SentinelKilnDB: A Large-Scale Dataset and Benchmark for OBB Brick Kiln Detection in South Asia Using Satellite Imagery Supplementary Information

Rishabh Mondal[1], Jeet Parab[2], Heer Kubadia[1], Shataxi Dubey[1], Shardul Junagade[1], Zeel B Patel[1], Nipun Batra[1]

[1]IIT Gandhinagar, India, [2]IIIT Surat, India

1 Data: Access and Datasheet

1.1 Dataset Access

The SentinelKilnDB datasets are available for downloading from Kaggle:
https://kaggle.com/datasets/3eb8e7201b14b158ed841718cb777c5b94a6a6375aaa8499c7376ec831f8d879

The Croissant metadata can be downloaded from Google Drive:
https://drive.google.com/file/d/1pxOOkVmaJ-YDG1p980UdIYE7ayy-s4yG/view?usp=drive$_{l}ink$

The Croissant metadata verifier can be downloaded from:
https://drive.google.com/file/d/1UfiikISgrsGmpySj4bb-IJRdKroHrdSp/view?usp=drive$_{l}ink$

1.2 Datasheet

1.2.1 Motivation

We provide a datasheet for our dataset based on the methodology outlined in "Datasheets for Datasets" by Timnit Gebru et al. [1]. The questions are presented in blue, with our corresponding responses shown in black.

For what purpose was the dataset created? Was there a specific task in mind?

This dataset was created for academic and research purposes to advance scientific understanding and support policy development on air quality and sustainability issues. The findings highlight important opportunities to improve regulatory compliance and encourage the adoption of cleaner technologies within the brick kiln sector, which is a significant contributor to regional air pollution.

Beyond its environmental relevance, this dataset is especially valuable for the fields of object detection and computer vision. It provides a large-scale, hand-validated collection of brick kiln locations annotated with oriented bounding boxes (OBBs) on freely available Sentinel-2 satellite imagery. The use of OBBs allows for precise localization and orientation of kilns, which is critical for accurately estimating their size, type, and emission impact. Because the dataset is based on low-resolution but globally accessible imagery, it presents unique challenges and opportunities for developing robust and scalable object detection models. Researchers can use it to test and improve algorithms for detecting small objects in complex, real-world environments. This makes the dataset a valuable benchmark for advancing machine learning techniques in remote sensing, environmental monitoring, and sustainable development.

Preprint. Under review.

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 Rishabh Mondal, Jeet Parab, Heer Kubadia, Shataxi Dubey, Shardul Junagade, Zeel B. Patel, and Nipun Batra and the team of Sustainability Lab, IIT Gandhinagar, India. More information about the lab is available at: https://sustainability-lab.github.io/

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 IIT Gandhinagar.

Any other comments?

We collected Sentinel-2 satellite images using bands B4 (Red), B3 (Green), and B2 (Blue), each at a 10-meter resolution, from September 2023 to February 2024. We chose this period because brick kilns in the Indo-Gangetic Plain are most active during winter, when agricultural activity is low and visibility conditions are optimal. To ensure image quality, we filtered images with only less than 1% cloud cover and used the QA60 band to mask cloudy pixels. We divided each image into 128 × 128 pixel patches using a sliding window of 0.0027° (about 30 pixels) to capture kilns near patch edges. Finally, we removed duplicate detections by matching their geographic coordinates.

1.2.2 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.

The dataset comprises 78,707 RGB images of brick kilns, each with a spatial resolution of 10 meters per pixel and dimensions of 128 × 128 pixels, captured by the Sentinel-2 satellite. Annotations are provided in three formats: axis-aligned bounding boxes, oriented bounding boxes, and the DOTA format.

Each annotation includes the category of the brick kiln and the coordinates of the bounding box.

1. The axis-aligned bounding box format includes the kiln category, the coordinates of the bounding box center, and its width and height, all expressed in normalized values.
2. The oriented bounding box format provides the kiln category along with the normalized coordinates of the four corners: top-left, top-right, bottom-left, and bottom-right.
3. The DOTA format specifies the four corner coordinates of the bounding box, the kiln category, and a difficulty level in the following structure: (x1, y1, x2, y2, x3, y3, x4, y4, class name, difficult).

