Disaster Change Captioning

Antonin Faure
Syrielle Montariol[*] **Antoine Bosselut**[†]
{firstname.name}@epfl.ch

Abstract

Understanding the change between two images is a key task in computer vision, encompassing sub-tasks such as change detection (CD) and answering questions (VQA) about that change, its causes, and consequences. A primary application domain is remote sensing using satellite images, particularly for disaster relief, urban tracking, agricultural development, and deforestation monitoring. Current datasets often focus on urban changes, with disaster relief datasets being limited in size and scope.

In this research, we propose a comprehensive dataset of paired multispectral aerial images with manually annotated pixel-level disaster-specific changes for over 60 disasters. This dataset is enriched with auxiliary information, including fine-grained terrain data from OpenStreetMap, ESA WorldCover, and textual information from news articles and disaster relief reports from ReliefWeb.

The primary focus of this work is on the task of change detection by providing baselines for this task and discussing the intensive efforts and challenges involved in creating and annotating the dataset. It also highlights the potential of integrating diverse data sources to enhance disaster analysis and sets the stage for future work in expanding the dataset and exploring additional tasks such as image classification, semantic segmentation, visual question answering, and report generation. The datasets and models will be made public to facilitate transfer learning and enable rapid progress in remote sensing applications.[1]

1 Introduction

Change detection in remote sensing imagery plays a pivotal role in disaster response and environmental monitoring. The ability to accurately identify and analyze changes in imagery before and after an event can significantly enhance disaster preparedness and recovery efforts. While various datasets and models exist for urban change detection, the domain of disaster response requires specialized approaches due to the unique nature of the data and the critical importance of timely and precise analysis. This study leverages a novel multimodal dataset, integrating multispectral aerial images with auxiliary data sources including terrain information and disaster-specific textual reports, to develop and evaluate models for disaster response tasks.

2 Related Work

2.1 Datasets

Natural disaster datasets can include various types of images such as ground-level images, satellite imagery, and aerial imagery (UAV). Each type of imagery offers unique advantages and challenges for disaster analysis.

(Rahnemoonfar et al., 2021) propose **FloodNet**, which consists of high-resolution images captured only after flood disasters. FloodNet supports three tasks: classification (binary: flooded or not), semantic segmentation (10 classes specific to flood disasters), and visual question answering (VQA). The dataset includes approximately 4500 question-image pairs categorized into four groups: “Simple Counting”, “Complex Counting”, “Yes/No”, and “Condition Recognition”. Additionally, 2343 images are annotated with 9 classes, such as building-flooded, building-non-flooded, road-flooded, road-non-flooded, water, tree, vehicle, pool, and grass.

The **xBD** dataset (Gupta et al., 2019) focuses on building damage assessment and includes around 700,000 building annotations across over 5,000 km² of imagery from 15 countries, covering 6 different types of disasters. The dataset provides pairs of images captured before and after the disasters.

AIDER (Kyrkou and Theocharides, 2019,

[*] Supervisor

[†] Professor

[1] Code, datasets and pre-trained models are available at https://github.com/epfl-nlp/disaster-change-captioning

2020) contains images for four disaster events: Fire/Smoke, Flood, Collapsed Building/Rubble, and Traffic Accidents, mainly comprising high-resolution aerial images. The primary task is disaster classification into these four categories.

The **Ida-BD** dataset (Roueche et al., 2022) includes 87 pre- and post-disaster satellite imagery pairs with very high resolution (0.5m/pixel) from Hurricane Ida 2021 in Louisiana, USA. Similar to the xBD dataset, it uses polygons to represent building segments and provides four damage categories.

RescueNet (Chowdhury et al., 2022) consists of 4494 high-resolution post-disaster UAV images from Hurricane Michael, annotated for tasks such as classification and semantic segmentation, with pixel-level annotations for 10 classes and different damage levels.

RESCUENET-VQA (Sarkar and Rahnemoonfar, 2023) includes 103,192 image-question-answer triplets derived from images captured after Hurricane Michael. It features 206 unique questions across 9 categories: Simple Counting, Complex Counting, Building Condition Recognition, Road Condition Recognition, Level of Damage, Risk Assessment, Density Estimation, Positional, and Change Detection.

The **AIST Building Change Detection (ABCD)** dataset includes post-tsunami images, focusing on building change detection.

(Chen et al., 2018) provides data from two sources, including crowdsourced annotated DigitalGlobe satellite imagery and data collected by FEMA, covering Hurricane Harvey with both raster (satellite and aerial imagery) and vector data (auxiliary building damage information).

