/Zhou-Hangyu/allclear).

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

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

16 2 Datasheet

17 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
18 black are our answers.
19

20 2.1 Motivation

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

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

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

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

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

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

33 **Any other comments?**

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

36 2.2 Composition

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

40 An individual instance in the benchmark dataset is a set of input images, target (clear) images, cloud
41 and shadow masks, land use and land cover maps, and metadata. The input images primarily consist
42 of Sentinel-2 images, while auxiliary sensor information such as Sentinel-1 and Landsat 8/9 may be
43 included if specified in the arguments. Additionally, the number of timestamps for the input images
44 can be 3, 6, or 12, indicating that the inputs contain images from different time frames, typically
45 covering approximately 30 days of image collection, given the average revisit time for Sentinel-2 is
46 5 days. The cloud and shadow masks are binary spatial maps for each input and target Sentinel-2
47 image. The land use and land cover maps correspond to the target images. The metadata includes
48 geolocation information such as latitude and longitude, as well as timestamps, sun elevation, sun
49 azimuth, and precomputed cloud coverage.

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

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

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

58 The dataset contains all instances from 23,742 ROIs (Regions of Interest) for the year 2022. It does
59 not include all ROIs around the world, but it is a representative subset. We believe the samples are

60 representative of the larger geographic coverage, as the ROI selection was balanced using land use
61 and land cover maps.

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

64 We describe an instance using an ordered pair $\langle I_1, I_2, I_3, T, M_1, M_2, M_3, M_T, DW, metadata \rangle$.
65 Specifically, there are three input cloudy images I_1, I_2, I_3 and a single target image T , each of spatial
66 size $\mathbf{R}^{256 \times 256}$. The number of channels is 13 for Sentinel-2, 2 for Sentinel-1, and 11 for Landsat-8/9.
67 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,
68 with the number of channels being 5. These channels represent the cloud probability, binary cloud
69 mask, and binary shadow mask with dark pixel thresholds of 0.2, 0.25, and 0.3. The DW indicates
70 the land cover and land use maps, which have the same spatial size and resolution, with nine classes
71 representing water, trees, grass, flooded vegetation, crops, shrub and scrub, built-up areas, bare land,
72 and snow and ice. The $metadata$ includes geolocation (latitude and longitude), sun elevation and
73 azimuth, and timestamps.

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

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

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

80 All the information is included in the instances.

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

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

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

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

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

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

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

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

107 **Does the dataset contain data that might be considered confidential (e.g., data that is pro-**
108 **ected by legal privilege or by doctor–patient confidentiality, data that includes the content of**
109 **individuals’ nonpublic communications)?** If so, please provide a description.

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

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

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

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

119 No, it does not identify any subpopulations.

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

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

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

128 No, it does not contain sensitive information.

129 **Any other comments?**

130 None.

131 **2.3 Collection Process**

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

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

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

142 The raw satellite images were collected using Google Earth Engine APIs².

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

145 The dataset is not a sample of a larger dataset.

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

148 The first authors are involved in the data collection process.

²<https://developers.google.com/earth-engine>

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

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

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

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

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

161 Our dataset does not contain information about individuals.

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

165 Our dataset does not contain information about individuals.

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

170 Our dataset does not contain information about individuals.

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

174 Our dataset does not contain information about individuals.

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

178 Our dataset does not contain information about individuals.

179 **Any other comments?**

180 None.

181 **2.4 Preprocessing/cleaning/labeling**

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

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

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

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

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

194 Not applicable.

195 **Any other comments?**

196 None.

197 **2.5 Uses**

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

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

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

202 N/A.

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

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

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

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

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

219 None.

220 **Any other comments?**

221 None.

222 **2.6 Distribution**

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

225 Yes, the dataset is publicly available on the internet.

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

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

230 **When will the dataset be distributed?**

231 The dataset is available (since June 2024).

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

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

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

242 Since our dataset is derived from Sentinel-2, Sentinel-1, and Landsat-8/9 images. Please also refer to
243 Sentinel terms of service³ and Landsat terms of service⁴.