The images were collected between September 2023 and February 2024 and are of high visual quality, with cloud cover below 1%.

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

There are 105,933 unique instances of brick kilns in the dataset. There are three categories of the brick kilns: Clamp-Fired Continuous Brick Kilns (CFCBK), Fixed Chimney Bull's Kilns (FCBK), and Zigzag kilns.

Due to an approximate overlap of 30 pixels between adjacent images, some kilns may appear in more than one image. Taking this into account, the distribution of kiln categories in the dataset is as follows:

The dataset contains:

1. 3328 CFCBK
2. 56,943 FCBK
3. 45,658 Zigzag kilns

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).

Our dataset covers a large area of 2.8 million square kilometers across South Asia, including parts of Bangladesh, Pakistan, Afghanistan, and ten states in India. It focuses on the Indo-Gangetic Plain, also known as the "brick belt," which is a densely populated and highly polluted region where brick kilns contribute 8–14% of the ambient air pollution. Brick kilns are especially prominent here because the region has high demand for bricks due to rapid urbanization and construction, and it has abundant clay resources ideal for brick manufacturing.

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

Yes, each instance is paired with a label from the three types: CFCBK, FCBK, or Zigzag.

Is any information missing from individual instances?

All relevant information for each instance is provided in Section 3.4 of the main paper.

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.

The dataset is split using a class-wise stratified approach to ensure balanced class representation in each set. The splits follow a 60:20:20 ratio for training, validation, and testing, respectively.

1. The training set includes 47,214 images and 47,214 annotation files. The training set contains 63,787 instances with 2,032 CFCBK class, 34,292 instances of FCBK class and 27,463 instances of Zigzag class
2. The validation set consists of 15,738 images and 15,738 annotation files, containing 21,042 instances in total, 649 from the CFCBK class, 11,339 from the FCBK class, and 9,054 from the Zigzag class
3. The test set also includes 15,738 images and 15,738 annotation files, with a total of 21,100 instances, 647 from the CFCBK class, 11,312 from the FCBK class, and 9,141 from the Zigzag class.

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

The dataset initially contained redundancies due to overlapping pixels between nearby brick kilns, caused by splitting large Sentinel-2 satellite images ($11{,}000 \times 11{,}000$ pixels) into smaller 128×128 pixel patches with a 30-pixel (0.0027°) overlap in both latitude and longitude. This overlap was experimentally determined to be sufficient for capturing brick kilns located near patch edges without losing important details. However, it also caused some kilns to appear in multiple patches. To address this, we applied a deduplication process based on geographic coordinates, matching and removing duplicate kiln instances to retain only unique entries. This careful preprocessing reduces redundancy and improves the dataset's accuracy and reliability. Additionally, the image size and sliding window parameters were optimized to balance coverage and computational efficiency, ensuring high-quality data for object detection and analysis.

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.

Yes, the dataset is self-contained, as we provide all the satellite images, labels, and metadata together. However, the imagery itself is sourced from the Sentinel-2 satellite, which is operated by the European Space Agency (ESA) and comes with its own licensing terms. While Sentinel-2 data is freely available for public use, users must comply with ESA's license conditions, which generally allow free use for research and non-commercial purposes. There are no fees or restrictions beyond the standard Sentinel-2 data policy. Users can download, store, and use the dataset to develop new object detection methods and improve the dataset, but should refer to ESA's guidelines for any specific usage requirements.

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' non-public communications)? If so, please provide a description.

No, Sentinel-2 imagery is free to use for non-commercial purposes and is publicly accessible.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

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

1.2.3 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 constructed using publicly available Sentinel-2 satellite imagery. In our previous research, Space to Policy [2], we identified and annotated 30,638 brick kilns across five Indo-Gangetic states—Punjab, Bihar, Haryana, Uttar Pradesh, and West Bengal—using high-resolution Planet imagery.

To integrate these annotations with the Sentinel-2 data, we overlaid the Planet imagery labels onto the corresponding cropped, geo-referenced Sentinel-2 images. Since Planet imagery is provided in the Web Mercator projection (EPSG:3857) and Sentinel-2 imagery uses the Universal Transverse Mercator (UTM) projection (EPSG:326XX, depending on the zone), we reprojected the oriented bounding box (OBB) coordinates from Web Mercator to the appropriate UTM zone of each Sentinel-2 tile. This transformation was performed using geolocation metadata and spatial reference conversion techniques.

