MultiEarth 2022 – Multimodal Learning for Earth and Environment Workshop and Challenge

Miriam Cha[1], Kuan Wei Huang[2], Morgan Schmidt[2], Gregory Angelides[1], Mark Hamilton[2],
Sam Goldberg[2], Armando Cabrera[3], Phillip Isola[2], Taylor Perron[2], Bill Freeman[2], Yen-Chen Lin[2],
Brandon Swenson[3], Jean Piou[1]

[1]MIT Lincoln Laboratory, [miriam.cha, gregangelides, jepiou]@ll.mit.edu
[2]MIT, [kwhuang, morgansc, markth, sgoldberg, phillipi, perron, billf, yenchenl]@mit.edu
[3]DAF MIT AI Accelerator, [armando.cabrera, brandon.swenson.2]@us.af.mil

Abstract

The Multimodal Learning for Earth and Environment Challenge (MultiEarth 2022) will be the first competition aimed at the monitoring and analysis of deforestation in the Amazon rainforest at any time and in any weather conditions. The goal of the Challenge is to provide a common benchmark for multimodal information processing and to bring together the earth and environmental science communities as well as multimodal representation learning communities to compare the relative merits of the various multimodal learning methods to deforestation estimation under well-defined and strictly comparable conditions. MultiEarth 2022 will have three sub-challenges: 1) matrix completion, 2) deforestation estimation, and 3) image-to-image translation. This paper presents the challenge guidelines, datasets, and evaluation metrics for the three sub-challenges. Our challenge website is available at `https://sites.google.com/view/rainforest-challenge`.

1. Introduction

Despite international efforts to reduce deforestation, the world loses, so far, an area of forest that is equivalent to the size of 40 football fields every minute [11]. Deforestation in the Amazon rainforest accounts for the largest share, contributing to reduced biodiversity, habitat loss, and climate change. Since much of the region is difficult to access, satellite remote sensing offers a powerful tool to track changes in the Amazon. However, obtaining a continuous time series of images is hindered by seasonal weather, clouds, smoke, and other inherent limitations of optical sensors. Synthetic aperture radar (SAR), which is insensitive to lighting and weather conditions, appears to be a well suited tool for the task, but SAR images are more difficult for humans to interpret than optical images. A key component of this challenge is to monitor the Amazon rainforest in all weather and lighting conditions using our multimodal remote sensing dataset, which includes a time series of multispectral and SAR images. The 2022 Multimodal Learning for Earth and Environment Challenge (MultiEarth 2022) will be the first competition aimed at monitoring the Amazon rainforest and predicting deforestation using multimodal representation learning methods.

While considerable research has been devoted to tracking changes in forests [4, 6, 8, 9], the analyses typically rely on passive, optical sensors such as the U.S. Geological Survey's Landsat, NASA's Moderate Resolution Imaging Spectroradiometer (MODIS), and the European Space Agency's Sentinel-2 [1]. These passive, optical sensors need an unobstructed and illuminated view of the scene in order to capture meaningful images. This limits their effectiveness in monitoring the Amazon rainforest with year-round cloud coverage. Beyond the Amazon, approximately 67 percent of Earth's surface is typically covered by clouds [7]. Therefore, using SAR for cloud-free observation will have a broad applicability.

MultiEarth 2022 will conduct the following sub-challenges to support the interpretation and analysis of the rainforest at any time and any weather conditions:

- **Matrix Completion Sub-Challenge**: given remote sensing images taken at different locations and dates, and in different modalities, participants are required to predict appearance at a novel [lon, lat, date, modality] query. Performance is measured on the following visual metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [10], Learned Perceptual Image Patch Similarity (LPIPS) [12], and Fréchet Inception Distance (FID) [5]. Detailed challenge description is provided in Section 3.1.

- **Deforestation Estimation Sub-Challenge**: beyond visual appearance, participants are required to classify whether a region is deforested or not at a novel [lon, lat, date, modality] query. Modality will be 'deforestation' for this sub-challenge. Participants will be given the multimodal remote sensing dataset along with the deforestation label maps. Performance is measured on the following metrics: pixel accuracy, F1 score, and Intersection over Union (IoU). Detailed description of this sub-challenge can be found in Section 3.2.