(Yuan et al., 2022) introduced the **CDVQA** dataset, which includes multi-temporal image-question-answer triplets generated using an automatic question-answer generation method.

LEVIR-CD (Chen and Shi, 2020) is a remote sensing building change detection dataset comprising 637 very high-resolution (VHR, 0.5m/pixel) Google Earth (GE) image patch pairs with a size of 1024 × 1024 pixels, with bitemporal images spanning 5 to 14 years.

(Pang et al., 2024) discusses large-scale pre-training datasets.

2.2 Models for VQA in Remote Sensing

Current-generation Visual Language Models (VLMs) are not well-adapted for remote sensing, which includes multispectral (like Sentinel-2), non-optical, or multi-temporal images (for tasks like change detection and captioning) (Zhang and Wang, 2024). Most remote sensing tasks involve some form of classification, which requires fine-tuning or at least in-context learning abilities, or segmentation.

2.3 Event Dataset Collection

João Pedro, who previously worked on this project last semester, collected an extensive dataset of 5045 unique events from various sources. These events cover multiple disaster categories, including Flood, Earthquake, Wildfire, Explosion (Industrial), and more. The sources include:

- EM-DAT: The Emergency Events Database
- Wikipedia
- Wikidata

3 Datasets

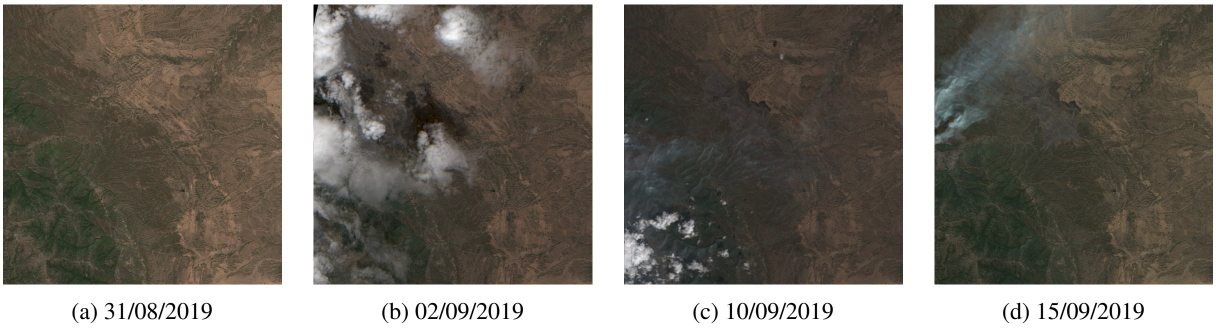
3.1 Sentinel-2 Images

The image collection process for our disaster-related dataset relies on the Copernicus Sentinel-2 Surface Reflectance (Level-2A) Harmonized data, accessible via Google Earth Engine (GEE). This dataset offers high-quality, multispectral imagery, which is vital for accurately capturing the characteristics of disaster events and their impacts on the environment. In addition, for each image we saved its metadata including the acquisition date, sensor details, cloud cover percentage, and the exact coordinates of the tiled areas.

3.1.1 Data Collection

Sources We exclusively use the *COPERNICUS/S2_SR_HARMONIZED*[2] collection from Sentinel-2, which starts from **March 28, 2017**. This dataset provides images with 12 spectral bands, offering valuable information across different wavelengths, including visible, near-infrared, and short-wave infrared. Each image's spatial resolution ranges from 10 meters, for the RGB bands, to 60 meters depending on the band. Due to the size constraints of direct downloads from GEE, we used the library *geemap* which allows large file processing by directly downloading all bands into a single .tif raw file.

[2] https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED


(a) 31/08/2019

(b) 02/09/2019

(c) 10/09/2019

(d) 15/09/2019

Figure 1: Sentinel-2 images temporal evolution for a Wildfire event

Temporal Coverage To adequately capture the state of an area before and after a disaster, we gather images within a **90-day window** centered on the disaster date. This period is divided to include 45 days before and 45 days after the event, allowing for an analysis of the preliminary conditions and the subsequent recovery or degradation. Within this timeframe, we prioritize acquiring the first 10 images closest to the disaster date to ensure we capture the most relevant changes.

Spatial Coverage For each disaster, the spatial extent of the imagery encompasses a $30km \times 30km = \mathbf{900km^2}$ area centered on the reported disaster location. This size is typically sufficient to cover the affected area and some of the surrounding regions, providing a comprehensive view of the disaster's impact.