To ensure precise alignment between the datasets, we conducted thorough visual inspections using Geemap [3] and validated the results quantitatively through Intersection over Union (IoU) metrics.

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[1].

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.

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 was built using satellite imagery from September 2023 to February 2024.

1.2.4 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.

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

We preprocessed the Sentinel-2 images, with detailed steps described in Section 3.2. The entire dataset was manually validated, and to the best of our knowledge, the data is highly accurate.

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 by the source of imagery.

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.

1.2.5 Uses

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?

This dataset can support a wide range of research and development tasks beyond its primary objective. Some potential use cases include:

- **Oriented Object Detection:** The dataset includes oriented bounding box (OBB) annotations, making it an excellent testbed for training and evaluating models that detect objects with arbitrary orientations—particularly relevant for remote sensing and aerial imagery.
- **Domain Generalization Research:** Covering diverse regions across South Asia, the dataset is ideal for studying how models trained in one geographic area generalize to others, advancing research in domain adaptation and generalization.
- **Super-Resolution Algorithm Development:** The high-resolution object annotations support the development and benchmarking of super-resolution algorithms aimed at enhancing the quality of low-resolution satellite images.
- **Self-Supervised Learning:** The dataset can be used to pretrain models in a self-supervised manner by leveraging the spatial and temporal patterns of geospatial data to learn useful feature representations without the need for explicit labels.
- **Active Learning Scenarios:** With a large volume of unlabelled satellite data available, the dataset is well-suited for simulating active learning tasks where models iteratively select the most informative samples for annotation.
- **Environmental Monitoring and Compliance:** By identifying and classifying brick kilns, the dataset can aid regulatory agencies in automatically monitoring pollution sources, contributing to data-driven compliance with environmental standards.

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

None.

1.2.6 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, we will make the dataset publicly available on the internet upon acceptance.

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

The dataset can be downloaded from Kaggle platform. The dataset currently does not have a DOI.

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 imagery, users are also requested to refer to the Sentinel Terms of Service.[2]

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.

1.2.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset is hosted and supported by Kaggle Platform.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Rishabh Mondal and Nipun Batra can be contacted via email (rishabh.mondal@iitgn.ac.in, and nipun.batra@iitgn.ac.in).

1.3 Super-Resolution Comparisons

Figure 1 presents a comparative visualization of super-resolution techniques applied to a specific patch as captured by Sentinel-2 imagery. The top row includes the original Sentinel-2 image (a), followed by higher-resolution imagery from ESRI Wayback (b) and Planet Labs (c), serving as external visual references. The second row features imagery from Google Earth (d), standard bilinear interpolation (e), and the SwinIR deep learning-based super-resolution model (f). The final row shows outputs from ESRGAN (g), HIT (h), and Stable Diffusion (i), representing various state-of-the-art generative and enhancement models.

This 3×3 grid offers a side-by-side comparison, highlighting visual differences in clarity, structure, and detail enhancement across traditional and modern super-resolution methods. It underscores the potential of advanced models to bridge the resolution gap between publicly available low-res satellite imagery and high-res commercial sources, which is critical for accurate remote sensing applications such as object detection and environmental monitoring.

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

(a) Original Sentinel Imagery (b) ESRI Wayback Imagery (c) Planet Labs Imagery

(d) Google Earth Imagery (e) Bilinear Interpolation (f) SwinIR

(g) ESRGAN (h) HIT (i) Stable Diffusion

Figure 1: Comparison of nine images over the same target Sentinel-2 tile: (i) original resolution, (ii) external sources (ESRI, Planet Labs, Google Earth), and (iii) super-resolution outputs from interpolation and deep learning methods (Bilinear, SwinIR, ESRGAN, HIT, Stable Diffusion).

2 Regional Brick Kiln Class Counts

Table 1 presents a detailed breakdown of brick kiln counts by kiln type—CFCBK, FCBK, and Zigzag—for all the states and provinces included in our dataset across South Asia. This comprehensive table complements the country-level summaries provided in the main paper by offering fine-grained regional insights. The dataset spans nine Indian states, eight administrative divisions in Bangladesh, four provinces in Pakistan, and all 34 provinces of Afghanistan.