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

247 No, there are no restrictions on the dataset.

248 **Any other comments?**

249 None.

250 **2.7 Maintenance**

251 **Who will be supporting/hosting/maintaining the dataset?**

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

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

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

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

258 No.

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

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

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

267 Our dataset does not contain information about individuals.

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

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

271 **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for**
272 **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>

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

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

278 **Any other comments?**

279 None.

280 **2.8 Author Statement**

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

283 **2.9 Hosting, Licensing, and Maintenance Plan**

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

287 **3 Dataset Curation**

288 We define a sample (i.e., an instance) from the AllClear dataset using an ordered pair
 289 $\langle I_1, I_2, I_3, T, M_1, M_2, M_3, M_T, DW, metadata \rangle$. Specifically, there are three input cloudy im-
 290 ages I_1, I_2, I_3 and a single target image T , each of spatial size $\mathbf{R}^{256 \times 256}$. The number of channels
 291 is 13 for Sentinel-2, 2 for Sentinel-1, and 11 for Landsat-8/9. We set Sentinel-2 to be the main
 292 sensor (i.e., we evaluate models’ performance on reconstructing Sentinel-2 images) and use the other
 293 satellites as auxiliary ones. The cloud and shadow masks for input M_1, M_2, M_3 and target M_T are all
 294 the same size as the inputs, with the number of channels being 5. These channels represent the cloud
 295 probability, binary cloud mask, and binary shadow mask with dark pixel thresholds of 0.2, 0.25, and
 296 0.3. Notably, the cloud and shadow masks are paired with and derived from Sentinel-2 input images
 297 only. The DW indicates the land cover and land use maps derived from Dynamic World (DW) V1
 298 algorithm [Brown et al., 2022], which have the same spatial size and resolution, with nine classes
 299 representing water, trees, grass, flooded vegetation, crops, shrub and scrub, built-up areas, bare land,
 300 and snow and ice. The $metadata$ includes geolocation (latitude and longitude), sun elevation and
 301 azimuth, and timestamps.

302 For the benchmark dataset, we ensured that every target image have a corresponding land cover
 303 map generated by Dynamic World to enable stratified evaluation. After removing instances without
 304 corresponding land cover maps, we found that 98 out of 3,796 original test ROIs were disqualified,
 305 so we moved them to the training split to maintain benchmark dataset quantity and quality. For
 306 benchmark evaluation, we notice that some ROIs can provide over 30 test instances while some ROIs
 307 only have single test instance as shown in Figure 1, and thus we decide to sample one instance for
 308 each ROI to avoid oversampling, resulting in 3,698 benchmark instances. Future works can include
 309 more test instances as an alternative to gain a more comprehensive evaluation on model performance.
 310 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)

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

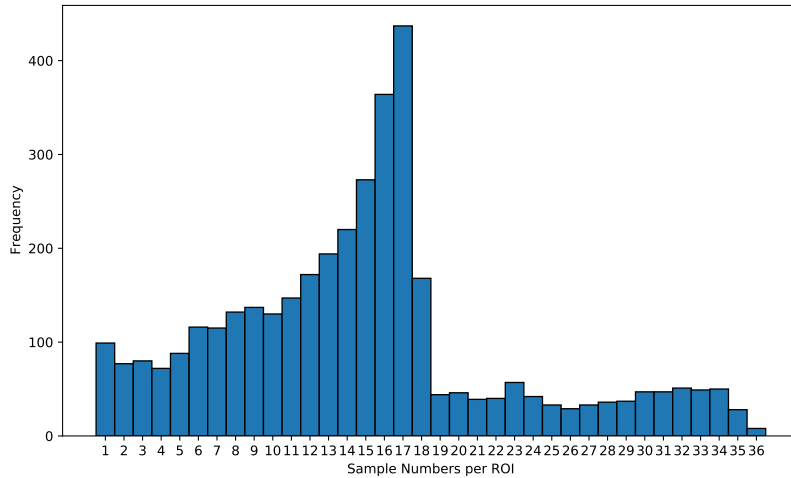


Figure 1: Histogram of the number of instance per test ROI.