Cloud Filtering & Image Quality Given the frequent cloud cover issues in optical satellite imagery, we apply strict cloud filtering criteria to select only those images where cloud cover is less than **10%** of the total area using the *QA60* band of Sentinel-2 representing the cloud mask. This step is crucial to ensure that the imagery is usable for visual assessments and algorithmic processing.

3.1.2 Data Processing

To generate a RGB preview we used the rasters of bands B2, B3, B4 and first apply a gamma correction $V = V^{\frac{1}{\gamma}}$ with $\gamma = 2$ followed by a min-max scaling to obtain a 0-255 scale. Additionally, the images were reprojected to the EPSG:3857[3] projection to ensure alignment with OSM data, which uses this projection.

3.1.3 Limitations

- **Event Retrieval:** Only a small proportions of events are retrieved (before 2019 almost all events fail to retrieve at least one image in the region across the 90 days period).
- **Image artifacts:** Some images have blacks segments due to the way sentinel orbit around the earth and doesn't collect data for every region on earth.
- **Geometric shifts:** Sometimes the images have 1 or 2 pixels of shift (due either to the API or the image processing that remap the image into the EPSG:3857 projection)
- **RGB Processing Variability:** The image processing of RGB preview isn't an exact process since the rasters data can vary a lot across images (some Sentinel-2 images can have pure black/white/red/blue/green pixels which are probably due to the satellite itself)

3.2 OpenStreetMap

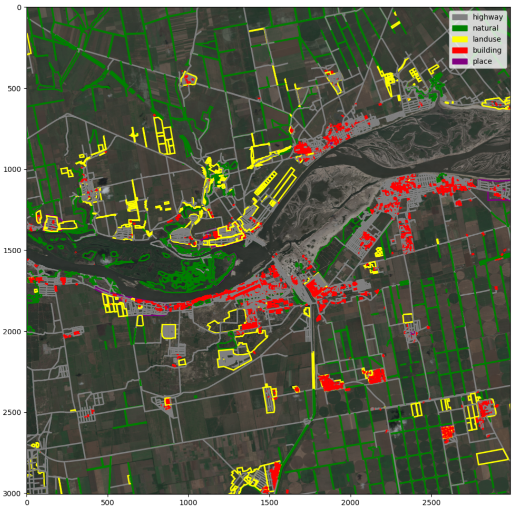
To enhance our dataset with contextual geographic information, we integrated data from OpenStreetMap (OSM). This integration provides detailed information about various geographic features that can significantly aid in the analysis of disaster impact.

3.2.1 Data Collection

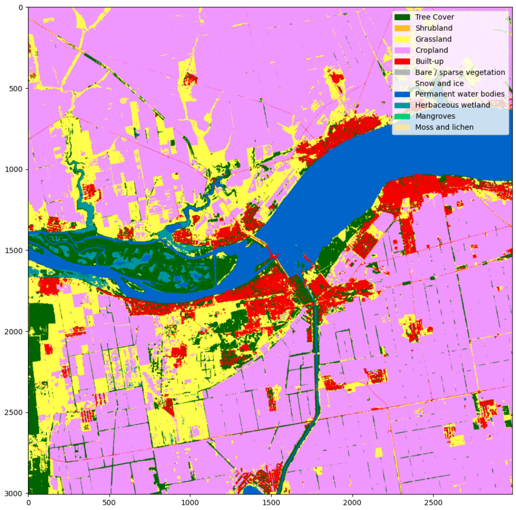
We used the *Overpass API* to query OSM for relevant geographic features within the bounding boxes of our disaster event areas. Our queries targeted features tagged under categories like highways, natural elements (e.g., forests, water bodies), land use (e.g., residential areas, orchards), buildings and administrative area. These categories were selected due to their relevance in disaster impact analysis.

The bounding boxes for our queries were determined based on the extent of the affected area for each disaster event. This ensures that the retrieved OSM data accurately corresponds to the regions of interest in our dataset.

[3] https://epsg.io/3857



(a) Example of OSM data overlay on a Flood event's satellite image.



(b) Example of WorlCover data of a Flood event

3.2.2 Data Processing

The raw OSM data extracted using Overpass API were processed to align with our disaster dataset's spatial and temporal resolutions. This involved converting vector data into raster format to match the Sentinel-2 imagery's 10-meter resolution.