Uttar Pradesh in India exhibits the highest concentration of kilns with a total of 19,910, followed by Bihar and West Bengal. Across Bangladesh, Dhaka and Chittagong divisions show the largest number of brick kilns. In Pakistan, Punjab province dominates with 10,473 kilns, while in Afghanistan, 609 kilns are spread across its provinces. Overall, the dataset includes 62,671 brick kilns, comprising 1,944 CFCBKs, 33,963 FCBKs, and 26,764 Zigzag kilns. This regional disaggregation helps highlight spatial patterns in kiln types, supporting more localized studies in air pollution monitoring, industrial activity, and policy design.

Table 1: Class-wise brick kiln counts for all states and provinces covered in our dataset.

Country	State/Region	CFCBK	FCBK	Zigzag	Total
India	Gujarat	57	961	45	1063
	Assam	5	1239	109	1353
	Jharkhand	17	1229	191	1437
	Haryana	1	159	2231	2391
	Rajasthan	34	2137	232	2403
	Punjab	1	444	2074	2519
	West Bengal	67	1575	2776	4418
	Bihar	58	2038	5392	7488
	Uttar Pradesh	1699	11669	6542	19910
Bangladesh	Sylhet	0	9	216	225
	Barisal	0	109	256	365
	Mymensingh	0	108	411	519
	Rangpur	0	112	894	1006
	Khulna	0	388	645	1033
	Rajshahi	0	251	880	1131
	Dhaka	2	92	1205	1299
	Chittagong	0	392	933	1325
Pakistan	Balochistan	1	63	1	65
	Sindh	1	788	2	791
	Khyber Pakhtunkhwa	0	846	2	848
	Punjab (PK)	1	8746	1726	10473
Afghanistan	34 Provinces	0	608	1	609
Total		**1944**	**33963**	**26764**	**62671**

3 Benchmark

3.1 Evaluation Setup:

For all tasks, we set the Intersection over Union (IoU) threshold to 50% and use a confidence threshold of 0.005 for detections. For task `T1`, we split the dataset into training, validation, and test sets in a

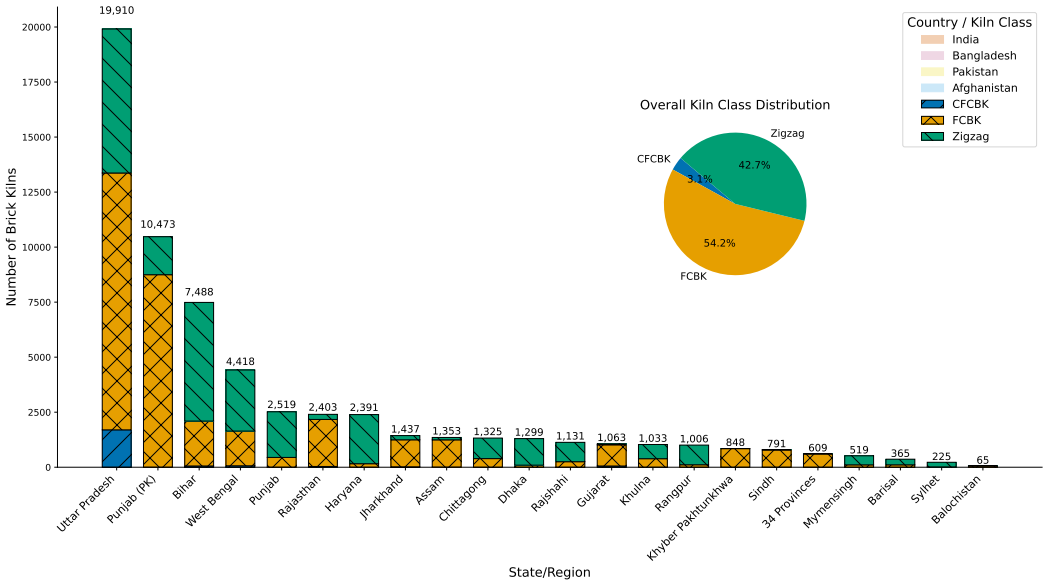

Figure 2: Visualization of the regional brick kiln counts across different states/regions, categorized by kiln classes (CFCBK, FCBK, and Zigzag). The stacked bar chart illustrates the distribution of each kiln class within each region, while the accompanying pie chart represents the overall proportion of each kiln class aggregated across all regions, providing a clear comparative overview of the kiln class distribution at both regional and aggregate levels.