- **Image-to-Image Translation Sub-Challenge**: participants are required to predict a set of possible cloud-free corresponding optical images given an input SAR image. For this sub-challenge, we provide an aligned dataset (e.g. $[\mathbf{x}, [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]]$) where an input SAR image $\mathbf{x}$ is paired to a set of ground truth optical images $[\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]$. Performance is evaluated based on $\sum_j \min_i \|f(\mathbf{x})_i - \mathbf{y}_j\|$ where $f(\cdot)$ is a prediction of $\mathbf{y}$ given $\mathbf{x}$, and f may make multiple predictions indexed by i. Detailed guideline of the sub-challenge is provided in Section 3.3.

To be eligible to participate in the challenge, every entry has to be accompanied by a paper presenting the results and the methods that created them, which will undergo peer-review. The organizers reserve the right to re-evaluate the findings, but will not participate in the Challenge.

2. Datasets

2.1. Multimodal Remote Sensing Dataset

Participants will receive a multimodal remote sensing dataset that consists of Sentinel-1, Sentinel-2, Landsat 5, and Landsat 8 as reported in Table 1. Sentinel-1 uses a synthetic aperture radar (SAR) instrument, which collects in two polarization bands: VV (vertical transmit/vertical receive) and VH (vertical transmit/horizontal receive). Sentinel-2, Landsat 5, and Landsat 8 use optical instruments, which measure in spectral bands in the visible and infrared spectra. We also include in the dataset the associated layers with cloud quality for Sentinel-2, Landsat 5, and Landsat 8 (i.e. QA60 and QA_PIXEL). Detailed band designations for each sensor can be found in Google Earth Engine Data Catalog[1].

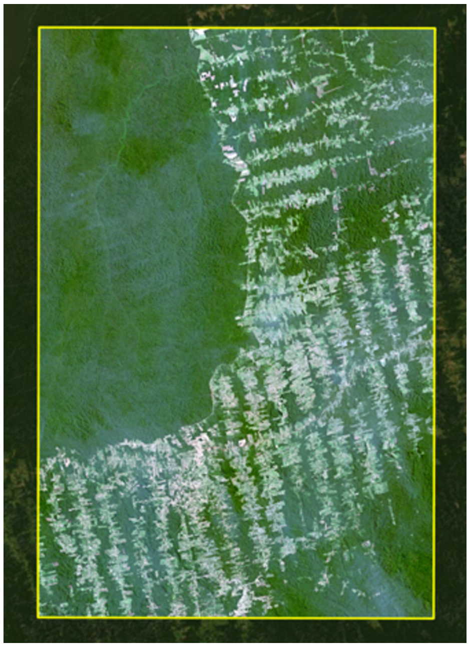

Figure 1. Study area in the Amazon, bounded by (4.39° S, 55.2° W), (4.39° S, 54.48° W), (3.33° S, 54.48° W) and (3.33° S, 55.2° W).

The region of interest is the rectangle bounded by the points (4.39° S, 55.2° W), (4.39° S, 54.48° W), (3.33° S, 54.48° W) and (3.33° S, 55.2° W) as shown in Figure 1. This area is further divided into squares centered on 54 × 37 = 1998 coordinate pairs by iterating in 0.02° increments in the latitude and longitude directions. Our dataset consists of 11,978,118 total training images. (Note that the image-to-image translation sub-challenge will use a subset of the total training images.) Test images are selected from the middle 7 columns (from longitudes -54.78 to -54.90), which are withheld as test region. The training images come from all bands and dates from the 4 sensor collections specified in Table 1 covering the area outside of the test region. Each image file name is in the format `{Collection}_{Band}_{Longitude}_{Latitude}_{Year}_{Month}_{Day}.tiff`. Sample images of Sentinel-1, Sentinel-2, Landsat 5, and Landsat 8 from our dataset are shown in Figure 2.