3.2.3 Limitations

Despite the richness of OSM data, several limitations exist:

- **Incomplete Coverage:** OSM data may not be comprehensive for all regions, particularly in less urbanized or less developed areas.
- **Update Frequency:** The data might not always be up-to-date, as OSM relies on voluntary contributions which can vary significantly in frequency and accuracy.
- **Variability in Detail:** The level of detail can vary widely between regions. Some areas may have highly detailed annotations, while others may lack critical geographic features.
- **Temporality:** We only took the latest data for the given region and not the data at the time of the event assuming their were no changes.

3.3 Basemap: ESA WorldCover v200

To complement the OSM data and provide detailed information about ground types, we utilized the ESA WorldCover v200 [4] dataset. This dataset offers global land cover maps at a 10-meter resolution, making it suitable for our disaster impact analysis.

3.3.1 Data Collection

The ESA WorldCover v200 data from 2020 were accessed through Google Earth Engine (GEE). This dataset provides detailed classifications of land cover types, such as forests, grasslands, urban areas, water bodies, and more. The global coverage and high resolution of this dataset make it an excellent resource for understanding the baseline land cover conditions before and after disasters.

3.3.2 Data Processing

The WorldCover data were processed to ensure spatial alignment with our Sentinel-2 imagery. This involved reprojecting the WorldCover data to match the EPSG:3857 projection used in our dataset and resampling it to a 10-meter resolution. This processing step ensures that the land cover data can be accurately overlaid on our satellite imagery for comprehensive analysis.

3.3.3 Limitations

While the ESA WorldCover dataset is highly valuable, it also has some limitations:

- **Temporal Discrepancy:** The dataset is from 2020, which may not perfectly align with the disaster events' timeframes. Assumptions were made that land cover types have not significantly changed within the short periods around each disaster event (urban development being the most probable source of change)
- **Resolution Constraints:** Although the 10-meter resolution is relatively high, certain

[4] https://developers.google.com/earth-engine/datasets/catalog/ESA_WorldCover_v200

small-scale changes or fine details might still be missed.

3.4 ReliefWeb

To further enrich our dataset, we incorporated textual information from ReliefWeb[5], a specialized platform that provides comprehensive reports and data on disasters worldwide. This data can be particularly useful for question-answering tasks by providing detailed context and descriptions of events.

3.4.1 Data Collection

Disaster Events Retrieval We queried the ReliefWeb API to retrieve data for events with the same types as those in our dataset and during the same period (2017-present). Each retrieved disaster event includes an identifier, type, name, description, and the date of occurrence.

Report Collection For each disaster event, we collected associated reports using the ReliefWeb API. These reports provide detailed textual information, including the body of the report, publication date, title, summary, and source. The reports are stored in a structured format, enabling easy integration and further analysis.

3.4.2 Limitations

- **Event Matching**: The matching between our dataset and ReliefWeb data is not exact, which can lead to inconsistencies.
- **Merging Datasets**: The inner join matching of our data with ReliefWeb data is relatively small, limiting the extent of enriched information available.

4 Annotations

To facilitate the development of accurate and reliable models for disaster impact analysis, we utilized the CVAT (Computer Vision Annotation Tool)[6] for annotating our dataset. This process involves manually labeling the images to highlight changes caused by various types of disasters. The annotation focuses on events that are prone to visible changes at a 10m/pixel resolution, excluding events like earthquakes where changes are typically not visible from low resolution satellite imagery.

The annotation process is extensive and requires meticulous attention to detail. Out of 1056 retrieved events, 300 where evaluated by hand out of which only **68 samples** have been deemed exploitable and annotated.

4.1 Visibility subdatasets

Based on the visibility of changes in the satellite images, we generated three types of samples:

1. **Visible**: The changes are clearly visible from the satellite images.
2. **Tiny-visible**: The changes are visible pixel-wise but only cover a tiny area.
3. **Not visible**: The change is not visible from the satellite image, but the affected entity/area is known (cross-referenced with news and Google Maps to locate the exact position) and annotated.

The distribution of the annotated samples is shown in Fig 4. We can note that:

- **Visible** (36 samples): Wildfires are the most visible type of disaster in our dataset, as their effects are often extensive and easily detectable from satellite imagery. Some floods also show clear visible changes, particularly in areas where water bodies overflow into surrounding regions.
- **Tiny-visible** (13 samples): These are mostly related to industrial fires, explosions, or small landslides, where the affected area is relatively small.
- **Not Visible** (19 samples): These typically involve fire or explosion events affecting a single building or a small area, usually in an urban environment.