60:20:20 ratio with balanced class distribution. For task T2, training is done on Uttar Pradesh data split similarly, and evaluation is performed on both in-region (Uttar Pradesh) and out-of-region data (Dhaka and Punjab). For task T3, various super-resolution methods are applied to Sentinel-2 images of the Delhi-NCR dataset, split 60:20:20, and the best detection model is tested on these enhanced images. Performance is measured using class-agnostic and class-aware mAP at 50% IoU across three classes. Experiments were run on a system with an NVIDIA A100 GPU.

3.1.1 Comparison of Detection Methods

This subsection presents a comparison of class-agnostic mean Average Precision at 50% IoU (mAP 50) for different object detection methods. We evaluate one-stage detectors, two-stage detectors, and transformer-based detectors such as DETR. Figure 6 summarizes the results, illustrating the strengths and weaknesses of each approach in detecting brick kilns.

3.1.2 Cross-Region Generalization Performance

To assess the generalization capability of our models, we conducted out-of-region experiments. Models trained on the Uttar Pradesh dataset were tested on other regions including Punjab (Pakistan) and Dhaka. The results, shown in Figure 7, demonstrate the robustness and limitations of the detection methods when applied to geographically distinct test sets.

3.1.3 Impact of Super-Resolution Techniques

This subsection examines the impact of various super-resolution techniques on the detection pipeline, particularly the YOLOv11L-OBB model. Figure 8 compares the class-agnostic mAP 50 performance when different super-resolution methods are applied, highlighting improvements and trade-offs in detection accuracy.

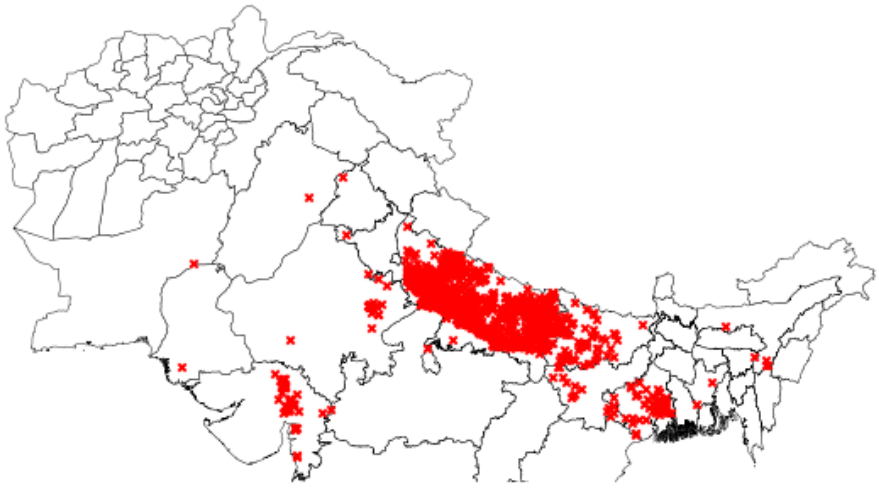

Figure 3: Spatial Distribution of CFCBK Brick Kilns in our dataset.