316 4 Experiments

317 4.1 Baseline evaluation on models pre-trained on STGAN dataset and Sen2_MTC dataset

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

322 4.2 Ablation studies on multi-modality

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

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

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

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

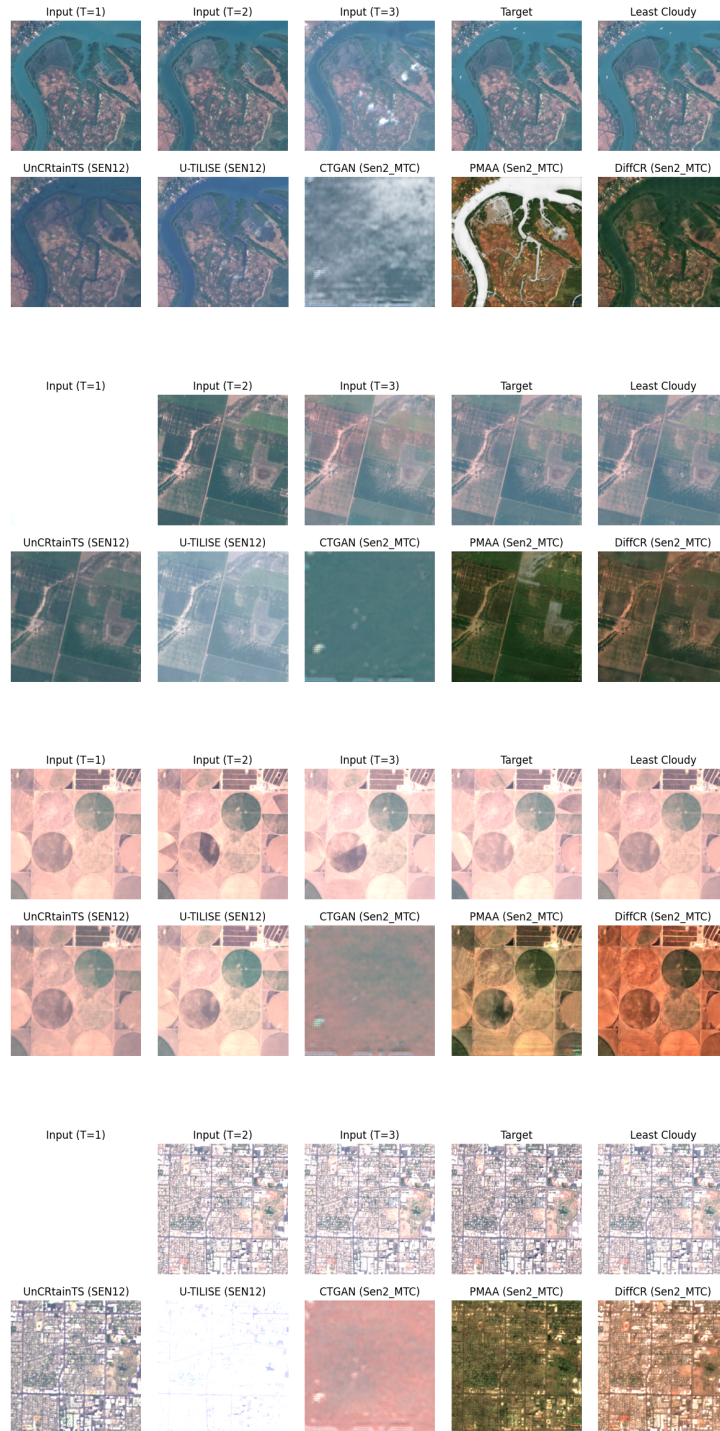


Figure 2: Qualitative comparison of the results from different baseline models. The results from four ROIs are shown, including three input images, the target image, the simple baseline result (i.e., Least Cloudy), and the outputs from previous pre-trained models. Specifically, we added the dataset that 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.

344 4.3 Correlation between Cloud Removal Quality and Cloud and Shadow Coverage

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

Sentinel-2	Sentinel-1	Landsat-8/9	Preproc.	PSNR (\uparrow)	SSIM (\uparrow)	SAM (\downarrow)	MAE (\downarrow)
✓	✓		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 underlined.