5 Change Detection

The task of change detection is crucial for analyzing the impact of disasters. For this task, we focus exclusively on the *visible* subdataset, as the other datasets do not provide relevant information for visual change detection.

5.1 Evaluated Models

We evaluated several state-of-the-art (SOTA) models on this task on models cited by (Jiang et al., 2023) or (Corley et al., 2024) that include:

- **FC-EF** (Caye Daudt et al., 2018): A single-stream network, where two images are concatenated as a single input and fed into a full convolutional network (FCN).

[5] https://reliefweb.int

[6] https://github.com/cvat-ai/cvat

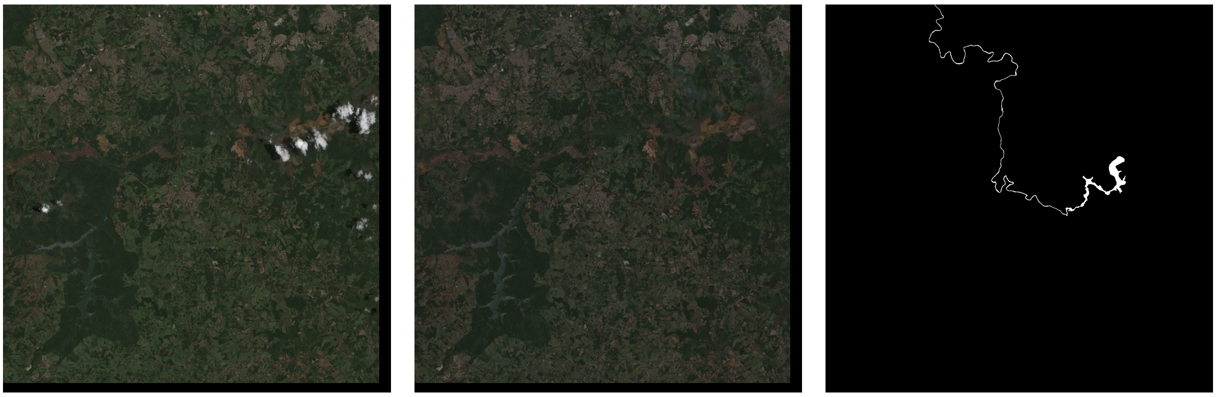

(a) Before (b) After (c) Annotated Change

Figure 3: Example of before, after, and annotated changes for a dam rupture (in the *tiny-visible* subset)

	Method	Precision	Recall	F1	OA	IoU
Pretrained	FC-EF	**96.90**	1.01	2.01	95.06	1.01
	FC-Siam-Diff	70.42	0.11	0.23	95.02	0.11
	FC-Siam-Conc	95.04	0.75	1.50	95.05	0.75
	BIT	77.69	2.30	4.48	95.09	2.29
	TinyCD	69.86	0.12	0.25	95.02	0.12
	SeCo	1.19	8.49	2.09	60.43	1.05
Trained	FC-EF (ep. 14)	28.68	50.81	**36.67**	93.00	**22.45**
	FC-Siam-Diff (ep. 19)	18.82	34.14	24.26	91.50	13.81
	FC-Siam-Conc (ep. 5)	87.27	14.39	24.71	**96.50**	14.10
	BIT (ep. 3)	15.14	12.79	13.86	93.66	7.45
	TinyCD (ep. 10)	8.97	**71.18**	15.94	70.09	8.66

Table 1: Change detection results on pretrained and trained models for the *visible* dataset test split

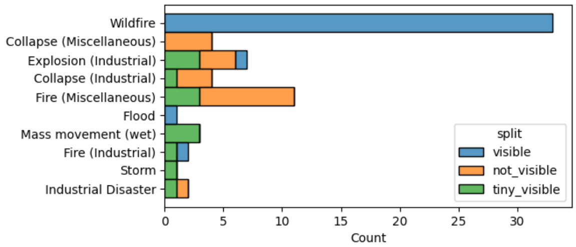


Figure 4: Distribution of disaster types across the three subdatasets

- **FC-Siam-Diff** (Caye Daudt et al., 2018): A dual-stream network, where two images are extracted features by using two FCN encoders, and the difference operation is first performed on the two image features, and the extracted difference features at different levels are input to the FCN decoder

- **FC-Siam-Conc** (Caye Daudt et al., 2018): A dual-stream network, where two images are extracted by two FCN encoders respectively, and the features are concatenated together and input to a FCN decoder.