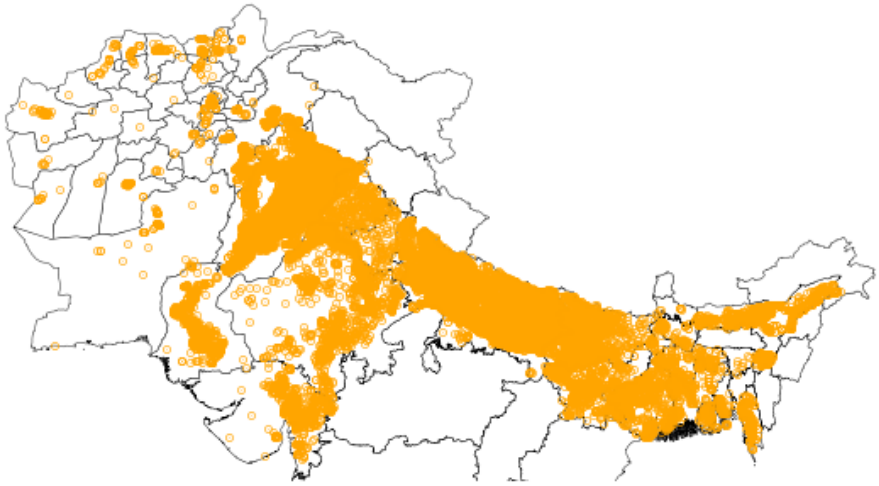

Figure 4: Spatial Distribution of FCBK Brick Kilns in our dataset.

4 Custom Developed Hand Validation Tools

Figure 9 shows a snippet of the custom hand-validation interface developed in-house. This interface overlays bounding box annotations on ESRI Wayback Imagery, highlighting a Zigzag brick kiln.

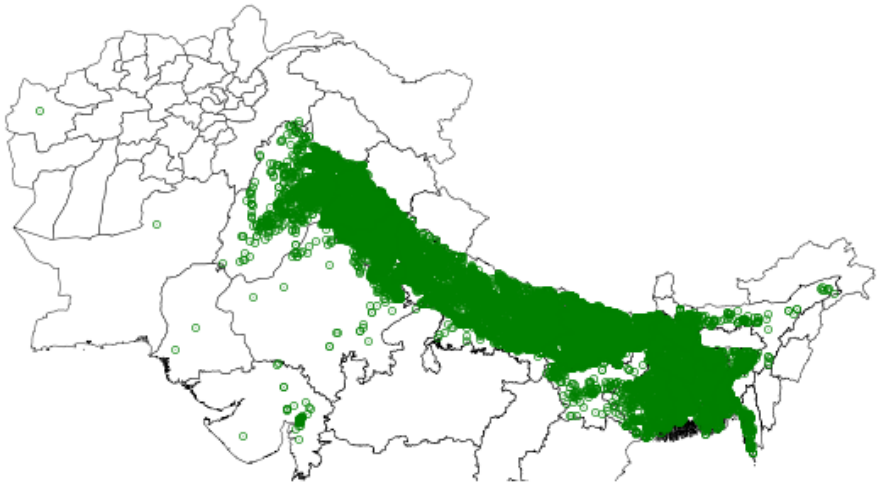

Figure 5: Spatial Distribution of Zigzag Brick Kilns in our dataset.