Satellite images from this dataset are downloaded using Google's Earth Engine platform. Each coordinate pair is projected from UTM (Universal Transverse Mercator) to EPSG (European Petroleum Survey Group). Then a 256×256 image centered at the coordinate is extracted for Sentinel-1 and Sentinel-2, and a 85×85 image centered at the coordinate is extracted for Landsat 5 and Landsat 8. The difference in pixels accounts for the 10-meter resolution Sentinel-1 and Sentinel-2 vs the 30-meter resolution Landsat 5 and Landsat 8. If the Landsat images are upsampled to 256×256, they will be geospatially aligned with the Sentinel images.

[1] Sentinel-1: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S1_GRD
Sentinel-2: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR
Landsat 5: https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LT05_C02_T1_L2
Landsat 8: https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1_L2

Sensor	Time	Bands	Resolution (m)	# Images	Link
Sentinel-1	2014-2021	VV, VH	10	859,627	link1 link2
Sentinel-2	2018-2021	B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B11, B12, QA60	10	5,395,559	link1 link2
Landsat 5	1984-2012	SR_B1, SR_B2, SR_B3, SR_B4, SR_B5, ST_B6, SR_B7, QA_PIXEL	30	3,550,368	link1 link2
Landsat 8	2013-2021	SR_B1, SR_B2, SR_B3, SR_B4, SR_B5, SR_B6, SR_B7, ST_B10, QA_PIXEL	30	2,172,564	link1 link2

Table 1. Overview of our multimodal remote sensing dataset that includes Sentinel-1, Sentinel-2, Landsat 5, and Landsat 8.

The dataset can be downloaded from the links provided in Table 1. Data is freely available for development, research, or educational purposes. For more details, please refer to Google Earth Engine License Agreement[2].

2.2. Deforestation Labels

Labeled data of deforestation for training and testing are manually labeled using cloud-free monthly mosaic satellite images from Planet [2]. The Planet imagery has 3.7 m spatial resolution and consists of 3 bands (RGB) with images from a single month joined together to achieve an image with the least amount of cloud cover. Most are cloud-free but some have a few clouds. Eleven time slices (months) are labeled over six years, from 2016 to 2021. They include: 08/2016; 07, 08/2017; 06, 08/2018; 07, 08/2019; 06, 08/2020; and 05, 08/2021. The August 2016 image contains a greater number of clouds and artifacts for the mosaic process. August 2021 is labeled by the authors and the remaining ten are labeled by the team at Scale AI [3]. The following guidelines are used by Scale AI to label deforested areas within polygons: 1) areas deforested for human use in any stage of regrowth; 2) forested areas 1 hectare or larger with unbroken forest canopy are not labeled; 3) roads are included as deforested if clearly visible; 4) rivers going through deforested areas are left unlabeled if their area is greater than 1 hectare; 5) clouds or image artifacts are not labeled; 6) snapping was used during labeling to connect polygons. For the August 2021 labeled data, our team does not use snapping but instead leave small spaces between polygons. Therefore, in the labeled dataset, unlabeled areas are mostly forest, but may also contain some clouds or rivers. Shapefiles with labeled polygons are converted into rasters with 0's for forested/other and 1's for deforested areas. Rasters are divided into georeferenced image chips (256×256 pixels) that correspond with the multimodal remote sensing dataset described in Section 2.1. We show sample images of deforestation label maps at various time slices in Figure 3.

3. Challenge Tasks

3.1. Matrix Completion Sub-Challenge

Problem Definition. Most studies of forest monitoring rely on remote sensing data collected by optical sensors [4,6,8,9]. However, it is often difficult, particularly in moist forest areas like the Amazon rainforest, to obtain cloud-free images. This sub-challenge focuses on filling in spatial, temporal, and modality gaps in remote sensing data, especially gaps created by unfavorable lighting, weather conditions, or other atmospheric factors. The Matrix Completion Sub-Challenge represents the sensory world as a huge tensor of measurements, in multiple modalities, sampled at different places and times. This tensor is typically very sparse, with most measurements missing, since satellites only photograph a tiny fraction of all possible wavelengths at all possible positions on the globe at all possible points in time. Participants are required to solve the tensor completion problem on this huge yet sparse multidimensional measurement tensor, and may develop and use methods from representation learning and generative modeling, or may consider other approaches.