- **BIT** (Chen et al., 2021): A dual-stream network, which extracts high-level features via convolutional networks and constructs semantic tokens by using a Transformer.

- **SeCo** (Mañas et al., 2021): An architecture that uses a ResNet encoder backbone for feature extraction with a U-Net decoder for segmentation tasks. It employs multiple projection heads to create distinct embedding subspaces, which are optimized for seasonal invariance and variance using contrastive learning on temporal and artificial augmentations.

- **TinyCD** (Codegoni et al., 2022): An architecture that uses an EfficientNet backbone to extract convolutional features to feed to a custom attention-based decoder network

It is important to note that all the pretrained baselines were trained on other datasets like LEVIR-CD (Chen and Shi, 2020), which is a building change detection dataset with a resolution of 1024 x 1024

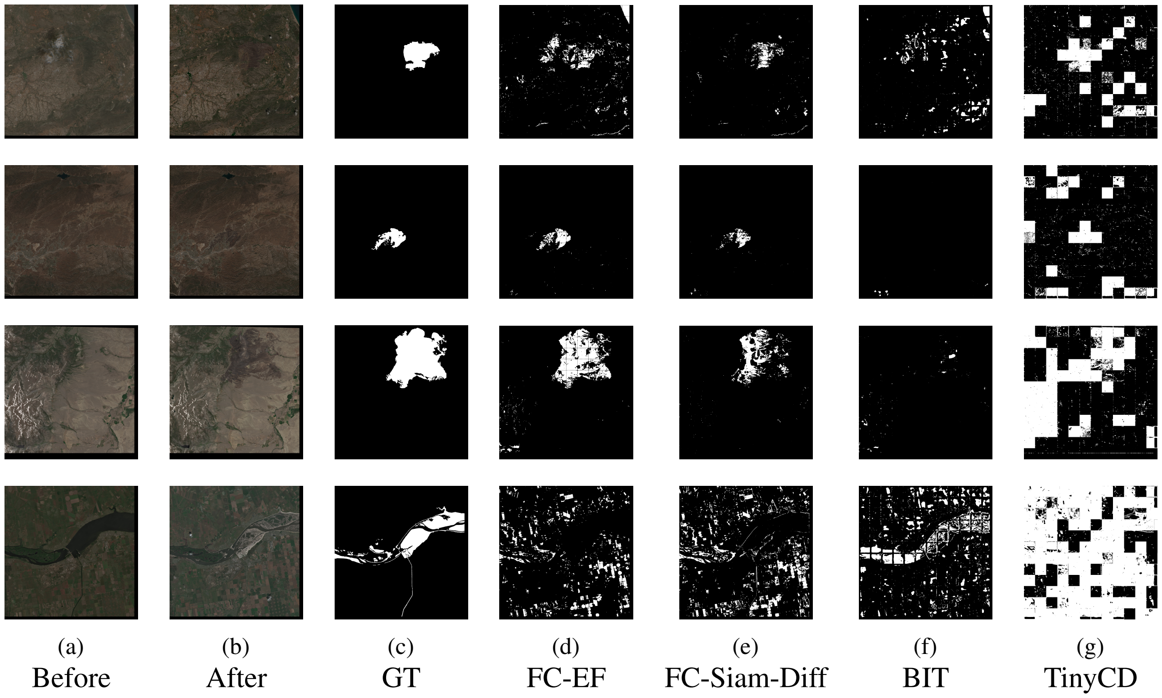


Figure 5: Visualization of change detection of *trained* models on some test set samples

pixels and 0.5m/pixel. Our dataset has a resolution of 10m/pixel, making direct transfer of models challenging even though the patch sizes are 256 x 256 pixels in both cases. The SeCo model had only its ResNet encoder pretrained, leading to non-reliable results for the UNet decoder (untrained) across different runs.

5.2 Training Setup

For our experiments, we trained these models on a split of our dataset. The dataset was divided as follows:

- Training set: 25 samples (70%)
- Validation set: 4 samples (10%)
- Test set: 7 samples (20%)

To enhance the diversity of the training set, data augmentation was applied by randomly flipping and rotating the images by 90° increments between epochs. Models were trained for 20 epochs with a batch size of 1 and patch size of 256x256, using the Adam optimizer with a weight decay of $1e-4$. The learning rate followed an exponential decay schedule with a decay rate of 0.95 per epoch. For the loss function, we used BCEWithLogitsLoss for SeCo and TinyCD models, and CrossEntropyLoss for the others, with dynamically computed class weights to handle class imbalance. The training was conducted on one NVIDIA GeForce GTX TITAN X (12GB VRAM).