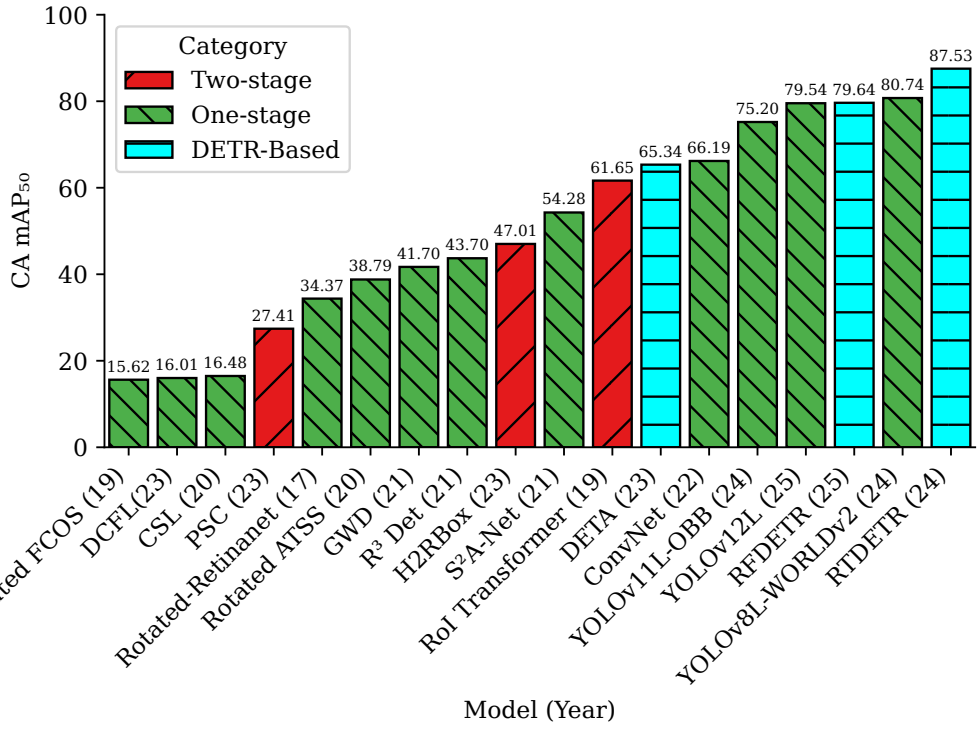

Figure 6: Comparison of class-agnostic mAP 50 for one-stage, two-stage, and transformer-based (DETR) object detection methods.

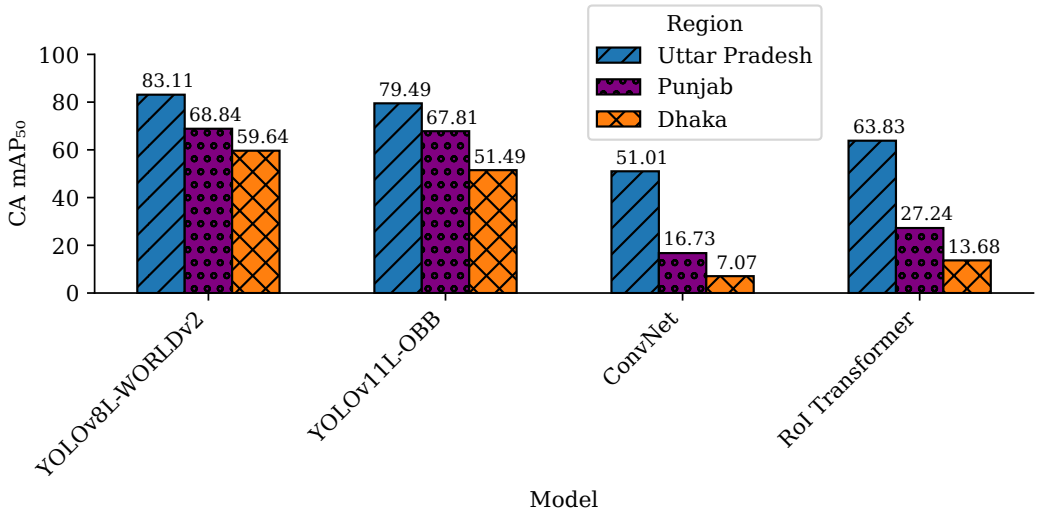

Figure 7: Comparison of class-agnostic mAP 50 for out-of-region performance experiment using the Uttar Pradesh dataset. Results are shown on test sets of Uttar Pradesh, Punjab (Pakistan), and Dhaka using the best-performing methods.