Data. Training set from the multimodal remote sensing dataset described in Section 2.1 is provided.

Metrics. To rank all submissions to this sub-challenge, imaging results entered by the participants will be evaluated using the following four metrics: Peak Signal-to-Noise Ratio (PNSR), Structural Similarity Index Measure (SSIM) [10], Learned Perceptual Image Patch Similarity (LPIPS) [12], and Fréchet Inception Distance (FID) [5].

Submission Format. For the test input, 2000 test queries will be provided as a list of lists *i.e.* $[[\text{lon}_0, \text{lat}_0, \text{date}_0, \text{modality}_0], \ldots, [\text{lon}_{1999}, \text{lat}_{1999}, \text{date}_{1999}, \text{modality}_{1999}]]$. Each test query is in the format [lon, lat, date, modality]. For example, $[-55.15, -4.11, 2021\_12\_04, \text{Landsat8\_SR\_B2}]$ will represent Landsat8_SR_B2_-55.15_-

[2] https://earthengine.google.com/terms/

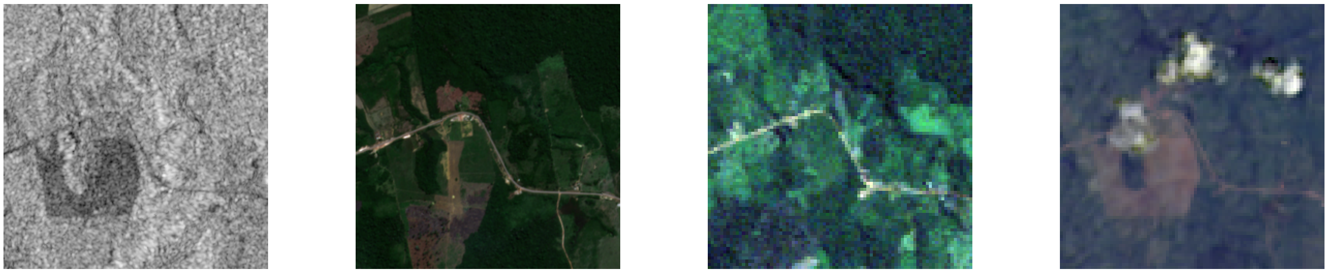
(a) Sentinel-1 (b) Sentinel-2 (c) Landsat 5 (d) Landsat 8

Figure 2. Image examples from our multimodal dataset (LAT/LON: -4.11/-55.14) (a) Sentinel-1 VV on 12/30/2021, (b) Sentinel-2 RGB (B4, B3, B2) on 8/4/2021, (c) Landsat 5 RGB (SR_B3, SR_B2, SR_B1) on 6/12/1998, and (d) Landsat 8 RGB (SR_B4, SR_B3, SR_B2) on 12/4/2021.

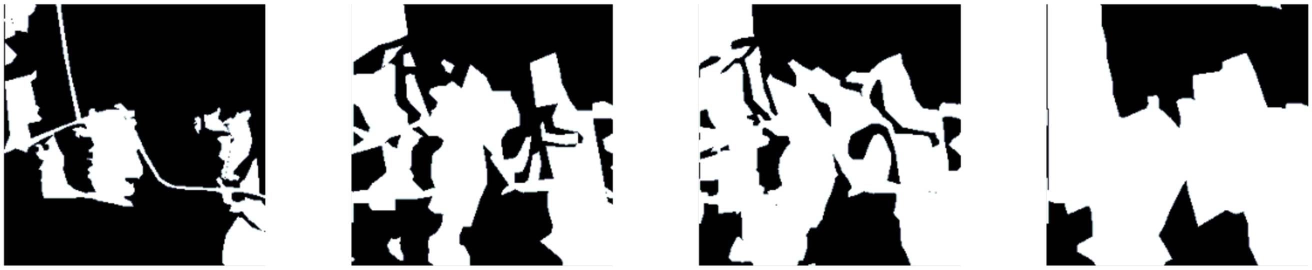
(a) 08/2016 (b) 07/2017 (c) 06/2018 (d) 08/2020

Figure 3. Example images of deforestation label maps corresponding to the location shown in Figure 2 (LAT/LON: -4.11/-55.14) at various time slices: 08/2016, 07/2017, 06/2018, and 08/2020.