5.3 Evaluation Metrics

We used five evaluation metrics to assess the performance of the change detection algorithms, based on (Chen et al., 2021) and defined as follows:

$$\mathit{Precision} = \frac{TP}{(TP+FP)} \tag{1}$$

$$\mathit{Recall} = \frac{TP}{(TP+FN)} \tag{2}$$

$$\mathit{F1} = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \tag{3}$$

$$\mathit{IoU} = \frac{TP}{(TP+FN+FP)} \tag{4}$$

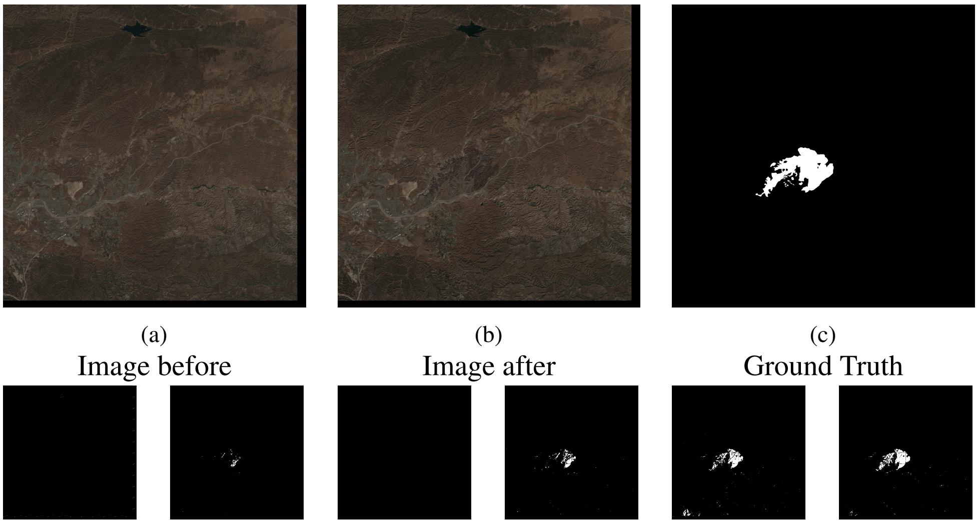
$$\mathit{OA} = \frac{(TP+TN)}{(TP+TN+FN+FP)} \tag{5}$$

Where TP, TN, FP, and FN represent the number of true positive, true negative, false positive, and false negative, respectively.

5.4 Results

The results for both pretrained and trained models are summarized in Table 1. Visual examples of the change detection results are shown in Fig. 5. To illustrate the training process, we present the F1 scores for the training and validation sets across 20 epochs.

All in all the TinyCD and BIT are not training as efficiently as the FCN methods (see Fig. 7) and all methods are not generalising well (see Fig. 8).



(a)
Image before

(b)
Image after

(c)
Ground Truth

(d) Epoch 1

(e) Epoch 2

(f) Epoch 3

(g) Epoch 4

(h) Epoch 5

(i) Epoch 14

Figure 6: Visualization of change detection on a *test* sample over epochs of the FC-EF model training

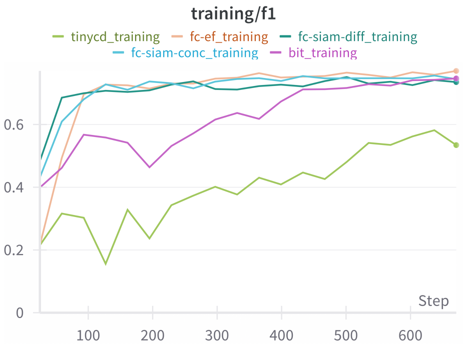


Figure 7: F1 score of the FC-EF model on the *training* set (across 20 epochs)

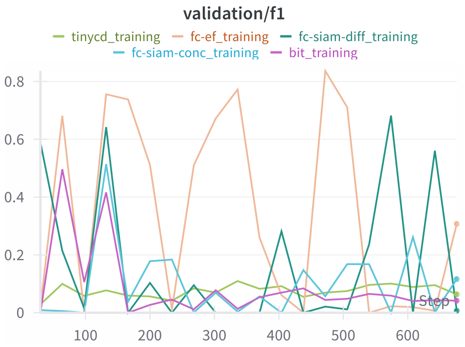


Figure 8: F1 score of the FC-EF model on the *validation* set (across 20 epochs)

However the FC-EF seems the most promising in early generalization across epochs as shown in Fig. 6 with a test sample.