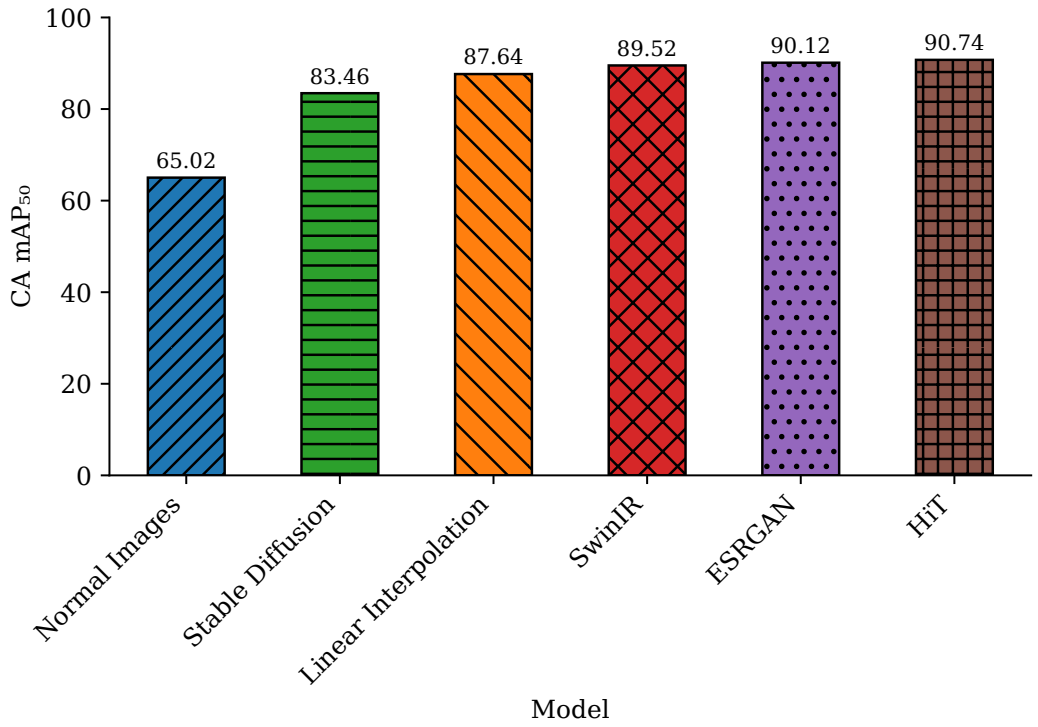

Figure 8: Class-agnostic mAP 50 comparison of different super-resolution methods applied to the YOLOv11L-OBB detection pipeline.

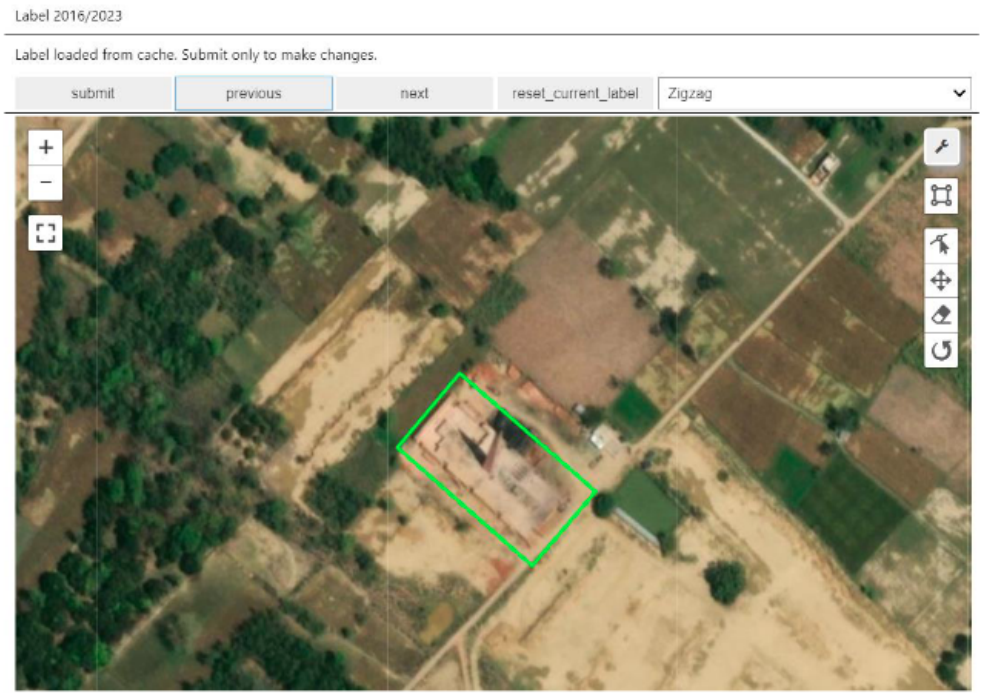

Figure 9: Snippet of the custom hand-validation interface developed in-house. The interface overlays bounding box annotations on ESRI Wayback Imagery, displaying a Zigzag brick kiln.

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[2] Zeel B Patel, Rishabh Mondal, Shataxi Dubey, Suraj Jaiswal, Sarath Guttikunda, and Nipun Batra. Space to policy: Scalable brick kiln detection and automatic compliance monitoring with geospatial data. *arXiv preprint arXiv:2412.04065*, 2024.

[3] Qiusheng Wu. geemap: A python package for interactive mapping with google earth engine. *Journal of Open Source Software*, 5(51):2305, 2020.