4.11_2021_12_04.tiff. For the test output, participants will submit in total 2,000 256×256 images, one 256×256 image for each input test query. Imagery related to the input test queries will be made available in our website and can be used to help generate the requested output.

3.2. Deforestation Estimation Sub-Challenge

Problem Definition. Beyond predicting visual appearance, this sub-challenge is aimed at estimating deforestation from the multimodal remote sensing dataset. As described in the Matrix Completion Sub-Challenge, the multimodal dataset can be represented as a multidimensional data matrix with missing entries. The goal of this sub-challenge is to perform a binary classification to predict whether a region is deforested or not. As solutions for the Matrix Completion Sub-Challenge can be naturally extended, we strongly encourage participants in the Matrix Completion Sub-Challenge to submit to the Deforestation Estimation Sub-Challenge.

Data. Participants will use the multimodal remote sensing dataset described in Section 2.1 to predict binary deforestation label maps in Section 2.2.

Metrics. Performance is measured on the following standard metrics: pixel accuracy, F1 score, and Intersection over Union (IoU).

Submission Format. For the test input, 1000 test queries will be provided as a list of lists *i.e.* $[[\text{lon}_0, \text{lat}_0, \text{date}_0, \text{modality}_0], \ldots, [\text{lon}_{999}, \text{lat}_{999}, \text{date}_{999}, \text{modality}_{999}]]$. Each test query is in the format [lon, lat, date, modality]. For example, $[-55.15, -4.11, 2021\_08\_01,$ deforestation$]$ will represent deforestation_-55.15_-4.11_2021_08.png. To have a consistent naming convention for the date (*i.e.* year_month_day), we add a nominal day label of "_01" to all deforestation estimation test queries. For the test output, participants will submit in total 1,000 256×256 binary masks, one 256×256 binary mask for each input test query. Imagery related to the input test queries will be made available in our website and can be used to help generate the requested output.

3.3. Image-to-Image Translation Sub-Challenge

Problem Definition. Obtaining a continuous time series of view of the Amazon rainforest is hindered by weather, clouds, smoke, and other inherent limitations of passive sensors (e.g. optical sensors) that rely on sunlight. Such limitations produce a major information gap in the Amazon rainforest that gets rain throughout the entire year. Synthetic aperture radar (SAR) is an active sensor that can collect images with relative invariance to weather and lighting conditions. However, visual interpretation of SAR images is not intuitive due to the large dynamic range, low spatial correlation, and radar-specific geometry distortion. To enhance SAR interpretability, this sub-challenge is aimed at modeling a distribution of possible electro-optical (EO) image outputs conditioned on a SAR input image. Here, EO image is a 3-channel RGB image from Sentinel-2. In Sentinel-2, RGB bands are represented as B4, B3, and B2, respectively. SAR image is a 2-channel Sentinel-1 image consisting of VV and VH bands. In this Image-to-Image Translation Sub-Challenge, participants need to model a distribution of potential results in a conditional generative modeling setting.

Data. For this sub-challenge, we provide JSON files specifying which Sentinel-2 EO images (B4, B3, and B2) correspond to which SAR images. We provide two JSON files: one for Sentinel-1 VV band[3] and another for Sentinel-1VH band [4]. The mappings in both files are identical. The aligned dataset will have the following format: $[\mathbf{x}, [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]]$ where a SAR image $\mathbf{x}$ is paired to a set of ground truth EO images $[\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]$. For each SAR image, all EO images of the same geographic region and which were collected within 7 days of the SAR image timestamp will be identified. On average 3 EO images are paired with each SAR image ($N \approx 3$).