6 Conclusion

This research developed a detailed dataset tailored for disaster impact analysis using multispectral aerial images from Sentinel-2, enriched with contextual data from OpenStreetMap (OSM), ESA WorldCover, and textual reports from ReliefWeb. The integration of these diverse data sources allows for a nuanced understanding of disaster impacts across different scenarios.

The annotation process proved to be highly time-intensive, with a stringent selection criterion that resulted in a small proportion of the initially retrieved samples being used. The dataset was divided into three visibility-based subsets: visible, tiny-visible, and not visible, facilitating specific analyses tailored to the characteristics of each subset.

A proof of concept for change detection tasks was performed. The limited scope of this initial exploration sets the stage for a future larger dataset that could support SOTA models training in change detection and VQA for disaster events.

7 Future Work

Looking ahead, the project will benefit from:

- Expanding the dataset annotations to include a wider array of disaster types and more samples.
- Streamlining the process for synchronizing event data with corresponding news and textual reports to enhance the dataset's reliability and comprehensiveness.

- Further developing the change detection framework by refining model training and exploring the use of Sentinel-2 full spectral data (12 bands) beyond the RGB channels, combined with the WorldCover layer.
- Implementing additional tasks such as visual question answering (VQA) to leverage the multimodal nature of the dataset fully.

References

R. Caye Daudt, B. Le Saux, and A. Boulch. 2018. Fully convolutional siamese networks for change detection. In *IEEE International Conference on Image Processing (ICIP)*.

Hao Chen, Zipeng Qi, and Zhenwei Shi. 2021. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14.

Hao Chen and Zhenwei Shi. 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10).

Sean Andrew Chen, Andrew Escay, Christopher Haberland, Tessa Schneider, Valentina Staneva, and Youngjun Choe. 2018. Benchmark dataset for automatic damaged building detection from post-hurricane remotely sensed imagery. *arXiv preprint arXiv:1812.05581*.

Tashnim Chowdhury, Robin Murphy, and Maryam Rahnemoonfar. 2022. Rescuenet: A high resolution uav semantic segmentation benchmark dataset for natural disaster damage assessment. *arXiv preprint arXiv:2202.12361*.

Andrea Codegoni, Gabriele Lombardi, and Alessandro Ferrari. 2022. Tinycd: A (not so) deep learning model for change detection. *arXiv preprint arXiv:2207.13159*.

Isaac Corley, Caleb Robinson, and Anthony Ortiz. 2024. A change detection reality check. *Preprint*, arXiv:2402.06994.

Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. 2019. Creating xbd: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Bo Jiang, Zitian Wang, Xixi Wang, Ziyan Zhang, Lan Chen, Xiao Wang, and Bin Luo. 2023. Vct: Visual change transformer for remote sensing image change detection. *Preprint*, arXiv:2310.11417.

Christos Kyrkou and Theocharis Theocharides. 2019. Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles. In *CVPR workshops*, pages 517–525.

Christos Kyrkou and Theocharis Theocharides. 2020. Emergencynet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:1687–1699.

Oscar Mañas, Alexandre Lacoste, Xavier Giro-i Nieto, David Vazquez, and Pau Rodriguez. 2021. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. *arXiv preprint arXiv:2103.16607*.

Chao Pang, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Xingxing Weng, Shuai Wang, Litong Feng, Gui-Song Xia, et al. 2024. H2rsvlm: Towards helpful and honest remote sensing large vision language model. *arXiv preprint arXiv:2403.20213*.

Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. 2021. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654.

David B Roueche, Jordan O Nakayama, Barbaros M Cetiner, K Sabarethinam, and T Kijewski-Correa. 2022. Hybrid framework for post-hazard building performance assessments with application to hurricanes. In *Proceedings of the 14th Americas Conference on Wind Engineering*.

Argho Sarkar and Maryam Rahnemoonfar. 2023. Rescuenet-vqa: A large-scale visual question answering benchmark for damage assessment. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 1150–1153. IEEE.

Zhenghang Yuan, Lichao Mou, Zhitong Xiong, and Xiao Xiang Zhu. 2022. Change detection meets visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13.

Chenhui Zhang and Sherrie Wang. 2024. Good at captioning, bad at counting: Benchmarking gpt-4v on earth observation data. *arXiv preprint arXiv:2401.17600*.