Fraction of Data	# data point	Preproc.	PSNR (\uparrow)	SSIM (\uparrow)	SAM (\downarrow)	MAE (\downarrow)
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

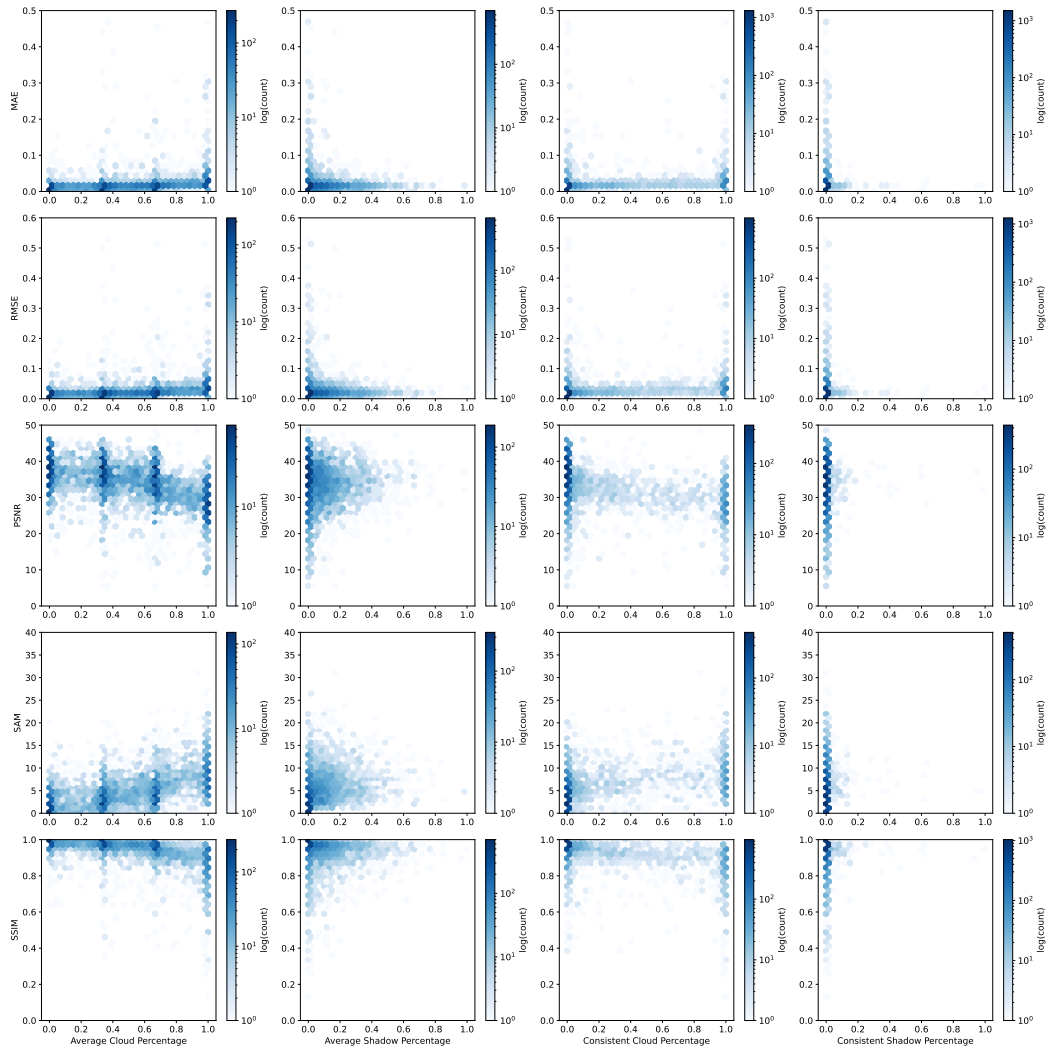


Figure 3: Correlation between cloud removal quality and cloud and shadow coverage of UnCRtainTS trained on full AllClear train set, evaluated on the AllClear test set. From left to right, the columns indicate average cloud coverage, average shadow mask coverage, consistent cloud coverage, and consistent 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.

357 **References**

- 358 Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks
359 Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, Simon
360 Ilyushchenko, et al. Dynamic world, near real-time global 10 m land use land cover mapping.
361 *Scientific Data*, 9(1):251, 2022.
- 362 Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach,
363 Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021.
- 364 Gi-Luen Huang and Pei-Yuan Wu. Ctgan: Cloud transformer generative adversarial network. In
365 *2022 IEEE International Conference on Image Processing (ICIP)*, pages 511–515. IEEE, 2022.