Metrics. Performance is evaluated based on following:

$$\sum_j \min_i \|f(\mathbf{x})_i - \mathbf{y}_j\| \quad (1)$$

where $f(\mathbf{x})_i$ is a set of possible EO images translated from a generative model $f(\cdot)$ conditioned on an input SAR image $\mathbf{x}$, and $\mathbf{y}_j$ is an EO image from the corresponding ground truth set. This metric also evaluates diversity of generated output images staying faithful to the diversity of the ground-truth EO data.

Submission Format. For testing, we will provide 5000 Sentinel-1 SAR images, where each SAR image is 256×256×2 and the 2 channels correspond to the Sentinel-1 VV and VH bands, respectively. For the test output, participants will submit in total 15,000 256×256×3 EO images, three translated 256×256×3 EO images for each input Sentinel-1 SAR image. The EO image channels correspond to the Sentinel-2 RGB channels (B4, B3, B2). This is a multimodal image-to-image translation problem where participants will generate three possible EO images given an input SAR image.

4. Conclusion

We introduce MultiEarth 2022 – the first open Multimodal Learning for Earth and Environment Challenge. It comprises three sub-challenges: 1) matrix completion, 2) deforestation estimation, and 3) image-to-image translation. This manuscript describes MultiEarth 2022's challenge conditions, data, evaluation metrics, and submission guideline. Only a few labeled, multimodal datasets including passive and active sensors have been publicly available, limiting the number of participants who can research and analyze Earth's surface at all times and in all weather conditions. We collect and disseminate a multimodal dataset that includes a continuous time series of Sentinel-1, Sentinel-2, Landsat 5 and Landsat 8, with deforestation labels.

5. Acknowledgments

The authors would like to thank Dr. M.K. Newey for helping the team labeling the August 2021 time slice data, and Scale AI, for its efforts to label, in a short period of time, the remaining ten time slices considered in this MultiEarth2022.

Research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] European Space Agency, 2015. Sentinels: Space for Copernicus. http://esamultimedia.esa.int.

[2] Planet Team. https://api.planet.com.

[3] Scale AI. https://scale.com.

[4] Maria Antonia Brovelli, Yaru Sun, and Vasil Yordanov. Monitoring forest change in the amazon using multi-temporal remote sensing data and machine learning classification on google earth engine. *ISPRS International Journal of Geo-Information*, 2020.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *NeurIPS*, 2017.

[6] Crismeire Isbaex and Ana Margarida Coelho. The potential of sentinel-2 satellite images for land-cover/land-use and forest biomass estimation: A review. In *Forest Biomass*. IntechOpen, Rijeka, 2021.

[7] Michael D. King, Steven Platnick, W. Paul Menzel, Steven A. Ackerman, and Paul A. Hubanks. Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites. *IEEE Transactions on Geoscience and Remote Sensing*, 2013.

[8] Thaís Almeida Lima, René Beuchle, Andreas Langner, Rosana Cristina Grecchi, Verena C. Griess, and Frédéric Achard. Comparing sentinel-2 msi and landsat 8 oli imagery for monitoring selective logging in the brazilian amazon. *Remote Sensing*, 2019.

[9] Fiona Ngadze, Kudzai Shaun Mpakairi, Blessing Kavhu, Henry Ndaimani, and Monalisa Shingirayi Maremba. Exploring the utility of sentinel-2 msi and landsat 8 oli in burned area mapping for a heterogenous savannah landscape. *PLOS ONE*, 2020.

[10] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

[11] World Resources Institute. 2017 Was the Second-Worst Year on Record for Tropical Tree Cover Loss. https://www.wri.org.

[12] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

[3] https://rainforestchallenge.blob.core.windows.net/dataset/sentinel_vv_image_alignment_train.json

[4] https://rainforestchallenge.blob.core.windows.net/dataset/sentinel_vh_image_alignment_train.json